Institute of Mathematical Statistics

COLLECTIONS

Volume 00

Optimality

The Third Erich L. Lehmann Symposium

Javier Rojo, Editor

Institute of Mathematical Statistics
Beachwood, Ohio, USA

Institute of Mathematical Statistics
Collections

Series Editor:
Anirban DasGupta

The production of the *Institute of Mathematical Statistics Collections* is managed by the
IMS Office: Rong Chen, Treasurer and
Elyse Gustafson, Executive Director.

Library of Congress Control Number: 0000000000

International Standard Book Number 0-000000-00-0

International Standard Serial Number 0749-2170

Copyright © 2009 Institute of Mathematical Statistics

All rights reserved

Printed in the United States of America

Contents

Preface
Javier Rojo v

Contributors to this volume
. viii

Scientific program
. ix

List of Participants
. xvii

Acknowledgement of referees' services
. xxii

PAPERS

Honoring Erich Leo Lehmann

Erich L. Lehmann, The Lehmann Symposia, and November 20^{th} 1917
Javier Rojo 1

The Honorable Erich L. Lehmann
Stephen Stigler 8

Optimality

Some History of Optimality
Erich L. Lehmann 11

An Optimality Property of Bayes' Test Statistics
Raghu Raj Bahadur and Peter J. Bickel 18

On the Non-Optimality of Optimal Procedures
Peter J. Huber 31

Semi-parametric Inference

Proportional Hazards Regression with Unknown Link Function
Wei Wang, Jane-Ling Wang and Qihua Wang 46

Semiparametric Models and Likelihood - The Power of Ranks
Kjell Doksum and Akichika Ozeki 66

Bootstrap

On Bootstrap Tests of Hypotheses
Wei-Yin Loh and Wei Zheng 93

Nonparametric Inference

Nonparametric Estimation for Lévy Models Based on Discrete-Sampling
José E. Figueroa-López 117

On the Estimation of Symmetric Distributions under Peakedness Order Constraints
Javier Rojo and José Batún-Cutz 147

Functional Data Analysis

A Functional Generalized Linear Model with Curve Selection in Cervical Precancer Diagnosis using Fluorescence Spectroscopy
Hongxiao Zhu and Dennis D. Cox 174

Nonparametric Estimation of Hemodynamic Response Function: A Frequency Domain Approach
Ping Bai, Young Truong and Xuemei Huang 191

Mixed Models, Posterior Means and Penalized Least-Squares
Yolanda Muñoz Maldonado 217

Probability

From Charged Polymers to Random Walk in Random Scenery
Xia Chen and Davar Khoshnevisan 237

Recovery of Distributions via Moments
Robert M. Mnatsakanov and Artak S. Hakobyan 252

Asymptotic Theory

Asymptotic Efficiency of Simple Decisions for the Compound Decision Problem
Eitan Greenshtein and Ya'acov Ritov 266

Large Sample Statistical Inference for Skew-Symmetric Families on the Real Line
Rolando Cavazos–Cadena and Graciela González–Farías 276

Multiple Comparison

Parametric Mixture Models for Estimating the Proportion of True Null Hypotheses and Adaptive Control of FDR
Ajit C. Tamhane and Jiaxiao Shi 304

Bayesian Decision Theory for Multiple Comparisons
Charles Lewis and Dorothy T. Thayer 326

Ad-hoc Networks

The Challenges of Model Objective Selection and Estimation for Ad-hoc Network Data Sets
Farinaz Koushanfar and Davood Shamsi 333

Finance

A Note on the Investment Proportions of a Minimum-Variance Equity Portfolio
Wilhelmine von Turk 347

Preface

The venue for the 3^{rd} Lehmann Symposium was the School of Engineering at Rice University from May 16^{th} through May 19^{th}, 2007. The collection of refereed papers included in this volume represents a selection of the papers submitted for publication. Most of the work was presented at the Symposium but there are some contributions that were submitted by participants who did not present their work during the conference.

All activities of the Symposium, except for a banquet held at the student center, were held in Duncan Hall – home of the Statistics Department. Duncan Hall's floor plan, with its open atrium, its main auditorium, and several conveniently located meeting rooms, allows for, and facilitates, interaction among the participants.

As it has been the tradition of the Symposia, the event opens with a session for young investigators. The purpose of initiating the Symposia in this way is to free the young investigators from this activity, and introduce them to other more senior investigators with the goal that the young investigators may more easily mingle with the group. For the third Lehmann Symposium the four young investigators were Yolanda Muñoz Maldonado, Brisa Sánchez, Farinaz Koushanfar, and José Enrique Figueroa-López. At the end, due to unforeseen circumstances, José Enrique was moved to the probability session. All four young investigators provided motivating talks and three of them submitted their work for this volume. All four have a bright future ahead of them.

It is also the tradition of the Symposia that the young investigators session is immediately followed by the first Plenary session and this spot has always been filled by Erich L. Lehmann. Erich provided a great lecture on the history of optimality. The rest of the program, I hope that the reader will agree with me, was excellent.

The papers presented here cover several areas: some of the works consider classical aspects of the discipline and others deal with contemporary aspects of the theory and applications of statistics. Thus, the reader will find a fascinating section dedicated to the subject of optimality. Lehmann, Bahadur and Bickel, and Huber

provide excellent discussions on various aspects of optimality. Semi-parametric and non-parametric inference, bootstrap tests of hypotheses, functional data analysis, asymptotic theory, ad-hoc networks, and finance are some of the areas represented in the volume. Intentionally, I left probability to the end. It has been a goal of the Symposium to have a probability component. It is felt that the perceived distancing of probability and statistics, even at the level of Ph.D. programs, cannot be healthy. Future Lehmann Symposia will continue to encourage a closer relationship between the two subjects.

The Symposium could not occur without the financial support of several generous contributors. Several institutions have provided support for the series of Symposia. I want, however, to acknowledge the explicit support and work of the individuals within those institutions responsible for securing the funding. Demissie Alemayehu of Pfizer has been a constant and faithful supporter of the Symposia and has been responsible for Pfizer's generous contributions to the 2^{nd} and 3^{rd} Lehmann Symposia. I also want to acknowledge the support of the National Science Foundation. Shulamith Gross, Grace Yang, and Gábor Székely have been instrumental in supporting the 2^{nd} and 3^{rd} Lehmann Symposia. Gary Rosner, from the U.T. MD Anderson Cancer Center, has worked to obtain MD Anderson Cancer support for the last two Symposia. Kathleen O'Hara at the Mathematical Sciences Research Institute (MSRI) and James O. Berger at the Statistics and Applied Mathematics Sciences Institute (SAMSI) provided financial support for the 3^{rd} Symposium. MSRI supported a proposal to hold the event at their facilities but plans changed. You can read the details in the first article of the volume. The efforts of Robert Hardy of the University of Texas School of Public Health, and Rudy Guerra of the Gulf Coast Consortia, are also gratefully acknowledged. Victor Pérez-Abreu, of the Centro de Investigación en Matemáticas (CIMAT) provided support for the 1^{st} and 2^{nd} Symposia.

I also want to give special recognition to my student Tuan S. Nguyen for his indefatigable efforts and continued help in the preparation of this volume.

Last, but not least, I want to acknowledge the support of Rice University through

the School of Engineering and its Department of Statistics for allowing me to engage in these activities.

Javier Rojo
Statistics Department
Rice University
Houston, Texas

January 30, 2009

Contributors to this volume

Bahadur, R. R., *University of Chicago*
Bai, P., *University of North Carolina at Chapel Hill*
Batún–Cutz, J., *Universidad Autónoma de Yucatán, Mérida, México*
Bickel, P. J., *University of California, Berkeley*

Cavazos–Cadena, R., *Universidad Autónoma Agraria Antonio Narro*
Chen, X., *University of Tennessee*
Cox, D. D., *Rice University*

Doksum, K., *University of Wisconsin, Madison*

Figueroa-López, J., *Purdue University*

González–Farías, G., *Centro de Investigación en Matemáticas A. C.*
Greenshtein, E., *Duke University*

Hakobyan, A. S., *West Virginia University*
Huang, X., *University of North Carolina at Chapel Hill*
Huber, P. J., *Switzerland*

Khoshnevisan, D., *University of Utah*
Koushanfar, F., *Rice University*

Lehmann, E. L., *University of California, Berkeley*
Lewis, C., *Fordham University*
Loh, W., *University of Wisconsin, Madison*

Mnatsakanov, R. M., *West Virginia University*
Muñoz Maldonado, Y., *Michigan Technological University*

Ozeki, A., *University of Wisconsin, Madison*

Ritov, Y., *Jerusalem, Israel*
Rojo, J., *Rice University*

Shamsi, D., *Rice University*
Shi, J., *Northwestern University*
Stigler, S., *University of Chicago*

Tamhane, A. C., *Northwestern University*
Thayer, D. T., *Educational Testing Service*
Truong, Y., *University of North Carolina at Chapel Hill*

Von Turk, W., *American Century Investmens*

Wang, J., *University of California, Davis*
Wang, Q., *Chinese Academy of Science and The University of Hong Kong*
Wang, W., *Harvard Medical School and Brigham and Women's Hospital*

Zheng, W., *University of Wisconsin, Madison*
Zhu, H., *Rice University*

SCIENTIFIC PROGRAM

The Third Erich L. Lehmann Symposium
May 16 - 19, 2007
Rice University

Symposium Chair and Organizer

Javier Rojo
Statistics Department, MS-138
Rice University
6100 Main Street
Houston, TX 77005

Plenary Speakers

Erich L. Lehmann University of California, Berkeley	*Some history of optimality*
Lawrence D. Brown The Wharton School University of Pennsylvania	*A unified view of regression, shrinkage, empirical Bayes, hierarchical Bayes, and random effects*
James O. Berger Duke University	*Some recent developments in Bayesian model selection*
Rodrigo Bañuelos Purdue University	*Isoperimetric bounds for Lévy processes*
Peter J. Bickel University of California, Berkeley	*The collapse of particle filters*
Stephen M. Stigler University of Chicago	*Karl Pearson and testing statistical hypotheses*
Peter J. Huber	*On the non-optimality of optimal procedures*
Willem R. van Zwet University of Leiden	*Statistics and the law: the case of the nonchalant nurse*

Invited Session Scientific Committee

Javier Rojo, Chair	**Rice University**
Jane Ling Wang	**University of California, Davis**
Rudy Guerra	**Rice University**
Juliet P. Shaffer	**University of California, Berkeley**
Wei-Yin Loh	**University of Wisconsin, Madison**
Peter J. Bickel	**University of California, Berkeley**
Kjell Doskum	**University of Wisconsin, Madison**
Yongzhao Shao	**New York University**
Demissie Alemayehu	**Pfizer and Columbia University**
James O. Berger	**Duke University and SAMSI**

Invited Sessions

Young Investigators

Javier Rojo, Organizer
Yolanda Muñoz Maldonado, Chair

Brisa N. Sánchez University of Michigan	*Residual-based diagnostics for structural equation models*
Yolanda Muñoz Maldonado UT Health Sc Center Houston	*Penalized least squares and frequentist and bayesian mixed-effects models*
Farinaz Koushanfar Rice University	*How challenging is the data set?*

Statistical Problems in the Analysis of Genomic and Magnetic Resonance Imaging Data

Wei-Yin Loh, Chair

Sunduz Keles University of Wisconsin, Madison	*Statistical issues arising in the study of transcription regulation*
Shaw-Hwa Lo Columbia University	*A method toward mapping of common traits*

Young K. Truong — *Spatio-temporal modeling for fMRI data*
UNC, Chapel Hill

Optimality in Bioinformatics: Theory vs Practice

Rudy Guerra, Chair

David Dahl — *Simultaneous inference for multiple testing and clustering via Dirichlet process mixture models*
Texas A&M University

Chad Shaw — *Using annotations in the analysis of genome scale data*
Baylor College of Medicine

Rudy Guerra — *Incorporating biological knowledge in gene expression analysis*
Rice University

Regularized Methods of Classification and Estimation of Nonparametric Regression and Covariance Matrices When Data is High Dimensional

Peter J. Bickel, Chair

Wei-Yin Loh — *Regression and variable selection in large p, small n problems*
University of Wisconsin

Debashis Paul — *Principal component analysis for structured high-dimensional data*
University of California, Davis

Ya'acov Ritov — *Consistent learning methods are approximately local*
Hebrew University

Probability, Levy Process, and Applications

Javier Rojo, Organizer
Rodrigo Bañuelos, Chair

Dennis Cox — *Multiscale models for chemical reaction processes*
Rice University

Davar Khoshnevisan — *On some applications of stable processes*
University of Utah

José Enrique Figueroa — *Non-parametric estimation for some models driven by Levy processes*
Purdue University

Multiplicity: Developments and Current Issues

Juliet P. Shaffer, Chair

Charles Lewis Fordham University	*Bayesian decision theory for multiple comparisons*
Helmut Finner Deutsches Diabetes-Zentrum	*Testing for equivalence in k sample models*
Ajit C. Tamhane Northwestern University	*A mixture model approach to estimating the number of true null hypotheses and adaptive control of FDR*

Recent Advances in Non- and Semi-parametric Modeling

Jane-Ling Wang, Chair

Kjell Doksum University of Wisconsin	*Semi-parametric models based on transformations and extremes*
Xihong Lin Harvard University	*Statistical challenges in analyzing mass spectrometry proteomic data*
Naisyin Wang Texas A&M University	*Analysis of hierarchical biomedical data using semiparametric models*

Statistical Inference for Population Substructures via Clustering, Mixture Models and Other Approaches

Demissie Alemayehu, **Yongzhao Shao**, Organizers
Yongzhao Shao, Chair

Bruce G. Lindsay Pennsylvania State University	*Modal inference: halfway between clustering and mixture analyses*
Yuewu Xu Fordham University	*A limit theory for likelihood ratio test under unidentifiability for general dependent processes*
Yongzhao Shao New York University	*Recent developments in likelihood theory with applications to testing homogeneity in finite mixture models and other models*

Semiparametric Models, Longitudinal Survival Data, False Discovery Rate, And Brain fMRI

Kjell Doksum, Chair

Jane-Ling Wang University of California at Davis	*Semi-parametric analysis of longitudinal data truncated by event-time*
Kam-Wah Tsui and Shijie Tang University of Wisconsin, Madison	*Simultaneous testing of multiple hypotheses using generalized p-values*
Chunming Zhang University of Wisconsin, Madison	*Semi-parametric detection of significant activation for brain fMRI*

Multiple Testing and Subgroup Analysis

James O. Berger, Chair

Juliet P. Shaffer University of California	*Multiplicity and subgroup analysis*
Peter Mueller U.T. M.D. Anderson Cancer Center	*The optimal discovery procedure and Bayesian decision rules*
M. J. Bayarri Valencia, Duke U. and SAMSI	*Bayesian and frequentist handling of multiple U. testing*

CONTRIBUTED PAPERS

Nancy L. Glenn, University of South Carolina, Columbia: *The GEM Algorithm*

Richard C. Ott, Mesa State College: *On the Operating Characteristics of Some Non-parametric Methodologies for the Classification of Distributions by Tail Behavior*

Xiaohui Wang, University of Texas-Pan American: *Classifications of Proteomic Mass Spectra and Other Curve Data*

Xiaohu Li, School of Mathematics and Statistics, Lanzhou University, People's Republic of China: *Stochastic Comparison on Conditional Order Statistics - Some New Results*

Robert Mnatsakanov, West Virginia University, *Some Asymptotic Properties of Varying Kernel Density Estimator*

Monnie McGee, Southern Methodist University, *A Distribution Free Summarization Method for Affymetrix GeneChip® Arrays*

Changxiang Rui, University of Arkansas, *Point and Block Prediction in Log-Gaussian random Fields: The Non-constant Mean Case*

Qiang Zhao, Texas State University, San Marcos, *Survival Analysis of Microarray Gene Expression Data Using Correlation Principal Component Regression*

Suhasini Subba Rao, Texas A&M University, *Normalised Least-squares estimation in time-varying arch models*

Santanu Chakraborty, UT Pan-American, *Parametric Inference on Zero-Inflated Poisson distribution and its variants*

Hongxiao Zhu, Rice University*, A Functional Generalized Linear Model with application to Cervical Pre-cancer Diagnosis using Fluorescence spectroscopy*

Xiaowei Wu, Rice University, *Some Estimation and Hypothesis Testing Problems in Fluctuation Analysis*

John Fresen, University of Missouri - Columbia, *On the Definition of Weak Convergence of a Sequence of Random Elements*

Victor De Oliveira, UT San Antonio*, Objective Bayesian Analysis of Spatial Data with Measurement Error*

Pang Du, Virginia Tech, *Smoothing spline frailty model*

Graciela Gonzalez, CIMAT, *Some important issues in inference under certain types of singularities*

The Third Erich L Lehmann Symposium
Javier Rojo, Organizer/Chair
Invited Program – All talks are in McMurtry Auditorium Duncan Hall 1055

	Wednesday May 16	Thursday May 17	Friday May 18	Saturday May 19
7:30	Breakfast 7:30 – 8:45 Duncan Hall	Breakfast Duncan Hall	Breakfast Duncan Hall	Breakfast Duncan Hall
8:45 9:00	Opening Remarks	Classification, nonparametric regression: high dimensional data	Recent Advances in Non and Semiparametric Modeling	Multiple Testing and Subgroup analysis
	Plenary Speaker **Erich L Lehmann**	Wei-Yin Loh Debashis Paul Ya'acov Ritov	Kjell Doksum Xihong Lin Naisyin Wang	Juliet P Shaffer Peter Mueller Susie Bayarri
10:00	Young Investigators			
10:30 – 10:50	Brisa N. Sánchez	Coffee Break 10:30 – 10:50	Coffee Break	Coffee Break
	Yolanda Muñoz Maldonado	**Plenary Speaker** 10:50	**Plenary Speaker** 10:50	**Plenary Speaker** 10:50
	Farinaz Koushanfar	**Rodrigo Bañuelos**	**Lawrence D Brown**	**Peter J Huber**
	Lunch 12:00 – 1:15 Duncan Hall	Lunch 11:50 – 1:15 Duncan Hall	Lunch 11:50 – 1:15 Duncan Hall	**Plenary Speaker 11:50** **Willem van Zwet**
1:15 – 3:00	Genomics and magnetic resonance	Probability, Lévy Processes and Applications	Inference for Substructures: Clustering, Mixtures, …	**CLOSING LUNCHEON DUNCAN HALL**
	Sunduz Keles	Dennis D. Cox	Bruce G. Lindsay	
	Shaw-Hwa Lo	Davar Khoshnevisan	Yuewu Xu	**1:00 – 2:30**
	Young K. Truong	Jose Enrique Figueroa	Yongzhao Shao	
3:00 – 3:20	Coffee Break	Coffee Break	Coffee Break	
3:20 4:20	**Plenary Speaker** **Peter J Bickel**	**Plenary Speaker** **Stephen M Stigler**	**Plenary Speaker** **James O Berger**	
4:20 – 6:05	Optimality in Bioinformatics: Theory vs Practice	Multiplicity: Developments and current issues	Semiparametrics, longitudinal survival, false discovery, fMRI	
	David B. Dahl	Charles Lewis	Jane-Ling Wang	
	Chad Shaw	Helmut Finner	Kam-Wah Tsui	
	Rudy Guerra	Ajit C. Tamhane	Chunming Zhang	
	CONTRIBUTED TALKS 6:05 – 7:05	Banquet Ballroom Ley Student Cr Cash Bar/Music 6:45 – 7:30 Dinner/Music 7:30 – 10:00	CONTRIBUTED TALKS 6:05 – 7:05	


Third Lehmann Symposium
May 16th-19th, 2007
Rice University, Houston Texas

List of Participants

Susan Alber
Department of Biostatistics
1515 Holcomber Blvd
Houston, TX 77030
salber@mdanderson.org

Veera Baladandayuthapani
1515 Holcombe Boulevard
Unit 447
Houston, TX 77030
veera@mdanderson.org

Rodrigo Bañuelos
Department of Mathematics
Purdue University
West Lafayette, IN 47906
banuelos@math.purdue.edu

Rosa Bañuelos
2630 Bissonnet St
Apt 167
Houston, TX 77005
banuelos@rice.edu

M.J. Bayarri
SAMSI
Research Triangle Park
NC 27709
susie.bayarri@uv.es

Dror Berel
4045 linkwood dr. #605
Houston, TX 77025
dror.berel@gmail.com

James Berger
PO Box 14006
Research Triangle Park
NC 27709
berger@samsi.info

Nancy Bickel
Statistics 367 Evans Hall
UC Berkeley
Berkeley, CA 94720
bickel@stat.berkeley.edu

Peter Bickel
Statistics 367 Evans Hall
UC Berkeley
Berkeley, CA 94720
bickel@stat.berkeley.edu

Lawrence D Brown
Statistics Dept, 400 Huntsman Hall
3730 Walnut St
Philadelphia, PA 19104
lbrown@wharton.upenn.edu

William F Bryant
6100 Main St.
Houston, TX 77005
wfb1@rice.edu

Christopher P Calderon
CAAM MS#138
Rice University
Houston, TX 77251
calderon@rice.edu

Santanu Chakraborty
Department of Mathematics
University of TX - Panamerican
Edinburg, TX 78539
schakraborty@utpa.edu

Jamie Chatman
7200 Almeda Rd, #818
Houston, TX 77054
jchatman@rice.edu

Dunlei Cheng
8080 North Central Expressway
Suite 500
Dallas, TX 75206
dunleic@baylorhealth.edu

Raj Chikara
2700 Bay Area Blvd.
Houston, TX 77058
chhikara@uhcl.edu

Dennis D Cox
Rice University MS-138
6000 Main St.
Houston, TX 77005
dcox@rice.edu

David B Dahl
Department of Statistics
3143 TAMU
College Station, TX 77843
dahl@stat.tamu.edu

Sujay Datta
Department of Statistics
3143 TAMU
College Station, TX 77843
sdatta@stat.tamu.edu

Victor De Oliveira
Management Sci Statistics
University of Texas
San Antonio, TX 78249
victor.deoliveira@utsa.edu

Kjell A Doksum
Statistics Department
University of Wisconsin
Madison, WI 53706
doksum@stat.wisc.edu

Pang Du
13000 Foxridge Ln
Apt J
Blacksburg, VA 24060
pangdu@vt.edu

Katherine Ensor
Rice University
6100 Main St.
Houston, TX 77005
ensor@rice.edu

Zhide Fang
Department of Mathematics
University of New Orleans
New Orleans, LA 70148
zfang@uno.edu

Jose Enrique Figueroa-Lopez
Department of Mathematics
Purdue University
West Lafayette, IN 47906
figueroa@stat.purdue.edu

Helmut Finner
German Diabetes Center
Duesseldorf, D-40225
DE
finner@ddz.uni-duesseldorf.de

Mujyanama B Frank
P.O BOX 117 Butare Rwanda
Butare, 250
RW
gifrank2003@yahoo.com

Daniel Fresen
Department of Mathematics
University of Missouri
Columbia, MO 65211
fresenj@MO.edu

John L Fresen
Department of Statistics
University of Missouri
Columbia, MO 65211
fresenj@MO.edu

Nancy L Glenn
1148 Rockwood Road
Columbia, SC 29209
nglenn@math.sc.edu

Graciela Gonzalez-Farias
Jalisco s/n
Mineral de Valenciana
Guanajuato, GTO 36240
farias@cimat.mx

Violeta Hennessey
1515 Holcombe Boulevard
Houston, TX 77030
violeta.g.hennessey@uth.hmc.edu

Christian Houdre
School of Mathematics
GA Institute of Technology
Atlanta, GA 30332
houdre@math.gatech.edu

Jianhua Hu
U.T. MD Anderson Cancer Center
1515 Holcombe Blvd
Houston, TX 77030
jhu@mdanderson.org

Jianhua Huang
Blocker Building
TX A&M University
xlhuang@mdanderson.org

Peter Huber
P. O. Box 198
CH-7250 Klosters
Klosters, CH
peterj.huber@bluewin.ch

Adarsh Joshi
3902 College Main
Apt 1411
Bryan, TX
adarsh@stat.tamu.edu

David Kahle
4102 Ascot Lane
Houston, TX 77092
dkahle@rice.edu

Yuliya Karpievitch
U.T. MD Anderson Cancer Center
1515 Holcombe Blvd
Houston, TX 77030
ykarpi@mdanderson.org

Demetrios Kazakos
3100 Cleburne St.
Houston, TX 77004
kazakosd@tsu.edu

Jerome Paul Keating
One UTSA Circle
The University of Texas
San Antonio, TX 78249
jerome.keating@utsa.edu

Sunduz Keles
Department of Statistics
University of Madison
Madison, WI 53705
keles@stat.wisc.edu

Davar Khoshnevisan
University of Utah
Department of Mathematics
Salt Lake City, UT 84112
davar@math.UT.edu

Farinaz Koushanfar
6100 S Main
Rice ECE Dept. MS 366
Houston, TX 77005
farinaz@rice.edu

Kalimuthu Krishnamoorthy
Dept of Mathematics
Louisiana State University
Lafayette, LA 70504
krishna@LA.edu

S.N. Lahiri
Statistics Department
TX A&M University
College Station, TX 77843
snlahiri@stat.tamu.edu

Mike Lecocke
14818 Personality
San Antonio, TX 78248
mlecocke@stmarytx.edu

J. Jack Lee
U.T. MD Anderson Cancer Center
1515 Holcombe Blvd.
Houston, TX 77030
jjlee@mdanderson.org

Jong Soo Lee
6315 Forbes Ave Apt 905
Pittsburgh, PA 15217
jslee@stat.cmu.edu

Erich L Lehmann
University of California
Department of Statistics
367 Evans Hall # 3860
Berkeley, CA 94720
shaffer@stat.berkeley.edu

Charles Lewis
Fordham University
441 East Fordham Road
Bronx, NY 10458
clewis@fordham.edu

Hua Li
7900 Cambridge ST 16-1E
Houston, TX 77054
hua.li@uth.tmc.edu

Xiaohu Li
6061 De Zavala RD
Apt 1016
San Antonio, TX 78249
mathxhli@hotmail.com

Xihong Lin
Statistics Department
Harvard University
655 Huntington Avenue
Building 2, Office 419
Boston, MA 2115
xlin@hsph.harvard.edu

Bruce George Lindsay
422 Thomas Building
University Park, PA 16802
bgl@psu.edu

Peng Liu
Department of Statistics
Iowa State University
Ames, IA 50011
PLIU@iastate.edu

Shaw-Hwa Lo
Statistics Department
Columbia University
New York, NY 10027
slo@stat.columbia.edu

Wei-Yin Loh
Department of Statistics
University of Wisconsin
Madison, WI 53706
loh@stat.wisc.edu

Joseph F Lucke
Center for Clinical Research
& Evidence-based Medicine
UTHSC Medical School
Houston, TX 77030
Joseph.F.Lucke@uth.tmc.edu

Denka Grudeva Markova
1806 S. 8th Street Apt 252
Waco, TX 76706
denka_markova@baylor.edu

Matthias Mathaes
Department of Statistics
Rice University
Houston, TX 77005
matze@rice.edu

Monnie McGee
Department of Statistics
3225 Daniel Ave
P.O. Box 750332
Dallas, TX 75275
mmcgee@smu.edu

Robert Musheg Mnatsakanov
P.O. Box 6330
Morgantown, WV 26506
rmnatsak@stat.wvu.edu

MaryAnn Morgan-Cox
One Bear Place #97140
Waco, TX 76798
MaryAnn_Morgan-Cox@baylor.edu

Peter Mueller
U.T. MD Anderson Cancer Center
1515 Holcombe Blvd
Houston, TX 77030
pm@wotan.mdacc.tmc.edu

Yolanda Muñoz Maldonado
Mathematics Department
Michigan Technological University
1400 Townsend Drive
Houghton, MI 49931
ymunoz@mtu.edu

Michael Naaman
7315 Brompton St.
APT. 368B
Houston, TX 77025
mikenaaman@rice.edu

Tuan S Nguyen
Department of Statistics
Rice University
6100 Main St
Houston, TX 77005
kcstevone@gmail.com

Linda Njoh
1806 S. 8th street
Apt 252
Waco, TX 76706
linda_njoh@baylor.edu

Richard C Ott
3505 North 12th St. Apt #D8
Grand Junction, CO 81506
rott@mesastate.edu

Haiying Pang
Department of Biostatistics
MD Anderson Cancer Center
Houston, TX 77030
haiying.pang@uth.tmc.edu

Galen Papkov
Department of Statistics
Rice University
Houston, TX 77005
gpapkov@rice.edu

Debashis Paul
Department of Statistics
University of California
Davis, CA 95616
debashis@wald.ucdavis.edu

Claudia Pedroza
U. T. School of Public Health
RAS E831
Houston, TX 77030
claudia.pedroza@uth.tmc.edu

Victor Pérez-Abreu
CIMAT
Apdo. Postal 402
Guanajuato, Gto. 36000
MX
pabreu@cimat.mx

Kenneth Pietz
Michael E. DeBakey
VA Medical Center (152)
2002 Holcombe Blvd.
Houston, TX 77030
kpietz@houston.rr.com

Darwin Poritz
968 Southern Pass Ct
Houston, TX 77062
darwin.h.poritz@nasa.gov

Jessica Pruszynski
2425 S. 21st St.
Apt. 104
Waco, TX 76706
jessica_pruszynski@baylor.edu

Lindsay Renfro
18 Barksdale Ave.
Waco, TX 76705
Lindsay_Renfro@baylor.edu

Peter Richardson
5213 Shady Maple Dr.
Kingwood, TX
peterr@bcm.tmc.edu

Rudolf H Riedi
Department of Statistics
Rice University
6100 Main St
Houston, TX 77005
riedi@rice.edu

Kristin Ritz
5440 N. Braeswood Blvd.
#956
Houston, TX 77096
kristinr@rice.edu

Javier Rojo
Statistics Department
Rice University
Houston, TX 77005
jrojo@rice.edu

Gary Rosner
Department of Biostatistics
UT MD Anderson Cancer
Center
1515 Holcombe Boulevard
Houston, TX 77030
glrosner@mdanderson.org

Changxiang Rui
7019 River Elms
San Antonio, TX
cxrui@uark.edu

Brisa N Sanchez
1420 Washington Heights
Ann Arbor, MI 48109
brisa@umich.edu

David W Scott
Department of Statistics
Rice University
Houston, TX 77005
scottdw@stat.rice.edu

John Seaman
121 Neely Rd
Hewitt, TX
john_w_seaman@baylor.edu

Juliet P Shaffer
University of California
Dept. of Statistics
367 Evans Hall # 3860
Berkeley, CA 94704
shaffer@stat.berkeley.edu

Yongzhao Shao
650 First Avenue
Fifth Floor, #538
New York, NY 10016
shaoy01@med.nyu.edu

Chad A Shaw
1 Baylor Plaza
Houston, TX 77030
cashaw@bcm.tmc.edu

Yu Shen
Department of Biostatistics
UT MD Anderson Cancer Center
Houston, TX 77030
yshen@mdanderson.org

Stephen Stigler
Statistics Dept
University of Chicago
Chicago, IL 60637
stigler@galton.uchicago.edu

Suhasini Subba Rao
Department of Statistics
TX A&M University
College Station, TX 77843
suhasini@stat.tamu.edu

Gabor J Szekely
The National Science Foundation
4201 Wilson Blvd
Arlington, VA 22201
gszekely@nsf.gov

Ajit C Tamhane
Department of Statistics
2008 Sheridan Road
Evanston, IL 60208
ajit@iems.northwestern.edu

Young K Truong
Department of Biostatistics
The University of North Carolina
Chapel Hill, NC 27599
truong@bios.unc.edu

Kam-Wah Tsui
Department of Statistics
University of Madison
Madison, WI 53706
kwtsui@stat.wisc.edu

Willem R van Zwet
Dept. of Mathematics
University of Leiden
P.O. Box 9512
Leiden 2341CB
vanzwet@math.leidenuniv.NL

Marco Vilela
Department of Biostatistics
UT MD Anderson Cancer Center
Houston, TX 77030
mvilela@mdanderson.org

Wilhelmine von Turk
43 Del Mar Avenue
Berkeley, CA 94708
wvonturk@sbcglobal.net

Jane-Ling Wang
Department of Statistics
University of California
Davis, CA 95616
wang@wald.ucdavis.edu

Naisyin Wang
Department of Statistics
TX A&M University
College Station, TX
77843
nwang@stat.tamu.edu

Xiaohui Wang
7715 N 4th Court
McAllen, TX 78504
xhwang@utpa.edu

Feifei Wei
8170 33rd Ave. S.
MS#21111R
Bloomington, MN 55425
feifei.wei@healthpartners.com

Talithia Williams
301 Platt Boulevard
Claremont, CA 91711
williams@math.hmc.edu

Xiaowei Wu
Department of Statistics
Rice University
Houston, TX 77005
xwwu@stat.rice.edu

Yuewu Xu
Fordham School of Business
113 West 60th Street
New York, NY 10023
yuxu@fordham.edu

Ya'acov Ritov
10 Rabi Benyamin
Jerusalem 96306
IL
yaacov@mscc.huji.ac.il

Jose-Miguel Yamal
5726 Portal Dr.
Houston, TX 77096
jmy@odin.mdacc.tmc.edu

Po Yang
Dept. of Mathematical Sciences
DePaul University
Chicago, IL 60614
pyang3@depaul.edu

Chunming Zhang
1300 University Avenue
University of Wisconsin
Madison, WI 53706
cmzhang@stat.wisc.edu

Qiang Zhao
601 University Drive
San Marcos, TX 78666
qiang.zhao@txstate.edu

Hongxiao Zhu
Department of Statistics
Rice University
Houston, TX 77005
hxzhu@stat.rice.edu

Acknowledgement of referees' services

The efforts from the following referees are gratefully acknowledged

Miguel Arcones
University of Binghamton
Binghamton, NY

Moulinath Banerjee
University of Michigan
Ann Arbor, MI

Camelia Bejan
Rice University
Houston, TX

Prabir Burman
University of California
Davis, CA

Patrick S Carmack
University of Texas
Southwestern Medical Ctr

Rolando Cavazos
Universidad Autónoma
Agraria Antonio Narro

William F. Christensen
Brigham Young University
Provo, UT

Dennis D Cox
Rice University
Houston, TX

Kjell Doksum
University of Wisconsin
Madison, WI

José Enrique Figueroa
Purdue University
West Lafayette, IN

Helmut Finner
Heinrich-Heine-University
Düsseldorf, DE

Chong Gu
Purdue University
Lafayette, IN

Rudy Guerra
Rice University
Houston, TX

Graciela González
CIMAT
México

Daniel Hernández
CIMAT
México

Jianhua Huang
Texas A&M University
College Station, TX

Ruben Juárez
University of Hawaii
Manoa, HI

Tomasz Kozubowski
University of Nevada
Reno, NV

Erich L. Lehmann
University of California
Berkeley, CA

Xiaoyan Leng
Wake Forest University
Winston-Salem, NC

Ming-Ying Leung
University of Texas
El Paso, TX

Regina Liu
Rutgers University
Piscataway, NJ

Wei-Yin Loh
University of Wisconsin
Madison, WI

Robert M. Mnatsakanov.
West Virginia University
Morgantown, WV

William Navidi
Colorado School of Mines
Golden, CO

Anna Panorska
University of Nevada
Reno, NV

Arthur Pewsey
University of Extremadura
Spain

Peihua Qiu
University of Wisconsin
Madison, WI

Frits Ruymgaart
Texas Tech University
Lubbock, TX

Sanat Sarkar (2)
Temple University
Philadelphia, PA

Juliet P. Shaffer
University of California
Berkeley

Yongzhao Shao
Iowa State University
Ames. IA

Stephen Stigler
University of Chicago
Chicago, IL

Young Troung
Univ of North Carolina
Chapel Hill, NC

Alex Tsodikov
University of Michigan
Ann Arbor, MI

Zhiliang Ying
Columbia University
New York, NY

Stephan R. Sain
NCAR
Boulder, CO

Jane-Ling Wang
University of California
Davis

Cun-Hui Zhang
Rutgers University
Piscataway, NJ

Nicola Sartori
University of Padova
Italy

Guosheng Yin
MD Anderson Cancer Ctr
Houston, TX

Hongyu Zhao
Yale University
New Haven, CT

IMS Collections
Vol. 0 (2009) 1–7
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000021

Erich L. Lehmann, The Lehmann Symposia, and November 20^{th} 1917

Javier Rojo[1]

Rice University

The Lehmann Symposia originated as a result of a conversation I had in the year 2001 with the, then, Director of the Centro de Investigación en Matemáticas (CIMAT), Victor Pérez-Abreu. We both felt that there was an urgent need to bring back into focus theoretical statistics and our proposed solution was a series of Symposia that could serve as a forum for some of the exciting theoretical work being done in statistics. The First Lehmann Symposium took place at CIMAT in May of 2002. Most of the participants were Mexican colleagues. The program can be seen at the site http://www.stat.rice.edu/lehmann/1st-Lehmann.html. The second Lehmann Symposium – http://www.stat.rice.edu/lehmann/ – was held in May of 2004 at the School of Engineering at Rice University. Initially, the venues for the Symposia would alternate between CIMAT and Rice University. However, for various reasons, some being financial, it was decided to hold the 3^{rd} Lehmann Symposium in the United States.

The original plans for the Third Lehmann Symposium were to hold the symposium at the Mathematical Sciences Research Institute (MSRI) in Berkeley during the month of November of 2007. The Third Symposium, however, ended up being held at Rice University for a second time during May of 2007. See http://www.stat.rice.edu/~jrojo/3rd-Lehmann/. I co-edited webcasts of the Second and Third Symposia, and these webcasts are freely available to the public. They can be found at the following sites:
http://webcast.rice.edu/webcast.php?action=details&event=408 — second symposium, and
http://webcast.rice.edu/webcast.php?action=details&event=1057 — third symposium.

But why was the venue for the Third Symposium changed from California back to Texas, and why was the date changed from November 20^{th}, 2007 to May 16^{th}, 2007? There were very good reasons for holding the opening of the Symposium on November 20^{th}, 2007. For example, November 20^{th}, 2007 was the silver anniversary of the greatest big game of all time. See, for example: http://www.alumni.berkeley.edu/KCAA_Multimedia/The_Play_1982.asp. Another good reason to start the Symposium on November 20^{th} was to co-celebrate, with our Mexican counterparts, the start of the first major 20^{th} century revolution. The Mexican revolution started on November 20^{th}, 1910 to remove the dictator Porfirio Díaz who had remained in power for 30 years. This revolution led to the Constitution of 1917 and the start of the Partido Revolucionario Institucional that held power until 2000 when a candidate from the Partido Acción Nacional, Vicente Fox, won the Presidential election. Francisco I. Madero, with the help of Francisco Villa, took over from Porfirio Díaz.

[1]Department of Statistics, MS-138; Rice University; 6100 Main Street; Houston, TX 77005;



Francisco Villa Porfirio Diaz Francisco I. Madero

The left and right images are taken from the Wikimedia Commons, and are available from the United States Library of Congress's Prints and Photographs Division under digital records ID npcc-19554 and ggbain-01887, respectively. The middle image is in the public domain and it was taken from Project Gutenberg, and is available from Enock (1912).

The Constitution of 1917, signed in February 5^{th} of that year, is the current Mexican Constitution. The year of 1917, as it turned out, was a very significant year in the history of the world.

Some of the Significant Events of 1917

February 5 - The Mexican Constitution is adopted by the Mexican Congress in the City of Queretaro, and this Constitution continues to be the Mexican Magna Carta.

February 7 - The United States ends the search for Pancho Villa. The Mexican expedition, as it is known, was led by General Pershing and had the objective of capturing Francisco (Pancho) Villa who had crossed over and ransacked Columbus, New Mexico. Pershing was unable to capture Villa, and after several encounters with various groups, including the Mexican army, Pershing and his forces returned to the United States in February 1917.

February 24 - President Woodrow Wilson is presented with a deciphered German telegram – the Zimmermann telegram, in which Germany offers to give American territory back to Mexico, if Mexico will declare war on the United States.

March 31 - The United States takes possession of the Virgin Islands after paying $25 million to Denmark.

April 6 - As a result of the deciphering of the Zimmermann telegram, the United States Congress declares war on Germany.

April 11 - Babe Ruth pitches for the Boston Red Sox and beats the New York Yankees 10-3, allowing 3 hits. On January 5th, 1920, Ruth was traded to the New York Yankees and thus started the Bambino Curse for the Red Sox. The curse was finally broken 84 years later when the Sox defeated the Cardinals in the 2004 World Series.

April 16 - Lenin arrives in Petrograd.

email: jrojo@rice.edu

May 17 - The paper "On the distribution of the correlation coefficient in small sample. Appendix I to the papers of "Student" and R. A. Fisher. A cooperative study", by H.E. Soper, A.W. Young, B.M. Cave, A. Lee and K. Pearson, (Biometrika 1917 11: 328-413; doi:10.1093/biomet/11.4.328), and the paper "I. Tables for estimating the Probability that the Mean of a unique Sample of Observations lies between -∞ and any given Distance of the Mean of the Population from which the Sample is drawn" by "Student" (Biometrika 1917 11: 414-417; doi:10.1093/biomet/11.4.414) are published. The former would include a criticism of Fisher's maximum likelihood principle that helped ignite a feud between Fisher and Pearson.

May 18 - The Selective Service Act passes the U.S. Congress giving the President the power of conscription.

July 4 - Petrograd Street demonstration - The Bolshevik revolution looms in the horizon.



Demonstrators at the crossroads of Nevsky Prospect and Sadovay Street in Petrograd fled as government troops fired into the crowd. The photograph was taken by Ayn Rand and is available at http://www.noblesoul.com/orc/bio/pics1.html.

November 7 - Bolshevik Revolution begins: The workers of St. Petersburg in Russia, led by the Bolsheviks and the Bolshevik leader Vladimir Lenin, attacked the ineffective Kerensky Provisional Government.

November 20 - Ukraine is declared a republic.

November 20, 1917 and Beyond

Amidst the shadows of war and civil unrest, a small burst of light began to shine in Strasbourg, France. Erich Leo Lehmann was born November 20th, 1917 – a mere 7 months after The United States entered the First World War.

Some years later, at the age of 16, he and his family went to live in Switzerland to avoid the Nazis. After five years in Switzerland, and two years in Cambridge, Erich L. Lehmann arrived in the United States in 1940 with a letter of introduction from the wife of Edmund Landau. Landau had passed away a couple of years earlier from a heart attack. The letter of introduction was for Richard Courant who was in New York and had been a colleague of Landau in Göttingen. After being asked by Courant if he wanted to stay in New York or live in the United States, Lehmann responded that he wanted to live in the United States and then followed Courant's



Erich Leo Lehmann in 1919

advice to go to an "up-and-coming university" in Berkeley, California. Upon his arrival in Berkeley in 1941, Erich Lehmann met with Griffith C. Evans who had been a mathematician at the Rice Institute, now Rice University. Evans had been brought to Berkeley to develop the mathematics department that was in disarray. Evans was an excellent mathematician and many of his contributions as a mathematician and administrator have been recorded in Morrey (1983) and Lehmann (2007). As an administrator, Evans was able to attract to Berkeley some of the best mathematicians of the time. With a broad vision for the mathematics department, Evans supported a three-week visit by R. A. Fisher to Berkeley. The visit did not go well. Reid (1992) writes in her book that, despite a generous endowed lectureship whose terms required the lecturer to spend their time on campus to interact with interested faculty, Fisher spent the first five days of his visit in San Francisco and went back to England a day earlier "standing up a dinner in his honor". Reid (1992) writes that according to Raymond T. Birge, chair of the Physics department at the time, "Fisher was the most conceited man he had ever met - 'and that is saying a lot with such competitors as Millikan et al!'" Birge put forth Neyman's name to Evans. Evans had never heard of Neyman but after some inquiring an offer was



Erich Leo Lehmann in 1924

made. Neyman accepted and a few years later, after an offer from Columbia became available, was able to negotiate with Evans for the creation of a separate statistics department. These and other fascinating details may be found in Lehmann (1993, 1996, and 2007) and Reid (1992) and other references in the bibliography.

During their first meeting, Evans offered him a probationary graduate student status and six months later a teaching assistantship in the Mathematics Department. With the advent of the Second World War, Evans suggested to change areas of study and consider a more useful subject. Either Physics or Statistics would be more useful than Mathematics. After completing the required course work, and after returning from Guam where he and Joseph Hodges had served, it was time for Erich L. Lehmann to begin work on a dissertation. A topic with a probabilistic flavor was proposed by Pao-Lu Hsu after consulting with Neyman. Progress was swift and as Lehmann prepared to write up some of the results, a reference led to other references that led to the painful discovery that the results so far obtained had been published a few decades earlier. At that time Neyman was invited as a member of a delegation to observe the Greek elections. Concerned with the disappointment of his student, and knowing that he might return until a few months

later, Neyman asked Hsu to consider providing a new problem for Lehmann's thesis. Hsu suggested a new problem that he had thought about and planned to work on, and a problem for which he already had some preliminary results. Lehmann came to know about Hsu's generosity some time later and had hoped to thank Hsu personally after Hsu's return to Berkeley from Columbia, but this would never happen as Hsu opted to return to China. With Neyman in Greece and Hsu back in China, Neyman suggested George Polya as a surrogate advisor. Frequent visits to Polya at Stanford finally yielded a thesis. A new problem presented itself in that Polya was not a faculty member at Berkeley. Fortunately, Neyman was able to return in time for the thesis defense. He had been asked to return to the United States as his services were no longer needed in Greece. In effect, he had been dismissed for insubordination. He had decided to investigate on his own the possibility that the elections had been rigged. It thus happened that in 1946 Erich L. Lehmann received the degree of Doctor of Philosophy. The title of his thesis: "Optimum Tests of a Certain Class of Hypotheses Specifying the Value of a Correlation Coefficient".

Erich L. Lehmann stayed in Berkeley as a young faculty member and the "Rest of the story" is well known. Besides his many influential publications, he was able to produce 41-plus Ph.D. students. The following table provides the names of the students and the year of graduation.

Colin Ross Blyth	1950	Gouri Kanta Bhattacharyya	1966
Fred Charles Andrews	1953	James Nwoye Adichie	1966
Allan Birnbaum	1954	Dattaprabhakar V. Gokhale	1966
Hendrik Salomom Konijn	1954	Frank Rudolph Hampel	1968
Balkrishna V. Sukhatme	1955	Wilhelmine von Turk Stefansky	1969
V. J. Chacko	1959	Louis Jaeckel	1969
Piotr Witold Mikulski	1961	Friedrich Wilhelm Scholz	1971
Madan Lal Puri	1962	Dan Anbar	1971
Krishen Lal Mehra	1962	Michael Denis Stuart	1972
Subha Bhuchongkul Sutchritpongsa	1962	Claude L. Guillier	1972
Shishirkumar Jogdeo	1962	Sherali Mavjibhai Makani	1972
Peter J. Bickel	1963	Howard J. M. D'Abrera	1973
Arnljot Høyland	1963	Hyun-Ju Yoo Jin	1974
R. Murty Ponnapalli	1964	Amy Poon Davis	1977
Milan Kumar Gupta	1964	Jan F. Bjørnstad	1978
Madabhushi Raghavachari	1964	William Paul Carmichael	1981
Vida Greenberg	1964	David Draper	1981
Kjell Andreas Doksum	1965	Wei-Yin Loh	1982
William Harvey Lawton	1965	Marc J. Sobel	1983
Shulamith Gross	1966	Javier Rojo	1984
Bruce Hoadley	1966		

So why, then, was the date changed from November 20^{th} to May 16^{th}? After all, it would have been a great way of celebrating Erich's wonderful 90 years of life. But that is precisely the issue. To the reader who does not know Erich L. Lehmann personally, holding a conference on his birthday, a conference that is named after him, would seem only natural. However, those close to him know very well that he is very modest and an event like the Symposium held on his birthday, would be rather uncomfortable for him. He thought that the meeting would turn into a birthday celebration and he would not have it that way. The Lehmann Symposia should be true to its beginnings: A meeting to showcase good theoretical work. Thus, it came to be that the venue and the date for the 3^{rd} Lehmann Symposium were changed. As it was not possible to celebrate his 90^{th} birthday with the symposium, this volume is dedicated to Erich's 90^{th} birthday. I am sure that, given the opportunity,

all his Ph.D. students listed above and colleagues around the world would join me in wishing Erich Leo Lehmann many more wonderful years! Our lives have been greatly enriched through our interactions, professional and social, with him.

Acknowledgements

The work of the author was partially supported by NSF Grant DMS-053246, NSA Grant H98230-06-1-0099, and NSF REU Grant DMS-0552590.

References

[1] Enock, Reginald C. (1912). *Mexico: Its Ancient and Modern Civilisation, History, Political Conditions, Topography, Natural Resources, Industries and General Development*. Hume, M. (ed.). London, T. Fisher Unwin.
[2] DeGroot, Morris H. (1986). A conversation with Erich L. Lehmann. *Statist. Sci.*, **1.2**, 243–258.
[3] Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association*, **88.424**, 1242–1249.
[4] Lehmann, E. L. (1993). Mentors and early collaborators: reminiscences from the years 1940-1956 with an epilogue. *Statist. Sci.*, **8.3**, 331–341.
[5] Lehmann, E. L. (1996). The creation and early history of the Berkeley statistics department. *Statistics, probability and game theory*, IMS LNMS, **30**, Inst. Math. Statist., Hayward, CA.
[6] Lehmann, E. L. (1997). Testing statistical hypotheses: the story of a book. *Statist. Sci.*, **12.1**, 48–52.
[7] Lehmann, E. L. (2004). Optimality and symposia: some history. In *The First Erich L. Lehmann Symposium - Optimality*. J. Rojo and V. Pérez-Abreu (Eds.) IMS LNMS, **44**, 1–10.
[8] Lehmann, E. L. (2007). *Reminiscences of a Statistician: The Company I Kept*. Springer, New York.
[9] Morrey, C. B. (1983). *Griffith Conrad Evans 1887–1973: A Biographical Memoir*. National Academy of Sciences, Washington D.C.
[10] Reid, C. (1982). *Neyman from Life*. Springer-Verlag, Berlin-Heidelberg-New York-Tokyo.

IMS Collections
Vol. 0 (2009) 8–10
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000015

The Honorable Erich L. Lehmann

Stephen Stigler

University of Chicago

The year 2007 marks a concurrence of important statistical anniversaries. It is the 350^{th} anniversary of the publication of the first printed work on mathematical probability, the short tract that Christian Huygens wrote following a visit to Paris, where he learned of the investigations of Fermat and Pascal. Also, 2007 is the 150^{th} year since the birth of Karl Pearson, father of the Chi-square test and much else. And related to both those events, it is also the year our teacher, friend, and colleague Erich Lehmann celebrates his 90^{th} birthday. Christian Huygen's tract served as the textbook on probability for over a half century, helping to form that subject. Karl Pearson inaugurated an important species of hypothesis testing. Both then have important similarities to Erich Lehmann. But rather than further explore those analogies immediately, I would like to characterize an important part of Erich's ongoing research by looking back to a more modern document.

The University of Chicago, rare among research universities, gives honorary degrees only in recognition of scholarly contributions of the highest order. We do not use those degrees to honor movie stars, philanthropists, or even heads of state (at least not over the past 80 years). There is a partial exception: we do so honor the departing Chair of our Board of Trustees. But that is the limit of the exceptions. We do not use this device to honor work done at Chicago; our major financial supporters are recognized in other ways; and discreet inquiries on behalf of politicians, celebrities, popular artists, and several heads of state have been politely turned aside, however meritorious they may have been on other grounds. Scholarship is the only coin of our realm.

One of the fields where this practice has been actively pursued is statistics. The first degree our newly formed department granted was an honorary Doctorate of Science to Ronald Fisher, June 13, 1952. This was followed over the next 31 years by degrees to Harold Hotelling (1955), Jerzy Neyman (1959), Maurice Bartlett (1966), John Tukey (1969), and Fred Mosteller (1973). In 1990, in anticipation of our University's centennial celebration beginning the following year, the Department undertook to resume this practice after several years by proposing an honorary degree for Erich Lehmann.

The procedure for granting honorary degrees at Chicago is a bit involved. After getting departmental agreement (an easy matter in this case) it is necessary to prepare a detailed case to be submitted to a cross-university committee appointed for this task. Exacting standards are upheld and only a fraction of proposals are given the nod of approval. Local legend has it that when someone proposed the Queen of England for a degree, it was turned back with a request for her list of publications. The procedure is like that for hiring a senior scholar from outside the university: several letters of recommendation must be solicited, the evidence must be assembled and carefully presented, all with an uncertain outcome. And unlike senior hires, even if the proposal is successful and the offer accepted, you do not get to keep the candidate!

In preparing Erich's case we solicited letters from a dozen of the top international statistical scholars over the past half-century. All of these were glowing testaments to an amazingly influential body of work, as well as to Erich's collegial role in helping to build modern mathematical statistics. To buttress the case we did a careful citation study to document quantitatively the pervasive influence of Erich's articles and books. To our eye there was no doubt, but all this needed to be presented to a demanding committee of some of the universitys best professors, regrettably none with more than a superficial knowledge of our subject. It is that memo which I will now present.

* * *

23 April 1990

Memo to Committee on Honorary Degrees

On behalf of the Department of Statistics, and with their unanimous endorsement, I wish to nominate Erich L. Lehmann for an Honorary Degree. We believe he would be an exceptionally appropriate candidate for a Centennial Honorary Degree at the convocation of October 1991; alternatively he could be considered for an earlier convocation, such as June 1991. Prior to his retirement in 1988, Lehmann was Professor of Statistics at the University of California at Berkeley, where he remains extremely active.

Great scientists can be roughly classified as one of two types: those whose primary achievement was a single brilliant discovery, and those who have over the course of a life's work constructed a discipline, a school. Lehmann has made many important discoveries, but he is more of the second type than the first. In the years after the Second World War, the dominant paradigm in American mathematical statistics was the decision theoretic school of Jerzy Neyman (Sc.D., Chicago, 1959) and Abraham Wald. Wald died tragically in 1950 and Neyman's attention was absorbed by other matters after the mid-1950's. While many hands played a role in the prospering of this approach over 1950-1980, it is arguable that the chief architect of the expansion of the paradigm, and the one most responsible for its immense international influence over that period, was Erich Lehmann.

There is an interesting historical parallel to Lehmann's role in this school of thought. In the 1650's Pascal and Fermat founded probability theory. Yet they taught few and published little. It was Christian Huygens's 1657 tract that was responsible for the form in which these ideas were disseminated for half a century, and their widespread application in mathematics, philosophy, and science. Similarly, it was Lehmann's mimeographed lecture notes of estimation (1950, widely circulated and reproduced, but only published as a book in 1983) and Lehmann's book of hypothesis testing (1959) that provided the form in which the Neyman-Wald approach came to dominate a major portion of the mathematical world. Not all of the results are Lehmann's (though many are), but the arrangement, the elegant seamless presentation, the coherency of the whole, are his in a way that has not been true in many sciences. To a large degree, Erich Lehmann created the curriculum of the world's graduate programs in mathematical statistics over the period 1960-1980. Lehmann has personally supervised over 50 Ph.D. dissertations, and he counts among his students some of the most influential statisticians in the United States, Europe, and Asia.

The focus of the school that is so associated with Lehmann is the application

of decision theory to statistical problems, the construction of a calculus of optimal statistical procedures. In one simple setting, parametric estimation, an objective criterion ("a loss function") is defined and an attempt is made within a framework of stochastic models to seek the best procedure, or at any rate to determine an order for the available procedures. The pitfalls of this approach are many: In all but the simplest problems, the mathematical difficulties can be immense, and the specification of models and objective criteria that capture the essence of the scientific problem is rarely straightforward. The success of Lehmann and his school has been due to the balance they have maintained in creating a system of mathematical structures of sufficient richness to encompass a large range of practical problems, yet not so amorphously vague that a true discipline could not be constructed around them. Mathematics flourishes in the detailed exploration of constrained spaces with widely accepted rules; statistics flourishes with the flexibility to treat the infinite variety of problems in the real world. Lehmann's genius has been his ability to reconcile these divergent goals and build a school that has enriched both sides.

* * *

To this was appended the letters and the citation study.

The proposal was successful; we were delighted when Erich accepted and both he and Julie attended the special Convocation on October 3, 1991, celebrating the University's Centennial. Their visit was the occasion for several parties, dinners, toasts, and culminated in a grand ceremony in Rockefeller Chapel [yes, we do recognize donors in other ways], attended by dignitaries representing the great universities of the world. The citation on Erich's degree read:

> Your research on the application of decision theory to statistical problems has helped create and organize modern mathematical statistics; your elegant treatises have guided the curricula of a majority of the nation's graduate programs and given shape to the discipline, and your teaching has inspired a generation of scholars.

Erich has received many other honors, of course, including an earlier honorary degree at the University of Leiden. And we have since 1991 given more honorary degrees; to Charles Stein (1992), Ulf Grenander (1994), Bradley Efron (1995), David Aldous (2000), Persi Diaconis (2003), and Grace Wahba (2007). The occasion of Erich's University of Chicago degree retains a special place in our memories. So too, to those of us who took his classes (in my case some 40 years ago) does the memory of his careful and supremely clear lectures, and his papers and books that helped shape the modern statistical world. For all that and more, we may toast this extraordinary scholar on the 90^{th} anniversary of his birth!

IMS Collections
Vol. 0 (2009) 11–17
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000009

Some History of Optimality

Erich L. Lehmann

University of California, Berkeley

Contents

1 Combination of Observations 11
2 Maximum Likelihood Estimation 11
3 The Neyman-Pearson Program 12
4 The Neyman-Pearson Theory of Hypothesis Testing 12
5 Wald's Optimality Criteria 13
6 The Hunt-Stein Theorem 13
7 Some Extension of the Neyman-Pearson Theory 14
8 Large Sample Optimality of Testing 14
9 Optimal Design 15
10 A Culture Clash 15
11 Tukey's Criticism 16
12 Conclusion 16
References 16

1. Combination of Observations

The earliest optimality considerations appear to be those of Laplace and Gauss at the beginning of the 19^{th} Century concerned with determining the best linear estimates of parameters in linear structures. Laplace calls these optimal estimates "the most advantageous" while Gauss refers to them as "the most plausible values." Various aspects of this problem were discussed throughout the 19^{th} Century under the heading "Combination of Observations." The version of the principal result generally accepted today is the so-called Gauss-Markov theorem. It states (in modern language) that the least squares estimates are the linear unbiased estimates with minimum variance. While restricted to linear estimates, the result is nonparametric in that it makes no assumptions about the distribution of the errors. For an account of this work see, for example, Stigler (1986), Hald (1998), or Chatterjee (2003).

2. Maximum Likelihood Estimation

Optimality next played an important part in Fisher's fundamental paper of 1922. In this paper (followed by a clarifying paper in 1925), Fisher considers estimation in parametric models, proposes the maximum likelihood estimator (MLE) as a generally applicable solution, and claims (but does not prove) that such estimators are consistent and asymptotically efficient efficient (i.e., that they minimize the asymptotic variance). Note that unlike the Gauss-Markov theorem, maximum likelihood estimation does assume that the distribution of the variables belongs to a given parametric family.

Maximum likelihood has become the most widely used method of estimation, and there has been an enormous amount of work connected with Fisher's claims concerning it. It has led to the discovery of superefficiency on the one hand and of second order efficiency on the other. Many counterexamples have been found (even to consistency), and a number of modifications (bias correction and replacement of the MLE by a consistent root of the likelihood equation, for example) have been proposed. The situation is complex, but under suitable restrictions Fisher's conjecture is essentially correct when the likelihood equation has a unique root. For a more precise statement, see Shao (1999), and for further discussion, for example, Efron (1982), Le Cam (1990), and Barndorff-Nielsen and Cox (1994).

3. The Neyman-Pearson Program

Least squares and maximum likelihood were first proposed on intuitive grounds and then justified by showing that they possessed certain optimality properties. Optimality as a deliberate program for determining good procedures was introduced in 1933 by Neyman and Pearson in a paper (on testing rather than estimation) appropriately called, "On the problem of the most efficient tests of statistical hypotheses." As they explain in the introduction:

> The main purpose of the present paper is to find a general method of determining tests which ... would be most efficient"[in the sense of minimizing the probability of erroneous conclusions].

In a certain sense this is the true start of optimality theory.

4. The Neyman-Pearson Theory of Hypothesis Testing

Neyman and Pearson (1933) implemented the above program by seeking, for any given situation, the test which, among all those controlling the probability of false rejection at a given level α, has the maximum power (and hence the minimum probability of false acceptance).

For testing a simple hypothesis against a simple alternative, they found the solution to this problem to be the likelihood ratio test. This result, which is mathematically quite elementary but has crucial statistical consequences, is known as the Neyman-Pearson Lemma.

It turns out that in some (very rare) cases the same test is most powerful against all alternatives under consideration. Such a uniformly most powerful (UMP) test is then the optimal solution to the given testing problem. Where a UMP test does not exist, additional criteria must be invoked.

For example, when nuisance parameters are present, Neyman and Pearson require that under the hypothesis the rejection probability be α for all values of the nuisance parameters. They call such rejection regions similar regions. As an important example, they show that the one-sided t-test is UMP among all similar regions.

For two-sided alternatives, one would not expect UMP tests to exist even without nuisance parameters. For such cases, Neyman and Pearson then impose the additional condition of unbiasedness, i.e., that the power of the test is $\geq \alpha$ for all alternatives. In follow-ups to their 1933 paper (1936 and 1938), they show, for example, that the two-sided t-test is UMP among all unbiased tests.

UMP similar or unbiased tests exist for important classes of testing problems concerning a single real-valued parameter (in the presence or not of nuisance parameters) but not for hypotheses such as

$$H : \theta_1 = \cdots = \theta_s$$

concerning several parameters.

A different condition, the principle of invariance (suggested by Hunt and Stein, unpublished), is successful in a number of important such multiparameter situations. If both the hypothesis and the class of alternatives remain invariant under a group G of transformations of the sample space, there does in these cases exist a UMP test among all tests invariant under G.

5. Wald's Optimality Criteria

A quite-different approach to optimality was initiated by Wald in his 1939 paper, "Contributions to the theory of statistical estimation and testing hypotheses," and was then developed further in a series of publications culminating in his 1950 book, "Statistical Decision Functions."

This approach was part of Wald's formulation of a general theory of decision procedures. Instead of seeking procedures that are uniformly optimal among some suitably restricted class of decision procedures, Wald proposes to minimize some global feature of the performance.

Specialized to hypothesis testing, these proposals reduce to:

i Maximize the average power, averaged with respect to some suitable weight function over the alternatives. For obvious reasons, Wald called such maximizing procedures Bayes solutions.

ii Maximize the minimum power over alternatives bounded away from the hypothesis.

6. The Hunt-Stein Theorem

An important connection between the earlier invariance approach and that of Wald is the Hunt-Stein theorem. It states that if a UMP invariant test exists under a group G satisfying certain conditions (called amenability), then this test also maximizes the minimum power over any invariant set of alternatives.[1]

To illustrate this theorem, consider the univariate linear hypothesis. Under the combined action of the groups of location, scale and orthogonal transformation, a UMP invariant test exists. Since these groups satisfy the Hunt-Stein conditions, the resulting test therefore maximizes the minimum power against all invariant sets of alternatives.

As a second example, consider the multivariate one-sample problem. Hotelling's T^2-test is UMP among all tests that are invariant under the group G of all nonsingular linear transformations. Since G is not amenable, the Hunt-Stein theorem does not apply. The resulting maximin problem poses considerable difficulties.

[1] The 1946 paper by Hunt and Stein, "Most stringent tests of statistical hypotheses," containing this theorem was never published. The theorem appeared in print for the first time in Lehmann, "Testing Statistical Hypotheses," John Wiley, 1959.

7. Some Extension of the Neyman-Pearson Theory

The Neyman-Pearson theory has been extended in a number of directions. The following are two extensions of the Neyman-Pearson Lemma, which is so basic to this theory.

(i) Sequential Analysis

During World War II, Wald proposed sequential procedures, as a way of obtaining good power with fewer observations. In particular, he suggested the probability ratio test for testing a simple hypothesis against a simple alternative. This test continues observation as long as the likelihood ratio remains between two fixed limits and takes the indicated decision (acceptance or rejection) as soon as it falls outside these limits.

In 1948, Wald and Wolfowitz proved the remarkable result that for testing a simple hypothesis against a simple alternative, the sequential probability ratio test, among all tests with the same (or smaller) probabilities of error, minimizes the expected number of observations both under the hypothesis and the alternative.

(ii) Robust Inference

All the work reported so far (except for the Gauss-Markov theorem) was carried out under the assumption of an underlying parametric model. In practice, such an assumption can be expected to hold at best approximately. As a more realistic formulation, Huber (1964) suggested replacing the assumption of a parametric model by that of a neighborhood of such a model.

In the following year, he obtained the analog of the Neyman-Pearson Lemma. For testing the neighborhood of a distribution P_0, the test maximizing the minimum power over the neighborhood of an alternative P_1 is a censored version of the likelihood ratio test of P_0 against P_1.

(iii) Multiple Testing

A very different extension of the Neyman-Pearson theory that is of great practical importance deals with the situation in which a number of hypotheses (sometimes a very large number) are being tested rather than just one. Unlike the work discussed so far, which is classical, the theory of multiple testing is an area of very active ongoing research.

The first problem here (before optimization) is to find a suitable generalization of the concept of significance level that would provide satisfactory control of the probability of false rejections. After this, maximin tests have been obtained, which require, however, not only unbiasedness and invariance but also a condition of monotonicity. Surveys of the current state of this work are provided by Shaffer (2004, 2006).

8. Large Sample Optimality of Testing

A small-sample theory of optimal estimation parallels that of optimal testing sketched in Sections 4-7, with concepts such as unbiasedness, equivariance (instead of invariance), and minimax variance (instead of maximin power), and will not be

discussed here. Asymptotic optimality for estimation goes back to Fisher (1922), as mentioned in Section 2, and is defined as minimum asymptotic variance.

For testing, asymptotic optimality is considerably more complex, both conceptually and technically. It was first studied by Wald in 1941. Consider testing a simple hypothesis $\theta = \theta_0$ against a simple alternative $\theta = \theta_1$. If we keep both θ_0 and θ_1 fixed, and carry out the tests at a fixed level α, the power of any reasonable test sequence will tend to 1. Thus any such test sequence will in a trivial sense be asymptotically UMP.

A more useful approach is obtained by considering a sequence of alternatives

$$\theta_n = \theta_0 + h/\sqrt{n} \tag{8.1}$$

For a sequence with fixed h, the power will typically tend to a limit between 0 and 1 as $n \to \infty$. As h varies from 0 to ∞, the limiting power will be an increasing function of h, going from α to 1; we shall call this the asymptotic power function. A sequence of tests can then be defined as being asymptotically most powerful (AUMP) if it maximizes the asymptotic power for all h.

Unlike the finite sample situation where UMP tests exist only rarely and then are unique, it turns out that AUMP tests exist under very weak assumptions, and that in fact many different AUMP tests exist for the same situation, among them the likelihood ratio test, the Wald test, and the locally most powerful (Rao) test. To distinguish them, one must resort to higher order asymptotics. (See, for example, Barndorff-Nielsen and Cox (1994)). An exposition of the first order theory due to Le Cam can be found in Lehmann and Romano (2005).

9. Optimal Design

In generalization of minimizing the variance of an estimator, optimal design theory is concerned with determining the design that minimizes some function of the covariance matrix of the best linear estimators of the parameters in question. In particular, D-optimality minimizes the determinant, E-optimality is a minimax criterion, and so on. After some isolated earlier efforts, the problem of optimal design was studied systematically by Kiefer in more than 40 papers between 1958 and his early death in 1981. They make up vol. III of his collected papers.

10. A Culture Clash

Not everyone was enthusiastic about optimality as a guiding principle. Criticism was highlighted at a 1958 meeting of the Royal Statistical Society at which Kiefer presented a survey talk on "Optimum Experimental Designs." It was a discussion paper, and the reaction was nearly uniformly negative. The core of the disagreement is stated clearly when Barnard quotes Kiefer as saying of procedures proposed in a paper by Box and Wilson (1951) that they "often [are] not even well-defined rules of operation." Barnard's reply:

> In the field of practical human activity, rules of operation which are not well defined may be preferable to rules which are.

The conflict is discussed by Henry Wynn in his introduction to a reprint of Kiefers paper in "Breakthroughs in Statistics," vol. I (Kotz and Johnson, Eds.). Wynn calls it "a clash of statistical cultures."

This clash is between the Wald school of abstraction and optimality on the one hand and what Tocher, in his discussion of Kiefer's paper, calls "the school of British experimental design – a very practical people," on the other.

11. Tukey's Criticism

Criticism of optimality was not confined to England. An outspoken American critic questioning the value and importance of optimization was John Tukey. His attitude is indicated by the titles of two philosophical papers in 1961 and 1962, which are titled respectively, "The tyranny of the best" and "Dangers of optimization."

Tukey's concern with optimality had its origin in the fact that at the time optimization had become the dominant interest of mathematical statistics. In the 1962 paper, he writes:

> Some [statisticians] seem to equate [optimization] to statistics an attitude which, if widely adopted, is guaranteed to produce a dried-up, encysted field with little chance of real growth.

12. Conclusion

That optimality considerations had become so dominant is explained by the historical situation. Periods of great innovation are followed by periods of consolidation, in which the new work is given a final (or, as Tukey says, encysted) form. Such a dichotomy is also discussed by Huber (1975). Thus, the revolutionary work of Student and Fisher was followed by the optimization approach of Neyman, Pearson, Wald and their students described in this paper.

Today we are faced with the opposite situation. We are confronted with new problems arising from immense data sets and involving problems of great complexity. Ad hoc solutions are proposed and tried out on a few examples. This is a natural first step, but eventually we will want to justify the solutions at which we have arrived intuitively and by trial and error. A theoretical underpinning will be provided and the conditions will be found under which these solutions are the best possible.

References

[1] Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics.* Chapman & Hall, London.

[2] Box, G. E. P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society (B)*, **13**, 1–45.

[3] Chatterjee, S. K. (2003). *Statistical Thought.* Oxford University Press.

[4] Efron, B. (1982). Maximum likelihood and decision theory. *Annals of Statistics,* **10**, 309–368.

[5] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London (A)*, **222**, 401–415.

[6] Fisher, R. A. (1925). Theory of statistical estimation. *Cambridge Philos. Soc.*, **22**, 700–725.

[7] Hald, A. (1998). *A History of Mathematical Statistics - from 1750 to 1930.* John Wiley, New York.

[8] Huber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73–101.

[9] Huber, P. (1965). A robust version of the probability ratio test. *Annals of Mathematical Statistics*, **36**, 1753–1758.

[10] Huber, P. (1975). Applications vs. abstraction: the selling out of mathematical statistics? In *Proc. Conference on Directions of Mathematical Statistics. Suppl. Adv. Prob.*, **7**, 84–89.

[11] Kiefer, J. C. (1959). Optimum experimental designs (with discussion). *Journal of the Royal Statistical Society (B)*, **21**,273–319. (Reprinted in Kotz and Johnson (1992). *Breakthroughs in Statistics, Vol. 1.*, Springer-Verlag.

[12] Kimball, G. E. (1958). A critique of operations research. *J. Wash. Acad. Sci.*, **48**, 33–37.
[13] Kotz, S. and Johnson, N. L. (Eds. 1992, 1997). *Breakthroughs in Statistics*, **1**. Springer-Verlag, New York.
[14] Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. of Calif. Publ. in Statist.*, **1**, 277–330.
[15] Le Cam, L. (1990). Maximum likelihood an introduction. *ISI Review*, **58**, 153–171.
[16] Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses* (3rd Ed.). Springer, New York.
[17] Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. (A)*, **231**, 289–337.
[18] Neyman, J. and Pearson, E. S. (1936, 1938). Contributions to the theory of testing statistical hypotheses. *Statist. Res. Memoirs*, **1**, 1–37; **2**, 25–57..
[19] Pearson, E. S. (1939). "Student" as a statistician. *Biometrika*, **30**, 210–250.
[20] Shaffer, J. P. (2004). Optimality results in multiple hypothesis testing. In *Proceedings of the First Lehmann Symposium.* J. Rojo and V. Pérez-Abreu (Eds). IMS LNMS, **44**, 11–35.
[21] Shaffer, J. P. (2006). Recent developments towards optimality in multiple hypothesis testing. In *Proceedings of the Second Lehmann Symposium.* J. Rojo (Ed). IMS LNMS, **49**, 16–32.
[22] Shao, J. (1999). *Mathematical Statistics.* Springer, New York.
[23] Stigler, S. (1986). *The History of Statistics.* Harvard University Press, Cambridge, MA.
[24] Tukey, J. W. (1961). Statistical and quantitative methodology. In *Trends in Social Science*, (D. P. Ray, Ed.). Philosophical Library, New York. (Reprinted in Vol. III of Tukey's Collected Works.)
[25] Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics.* Reprinted in Vol. III of Tukey's Collected Works.
[26] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.*, **54**, 426–482.
[27] Wald, A. (1950). *Statistical Decision Functions.* John Wiley, New York.
[28] Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, **19**, 326–339.

IMS Collections
Vol. 0 (2009) 18–30
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000001

An Optimality Property of Bayes' Test Statistics

Raghu Raj Bahadur and Peter J. Bickel[1,*]

University of Chicago and Imperial College, London

Dedicated to Erich Lehmann on his 90th Birthday

Contents

1 Introduction 19
2 A Generalization of a Theorem of Bahadur 19
3 General Assumptions and a Useful Lemma 21
4 The Main Theorem 26
5 Optimality of Minimax Tests 29
References 30

Preface:

This paper dates back to the late 60's when I collaborated with Raj Bahadur, who is unfortunately no longer with us. The reason it has not appeared until now is that he felt it had to be accompanied by a number of multivariate examples. We both went on to other things; the examples were not worked out although we both knew of the existence of some of them. So why is this paper appearing here (with the approval of Steve Stigler, an executor of the Bahadur estate)? First, in addition to attesting to Erich's continued vital presence, it gives me the opportunity of paying a tribute to Bahadur, who was a friend of both of ours. Second, it is an interesting reminder of how writing styles have changed on the whole I think for the better – from rigorous abstract formulation and mathematically rigorous presentation to more motivation and a lot of hand waving. Third, and most importantly, the result is an example of what I think both Erich and I consider an important endeavor, the reconciliation of the Bayesian and frequentist points of view (in context of now rather unfamiliar asymptotics). In an important paper in the 5^{th} Berkeley Symposium [4], Bahadur showed that the maximum likelihood ratio statistic possessed an optimality property from the view of a large deviation based frequentist comparison of tests he introduced in 1960 [1]. Our paper shows that this property is shared by Bayes test statistics for reasonable priors and conjectures that a corresponding Bayesian optimality property holds for the maximum likelihood ratio statistic. If true this can be viewed as the large deviation analogue of the well-known Bernstein von Mises' theorem – see Lehmann and Casella [10] p.489, which establishes the equivalence at the $n^{-\frac{1}{2}}$ scale of Bayesian and maximum likelihood estimates. Establishing this conjecture is left as a challenge to the reader.

[1]On leave from University of California, Berkeley (1965-66).
*Prepared with partial support of N.S.F. Grant G.P. 2593.

Given the historical interest I have not changed the text save for typos and only brought references up to date.

1. Introduction

In [4] one of the authors established the optimality of the classical likelihood ratio test statistic in terms of a method of stochastic comparison previously introduced by him in [1], [2], and [3].

In the main theorem of this paper, Theorem 2 of Section 4, we show that this property is shared by Bayes test statistics (averages of likelihood ratios with respect to probability measures on the parameter space) under conditions which are slightly different from, and in some respects weaker than those given in [4]. These assumptions are given and discussed in Section 3. Section 5 contains a theorem establishing the asymptotic optimality of minimax tests under appropriate restrictions.

In Section 2 we give a strengthening Theorem 1 of [4], which established a lower bound for the slope of any family of tests in terms of the Kullback-Leibler information numbers. The proof given here drops assumption 1 of [4] and weakens assumption 2 considerably. This argument seems to give some insight into the necessity of an assumption such as our modification of assumption 2 of [4].

2. A Generalization of a Theorem of Bahadur

Even as in [4] we let X be an abstract space, $\mathcal{A}$ a field on X, P_θ, $\theta \in \Theta$, a set of probability measures on $(X, \mathcal{A})$, and Θ_0 a given subset of Θ. For any θ, θ' we define,

$$K(\theta, \theta') = -\int_X \log \frac{dP_{\theta'}}{dP_\theta}(x) dP_\theta(x) \ , \tag{2.1}$$

where $\frac{dP_{\theta'}}{dP_\theta}$ is the ratio of the Radon Nikodym derivatives of P_θ, $P_{\theta'}$ with respect to (say) $P_\theta + P_{\theta'}$ and $0/0$ is by convention equal to 1. Also let,

$$J(\theta) = \inf\{K(\theta, \theta') : \theta' \in \Theta_0\} \ . \tag{2.2}$$

It is well known that (cf. [4]), $0 \leq K(\theta, \theta') \leq \infty$, and necessarily the same is true of $J(\theta)$. Following [4], let T_n be any sequence of extended real valued measurable functions of the infinite product space $(X^\infty, \mathcal{A}^\infty)$ such that T_n is a function of the first n co-ordinates only. Denote the cumulative distribution of T_n, when θ obtains, by $F_n(t, \theta)$, i.e.,

$$F_n(t, \theta) = P_\theta\big[T_n(s) < t\big] \ , \tag{2.3}$$

where P_θ now denotes the infinite product measure extension of P_θ to X^∞. Finally, let,

$$L_n(s) = \sup\big\{1 - F_n\big(T_n(s), \theta\big) : \theta \in \Theta_0\big\} \ . \tag{2.4}$$

We assume that L_n is measurable. This for instance holds if $F_n(T_n, \theta)$ is a separable stochastic process in θ for $\theta \in \Theta_0$.

We can now state and prove, in the above framework,

Theorem 1. *If*

$$\int_X \Big[\log\Big(\frac{dP_\theta}{dP_{\theta'}}\Big)\Big]^2 dP_\theta(x) < \infty \ ,$$

for every $\theta \in \Theta - \Theta_0$, $\theta' \in \Theta_0$ *such that* $K(\theta, \theta') < \infty$, *then*

$$\liminf_n \frac{1}{n} \log L_n(s) \geq -J(\theta) \tag{2.5}$$

with P_θ *probability 1 for every* $\theta \in \Theta - \Theta_0$.

Proof. Fix $\theta \in \Theta - \Theta_0$. Assume the Theorem has been proved for Θ_0 simple. Clearly we can suppose $J(\theta) < \infty$ and can find $\{\theta_m\}$ with $K(\theta, \theta_m) < \infty$ and $K(\theta, \theta_m) \to J(\theta)$. But then

$$\frac{1}{n} \log L_n(s) \geq \frac{1}{n} \log \big(1 - F_n(T_n(s), \theta_m)\big) \ . \tag{2.6}$$

By our assumption of the theorem for Θ_0 simple, we have

$$\liminf_n \frac{1}{n} \log \big(1 - F_n(T_n(s), \theta_m\big) \geq -K(\theta, \theta_m) \tag{2.7}$$

with probability 1. Inequalities (2.6) and (2.7) then imply (2.5). If $\Theta_0 = \{\theta_0\}$,

$$P_\theta\big[\liminf_n \frac{1}{n} \log L_n(s) \geq -J(\theta)\big] = 1$$

if and only if,

$$P_\theta\big[1 - F_n(T_n, \theta_0) < a^n \exp -nK(\theta, \theta_0) \text{ infinitely often } \big] = 0 \tag{2.8}$$

for every $0 \leq a \leq 1$. Fix a. Let $A_n = \big[1 - F_n(T_n, \theta_0) < a^n \exp\{-nK(\theta, \theta_0)\}\big]$. Then,

$$P_{\theta_0}(A_n) \leq a^n \exp\{-nK(\theta, \theta_0)\} \ . \tag{2.9}$$

By the Neyman-Pearson lemma there exists c_n, such that,

$$P_{\theta_0}\Big\{\sum_{j=1}^n \log \frac{dP_\theta}{dP_{\theta_0}}(x_j) > nc_n\Big\} \leq a^n \exp\{-nK(\theta, \theta_0)\} \tag{2.10}$$

and,

$$P_{\theta_0}\Big\{\sum_{i=1}^n \log \frac{dP_\theta}{dP_{\theta_0}}(x_i) > nc_n\Big\} \geq P_{\theta_0}(A_n) \ . \tag{2.11}$$

We require,

Lemma 1. *If* (2.10) *holds for all* n, *then there exists an* $\varepsilon > 0$ *such that* $\liminf_n c_n \geq K(\theta, \theta_0) + \varepsilon$.

Proof. By a Theorem of Chernoff [7],

$$\frac{1}{n} \log P_{\theta_0}\Big\{\sum_{i=1}^n \log \frac{dP_\theta}{dP_{\theta_0}}(x_i) > nz\Big\} \to \inf_t \log H(t, z) \ , \tag{2.12}$$

for

$$z \geq \int_X \log \frac{dP_\theta}{dP_{\theta_0}}(x) dP_{\theta_0}(x) \ ,$$

where

$$H(t,z) = e^{-tz} \int_X \left(\frac{dP_\theta}{dP_{\theta_0}}\right)^t (x) dP_{\theta_0}(x) \ .$$

By the theory of the Laplace transform, $0 \le H(t,z) \le \infty$, $H(t,z)$ is strictly convex in t wherever it is finite and if $\inf_t H(t,z) < 1$, the infimum is obtained for a unique $t(z)$ given by the solution of the equation

$$z = \frac{\int \log \frac{dP_\theta}{dP_{\theta_0}}(x) \left[\frac{dP_\theta}{dP_{\theta_0}}(x)\right]^t dP_{\theta_0}(x)}{\int \left[\frac{dP_\theta}{dP_{\theta_0}}(x)\right]^t dP_{\theta_0}(x)} \ . \tag{2.13}$$

It is easily seen that if $z_0 = K(\theta, \theta_0)$, then $t(z_0) = 1$, and,

$$\inf_t \log H(t, z_0) = -K(\theta, \theta_0) \ . \tag{2.14}$$

From (2.10), (2.12) and (2.14) we can immediately conclude that $\liminf_n c_n \ge K(\theta, \theta_0)$. But, in fact, by the implicit function theorem as $z \to K(\theta, \theta_0)$ we have $t(z) \to 1$, and $\log H\big(z, t(z)\big) \to K(\theta, \theta_0)$, by dominated convergence. Choose $z_1 > z_0$ such that $H(z_1, t(z_1) > a \exp\{-K(\theta, \theta_0)\}$. Then $\varepsilon = z_1 - z_0$ will do for the Lemma.

It now follows from the basic assumption of the theorem, by a result of Erdös, Hsu and Robbins [9] that,

$$\sum_n P_{\theta_0}\Big[\sum_{i=1}^n \log \frac{dP_\theta}{dP_{\theta_0}}(x_i) \ge n c_n\Big] < \infty \ , \tag{2.15}$$

and this by (2.11) and the Borel Cantelli lemma suffices for (2.8) and the theorem to hold.

Remarks. 1. Erdös has shown in [9] that our second moment assumption is necessary as well as sufficient for (2.15) to hold. Although, of course, (2.15) is not necessary for (2.8) the relative arbitrariness of the A_n apart from condition (2.9) would suggest that the Theorem may be false if some condition such as the one imposed does not hold.

2. As in [4], if we define $N(\varepsilon, s)$ = least positive m such that $L_n \ge \varepsilon$ for all $n \ge m$ and ∞ otherwise, we have under the assumptions of our theorem 1,

$$\liminf_{\varepsilon \to 0} \frac{N(\varepsilon, s)}{-\log \varepsilon} \ge \frac{1}{J(\theta)} \text{ a.s. } P_\theta \ . \tag{2.16}$$

3. General Assumptions and a Useful Lemma

Before giving further structural assumptions needed for Sections 4 and 5 we prove a simple general lemma already implicit in [4] stating a useful sufficient condition for a sequence $\{T_n\}$ to be optimal. We say $\{T_n\}$ is asymptotically optimal if,

$$\lim_n \frac{1}{n} \log L_n(s) = -J(\theta) \tag{3.1}$$

with P_θ probability 1 for all $\theta \in \Theta - \Theta_0$. Then (3.1) implies, (cf. [4])

$$\lim_{\varepsilon \to 0} \frac{N(\varepsilon, s)}{-\log \varepsilon} = \frac{1}{J(\theta)} \text{ a.s. } P_\theta \ . \tag{3.2}$$

Lemma 2. *If the conclusion of Theorem 1 holds and*

$$i) \ \liminf_n T_n \geq J(\theta) \ a.s. \ P_\theta$$
$$ii) \ \limsup_n \log \big(1 - G_n(t)\big) \leq -t$$

where $G_n(t) = \inf \big\{F_n(t, \theta_0) : \ \theta_0 \in \Theta_0\big\}$ *and* θ *ranges over* $\Theta - \Theta_0$, *then* $\{T_n\}$ *is asymptotically optimal.*

Proof. It clearly suffices to show that

$$\limsup_n \frac{1}{n} \log L_n(s) \leq -J(\theta) \tag{3.3}$$

with P_θ probability 1. But since $L_n(s) = 1 - G_n(T_n)$ and $1 - G_n(t)$ is monotone decreasing, i) and ii) obviously imply (3.3).

We begin by giving nine general assumptions which are sufficient to ensure the validity of Theorem 2 of the main section.

Assumption 1. There exists a c finite measure μ on $(X, \mathcal{A})$ which dominates the family $\{P_\theta\}$. We denote the density of P_θ with respect to μ by $f(x, \theta)$. Then,

$$\frac{dP_\theta(x)}{dP_{\theta'}} = \frac{f(x, \theta)}{f(x, \theta')} \text{ a.e. } P_\theta + P_{\theta'} \ .$$

Assumption 2. Θ is a metric space. The topological Borel field on Θ is denoted by $\mathcal{B}$. $f(x, \theta)$ is bimeasurable in (x, θ) on $(X \times \Theta, \mathcal{A} \times \mathcal{B})$.

Assumption 3. We are given a probability measure ν on $(\Theta, \mathcal{B})$, $\Theta_0 \in \mathcal{B}$, and $\nu(\Theta_0) > 0$. Moreover, if $S(\theta, d)$ is the open sphere of centre θ and radius d, $\nu\big\{S(\theta, d) \cap [\Theta - \Theta_0]\big\} > 0$ for all $\theta \in \Theta - \Theta_0$ and $d > 0$.

Assumption 4. There exists a suitable metric compactification $\bar{\Theta}_0$ of Θ_0 (viz [4]). That is, we first define $\hat{S}(\theta, d)$ to be the sphere of radius d and centre θ in $\bar{\Theta}_0$ and then take,

$$g_0(x, \theta, d) = \sup \big\{f(x, \lambda) : \ \lambda \in \hat{S}(\theta, d) \cap \Theta_0\big\} \ . \tag{3.4}$$

We assume g_0 is measurable in x for d sufficiently small and define,

$$g_0(x, \theta, 0) = \lim_{d \to 0} g_0(x, \theta, d) \ . \tag{3.5}$$

The final assumption, (see [4]) is,

$$E_\theta \left(\frac{g(x, \theta', 0)}{f(x, \theta)} \right) \leq 1 \ , \tag{3.6}$$

where $E_\theta\big(h(x)\big)$ denotes $\int_X h(x)dP_\theta(x)$ for any integrable function h.

Assumption 5. Define for all $\theta' \in \bar{\Theta}_0$, $\theta \in \Theta - \Theta_0$,

$$\bar{K}(\theta,\theta') = -E_\theta\left(\frac{\log g_0(x,\theta',0)}{f(x,\theta)}\right) . \tag{3.7}$$

(3.6) and Jensen's inequality guarantee $0 \le \bar{K} \le \infty$. Assume,

$$J(\theta) = \inf\{K(\theta,\theta') : \theta' \subset \bar{\Theta}_0\} . \tag{3.8}$$

Assumption 6.

$$E_\theta\left(\log\frac{g_0(x,\theta',d)}{f(x,\theta)}\right) < \infty , \tag{3.9}$$

for all $\theta \in \Theta - \Theta_0$, $\theta' \in \Theta_0$. As in, [4] p.22, this is equivalent to,

$$\lim_{d\to 0} E_\theta\left(\frac{\log g_0(x,\theta',d)}{f(x,\theta)}\right) \le -K(\theta,\theta') . \tag{3.10}$$

Assumption 7. Define,

$$\eta(x,\theta,d) = \inf\left\{\log\frac{f(x,\lambda)}{f(x,\theta)} :\ \lambda \in S(\theta,d)\cap[\Theta-\Theta_0]\right\} . \tag{3.11}$$

Assume that η is a measurable function of x for d sufficiently small and that,

$$\lim_{d\to 0} E_\theta\big(\eta(x,\theta,d)\big) = 0 \tag{3.12}$$

for all $\theta \in \Theta - \Theta_0$.

Assumption 8. Define, for $\theta' \in \Theta_0$,

$$\gamma(x,\theta',d) = \log\inf\left\{\frac{f(x,\lambda)}{f(x,\theta')} :\ \lambda \in S(\theta',d)\cap\Theta_0\right\} , \tag{3.13}$$

$$\varphi(t,\theta',d) = E_{\theta'}\big(\exp\{-t\gamma(x,\theta',d)\}\big) . \tag{3.14}$$

For every $0 < \rho < 1$, $\beta > 0$, there exists $d(\theta',\rho,\beta)$ such that

$$\inf_t e^{-t\beta}\varphi\big(t,\theta',d(\theta',\rho,\beta)\big) \le \frac{\rho}{2} .$$

Assumption 9.

$$\inf\left\{\nu\big[S\big(\theta',d(\theta',\rho,\beta)\big)\cap\Theta_0 :\ \theta' \in \Theta_0\big]\right\} = m(\rho,\beta) > 0 .$$

We shall now examine these assumptions in turn giving where necessary stronger, but more easily checkable conditions, which we shall denote by primes. Thus, (say) assumption 4' will imply assumption 4 and the conclusion of theorem 2 will continue

to hold if 4 is replaced by 4'. The more important of these useful weakenings of Theorem 2 will be isolated as a Corollary.

Assumption 1 is self-explanatory and clearly cannot be weakened appreciably.

The requirement that Θ be a metric space can clearly be dropped and replaced by the requirement that Θ be a topological space and $\mathcal{B}$ the topological Borel field. However, the notational convenience involved in being able to define quantities in terms of spheres of a given radius rather than neighbourhood bases seems well worth the loss of generality. On the other hand, assumption 2 is obviously satisfied if we have the usual,

Assumption 2'. Θ is a subset of k dimensioinal Euclidean space with the usual metric topology. $\mathcal{B}$ is the Borel σ field and $f(x,\theta)$ is bimeasurable.

Weakenings of assumption 3 do not fit readily into this program, but we mention that we can drop the requirement that ν be a probability (finite) measure if the following two conditions hold, as well as the second part of assumption 3:

(1) There exists N such that, $\int_\Theta \prod_{i=1}^N f(x_i,\lambda)\nu(d\lambda) < \infty$ a.s. P_θ and
(2) $\int_{\Theta_0} \prod_{i=1}^N f(x_i,\lambda)\nu(d\lambda) > 0$, $\int_{\Theta-\Theta_0} \prod_{i=1}^N f(x_i,\lambda)\nu(d\lambda) > 0$ a.s. P_θ for all $n \geq N$.

For details of the proof of i) of Lemma 2 for $\bar{T}_n$ under these assumptions we refer to [6]. The basic idea of this generalization is to consider the process of observation as really starting after N with prior distribution, the posterior distribution of θ given $x_1,\ldots,x_N$, which by (1) is a true probability distribution. In fact, we can in general make dependent ν on the observations all along if we modify the second part of assumption 3 and assumption 9 suitably. The generalization given above is of interest in the case when reasonable tests arise from improper "priors", e.g. Lebesgue measure.

The most natural replacement of assumption 4 is of course assuming that Θ_0 is already compact. In this case, we have, **Assumption 4'.** Θ_0 is compact.

We can then drop 5, but must replace 6 by its equivalent form,

Assumption 6'. $\lim_{d\to 0} E_\theta \log\left(\frac{g_0(x,\theta',d)}{\rho(x,\theta)}\right) \leq -K(\theta,\theta')$ for all $\theta \in \Theta - \Theta_0$, $\theta' \in \Theta_0$.

Measurability of g_0 must still of course be invoked. Assumption 6 may replace 6' if $f(x,\theta)$ is continuous in θ for almost all x. A less stringent modification in some senses which is most useful is combining 4, 5, and 6 with 2' to give:

Assumption (4,5,6)" Assumption 2' holds and

(a) $\lim_{d\to 0} E_\theta\Big(\log \frac{g_0(x,\theta',d)}{f(x,\theta)}\Big) \leq -K(\theta,\theta')$ for $\theta \in \Theta - \Theta_0$, $\theta' \in \Theta_0$.
(b) $\lim_{d\to\infty} E_\theta\left(\log\sup\left\{\log\frac{f(x_\lambda,\lambda)}{f(x,\theta)} : \lambda \in \Theta_0 \| \lambda\| \geq d\right\}\right) \leq -J(\theta)$, for all $\theta \in \Theta - \Theta_0$, where $\|\cdot\|$ is the usual Euclidean norm.

This assumption is clearly equivalent to 5 and 6 if in 4 we take $\bar{\Theta}_0$ to be the closure of Θ_0 in the one point compactification of R.

Assumption 7 is most readily replaced by,

Assumption 7'. $f(x,\theta)$ is continous in θ for almost all $x(\mu)$ and, $E_\theta\big(\eta(x,\theta,d)\big) < \infty$ for some $d > 0$ for each $\theta \in \Theta - \Theta_0$.

Assumption 7' and the dominated convergence theorem readily imply 7. We need not require measurability of g_0 in this case in view of 4 or 2' since Θ_0 being a subset of a separable metric space is separable. The same is true of η if 2' holds or more generally if Θ is a separable metric space.

A useful substitute for assumption 8 is

Assumption 8'. There exists an $M < \infty$ such that for every $0 < T < \infty$ we can find a $d^x(\theta, T)$ with $\varphi(T, \theta', d^x) \leq M$. Assumption 8' easily implies 8 since, then,

$$\inf_t e^{-t\beta}\varphi(T, \theta', d^x) \leq e^{-T\beta}M < \frac{\rho}{2}\ , \tag{3.15}$$

for T sufficiently large.

In many situations 8' is most easily verified by showing that,

$$\varphi(t, \theta', d) \to 1\ , \tag{3.16}$$

uniformly on compacts in t as $d \to 0$. This in turn is implied by,

Assumption 8". $f(x, \theta)$ is continuous in θ for almost all $x(\mu)$ and for every $0 < T < \infty$, $\varphi(T, \theta', d) < \infty$ for d sufficiently small.

Assumption 8" implies (3.16) by way of the dominated convergence theorem if we remark that $\varphi(t, \theta', d)$ is monotone increasing in t for every fixed θ', d.

Finally, we can replace assumption 9 by,

Assumption 9'. $d(\theta, \rho, \beta)$ is independent of θ', assumption 4' holds, and $\nu\big[S(\theta', d) \cap \Theta_0\big] > 0$ for all $\theta' \in \Theta_0$.

To show that 9' implies 9 we need only prove that,

$$\inf\big\{\nu\big[S(\theta', d) \cap \Theta_0\big] :\ \theta' \in \Theta_0 f > 0\ . \tag{3.17}$$

Then, if $\theta'_n \to \theta'$ and, $S^*(\theta, d) = S(\theta, d) \cap \Theta_0$

$$\begin{aligned}\nu\big[S^*(\theta', d)\big] - \nu\big[S^*(\theta'_n, d)\big] &= \nu\big[\lambda :\ \delta(\lambda, \theta') < d, \quad \delta(\lambda, \theta'_n) \geq d,\ \lambda \in \Theta_0\big] \\ &\quad - \nu\big[\lambda :\ \delta(\lambda, \theta') \geq d, \quad \delta(\lambda, \theta'_n) < d,\ \lambda \in \Theta_0\big].\end{aligned} \tag{3.18}$$

Clearly, if $n \to \infty$, the first term of the above difference tends to 0 since the set whose measure is computed tends to the empty set. Therefore, for fixed d, $\nu\big[S^*(\theta', d)\big]$ is a lower semi-continuous function of θ' on Θ_0 and 2', and the third part of 9', imply (3.17).

This completes our roster of simplifying assumptions. Clearly if further restrictions are put on $f(x, \theta)$ the verification of most can be very easy. A strong form of Lemma 3, $\tau_n \to J(\theta)$, is given under very weak assumptions in [5] if $f(x, \theta)$ is of exponential type, (Theorem 4.1). If the conditions 1-7 of this paper are made two sided (viz. [5] Theorem 4.2) one can obtain this strengthening of Lemma 3 in general. In fact, the conditions detailed in theorem 4.2 of [5] are somewhat restrictive versions of our conditions 2'–7'.

The assumptions required by this paper but not by [4] are 7, 8, and 9. On the other hand, assumption 6 of that paper and the fact that Θ can be suitably compactified (rather than just Θ_0) is not required by us. Since simple hypotheses tend to be somewhat more common than simple alternatives this seems a gain. Otherwise the non structural assumptions of this paper and [4] coincide.

4. The Main Theorem

Given ν in assumption 3 we now define,

$$\bar{T}_n = \frac{1}{n}\log\frac{\int_{\Theta-\Theta_0}\prod_{i=1}^n f(x_i,\lambda)\ \nu(d\lambda)}{\int_{\Theta_0}\prod_{i=1}^n f(x_i,\lambda)\ \nu(d\lambda)}\ . \tag{4.1}$$

By assumption 2, $\bar{T}_n$ is well defined $\left(\frac{\infty}{\infty}=1,\ \frac{0}{0}=1\right)$. In fact, $\bar{T}_n$ is a version of the test statistic a Bayesian with prior ν would use to test $H:\theta\in\Theta_0$, rejecting for large values of $\bar{T}_n$. We can now state the principal theorem of the paper.

Theorem 2. *If assumptions 1–9 and the conclusion of Theorem 1 holds, then $\{\bar{T}_n\}$ is asymptotically optimal.*

Proof. The proof preceeds by way of some lemmas. We have first,

Lemma 3. *Under assumptions 1–7, $\{\bar{T}_n\}$ satisfies condition i) of Lemma 2.*

Proof. Suppose that $\theta\in\Theta-\Theta_0$ holds. Define,

$$U_n(s,\theta)=\int_{\Theta-\Theta_0}\prod_{i=1}^n\frac{f(x_i,\lambda)}{f(x_i,\theta)}\nu(d\lambda) \tag{4.2}$$

$$V_n(s,\theta)=\int_{\Theta_0}\prod_{i=1}^n\frac{f(x_i,\lambda)}{f(x_i,\theta)}\nu(d\lambda)\ . \tag{4.3}$$

Then,

$$\bar{T}_n=\frac{1}{n}\log\frac{U_n(s,\theta)}{V_n(s,\theta)}\ . \tag{4.4}$$

We show first,

$$\limsup_n\frac{1}{n}\log V_n(s,\theta)\le -J(\theta) \tag{4.5}$$

a.s. P_θ. Note that,

$$\frac{1}{n}\log V_n(s,\theta)\le\frac{1}{n}\log\left(\sup\left\{\frac{1}{n}\sum_{i=1}^n\log\frac{f(x_i,\lambda)}{f(x_i,\theta)}\ :\ \lambda\in\Theta_0\right\}\right)\ . \tag{4.6}$$

And the right hand side equals $R_0(\Theta_0,\theta)$ in the notation of [4] which converges to $-J(\theta)$ a.s. P_θ by Lemma 4 of that paper. An examination of the proof of this Lemma will show that only our assumptions 1–6 are used.

To establish the Lemma we need now only show,

$$\liminf_n\frac{1}{n}\log U_n(s,\theta)\ge 1 \text{ a.s. } P_\theta. \tag{4.7}$$

By assumption 7 we can find $d_1(\theta,\varepsilon)$ such that,

$$E_\theta\big(\eta(X_i,\theta,d_1)\big)\ge-\varepsilon\ . \tag{4.8}$$

But,

$$\frac{1}{n}\log U_n \geq \log \nu\big[S(\theta,d_1)\cap[\Theta-\Theta_0]\big] + \inf\Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i,\lambda)}{f(X_i,\theta)} \;:\; \lambda\in S(\theta,d_1)\cap[\Theta-\Theta_0]\Big\}$$

$$\geq \frac{1}{n}\sum_{i=1}^{n}\eta(x_1,\theta,d_1)+\log\nu\big[S(\theta,d_1)\cap[\Theta-\Theta_0]\big]\ . \tag{4.9}$$

Letting $n\to\infty$ and then $\varepsilon\to 0$, (4.9), assumption 3, and the strong law of large numbers imply (4.7). The Lemma follows.

We complete the proof of the Theorem by way of two further Lemmas.

Lemma 4. *Under the first part of assumption 3, for all* $n,t,\theta'\in\Theta_0$,

$$P_{\theta'}\big[\frac{1}{n}\log U_n(s,\theta')\geq t\big]\leq e^{-nt}. \tag{4.10}$$

Proof.

$$P_{\theta'}\big[\frac{1}{n}\log U_n(s,\theta')\geq t\big]=\int_W\prod_{i=1}^{n}f(x_i,\theta')\ \mu(dx_1)\dots\mu(dx_n)$$

$$\text{where } W=\Big[s:\prod_{i=1}^{n}f(x_i,\theta')\leq e^{-nt}\int_{\Theta-\Theta_0}\prod_{i=1}^{n}f(x_i,\lambda)\nu(d\lambda)\Big]\ .$$

Thus our probability is bounded above by

$$e^{-nt}\int_{X^n}\int_{\Theta-\Theta_0}\prod_{i=1}^{n}f(x_i,\lambda)\nu(d\lambda)\ \mu(dx_1)\dots\mu(dx_n)\ . \tag{4.11}$$

But the right hand side of (4.11), by Fubini's theorem, is

$$e^{-nt}\int_{\Theta-\Theta_0}\int_{X^n}\prod_{i=1}^{n}f(x_i,\lambda)\ \mu(dx_1)\dots\mu(dx_n)\nu(d\lambda)\leq e^{-nt}\ .$$

Lemma 5. *Under assumptions 8 and 9 if* $\theta'\in\Theta_0$, *for every* $0<\rho<1$, $\beta>0$, *there exist* $N(\rho,\beta)$ *such that for* $n\geq N(\rho,\beta)$

$$P_{\theta'}\big[\frac{1}{n}\log V_n(s,\theta')\geq-\beta\big]\leq\rho^n\ . \tag{4.12}$$

Proof. Choose $d(\theta',\rho,\beta)$ as in assumptions 8 and 9. Then since

$$\frac{1}{n}\log V_n(S,\theta')\geq\log\nu\big[S(\theta',d(\theta',\rho,\beta)\cap\Theta_0\big] + \inf\Big\{\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(x_i,\lambda)}{f(x_i,\theta')} \;:\; \lambda\in S(\theta',d(\theta',\rho,\beta)\cap\Theta_0\Big\}$$

$$\geq\log\nu\Big[S\big(\theta',d(\theta',\rho,\beta)\big)\cap\Theta_0\Big]+\frac{1}{n}\sum_{i=1}^{n}\gamma\big(x_i,\theta',d(\theta',\rho,\beta)\big)\ , \tag{4.13}$$

we have,

$$P_{\theta'}\big[\frac{1}{n}\log V_n(s,\theta') \le -\beta\big] \le P_{\theta'}\big[\sum_{i=1}^{n}\gamma\big(x_i,\theta',d(\theta',\rho,\beta)\big) \le -n\beta - \log m(\rho,\beta)\big]\ . \tag{4.14}$$

By Lemma 1 of [4]

$$\begin{aligned} &P_{\theta'}\big[\frac{1}{n}\sum_{i=1}^{n}\gamma\big(x_i,\theta',d(\theta',\rho,\beta)\big) \le -n\beta - \log m(\rho,\beta)\big] \\ &\qquad \le \inf_t\Big\{\exp\big[-\beta t - \log m(\rho,\beta)\frac{t}{n}\big]\varphi\big(t,\theta',d(\theta',\rho,\beta)\big)\Big\}^n \end{aligned} \tag{4.15}$$

for $n \ge -\frac{\log m(\rho,\beta)}{\beta}$ which is finite by assumption 9. If $\inf_t e^{-\beta t}\varphi\big(t,\theta',d(\theta',\rho,\beta)\big)$ is attained for $t = t_0$ which is strictly positive by lemma 1 of [4], we can choose,

$$N(\rho,\beta) \le \max\Big(-\frac{\log m(\rho,\beta)}{\beta}, -\frac{t_0\log m(\rho,\beta)}{\log 2}\Big) < \infty\ .$$

Now,

$$\begin{aligned} P_{\theta'}\big[\frac{1}{n}\log\bar{T}_n \ge t\big] &\le P_{\theta'}\big[\frac{1}{n}\log U_n(s,\theta') \ge -\beta, \frac{1}{n}\log V_n(s,\theta') \ge -\beta\big] \\ &\quad + P_{\theta'}\big[\frac{1}{n}\log V_n(s,\theta') \le -\beta\big]\ . \end{aligned} \tag{4.16}$$

By Lemmas 4 and 5 the right hand side of (4.16) is bounded above by $e^{-n(t-\beta)} + \rho^n$ for $n \ge N(\rho,\beta)$. Hence,

$$\begin{aligned} \limsup_n \frac{1}{n}\log\sup\big\{P_{\theta'}[\bar{T}_n \ge t] :\ \theta \in \Theta_0\big\} &\le \limsup_n \frac{1}{n}\log\big(e^{-n(t-\beta)} + \rho^n\big) \\ &= \max[-(t-\beta), \log\rho]\ . \end{aligned} \tag{4.17}$$

Letting $\rho \to 0$ first and then $\beta \to 0$, we find that ii) of Lemma 2 is satisfied by $\bar{T}_n$ and the Theorem is proved. Gathering the most useful of the "prime" assumptions together we state,

Corollary 1. *If assumptions 1, 2', 3, (456)'', 7', 8", and 9 hold, then the conclusion of Theorem 2 is valid.*

The most immediate field of application of Corollary 1 is when $f(x,\theta)$ is the density of a Koopman-Darmois (exponential) family.

$$f(x,\theta) = e^{\theta.t(x)} \tag{4.18}$$

where $\theta = (\theta_1,\ldots,\theta_k)$, $t(x) = \big(t_1(x),\ldots,t_k(x)\big)$ and the θ_j, t_j are real. Assumptions 1, 2', (456)'', 7' and 8" are then automatically satisfied and we need only impose conditions 3, and 9 on ν. If Θ_0 is compact, 9' is automatic and 3 is all that is needed.

5. Optimality of Minimax Tests

The main result of this section is an immediate consequence of the following Lemma. We retain the notation of the previous section, defining only

$$\bar{F}_n(t,\theta) = P_\theta[\bar{T}_n < t] \ . \tag{5.1}$$

$$\bar{G}_n(t) = \inf\left\{\bar{F}_n(t,\theta') : \ \theta' \in \Theta_0\right\} \ . \tag{5.2}$$

Lemma 6. *Suppose that there exists a measurable subset S of Θ_0 such that,*
iii) $1-\bar{F}_n(t,\theta')$ *is a constant on S (for fixed n, t) as a function of θ' .*
iv) $\sup\left\{1-\bar{F}_n(t,\theta') : \ \theta' \in S\right\} = 1-\bar{G}_n(t)$.
v) $\nu[\Theta_0 - S] = 0$.
Then,

$$\limsup_n \frac{1}{n}\log\left(1-\bar{G}_n(t)\right) \le -t \ . \tag{5.3}$$

Proof.

$$\left(1-\bar{G}_n(t)\right) = \sup\left\{1-\bar{F}_n(t,\theta') : \ \theta' \in S\right\} = \int_{\Theta_0} P_{\theta'}[\bar{T}_n \ge t]\nu(d\theta') \tag{5.4}$$

by iii), iv), v). Now, let

$$C_n = \Big[\int_{\Theta_0}\prod_{i=1}^n f(x_i,\lambda)\nu(d\lambda) \le e^{-nt}\int_{\Theta-\Theta_0}\prod_{i=1}^n f(x_i,\lambda)\nu(d\lambda)\Big] \ .$$

Then,

$$\begin{aligned}\int_{\Theta_0} P_{\theta'}[\bar{T}_n < t]\nu(d\theta') &= \int_{\Theta_0}\int_{C_n}\prod_{i=1}^n f(x_i,\lambda)\mu(dx_1)\dots\mu(dx_n)\nu(d\lambda)\\ &= \int_{C_n}\int_{\Theta_0}\prod_{i=1}^n f(x_i,\lambda)\mu(dx_1)\dots\mu(dx_n)\nu(d\lambda)\\ &\le e^{-nt}\int_{C_n}\int_{\Theta-\Theta_0}\prod_{i=1}^n f(x_i,\lambda)\nu(d\lambda)\mu(dx_1)\dots\mu(dx_n) \le e^{-nt} \ . \end{aligned} \tag{5.5}$$

We formulate,

Assumption 10. For every $n,\ t$

a) $\nu\Big[\theta\in\Theta-\Theta_0 : \ \sup\left\{\bar{F}_n(t,\lambda) : \ \lambda\in\Theta-\Theta_0\right\} > \bar{F}_n(t,\theta)\Big] = 0.$
b) $\nu\Big[\theta' : \ \bar{G}_n(t) < F_n(t,\theta') : \ \theta'\in\Theta_0\Big] = 0.$

We can now state,

Theorem 3. *If assumption 10 holds the test which rejects if $\bar{T}_n \ge t$ is minimax for $H:\theta\in\Theta_0$ vs $K:\theta\in\Theta-\Theta_0$ at level $1-\bar{G}_n(t)$. If assumptions 1–7 and assumption 10b) hold, then the sequence of test statistics $\{\bar{T}_n\}$ is asymptotically optimal.*

Proof. The first part of the Theorem is classical (c.f. Lehmann [8] p.327). The second part is an immediate consequence of Theorem 2 and Lemma 5 since assumption 10b) is equivalent to iii), iv), v).

This Theorem is of interest in connecting the classical finite sample optimality results with stochastic comparison. The most immediate application of this result is in the one-parameter exponential family where minimax tests of (say) $H : \theta \leq \theta_0$ vs $K : \theta = \theta_1 > \theta$ are. Bayes tests with respect to two point distributions satisfy assumption 10 (cf. [8]). Unfortunately proving optimality directly is trivial in this case. More interesting candidates are in the normal situation the t-statistic and the S^2 statistic used in testing $H : \mu < \mu_0$ and $H : \sigma \leq \sigma_0$ when μ, and σ are respectively unknown. Although we are here presented with a situation which does not quite fall under Theorem 3, (ν satisfying 10a,b depends on n, (cf. [8], p.94), one can easily check the conclusion of Lemma 2 directly and then apply Lemma 5 to obtain optimality.

Finally, it may be interesting to see that from a quasi Bayesian point of view, if stochastic comparison is defined in terms of the observed expected level of significance, (where the expectation is taken under the prior) then Lemma 5 and assumptions 1–7 and an analogue of Theorem 1 guarantee the optimality of $\bar{T}_n$ in this sense. Formally for any sequence $\{T_n\}$ we would then consider not L_n but,

$$L_n^*(s) = \big(1 - G_n^*(T_n)\big)$$

where $G_n^*(t) = \int_{\Theta_0} F_n(t, \theta')\nu(d\theta')$.

The analogue of the conclusion of Theorem 1 needed would be that a.s. P_θ,

$$\liminf_n \frac{1}{n} \log L_n^* \geq -J(\theta) \ .$$

References

[1] Bahadur, R. R. (1960). Stochastic comparison of tests. *Ann. Math. Stat.*, **31** 276–295.
[2] Bahadur, R. R. (1960). Asymptotic efficiency of tests and estimates. *Sankhya*, **22** 229–252.
[3] Bahadur, R. R. (1960). Simultaneous comparison of the optimum and sign tests of a normal mean. In *Contributions to Probability and Statistics*. Stanford Univ. Press, 79–88.
[4] Bahadur, R. R. (1965). An optimal property of the likelihood ratio statistic. In *Proceedings of the 5^{th} Berkeley Symposium*, **1**, 13–26.
[5] Bickel, P. J. and Yahav, J. A. (1965). Asymptotically pointwise optimal procedures in sequential analysis. In *Proceedings of the 5^{th} Berkeley Symposium*, **1**, 401–415.
[6] Bickel, P. J. and Yahav, J. A. (1969). Asymptotic theory of Bayes solution. *Prob. Theory and Related Fields*, **11**, 257–276.
[7] Chernoff, H. (1952). A measure of asymptotic efficiency for tests of hypothesis based on the sum of observations. *Ann. Math. Stat.*, **23**, 493–507.
[8] Erdös, P. (1949). On a theorem of Hsu and Robbins. *Ann. Math. Stat.*, **20**, 286–291.
[9] Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. J. Wiley and Sons.

IMS Collections
Vol. 0 (2009) 31–45
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000006

On the Non-Optimality of Optimal Procedures

Peter J. Huber[1]

Abstract: This paper discusses some subtle, and largely overlooked, differences between conceptual and mathematical optimization goals in statistics, and illustrates them by examples.

Contents

1 Introduction 31
2 On Optimization and Models 32
3 Classical Mathematical Statistics and Decision Theory 32
4 Tukey's 1962 Paper 32
5 Pitfalls of Optimality 33
6 Examples from Classical Statistics 34
6.1 The Fuzzy Concepts Syndrome 34
6.2 The Straitjacket Syndrome 36
6.3 The Scapegoat Syndrome 36
7 Problems with Optimality in Robustness 37
7.1 Asymptotic Robustness for Finite $\varepsilon > 0$ 37
7.2 Finite Sample Robustness for Finite $\varepsilon > 0$ 38
7.3 Asymptotic Robustness for Infinitesimal ε 38
7.4 Optimal Breakdown Point 39
8 Design Issues 40
8.1 Optimal Designs 40
8.2 Regression Design and Breakdown 41
9 Bayesian Statistics 42
10 Concluding Remarks 44
References 44

1. Introduction

First, we shall identify those parts of statistics that rely in a crucial fashion on optimization. The most conspicuous among them are: classical mathematical statistics, decision theory, and Bayesian statistics.

Classical mathematical statistics was created by R. A. Fisher (1922), in a paper concerned with estimation, and by J. Neyman and E. S. Pearson (1933), in a paper concerned with testing. It was brought to completion by E. L. Lehmann in his lecture notes (1949, 1950); those notes later grew into two books (1959, 1983). Around the same time when Lehmann produced his lecture notes, A. Wald (1950) expanded the scope of mathematical statistics by creating statistical decision theory.

[1]POB 198, 7250 Klosters, Switzerland, email: peterj.huber@bluewin.ch
Keywords and phrases: optimality, superefficiency, optimal robustness, breakdown point, optimal design, Bayesian robustness

The central concerns of classical mathematical statistics were *efficiency* in estimation (i.e. minimum variance), and *power* in testing problems, both being optimality concerns. Decision theory confirmed the central interest in optimality, but shifted the emphasis to *admissibility* and *minimaxity*.

A heavy slant towards optimality, of a different origin, holds also for Bayesian statistics. For a given model, consisting of a prior distribution and a family of conditional distributions, the Bayes formula by definition gives the "best" procedure; it is admissible in decision theoretic terminology.

The above-mentioned three areas of statistics appear to be the only ones where optimality is central to the theory. Elsewhere, optimality seems to provide mere icing on the cake. Note that the papers of Fisher and of Neyman-Pearson imprinted subsequent generations of statisticians with an (often uncritical) love of optimality. By 1960, as a young mathematical statistician you would not dare submit a new procedure to a reputable journal, unless you could prove some optimality property. (Later, there was a reversal, and too many statistical algorithms may have slipped through the editorial gates with enthusiastic but inadequately substantiated claims.)

2. On Optimization and Models

Mathematical optimization always operates on some model. Models are simplified approximations to the truth; the hope is that optimality at the model translates into approximate optimality at the true situation. In the sciences, the main purpose of models is different: they are to assist our conceptual understanding, and to help with communication of ideas.

In traditional statistics there is no methodology for assessing the *adequacy* of a model. At best, traditional statistics can reject a model through a goodness-of-fit test — and Bayesian statistics cannot even do that. A non-rejected model is not necessarily adequate, and even more embarrassing, a rejected model sometimes may provide a perfectly adequate approximation.

3. Classical Mathematical Statistics and Decision Theory

Classical mathematical statistics provides a clean theory under very restrictive assumptions, such as restricting the underlying models to exponential families, or the procedures to unbiasedness or invariance.

Decision theory clarified the classical views and reduced the dependence on restrictions. It also opened new areas, in particular optimal design theory (Kiefer 1959). But on the whole, decision theory was less successful than originally hoped. The two principal success stories are the Stein estimates (James and Stein 1961), relating to admissibility, and robustness (Huber 1964), relating to minimaxity.

4. Tukey's 1962 Paper

In his long 1962 paper "The Future of Data Analysis", while ostensibly talking about his personal predilections, Tukey actually redefined the field of statistics. Tukey's central theme was his emphasis on *judgment* (Section 7). At the same time, he played down the importance of mathematical rigor and optimality (Sections 5 and 6). Possibly the most important issue worked out in his long and multifaceted paper was that there is more to theoretical statistics than mathematical

statistics. This reminds one of Clausewitz (1832), who castigated the theorists of military strategy of his time because they "considered only factors that could be mathematically calculated".

In his paper, Tukey eschewed models. Why? Perhaps because in traditional statistics models erroneously are considered as *substitutes* for the truth, rather than as *simplified approximations*. Note in particular his quote of Martin Wilk at the end of Section 4: "The hallmark of good science is that it uses models and 'theory' but never believes them".

Tukey of course was not the first to question the role of models and of optimality. Statistical methods based on ranks and nonparametrics had become popular precisely because they avoided dependence on uncertain models and were valid under weaker assumptions, even if they lacked the flexibility and wide applicability of the parametric approaches.

But the problems with models and optimality go deeper. They have less to do with the idealized models *per se*, but more with the procedures optimized for them.

5. Pitfalls of Optimality

There are four basic pitfalls, into which mathematically optimal procedures can be trapped:

(i) the Fuzzy Concepts Syndrome:
sloppy translation of concepts into mathematics,
(ii) the Straitjacket Syndrome:
overly restrictive side conditions,
(iii) the Scapegoat Syndrome:
confuse the model with the truth,
(iv) the Souped-Up Car Syndrome:
optimize speed and produce a delicate gas-guzzler.

These pitfalls affect distinct, very different aspects of statistics, namely: (i) concepts, (ii) procedures, (iii) models, and (iv) target functions. The list of course is not exhaustive. The pitfalls shall be discussed with the help of ten examples:

Classical:
(1) superefficiency
(2) unbiasedness, equivariance
(3) efficiency at the model
Robustness:
(4) asymptotics for finite ε
(5) finite n, finite ε
(6) asymptotics for infinitesimal ε
(7) optimal breakdown point
Design:
(8) optimal designs
(9) regression design and breakdown
Bayesian statistics:
(10) Bayesian robustness

6. Examples from Classical Statistics

The three "classical" examples (1)-(3) neatly illustrate the first three pitfalls.

6.1. The Fuzzy Concepts Syndrome

Problems caused by the Fuzzy Concepts Syndrome mostly are relics from earlier development stages of statistical theories. In a conference on Directions for Mathematical Statistics, I had argued (Huber 1975b): "In statistics as well as in any other field of applied mathematics [...] one can usually distinguish (at least) three phases in the development of a problem. In Phase One, there is a vague awareness of an area of open problems, one develops *ad hoc* solutions to poorly posed questions, and one gropes for the proper concepts. In Phase Two, the right concepts are found [...]. In Phase Three, the theory begins to have a life of its own, [...] and its boundaries of validity are explored by leading it *ad absurdum*; in short, it is squeezed dry." In the 1970s there had been widespread anxiety about the future of mathematical statistics. As a deeper reason for this anxiety I had proposed the diagnosis that too many of the then current activities belonged to the later phases of Stage Three.

In the groping phase, somewhat reckless heuristics can be beneficial. The concepts inevitably are fuzzy, and correspondingly, they are translated into mathematics in a sloppy fashion. But recklessness, fuzziness and sloppiness should be cut down at the latest at the beginning of the squeezing phase (the "consolidation phase", in Erich Lehmann's terminology). Though, in the later phases it is tempting to concentrate on the mathematical formalism and to neglect a re-examination of its conceptual origins. And admittedly, even in the mathematical formalism, any attempts to eliminate sloppiness in its entirety will lead to an admirable, but non-teachable theory, as already Whitehead and Russell with their monumental *Principia Mathematica* (1910–13) have demonstrated.

In mathematical statistics, asymptotics is exceptionally prone to sloppiness. Details notoriously are not adequately elaborated. Indeed, the expression "asymptotic theory" itself is used misleadingly. In standard mathematical usage asymptotic theory ordinarily is concerned with asymptotic expansions. Statistics knows such expansions too (e.g. Edgeworth expansions), but mostly, "asymptotic theory" denotes what more properly should be called "limiting theory". A few examples follow.

- Remainder terms? With asymptotic expansions, the first neglected term gives an indication of the size of the error. In statistics, asymptotic results hardly ever are complemented by remainder terms, however crude. That is, whatever the actual sample size is, we never know whether an asymptotic result is applicable.
- What kind of asymptotics is appropriate? In regression, for example, we have n observations and p parameters. Should the asymptotics be for p fixed, $n \to \infty$, or for $p/n \to 0$, or for what?
- Order of quantifiers and limits? Usually, one settles on an order that makes proofs easy.

Example 1. Perhaps the most illuminating case of the Fuzzy Concepts Syndrome has to do with superefficiency. There is a famous pathological example due to Hodges (see LeCam 1953). Assume that the observations $(x_1, \ldots, x_n)$ are i.i.d. nor-

mal $\mathcal{N}(\theta, 1)$. Estimate θ by

$$T_n = \bar{x}, \quad \text{if } |\bar{x}| \geq n^{-1/4}$$
$$T_n = \bar{x}/2, \text{ if } |\bar{x}| < n^{-1/4}.$$

Then T_n is consistent for all θ, with asymptotic variance n^{-1} for $\theta \neq 0$, but $\frac{1}{4}n^{-1}$ for $\theta = 0$. That is, the estimate T_n is efficient everywhere, but superefficient at 0. See Lehmann (1983, p. 405–408) for a discussion of various responses to the unpleasantness caused by Hodges' example.

Informally, asymptotic efficiency means that in large samples the variance of the estimate approaches the information bound, and this for all θ. Everyday language is notoriously ambiguous about the order of the quantifiers. For example we may spell out asymptotic efficiency as:

$$(\forall \varepsilon > 0)\ (\forall \theta)\ (\exists n_0)\ (\forall n > n_0)\ \{T_n \text{ is } \varepsilon\text{-efficent}\}, \tag{1}$$

where we define ε-efficiency by, say,

$$\{T_n \text{ is } \varepsilon\text{-efficent}\} \ = \ \left\{E_\theta \left(n(T_n - \theta)^2\right) < 1/I(\theta) + \varepsilon\right\}. \tag{2}$$

But then, for any fixed n, T_n might be arbitrarily bad for some θ. Therefore, since we do not know the true value of θ, we never will know whether an estimate satisfying (1) is any good, however large n is. In other words: while the definition of asymptotic efficiency may be technically in order, it is conceptually inacceptable.

On closer inspection we conclude that the order of quantifiers in (1) does not correspond to our intuitive concept of asymptotic efficiency. An improved version is obtained by interchanging the second and third quantifiers:

$$(\forall \varepsilon > 0)\ (\exists n_0)\ (\forall \theta)\ (\forall n > n_0)\ \{T_n \text{ is } \varepsilon\text{-efficent}\}. \tag{3}$$

It turns out that this version excludes superefficiency.

But version (3) still is negligent. Conceptually, unbounded loss functions are unsatisfactory. Technically, the awkward fact is that for very long-tailed distributions, the expectation in (2) may fail to be finite for all n and all "reasonable" estimators (i.e. for all estimators T_n whose value is contained in the convex hull of the observations, cf. Huber 1972, p. 1047), while the limiting distribution exists and has a finite variance. To obtain a definition of asymptotic efficiency working more generally we might rewrite (3) to

(4)

$$(\forall c > 0)(\forall \varepsilon > 0)(\exists n_0)(\forall \theta)(\forall n > n_0)\left\{E_\theta\left(\left(\left[\sqrt{n}(T_n - \theta)\right]_{-c}^{+c}\right)^2\right) < 1/I(\theta) + \varepsilon\right\}.$$

Here, $[x]_a^b = max(a, min(b, x))$. Of course, (4) is not yet the final word; for example, we might want to replace the global uniform bound by a local one.

In my opinion Hodges' example should not be considered as a local improvement of the standard estimate, comparable to the James-Stein estimate, but rather as an ingenious spotlight on a conceptual inadequacy of the traditional formalization of asymptotic efficiency. This interpretation is not new. In particular, the crucial technical result, namely that one-sided locally uniform bounds suffice to prevent superefficiency, had been published in an abstract more than 40 years ago (Huber 1966). But I never had found a congenial outlet for the philosophical side of the result; it took the present symposium to provide one.

6.2. The Straitjacket Syndrome

Example 2. Classical examples of the Straitjacket Syndrome, that is of overly restrictive side conditions on the procedures, are well known and do not need a detailed discussion. One is furnished by unbiasedness: unbiased estimates may not exist, or they may be nonsensical, cf. Lehmann (1983, p. 114). Other examples occur with invariance (more properly: equivariance): equivariant estimates may be inadmissible (Stein estimation).

6.3. The Scapegoat Syndrome

This subsection is concerned with excessive reliance on idealized models. The word "scapegoat" refers to the pious belief that the gods of statistics will accept the model as a substitute for the real thing.

As statisticians, we should always remember that models are simplified approximations to the truth, not the truth itself. Sometimes they are not even that, namely when they are chosen for ease of handling rather than for adequacy of representation; typical examples are conjugate priors in Bayesian statistics. The following eye-opening example gave rise to robustness theory.

Example 3. In 1914, Eddington had advocated the use of mean absolute deviations, against root-mean-square (RMS) deviations, as estimates of scale. Fisher (1920) objected and showed that for normal errors RMS deviations are 12% more efficient. Tukey (1960) then pointed out that for the contaminated normal error model

$$F(x) = (1-\varepsilon)\Phi(x/\sigma) + \varepsilon\Phi(x/(3\sigma)) \tag{5}$$

mean absolute deviations are more efficient for all $0.002 < \varepsilon < 0.5$.

The unfortunate fact is that errors in real data typically are better approximated by a contamination model with a contamination rate ("gross error rate") in the range $0.01 < \varepsilon < 0.1$, than by the normal model.

The main lesson to be learnt from the Eddington–Fisher–Tukey example is that the standard normal error model may be quite accurate, especially in the center of the distribution. The problem is that the tail behavior of real data, to which the traditional estimates are highly sensitive, usually is rather indeterminate and difficult to model. The mistake of Fisher (and others) had been to treat the standard model as the exact truth.

We note a few conclusions from such examples:

- Optimality results put in evidence what can (and what cannot) be achieved in an *ideal* world.
- Notoriously, optimal procedures are unstable under *small deviations* from the ideal situation.
- The task thus is to find procedures that achieve *near optimality* under the ideal situation, but that are more *stable* under small deviation.

In 1964, I had begun to implement a program suggested by this under the name of robustness. The guiding ideas were:

– Keep the optimality criterion (asymptotic variance, ...).
– Formalize small deviations (ε-contamination, ...).

– Find best sub-optimal procedures (best in a minimax sense).

The robustness notion I had adopted corresponds to Tukey's 1960 version. Though, this is not the unique interpretation of robustness occurring in the literature. In the 1970's, under Tukey's influence, there was a semantic shift, adopted by many, namely that the purpose of robustness was to provide procedures with a strong performance for a *widest possible* selection of heavy-tailed distribution.

But I still prefer the original 1960 version. In particular, I hold that robustness should be classified with *parametric* procedures, and that *local stability* in a neighborhood of the parametric model is the basic, overriding requirement.

7. Problems with Optimality in Robustness

Robustness had been designed to safeguard against pitfalls of optimal procedures. But optimal robustness is vulnerable to the very same pitfalls, and there are even a few new variants. The conceptual problem mentioned below in Example 4, and its solution described in Example 5, both have received less resonance in the robustness literature than they would have deserved. While the influence function without doubt is the most useful heuristic tool of robustness, one ought to be aware that optimality results based on it are no better than heuristic (Example 6).

7.1. Asymptotic Robustness for Finite $\varepsilon > 0$

Example 4. In the decision theoretic formalization of my 1964 paper I had imposed an unpleasant restriction on Nature by allowing only symmetric contaminations. The reason for this was that asymmetric contamination causes a bias term of the order $O(1)$. Asymptotically, this bias then would overpower the random variability of the estimates (which typically is of the order $O(n^{-1/2})$). Automatically, this would have led to the relatively inefficient sample median as the asymptotically optimal estimate. On the other hand, for the sample sizes and contamination rates of practical interest, the random variability usually is more important. Simultaneously, the symmetry assumption had permitted to extend the parameterization to the entire ε-neighborhood and thereby had made it possible to maintain a standard point-estimation approach.

The assumption of exact symmetry is repugnant, it violates the very spirit of robustness. Though, restrictions on the distributions are much less serious straitjackets than restrictions on the procedures (such as unbiasedness). The reason is that after performing optimization under symmetry restrictions, one merely has to check that the resulting asymptotically "optimal" estimate remains nearly optimal under more realistic asymmetric contaminations, see Huber (1981, pp. 104–106).

Curiously, people have worried (and still continue to worry!) much more about the symmetry straitjacket than about a conceptually much more serious problem. That problem is that 1% contamination has entirely different effects in samples of size 10 or 1000. Thus, asymptotic optimality theory need not be relevant at all for modest sample sizes and contamination rates, where the expected number of contaminants is small and may fall below 1. Fortunately, this question could be settled through an exact finite sample theory – see the following example. This theory also put to rest the problem of asymmetric contamination.

7.2. Finite Sample Robustness for Finite $\varepsilon > 0$

Example 5. To resolve the just mentioned conceptual problem, one needs a finite sample robustness theory valid for finite $\varepsilon > 0$. Rigorous such theories were developed early on, see Huber (1965) for tests and Huber (1968) for estimation. The latter covers the same ground as the original asymptotic robustness theory, namely single parameter equivariant robust estimation. Gratifyingly, it leads to procedures that are qualitatively and even quantitatively comparable to the M-estimators obtained with the asymptotic approach.

This finite sample approach to robustness does not make any symmetry assumptions and thus also avoids the other objections that have been raised against asymptotic robustness theory. In particular, by aiming not for point estimates, but for minimax interval estimates, it bypasses the parameterization and asymmetry problems. Despite its conceptual importance, the finite sample theory has attained much less visibility than its asymptotic and infinitesimal cousins. I suspect the reason is that the approach through an unconventional version of interval estimates did not fit into established patterns of thought. In the following I shall sketch the main ideas and results; for technical details see Huber (1968).

Just as in the original asymptotic theory, we consider the one-parameter location problem and assume that the error distribution is contained in an ε-neigborhood of the standard normal distribution. The optimally robust finite sample estimator turns out to be an M-estimate T defined by

$$\sum \psi(x_i - T) = 0, \tag{6}$$

where $\psi(x) = [x]_{-k}^{k} = max(-k, min(k, x))$ for some $k > 0$. But instead of minimizing the maximal asymptotic variance, this estimator is optimal in the sense that it minimizes the value α for which one can guarantee

$$P\{T < \theta - a\} \le \alpha, \quad P\{T > \theta + a\} \le \alpha \tag{7}$$

for all θ and all distributions in the ε-neighborhood.

We have three free parameters, n, ε and a. Interestingly, the characteristic parameter k of the ψ-function depends only on ε and a, but not on the sample size n. In (7), instead of minimizing α for fixed a, we might alternatively minimize a for fixed α. The asymptotic theory can be linked to these exact finite sample optimality results in several different fashions. In particular, if we let $n \to \infty$, but keep both α and k fixed, then a and ε of the optimally robust estimates go to 0 at the rate $O(n^{-1/2})$. Conceptually, ε-neighborhoods shrinking at a rate $O(n^{-1/2})$ make eminent sense, since the standard goodness-of-fit tests are just able to detect deviations of this order. Larger deviations should be taken care of by diagnostics and modeling, while smaller ones are difficult to detect and should be covered (in the insurance sense) by robustness.

7.3. Asymptotic Robustness for Infinitesimal ε

Example 6. Parametric families more general than location and scale are beyond the scope of the above approaches to robustness. Hampel proposed to attack them via gross error sensitivity: minimize asymptotic variance at the model, subject to a bound on the influence function (see Hampel 1974, and Hampel *et al.* 1986). This

approach is infinitesimal in nature and stays strictly at the parametric model. In essence, it is concerned only with the limiting case $\varepsilon = 0$.

Heuristically, it combines two desirable properties of robust estimates: good efficiency at the model, and low gross error sensitivity. However, a bound on the latter *at the model* does not guarantee robustness (local stability in a neighborhood of the model), there are counter examples with L-estimates, see Huber (1981, pp. 71–72). Thus, the conceptual basis of this approach is weak. Even if it should yield robust procedures, we have no guarantee that they are approximately optimal for non-zero ε, and we thus have to pray to the statistical gods that they will accept an infinitesimal scapegoat. As a minimum, one ought to check the breakdown point of procedures constructed by this method.

There is a conceptually more satisfactory, but technically more complicated alternative approach via shrinking neighborhoods: while $n \to \infty$, let $\varepsilon \to 0$ at the rate $O(n^{-1/2})$. This particular asymptotic theory had been motivated by the finite sample approach of Example 5. It was introduced by C. Huber-Carol in her thesis (1970) and later exploited by Rieder in several papers, culminating in his book (1994). The limiting results are comparable to those obtained with the infinitesimal approach, and like these, in the location parameter case they agree with those obtained in Example 4.

The principal heuristic appeal of the shrinking neighborhood approach is that in the location case it yields a sequence of estimates that have a well-defined optimality property for each n. We therefore can hope that in the general case it yields a sequence of estimates that are approximately optimal for non-zero ε. But to be honest, we have no way to check whether the heuristic arguments reliably carry beyond the location case. That is, we may run into a fifth pitfall: overly optimistic heuristics.

7.4. Optimal Breakdown Point

Hampel, at that time a student of Erich Lehmann, in his Ph.D. thesis (1968) had introduced the breakdown point by giving it an asymptotic definition. Conceptually, this may have been misleading, since the notion is most useful in small sample situations, see Donoho and Huber (1983). With large samples and high contamination rates you may have enough data to interpret the information contained in the contamination part. Therefore, rather than blindly using high breakdown point procedures, you may spend your efforts more profitably on an investigation of mixture models.

Example 7. All standard regression estimates, including the one based on least absolute deviations (the L_1-estimate, which generalizes the highly robust sample median), are sensitive to observations sitting at influential positions ("leverage points"). A single bad observation at an extreme leverage point may cause breakdown. Clearly, a higher breakdown point would be desirable. How large can it be made, and how large should it be? Via projection pursuit methods it is indeed possible to approach a breakdown point of 1/2 in large samples, provided the data are in general position (i.e., under the idealized, uncorrupted model no p rows of the n-by-p design matrix are linearly dependent, and thus any p observations give a unique determination of θ). This is a result of considerable theoretical interest.

Unfortunately, all estimators that try to optimize the breakdown point seem to run into the Souped-up Car Syndrome. The first among them was the LMS-estimate (Hampel 1975, Rousseeuw 1984).

The LMS- (Least Median of Squares) estimate of θ modifies the least squares approach by minimizing the *median* instead of the *mean* of the squared residuals:

$$\text{median}\left\{(y_i - x_i^T\theta)^2\right\}. \tag{8}$$

If the data points are in general position, its breakdown point is $([n/2]-p+2)/n \to 1/2$. But it has the following specific drawbacks:

– Its efficiency is low: the dispersion of the estimate decreases at the rate $n^{-1/3}$, instead of $n^{-1/2}$.

– Its computational complexity increases exponentially with p.

– What if the points are not in general position?

My conclusion is that an asymptotic theory for large p and n does not make much sense under such circumstances, and a small sample theory is not available.

S-estimates were introduced by Rousseeuw and Yohai (1984) to overcome some of these shortcomings. They estimate θ by minimizing a suitable robust M-estimate of the scale of the residuals. Under suitable regularity conditions their breakdown point also approaches 1/2 in large samples, and they reach a high efficiency at the ideal model, with a dispersion converging at the rate $n^{-1/2}$. Unfortunately, S-estimators suffer from a serious flaw which probably cannot be removed, namely that uniqueness and continuity can only be proved under certain conditions, see Davies (1993, Section 1.6).

Moreover, Davies (*ibid.*) points out that all known high breakdown point estimators of regression are inherently unstable. Paradoxically, it thus seems that in order to achieve an optimal regression breakdown point we may have to sacrifice robustness.

8. Design Issues

8.1. Optimal Designs

Example 8. Assume that the task is to fit the best possible straight line to data originating from an exactly linear function. Then the optimal regression design puts all observations on the extreme points of the segment where observations can be made.

However, a possibly more realistic version of this task is to fit the best straight line to an *approximately* linear function. In either case, one would want to make something like the expectation of the integrated mean square error as small as possible. Of course, usually one does not know much about the small deviations from linearity one might have to cope with (and does not care about them, so long as they are small). Already Box and Draper (1959, p. 622) had recognized the crux of the situation and had pointed out: "The optimal design in typical situations in which both variance and bias occur is very nearly the same as would be obtained if *variance were ignored completely* and the experiment designed so as to *minimize bias alone*."

In other words, the "naive" design, which distributes the observations evenly over the accessible segment, in such a situation should be very nearly optimal, since it minimizes the integrated squared bias of the fit. Apart from that, it has an advantage over the optimal design since its redundancy allows to check the linearity assumption.

These aspects have been made precise in a minimax sense by Huber (1975a). The most surprising fact emerging from this study was: there is a range where the

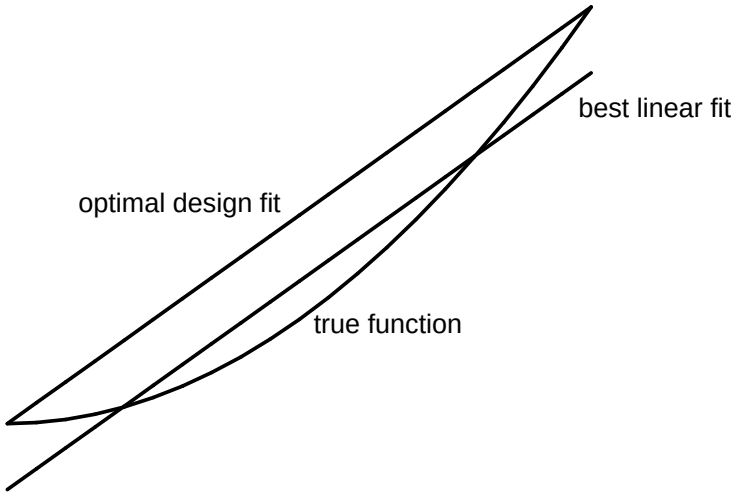


FIG 1. *Optimal design fit (based on the theoretically optimal design) and best linear fit (minimizing the integrated squared error), to a not-quite linear function. Observational errors are neglected in this figure.*

deviation from linearity is slight enough to stay below statistical detectability, yet large enough so that the "naive" design will outperform the "optimal" design and give a better linear approximation to the true function. Even though the effects are much less dramatic than in Example 3, we evidently have run here into another example of the Scapegoat Syndrome.

8.2. Regression Design and Breakdown

Example 9. In higher dimensions, generalizing the preceding example, an optimal linear regression design would place an equal number of m observations onto each of the $(p+1)$ corners of a p-dimensional simplex. Technically, such a design is optimal, but again, it lacks redundance.

For such a design the best possible breakdown point is

$$\varepsilon^* = \lceil m/2 \rceil / (m(p+1)) \approx 1/(2(p+1)). \tag{9}$$

This breakdown point is attained by the L_1-estimate (calculate the median at each corner). The so-called high-breakdown point LMS- and S-estimates cannot do any better.

But already an arbitrarily small jittering of the design points will bring them into general position. Then the breakdown point of LMS and S is close to $1/2$. How can this happen?

On closer inspection we see that the high breakdown point of LMS- and S-estimates is achieved by extrapolation: at each corner, you put more faith in the value extrapolated from the mp observations clustering near the far-away other p corners, than in the m local values. The fitted hyperplane thus not only loses efficiency, but becomes sensitive to small errors affecting a majority of the observations, such as rounding.

The conclusion is that high breakdown point regression is not necessarily robust. We have a clear case of the Souped-up Car Syndrome: both extremes, optimal

design and optimal breakdown point, lead to estimates with undesirable properties, and a compromise is called for. A quantitative, *design-dependent* theory of robust regression would seem to be needed. The customary assumption underlying all high breakdown point regression theories in the published literature, namely that the regression carrier is a random sample from a suitable multi-dimensional continuous distribution, in my opinion is much too narrowly conceived.

9. Bayesian Statistics

Example 10. What is Bayesian robustness? Bayesian statistcs has a built-in problem with the Scapegoat Syndrome, that is, with over-reliance on the model; this problem becomes acute in connection with robustness. By definition, Bayes procedures are optimal for the chosen model, consisting of a prior $\alpha(\theta)$ and a family of conditional densities $f(x,\theta)$. Instability, and conversely robustness, thus are properties of the *model*. This was emphasized in 1978 by George Box in an illuminating, facetious but profound oral interchange with John Tukey at an ARO meeting on Robustness in Statistics (Launer and Wilkinson, 1979). Box maintained that, after all, *he* had invented the term (see Box 1953), and that he could define it as he pleased, and that in his opinion robustness was to be achieved by choosing a proper model, not by tampering with the data (by trimming or Winsorizing) as Tukey was wont to do. He did not elaborate on how to choose such a model.

The philosophical problem of Bayesian statistics is that it is congenitally unable to separate the model, the underlying true situation, and the statistical procedure. It acts as if the model were exactly true, and it then uses the corresponding optimal procedure. A fundamentalist Bayesian, for whom probabilities exist only in the mind, will not be able to see that there is a problem of the Scapegoat type; it takes a pragmatist like George Box to be aware of it.

I shall now attempt to sketch a way around this Bayesian Scapegoat Sydrome. The central question is: what is a *robust model*? *Ad hoc* parametric supermodels, which sometimes are advertised as *the* Bayesian approach to robustness, do not guarantee robustness. There are no reliable guidelines to select such models, and the resulting procedures may suffer from instabilities.

If we proceed pragmatically, then, as a minimum requirement, the statistical conclusions from the model ought to be insensitive to occasional outliers. Sensitivity studies *à la* Berger, that is: admit that the specifications are inaccurate and find the range of implied conclusions (see Wolpert 2004, p. 212), may reveal the presence of outliers: if there are outliers, small changes in the tails of the model $f(x,\theta)$ can produce large effects. Also, they may reveal conflicts between the prior and the observational evidence: if the observational evidence points to a θ far from the center of the prior, small changes in the tails of the latter can produce large effects. Thus, if a sensitivity analysis shows that the range of implied conclusions is narrow, any model in the uncertainty range will do. If not, we better choose a robust model. But then, why not choose a robust model right away? Unfortunately, sensitivity studies do not help us find a robust model.

The following is a proposal for an informal portmanteau definition of robustness, covering both Bayesian and non-Bayesian statistics:

> *Uncertain parts of the evidence should never have overriding influence on the final conclusions.*

This is supposed to apply not only to questionable data (outliers), but also to uncertainties in the model densities $f(x,\theta)$ and to uncertainties in the prior $\alpha(\theta)$,

and even to vagueness in the specification of the target loss function.

How to implement such a loose definition? The first two of the above four requirements are interconnected and tricky to separate: insensitivity to dubious data features (gross errors), and insensitivity to uncertain model specifications. I claim that the following implementation should do the job for both aspects: Choose a model $f(x,\theta)$ *within the uncertainty range,* such that the conclusions are insensitive to gross errors. This has to be made precise.

The mode of the posterior density solves

$$\alpha'(\theta)/\alpha(\theta) + \sum f'(x_i,\theta)/f(x_i,\theta) = 0, \tag{10}$$

where the prime denotes the derivative with respect to θ. For a flat prior, the mode of the posterior coincides with the maximum likelihood estimate.

As Freedman (1963) has expressed it, there is a "striking and mysterious fact", namely that asymptotically, Bayes and M.L. estimates behave similarly: they not only have the same asymptotic distribution, but if the true underlying distribution belongs to the parametric family, the Bayesian posterior distribution, centered at the M.L. estimate and scaled by $n^{-1/2}$, is asymptotically normal and coincides with the asymptotic distribution of the M.L. estimate, centered at the true θ and also scaled by $n^{-1/2}$. See also LeCam (1957); the result apparently goes back to Bernstein and von Mises.

Thus, if we are willing to adopt the infinitesimal approach via gross error sensitivity (see Example 6), asymptotic robustness ideas should carry over from non-Bayesian M-estimates. Though, Hampel's approach through gross error sensitivity does not apply without some caveats, since it does not automatically lead to ψ-functions that are logarithmic derivatives of probability densities (which is a necessary side condition in the Bayes context — another example of a straitjacket). Finite ε-neighborhoods need somewhat more work. Assume that the M- and Bayes estimates both are calculated on the basis of the least favorable density (instead of the unknown true underlying distribution, which is supposed to lie anywhere in the given ε-neighborhood). Then, the M- and Bayes estimates still have the same asymptotically normal distribution, but the equivalence with the asymptotic posterior is lost. Though, in the one-dimensional location case it can be shown that the asymptotic variance of the posterior then lies between the asymptotic variance of the M-estimate and the upper bound for that variance obtained from the least favorable distribution (to be published in the forthcoming 2nd edition of my *Robust Statistics*). — As an amusing aside on the subject of pitfalls, I might mention that the usual applications in econometrics of one the formulas relevant in this context (the so-called "sandwich formula") go so far beyond its original intention that they deserve an honorable mention in the category of overly optimistic heuristics, see Freedman (2006).

In short, the heuristic conclusion, deriving from hindsight based on non-Bayesian robustness, thus is that f'/f ought to be bounded. In (10) the prior acts very much like a distinguished additional observation. Thus, in analogous fashion, also α'/α ought to be bounded. In both cases, the bounds should be chosen as small as feasible. Ordinarily, these bounds are minimized by the least informative distributions, with Fisher information used as measure of information. Thus, a possible optimization goal can be expressed:

> *A most robust Bayesian model can be found by choosing α and f to be least informative within their respective (objective or subjective) uncertainty ranges.*

For all practical purposes this is the same recipe as the one applying to the non-Bayesian case. But like there, it is difficult to implement once one wants to go beyond the location case. And if it is adopted overly literally, we might even get trapped in one the pitfalls of optimality.

10. Concluding Remarks

In the 1970s statistical theory clearly had been in the third, consolidation or "squeezing" phase of the development cycle. At present, we seem to have entered a new cycle and seem to be in the middle of a new "groping" phase, trying to get conceptual and theoretical handles on challenging problems involving extremely large and complexly structured data sets.

I hope that this time the laxness of the groping phase will be eliminated in time, and will not be cemented into place during the consolidation phase. Perhaps it may help to keep in mind the following aphorisms on optimality and optimization. They are not new, they are re-iterating sentiments already expressed by Tukey in 1962. Those sentiments subsequently had been studiously ignored by most statisticians. I hope that this time they will fare better.

- Optimality *results* are important: they show what can (and what cannot) be achieved under ideal conditions, and in particular they show whether a given procedure still has worthwhile potential for improvement.
- Optimal *procedures* as a rule are too dangerous to be used in untempered form.
- Beware of sloppy asymptotics.
- Never confuse the idealized model with the truth.
- Do not optimize one aspect to the detriment of others.
- There are no clear-cut rules on how the tempering of optimal procedures should be done — compromises are involved, and one must rely on human judgment. But if one insists on a mathematical approach, minimizing Fisher information within a subjective uncertainty range often will do a good job, both for Bayesians and non-Bayesians.

References

[1] Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, **40**, 318–335.

[2] Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a response surface design. *J. Amer. Statist. Ass.*, **54**, 622–654.

[3] Clausewitz, C. von (1832). *Vom Kriege*. 19th edition (1991). Dümmler Verlag, Bonn.

[4] —— (1984). *On War*. Edited and translated by M. Howard and P. Paret. Princeton University Press, Princeton, NJ.

[5] Davies, P. L. (1993). Aspects of robust linear regression. *The Annals of Statistics*, **21**, 1843–1899.

[6] Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*. Bickel, P. J., Doksum, K. A., and Hodges, J. L. (Eds.) Wadsworth, Belmont, CA.

[7] Eddington, A. S. (1914). *Stellar Movements and the Structure of the Universe*. Macmillan, London.

[8] Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error and the mean square error. *Monthly Not. Roy. Astron. Soc.*, **80**, 758–770.

[9] —— (1922). On the mathematical foundation of theoretical statistics. *Philos. Trans. Roy. Soc. London, Ser. A*, **222**, 309–368.

[10] Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Annals of Mathematical Statistics*, **34**, 1386–1403.

[11] —— (2006). On the so-called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician*, **60**, 209–302.

[12] Hampel, F. R. (1968). Contributions to the theory of robust estimation. *Ph. D. Thesis*. University of California, Berkeley.
[13] ——(1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Ass.*, **62**, 1179–1186.
[14] ——(1975), Beyond location parameters: Robust concepts and methods. Proc. 40th Session I. S. I., Warsaw 1975. *Bull. Int. Statist. Inst.*, **46**, Book 1, 375–382.
[15] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics. The Approach Based on Influence*. John Wiley, New York.
[16] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.
[17] ——(1965). A robust version of the probability ratio test. *Ann. Math. Statist.*, **36**, 1753–1758.
[18] ——(1966). Strict efficiency excludes superefficiency, (Abstract). *Ann. Math. Statist.*, **37**, 1425.
[19] ——(1968). Robust confidence limits. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, **10**, 269–278.
[20] ——(1972). Robust statistics: A review. *Ann. Math. Statist.*, **43**, 1041–1067.
[21] ——(1975a). Robustness and designs. In *A Survey of Statistical Design and Linear Models*, J. N. Srivastava (Ed.) North Holland, Amsterdam.
[22] ——(1975b). Application vs. abstraction: the selling out of mathematical statistics. *Suppl. Adv. Appl. Prob.*, **7**, 84–89.
[23] ——(1981). *Robust Statistics*. John Wiley, New York.
[24] Huber-Carol, C. (1970). Etude asymptotique de tests robustes. *Ph.D. Thesis*. Eidgen, Technische Hochschule, Zürich.
[25] James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **I**, 311–319.
[26] Kiefer, J. (1959). Optimum experimental designs. *J. Roy. Statist. Soc. Ser. B* **21**, 272–319.
[27] Launer, R. L. and Wilkinson, G. N. (Eds.) (1979). *Proc. ARO Workshop on Robustness in Statistics*, April 11–12, 1978, Academic Press, New York.
[28] LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. Calif. Publ. Statist.*, **1**, 277–330.
[29] —— (1957). Locally asymptotically normal families of distributions. *Univ. Calif. Publ. Statist.*, **3**, 37–98.
[30] Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. John Wiley, New York.
[31] —— (1983). *Theory of Point Estimation*. John Wiley, New York.
[32] Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London, Ser. A*, **231**, 289–337.
[33] Rieder, H. (1994). *Robust Asymptotic Statistics*. Springer, Berlin.
[34] Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, **79**, 871–880.
[35] Rousseeuw, P. J. and Yohai, V.J. (1984). Robust regression by means of S-Estimators. In *Robust and Nonlinear Time Series Analysis*. Franke, J., Härdle, W. and Martin, R. D. (Eds.) Lecture Notes in Statistics, **26**. Springer, New York.
[36] Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics*, Olkin, I. (Ed.) Stanford University Press, Stanford, Calif.
[37] —— (1962). The future of data analysis. *Ann. Math. Statist.*, **33**, 1–67.
[38] Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.
[39] Whitehead, A. N. and Russell, B. (1910–13). *Principia Mathematica*, 3 vols. Cambridge University Press.
[40] Wolpert, R. L. (2004). A conversation with James O. Berger. *Statistical Science*, **19**, 205–218.

IMS Collections
Vol. 0 (2009) 46–65
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000018

Proportional Hazards Regression with Unknown Link Function

Wei Wang[1], **Jane-Ling Wang**[2] **and Qihua Wang**[3]

Harvard Medical School and Brigham and Women's Hospital; University of California, Davis; Chinese Academy of Science and The University of Hong Kong

Abstract: Proportional hazards regression model assumes that the covariates affect the hazard function through a link function and an index which is a linear function of the covariates. Traditional approaches, such as the Cox proportional hazards model, focus on estimating the unknown index by assuming a known link function between the log-hazard function and covariates. A linear link function is often employed for convenience without any validation. This paper provides an approach to estimate the link function, which can then be used to guide the choice of a proper parametric link function. This is accomplished through a two-step algorithm to estimate the link function and the effects of the covariates iteratively without involving the baseline hazard estimate. The link function is estimated by a smoothing method based on a local version of partial likelihood, and the index function is then estimated using a full version of partial likelihood. Asymptotic properties of the non-parametric link function estimate are derived, which facilitates model checking of the adequacy of the Cox Proportional hazards model. The approach is illustrated through a survival data and simulations.

Contents

1 Introduction 46
2 Estimation Procedure and Main Results 48
2.1 Algorithm and Computational Issues 49
2.2 Main Results 51
2.3 Model Checking and Selection 51
3 Simulation Studies 52
4 Data Analysis 54
5 Conclusion and Future Research 57
Acknowledgement 59
Appendix 60
References 65

1. Introduction

Proportional hazards regression model has played a pivotal role in survival analysis since Cox proposed it in 1972. Let T represent survival time and Z its associate

[1]Harvard Medical School and Brigham and Women's Hospital, Boston, USA, email: wwang4@partners.org
[2]University of California, Davis, USA, Apartado Postal 402, email: wang@wald.ucdavis.edu
[3]Chinese Academy of Science, Beijing, China, email: qhwang@amss.ac.cn and University of Hongkong, Hongkong, email: qhwang@hku.hk
AMS 2000 subject classifications: Primary 62G05; secondary 62N02
Keywords and phrases: Partial likelihood, local partial likelihood, Nonparametric smoothing, Dimension reduction

covariate vector. Under the proportional hazards model, the hazard function for T, given a particular value z for the covariate Z, is defined as

$$\lambda\{t \mid z\} = \lambda_0(t) \exp\{\psi({\beta_0}^T z)\}, \tag{1}$$

where $\lambda_0(t)$ is an unknown baseline hazard function corresponding to $z = (0, \cdots, 0)$, and $\psi(\cdot)$ is called the link function with $\psi(0) = 0$. With fully specified link function ψ, the partial likelihood method was introduced in [4, 5] to estimate the regression parameters, β_0, with the option to accommodate censored data. The most common choice for ψ is the identity function, which corresponds to the time-honored Cox model. In reality the link function is unknown and needs to be estimated. This is especially useful to validate a preferred choice, as an erroneous link function could dramatically distort risk assessment or interpretation of odds ratios. When the link function is known, such as in the Cox model, model (1) is a special case of the transformation model first proposed in [7] and subsequently studied in [6], [3] etc. Our goal in this paper is to consider model (1) with an unknown link function. This problem was first studied in an unpublished Ph.D. thesis [19]. However, the procedure there was less efficient and we propose an improved estimate, studying its asymptotic properties.

Previous work focuses on the special case when the covariate is one-dimensional, or equivalently when β is known in (1). Under this special one-dimensional case, a local partial likelihood technique in [18] and a variation of the local scoring algorithm of [12] can be used to estimate the unknown link function in (1). Gentleman and Crowley [10] proposed a local version of the full likelihood instead of partial likelihood by alternating between estimating the baseline hazard function and estimating the covariate effects. The local likelihood methods in these papers were based on data whose covariate values fall in a neighborhood of the targeted location. Fan, Gijbels and King [8] used instead a local polynomial method to approximate the local partial likelihood, and derived rigorous asymptotic results for their approach. Simulation studies there showed that the local partial likelihood method is comparable to the local full likelihood method in [10]. Two spline approaches have also been considered, with smoothing splines resulting from a penalized partial likelihood in O'Sullivan [16] and regression splines from [17].

While the aforementioned approaches can be easily extended to q-dimensional covariates by estimating a multivariate unknown link function $\psi(z_1, \cdots, z_q)$, such nonparametric approaches are subject to the curse of dimensionality and may not be suitable for $q \geq 3$. Moreover, the resulting model would be different from model (1), which has the attractive dimension reduction feature that the covariate information is succinctly summarized in a single index and is a nonparametric extension of the Cox proportional hazards model. Model (1) could also be used as an exploratory tool to guide the choice of a suitable parametric link function.

A two-step iterative algorithm to estimate the link function and the covariate effects is proposed in Section 2. In the first step, an initial estimate of the regression parameter β is plugged in model (1) so that the link function can be estimated by a smoothing method based on a local version of partial likelihood ([4, 5]). The second step involves updating the regression parameters using the full partial likelihood with the estimated link function in step 1 inserted. These two steps will be iterated until the algorithm converges. Asymptotic results for the link estimators are stated in Section 2. In particular, Theorem 2 provides the building blocks to check the link function and inference for the individual risk, $\psi(\beta^T z)$. It also reveals that the nonparametric estimate of the link function is as efficient as the one for model

(1) but with a known regression parameter β. Thus, there is no efficiency loss to estimate the link function even if β is unknown in our setting. This is also reflected in the simulation studies in Section 3. The approach in Section 2 is further illustrated in Section 4 through a data set from the Worcester Heart Attach Study. All the proofs of the main results are relegated to an appendix.

We remark here that [15] also studied model (1) with a different approach. They assumed that the link function is in a finite dimensional subspace spanned by polynomial spline bases functions and the dimension of this subspace is known. This leads to a flexible parametric model where the spline coefficients corresponding to the link functions and β can then be estimated directly through traditional partial likelihood approaches. While this has the benefit of simplicity as everything is in the parametric framework, it tends to underestimate the standard errors of the estimates. Two sources of bias arise, one derives from the fact that in reality the number of spline bases depends on the data and is a function of the sample size, so the standard errors are underestimated by the simple parametric inference. In addition, the link estimation might be biased, as in theory an infinite number of spine bases might be required to span the unknown link function. These biases could significantly affect the asymptotic results. In contrast, our approach provides correct asymptotic theory and efficient estimation of the link function.

2. Estimation Procedure and Main Results

Since there are three different unknown parameters, $\lambda_0(\cdot)$, $\psi(\cdot)$ and β in model (1), we need to impose some conditions to ensure identifiability. To identify λ_0, it suffices to set $\psi(v) = 0$ at some point v, a common choice is $v = 0$. Since only the direction of β is identifiable if ψ is unknown, we assume that $\|\beta\|$=1 (here $\|\cdot\|$ represents the Euclidean norm) and that the sign of the first component of β is positive. As for the sampling plan, we assume an independent censoring scheme, in which the survival time T and censoring time C are conditionally independent, given the covariate vector Z. Let $X = min(T, C)$ be the observed event-time and $\Delta = I\{T \le C\}$ be the censoring indicator. The data $\{X_i, Z_i, \delta_i\}$ is an i.i.d. sample of $\{X, Z, \Delta\}$. We use the notation $t_i < \cdots < t_N$ to denote the N distinctive ordered failure times, and (j) to denote the label of the item failing at time t_j. The risk set at time t_j is denoted by $\mathcal{R}_j = \{i : X_i \ge t_j\}$.

For a fixed parametric value β, one can estimate the link function $\psi(\cdot)$ by any smoothing method, such as those cited in Section 1 when β is assumed known. We adopt the local partial likelihood approach in [8] and assume, for a given point v, that the p-th order derivative of $\psi(v)$ at point v exists. A Taylor expansion for $\beta^T Z$ in a neighborhood of v then yields,

$$\begin{aligned}\psi(\beta^T Z) &\approx \psi(v) + \psi^{'}(v)(\beta^T Z - v) + \cdots + \frac{\psi^{(p)}(v)}{p!}(\beta^T Z - v)^P \\ &= \psi(v) + (\beta^T \mathbf{Z})^T \gamma(v),\end{aligned} \tag{2}$$

where $\gamma(v) = \{\psi'(v), \cdots, \psi^{(p)}(v)/p!\}^T$ is the $p-$dimensional vector associated with the derivatives of ψ and $\beta^T \mathbf{Z} = \{\beta^T Z - v, \cdots, (\beta^T Z - v)^p\}^T$.

Let K be a kernel function, h be a bandwidth, and define $K_h(u) = h^{-1}K(u/h)$. Applying kernel weights to the logarithm of the global partial likelihood

$$\sum_{j=1}^{N} \psi(\beta^T Z_{(j)}) - \log\left\{ \sum_{i \in \mathcal{R}_j} \exp\Big\{\psi(\beta^T Z_{(j)})\Big\} \right\}$$

and replacing $\psi(\beta^T z)$ by the local approximation in (2), we arrive at (similarly to [8]) the local version of the log partial likelihood:

$$\sum_{j=1}^{N} K_h\big(\beta^T Z_{(j)} - v\big)\Bigg[\big(\beta^T \mathbf{Z}_{(j)}\big)^T \gamma(v) - \log\Bigg\{\sum_{i\in\mathcal{R}_j} \exp\Big\{\big(\beta^T \mathbf{Z}_i\big)^T \gamma(v)\Big\} K_h\{\beta^T Z_i - v\}\Bigg\}\Bigg], \tag{3}$$

where $\beta^T \mathbf{Z}_i$ and $\beta^T \mathbf{Z}_{(j)}$ are defined as $\beta^T \mathbf{Z}$ with $\mathbf{Z}$ replaced by $\mathbf{Z}_i$ and $\mathbf{Z}_{(j)}$ respectively. It can be shown that the local log partial likelihood in (3) is strictly concave with respect to $\gamma(\cdot)$, so for a fixed β, it has a unique maximizer with respect to γ. Let $\hat{\gamma}(v)$ be the local partial likelihood estimate with $\hat{\gamma}_k(v)$ denoting its k-th component, then $\psi^{(k)}(v)$ can be estimated by $\hat{\psi}^{(k)}(v) = k!\ \hat{\gamma}_k(v)$, for $k = 1, \cdots, p$. In principle, one could maximize (3) with respect to both β and γ, and this corresponds to maximizing the real local log likelihood. But we choose to maximize (3) only with respect to γ for a fixed estimated value of β, and this corresponds to maximizing a pseudo local log likelihood as the true β in (3) is replaced by an estimate. There are two reasons for our choice. First (3) is concave in γ, but not necessarily in β. Second, maximizing with respect to both parameters is probably not worth the additional computational cost, as the local likelihood procedures mainly serve as a smoother and the choice of the smoother is usually not crucial.

To estimate the link function, we use

$$\hat{\psi}(v) = \int_0^v \hat{\psi}'(w)dw,$$

where $\hat{\psi}'(v)$ is the first component of $\hat{\gamma}(v)$ at the last iteration step. There are several ways to approximate this integral, such as the trapezoidal rule or Gaussian quadrature. For computational simplicity, we apply the trapezoidal rule in the simulation studies, as suggested in [18], and this appears to be satisfactory.

2.1. Algorithm and Computational Issues

The procedure described in the previous subsection requires a certain choice of β in equation (2). This can be done either independently or iteratively as once an estimate of ψ is obtained, one can then estimate β through the global partial likelihood. An iterative algorithm, as shown below, can be established by alternatingly updating the estimates for β and ψ. Such an iteration procedure may improve the link estimate as a better estimate of β will lead to a better estimate of ψ.

Step 1. (a) Assign a nonzero initial value to β, and call it $\hat{\beta}$.

(b) For a given v, plug $\hat{\beta}$ into the pseudo log local partial likelihood and maximize

$$\sum_{j=1}^{N} K_h\{\hat{\beta}^T Z_{(j)} - v\} \cdot \Bigg[[\hat{\beta}^T \mathbf{Z}_{(j)}]^T \gamma(v) - \log\Bigg\{\sum_{i\in\mathcal{R}_j} \exp\{[\hat{\beta}^T \mathbf{Z}_i]^T \gamma(v)\} K_h\{\hat{\beta}^T Z_i - v\}\Bigg\}\Bigg]$$

with respect to $\gamma(v)$ to get the estimate $\hat{\gamma}(v)$.

(c) Obtain the values of $\hat{\gamma}(v)$, for $v = \hat{\beta}^T Z_i, i = 1, \cdots, n$.

(d) Apply the trapezoidal rule to obtain $\{\hat{\psi}(\hat{\beta}^T Z_i) : i = 1, \cdots, n\}$.

Step 2. Plug $\hat{\psi}(\cdot)$ into the log (global) partial likelihood

$$l_G(\beta, \hat{\psi}) = \sum_{j=1}^{N} \left[\hat{\psi}(\beta^T Z_{(j)}) - \log \left\{ \sum_{i \in \mathcal{R}_j} \exp\{\hat{\psi}(\beta^T Z_i)\} \right\} \right],$$

and maximize it with respect to β to update the estimate $\hat{\beta}$. We use the angle between two estimated $\hat{\beta}$ at two consecutive iterations as the convergence criterion.

Remark 1. The Newton-Raphson method is used to find the estimators in Step 1 and 2. The initial value of β can be set in different ways but cannot be zero as a nonzero value is needed in step 1 to estimate the link function. However, this restriction does not exclude the final β-estimate to be zero or close to zero. A simple choice is to fit the usual Cox model and use this estimator in the first step. To accelerate the computation, one can also use alternative estimates as described below.

Remark 2. It is possible to accelerate the computation by using a $\sqrt{n}$-consistent initial estimator, as Theorems 1 and 2 below imply that no iteration is required for the link estimate and that it will converge efficiently at the usual nonparametric rate. Namely, the link function can be estimated with the same efficiency as when β is known. In practice, we find that one iteration helps to improve the numerical performance but further iteration is usually not necessary. There are two choices for a $\sqrt{n}$-consistent initial estimator, one is the estimator in [2] that extends the sliced inverse regression (SIR) approach to censored data. Specifically, this approach requires a certain design condition as listed in (2.3) there but has the advantage that it leads to a $\sqrt{n}$-consistent estimator for β without the need to estimate the link function. Another initial estimate which does not rely on the design conditions (2.3) in [2] is provided in a Ph. D. thesis [19]. Specifically, this involves replacing the ψ function in step 2 above by its local version (2), which leads to the cancelation of the term ψ and results in a version of log (global) partial likelihood that involves only the derivative $\gamma(v)$ of ψ but not ψ itself. Thus, Step 2 above is replaced by

Step 2*. Maximize the following approximate log (global) partial likelihood with respect to β:

$$\sum_{j=1}^{N} \left[[\beta^T Z_{(j)}]^T \hat{\gamma}\{\hat{\beta}^T Z_{(j)}\} - \log \left\{ \sum_{i \in \mathcal{R}_j} \exp\Big([\beta^T Z_i]^T \hat{\gamma}\{\hat{\beta}^T Z_i\} \Big) \right\} \right].$$

This approximation may result in some efficiency loss, but has computational advantages over the estimate in Step 2, since we do not need to estimate $\psi(v)$ and thus can skip Step 1(d). The resulting estimate for β was shown in [19] to be $\sqrt{n}$-consistent, consequently an ideal choice as the initial estimate for β.

Remark 3. In step 1, the local log partial likelihood in (3) is replaced by a pseudo log partial likelihood with β replaced by $\hat{\beta}$. As this $\hat{\beta}$ approaches β, the link estimate resulting from maximizing the pseudo log partial likelihood can be expected to approach the true link function at the usual nonparametric rate. This is because the parametric estimate $\hat{\beta}$ converge to its target at the $\sqrt{n}$–rate, which is faster than the nonparametric link estimate. A rigorous proof is provided in Theorem 1 and Theorem 2.

Remark 4. For large sample sizes, it is unnecessary to estimate the link function at each data point. An alternative way is to estimate the link function at equal-distance grid points, then use interpolation or smoothing methods to obtain the estimated value at each data point. Our simulation results show that this short-cut is computationally economical while retaining similar accuracy.

2.2. Main Results

Let $f(\cdot)$ be the probability density of $\beta^T Z$, for a given v, let $P(t \mid v) = P(X \geq t \mid \beta^T Z = v)$, $Y(t) = I\{X \geq t\}$, $H = diag\{h, \cdots, h^p\}^T$ and $\mathbf{u} = \{u, \cdots, u^p\}^T$.

Theorem 1. *Under conditions (C1)-(C5) in the Appendix, for any $\sqrt{n}$ consistent estimator $\hat{\beta}$ of the true parameter β_0, let $\hat{\gamma}(\cdot)$ be the corresponding estimator for the derivatives $\gamma_0(\cdot)$ of the true link ψ and $\hat{\psi}(v) = \int_0^v \hat{\psi}'(w)dw$, where $\hat{\psi}'(\cdot)$ is the first component of $\hat{\gamma}(\cdot)$. If $h \to 0$, $nh/\log n \to \infty$, $nh^4 \to \infty$ then*

$$\sup_v |\hat{\gamma}(v) - \gamma_0(v)| \to_p 0,$$

and

$$\sup_z |\hat{\psi}(\hat{\beta}^T z) - \psi({\beta_0}^T z)| \to_p 0$$

Theorem 2. *Under the conditions in Theorem 1 and for bounded nh^{2p+3},*

$$(a)\sqrt{nh}\left\{H(\hat{\gamma}(v) - \gamma_0(v)) - \frac{\psi^{(p+1)}(v)}{(p+1)!}A^{-1}bh^{p+1}\right\} \to_D N\left\{0, \frac{\sigma^2(v)}{f(v)}A^{-1}DA^{-1}\right\}.$$

Furthermore, we have

$$(b)\sqrt{nh}\left\{H(\hat{\gamma}(\hat{\beta}^T z) - \gamma_0({\beta_0}^T z)) - \frac{\psi^{(p+1)}({\beta_0}^T z)}{(p+1)!}A^{-1}bh^{p+1}\right\} \to_D N\left\{0, \frac{\sigma^2({\beta_0}^T z)}{f({\beta_0}^T z)}A^{-1}DA^{-1}\right\},$$

where $A = \int \mathbf{u}\mathbf{u}^T K(u)du - \nu_1\nu_1^T$, $b = \int u^{p+1}(\mathbf{u} - \nu_1)K(u)du$, $D = \int K^2(u)(\mathbf{u} - \nu_1)^{\otimes 2}du$, $\nu_1 = \int \mathbf{u}K(u)du$, and $\sigma^2(v) = E\{\delta|\beta^T Z = v\}^{-1}$.

Theorem 1 establishes the uniform consistency of the local partial likelihood estimator of γ_0 and Theorem 2 provides the joint asymptotic normality of the derivative estimators. The limiting distribution of $\hat{\gamma}$ is identical to the one in [8], where β is assumed to be known. Thus, there is no efficiency loss as long as β can be estimated at the usual $\sqrt{n}$-rate.

2.3. Model Checking and Selection

While an estimated link function is of interest to correctly reflect the risk associated with a covariate, a parametric link function is often preferable to a nonparametric one to lend a parsimonious model with more interpretable results. Thus, a main incentive to estimate the link function could be for exploratory model selection to facilitate the choice of a proper parametric link function in the proportional hazards

model (1). If so, the β estimate in Step 2 only aids in the link estimation and need not be the end product. Once a suitable link function has been selected, Theorem 2 can be used for model checking. For instance, to check the identity link function under the Cox model, one can test H_0: $\psi'(v) = 1$. Since the first component of $\gamma(v)$ is $\psi'(v)$, a local polynomial of order $p = 2$ is usually employed to estimate such a derivative, and the resulting asymptotic distribution of the corresponding estimate is given below.

Corollary 1. *Under the condition of Theorem 2, and with $p = 2$ there, we have*

$$\sqrt{nh^3}\left(\hat{\psi}'(v) - \psi'(v) - \frac{1}{6}\psi^{(3)}(v)h^2\frac{\int u^4K(u)du}{\int u^2K(u)du}\right) \to_D N\left(0, \frac{\sigma^2(v)\int u^2K^2(u)du}{f(v)\int u^2K(u)du}\right).$$

Corollary 1 facilitates the construction of testing procedures and asymptotic simultaneous confidence bands for the link function, but rigorous asymptotic theory requires much further work and is not available yet. In principle, one could check the appropriateness of the link function at all data points v that falls in the range of $\beta^T Z$. Since the true value of β is unknown, it is natural to replace it with an estimate. However, one must bear in mind the precision of this estimate as well as the low precision of $\psi'(v)$ for v in the boundary region of $\hat{\beta}^T Z$. Here boundary region is defined as within one bandwidth of the data range, where a smoothing procedure is employed. Since the bandwidth h is usually of a higher order than $n^{-\frac{1}{2}}$, the anticipated rate of convergence for $\hat{\beta}$, we recommend to restrict inference on $\psi'(v)$ for v that is in the interior and at least one bandwidth h away from either boundary of the range of $\hat{\beta}$.

Short of such a rigorous inference procedure for model checking, pointwise confidence intervals have often been used as a substitute for exploratory purposes. In the example in Section 4, we illustrate how to check the appropriateness of the Cox model, i.e. identity link function, using pointwise confidence intervals developed from Corollary 1. Readers should bear in mind that this is only an exploratory data analysis tool rather than a formal inference procedure.

3. Simulation Studies

To see how the algorithm in Section 2 works for the proposed model, we conducted simulation studies where a quadratic link function $\psi(\beta^T Z) = (\beta^T Z)^2$ with $\beta = (1, 3)^T$ and a constant baseline, $\lambda_0 = 0.005$, were employed. The design for the two-dimensional covariate, $Z = (Z_1, Z_2)^T$ is: $Z_1 \sim U(-1, 1)$ and Z_2 is a truncated $N(0, 1)$ with values in $[-1, 1]$. Parameters of β were chosen in such a way that the simulation generates a reasonable signal to noise ratio (cf. Fig 1). If we take ε to have the standard exponential distribution, exp(1), the resulting hazard function will be $\lambda_0 \exp\{\psi(\beta^T Z)\}$, and survival times from this model can be generated as $T = \exp\{\psi(-\beta^T Z)\}\varepsilon/\lambda_0$. Different uniform distributions were utilized to generate three independent censoring times so that the censoring rates were 0%, and roughly 25% and 50%. The Epanechnikov kernel was adopted in the link estimation. Two sample sizes, 200 and 50, were selected to see whether the methods are flexible for moderate to small samples. For $n = 200$ we used 25 equal-distance grid points to estimate the link function to save computational time as elucidated in Remark 4 of Section 2. Piece-wise spline interpolation was then used to get the link estimate during each iteration of the algorithm.

Due to the complication from the identifiability problem, the link function can only be identified up to a constant. Thus, $\hat{\psi}(v)$ and $\hat{\psi}(v) + c$, for any constant

TABLE 1

Comparison of different methods for estimating $\psi(\beta^T Z) = (\beta^T Z)^2$, where $\beta = (1,3)^T$ and $Z = (Z_1, Z_2)^T$ with $Z_1 \sim U(-1,1)$ and $Z_2 \sim N(0,1)$ (truncated at [-1,1]), n=200. The numbers before and after "/" are the means and standard deviations of d based on 100 simulations.

		h^*				Optimal
Censoring	†	0.1	0.2	0.3	0.4	MSE
No	1	3.152/0.181	3.152/0.181	3.152/0.181	3.152/0.181	9.970
	4	0.191/0.114	0.191/0.114	0.191/0.114	0.191/0.114	0.049
	2	1.234/2.554	0.439/0.142	0.843/0.150	1.417/0.151	0.213
	3	1.210/3.522	0.490/0.144	0.907/0.168	1.476/0.168	0.261
25%	1	3.151/0.181	3.151/0.181	3.151/0.181	3.151/0.181	9.961
	4	0.201/0.112	0.201/0.112	0.201/0.112	0.201/0.112	0.053
	2	1.256/2.548	0.425/0.157	0.684/0.172	1.197/0.178	0.206
	3	0.981/1.991	0.468/0.156	0.746/0.185	1.256/0.194	0.244
50%	1	3.149/0.181	3.149/0.181	3.149/0.181	3.149/0.181	9.947
	4	0.210/0.117	0.210/0.117	0.210/0.117	0.210/0.117	0.056
	2	1.361/2.529	0.525/0.215	0.535/0.161	0.729/0.210	0.322
	3	1.236/2.138	0.536/0.201	0.571/0.169	0.808/0.226	0.327

† Method 1 is under identity link and unknown β. Method 2 is under unknown link and true β. Method 3 is under unknown link and unknown β. Method 4 is under quadratic link and unknown β.

c, are considered to be equivalent procedures, and any measures of performance would declare these two procedures identical. This points to selecting a measure which measures the variation instead of the real difference. We adopt a measure proposed in [10] which is the standard deviation of the differences between the fitted values $\hat{\psi}(\hat{\beta}^T Z)$ and the true values $\psi(\beta^T Z)$ at all data points. More specifically, this measure, denoted by d, is the standard deviation of the difference $\{\hat{\psi}(\hat{\beta}^T Z) - \psi(\beta^T Z) : i = 1, \cdots, n\}$. We report in Table 1 the average values for this measure and its standard deviation based on 100 simulation runs.

Since at each estimating step, $\hat{\beta}$ was updated and the range of $\hat{\beta}^T Z$ might be different, we used a bandwidth h^*, which took a certain portion of the range of $\hat{\beta}^T Z$. For instance, an $h^* = 0.3$ means that the actual bandwidth is 0.3 times the range of the values of $\hat{\beta}^T Z$. Various bandwidths were explored, but we report only the results for bandwidth h^* varying from 0.1 to 0.4 (with 0.1 increment) times the data range of $\beta^T Z$ at each iteration stage. Results for other bandwidths were inferior and are not reported here.

Four procedures were compared and the results for $n = 200$ are shown in Table 1. Method 1 assumes that the link is identity (which is incorrect here) and the regression coefficient estimate $\hat{\beta}$ is therefore the Cox estimate based on the partial likelihood estimate. The aim is to see the effect of erroneously assuming the conventional Cox proportional hazards model. Method 2 assumes that β is known and estimates the unknown link function as in [8]. Method 3 is the new procedure where both the link function and regression coefficient β are estimated. Method 4 assumes that the true quadratic link function is known and the regression coefficient estimate, $\hat{\beta}$, is the partial likelihood estimate. The comparisons for the distance d are reported in Table 1. The results of the best procedures together with the corresponding optimal bandwidths are highlighted with boxes. It is not surprising that the best results came from the procedures with true quadratic link function and unknown β. Our estimators are close to those from method 2 ([8]) with known β, while the estimators based on the identity link model have much larger d.

To demonstrate the effect of an erroneous link function on regression estimates,

TABLE 2

Differences between the estimated $\hat{\beta}_2$ and the true β_2 where $\hat{\beta}_1$ is set equal to the true β_1, $\psi(\beta^T Z) = (\beta^T Z)^2$, where $\beta = (1,3)^T$ and $Z = (Z_1, Z_2)^T$ with $Z_1 \sim U(-1,1)$ and $Z_2 \sim N(0,1)$ (truncated at [-1,1]), n=200. The numbers before and after "/" are the biases and standard deviations of the $\hat{\beta}_2$ based on 100 simulations.

		h*				Optimal	Optimal
Censoring	†	0.1	0.2	0.3	0.4	MSE‡	MSE§
No	1	5.344/50.847	5.344/50.847	5.344/50.847	5.344/50.847	2613.949	11742.108
	4	-0.015/0.142	-0.015/0.142	-0.015/0.142	-0.015/0.142	0.021	0.703
	3	0.023/0.276	-0.051/0.154	-0.096/0.158	-0.215/0.190	0.026	0.968
25%	1	-1.229/24.742	-1.229/24.742	-1.229/24.742	-1.229/24.742	613.684	11320.063
	4	-0.013/0.154	-0.013/0.154	-0.013/0.154	-0.013/0.154	0.024	0.801
	3	0.038/0.307	-0.055/0.170	-0.089/0.171	-0.162/0.195	0.032	1.144
50%	1	-1.472/6.486	-1.472/6.486	-1.472/6.486	-1.472/6.486	44.237	11266.339
	4	-0.006/0.170	-0.006/0.170	-0.006/0.170	-0.006/0.170	0.029	0.968
	3	0.055/0.371	-0.050/0.188	-0.075/0.181	-0.110/0.190	0.038	1.362

† Method 1 is under identity link and unknown β. Method 4 is under true link and unknown β. Method 3 is under unknown link and unknown β.
‡ is the optimal MSE for $\hat{\beta}_2$ when $\hat{\beta}_1$ is set to be 1. § is the optimal MSE for the angles between $\hat{\beta}$ and the true β.

we report in Table 2 the results of the various estimates for β. Since there is no regression parameter estimate for method 2, only three procedures are compared in Table 2. There are several ways to compare the regression estimates, one way is to set the first component of β to the true value, then compare the difference between $\hat{\beta}$ and β based on the second component. Table 2 shows the results of the difference between the true β_2 and the estimate $\hat{\beta}_2$ for various procedures. The best procedures for the profile estimator in method 3 are shown in boxes under the optimal bandwidths. Another way to compare the different estimators is to calculate the angles between these estimators and the true parameter. To save space only the MSEs based on the optimal bandwidths are listed in the last column of Table 2 for the angle measure (degrees). Based on both optimal MSE measures reported in the last two columns of Table 2, the differences between the new profile estimators and the true parameters are way smaller than those from the identity link model, and reasonably close to those under the true link.

We can see that using the wrong link will lead to huge bias and MSE under all censoring patterns. The average angles between the β estimates assuming identity link and the true parameters are around 90°, which suggests that the β estimates with identity link are perpendicular to the true parameter space, indicating a total inability to estimate the regression parameter. This is in addition to the problem that the link function itself has been misspecified. Both underscore the importance to check the shape of the link function at the beginning of data analysis.

Four typical simulated curves are shown in Fig 1. The procedure (method 4) with known quadratic link function and unknown β performed the best. The procedure (method 2) with known β and our procedure captured the shape of the true curve well, but the procedure (method 1) based on the Cox model failed to capture the shape of the link function. Results for $n = 50$ summarized in Table 3 are consistent with the findings for $n = 200$ in Table 1.

4. Data Analysis

In this section, we illustrate the proposed model and estimation algorithm in Section 2 through the Worcester Heart Attack Study (WHAS) data. One of the goals of this study is to identify factors associated with the survival rates following hospital admission for acute myocardial infarction. The main data set has more than 11000

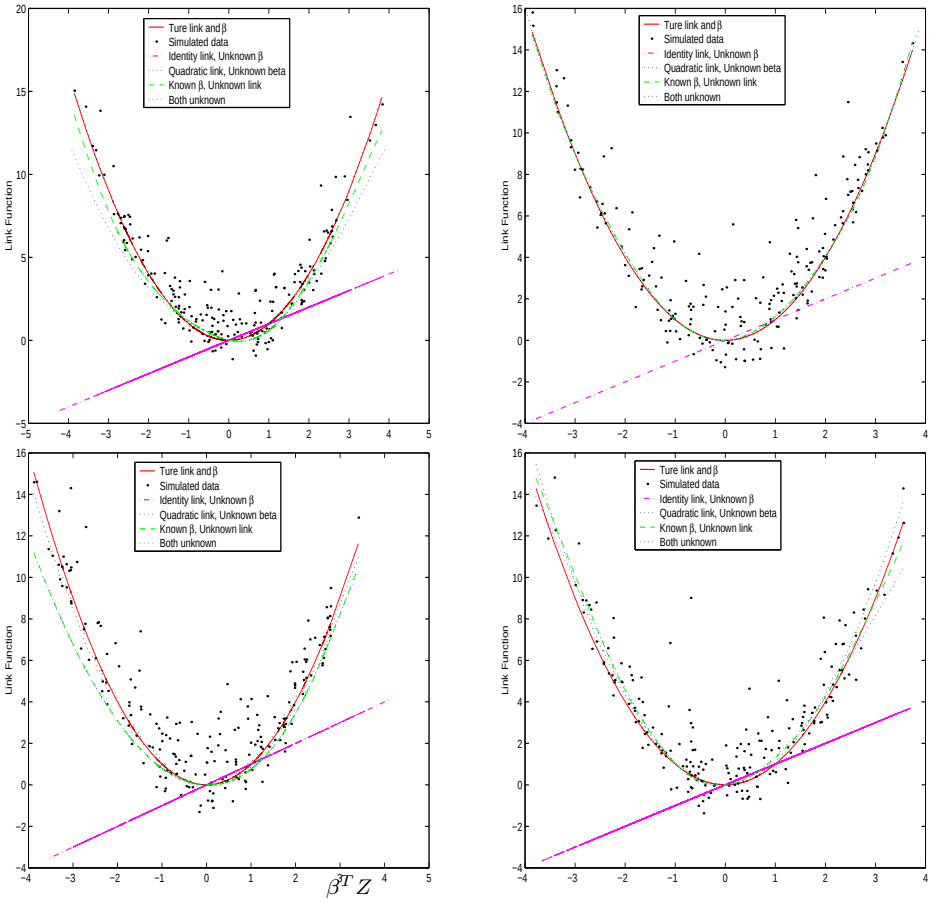


FIG 1. *Four typical simulated sets of data and estimated curves from the model* $\psi(\beta^T Z) = (\beta^T Z)^2$ *with* $\beta = (1,3)^T, Z = (Z_1, Z_2)^T$ *where* $Z_1 \sim U(-1,1)$ *and* $Z_2 \sim N(0,1)$ *(truncated at [-1,1]), 25% censoring and* $h^* = 0.2$. *The simulated data is indicated as black dots. The true curve* $\psi(\cdot)$ *is indicated as a red solid line, and the estimated curves* $\hat{\psi}(\cdot)$ *as a magenta dash line with identity link and unknown* β, *as a black dash-dotted line with quadratic link and unknown* β, *as a green dotted line with unknown link and true* β, *and as a blue dash-dotted line with both unknown link and unknown* β.

TABLE 3

Comparison of different methods for estimating $\psi(\beta^T Z) = (\beta^T Z)^2$, *where* $\beta = (1,3)^T$ *and* $Z = (Z_1, Z_2)^T$ *with* $Z_1 \sim U(-1,1)$ *and* $Z_2 \sim N(0,1)$ *(truncated at [-1,1]), n=50. The reported quantities are the smallest MSEs under optimal bandwidths.*

	No Censoring		25% Censoring		50% Censoring	
Method	d	Optimal h^*	d	Optimal h^*	d	Optimal h^*
Identity link and unknown β	9.199	-	9.168	-	9.102	-
Quadratic link and unknown β	0.184	-	0.203	-	0.295	-
Unknown link and true β	0.808	0.3	0.833	0.3	1.359	0.4
Unknown link and unknown β	0.945	0.4	1.286	0.4	1.423	0.4

admissions, but we used only a random sample of 500 patients as listed in [14]. This data set is chosen because the proportionality assumption has been carefully examined and is reasonably satisfied. Our goal here is to check the adequacy of the identity link function in the Cox proportional hazards regression model.

There were more than 10 covariates in the data set. After detailed model selection procedure, Hosmer et al. [14] included 6 variables age (AGE), initial heart rate (HR), initial diastolic blood pressure (DIASBP), body mass index (BMI), gender (GENDER), congestive heart complications (CHF), and the interaction between age and gender (AGEGENDER) in their model. After examining the linearity assumption using fractional polynomials, they decided to apply a two-term fractional polynomial model to the variable BMI. We thus begin with the univariate covariate BMI.

We tried different bandwidths and found similar patterns of the estimated link functions. In Fig 2 we report two of the results, which exhibit reasonable level of smoothness. The estimated link function in Fig 2 suggests clear nonlinearity. We then constructed a 95% point-wise approximated confidence interval of $\gamma(v)$ (Fig 3) to see whether it would cover the constant function 1. The results suggest that the estimated link functions have some curvature and further investigation is needed.

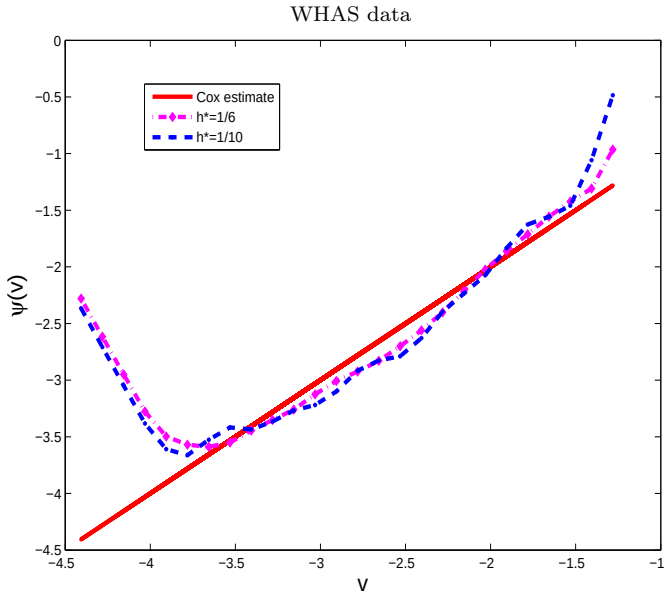


FIG 2. *The estimated link function, under two bandwidths, for the WHAS data with BMI as a covariate.*

Next we applied the proposed procedure to the multivariate model with all 7 covariates. We tried different bandwidths ranging from 1/10, 1/8, 1/7, 1/6, 1/5, 1/4 ,1/3, to 1/2 of the single index range, and plotted the results of three bandwidths in Fig 4. The estimated link function for $h^* = 1/4$ appears oversmoothed but all three estimates exhibit two bumps. The 95% point-wise approximated confidence intervals for $\gamma(v)$ as shown in Fig 5 also reveal curvature away from the constant $(= 1)$ horizontal line. Although it is arguable that the Cox model could be rejected at a low level, the suitability of an identity link function seems questionable.

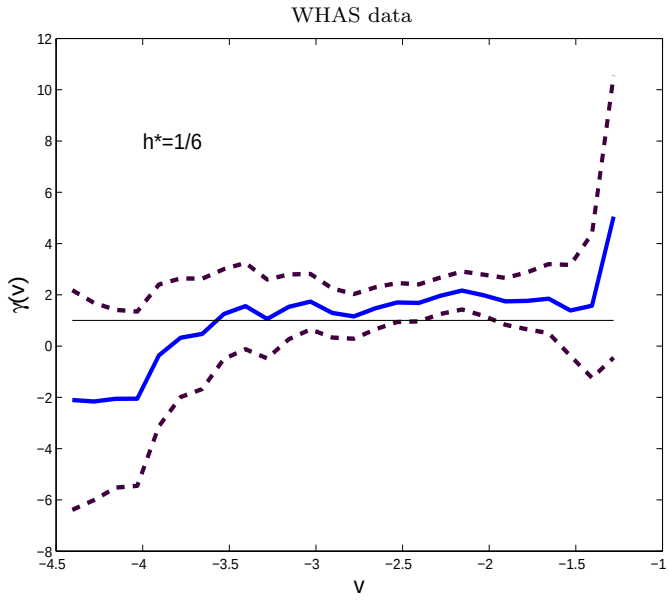


FIG 3. *The estimated confidence interval of γ at bandwidth $h^* = 1/6$ for the WHAS data based on BMI.*

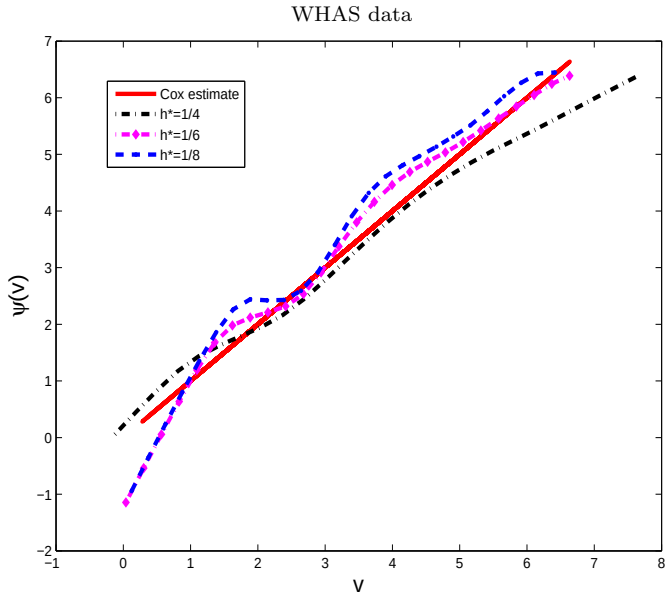


FIG 4. *The estimated link function for the WHAS data under three different bandwidths.*

5. Conclusion and Future Research

The proposed estimating procedures for the extended proportional hazards regression model with unknown link function and multi-dimensional covariates seem to be reliable for moderate to large sample sizes. Once the link function and the parameters of the index have been established, one can proceed to estimate the unknown

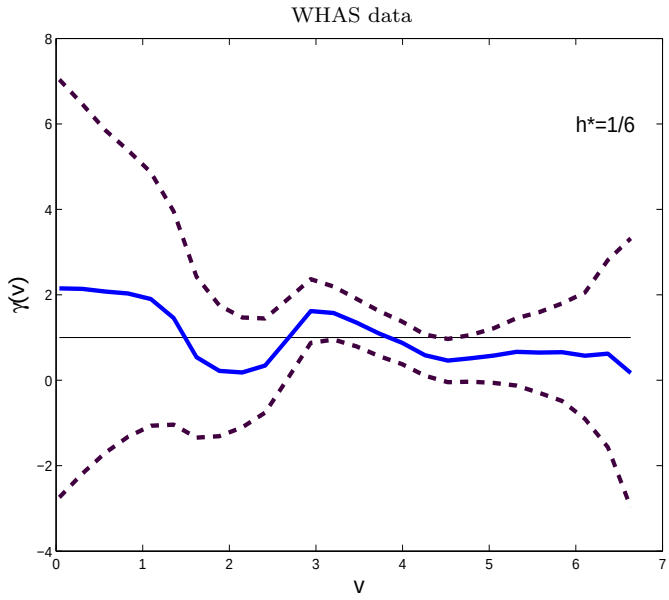


FIG 5. *The estimated confidence interval of γ for the WHAS data at bandwidth $h^* = 1/6$.*

baseline hazard function in model (1) using a Breslow-type estimate ([1]).

The cost of a misspecified link function has been demonstrated through the simulation studies in Section 3. As a consequence, the risk of an individual may be misinterpreted. It is thus important to at least estimate the link function in the initial model fitting stage as a model checking tool or guidance to a suitable class of parametric link functions. A rigorous test of parametric link function will be a worthwhile future project, as is the asymptotic theory for simultaneous inference of the link function and regression parameters.

The choice of automatic smoothing parameters, the bandwidth h in this case, is a challenging problem for proportional hazards model when a likelihood based smoother, such as the local partial likelihood estimate, is employed in the link estimate. This is because the components of the partial likelihood are dependent, hence the usual automatic bandwidth selection methods for linear models are not applicable here. The usual least square cross-validation procedure in nonparametric regressions also cannot be easily adapted to hazard based models such as the proportional hazards model. An alternative criterion, less computational intensive than cross-validation methods, was proposed in [18], based on a variation of the Akaike's information criterion for span selection, when the nearest neighborhood method was used for smoothing. However, the interpretation of AIC is not clear here since partial likelihood involves dependent components. The authors also acknowledged that the asymptotic correctness of the AIC criterion has not been established. Thus, automatic bandwidth choice remains an open question when the link function is being estimated. Meanwhile, we recommend to try several bandwidths and choose one that yields a moderately smooth link function as we did for the WHAS data. This subjective choice based on the visual degree of smoothness is commonly adopted as an ad hoc tool.

While this paper deals with time-independent covariates, it would be desirable to extend model (1) to time-dependent covariates as well. One complication is that the

entire history of the covariate process would be required or some kind of imputation needs to be performed to get even an initial estimate of β. Preliminary results were reported in [20] by imputing the covariate process through a functional principal components approach, and then proceeding with the estimation of the survival components at the second stage. Such a two-stage procedure is prone to bias as is well known in the joint modelling literature. Further investigations to correct the bias would be desirable, and joint modelling the longitudinal and survival process offers some hope if one can resolve the additional complication of an unknown link function. This is yet another worthwhile project to pursue in the future.

Acknowledgement

The work of Jane-Ling Wang and Wei Wang was completed with the partial support of National Science Foundation grants DMS04-06430, National Institutes of Health grant R01-DK-45939, and National Institute of Allergy and Infectious Diseases grant AI24643. Qihua Wang's research was supported by the National Science Fund for Distinguished Young Scholars in China (10725106), the National Natural Science Foundation of China (10671198), the National Science Fund for Creative Research Groups in China and the Research Grants Council of the Hong Kong (HKU 7050/06P). The authors are grateful for the invaluable suggestions of two reviewers and the editor.

Appendix

Let $f(\cdot)$ be the probability density of $\beta^T Z$, for a given v, let $P(t \mid v) = P(X \geq t \mid \beta^T Z = v)$, $\Lambda(t,v) = \int_0^t P(u \mid v)\lambda_0(u)du$, $Y(t) = I\{X \geq t\}$, $Y_i(t) = I\{X_i \geq t\}$, $H = diag\{h, \cdots, h^p\}^T$ and $\mathbf{u} = \{u, \cdots, u^p\}$.

We begin with some regularity conditions needed for the results.

(C1) $K \geq 0$ is a bounded density with compact support, and it has bounded first and second derivative.

(C2) $\psi(\cdot)$ has a continuous $(p+1)$th derivative around v.

(C3) The density $f(\cdot)$ of $\beta^T Z$ is continuous at point v and $\inf_w f(w) > 0$.

(C4) The conditional probability $P(t \mid \cdot)$ is equicontinuous at v.

(C5) $\int_0^\tau \lambda_0(u)du < \infty$.

Denote

$s_0(\beta, \psi, u) = E\big[Y(u)\exp\{\psi(\beta^T Z)\}\big]$,

$s_1(\beta, \psi, u) = E\big[Y(u)\exp\{\psi(\beta^T Z)\}\psi'(\beta^T Z)Z\big]$,

$s_2(\beta, \psi, u) = E\big[Y(u)\exp\{\psi(\beta^T Z)\}\big\{\psi''(\beta^T Z) + [\psi'(\beta^T Z)]^2\big\}ZZ^T\big]$,

$s_2^*(\beta, \psi, u) = E\big[Y(u)\exp\{\psi(\beta^T Z)\}[\psi'(\beta^T Z)]^2 ZZ^T\big]$.

(C6) The functions $s_r, r = 0, 1,$and 2, and s_2^* are bounded and s_0 is bounded away from 0 on $\mathcal{B} \times [0, \tau]$; the family of functions $s_r(\cdot, \psi, u)$ and $s_2^*(\cdot, \psi, u)$ is an equicontinuous family at β_0.

The following lemma is used repeatedly in the later proofs. The proof will be omitted and can be found in [19].

Lemma 1. *Let* $c_n(\beta_0, t) = n^{-1}\sum_{i=1}^n Y_i(t)g({\beta_0}^T Z_i)K_h({\beta_0}^T Z_i - v)$ *and* $c(t) = f(v)g(v)P(t \mid v)\int K(u)du$, *Under conditions (C1) and (C4), if* $g(\cdot)$ *is continuous at the point* v, *then*

$$\sup_{0 \leq t \leq \tau} |c_n(\beta_0, t) - c(t)| \to_P 0,$$

provided that $h \to 0, nh/\log n \to \infty, 0 < \tau \leq +\infty$.

If furthermore, $\hat{\beta}$ *is a* $\sqrt{n}$*-consistent estimate of* β_0 *and* $nh^4 \to \infty$, *then*

$$\sup_{0 \leq t \leq \tau} |c_n(\hat{\beta}, t) - c(t)| \to_P 0.$$

Proof of Theorem 1

Proof. For notation simplicity, we will use γ to represent $\gamma(v)$ and γ_0 for $\gamma_0(v)$.

The log local partial likelihood function at point v is given as

$$l\{\beta, \gamma, v\} = \frac{1}{n}\sum_{j=1}^N K_h(\beta^T Z_{(j)} - v) \left[\big[(\beta^T \mathbf{Z}_{(j)})^T\gamma\big] - \log\Big\{\sum_{i \in \mathcal{R}_j} \exp\{(\beta^T \mathbf{Z}_i)^T\gamma\}K_h(\beta^T Z_i - v)\Big\}\right].$$

Using counting process notation $N(t) = I\{X \leq t, \delta = 1\}$ and $N_i(t) = I\{X_i \leq t, \delta_i = 1\}$, under the independent censoring,

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)\exp\{\psi({\beta_0}^T Z_i)\}\lambda_0(u)du,$$

is a martingale with respect to the filtration $\mathcal{F}_t = \sigma\{N(u), I_{\{X \leq u, \delta = 0\}} : 0 \leq u \leq t\}$.

The empirical counterpart of $l\{\beta,\gamma,v\}$ up to time t is

$$l_n(\beta,\gamma,t,v)=\int_0^t \frac{1}{n}\sum_{i=1}^n K_h(\beta^T Z_i - v)\Big[\big[(\beta^T\mathbf{Z}_i)^T\gamma\big] - \log\big\{\sum_i Y_i(u)\exp\{(\beta^T\mathbf{Z}_i)^T\gamma\}K_h(\beta^T Z_i - v)\big\}\Big]dN_i(u).$$

Denote $S_{h,0}(\beta,\gamma,u,v)=\frac{1}{n}\sum_{i=1}^n K_h(\beta^T Z_i - v)Y_i(u)\exp\{(\beta^T\mathbf{Z}_i)^T\gamma\}$.

Let $\hat\beta$ be a $\sqrt{n}-$consistent estimate of the true parameter β_0, and $\hat\gamma$ be the corresponding estimate of the true γ_0, we can write

$$\begin{aligned}
&l_n(\hat\beta,\gamma,\tau,v)-l_n(\beta_0,\gamma_0,\tau,v)\\
=&\int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h({\beta_0}^T Z_i - v)\\
&\qquad\times\Big[\big[(\beta_0^T\mathbf{Z}_i)^T\gamma-(\beta_0^T\mathbf{Z}_i)^T\gamma_0\big]-\log\frac{S_{h,0}(\beta_0,\gamma,u,v)}{S_{h,0}(\beta_0,\gamma_0,u,v)}\Big]dM_i(u)\\
+&\int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h(\hat\beta^T Z_i - v)\Big[\big[(\hat\beta^T\mathbf{Z}_i)^T\gamma-(\beta_0^T\mathbf{Z}_i)^T\gamma\big]-\log\frac{S_{h,0}(\hat\beta,\gamma,u,v)}{S_{h,0}(\beta_0,\gamma,u,v)}\Big]dM_i(u)\\
+&\int_0^\tau \frac{1}{n}\sum_{i=1}^n \Big(K_h(\hat\beta^T Z_i - v)-K_h({\beta_0}^T Z_i - v)\Big)\\
&\qquad\times\Big[(\beta_0^T\mathbf{Z}_i)^T\gamma-\log S_{h,0}(\beta_0,\gamma,u,v)\Big]dM_i(u)\\
+&\int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h({\beta_0}^T Z_i - v)\Big[\big[(\beta_0^T\mathbf{Z}_i)^T\gamma-(\beta_0^T\mathbf{Z}_i)^T\gamma_0\big]-\log\frac{S_{h,0}(\beta_0,\gamma,u,v)}{S_{h,0}(\beta_0,\gamma_0,u,v)}\Big]\\
&\qquad\times Y_i(u)\exp\{\psi({\beta_0}^T Z_i)\}\lambda_0(u)du\\
+&\int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h(\hat\beta^T Z_i - v)\Big[\big[(\hat\beta^T\mathbf{Z}_i)^T\gamma-(\beta_0^T\mathbf{Z}_i)^T\gamma\big]-\log\frac{S_{h,0}(\hat\beta,\gamma,u,v)}{S_{h,0}(\beta_0,\gamma,u,v)}\Big]\\
&\qquad\times Y_i(u)\exp\{\psi({\beta_0}^T Z_i)\}\lambda_0(u)du\\
+&\int_0^\tau \frac{1}{n}\sum_{i=1}^n \Big(K_h(\hat\beta^T Z_i - v)-K_h({\beta_0}^T Z_i - v)\Big)\Big[(\beta_0^T\mathbf{Z}_i)^T\gamma-\log S_{h,0}(\beta_0,\gamma,u,v)\Big]\\
\equiv&X_n(\beta_0,\gamma,\tau,v)+I+II+A_n(\beta_0,\gamma,\tau,v)+III+IV.
\end{aligned}$$

Under the regularity conditions and from Lemma 1, it can be shown that

(1) $X_n(\beta_0,\gamma,\tau,v)$ is a locally square integrable martingale with the predictable variation process

$$\begin{aligned}
&\langle X_n(\beta_0,\gamma,\tau,v),X_n(\beta_0,\gamma,\tau,v)\rangle\\
=&\int_0^\tau \frac{1}{n^2}\sum_{i=1}^n K_h^2({\beta_0}^T Z_i - v)\Big[(\beta_0^T\mathbf{Z}_i)^T(\gamma-\gamma_0)-\log\frac{S_{h,0}(\beta_0,\gamma,u,v)}{S_{h,0}(\beta_0,\gamma_0,u,v)}\Big]^2\\
&\qquad\times Y_i(u)\exp\{\psi({\beta_0}^T Z_i)\}\lambda_0(u)du\\
=&O_p(\frac{1}{nh}).
\end{aligned}$$

(2) $A_n(\beta_0,\gamma,\tau,v) \to_p$

$$f(v)\exp\{\psi(v)\}\Lambda(\tau,v) \times \left[(\int \mathbf{u}K(u)du)^T H(\gamma-\gamma_0) - \log\{\int \exp\{\mathbf{u}^T H(\gamma-\gamma_0)\}K(u)du\}\right] + o_p(1) \equiv A(\beta_0,\gamma,\tau,v) + o_p(1),$$

(3) $I = O_p(\frac{1}{nh})$, $II = O_p(\frac{1}{nh^2})$, $III = O_p(\frac{1}{\sqrt{nh^2}})$, and $IV = O_p(\frac{1}{\sqrt{nh^4}})$.

This means $X_n(\beta_0,\gamma,\tau,v)$, I, II, III and IV converge to zero at a faster rate than $A_n(\beta_0,\gamma,\tau,v)$. By Lemma 8.2.1(2) in [9], $l_n(\hat\beta,\gamma,\tau,v) - l_n(\beta_0,\gamma_0,\tau,v)$ has the same limiting distribution as $A_n(\beta_0,\gamma,\tau,v)$. Thus, we have

$$l_n(\hat\beta,\gamma,\tau,v) - l_n(\beta_0,\gamma_0,\tau,v) \to_p A(\beta_0,\gamma,\tau,v). \tag{4}$$

It is obvious that $A(\beta_0,\gamma,\tau,v)$ is strictly concave, with a maximum at $\gamma=\gamma_0$. Hence, the right-hand side of (4) is maximized at $\gamma=\gamma_0$. The left-hand side of (4) is maximized at $\gamma=\hat\gamma$, since $\hat\gamma$ maximizes $l_n(\hat\beta,\gamma,\tau,v)$. Therefore, $\sup_v \left|\hat\gamma(v) - \gamma_0(v)\right| \to_p 0$.

By Dominated Convergence Theorem, we have

$$\sup_v \left|\hat\psi(v) - \psi(v)\right| \to_p 0. \tag{5}$$

This implies

$$\sup_z \left|\hat\psi(\hat\beta^T z) - \psi({\beta_0}^T z)\right| \le \sup_z \left|\hat\psi(\hat\beta^T z) - \psi(\hat\beta^T z)\right| + \sup_z \left|\psi(\hat\beta^T z) - \psi({\beta_0}^T z)\right| \to_p 0,$$

where the second term converges to zero by continuity of ψ and $\sqrt{n}$-consistency of $\hat\beta$. Theorem 1 is thus proved.

Proof of Theorem 2

Proof. Let $\eta = H\gamma$, we can write the log local partial likelihood function in terms of β and η

$$l_n(\beta,\eta,\tau,v) = \int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h(\beta^T Z_i - v)\Big[(\beta^T\mathbf{Z}_i)^T H^{-1}\eta - \log\{\sum_i^n Y_i(u)\exp\{(\beta^T\mathbf{Z}_i)^T H^{-1}\eta\}K_h(\beta^T Z_i - v)\}\Big]dN_i(u).$$

Accordingly let $S_{h,0}(\beta,\eta,u,v) = \frac{1}{n}\sum_{i=1}^n K_h(\beta^T Z_i - v)Y_i(u)\exp\{(\beta^T\mathbf{Z}_i)^T H^{-1}\eta\}$, and $S_{h,1}(\beta,\eta,u,v) = \frac{1}{n}\sum_{i=1}^n K_h(\beta^T Z_i - v)Y_i(u)\exp\{(\beta^T\mathbf{Z}_i)^T H^{-1}\eta\}(\beta^T\mathbf{Z}_i)^T H^{-1}$, and for a $\sqrt{n}$-consistent estimate $\hat\beta$ of β_0, Lemma 1 implies

$$\sup_{0\le u<\tau}\left|\frac{S_{h,1}(\hat\beta,\eta,u,v)}{S_{h,0}(\hat\beta,\eta,u,v)} - \nu_1\right| \to_p 0,$$

where $\nu_1 = \int \mathbf{u}K(u)du$.

The derivative of $l_n(\beta,\eta,\tau,v)$ with respect to η evaluated at $\hat{\beta}$ and $\eta_0 = H\gamma_0$ is

$$\begin{aligned}
&l_n'(\hat{\beta},\eta_0,\tau,v)\\
=&\int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v)\left[(\hat{\beta}^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\hat{\beta},\eta_0,u,v)}{S_{h,0}(\hat{\beta},\eta_0,u,v)}\right] dN_i(u)\\
=&\int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v)\left[(\hat{\beta}^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\hat{\beta},\eta_0,u,v)}{S_{h,0}(\hat{\beta},\eta_0,u,v)}\right] dM_i(u)\\
+&\int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v)\left[(\hat{\beta}^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\hat{\beta},\eta_0,u,v)}{S_{h,0}(\hat{\beta},\eta_0,u,v)}\right]\\
&\qquad\qquad \times Y_i(u)\exp\{\psi({\beta_0}^T Z_i)\}\lambda_0(u)du\\
\equiv& U_n(\hat{\beta},\eta_0,\tau,v) + B_n(\hat{\beta},\eta_0,\tau,v).
\end{aligned}$$

The first term

$$\begin{aligned}
&U_n(\hat{\beta},\eta_0,\tau,v)\\
=&\int_0^t \frac{1}{n}\sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v)\left[(\hat{\beta}^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\hat{\beta},\eta_0,u,v)}{S_{h,0}(\hat{\beta},\eta_0,u,v)}\right] dM_i(u)\\
=&\int_0^t \frac{1}{n}\sum_{i=1}^n K_h({\beta_0}^T Z_i - v)\left[(\beta_0^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\beta_0,\eta_0,u,v)}{S_{h,0}(\beta_0,\eta_0,u,v)}\right] dM_i(u)\\
+&\int_0^t \frac{1}{n}\sum_{i=1}^n \Big(K_h(\hat{\beta}^T Z_i - v) - K_h({\beta_0}^T Z_i - v)\Big)\\
&\qquad\qquad \times \left[(\beta_0^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\beta_0,\eta_0,u,v)}{S_{h,0}(\beta_0,\eta_0,u,v)}\right] dM_i(u)\\
+&\int_0^t \frac{1}{n}\sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v)\\
&\qquad\qquad \times \left[(\hat{\beta}^T\mathbf{Z}_i - \beta_0^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\hat{\beta},\eta_0,u,v)}{S_{h,0}(\hat{\beta},\eta_0,u,v)} - \frac{S_{h,1}(\beta_0,\eta_0,u,v)}{S_{h,0}(\beta_0,\eta_0,u,v)}\right] dM_i(u)\\
\equiv& U_n(\beta_0,\eta_0,\tau,v) + V + VI.
\end{aligned}$$

It is clear that $\sqrt{nh}U_n(\beta_0,\eta_0,t,v)$ is a martingale with predictable variation

$$\begin{aligned}
&\langle\sqrt{nh}U_n(\beta_0,\eta_0,t,v), \sqrt{nh}U_n(\beta_0,\eta_0,t,v)\rangle\\
=&\frac{nh}{n^2}\sum_{i=1}^n \int_0^\tau K_h^2({\beta_0}^T Z_i - v)\left[(\beta_0^T\mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\beta_0,\eta_0,u,v)}{S_{h,0}(\beta_0,\eta_0,u,v)}\right]^{\otimes 2}\\
&\qquad\qquad \times Y_i(u)\exp\{\psi({\beta_0}^T Z_i)\}\lambda_0(u)du.\\
=&f(v)\exp\{\psi(v)\}\Lambda(t,v)\int K^2(u)(\mathbf{u}-\nu_1)^{\otimes 2}du + o_p(1) \equiv \Sigma_U(t,v) + o_p(1),
\end{aligned}$$

where the last step follows from Lemma 1.

The Lindberg conditions are satisfied (see [19] for details), we have thus proven that

$$\sqrt{nh}U_n(\beta_0,\eta_0,\tau,v) \to_D N\big(0,\Sigma_U(\tau,v)\big). \tag{6}$$

As for the term V and VI, similarly to the proof of Theorem 1, we have

$$V = \int_0^\tau \frac{1}{n}\sum_{i=1}^n \Big(K_h(\hat{\beta}^T Z_i - v) - K_h({\beta_0}^T Z_i - v)\Big) \times \left[(\beta_0^T \mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\beta_0, \eta_0, u, v)}{S_{h,0}(\beta_0, \eta_0, u, v)}\right] dM_i(u) = O_p(\frac{1}{nh^2}), \tag{7}$$

and

$$VI = \int_0^\tau \frac{1}{n}\sum_{i=1}^n K_h(\hat{\beta}^T Z_i - v) \times \left[(\hat{\beta}^T \mathbf{Z}_i - \beta_0^T \mathbf{Z}_i)^T H^{-1} - \frac{S_{h,1}(\hat{\beta}, \eta_0, u, v)}{S_{h,0}(\hat{\beta}, \eta_0, u, v)} - \frac{S_{h,1}(\beta_0, \eta_0, u, v)}{S_{h,0}(\beta_0, \eta_0, u, v)}\right] dM_i(u) =O_p(\frac{1}{nh}). \tag{8}$$

Applying Lemma 1 again and by Taylor expansion we get

$$B_n(\hat{\beta}, \eta_0, \tau, v) = f(v)\exp\{\psi(v)\}\frac{\psi^{(p+1)}(v)}{(p+1)!}\Lambda(\tau, v)\int K(u)(\mathbf{u} - \nu_1)u^{p+1}duh^{p+1} + o_p(h^{p+1}) + O_p(\frac{1}{\sqrt{n}}) =b(\tau, v) + o_p(h^{p+1}) + O_p(\frac{1}{\sqrt{n}}). \tag{9}$$

We have thus shown that, under (6), (7), (8) and (9),

$$\sqrt{nh}l_n'(\hat{\beta}, \eta_0, \tau, v) \to_D N\Big(b(\tau, v), \Sigma_U(\tau, v)\Big). \tag{10}$$

Next we focus on the property of the second derivative $l_n''(\hat{\beta}, \eta, t, v)$. Let $\hat{\eta} = H\hat{\gamma}$, by Taylor expansion and Lemma 1 we have

$$0 = l_n'(\hat{\beta}, \hat{\eta}, \tau, v) = l_n'(\hat{\beta}, \eta_0, \tau, v) + l_n''(\hat{\beta}, \eta^{**}, \tau, v)(\hat{\eta} - \eta_0), \tag{11}$$

where η^{**} lies in between $\hat{\eta}$ and η_0. Theorem 1 implies $\hat{\eta} \to_p \eta_0$, hence $\eta^{**} \to_p \eta_0$.

Using condition $(C1)$ and boundedness of $\hat{\beta}^T Z$, we arrive at

$$l_n''(\hat{\beta}, \eta^{**}, \tau, v) = l_n''(\hat{\beta}, \eta_0, \tau, v) + o_p(1) = \Sigma_l(\tau, v) + o_p(1). \tag{12}$$

By (10), (11), (12) and Slutsky's theorem,

$$\sqrt{nh}(\hat{\eta} - \eta_0) = \sqrt{nh}\Big[-l_n''(\hat{\beta}, \eta^{**}, \tau, v)^{-1} l_n'(\hat{\beta}, \eta_0, \tau, v)\Big] + o_p(1) \to_D N\Big(b(\tau, v), \Sigma_l(\tau, v)^{-1}\Sigma_U(\tau, v)\Sigma_l(\tau, v)^{-1}\Big).$$

Simple calculations lead to the result in Theorem 2.

References

[1] Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, **30**, 89–99.

[2] Chen, C. H. and Li, K. C. and Wang, J. L. (1999). Dimension reduction for censored regression data. *The Annals of Statistics*, **27**, 1–23.

[3] Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, **82**, 835–845.

[4] Cox, D. R. (1972). Regression models and life-table (with discussion). *Journal of the Royal Statistical Society, Ser.*, **B. 4**, 187–220.

[5] Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.

[6] Dabrowska, D. M. and Doksum, K. A. (1988). Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics*, **15**, 1–23.

[7] Doksum, K. A. (1987). An extension of partial likelihood methods for proportional hazard models to general transformation models. *The Annals of Statistics*, **15**, 325–345.

[8] Fan, J., Gijbels, I. and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *The Annals of Statistics*, **25**, 1661–1690.

[9] Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, Inc., New York.

[10] Gentleman, R. and Crowley, J. (1991). Local full likelihood estimation for the proportional hazards model. *Biometrics*, **47**, 1283–1296.

[11] Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, **420**, 942–951.

[12] Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, **3**, 297–318.

[13] Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, **46**, 1005–1016.

[14] Hosmer, D. W., Lemeshow, S. and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition*. John Wiley & Sons, Inc., New York.

[15] Huang, J. and Liu, L. (2006). Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics*, **62**, 793–802.

[16] O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing*, **9**, 531–542.

[17] Sleeper, L. A. and Harrington, D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, **85**, 941–949.

[18] Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, **82**, 559–567.

[19] Wang, W (2001). Proportional hazards model with unknown link function and applications to time-to-event longitudinal data. *Ph.D. Thesis*.

[20] Wang, W (2004). Proportional hazards regression with unknown link function and time-dependent covariates. *Statistica Sinica*, **14**, 885–905.

IMS Collections
Vol. 0 (2009) 66–92
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000003

Semiparametric Models and Likelihood - The Power of Ranks

Kjell Doksum[1,*] **and Akichika Ozeki**[2,*]

University of Wisconsin, Madison

Abstract: We consider classes of models related to those introduced by Lehmann in 1953 and Sklar in 1959. Recently developed algorithms for finding profile NP likelihood procedures are discussed, extended and implemented for such models by combining them with the MM algorithm. In particular we consider statistical procedures for a regression model with proportional expected hazard rates, and for transformation models including the normal copula. A variety of likelihoods introduced to deal with semiparametric models are considered. They all generate rank results, not only tests, but also estimates, confidence regions, and optimality theory, thereby, to paraphrase Lehmann (1953), demonstrating "the power of ranks".

Contents

1 Introduction 67
1.1 Lehmann Type Models. Cox Regression 67
1.2 Sklar Type Models. Copula Regression 67
2 Proportional Hazard and Proportional Expected Hazard Rate Models . . 68
3 Rank, Partial and Marginal Likelihood 70
4 Profile NP Likelihood 72
5 Profile NP Likelihood for the PEHR Model 73
5.1 The MM Algorithm 74
5.2 The MM Algorithm for the PEHR Model with $\theta \geq 0$ (SINAMI with $\theta \geq 0$) 74
5.3 The MM Algorithm for the SINAMI Model with $\theta \leq 0$ 75
5.4 The MM Algorithm for the SINAMI Model with $\theta \in R$ 75
5.5 Profile NPMLE Implementation 75
5.6 Estimation of the Variance of the Profile NPMLE 77
6 Simulation Results 77
6.1 PEHR Model Estimates 77
6.2 Model Fit for Misspecified Model 78
7 Estimation in the Normal Copula Model 78
7.1 The One Covariate Case 78
7.2 The Multivariate Covariate Case 81
8 Transformation and NP Models 81
8.1 Simulation Results 84

[1]Department of Statistics, University of Wisconsin, Madison, WI 53706, email: doksum@stat.wisc.edu
*Supported in part by NSF grant DMS-0505651.
[2]Department of Statistics, University of Wisconsin, Madison, WI 53706, email: ozeki@stat.wisc.edu
AMS 2000 subject classifications: Primary 62G05, 62G20; secondary 62N02
Keywords and phrases: Lehmann model, proportional hazard model, profile NP likelihood, nonparametric maximum likelihood, MM algorithm, copula models, Box-Cox models

8.1.1 Correctly Specified Model 84
8.1.2 Misspecified Model 84
Acknowledgements 91
References 91

1. Introduction

We will focus on statistical inference for models where the distribution of the data can be expressed as a parametric function of unknown distribution functions.

1.1. Lehmann Type Models. Cox Regression

Suppose T is a random variable with a continuous distribution function F. For testing the null hypothesis $H_0 : F = F_0$, Lehmann (1953) considered alternatives of the form

$$F_\theta(\cdot) = C_\theta(F_0(\cdot)), \tag{1.1}$$

for some continuous distribution $C_\theta(\cdot)$ on $[0, 1]$, which is known except for the parameter θ. We consider the problem of estimating θ when $F_0(\cdot)$ is an unknown baseline distribution. In this case, if $T_1, \cdots, T_n$ are independent with $T_i \sim C_{\theta_i}(F_0(\cdot))$ and we set $U_i = F_0(T_i)$, then U_i has distribution $C_{\theta_i}(\cdot)$ Moreover $R_i \equiv Rank(T_i) = Rank(U_i)$, which implies that the distribution of any statistical method based on $R_1, \cdots, R_n$ will not depend on F_0.

For regression experiments with observations $(T_i, \boldsymbol{x}_i), i = 1, \cdots, n$, where T_i is a response and $\boldsymbol{x}_i$ is a vector of nonrandom covariates, Cox (1972) considered the parametrization $\theta_i = g(\boldsymbol{\beta}^T \boldsymbol{x}_i)$ with $g(\cdot)$ a known function and $\boldsymbol{\beta}$ a vector of regression coefficients. He considered statistical inference procedures based on the Cox (1972, 1975) partial likelihood in very general frameworks. These procedures are based on generalized ranks and show how powerful ranks are in generating statistical inference procedures.

In this paper we consider a special case of (1.1) obtained from the Lehmann models $[F_0(t)]^N$ and $1 - [1 - F_0(t)]^N$ by letting N be a zero truncated Poisson variable whose parameter depends on covariates and regression coefficients. We call this model "SINAMI" after SIbuya (1968) and NAbeya and MIura (1972). For a subset of the parameter space, the model has proportional expected hazard rate (PEHR). We show that semiparametric likelihood methods for the SINAMI model give more weight to intermediate survival times than the Cox proportional hazard model which heavily weights long survival times. Recently developed algorithms for finding profile nonparametric maximum likelihood estimates (profile NPMLE's) are combined with the MM algorithm to produce estimates. In the two sample case, we carry out a Monte Carlo comparison of the NPMLE with a parametric MLE and a method of moment (MOM) estimate of the two sample parameter. The NPMLE is nearly unbiased but only about 70% as efficient in terms of root MSE as the parametric estimate if the parametric model is true. The MOM estimate is slightly less efficient than the NPMLE.

1.2. Sklar Type Models. Copula Regression

Suppose X and Y are random variables with continuous joint distribution $H(\cdot, \cdot)$ and marginals $F_1(\cdot)$ and $F_2(\cdot)$. Sklar (1959) considered models that include models

of the form

$$H_\theta(\cdot,\cdot) = C_\theta(F_1(\cdot), F_2(\cdot)), \tag{1.2}$$

for some continuous distribution $C_\theta(\cdot,\cdot)$ on $[0,1]\times[0,1]$, which is known except for the parameter θ. We consider the problem of estimating θ when $F_1(\cdot)$ and $F_2(\cdot)$ are unknown baseline distributions. If we set $U = F_1(X)$, $V = F_2(Y)$, then (U,V) has distribution $C_\theta(\cdot,\cdot)$, and $C_\theta(\cdot,\cdot)$ is called a *copula*. Note that if $(X_1,Y_1),\cdots,(X_n,Y_n)$ are independent with $(X_i,Y_i) \sim H_{\theta_i}(\cdot,\cdot)$, then $R_i \equiv Rank(X_i) = Rank(F_1(X_i))$ and $S_i \equiv Rank(Y_i) = Rank(F_2(Y_i))$, which shows that the distribution of any statistical method based on these ranks will not depend on (F_1,F_2). This model extends in the natural way to the d-dimensional case.

In this paper we consider the bivariate normal copula model where $C_\theta(u,v) = \Phi_\theta(\Phi^{-1}(u),\Phi^{-1}(v))$ with C_θ the bivariate $N(0,0,1,1,\theta)$ distribution. We also consider the multivariate normal copula model and show that in regression experiments it can be used to construct a "transform both sides regression" transformation model (copula regression model.) Klaassen and Wellner (1997) have shown that the normal scores correlation coefficient is semiparametrically efficient for the bivariate normal copula. We use simulations to compare this estimate with the profile MLE for the transform both sides Box-Cox regression model and a nonparametric estimate based on splines thereby augmenting the comparisons made by Zou and Hall (2002). The normal scores estimate is nearly as efficient as the parametric MLE for estimating median regression when the transform both sides Box-Cox model is correct. We also consider the performance of the estimates for models outside the copula regression model and find that the normal scores based estimate of median regression is remarkably robust with respect to both bias and variance. On the other hand, the profile MLE of median regression derived from the transform both sides Box-Cox model is very sensitive to deviations from the model. The nonparametric spline estimate is the best for extreme deviations from the copula regression model.

2. Proportional Hazard and Proportional Expected Hazard Rate Models

Interesting special cases of (1.1) are obtained by considering the distributions of $T_1 = \min(T_{01},\cdots,T_{0k})$ and $T_2 = \max(T_{01},\cdots,T_{0k})$ where $T_{01},T_{02},\cdots$ are i.i.d. as $T_0 \sim F_0$. Then, for $k \geq 1$,

$$T_1 \sim F_1(t) = 1-[1-F_0(t)]^k, \tag{2.1}$$

and

$$T_2 \sim F_2(t) = [F_0(t)]^k, \tag{2.2}$$

with $t > 0, k = 1,2,\cdots$. More general forms (Lehmann (1953); Savage (1956)) are

$$T_1 \sim F_1(t) = 1-[1-F_0(t)]^\Delta, \tag{2.3}$$

and

$$T_2 \sim F_2(t) = [F_0(t)]^\Delta, \tag{2.4}$$

with $t > 0, \Delta > 0$. Here (2.3) can be derived by considering two-sample models where the two samples follow distributions of the form (2.1) with different k's (Bickel and Doksum (2007), Problem 1.1.12.)

For T_1, the hazard rate is

$$\lambda(t) \equiv \frac{f(t)}{1-F(t)} = \triangle \frac{f_0(t)}{1-F_0(t)} \equiv \triangle \lambda_0(t). \tag{2.5}$$

In regression experiments, we set $\triangle_i = g(\boldsymbol{\beta}^T \boldsymbol{x}_i)$, and note that (2.5) is the Cox proportional hazard (PH) model (Cox (1972)).

Nabeya and Miura (1972) proposed replacing k in (2.1) and (2.2) by a random variable. In particular, they considered $T_1 = \min(T_{01}, \cdots, T_{0N})$, where N is independent of $T_{01}, T_{02}, \cdots$, and has a zero truncated Poisson(θ) distribution with $\theta > 0$. They also considered $T_2 = \max(T_{01}, \cdots, T_{0M})$, $T_{0i} \sim F_0$ where M is independent of $T_{01}, T_{02}, \cdots$, and has a zero truncated Poisson($-\theta$) distribution with $\theta < 0$.

Using Sibuya (1968), they found

$$T_1 \sim F_1(t) = \frac{1-e^{-\theta F_0(t)}}{1-e^{-\theta}}, \quad \theta > 0, \tag{2.6}$$

$$T_2 \sim F_2(t) = \frac{1-e^{-\theta F_0(t)}}{1-e^{-\theta}}, \quad \theta < 0. \tag{2.7}$$

Combining (2.6) and (2.7), we get

$$\begin{aligned} T \sim F(t) = & \frac{1-e^{-\theta F_0(t)}}{1-e^{-\theta}}, \quad \theta \neq 0, \\ = & \ F_0(t), \qquad\qquad \theta = 0. \end{aligned} \tag{2.8}$$

Note that model (2.6) is a mixture of proportional hazard models for individuals with the same baseline hazard rate $\lambda_0(\cdot)$ but different hazard factors $\triangle$ in the factorization (2.5) of the hazard rate. Let $\lambda(t;k)$ denote the hazard rate of T_1 given $N = k$; then by (2.5)

$$E\lambda(t;N) = \sum_{k=1}^{\infty} k\lambda_0(t) p_\theta(k) = \tau(\theta)\lambda_0(t), \ \theta > 0, \tag{2.9}$$

where $p_\theta(x)$ is the zero truncated Poisson(θ) probability and

$$\tau(\theta) = E(N) = \frac{\theta}{1-\exp(-\theta)}. \tag{2.10}$$

Thus (2.6) is a model with proportional expected hazard rate. Note that (2.8) does not have this property for $\theta < 0$. We will refer to (2.6) and (2.8) as the PEHR and SINAMI models, respectively.

***Remark* 2.1 :** In regression experiments, the traditional frailty models are also constructed by introducing a random element in the PH model. However, these models are different from the PEHR and SINAMI models. To see this recall that in the frailty model the conditional hazard rate given the covariate vector $\boldsymbol{x}$ (see Oakes (1992)) for the history and interpretation of frailty models) is of the form

$$\lambda_W(t|\boldsymbol{x}) = \lambda_0(t) W \exp[\boldsymbol{\beta}^T \boldsymbol{x}], \tag{2.11}$$

where W is a random effect that incorporates potential unobservable covariates that represent frailties. Semiparametric optimality theory for model (2.11) has been developed by Kosorok, Lee, and Fine (2004).

Consider model (2.5) with $\Delta = N$ and N a zero truncated Poisson(θ) random variable with $\theta = g(\boldsymbol{\beta}^T\boldsymbol{x})$, i.e., the conditional hazard rate given $\boldsymbol{x}$ is

$$\lambda^{(N)}(t|\boldsymbol{x}) = N\lambda_0(t). \tag{2.12}$$

Here N plays the role of $W\exp[\boldsymbol{\beta}^T\boldsymbol{x}]$ in (2.11). However, (2.11) and (2.12) are different because N is an integer and $W\exp[\boldsymbol{\beta}^T\boldsymbol{x}]$ is not when $\boldsymbol{\beta} \neq 0$. In model (2.12), N represents the effect of both observed covariates and frailties. In deriving the distribution function (2.6), the unobservable covariates are averaged out, that is, we compute $P(T \leq t) = E[P(T \leq t|N)]$.

***Remark* 2.2 :** Model (2.8) was considered by Bell and Doksum (1966), Example 5.2 and Table 8.1) and Ferguson (1967, p.257, Problem 5.7.7) without any of the above interpretations. Nabeya and Miura (1972) did not use any proportional hazard or frailty interpretation. These concepts had not been invented yet.

Fig.1 gives a plot of the relative hazard rate $\lambda(t|x = 1)/\lambda(t|x = 0)$ with $\theta = -3, -1.5, 1.5, 3$ for model (2.8) with $\theta = \beta x$, $F_0(t) = 1 - \exp(-t)$, $t > 0$, and

$$\lambda(t|x) = \frac{\theta f_0(t)}{1 - e^{-\theta(1-F_0(t))}}. \tag{2.13}$$

In the PEHR and SINAMI models, the hazard ratio between two covariate values converge to unity as time increases. This explains why the likelihoods for these models give less weight to long survival times than the likelihood for the Cox model (see Section 3). The hazard rate is decreasing for the PEHR model for any continuous F_0.

3. Rank, Partial and Marginal Likelihood

In regression experiments, we observe $(T_i, \boldsymbol{x}_i)$, $i = 1, \cdots, n$, where $T_1, \cdots, T_n$ are independent responses and $\boldsymbol{x}_i$ is a nonrandom covariate vector. In the proportional hazard model, it is customary to use model (2.5) for T_i with $\triangle_i = \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ because $\triangle_i$ needs to be positive. In the PEHR model, $\theta_i = \boldsymbol{\beta}^T\boldsymbol{x}_i$ is a possible parametrization, but $\theta_i = \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ could also be used. Let $\boldsymbol{R} = (R_1, \cdots, R_n)$ where $R_i = \text{Rank}(T_i)$, then $l_{\boldsymbol{r}}(\boldsymbol{\beta}) = P(\boldsymbol{R} = \boldsymbol{r})$ is the *rank likelihood* (Hoeffding (1951)).

We first consider the one covariate case. Using the rank likelihood, the locally most powerful (LMP) rank test statistic for $H_0 : \beta = 0$ versus $H_1 : \beta > 0$ is (approximately) for the Cox model (Savage (1956), 1957), Cox (1964)), Oakes and Jeong (1998)):

$$\sum_{i=1}^{n}[-\log(1 - \frac{R_i}{n+1})](x_i - \bar{x}) \quad \text{(Savage or log rank)}, \tag{3.1}$$

and for the PEHR and SINAMI models, the LMP rank test statistics is (Bell and Doksum (1966), Ferguson (1967), Nabeya and Miura (1972)):

$$\sum_{i=1}^{n}\frac{R_i}{n+1}(x_i - \bar{x}) \quad \text{(Wilcoxon type)}, \tag{3.2}$$

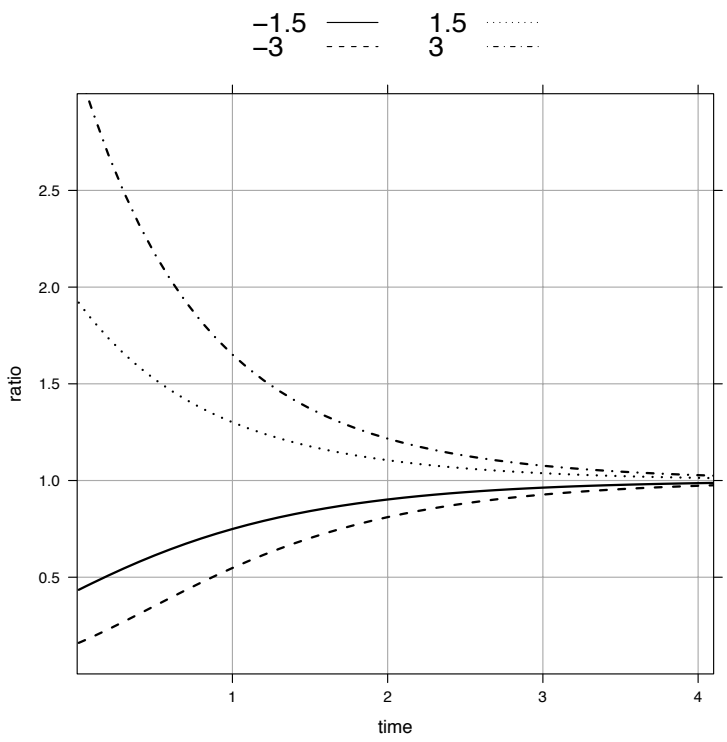


FIG 1. *SINAMI and PEHR hazard ratios for* $\theta = -3, -1.5, 1.5, 3$.

where R_i =Rank(T_i). The log rank statistic gives more weights to large observations, that is, in survival analysis, to those that live longer, while the Wilcoxon statistics is even handed.

In order to compare how much relative weight is given to the small, in between, and large observed survival times for the PH and PEHR models, we next consider the rank likelihood for d covariates. Note that if $h(\cdot)$ is decreasing, then Rank$(h(T_i)) = n + 1 - R_i$. For the proportional hazard model, transform T_i by $U_i = 1 - F_0(T_i)$, then by (2.3) we have $f_{U_i}(u) = \triangle_i u^{\triangle_i - 1}, 0 < u < 1$. Hoeffding (1951) formula shows,

$$\text{Rank lkhd} = \Pi_{i=1}^n \triangle_i \int \cdots \cdots \int_{0<u_1<u_2<\cdots<u_n<1} \Pi_{i=1}^n u_i^{\delta_i - 1} du_1 \cdots du_n, \tag{3.3}$$

where $\delta_i = \triangle_{b_i}$ and b_i= index on the T with rank $n + 1 - i$= reverse anti-rank. It follows that

$$\text{Rank lkhd} \propto \Pi_{i=1}^n \frac{\triangle_i}{\sum_{k:T_k \geq T_{(i)}} \triangle_k}, \tag{3.4}$$

that is, the familiar Cox (1972, 1975) *partial likelihood* formula. Here $\{k : T_k \geq T_{(i)}\}$= patients at risk at time $T_{(i)}$ where $T_{(i)}$ is the ith ordered survival time. Kalbfleich and Prentice (1973, 2002) called the rank likelihood the *marginal likelihood* and extended it to censored data.

For the PEHR model, transform T_i by the decreasing function $U_i = a^{-1}\{\exp\{-F_0(T_i)\} - b\}$, then, we have $f_i(u) = a\tau_i(au + b)^{\theta_i - 1}$, $0 < u < 1$,

where $b = e^{-1}$, $a = 1 - b$, and $\tau_i \equiv \tau(\theta_i)$. Let $\gamma_i = \theta_{b_i}$, b_i=index on the T with rank $n + 1 - i$. Then,

$$\text{Rank lkhd} \propto (\Pi_{i=1}^n \tau_i) \int \cdots \cdots \int_{0<u_1<u_2<\cdots<u_n<1} \Pi_{i=1}^n (au_i + b)^{\gamma_i - 1} du_1 \cdots du_n. \tag{3.5}$$

If we perform the integration, we find that the likelihood for the PEHR model is similar to the likelihood for the Cox model except that in addition to terms involving $\{k : T_k \geq T_{(i)}\}$, $i = 1, \cdots, n$, it includes terms involving $\{k : T_{(i)} \leq T_k \leq T_{(j)}\}$, $i = 1, \cdots, n$, $j = 1, \cdots, n$, $i \neq j$. That is, the PEHR likelihood gives more weight to the intermediate survival times than the Cox likelihood.

Computationally, the Cox rank likelihood is easier than the PEHR rank likelihood. However, we can handle the PEHR rank likelihood with available algorithms and software (e.g. MATLAB.) More generally, $F(x) = C_\theta(F_0(x))$ type models, originally considered by Lehmann (1953), can be handled effectively by considering the profile NP likelihood of the next section (e.g. Tsodikov and Gabribotti (2007)), Zeng and Lin (2007)).

4. Profile NP Likelihood

Andersen, Borgan, Gill, and Keiding (1996), Bickel, Klaassen, Ritov, and Wellner (1993, 1996), van der Vaart (1998), Murphy and van der Vaart (2000), Tsodikov and Garibotti (2007), Zeng and Lin (2007) and many others considered the problem of finding the MLE of all the parameters in a semiparametric model. It is useful to divide the procedure into two steps by grouping parameters into two groups. Suppose the distribution function of T is of the form $P(T \leq t) = F(\theta, \eta(t))$, where $\theta \in R^d$ and $\eta(\cdot)$ is a nondecreasing function. If we assume temporarily that $\eta(\cdot)$ has a positive derivative $\eta'(t)$ for $t \in \{t_1, \cdots, t_n\}$, then the likelihood is

$$\prod_{i=1}^n \eta'(t_i) f(\theta, \eta(t_i)),$$

where $f(\theta, \eta) = \partial F(\theta, \eta)/\partial \eta$. The NP likelihood we consider is of the form

$$L_{NP}(\theta, \eta) = \prod_{i=1}^n \eta\{t_i\} f(\theta, \eta[t_i]),$$

where $\eta[t_i] = \sum_{j \leq i} \eta\{t_i\}$ is a step function with positive jumps $\eta\{t_i\}$ at the data points t_i, $i = 1, \cdots, n$.

We assume that

$$af(\theta, a) \to 0 \quad as\ a \to \infty$$

Next we fix θ, and define $\hat{\eta}_\theta\{t_i\}$ as

$$\hat{\eta}_\theta\{.\} = ARG\ MAX_{\eta\{.\}} L_{NP}(\theta, \eta). \tag{4.1}$$

Set

$$PROF\ NPLIK = l(\theta) = MAX_{\eta\{.\}} L_{NP}(\theta, \eta), \tag{4.2}$$

and solve

$$\hat{\theta} = ARG\ MAX\ l(\theta). \tag{4.3}$$

Next estimate $\eta\{\cdot\}$ as $\hat{\eta}_\theta\{\cdot\}$. In the Lehmann model (1.1), the NP likelihood is

$$\prod_{i=1}^{n} F_0\{x_i\} C'_\theta(\sum_{j\le i} F_0\{x_j\}),$$

The method is similar to finding the empirical MLE, Owen (1988, 2001), and profile (partial) MLE's as in Andersen, et al. (1996), and Murphy and van der Vaart (2000).

***Remark* 4.1 :** Note that when $P(T \le t) = F(\theta, \eta(t))$, (4.1), (4.2), and (4.3) do not depend on the values of $t_1, \cdots, t_n$. In regression experiments, they will depend on the ranks of $t_1, \cdots, t_n$. For example, see (4.4) and (5.1). This is in contrast to the Hodges and Lehmann (1963) approach that uses estimating equations based on rank test statistics to obtain estimates of parameters. In this Hodges-Lehmann "rank inversion" approach, estimates are functions of the "raw" data rather than the ranks.

As an example that will guide the algorithm for the PEHR model, consider the Cox model. Set $\Lambda(t) = -\log(1 - F_0(t))$, then,

$$L_{NP}(\boldsymbol{\beta}, \Lambda) = \prod_{i=1}^{n} e^{\beta^T x_i} \Lambda\{t_i\} e^{-(e^{\beta^T x_i})\Lambda[t_i]},$$

where

$$\Lambda[t_i] = \sum_{j:t_j \le t_i} \Lambda\{t_j\}.$$

Using calculus, we find

$$\hat{\Lambda}_\beta\{t_i\} = ARG\ MAX_{\Lambda\{t_i\}}\ L_{NP}(\boldsymbol{\beta}, \Lambda) = (\sum_{j:t_j \ge t_i} e^{\beta^T x_j})^{-1},$$

and

$$l(\boldsymbol{\beta}) = PROF\ NPLIK = \prod_{i=1}^{n} \frac{e^{\beta^T x_i}}{\sum_{j:t_j \ge t_i} e^{\beta^T x_j}}. \tag{4.4}$$

This is exactly the same as the rank, the partial, and the marginal likelihood.

5. Profile NP Likelihood for the PEHR Model

Consider model (2.6) with $\theta > 0$. Set $\tau(\theta) = \theta[1 - e^{-\theta}]^{-1}$, then

$$L_{NP}(\boldsymbol{\theta}, F_0) = [\prod_{i=1}^{n} \tau(\theta_i)] \prod_{i=1}^{n} F_0\{t_i\} e^{-\theta_i F_0[t_i]}, \tag{5.1}$$

where $F_0[t_i] = \sum_{j:t_j \le t_i} F_0\{t_j\}$. Now set $p_i \equiv F_0\{t_i\}$ and maximize with respect to $p_1, \cdots, p_n$ with θ fixed. The maximization problem looks very similar to Cox model maximization except for the constraint $\sum p_i = 1$. We handle this constraint by writing $F_0(t) = 1 - \exp[-\Lambda(t)]$ with $\Lambda(t)$ unconstrained except for $\Lambda(t) \ge 0$ and by using a new approach based on the MM algorithm.

5.1. The MM Algorithm

Lang, Hunter and Yang (2000) introduced a concept called the MM algorithm. Its idea is that instead of maximizing a complicated original objective function, use a simpler surrogate function so that each iteration is faster and guarantees that the original objective function increases. Given the original objective function $l(\boldsymbol{h})$ for a maximization problem, a surrogate function $g(\boldsymbol{h}|\boldsymbol{h}_{old})$ must satisfy two properties:

$$l(\boldsymbol{h}_{old}) = g(\boldsymbol{h}_{old}|\boldsymbol{h}_{old}), \tag{5.2}$$

$$l(\boldsymbol{h}) \geq g(\boldsymbol{h}|\boldsymbol{h}_{old}). \tag{5.3}$$

The EM algorithm is a special case of the MM algorithm. A practical implementation issue of the MM algorithm is that we have to find a nice surrogate function case by case.

Now we construct a surrogate function based on Tsodikov (2003). Suppose we can write $l(\cdot)$ in the form $l(\boldsymbol{h}) = B(\boldsymbol{h}) - A(\boldsymbol{h})$ for some parameter vector $\boldsymbol{h} > \boldsymbol{0}$, where A and B are differentiable concave functions. Then by the concavity property,

$$g(\boldsymbol{h}|\boldsymbol{h}_{old}) = B(\boldsymbol{h}) - A(\boldsymbol{h}_{old}) - \nabla^T A(\boldsymbol{h}_{old})(\boldsymbol{h} - \boldsymbol{h}_{old}), \tag{5.4}$$

where $\nabla^T A(\boldsymbol{h}) = \partial A/\partial \boldsymbol{h}$ is the gradient of A, satisfies (5.2) and (5.3), and $g(\boldsymbol{h}|\boldsymbol{h}_{old})$ is a surrogate function for $l(\boldsymbol{h})$. Differentiating (5.4) gives

$$\nabla^T B(\boldsymbol{h}_{new}) = \nabla^T A(\boldsymbol{h}_{old}). \tag{5.5}$$

Solve (5.5) for $\boldsymbol{h}_{new}$. Iterate the procedure until there is a minimal change in $\boldsymbol{h}$.

5.2. The MM Algorithm for the PEHR Model with $\boldsymbol{\theta} \geq \boldsymbol{0}$ (SINAMI with $\boldsymbol{\theta} \geq \boldsymbol{0}$)

Let $\theta_i = g(\boldsymbol{\beta}^T \boldsymbol{x}_i)$, for some known function $g(\cdot) \geq 0$. When $\theta_i = 0$, the distribution function of T_i is $F_0(t)$. Let $F_0(t) = 1 - \exp[-\Lambda(t)]$, $h_k = \Lambda\{t_k\}$, and $\Lambda[t_i] = \sum_{k=1}^{i} h_k$ with $\Lambda(t) \geq 0$ and $h_k \geq 0$. Then for a temporarily fixed numerical vector $\boldsymbol{\beta}$,

$$\begin{aligned} l(\boldsymbol{h}) = log[L_{NP}(\boldsymbol{\beta}, \boldsymbol{h})] = &\sum_{i=1}^{n} \log \tau(\theta_i) + \sum_{i=1}^{n} \log h_i \\ &- \sum_{i=1}^{n} [\sum_{k=1}^{i} h_k + \theta_i (1 - \exp(-\sum_{k=1}^{i} h_k))]. \end{aligned} \tag{5.6}$$

Now we can write $l(\boldsymbol{h}) = B(\boldsymbol{h}) - A(\boldsymbol{h})$ with

$$B(\boldsymbol{h}) = \sum_{i=1}^{n} \log h_i, \tag{5.7}$$

$$A(\boldsymbol{h}) = \sum_{i=1}^{n} [\sum_{k=1}^{i} h_k + \theta_i (1 - \exp(-\sum_{k=1}^{i} h_k))]. \tag{5.8}$$

Here we may ignore $\Sigma \log \tau(\theta_i)$ because we maximize (5.6) w.r.t. $\boldsymbol{h}$.

$B(\boldsymbol{h})$ and $A(\boldsymbol{h})$ are concave, because for $0 \leq t \leq 1$, $B(t\boldsymbol{h}_a + (1-t)\boldsymbol{h}_b) \geq tB(\boldsymbol{h}_a)+(1-t)B(\boldsymbol{h}_b)$ and by mathematical induction, $A(t\boldsymbol{h}_a+(1-t)\boldsymbol{h}_b) \geq tA(\boldsymbol{h}_a)+(1-t)A(\boldsymbol{h}_b)$ hold. Note that

$$\partial B/\partial h_j = 1/h_j, \ j = 1, ..., n, \tag{5.9}$$

$$\partial A/\partial h_j = \sum_{i=1}^{n}(1+\theta_i(1-\exp(-\sum_{k=1}^{i} h_k)))1(j \leq i). \tag{5.10}$$

Using (5.5), (5.9), and (5.10), update $h_j, \ j = 1, \cdots, n$, at the same time,

$$h_{j,new} = [\sum_{i=1}^{n}(1+\theta_i(1-\exp(-\sum_{k=1}^{i} h_{k,old})))1(j \leq i)]^{-1}. \tag{5.11}$$

Iterate (5.11) until there is a minimal change in $l(\hat{\boldsymbol{h}}_{new})$; call the result $\hat{\boldsymbol{h}}_A$ ($\boldsymbol{A}$pproximated profile NPMLE). Note that we call $\hat{\boldsymbol{h}}_A$ approximated profile NPMLE because $\hat{\boldsymbol{h}}_A$ is obtained by fixing $\boldsymbol{\beta}$. This approximation is necessary because there is no closed form $\hat{\boldsymbol{h}}$ w.r.t. $\boldsymbol{\beta}$. Next set $l(\boldsymbol{\beta}) = log[L_{NP}(\boldsymbol{\beta}, \hat{\boldsymbol{h}}_A)]$ and maximize w.r.t. $\boldsymbol{\beta}$.

5.3. The MM Algorithm for the SINAMI Model with $\theta \leq 0$

Consider model (2.7) with $\theta_i = g(\boldsymbol{\beta}^T\boldsymbol{x}_i)$, for some known function $g(\cdot) \leq 0$. In this case we can use the algorithm of Section 5.2. To see this, suppose T satisfies model (2.7) with parameter $\theta_2 < 0$. Set $V = 1 - F_0(T)$, then V satisfies model (2.6) with parameter $\theta_1 = -\theta_2$. Moreover, the rank of $1 - F_0(T_i)$ is $n + 1 - R_i$.

5.4. The MM Algorithm for the SINAMI Model with $\theta \in R$

Consider model (2.7) with $\theta_i = g(\boldsymbol{\beta}^T\boldsymbol{x}_i)$, for some known function $g(\cdot) \in R$. In this case we can not use the transformation in Section 5.3 because it changes the likelihood and the monotonicity of the likelihood as a function of θ does not necessarily hold. Instead, we modify the algorithm as follows: If the value $\hat{\theta}_j$ in the jth iteration is positive, use the MM algorithm in Section 5.2. to find $\hat{\boldsymbol{h}}_j$. If $\hat{\theta}_j < 0$, then (5.6) implies that finding the maximizer $\boldsymbol{h}$ is a convex optimization problem which produces $\hat{\boldsymbol{h}}_j$.

5.5. Profile NPMLE Implementation

Successful convergence of the MM algorithm depends on a good starting point $(\hat{\theta}, \hat{\boldsymbol{h}})$. We consider the two sample problem:

$$T_{0,i} \sim F_0(t), \ i = 1, ..., n_0, \tag{5.12}$$

$$T_{1,i} \sim F(t) = \begin{cases} \frac{1-e^{-\theta F_0(t)}}{1-e^{-\theta}}, & (\theta > 0), \ i = 1, ..., n_1, \\ F_0(t), & (\theta = 0), \ i = 1, ..., n_1, \end{cases} \tag{5.13}$$

where $F_0(\cdot)$ is an unknown distribution with density f_0. Note that the density of $F(t)$ is

$$f(t;\theta) = \begin{cases} \tau(\theta) f_0(t) e^{-\theta F_0(t)}, & (\theta > 0), \\ f_0(t), & (\theta = 0). \end{cases} \tag{5.14}$$

We use an algorithm to find $(\hat{\theta}, \hat{\boldsymbol{h}})$ where $\theta = \beta$ in this case. For fixed F_0, (5.14) gives an MOM estimating equation for θ. We plug in an estimate $\hat{F}_0$ for F_0, and use the following algorithm:

$$step\ (1) : \hat{F}_0 \rightarrow\ \ step\ (2) : \hat{\theta} \rightarrow\ \ step\ (3) : (\hat{\theta}_A \leftrightarrow \hat{\boldsymbol{h}}_A)_{until\ \hat{\theta}_A\ converges.}$$

Here $\hat{\theta}_A$ and $\hat{h}_A$ are approximated profile NPMLE's from Section 5.2. The details are as follows:

***Step* (1) :** Compute the empirical distribution $\hat{F}_0(t)$ based on $T_{0,i}$ only:

$$\hat{F}_{0,[0]}(t) \equiv \hat{F}_0(t) = \frac{1}{1+n_0} \sum_{i=1}^{n_0} 1(T_{0,i} \le t). \tag{5.15}$$

Here $\hat{F}_0(t) \rightarrow_{a.s.} F_0(t)$ uniformly in t as $n_0 \rightarrow \infty$. The subscript [0] indicate iteration zero (starting point) for *step* (3). Note that the one-to-one relation between $\hat{F}_0$ and $\hat{\Lambda}_0$ is used to obtain $\hat{\boldsymbol{h}}$ by solving for $\boldsymbol{h}$ in the equations:

$$\begin{aligned} \hat{F}_0(t_i) \equiv &\quad \hat{F}_{0,i} = \exp(-\hat{\Lambda}_i),\ where \\ \hat{\Lambda}[t_i] \equiv &\quad \hat{\Lambda}_i = \sum_{k=1}^{i} \hat{h}_k,\ i = 1, ..., n_0. \end{aligned} \tag{5.16}$$

***Step* (2) :** Solve for $\hat{\theta}$ based on $\hat{F}_{0,[0]}(t)$:

$$\begin{aligned} \bar{T}_1 = &\quad \tau(\theta) \int y e^{-\theta \hat{F}_{0,[0]}(y)} d\hat{F}_{0,[0]}(y) \\ = &\quad \tau(\theta) \sum_{i=1}^{n_0} T_{0,(i)} \exp(-\theta \frac{i}{n_0}) \frac{1}{n_0}, \end{aligned} \tag{5.17}$$

where $T_{0,(i)}$ is an order statistics of $T_{0,1}, \cdots, T_{0,n_0}$ and $\bar{T}_1 = \sum_{i=1}^{n_1} T_{1,i}/n_1$. The solution is uniquely determined because the distribution function (5.13) is monotone increasing in θ and hence its mean is monotone decreasing in θ. If a model is $\theta \ge 0$ and $\hat{\theta} < 0$, set $\hat{\theta} = 0$. If a model is $\theta < 0$ and $\hat{\theta} > 0$, set $\hat{\theta} = 0$.

***Step* (3) :** Compute $\hat{\theta}_A$ and $\hat{\boldsymbol{h}}_A$ as follows:

$$(\hat{\theta}_A \leftrightarrow \hat{\boldsymbol{h}}_A)_{until\ \hat{\theta}_A\ converges.} \tag{5.18}$$

The first iteration $\hat{\theta}_{A,[1]}$, is obtained by maximizing (5.6) with $\hat{\theta}$ as a starting point, i.e., $\hat{\theta}_{A,[0]} = \hat{\theta}$ and with fixed $\hat{\boldsymbol{h}}_{[0]}$ obtained from (5.16) and $\hat{F}_{0,[0]}$, i.e.,

$$\hat{\theta}_{A,[1]} = \arg\max_{\theta} \{l(\theta, \hat{\boldsymbol{h}}_{[0]}) : \theta \ge 0\}. \tag{5.19}$$

Then by the MM algorithm in Section 5.2 with $\boldsymbol{\beta}_0 = \theta_0 = \hat{\theta}_{A,[1]}$ (see (5.11)), obtain $\hat{\boldsymbol{h}}_{[1]}$ using the starting point $\hat{\boldsymbol{h}}_{[0]}$. Next obtain

$$\hat{\theta}_{A,[2]} = \arg\max_{\theta} \{l(\theta, \hat{\boldsymbol{h}}_{A,[1]}) : \theta \geq 0\}, \tag{5.20}$$

with starting point $\hat{\theta}_{A,[1]}$. Then by the MM algorithm in Section 5.2 with $\boldsymbol{\beta}_0 = \theta_0 = \hat{\theta}_{A,[2]}$, obtain $\hat{\boldsymbol{h}}_{A,[2]}$ with starting point $\hat{\boldsymbol{h}}_{A,[1]}$. Repeat the procedure to get $\hat{\theta}_{A,[j]}$ and $\hat{\boldsymbol{h}}_{A,[j]}$ until convergence, i.e., $|\hat{\theta}_{A,[j]} - \hat{\theta}_{A,[j-1]}| < \epsilon$ for some small ϵ.

Numerical optimizations for $\hat{\theta}$ and $\hat{\theta}_A$ are carried out by the MATLAB fmincon() function.

***Remark* 5.1 :** For fixed $\hat{\boldsymbol{h}}$, $l(\theta, \hat{\boldsymbol{h}})$ is strictly concave and have a unique maximum.

***Remark* 5.2 :** The estimate of $\boldsymbol{\beta}$ in the Cox model that we have discussed is asymptotically optimal in the semiparametric sense (Begun, Hall, Huang, and Wellner (1983), Bickel et al. (1993, 1998), van der Vaart (1998)), Murphy and van der Vaart (2000)). These references and others give results that can be used to check the semiparametric asymptotic optimality of the profile NPMLE in the PEHR model.

***Remark* 5.3 :** *Transformation models.* We can show that the PEHR is a special case of transformation models as follows: Let F_λ be the exponential (λ) distribution function and define

$$G_0(y|\boldsymbol{x}) = \frac{1 - e^{-\theta F_\lambda(y)}}{1 - e^{-\theta}}, \quad y > 0, \tag{5.21}$$

where $\theta = g(\boldsymbol{x}, \boldsymbol{\beta})$. Let ψ be an increasing function from $[0, \infty)$ to $[0, \infty)$ and define the transformation model

$$G(y|\boldsymbol{x}) = G_0(\psi(y)|\boldsymbol{x}). \tag{5.22}$$

This model is of the form (2.6) with $F_0 = F_\lambda \psi(t)$. Klaassen (2007) gives results for general transformation models that can be used to check semiparametric asymptotic efficiency of estimates of $\boldsymbol{\beta}$ in the model (2.6).

5.6. Estimation of the Variance of the Profile NPMLE

Hypothesis tests and confidence intervals require standard errors (estimates of the standard deviation) of $\hat{\theta}_A$. An algorithm developed by Tsodikov and Garibotti (2007) combined with the preceding algorithm allows us to compute the profile information matrix which is the observed information matrix derived from the profile likelihood. This provides standard errors $SE(\hat{\theta}_A)$ of $\hat{\theta}_A$.

6. Simulation Results

6.1. PEHR Model Estimates

Monte Carlo (MC) simulation for model (5.13) with F_0 equal to the exponential distribution EXP(1) is based on 1000 Monte Carlo samples with n=100, 200, and

θ	2			3		
n	100	200	300	100	200	300
$E[\hat{\theta}_{MOM}]$	1.97	2.00	2.02	3.05	3.00	3.01
$E[\hat{\theta}_A]$	1.99	2.00	2.01	3.07	3.00	3.01
$E[\hat{\theta}_{PAR}]$	2.02	2.00	2.01	3.03	3.01	3.00
$SD(\hat{\theta}_{MOM})$	0.819	0.551	0.479	0.972	0.635	0.506
$SD(\hat{\theta}_A)$	0.788	0.541	0.466	0.960	0.618	0.497
$SD(\hat{\theta}_{PAR})$	0.541	0.387	0.317	0.605	0.431	0.335
$E[SE(\hat{\theta}_A)]$	0.752	0.527	0.430	0.837	0.578	0.471
$MSE[\hat{\theta}_{MOM}]$	0.671	0.303	0.230	0.949	0.403	0.256
$MSE[\hat{\theta}_A]$	0.621	0.292	0.217	0.926	0.383	0.248
$MSE[\hat{\theta}_{PAR}]$	0.289	0.150	0.101	0.367	0.186	0.112

TABLE 1

PEHR model simulation estimates (MC=1000, $\theta = 2, 3$)

300. $T_{0,i} \sim EXP(1)$, $i = 1, \cdots, n_0$, $T_{1,i} \sim PEHR(\theta = 2,\ or\ 3)$, $i = 1, \cdots, n_1$, $n_0 = n_1 = n/2$, iid. We compute Monte Carlo estimates of the expected values, standard deviations (SD's), and MSE's of $\hat{\theta}_{MOM}$, $\hat{\theta}_A$, and $\hat{\theta}_{PAR}$ where $\hat{\theta}_{PAR}$ is the parametric model MLE, obtained by assuming that F_0 is known and equal to the EXP(1) distribution.

We also compute the Monte Carlo estimates of the expected values $E[SE\ (\hat{\theta}_A)]$ of the standard errors computed as described in Section 5.5. Table 1 summarizes the result.

Overall $\hat{\theta}_{MOM}$, $\hat{\theta}_A$, and $\hat{\theta}_{PAR}$ have almost no bias in the estimation of θ =2 or 3. As expected, the parametric model estimate $\hat{\theta}_{PAR}$ has the smallest MSE. The approximated profile NPMLE $\hat{\theta}_A$ has a smaller MSE than $\hat{\theta}_{MOM}$, but the difference is small. The approximation to SD($\hat{\theta}_A$) is very good and improves as the sample size increases.

6.2. Model Fit for Misspecified Model

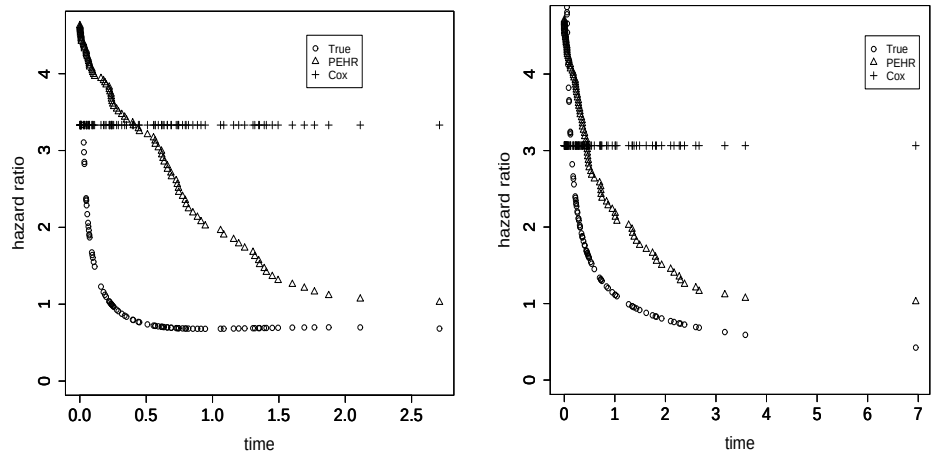
Next consider a model that is neither a Cox PH model nor a PEHR model: i.e., $T_{0,i} \sim EXP(1)$ and the true model for $T_{1,i}$ is : Case 1, Gamma(shape=0.5, scale=0.5), and the : Case 2, Weibull(shape=0.5, scale=0.2). Here the target values θ and $\boldsymbol{h}$ are those that minimize the Kullback-Leibler divergence between the true distribution and the model class of distributions (Doksum, Ozeki, Kim and Neto (2007)) .

Fig. 2 shows that PEHR gives better fit than the Cox model.

7. Estimation in the Normal Copula Model

7.1. The One Covariate Case

Assume that the pair (X,Y) has a joint density f(x,y) with respect to Lebesgue measure on R^2 and a joint distribution function F(x,y). Let F_1 and F_2 be the marginal distribution functions of X and Y, respectively, and let Φ denote the standard normal distribution function. Consider the transformations $X \to Z = \Phi^{-1}(F_1(X))$, $Y \to W = \Phi^{-1}(F_2(Y))$. Then the marginal distributions of Z and W are standard normal. The bivariate normal copula model $\mathcal{F}$ is defined by the



(*a*) *True model* : Gamma (*shape* = 0.5, *scale* = 0.5).

(*b*) *True model* : Weibull (*shape* = 0.5, *scale* = 0.2).

FIG 2. *Cox and PEHR estimated hazard ratio when* $F_0 \sim EXP(1)$ *and* $F \sim$ *Gamma or Weibull.*

assumption that the joint distribution of (Z,W) is bivariate normal with zero mean, unit variance, and correlation coefficient ρ. That is,

$$\mathcal{F} = \{F \ : (\Phi^{-1}(F_1(X)), \Phi^{-1}(F_2(Y))) \sim N(0,0,1,1,\rho)\},$$

where F_1 and F_2 are the marginals of F. Let $\{(X_i, Y_i), i = 1, 2, \cdots, n\}$ be independent and identically distributed with distribution function $F \in \mathcal{F}$, and set $Z_i = \Phi^{-1}(F_1(X_i))$, $W_i = \Phi^{-1}\ (F_2(Y_i))$, $i = 1, 2, \cdots, n$. If we (temporarily) assume that F_1 and F_2 are known, then because E(ZW) = ρ, a method of moments "estimate" of ρ, is $r_{MOM} = n^{-1}\sum_{i=1}^{n} Z_i W_i$. The asymptotic distribution of $\sqrt{n}(r_{MOM} - \rho)$ is $N(0, 1 + \rho^2)$ when $F \in \mathcal{F}$. Assuming F_1 and F_2 known, the asymptotic distribution of $\sqrt{n}(r_{MLE} - \rho)$, where r_{MLE} is the maximum likelihood "estimate" of ρ is $N(0, (1-\rho^2)^2/(1+\rho^2))$. The asymptotic variance $(1-\rho^2)^2/(1+\rho^2)$ of r_{MLE} is smaller than the asymptotic variance $(1-\rho^2)^2$ of the usual Pearson correlation coefficient r_P and much smaller than the asymptotic variance $1 + \rho^2$ of r_{MOM}.

Note that $\mathcal{F}$ is invariant under coordinate-wise increasing transformations. That is, if $(X, Y) \sim F \in \mathcal{F}$ and $U = h_1(X), V = h_2(Y)$ with h_1 and h_2 increasing, then the distribution G of (U,V) is in $\mathcal{F}$. If we want methods that are invariant under such transformations, we must use statistics based on the ranks defined in Section 1.

Suppose next that F_1 and F_2 are unknown. It may then make sense to replace the ordered Z's and W's by their expected values. This leads to the Fisher and Yates (1938) or normal scores $E(Z_{(i)})$, $i = 1, \cdots, n$ where $Z_{(1)}, \cdots, Z_{(n)}$ are N(0,1) order statistics. We write $a(i) = E(Z_{(i)})$. An accurate approximation to $E(Z_{(i)})$ is $\Phi^{-1}[(i - 3/8)/(n + 1/4)]$, e.g. Cox (2006).

Let $Z_i^{'} = a(R_i)$, $W_i^{'} = a(S_i)$ where R_i and S_i are the ranks of X_i and Y_i when the X's and Y's are ranked separately. Then we obtain estimates $\hat{\rho}_{MOM}$, $\hat{\rho}_{MLE}$, and $\hat{\rho}_P$ of ρ when F_1 and F_2 are unknown by replacing Z_i and W_i by $Z_i^{'}$ and $W_i^{'}$ in r_{MOM}, r_{MLE} and r_P. In this case $\hat{\rho}_{MOM}, \hat{\rho}_P$ are nearly identical and asymptotically equivalent, but they are different from $\hat{\rho}_{MLE}$. We will use $\hat{\rho}_P$ because it is slightly

less biased, and denote it by $\hat{\rho}_{NS}$ where NS signifies normal scores. Thus

$$\hat{\rho}_{NS} = \sum Z_i' W_i' / \sum a^2(i). \tag{7.1}$$

It follows from Bhuchongkul (1964) that based on the rank likelihood, $\hat{\rho}_{NS}$ is, uniformly in F_1 and F_2, a locally most powerful test statistics in the bivariate normal copula model. Zou and Hall (2002) gave an asymptotic extension of this result. They also computed the rank likelihood estimate of ρ in the bivariate normal copula model using an improved version of the likelihood sampler in Doksum(1987).

Klaassen and Wellner (1997) found $\sqrt{n}(\hat{\rho}_{NS}-\rho) \to_d N(0,(1-\rho^2)^2)$ in the copula model $\mathcal{F}$ with F_1 and F_2 unknown; the same as for the Pearson correlation in the bivariate normal model. In fact, in a bivariate normal$(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$ model, r_P is the MLE, and r_P and $\hat{\rho}_{NS}$ are asymptotically optimal in the parametric sense.

A fourth possible estimate is the profile NP estimate obtained by fixing ρ and replacing F_1 and F_2 by step functions with jumps $\{p_i\}$ and $\{q_i\}$ at (X_i,Y_i) in the log likelihood for the normal copula model. That is, ignoring constants (ρ is fixed), we maximize

$$\begin{aligned} l(\boldsymbol{p},\boldsymbol{q}) = \sum_i \{ & \log p_i + \log q_i \\ & + \frac{1}{2}((\Phi^{-1}(\sum_{k:X_k\le X_i} p_k))^2 + \frac{1}{2}(\Phi^{-1}(\sum_{k:Y_k\le Y_i} q_k))^2 \\ & + \frac{1}{2}(1-\rho^2)^{-1}[(\Phi^{-1}(\sum_{k:X_k\le X_i} p_k))^2 \\ & - 2\rho\Phi^{-1}(\sum_{k:X_k\le X_i} p_k)\Phi^{-1}(\sum_{k:Y_k\le Y_i} q_k) + (\Phi^{-1}(\sum_{k:Y_k\le Y_i} q_k))^2]\}, \end{aligned} \tag{7.2}$$

w.r.t. $(\boldsymbol{p},\boldsymbol{q})$ where $\sum p_i = 1$ and $\sum q_i = 1$. Then given $(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}})$, maximize the log likelihood w.r.t. ρ, which gives $\hat{\rho}_{PROF}$, a profile NPMLE.

***Remark* 7.1 :** An estimate $\hat{\theta}$ of a parameter $\theta \in R$ in a semiparametric model is regular if $\sqrt{n}(\hat{\theta}-\theta) \to_d N(0, V_{\hat{\theta}}(\theta,\eta))$ for some asymptotic variance $V_{\hat{\theta}}(\theta,\eta)$ and if $\hat{\theta}$ satisfies additional regularity conditions given in Bickel et al. (1993, 1998). For $F \in \mathcal{F}$, $V_{r_P}(\rho,F)$ depends on F (e.g. Bickel and Doksum (2007), Example 5.3.6), while $V_{\hat{\rho}_{NS}}(\rho,F)$ does not, as shown by Klaassen and Wellner (1997). Klaassen and Wellner (1997) go on to argue that $(1-\rho^2)^2$ is a semiparametric asymptotic variance lower bound for the class $\mathcal{S}$ of all regular estimates of ρ. Thus $\hat{\rho}_{NS}$ is semiparametrically optimal in the minimax sense:

$$\sup\{V_{\hat{\rho}_{NS}}(\rho,F) : F \in \mathcal{F}\} = \inf_{\hat{\rho}\in\mathcal{S}} \sup\{V_{\hat{\rho}}(\rho,F) : F \in \mathcal{F}\} \tag{7.3}$$

***Remark* 7.2 :** Recall that $\hat{\rho}_{NS}$ was obtained by inserting normal scores Z_i' and W_i' in the MOM estimate for the model with F_1 and F_2 known, and that the MLE r_{MLE} for this model has variance $(1-\rho^2)^2/(1+\rho^2)$. Klaassen and Wellner (1997) have shown that the approximate MLE $\hat{\rho}_{MLE}$ obtained from r_{MLE} by replacing (Z_i,W_i) with (Z_i',W_i') is semiparametrically optimal in the same sense as $\hat{\rho}_{NS}$. Because the distribution of the ranks do not depend on F_1 and F_2, this implies that $\hat{\rho}_{NS}$ and $\hat{\rho}_{MLE}$ are asymptotically equivalent for every $F \in \mathcal{F}$. We conjecture that $\hat{\rho}_{PROF}$ is also asymptotically optimal and equivalent to $\hat{\rho}_{NS}$.

***Remark* 7.3 :** The asymptotic distribution of $\hat{\rho}_{NS}$ when the distribution of (X,Y) is not in $\mathcal{F}$ can be obtained from Ruymgaart, Shorack, and Van Zwet (1972) and Ruymgaart (1974) .

7.2. The Multivariate Covariate Case

The normal copula model in the multivariate case is defined as follows: Let $Y \sim G$, $X_j \sim F_j$, $\boldsymbol{h}(\boldsymbol{X}) = (h_1(X_1), \cdots, h_d(X_d))$, where h_j, $j = 0, \cdots, d$ are increasing functions defined by

$$h_0(Y) = \Phi^{-1}(G(Y)), \tag{7.4}$$

$$h_j(X_j) = \Phi^{-1}(F_j(X_j)). \tag{7.5}$$

The distribution of the untransformed variables $(\boldsymbol{X}, Y)$ is a *copula model* if we assume that $(\boldsymbol{h}(\boldsymbol{X}), h_0(Y))$ is multivariate normal with 0 means and unit variances.

8. Transformation and NP Models

Consider a regression experiment with response Y and a random covariate vector $\boldsymbol{X} = (X_1, \cdots, X_d)^T$. We will extend the normal scores estimate $\hat{\rho}_{NS}$ of Section 7 to the d dimensioned case and compare it with estimates appropriate for parametric and nonparametric models. In the *copula regression* model of Section 7.2, we can write

$$h_0(Y) = \beta^T \boldsymbol{h}(\boldsymbol{X}) + \epsilon,\ \epsilon \sim N(0, \sigma^2), \tag{8.1}$$

where $\boldsymbol{\beta}$ is the set of regression coefficients when regressing $h_0(Y)$ on $\boldsymbol{h}(\boldsymbol{X})$. The transform both sides Box-Cox model is based on (8.1) with $h_0(Y) = Y^{(\lambda_{d+1})}$, $h_j(X_j) = X_j^{(\lambda_j)}$, $j = 1, \cdots, d$, where $t^{(\lambda)} = (t^\lambda - 1)/\lambda$. Thus for this case, we can write

$$Y^{(\lambda_{d+1})} = \alpha + \boldsymbol{\beta}^T \boldsymbol{X}^{(\boldsymbol{\lambda})} + \epsilon,\ \epsilon \sim N(0, \sigma^2). \tag{8.2}$$

We first consider a procedure for estimating the parameters in model (8.2):

I Profile Likelihood for a multivariate model.

Hernandez and Johnson (1980) considered the one sample multivariate Box-Cox transformation model. This was adopted to regression by Doksum, Ozeki, Kim and Neto (2007). We regard $(Y^{(\lambda_{d+1})}, \boldsymbol{X}^{(\boldsymbol{\lambda})})$ as a d+1 multivariate normal $(\boldsymbol{\mu}, \Sigma)$ vector. Regressing $Y^{(\lambda_{d+1})}$ on $\boldsymbol{X}^{(\boldsymbol{\lambda})}$ leads to (8.2). We fix $\xi \equiv (\boldsymbol{\lambda}, \lambda_{d+1})$ and estimate the parameters in the normal model by maximizing the likelihood thereby obtaining the familiar normal theory estimates $(\hat{\boldsymbol{\mu}}(\xi), \hat{\Sigma}(\xi))$. We plug these into the likelihood and obtain the profile likelihood $l(\xi)$, which we maximize to get $\hat{\xi}$ and the final estimates $(\hat{\boldsymbol{\mu}}(\hat{\xi}), \hat{\Sigma}(\hat{\xi}))$. These are the usual linear model estimates with Y_i, X_{ij} replaced by $Y_i^{(\hat{\lambda}_{d+1})}$, $X_{ij}^{(\hat{\lambda}_j)}$. Similarly, the estimate of $\boldsymbol{\beta}$ in (8.2) is the usual linear model estimate with Y_i, X_{ij} replaced by $Y_i^{(\hat{\lambda}_{d+1})}$, $X_{ij}^{(\hat{\lambda}_j)}$.

***Remark* 8.1 :** We also considered the maximum likelihood estimates of the parameters $\boldsymbol{\beta}, \sigma^2$, and $(\boldsymbol{\lambda}, \lambda_{d+1})$ in model (8.2). This approach has the problem that if we want to test $H_0 : \beta_j = 0$, then λ_j is not identifiable under H_0. Approach I does not have this problem. This is one case where likelihood and profile likelihood are very different. The algorithm for this MLE often failed to converge. When it did converge, it produced results close to those of method I. We omit the details.

***Remark* 8.2 :** As pointed out by Zou and Hall (2002), when d=1, the MLE of ρ in the Box-Cox transformation model with unknown transformation parameters and standardized transformations have the same efficiency as the MLE for the model with known transformation parameters because this Box-Cox model is between the bivariate normal model and the bivariate normal copula model and the MLE's in these models have the same asymptotic variance $(1-\rho^2)^2$. The result that the efficiency is the same whether or not the λ's are known in this Box-Cox model was also obtained by Wong (1981). This result is very different from the results of Bickel and Doksum (1981) regarding the estimation of regression coefficients.

***Remark* 8.3 :** Consider the transformation model

$$h_0(Y) = \boldsymbol{\beta}^T \boldsymbol{X} + \epsilon, \ \epsilon \sim N(0, \sigma), \tag{8.3}$$

where $\boldsymbol{X}$ is a vector of random covariates and $h(\cdot)$ is increasing. In this case we can consider the rank estimate $\hat{\boldsymbol{\beta}}_R$ obtained by maximizing the rank likelihood $l_{\boldsymbol{r}}(\boldsymbol{\beta}) = P(\boldsymbol{R} = \boldsymbol{r})$ defined in Section 3. The results of Bickel and Ritov (1997) imply that in a certain sense $\hat{\boldsymbol{\beta}}_R$ is semiparametrically optimal for model (8.3). However the normal scores estimate of $\hat{\boldsymbol{\beta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{a}$, where $\boldsymbol{a} = (a(S_1), \cdots, a(S_n))^T$ and $\boldsymbol{x}$ is a vector of nonrandom covariates, is not asymptotically optimal unless $|\boldsymbol{\beta}|/\sigma$ tends to zero at a certain rate as $n \to \infty$ (Doksum (1987)). MC methods for $\hat{\boldsymbol{\beta}}_R$ is introduced in Bickel and Doksum (2009), Section 10.5.

We next introduce a semiparametric approach for the copula regression model and a nonparametric regression approach.

II Normal score substitution.

The model (8.1) with h_j, $j = 0, \cdots, d$, satisfying (7.4) and (7.5) is invariant under increasing transformations. As in the d=1 case, this leads to using the ranks $\{S_i\}$ of the Y's and the ranks $\{R_{ij} : i = 1, \cdots, n\}$ of X_{ij} among $\{X_{ij} : i = 1, \cdots, n, j = 1, \cdots, d\}$. Because the distribution of the ranks is invariant under increasing transformations, for rank methods, model (8.1) is equivalent to

$$Y' = \boldsymbol{\alpha}^T \boldsymbol{X}' + \epsilon', \tag{8.4}$$

where $X'_j \sim N(0,1)$ and ϵ' are independent, $Y' \sim N(0,1)$ and $\boldsymbol{\alpha}^T = \Sigma^{-1}\boldsymbol{\rho}$ with $\boldsymbol{\rho} = (Corr(X'_1, Y'), \cdots, Corr(X'_d, Y'))^T$ and Σ the correlation matrix of $\boldsymbol{X}' = (X'_j)_{d\times 1}$. Here Σ is assumed to be nonsingular. Based on the distribution of the ranks, $\alpha_1, \cdots, \alpha_d$ are identifiable parameters in model (8.4). These parameters represents the relative importance of the X_j's.

The normal scores $Z'_{ij} = a(R_{ij})$ and $W'_i = a(S_i)$ have approximately the same distribution as the unobservable X'_{ij} and Y'_i in model (8.4). Because $E(Y'|\boldsymbol{x}') = \boldsymbol{\alpha}^T\boldsymbol{x}'$, if we replace X'_{ij} and Y'_i with Z'_{ij} and W'_i, we find that an approximate

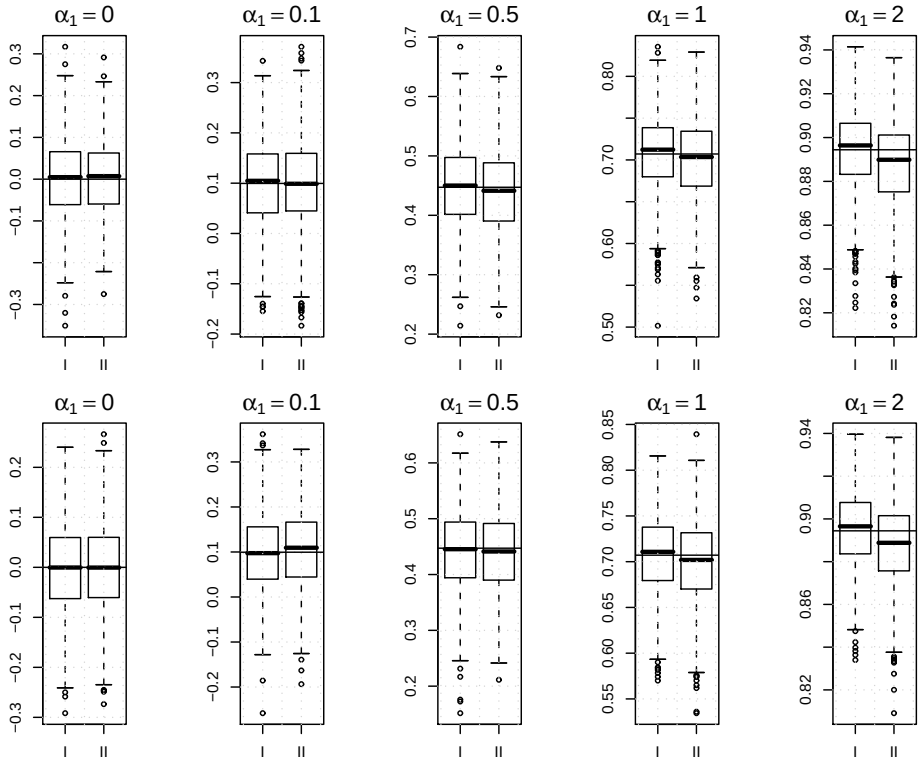


FIG 3. *$\hat{\rho}$ boxplots for correctly specified model (8.7) with n=128. Top 5: $\lambda_1 = \lambda_2 = 0.5$. Bottom 5: $\lambda_1 = \lambda_2 = 1$. The horizontal line gives the true value of ρ.*

method of moments estimate of α is

$$\hat{\boldsymbol{\alpha}} = (Z_D^T Z_D)^{-1} Z_D^T \boldsymbol{W'}, \tag{8.5}$$

where Z_D is the no intercept design matrix $(Z'_{ij})_{n\times d'}$, d' is the rank of the matrix (Z'_{ij}), and $\boldsymbol{W'} = (W'_1, \cdots, W'_n)^T$. Any subset of variables $X_j : j \in J$ with the same ranks, say $R_{1J}, \cdots, R_{nJ}$, is collapsed into one variable denoted as X_J with ranks $R_{1J}, \cdots, R_{nJ}$ to avoid singularity. Based on Klaassen and Wellner (1997), we conjecture that $\hat{\boldsymbol{\alpha}}$ is semiparametrically efficient for the multivariate normal copula model.

III Nonparametric estimation.

We next introduce a nonparametric approach. We consider the model

$$Y = m(\boldsymbol{X}) + \epsilon, \tag{8.6}$$

where $m()$ is unknown and ϵ has median zero. To estimate $m()$, we use a cubic B-spline and the R function smooth.spline(). The number of knots are automatically selected (less than the number of observations n.) The smoothing parameter is chosen by generalized cross validation (GCV.)

8.1. Simulation Results

We consider the d=1 case and consider the properties of estimates of $\rho = Corr(h_1(X), h_0(Y))$. In this case, the method II estimate is $\hat{\rho}_{NS}$.

8.1.1. Correctly Specified Model

The true model satisfies

$$Y^{(\lambda_2)} = \alpha_0 + \alpha_1 X^{(\lambda_1)} + \epsilon, \tag{8.7}$$

where $\epsilon \sim N(0, \sigma^2)$ and $X^{(\lambda_1)} \sim N(\mu_1, \sigma_0^2)$ are independent. This model is a subset of the normal copula model $\mathcal{F}$ with

$$F_1(x) = \Phi(\frac{x^{(\lambda_1)} - \mu_1}{\sigma_0}),\ F_2(y) = \Phi(\frac{y^{(\lambda_2)} - \mu_2}{\sigma_2}), \tag{8.8}$$

where $\mu_2 = EY^{(\lambda_2)}$, and $\sigma_2^2 = VarY^{(\lambda_2)}$. We use 1000 MC trials and take $\sigma^2 = 1$, $\sigma_0^2 = 1$, $(\lambda_1,\ \lambda_2) \in \{(0.5, 0.5), (1, 1)\}$, $\alpha_0 = 6$, $\alpha_1 = \in \{0, 0.1, 0.5, 1, 2\}$, and $\mu_1 = 5$.

Fig. 3 shows that methods I and II have similar properties for $\alpha_1 \leq 0.5$. For larger α_1, the normal scores estimate has a downward bias which is negligible for $n \geq 500$ (not shown here.) Method I converges all the time with the constraint $-\mathbf{4} \leq \boldsymbol{\lambda} \leq \mathbf{4}$. Method II does not involve any optimization and hence converges all the time.

8.1.2. Misspecified Model

We simulate the data from

$$Y^{(\lambda_2)} = (1 - \gamma)(\alpha_0 + \alpha_1 X^{(\lambda_1)}) + \gamma[L(X)] + \epsilon, \tag{8.9}$$

where $L()$ is a nonlinear function. Thus the model is a Box-Cox model when $\gamma = 0$, but when $\gamma > 0$, we are checking the performance of the methods when the model generating the methods are misspecified.

For comparisons of methods we need a parameter that makes sense for all three methods. One such parameter is

$$m(x) = Median(Y|X = x). \tag{8.10}$$

We consider the 25th, 50th, and 75th population quantiles of X, i.e., our parameters of interest are $m(x_{0.25})$, $m(x_{0.50})$, and $m(x_{0.75})$.

Methods I and II are based on models of the form

$$h_0(Y) = g(\boldsymbol{X}, \boldsymbol{\beta}) + \epsilon, \tag{8.11}$$

where $h_0(\cdot)$ is an increasing function. If X and ϵ are independent and $median(\epsilon) = 0$, then

$$m(x) = h_0^{-1}(g(\boldsymbol{X}, \boldsymbol{\beta})). \tag{8.12}$$

For method I, the MLE of m(x) in model (8.7) is,

$$\hat{m}(x) = (\hat{\lambda}_2[\hat{\beta}_0 + \hat{\beta}_1 \frac{x^{\hat{\lambda}_1} - 1}{\hat{\lambda}_1}] + 1)^{1/\hat{\lambda}_2}. \tag{8.13}$$

For method II, write model (8.1) as

$$h_0(Y) = \rho h_1(X) + \epsilon,\ \epsilon \sim N(0, \sigma^2), \tag{8.14}$$

where $\rho \equiv \rho(h_1(X), h_0(Y))$ is the correlation coefficient. Then by (8.12),

$$m(x) = h_0^{-1}(\rho h_1(x)), \tag{8.15}$$

where

$$h_0^{-1}(t) = F_2^{-1}(\Phi(t)). \tag{8.16}$$

It follows that

$$m(x) = {F_2}^{-1}(\Phi(\rho\Phi^{-1}(F_1(x))), \tag{8.17}$$

and by replacing F_1 and F_2 by their empiricals, a natural estimate of m(x) is

$$\hat{m}(x) = y_{([n\Phi(\hat{g}(x))])}, \tag{8.18}$$

where $\hat{g}(x) = \hat{\rho}_{NS}\Phi^{-1}(\hat{F}_1(x))$ and [ ] is the greatest integer function.

For method III, we use the smoothing spline estimate of $E(Y|X = x)$ described earlier. In our models with normal errors, $E(Y|X = x)$ coincide with the conditional median m(x).

In the simulation we use model (8.9) with $\epsilon \sim N(0, \sigma^2)$ and $X^{(\lambda_1)} \sim N(\mu_1, \sigma_0^2)$ independent,

$$L(t) = \alpha_0 + \alpha_1\mu_1 - 1.25 + 2.5[1 + \exp(-10(t - \mu_1))]^{-1}, \tag{8.19}$$

$\sigma^2 \in \{0.01, 0.1, 0.5, 1\}$, $\sigma_0^2 = 1$, $(\lambda_1,\ \lambda_2) \in \{(0.5, 0.5), (1, 1)\}$, $\alpha_0 = 6.25$, $\alpha_1 = \in \{0, 0.5, 1, 2\}$, $\mu_1 = 5$, and $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$. The sample size is n=512. There are 1000 MC trials.

Fig. 4, 5, 6, and 7 are boxplots of $\hat{m}(x_{0.25})$, $\hat{m}(x_{0.50})$, and $\hat{m}(x_{0.75})$ with the setting $(\lambda_1, \lambda_2, \alpha_1, \sigma^2) = (1, 1, 0.5, 0.5)$, $(\lambda_1, \lambda_2, \alpha_1, \sigma^2) = (1, 1, 0.5, 1)$, $(\lambda_1, \lambda_2, \alpha_1, \sigma^2) = (0.5, 0.5, 0.5, 0.5)$, and $(\lambda_1, \lambda_2, \alpha_1, \sigma^2) = (0.5, 0.5, 0.5, 1)$ respectively. Fig.8-11 give MSE's for the estimates $\hat{m}(x_{0.25})$, $\hat{m}(x_{0.50})$, and $\hat{m}(x_{0.75})$.

We see that method I is the best when model (8.2) is correct, that is, $\gamma = 0$. However when the model is increasingly misspecified, i.e., as γ increases, its absolute bias increases which leads to low MSE performance.

Method II is overall best in terms of MSE when $\lambda_1 = \lambda_2 = 0.5$ (Fig 6 and 7). When $\gamma = 0$, it is unbiased and its variance is between Method I and Method III (Fig 4,5,6, and 7).

Method III is overall best in terms of MSE when $\lambda_1 = \lambda_2 = 1$ and the model is badly misspecified. It's smaller bias makes up for its large variance in this case. But its MSE suffers at and near model (8.2) (Fig 6, 7, $\gamma = 0$).

In summary, the normal score procedure performs very well at and close to a copula model. For n large, this is to be expected from the results of Klaassen and Wellner (1997). The normal scores estimate is competitive with the Box-Cox estimate in the transform both sides Box-Cox model.

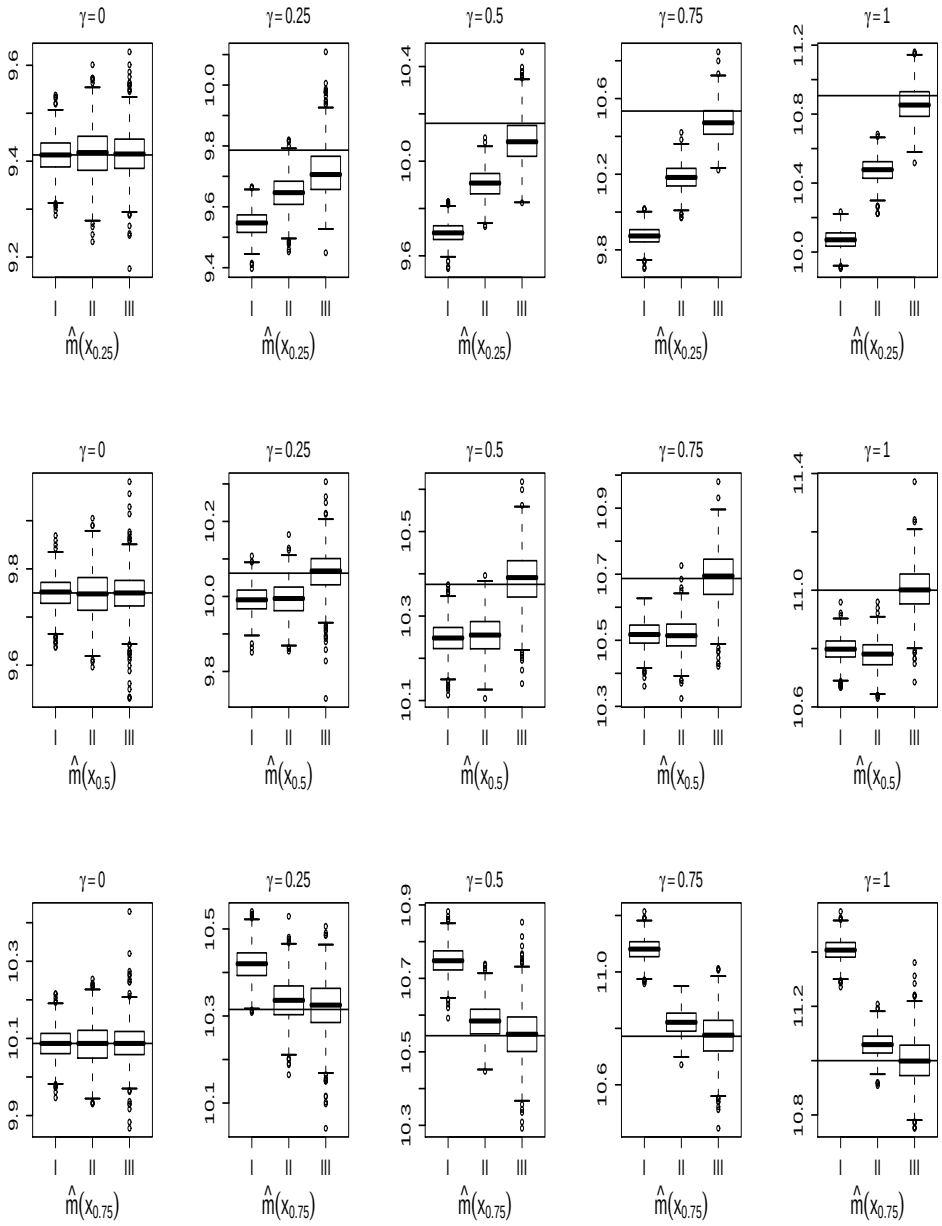


FIG 4. *Boxplots of the three estimates of median regression m(x) for model (8.9) with* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(1, 1, 0.5, 0.5). I: profile MLE, II: normal scores, and III: NP, spline. The true value of m(x) is the solid line.*

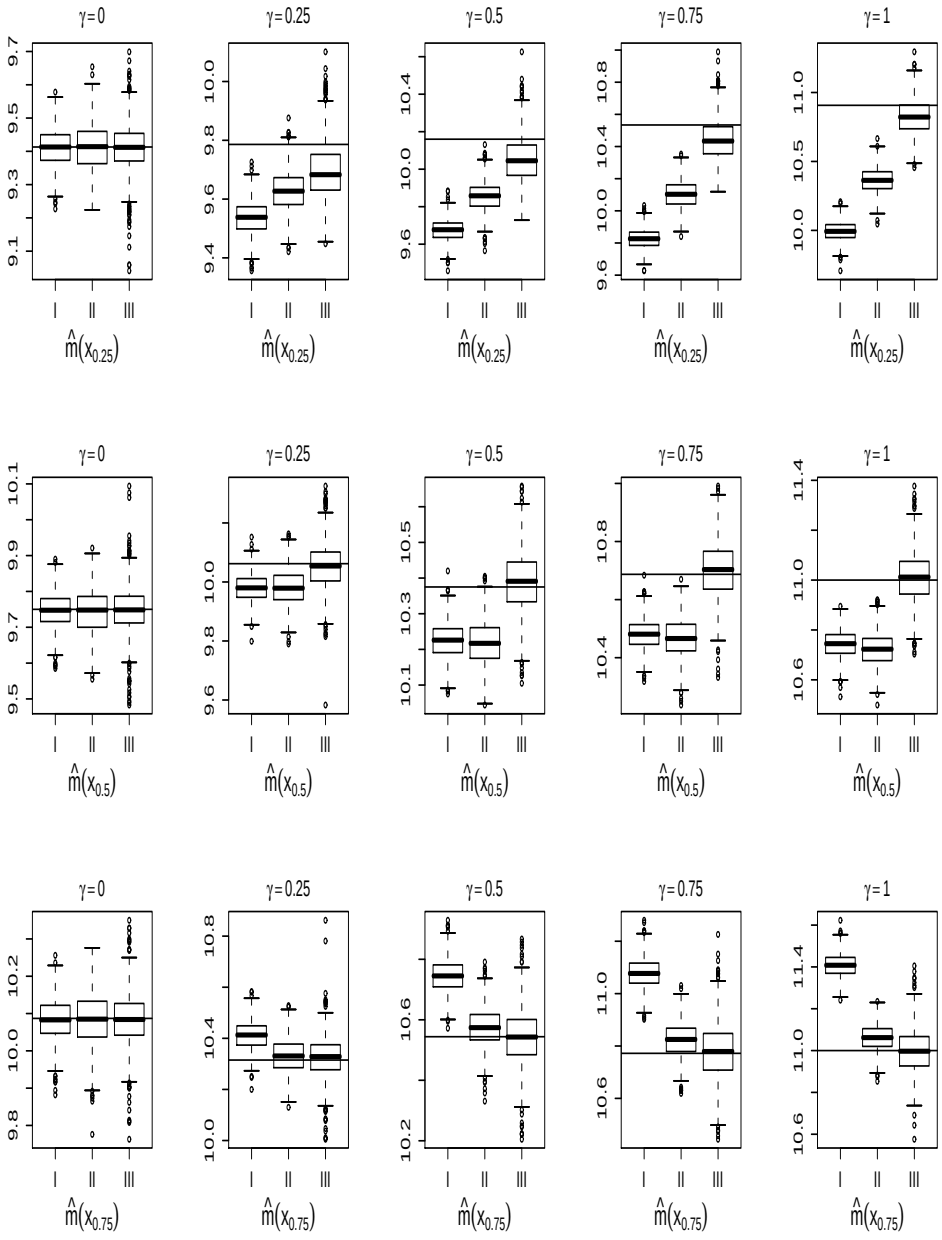


FIG 5. *Boxplots of the three estimates of median regression m(x) for model (8.9) with* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(1, 1, 0.5, 1). I: profile MLE, II: normal scores, and III: NP, spline. The true value of m(x) is the solid line.*

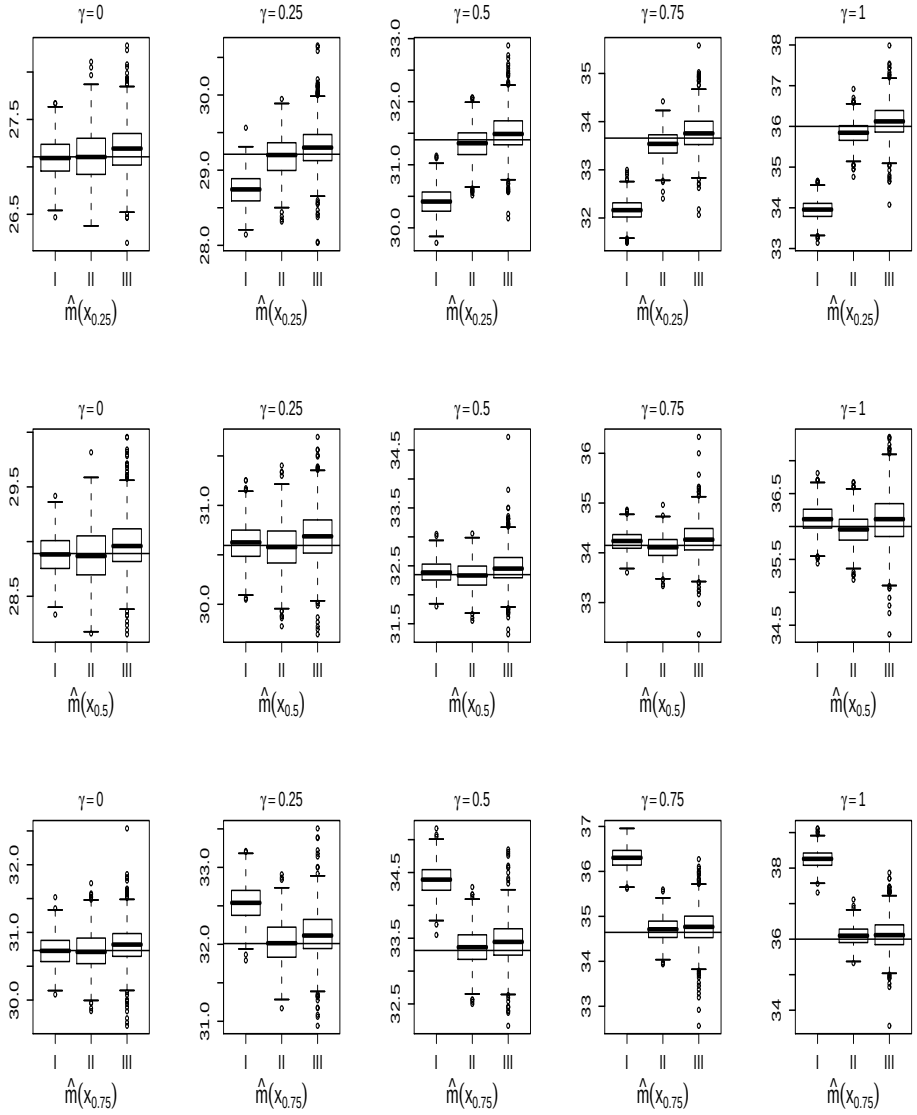


FIG 6. *Boxplots of the three estimates of median regression m(x) for model (8.9) with* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(0.5, 0.5, 0.5, 0.5). I: profile MLE, II: normal scores, and III: NP, spline. The true value of m(x) is the solid line.*

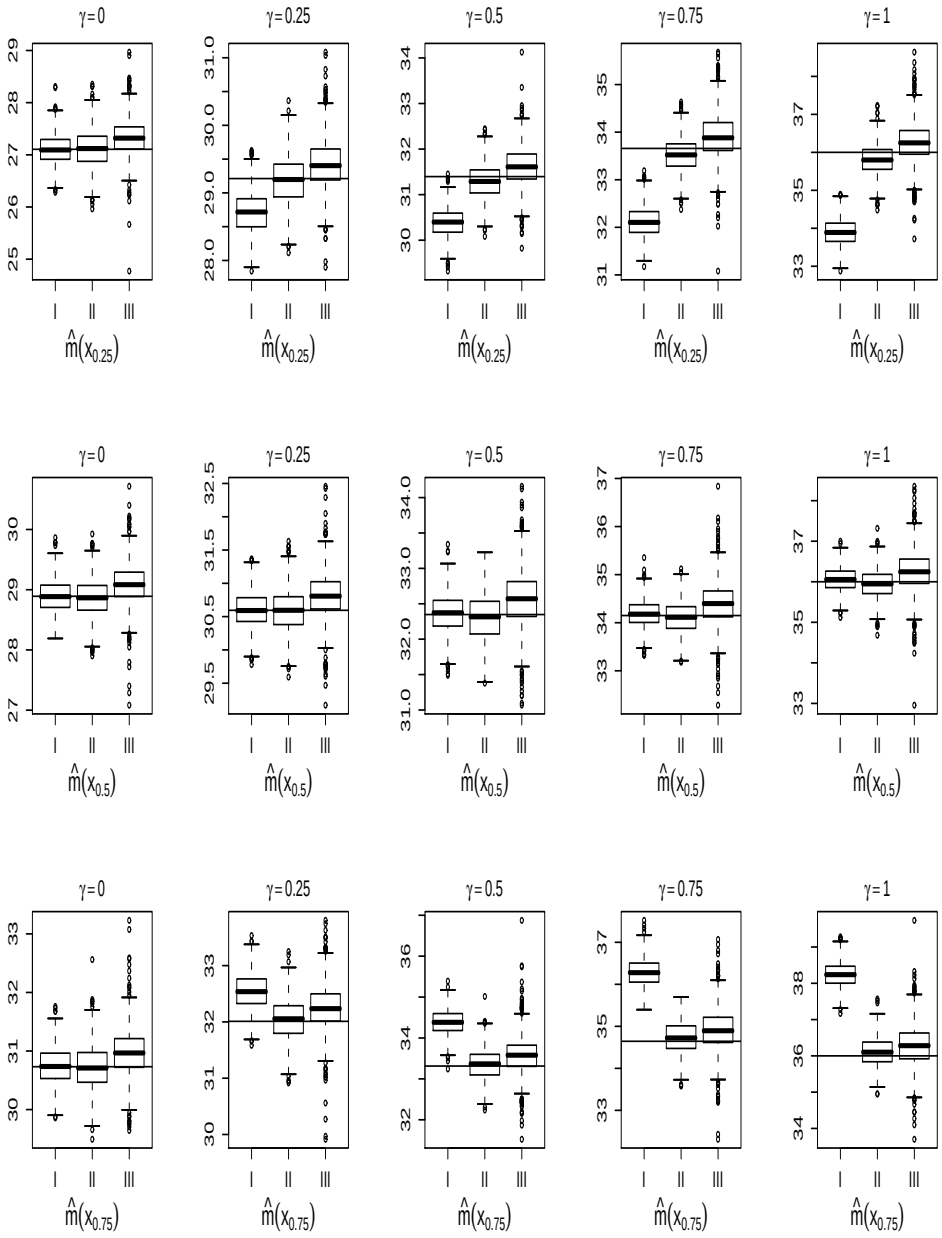


FIG 7. *Boxplots of the three estimates of median regression m(x) for model (8.9) with* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(0.5, 0.5, 0.5, 1). I: profile MLE, II: normal scores, and III: NP, spline. The true value of m(x) is the solid line.*

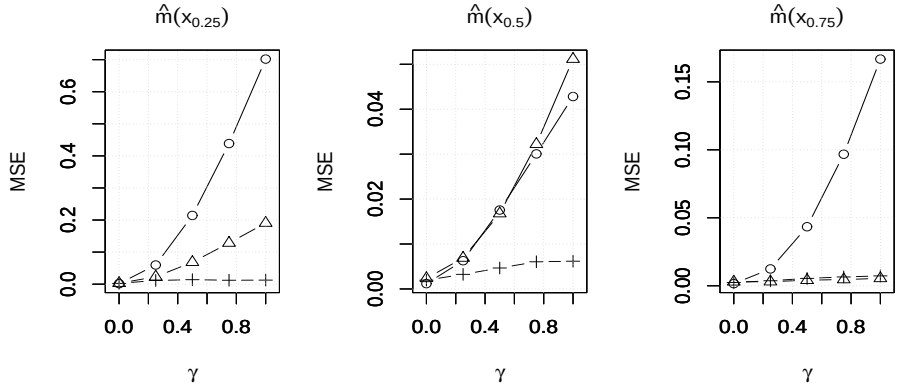


FIG 8. *MSE of the three estimates of m(x) as a function of the misspecification parameter* γ *for* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(1, 1, 0.5, 0.5).* $\bigcirc = I$, $\triangle = II$, $+ = III$.

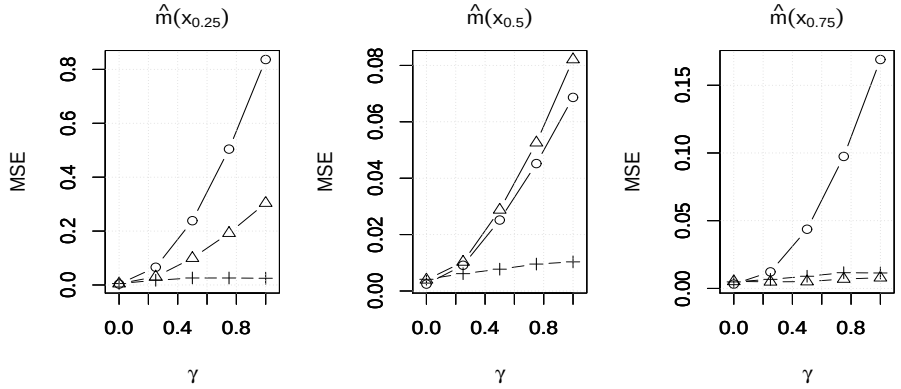


FIG 9. *MSE of the three estimates of m(x) as a function of the misspecification parameter* γ *for* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(1, 1, 0.5, 1).* $\bigcirc = I$, $\triangle = II$, $+ = III$.

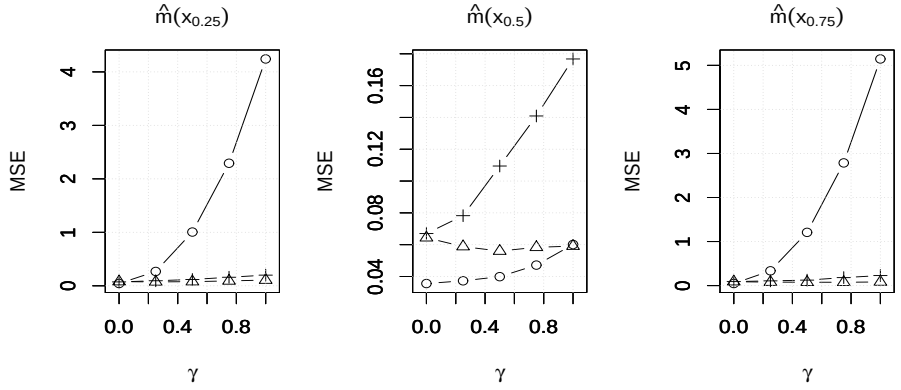


FIG 10. *MSE of the three estimates of m(x) as a function of the misspecification parameter* γ *for* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(0.5, 0.5, 0.5, 0.5).* $\bigcirc = I$, $\triangle = II$, $+ = III$.

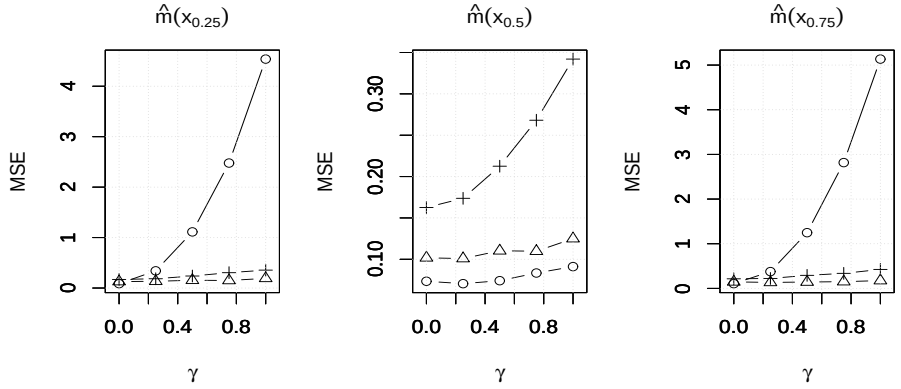


FIG 11. *MSE of the three estimates of m(x) as a function of the misspecification parameter* γ *for* $(\lambda_1, \lambda_2, \alpha_1, \sigma^2)$*=(0.5, 0.5, 0.5, 1).* $\bigcirc = I$, $\triangle = II$, $+ = III$.

Acknowledgements

We thank Peter Bickel, Aad van der Vaart and Jon Wellner for helpful comments.

References

[1] ANDERSEN, P. K., BORGAN, O., GILL, R. D. and KEIDING N. (1996). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.
[2] BEGUN, J. M., HALL, W. J., HUANG, WEI-MIN and WELLNER, J. A. (1983). Information and Asymptotic Efficiency in Parametric-Nonparametric Models. *The Annals of Statistics*, **11**, 2, 432–452.
[3] BELL, C. B. and DOKSUM, K. A. (1966). "Optimal" One-Sample Distribution-Free Tests and Their Two-Sample Extensions. *The Annals of Mathematical Statistics*, **36**, 1, 120–132.
[4] BHUCHONGKUL, S. (1964). A Class of Nonparametric Tests for Independence in Bivariate Populations. *Ann. Math. Statist.*, **35**,1, 138–149.
[5] BICKEL, P. J. and DOKSUM, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, **76**, 296–311.
[6] BICKEL, P. J. and DOKSUM, K. A. (2007). *Mathematical Statistics: Basic Ideas and Selected Topics, 2nd ed. Vol I*, Updated Printing. Pearson Prentice Hall, Upper Saddle River, NJ.
[7] BICKEL, P. J. and DOKSUM, K. A. To appear (2009). *Mathematical Statistics: Basic Ideas and Selected Topics Vol II*. Pearson Prentice Hall, Upper Saddle River, NJ.
[8] BICKEL, P. J., KLAASSEN, C. A., RITOV, Y. and WELLNER, J. A. (1993, 1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.
[9] BICKEL, P. J. and RITOV, Y. (1997). Local asymptotic normality of ranks and covariates in transformation models. In *Festschrift for Lucien Le Cam (D. Pollard and G. L. Yang, eds.)* Springer, New York.
[10] COX, D. R. (1964). Some applications of exponential ordered scores. *J. R. Stat. Soc.*, **B.26**, 103–110.
[11] COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc.*, **B.34**, 187–220.
[12] COX, D. R. (1975). Partial likelihood. *Biometrika.*, **62.2**, 269–276.
[13] COX, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge, 40.
[14] DOKSUM, K. A. (1987). An extension of partial likelihood methods from proportional hazard models to general transformation models. *Annals of Statistics*, **15**, 325–345.
[15] DOKSUM, K. A., OZEKI, A., KIM, J. and NETO, E. C. (2007). Thinking outside the box: Statistical inference based on Kullback-Leibler Empirical Projections. *Statistics and Probability Letters*. **77**, 1201–1213.
[16] FISHER, R. A. and YATES, F. (1938). *Statistical Tables for Biological, Agricultural and Medical Research*. (5th edition, 1957) Oliver and Boyd, Edinburgh.
[17] FERGUSON, T. S. (1967). *Mathematical Statistics. A Decision Theoretic Approach*. New York and London, Academic Press.
[18] HAJEK, J. and SIDAK, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

[19] Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Stat*, **34**, 598-611.
[20] Hoeffding, W. (1951). 'Optimum' nonparametric tests. *Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, Univ. Calif. Press, 83–92.
[21] Hernandez, F. and Johnson, R. A. (1980). The large-sample behavior of transformations to normality. *Journal of the American Statistical Association*, **75**, 855–861.
[22] Kalbfleich, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, **60**, 267–278.
[23] Kalbfleich, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of failure Time Data*, 2nd edition. John Wiley and Sons, Inc., Hoboken, New Jersey.
[24] Klaassen, C. A. J. (2007). A Sturm-Liouville Problem in Semiparametric Transformation Models. In *Advances In Statistical Modeling And Inference. Essays in Honor of Kjell A Doksum*. Editor, Vijay Nair. World Scientific Pub Co Inc., New Jersey.
[25] Kosorok, M. R., Lee B. L. and Fine J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *The Annals of Statistics*, **32**, No. 4, 1448–1491.
[26] Klaassen, C. A. J. and Wellner, J. A. (1997). Efficient Estimation in the Bivariate Normal Copula Model: Normal Margins Are Least Favourable. *Bernoulli*, **3**, No. 1, 55–77.
[27] Lange, K., Hunter, D. R. and Yang, I. (2000). Optimization Transfer Using Surrogate Objective Functions. *Journal of Computational and Graphical Statistics*, **9**, No. 1, 1–20.
[28] Lehmann, E. L. (1953). The power of rank tests. *Ann. Math. Statist.*, 24, 23–43.
[29] Murphy, S. A. and van der Vaart, A. W. (2000). On Profile Likelihood. *JASA*, **9**, No. 450, 449–465.
[30] Nabeya, S. and Miura, R. (1972). Locally Most Powerful Rank Tests for Lehmann mixture alternatives. *Technical Report*. University of California, Berkeley.
[31] Oakes, D. (1992). Bivariate survival models induced by frailties. *JASA*, **84**, 487–493.
[32] Oakes, D. and Jong-Hyeon Jeong (1998). Frailty Models and Rank Tests. *Lifetime Data Analysis*, **4**, 3, 209–228.
[33] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, No.2, 237–249.
[34] Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall.
[35] Ruymgaart, F. H. (1974). Asymptotic Normality of Nonparametric Tests for Independence. *Ann. Statist.*, **2**, 892–910.
[36] Ruymgaart, F. H., Shorack, G. R. and Van Zwet, W. R. (1972). Asymptotic Normality of Nonparametric Tests for Independence. *Ann. Math. Statist.*, **43**, 1122–1135.
[37] Savage, I. R. (1956). Contributions to the theory of rank orders statistics: the two-sample case. *Ann. Math. Stat.*, **27**, 590–615.
[38] Savage, I. R. (1957). Contributions to the Theory of Rank Order Statistics–The "Trend" Case. *Ann. Math. Stat.*, **28**, 4, 968–977.
[39] Sibuya, M. (1968). Generating Doubly Exponential Random Numbers. *Annals of the Institute of Statistical Mathematics, Tokyo, Supplement V*, 1–7.
[40] Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges. *L'Institut de Statistique de L'Universite de Paris*. **8**, 229–231.
[41] Tsodikov, A. (2003). Semiparametric models: a generalized self-consistency approach. *J. R. Statist. Soc.*, **B.65**, 3, 759–774.
[42] Tsodikov, A. and Garibotti, G. (2007). Profile information matrix for nonlinear transformation models. *Lifetime Data Analysis*, **13**, 139–159.
[43] van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press, Cambridge, UK.
[44] Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Statist.*, **69**, 4, 1–30.
[45] Zou, K. H. and Hall, W. J. (2002). On estimating a transformation correlation coefficient. *Journal of Applied Statistics*, **29**, 745–760.
[46] Wong, C. W. (1981). Transformation of Independent Variables in Regression Models. *Ph.D. Thesis*. Department of Statistics, University of Berkeley, CA.

IMS Collections
Vol. 0 (2009) 93–116
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000010

On Bootstrap Tests of Hypotheses

Wei-Yin Loh[1,*] **and Wei Zheng**[2,*]

University of Wisconsin–Madison

Abstract: The size of the bootstrap test of hypotheses is studied for the normal and exponential one and two-sample problems. It is found that the size depends not only on the problem, but on the choice of test statistic and the nominal level. In some special cases, the bootstrap test is UMP, but in other cases, it can be totally useless, such as being completely randomized or rejecting the null hypothesis with probability one. More importantly, the size is usually greater than the nominal level, even in the limit as the sample size goes to infinity.

Contents

1 Introduction 94
2 Testing a Normal Mean 94
2.1 Known Variance 95
2.1.1 Sample Mean Statistic 95
2.1.2 Standard Likelihood Ratio Statistic 95
2.1.3 Cox Likelihood Ratio Statistic 96
2.2 Unknown Variance 97
2.2.1 Standard Likelihood Ratio Statistic 98
2.2.2 Cox Likelihood Ratio Statistic 100
3 Testing a Normal Variance, Mean Unknown 101
3.1 $H_0 : \sigma^2 \leq 1$ vs. $H_1 : \sigma^2 > 1$ 101
3.1.1 Standard Likelihood Ratio Statistic 101
3.1.2 Cox Likelihood Ratio Statistic 103
3.2 $H_0 : \sigma^2 \geq 1$ vs. $H_1 : \sigma^2 < 1$ 104
3.2.1 Standard Likelihood Ratio Statistic 104
3.2.2 Cox Likelihood Ratio Statistic 105
4 Testing Difference of Two Normal Means 106
4.1 Known Variances 106
4.1.1 Difference of Means Statistic 106
4.1.2 Standard Likelihood Ratio Statistic 107
4.1.3 Cox Likelihood Ratio Statistic 107
4.2 Unknown but Equal Variances 108
4.2.1 Difference of Means Statistic 108
4.2.2 Standard Likelihood Ratio Statistic 110

[1]Department of Statistics, University of Wisconsin–Madison, 1300 University Avenue, WI 53706, email: loh@stat.wisc.edu

[2]Department of Statistics, University of Wisconsin–Madison, 1300 University Avenue, WI 53706, email: zheng@stat.wisc.edu

*This material is based upon work partially supported by the National Science Foundation under grant DMS-0402470 and the U.S. Army Research Laboratory and the U.S. Army Research Office under grant W911NF-05-1-0047.

AMS 2000 subject classifications: Primary 62F03, 62F40; secondary 62F05

Keywords and phrases: likelihood ratio, maximum likelihood, normal distribution, exponential distribution, resampling, size, significance level, uniformly most powerful

4.2.3 Cox Likelihood Ratio Statistic 111
5 Testing an Exponential Location Parameter 111
5.1 $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$ 111
5.1.1 Standard Likelihood Ratio Statistic 112
5.1.2 Cox Likelihood Ratio Statistic 113
5.2 $H_0 : \theta \geq 0$ vs. $H_1 : \theta < 0$ 113
5.2.1 Standard Likelihood Ratio Statistic 113
5.2.2 Cox Likelihood Ratio Statistic 113
6 Conclusion 114
Appendix 114
References 116

1. Introduction

Owing to its practical convenience and wide applicability, the bootstrap method [7] is used to test statistical hypotheses in many research studies. A sample of recent applications includes evolutionary molecular biology [1], genetic structure [2], gene frequency [11], cancer epidemiology [8], microscopy [3], quality of life [12], economic cycles [5], livestock management [9], and meat demand [6]. Despite its popularity, however, there have been few detailed studies of the theoretical validity of the bootstrap for hypothesis testing. This article addresses this issue for some simple parametric problems where the bootstrap null distributions can be studied analytically. Specifically, we consider one and two-sample problems involving normally and exponentially distributed observations. Our goal is to determine the finite-sample or limiting sizes of the bootstrap tests and compare them with those of the traditional tests.

First, we recall some definitions. Let $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ be a vector of n independent observations from F_μ. In the bootstrap method, we first find an estimate $\hat{\mu}_0$ of μ under H_0 and estimate F_μ with $\hat{F} = F_{\hat{\mu}_0}$. Given a test statistic $S = S(\mathbf{X}_n)$ for which large values lead to rejection of H_0, let G_μ denote the distribution function of S. Let $\mathbf{X}_n^* = (X_1^*, X_2^*, \ldots, X_n^*)$ be a vector of n independent observations from $\hat{F}$ and define $S^* = S(\mathbf{X}_n^*)$. The distribution function $\hat{G} = G_{\hat{\mu}_0}$ of S^* is the bootstrap distribution function of S, i.e., $\hat{G}$ is the distribution of S under $\hat{F}$.

For any nominal level of significance α $(0 < \alpha < 1)$, let $c_\alpha(\hat{\mu}_0)$ be the upper-α quantile of $\hat{G}$. Thus $c_\alpha(\hat{\mu}_0)$ is the smallest value such that $\hat{G}(c_\alpha(\hat{\mu}_0)) \geq 1 - \alpha$. The nominal level-$\alpha$ bootstrap test rejects H_0 with probability 1 if $S > c_\alpha(\hat{\mu}_0)$, and with probability $[\alpha - 1 + \hat{G}(c_\alpha(\hat{\mu}_0))]/[\hat{G}(c_\alpha(\hat{\mu}_0)) - \hat{G}(c_\alpha(\hat{\mu}_0)-)]$ if $S = c_\alpha(\hat{\mu}_0)$ and $\hat{G}(c_\alpha(\hat{\mu}_0)) > \hat{G}(c_\alpha(\hat{\mu}_0)-)$.

2. Testing a Normal Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, a normal distribution with mean μ and variance σ^2. Let $\phi(x)$ and $\Phi(x)$ denote the density and distribution functions of the $N(0, 1)$ distribution and let z_α be its upper-α critical value, that is, $1 - \Phi(z_\alpha) = \alpha$. Consider testing

$$H_0 : \mu \leq 0 \quad \text{vs.} \quad H_1 : \mu > 0. \tag{2.1}$$

2.1. Known Variance

We assume without loss of generality that $\sigma^2 = 1$. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. The unrestricted MLE of μ is $\hat{\mu} = \bar{X}_n$. Let $\hat{\mu}_i$ be the MLE of μ under H_i $(i = 0, 1)$. Then $\hat{\mu}_0 = \bar{X}_n I(\bar{X}_n < 0)$, $\hat{\mu}_1 = \bar{X}_n I(\bar{X}_n > 0)$, and $X_1^*, X_2^*, \ldots, X_n^*$ is a bootstrap random sample drawn from $N(\hat{\mu}_0, 1)$. Let $\bar{X}_n^* = n^{-1}\sum_{i=1}^n X_i^*$.

2.1.1. Sample Mean Statistic

Theorem 2.1. *If $0 < \alpha \le 1/2$, the bootstrap test based on $\bar{X}_n$ is uniformly most powerful (UMP), but if $1/2 < \alpha < 1$, the test rejects H_0 with probability 1.*

Proof. Recall that the UMP test rejects H_0 if $\bar{X}_n \ge z_\alpha n^{-1/2}$. Since $\bar{X}_n^*$ is normal with mean $\hat{\mu}_0$ and variance n^{-1}, its critical value is $c_\alpha(\hat{\mu}_0) = \hat{\mu}_0 + z_\alpha n^{-1/2} = \bar{X}_n I(\bar{X}_n < 0) + z_\alpha n^{-1/2}$. Therefore the bootstrap test rejects H_0 if $\bar{X}_n I(\bar{X}_n > 0) \ge z_\alpha n^{-1/2}$. If $0 < \alpha \le 1/2$, then $z_\alpha \ge 0$ and the bootstrap test is the UMP test. If $1/2 < \alpha < 1$, then $z_\alpha < 0$ and the test rejects H_0 w.p.1.

2.1.2. Standard Likelihood Ratio Statistic

Given the data and values μ_0 and μ_1, let

$$L(\mu_0, \mu_1, \mathbf{X}_n) = \log\left\{\prod_{i=1}^n \phi(x_i - \mu_1) \Big/ \prod_{i=1}^n \phi(x_i - \mu_0)\right\}.$$

A general statistic for testing H_0 is the log-likelihood ratio

$$L(\hat{\mu}_0, \hat{\mu}, \mathbf{X}_n) = \log\left\{\sup_\mu \prod_{i=1}^n \phi(x_i - \mu) \Big/ \sup_{\mu \in H_0} \prod_{i=1}^n \phi(x_i - \mu)\right\}.$$

Throughout this article, we let Z denote the standard normal variable and $z_\alpha^+ = \max(z_\alpha, 0)$. We need the following lemma whose proof is given in the Appendix.

Lemma 2.1. *Let $\theta \ge 0$. For fixed $0 < \alpha < 1$, the function*

$$P(|Z + \theta| > z_\alpha^+) - (1 - \alpha)E\{\Phi(Z + \theta)^{-1} I(Z + \theta > z_\alpha^+)\} \tag{2.2}$$

is maximized at $\theta = 0$ with maximum value

$$\min(2\alpha, 1) + (1 - \alpha)\log\{1 - \min(\alpha, 1/2)\} \tag{2.3}$$

which is greater than α for all $0 < \alpha < 1$.

Theorem 2.2. *The size of the bootstrap test based on the standard likelihood ratio is* $\min(2\alpha, 1) + (1 - \alpha)\log\{1 - \min(\alpha, 1/2)\}$.

Proof. Since

$$\begin{aligned} n^{-1}L(\hat{\mu}_0, \hat{\mu}, \mathbf{X}_n) &= \bar{X}_n(\hat{\mu} - \hat{\mu}_0) - (\hat{\mu}^2 - \hat{\mu}_0^2)/2 \\ &= \bar{X}_n(\bar{X}_n - \hat{\mu}_0) - (\bar{X}_n^2 - \hat{\mu}_0^2)/2 \\ &= (\bar{X}_n - \hat{\mu}_0)^2/2 \\ &= \bar{X}_n^2 I(\bar{X}_n > 0)/2 \end{aligned}$$

the test rejects H_0 if $S = \bar{X}_n I(\bar{X}_n > 0) \ge c_\alpha(\hat{\mu}_0)$, where the critical value is to be determined. Let $S^* = \bar{X}_n^* I(\bar{X}_n^* > 0)$ and consider two cases.

1. $\bar{X}_n > 0$. Then $S > 0$, $\hat{\mu}_0 = 0$, and $\bar{X}_n^*$ has a $N(0, n^{-1})$ distribution. For any $x \geq 0$, $P(S^* \leq x) = P(\bar{X}_n^* \leq x) = \Phi(xn^{1/2})$. Therefore if $0 < \alpha < 1/2$, $c_\alpha(\hat{\mu}_0) = z_\alpha n^{-1/2}$. Otherwise, if $\alpha \geq 1/2$, then $c_\alpha(\hat{\mu}_0) = 0$ and the bootstrap test rejects H_0 w.p.1. Thus for all $0 < \alpha < 1$, $c_\alpha(\hat{\mu}_0) = z_\alpha^+ n^{-1/2}$.
2. $\bar{X}_n \leq 0$. Then $S = 0$, $\hat{\mu}_0 = \bar{X}_n \leq 0$, and S^* has a $N(\bar{X}_n, n^{-1})$ distribution left-truncated at 0 with $P(S^* = 0) = P(\bar{X}_n^* \leq 0) = \Phi(-n^{1/2}\bar{X}_n)$. Thus

$$\begin{aligned} c_\alpha(\hat{\mu}_0) &= \begin{cases} \bar{X}_n + n^{-1/2} z_\alpha, & \text{if } \bar{X}_n + n^{-1/2} z_\alpha > 0 \\ 0, & \text{if } \bar{X}_n + n^{-1/2} z_\alpha \leq 0. \end{cases} \\ &= (\bar{X}_n + n^{-1/2} z_\alpha)^+. \end{aligned}$$

Since $S = 0$, the bootstrap test never rejects H_0 if $\bar{X}_n + n^{-1/2} z_\alpha > 0$. Otherwise, the test is randomized and rejects H_0 with probability $\{\alpha - 1 + \Phi(-n^{1/2}\bar{X}_n)\}/\Phi(-n^{1/2}\bar{X}_n)$.

Thus for $0 < \alpha < 1$,

$$\begin{aligned} P\{\text{Reject } H_0\} &= P\{\text{Reject } H_0, \bar{X}_n > 0\} + P\{\text{Reject } H_0, \bar{X}_n < 0\} \\ &= P(S > z_\alpha^+, \bar{X}_n > 0) \\ &\quad + P\{\text{Reject } H_0, \bar{X}_n + n^{-1/2} z_\alpha \leq 0, \bar{X}_n < 0\} \\ &= P(\bar{X}_n > z_\alpha^+ n^{-1/2}) \\ &\quad + E[\{\alpha - 1 + \Phi(-n^{1/2}\bar{X}_n)\}/\Phi(-n^{1/2}\bar{X}_n)]\, I(-n^{1/2}\bar{X}_n \geq z_\alpha^+) \\ &= P(|W| > z_\alpha^+) - (1-\alpha) E\{\Phi(W)^{-1}\, I(W > z_\alpha^+)\} \end{aligned}$$

where W is normally distributed with mean $-n^{1/2}\mu$ and variance 1. By Lemma 2.1, the supremum of the rejection probability under H_0 is attained when $\mu = 0$ and is given by (2.3). Figure 1 shows a plot of this function.

2.1.3. Cox Likelihood Ratio Statistic

Cox (1961) proposed the following alternative likelihood ratio statistic for testing separate families of hypotheses:

$$L(\hat{\mu}_0, \hat{\mu}_1, \mathbf{X}_n) = \log \left\{ \sup_{H_1} \prod_{i=1}^n \phi(x_i - \mu) \Big/ \sup_{H_0} \prod_{i=1}^n \phi(x_i - \mu) \right\}.$$

For the current problem,

$$\begin{aligned} L(\hat{\mu}_0, \hat{\mu}_1, \mathbf{X}_n) &= n\{\bar{X}_n(\hat{\mu}_1 - \hat{\mu}_0) - (\hat{\mu}_1^2 - \hat{\mu}_0^2)/2\} \\ &= n(\bar{X}_n|\bar{X}_n| - \bar{X}_n|\bar{X}_n|/2) \\ &= n\bar{X}_n^2 \operatorname{sgn}(\bar{X}_n)/2. \end{aligned}$$

Therefore rejecting H_0 for large values of $L(\hat{\mu}_0, \hat{\mu}_1, \mathbf{X}_n)$ is equivalent to rejecting for large values of $\bar{X}_n$, and the next theorem follows directly from Theorem 2.1.

Theorem 2.3. *If $0 < \alpha \leq 1/2$, the bootstrap test based on the Cox likelihood ratio has size α and is UMP. If $1/2 < \alpha < 1$, it rejects H_0 with probability 1.*

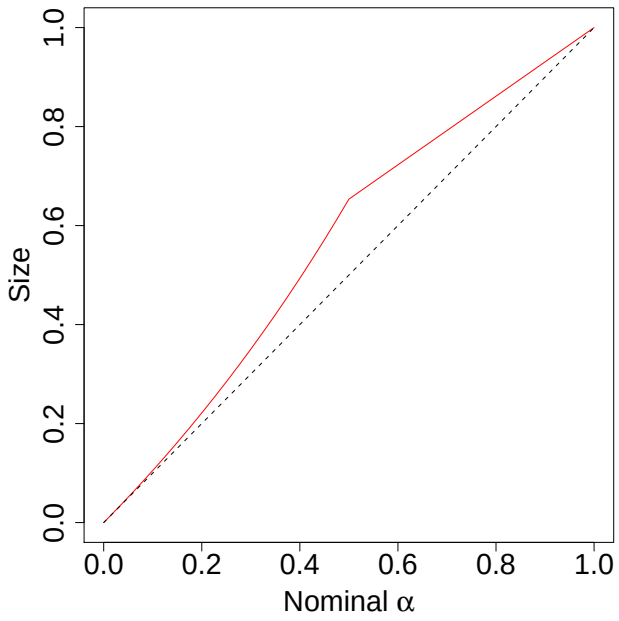


FIG 1. *Size (2.3) of bootstrap test for the normal mean based on the standard likelihood ratio, for known σ. The dashed line is the identity function.*

2.2. Unknown Variance

Now suppose we test the hypotheses (2.1) without assuming that σ is known. The log-likelihood function is

$$l(\mu, \sigma) = -n \log \sigma - \sum_i (X_i - \mu)^2/(2\sigma^2) - (n/2)\log(2\pi)$$

and its derivatives are $\partial l/\partial\mu = -\sigma^{-2}\sum(X_i - \mu)$ and $\partial l/\partial\sigma = -n\sigma^{-1} + \sigma^{-3}\sum(X_i - \mu)^2$. Hence the unrestricted and restricted (under H_0 and H_1) maximum likelihood estimates (MLEs) of μ and σ^2 are, respectively,

$$\begin{aligned} \hat{\mu} &= \bar{X}_n & \hat{\sigma}^2 &= n^{-1}\sum(X_i - \bar{X}_n)^2 \\ \hat{\mu}_0 &= \bar{X}_n I(\bar{X}_n < 0) & \hat{\sigma}_0^2 &= n^{-1}\sum(X_i - \hat{\mu}_0)^2 \\ \hat{\mu}_1 &= \bar{X}_n I(\bar{X}_n > 0) & \hat{\sigma}_1^2 &= n^{-1}\sum(X_i - \hat{\mu}_1)^2 \end{aligned}$$

giving the log-likelihood ratio statistics:

Standard: $n\log(\hat{\sigma}_0/\hat{\sigma}) = (n/2)\log\{\sum(X_i - \hat{\mu}_0)^2/\sum(X_i - \bar{X}_n)^2\}$
Cox: $n\log(\hat{\sigma}_0/\hat{\sigma}_1) = (n/2)\log\{\sum(X_i - \hat{\mu}_0)^2/\sum(X_i - \hat{\mu}_1)^2\}$.

The corresponding bootstrap tests reject H_0 for large values of $\sum(X_i - \hat{\mu}_0)^2/\sum(X_i - \bar{X}_n)^2$ and $\sum(X_i - \hat{\mu}_0)^2/\sum(X_i - \hat{\mu}_1)^2$, respectively.

2.2.1. Standard Likelihood Ratio Statistic

Let

$$T_n = n^{1/2}\bar{X}_n \left\{ \sum (X_i - \bar{X}_n)^2/(n-1) \right\}^{-1/2}. \tag{2.4}$$

The standard log-likelihood ratio statistic is

$$\begin{aligned}
\frac{\sum (X_i - \hat{\mu}_0)^2}{\sum (X_i - \bar{X}_n)^2} &= \frac{\sum \{X_i - \bar{X}_n I(\bar{X}_n < 0)\}^2}{\sum (X_i - \bar{X}_n)^2} \\
&= \begin{cases} 1, & \text{if } \bar{X}_n < 0 \\ \sum X_i^2 / \sum (X_i - \bar{X}_n)^2, & \text{if } \bar{X}_n \geq 0 \end{cases} \\
&= \begin{cases} 1, & \text{if } \bar{X}_n < 0 \\ 1 + n\bar{X}_n^2 / \sum (X_i - \bar{X}_n)^2, & \text{if } \bar{X}_n \geq 0 \end{cases} \\
&= \begin{cases} 1, & \text{if } \bar{X}_n < 0 \\ 1 + (n-1)^{-1} T_n^2, & \text{if } \bar{X}_n \geq 0. \end{cases}
\end{aligned}$$

Thus H_0 is rejected for large values of $S = T_n I(T_n > 0)$. Let $t_{\nu,\delta}$ denote the noncentral t-distribution with ν degrees of freedom and noncentrality parameter δ and let $t_{\nu,\delta,\alpha}$ denote its upper-α critical point.

Lemma 2.2. *For any ν and α, $t_{\nu,\delta,\alpha}$ is an increasing function of δ.*

Proof. Let Z denote a standard normal variable independent of χ^2_ν. Since

$$\begin{aligned}
P(t_{\nu,\delta} \leq x) &= P\left(\frac{Z + \delta}{\sqrt{\chi^2_\nu / \nu}} \leq x \right) \\
&= P\left(\frac{Z}{\sqrt{\chi^2_\nu / \nu}} \leq x - \frac{\delta}{\sqrt{\chi^2_\nu / \nu}} \right)
\end{aligned}$$

we see that $P(t_{\nu,\delta} \leq x)$ is a decreasing function of δ. Therefore $t_{\nu,\delta,\alpha}$ is an increasing function of δ.

Theorem 2.4. *If σ is unknown, the size of the nominal level-α test of $H_0 : \mu \leq 0$ vs. $H_1 : \mu > 0$ based on the standard likelihood ratio has lower bound*

$$\min(\alpha, 1/2) + E\left\{ \frac{\alpha - 1 + \Phi\left(-t_{n-1}\sqrt{n/(n-1)}\right)}{\Phi\left(-t_{n-1}\sqrt{n/(n-1)}\right)} I\left(t_{n-1}\sqrt{\frac{n}{n-1}} < -z^+_\alpha \right) \right\}$$

where t_{n-1} has a (central) t-distribution with $n-1$ degrees of freedom. As $n \to \infty$, the bound tends to (2.3), the size for the case where σ is known and n is finite.

Proof. Again, consider two cases.

1. $\bar{X}_n > 0$. Then $S > 0$ and $\hat{\mu}_0 = 0$. The bootstrap distribution of T^*_n is a central t_{n-1}-distribution and that of S^* is a central t_{n-1}-distribution left-truncated at 0. If $0 < \alpha < 1/2$, the test rejects H_0 whenever $T_n > t_{n-1,0,\alpha}$. Otherwise, if $\alpha \geq 1/2$, the test rejects H_0 with probability 1.
2. $\bar{X}_n < 0$. Then $S = 0$, $\hat{\mu}_0 < 0$, and S^* has a left-truncated noncentral $t_{n-1,\delta}$-distribution with $n-1$ degrees of freedom and noncentrality parameter

$$\delta = n^{1/2}\hat{\mu}_0/\hat{\sigma}_0 = n\bar{X}_n / \sqrt{\sum (X_i - \bar{X}_n)^2} = T_n \sqrt{n/(n-1)} \tag{2.5}$$

and probability $P(\bar{X}_n^* \le 0) = P\{n^{1/2}(\bar{X}_n^* - \hat{\mu}_0)/\hat{\sigma}_0 \le -n^{1/2}\hat{\mu}_0/\hat{\sigma}_0\} = \Phi(-\delta)$ at 0.

If $t_{n-1,\delta,\alpha} > 0$, the bootstrap test does not reject H_0 because $S = 0$. Otherwise, if $t_{n-1,\delta,\alpha} \le 0$, the test is randomized and rejects H_0 with probability $\{\alpha - 1 + \Phi(-\delta)\}/\Phi(-\delta)$. Note that the event $t_{n-1,\delta,\alpha} \le 0$ occurs if and only if $\alpha \ge P(T_n^* > 0 \,|\, \hat{\mu}_0, \hat{\sigma}_0)$. But

$$P(T_n^* > 0 \,|\, \hat{\mu}_0, \hat{\sigma}_0) = P(\bar{X}_n^* > 0 \,|\, \hat{\mu}_0, \hat{\sigma}_0) = 1 - \Phi(-\delta).$$

Therefore $t_{n-1,\delta,\alpha} \le 0$ if and only if $\delta \le -z_\alpha$.

Let $P_{\eta,\tau}$ denote probabilities when $\mu = \eta$ and $\sigma = \tau$. The size of the test for $0 < \alpha < 1$ is

$$\begin{aligned}
&\sup_{H_0} P_{\mu,\sigma}\{\text{Reject } H_0\} \\
&\quad = \sup_{H_0}[P_{\mu,\sigma}\{\text{Reject } H_0,\, \bar{X}_n > 0\} + P_{\mu,\sigma}\{\text{Reject } H_0,\, \bar{X}_n < 0\}] \\
&\quad = \sup_{H_0} P_{\mu,\sigma}[\{T_n I(T_n > 0) > t_{n-1,0,\alpha},\, \bar{X}_n > 0\} \\
&\qquad + P_{\mu,\sigma}\{\text{Reject } H_0,\, t_{n-1,\delta,\alpha} \le 0,\, \bar{X}_n < 0\}] \\
&\quad = \sup_{H_0}[P_{\mu,\sigma}\{T_n > \max(t_{n-1,0,\alpha}, 0)\} \\
&\qquad + E_{\mu,\sigma}[\{\alpha - 1 + \Phi(-\delta)\}/\Phi(-\delta)]\, I\{\delta < \min(-z_\alpha, 0)\}] \\
&\quad \ge P_{0,1}\{T_n > \max(t_{n-1,0,\alpha}, 0)\} + E_{0,1}[\{\alpha - 1 + \Phi(-\delta)\}/\Phi(-\delta)\, I\{\delta < -z_\alpha^+\}] \\
&\quad = \min(\alpha, 1/2) + E\left\{ \frac{\alpha - 1 + \Phi\left(-t_{n-1}\sqrt{\frac{n}{n-1}}\right)}{\Phi\left(-t_{n-1}\sqrt{\frac{n}{n-1}}\right)}\, I\left(t_{n-1}\sqrt{\frac{n}{n-1}} < -z_\alpha^+\right) \right\}
\end{aligned}$$

by equation (2.5). Since $t_{n-1} \to Z$ in distribution as $n \to \infty$, where Z is a standard normal variable,

$$\begin{aligned}
&\lim_{n\to\infty} E\left\{ \frac{\alpha - 1 + \Phi\left(-t_{n-1}\sqrt{\frac{n}{n-1}}\right)}{\Phi\left(-t_{n-1}\sqrt{\frac{n}{n-1}}\right)}\, I\left(t_{n-1}\sqrt{\frac{n}{n-1}} < -z_\alpha^+\right) \right\} \\
&\quad \to E\left\{ \frac{\alpha - 1 + \Phi(-Z)}{\Phi(-Z)}\, I(Z < -z_\alpha^+) \right\} \\
&\quad = (\alpha - 1)\int_{-\infty}^{-z_\alpha^+} \phi(z)/\Phi(-z)\, dz + \Phi(-z_\alpha^+) \\
&\quad = (\alpha - 1)\int_{z_\alpha^+}^{\infty} \phi(z)/\Phi(z)\, dz + \min(\alpha, 1/2) \\
&\quad = (1 - \alpha)\log\Phi(z_\alpha^+) + \min(\alpha, 1/2) \\
&\quad = (1 - \alpha)\log\{\max(1 - \alpha, 1/2)\} + \min(\alpha, 1/2) \\
&\quad = (1 - \alpha)\log\{1 - \min(\alpha, 1/2)\} + \min(\alpha, 1/2)
\end{aligned}$$

Thus the limiting size is $2\min(\alpha, 1/2) + (1 - \alpha)\log\{1 - \min(\alpha, 1/2)\} > \alpha$.

2.2.2. Cox Likelihood Ratio Statistic

Theorem 2.5. *If σ^2 is unknown, the size of the bootstrap test of (2.1) based on the Cox likelihood ratio has lower bound*

$$\min(\alpha, 1/2) + P(t_{n-1,t_{n-1}\sqrt{n/(n-1)},\alpha} < t_{n-1} < 0) \geq \alpha.$$

Proof. The Cox log-likelihood ratio statistic is

$$\begin{aligned}\frac{\sum(X_i-\hat{\mu}_0)^2}{\sum(X_i-\hat{\mu}_1)^2} &= \frac{\sum\{X_i-\bar{X}_nI(\bar{X}_n<0)\}^2}{\sum\{X_i-\bar{X}_nI(\bar{X}_n>0)\}^2}\\ &= \begin{cases}\{1+(n-1)T_n^2\}^{-1}, & \text{if } \bar{X}_n<0\\ 1, & \text{if } \bar{X}_n=0\\ 1+(n-1)T_n^2, & \text{if } \bar{X}_n>0\end{cases}\end{aligned}$$

where T_n is defined in (2.4). Thus rejecting for large values of the statistic is equivalent to rejecting for large values of $S = T_n$.

1. $\bar{X}_n > 0$. Then $\hat{\mu}_0 = 0$, $T_n > 0$, and T_n^* has a central t-distribution with $n-1$ degrees of freedom. Thus the test rejects H_0 if $T_n > t_{n-1,0,\alpha}$. If $1/2 \leq \alpha < 1$, then $t_{n-1,0,\alpha} \leq 0$ and the test rejects w.p.1.
2. $\bar{X}_n < 0$. Then $\hat{\mu}_0 < 0$, $T_n < 0$, and T_n^* has a noncentral t-distribution with $n-1$ degrees of freedom and noncentrality parameter δ given in (2.5). Hence H_0 is rejected if $T_n > t_{n-1,\delta,\alpha}$. Since $T_n < 0$, rejection occurs only if $t_{n-1,\delta,\alpha} < 0$.

If $0 < \alpha < 1/2$,

$$\begin{aligned}\sup_{H_0} P_{\mu,\sigma}(\text{Reject } H_0) &= \sup_{H_0}[P_{\mu,\sigma}\{\text{Reject } H_0,\, \bar{X}_n>0\}\\ &\quad + P_{\mu,\sigma}\{\text{Reject } H_0,\, \bar{X}_n<0\}]\\ &= \sup_{H_0}[P_{\mu,\sigma}(T_n > t_{n-1,0,\alpha}) + P_{\mu,\sigma}(t_{n-1,\delta,\alpha} < T_n < 0)]\\ &\geq P_{0,1}(T_n > t_{n-1,0,\alpha}) + P_{0,1}(t_{n-1,\delta,\alpha} < T_n < 0)\\ &= \alpha + P(t_{n-1,\delta,\alpha} < t_{n-1} < 0)\\ &\geq \alpha.\end{aligned}$$

If $1/2 \leq \alpha < 1$,

$$\begin{aligned}\sup_{H_0} P_{\mu,\sigma}(\text{Reject } H_0) &= \sup_{H_0}[P_{\mu,\sigma}\{\text{Reject } H_0,\, \bar{X}_n>0\}\\ &\quad + P_{\mu,\sigma}\{\text{Reject } H_0,\, \bar{X}_n<0\}]\\ &= \sup_{H_0}[P_{\mu,\sigma}(\bar{X}_n > 0) + P_{\mu,\sigma}(t_{n-1,\delta,\alpha} < T_n < 0)]\\ &\geq P_{0,1}(\bar{X}_n > 0) + P_{0,1}(t_{n-1,\delta,\alpha} < T_n < 0)\\ &= 1/2 + P(t_{n-1,\delta,\alpha} < t_{n-1} < 0)\\ &> 1/2 + P(t_{n-1,0,\alpha} < t_{n-1} < 0)\\ &= \alpha\end{aligned}$$

by Lemma 2.2.

3. Testing a Normal Variance, Mean Unknown

Let χ^2_ν denote a chi-squared random variable with ν degrees of freedom, $\chi^2_{\nu,\alpha}$ its upper-α point, and $\Psi_\nu(.)$ its cumulative distribution function.

Lemma 3.1. *$\Psi_{n-1}(n^2/\chi^2_{n-1,\alpha}) \to \alpha$ and $\Psi_{n-1}(n) \to 1/2$ as $n \to \infty$.*

Proof. Let $Z_1, Z_2, \ldots$ be independent $N(0,1)$ variables. Then

$$\begin{aligned}
\Psi_{n-1}(n^2/\chi^2_{n-1,\alpha}) &= P\left(\sum_{i=1}^{n-1} Z_i^2 \le n^2/\chi^2_{n-1,\alpha}\right) \\
&= P\left(\frac{\sum_{i=1}^{n-1}(Z_i^2-1)}{\sqrt{2(n-1)}} \le \sqrt{\frac{n-1}{2}}\left\{\frac{n^2}{(n-1)^2\chi^2_{n-1,\alpha}} - 1\right\}\right) \\
&\approx \Phi\left(\sqrt{(n-1)/2}\left\{\frac{n^2}{(n-1)^2\chi^2_{n-1,\alpha}} - 1\right\}\right) \text{ as } n \to \infty.
\end{aligned}$$

By the Wilson-Hilferty (1931) approximation, $\nu/\chi^2_{\nu,\alpha} = 1 - z_\alpha(2/\nu)^{1/2} + o(\nu^{-1})$. Therefore

$$\sqrt{(n-1)/2}\left\{\frac{n^2}{(n-1)^2\chi^2_{n-1,\alpha}} - 1\right\} \to -z_\alpha$$

which yields the first result. The second result is similarly proved.

Let $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ be a vector of n independent observations from $N(\mu, \sigma^2)$, with μ and σ unknown, and let $\hat\sigma^2 = n^{-1}\sum_{i=1}^n (X_i - \bar X_n)^2$ denote the unrestricted MLE of σ^2.

3.1. $H_0 : \sigma^2 \le 1$ vs. $H_1 : \sigma^2 > 1$

Let $\hat\sigma_i^2$ be the MLE of σ^2 under H_i $(i = 0, 1)$. Then $\hat\sigma_0^2 = \min(\hat\sigma^2, 1)$ and $\hat\sigma_1^2 = \max(\hat\sigma^2, 1)$. Define the log-likelihood ratio

$$M(\mu_0, \mu_1, \sigma_0, \sigma_1, \mathbf{X}_n) = \log\left(\frac{\prod_i \sigma_1^{-1}\phi\{\sigma_1^{-1}(x_i - \mu_1)\}}{\prod_i \sigma_0^{-1}\phi\{\sigma_0^{-1}(x_i - \mu_0)\}}\right).$$

3.1.1. Standard Likelihood Ratio Statistic

Theorem 3.1. *The size of the bootstrap test based on the standard likelihood ratio for testing $H_0 : \sigma^2 \le 1$ vs. $H_1 : \sigma^2 > 1$, with μ unknown, is bounded below by*

(3.1)

$$\min\{\alpha, 1-\Psi_{n-1}(n)\} + E\left[\frac{\alpha - 1 + \Psi_{n-1}(n^2/\chi^2_{n-1})}{\Psi_{n-1}(n^2/\chi^2_{n-1})}\, I\left\{\chi^2_{n-1} \le \min\left(n, \frac{n^2}{\chi^2_{n-1,\alpha}}\right)\right\}\right].$$

Proof. The standard log-likelihood ratio statistic is $M(\hat\mu, \hat\mu, \hat\sigma_0, \hat\sigma, \mathbf{X}_n)$ and

$$\begin{aligned}
2n^{-1}M(\hat\mu, \hat\mu, \hat\sigma_0, \hat\sigma, \mathbf{X}_n) &= \log(\hat\sigma_0^2\hat\sigma^{-2}) + \hat\sigma^2(\hat\sigma_0^{-2} - \hat\sigma^{-2}) \\
&= \hat\sigma^2\hat\sigma_0^{-2} - \log(\hat\sigma^2\hat\sigma_0^{-2}) - 1 \\
&= \begin{cases} 0, & \text{if } \hat\sigma^2 \le 1 \\ \hat\sigma^2 - \log(\hat\sigma^2) - 1, & \text{otherwise.} \end{cases}
\end{aligned}$$

Since the function $x - \log(x) - 1$ increases monotonically from 0 for $x > 1$, rejecting for large values of the statistic is equivalent to rejecting for large values of

$$S = n \max(\hat{\sigma}^2, 1) = \max\left\{\sum (X_i - \bar{X}_n)^2, n\right\}.$$

Let S^* denote the bootstrap version of S under resampling from $N(\bar{X}_n, \hat{\sigma}_0^2)$. To find the critical point of the distribution of S^*, consider two cases.

1. $\hat{\sigma}^2 > 1$. Then $S > n$, $\hat{\sigma}_0^2 = 1$, and the distribution of S^* is χ^2_{n-1} left-truncated at n, i.e., it has probability mass $\Psi_{n-1}(n)$ at n. If $0 < \alpha < 1 - \Psi_{n-1}(n)$, the critical point of the bootstrap distribution is $\chi^2_{n-1,\alpha}$. Otherwise, the critical point is n and the test rejects H_0 w.p.1.
2. $\hat{\sigma}^2 \le 1$. Then $S = n$, $\hat{\sigma}_0^2 = \hat{\sigma}^2$, and the distribution of S^* is $\hat{\sigma}^2\chi^2_{n-1}$ left truncated at n. Thus the test does not reject H_0 if $\hat{\sigma}^2\chi^2_{n-1,\alpha} > n$. On the other hand, if $\hat{\sigma}^2\chi^2_{n-1,\alpha} \le n$, then the critical point is n and the test rejects H_0 randomly with probability $\{\alpha - 1 + \Psi_{n-1}(n\hat{\sigma}_0^{-2})\}/\Psi_{n-1}(n\hat{\sigma}_0^{-2})$.

Since $\alpha < 1 - \Psi_{n-1}(n)$ if and only if $n < \chi^2_{n-1,\alpha}$, we have

$$\begin{aligned}
P_{\mu,\sigma}(\text{Reject } H_0,\, \hat{\sigma}^2 > 1) &= \begin{cases} P_{\mu,\sigma}(S > \chi^2_{n-1,\alpha},\, n\hat{\sigma}^2 > n), & \text{if } n < \chi^2_{n-1,\alpha} \\ P_{\mu,\sigma}(n\hat{\sigma}^2 > n), & \text{otherwise} \end{cases} \\
&= \begin{cases} P_{\mu,\sigma}(n\hat{\sigma}^2 > \chi^2_{n-1,\alpha}), & \text{if } n < \chi^2_{n-1,\alpha} \\ P_{\mu,\sigma}(n\hat{\sigma}^2 > n), & \text{otherwise} \end{cases} \\
&= \begin{cases} 1 - \Psi_{n-1}(\sigma^{-2}\chi^2_{n-1,\alpha}), & \text{if } n < \chi^2_{n-1,\alpha} \\ 1 - \Psi_{n-1}(n\sigma^{-2}), & \text{otherwise} \end{cases} \\
&= 1 - \Psi_{n-1}(\sigma^{-2}\max\{n, \chi^2_{n-1,\alpha}\}).
\end{aligned}$$

and

$$\begin{aligned}
&P_{\mu,\sigma}(\text{Reject } H_0,\, \hat{\sigma}^2 \le 1) = P_{\mu,\sigma}(\text{Reject } H_0,\, \hat{\sigma}^2\chi^2_{n-1,\alpha} \le n,\, n\hat{\sigma}^2 \le n) \\
&= E_{\mu,\sigma}\left[\frac{\alpha - 1 + \Psi_{n-1}(n\hat{\sigma}_0^{-2})}{\Psi_{n-1}(n\hat{\sigma}_0^{-2})} I(\hat{\sigma}_0^2\chi^2_{n-1,\alpha} \le n,\, n\hat{\sigma}^2 \le n)\right] \\
&= E\left[\frac{\alpha - 1 + \Psi_{n-1}(n^2\sigma^{-2}/\chi^2_{n-1})}{\Psi_{n-1}(n^2\sigma^{-2}/\chi^2_{n-1})} I(\chi^2_{n-1} \le \sigma^{-2}\min\{n, n^2/\chi^2_{n-1,\alpha}\})\right].
\end{aligned}$$

The choice $\sigma^2 = 1$ yields the lower bound

$$\begin{aligned}
&\sup_{H_0} P_{\mu,\sigma}(\text{Reject } H_0) \\
&\ge P_{\mu,1}(\text{Reject } H_0,\, \hat{\sigma}^2 > 1) + P_{\mu,1}(\text{Reject } H_0,\, \hat{\sigma}^2 \le 1) \\
&= 1 - \Psi_{n-1}(\max\{n, \chi^2_{n-1,\alpha}\}) \\
&\quad + E\left[\frac{\alpha - 1 + \Psi_{n-1}(n^2/\chi^2_{n-1})}{\Psi_{n-1}(n^2/\chi^2_{n-1})} I\left\{\chi^2_{n-1} \le \min\left(n, \frac{n^2}{\chi^2_{n-1,\alpha}}\right)\right\}\right] \\
&= \min\{\alpha, 1 - \Psi_{n-1}(n)\} \\
&\quad + E\left[\frac{\alpha - 1 + \Psi_{n-1}(n^2/\chi^2_{n-1})}{\Psi_{n-1}(n^2/\chi^2_{n-1})} I\left\{\chi^2_{n-1} \le \min\left(n, \frac{n^2}{\chi^2_{n-1,\alpha}}\right)\right\}\right].
\end{aligned}$$

Figure 2 shows graphs of the lower bound (3.1) for $n = 5, 10, 100$, and 500.

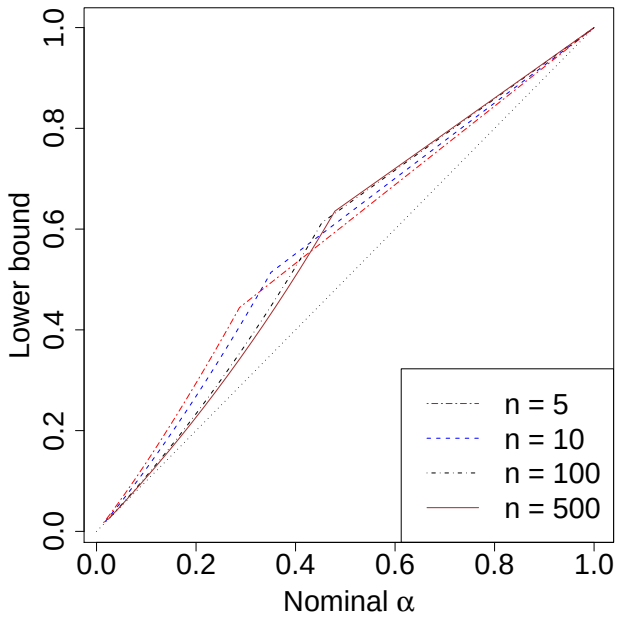


FIG 2. *Lower bounds (3.1) on the size of the bootstrap test of* $H_0 : \sigma^2 \le 1$ *vs.* $H_1 : \sigma^2 > 1$ *based on the standard likelihood ratio, for* $n = 5, 10, 100,$ *and 500. The 45-degree line is the identity function.*

3.1.2. Cox Likelihood Ratio Statistic

Theorem 3.2. *If* μ *is unknown, the bootstrap test of* $H_0 : \sigma^2 \le 1$ *vs.* $H_1 : \sigma^2 > 1$ *based on the Cox likelihood ratio has size* α *and is UMP for* $\chi^2_{n-1,\alpha} > n$*. It rejects* H_0 *w.p.1 for other values of* α*.*

Proof. The Cox log-likelihood ratio statistic is $M(\hat\mu, \hat\mu, \hat\sigma_0, \hat\sigma_1, \mathbf{X}_n)$. Since

$$\hat\sigma_0^2\,\hat\sigma_1^{-2} = \begin{cases} \hat\sigma^2, & \text{if } \hat\sigma^2 \le 1 \\ \hat\sigma^{-2}, & \text{if } \hat\sigma^2 > 1 \end{cases}$$

and

$$\hat\sigma_0^{-2} - \hat\sigma_1^{-2} = \begin{cases} \hat\sigma^{-2} - 1, & \text{if } \hat\sigma^2 \le 1 \\ 1 - \hat\sigma^{-2}, & \text{if } \hat\sigma^2 > 1 \end{cases}$$

we have

$$\begin{aligned} 2n^{-1}M(\hat\mu, \hat\mu, \hat\sigma_0, \hat\sigma_1, \mathbf{X}_n) &= \log(\hat\sigma_0^2\hat\sigma_1^{-2}) + \hat\sigma^2(\hat\sigma_0^{-2} - \hat\sigma_1^{-2}) \\ &= \begin{cases} \log(\hat\sigma^2) - \hat\sigma^2 + 1, & \text{if } \hat\sigma^2 \le 1 \\ -\log(\hat\sigma^2) + \hat\sigma^2 - 1, & \text{if } \hat\sigma^2 > 1 \end{cases} \end{aligned}$$

which is strictly increasing in $\hat\sigma^2$. Therefore rejecting for large values of the statistic is equivalent to rejecting for large values of $\hat\sigma^2$. Since the bootstrap null distribution of $n\hat\sigma^2$ is $\hat\sigma_0^2\chi^2_{n-1}$, the bootstrap critical point of $\hat\sigma^2$ is $n^{-1}\hat\sigma_0^2\chi^2_{n-1,\alpha}$. Thus the bootstrap test rejects H_0 if $\hat\sigma^2\,\hat\sigma_0^{-2} > n^{-1}\chi^2_{n-1,\alpha}$, or equivalently, $\max(\hat\sigma^2, 1) > n^{-1}\chi^2_{n-1,\alpha}$. If α is so large that $n^{-1}\chi^2_{n-1,\alpha} \le 1$, the bootstrap test rejects H_0 regardless of the data. On the other hand, if $n^{-1}\chi^2_{n-1,\alpha} > 1$, the test rejects H_0 if $\hat\sigma^2 > n^{-1}\chi^2_{n-1,\alpha}$, which coincides with the UMP test [10, p. 88].

3.2. $H_0 : \sigma^2 \geq 1$ vs. $H_1 : \sigma^2 < 1$

Next suppose we reverse the hypotheses and test $H_0 : \sigma^2 \geq 1$ versus $H_1 : \sigma^2 < 1$. Then $\hat{\sigma}_0^2 = \max(\hat{\sigma}^2, 1)$ and $\hat{\sigma}_1^2 = \min(\hat{\sigma}^2, 1)$.

3.2.1. Standard Likelihood Ratio Statistic

Theorem 3.3. *For any $0 < \alpha < 1$ and μ unknown, the size of the bootstrap test for $H_0 : \sigma^2 \geq 1$ vs. $H_1 : \sigma^2 < 1$, based on the standard likelihood ratio, is bounded below by*

(3.2)
$$\min\{\alpha, \Psi_{n-1}(n)\} + E\left[\frac{\alpha - \Psi_{n-1}(n^2/\chi^2_{n-1})}{1 - \Psi_{n-1}(n^2/\chi^2_{n-1})}\, I\left(\chi^2_{n-1} \geq \max\{n, n^2/\chi^2_{n-1,1-\alpha}\}\right)\right].$$

Proof. Direct computation yields

$$2n^{-1}M(\hat{\mu}, \hat{\mu}, \hat{\sigma}_0, \hat{\sigma}, \mathbf{X}_n) = \begin{cases} \hat{\sigma}^2 - \log(\hat{\sigma}^2) - 1, & \text{if } \hat{\sigma}^2 \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Since the function $x - \log(x) - 1$ decreases monotonically for $0 < x \leq 1$, the test rejects H_0 for small values of $S = n\min(\hat{\sigma}^2, 1)$. Let S^* denote the bootstrap version of S under resampling from $N(\bar{X}_n, \hat{\sigma}_0^2)$.

1. $\hat{\sigma}^2 < 1$. Then $\hat{\sigma}_0 = 1$ and the distribution of S^* is χ^2_{n-1} right-truncated at n, with probability mass $1 - \Psi_{n-1}(n)$ there. If $0 < \alpha < \Psi_{n-1}(n)$, the critical point of the bootstrap distribution is $\chi^2_{n-1,1-\alpha}$. Otherwise, the critical point is n and the test rejects w.p.1, because $S < n$.
2. $\hat{\sigma}^2 \geq 1$. Then $\hat{\sigma}_0^2 = \hat{\sigma}^2$, $S = n$ and the distribution of S^* is $\hat{\sigma}^2\chi^2_{n-1}$ right-truncated at n. The test does not reject H_0 if $\hat{\sigma}^2\chi^2_{n-1,1-\alpha} < n$. Otherwise, if $\hat{\sigma}^2\chi^2_{n-1,1-\alpha} \geq n$, the test rejects H_0 with probability $\{\alpha - \Psi_{n-1}(n\hat{\sigma}^{-2})\}/\{1 - \Psi_{n-1}(n\hat{\sigma}^{-2})\}$.

Since $\alpha < \Psi_{n-1}(n)$ if and only if $\chi^2_{n-1,1-\alpha} < n$,

$$\begin{aligned} & P_{\mu,\sigma}(\text{Reject } H_0,\, \hat{\sigma}^2 < 1) \\ &= \begin{cases} P_{\mu,\sigma}(S < \chi^2_{n-1,1-\alpha},\, n\hat{\sigma}^2 < n), & \text{if } 0 < \alpha < \Psi_{n-1}(n) \\ P_{\mu,\sigma}(n\hat{\sigma}^2 < n), & \text{otherwise} \end{cases} \\ &= \begin{cases} P_{\mu,\sigma}(n\hat{\sigma}^2 < \chi^2_{n-1,1-\alpha},\, n\hat{\sigma}^2 < n), & \text{if } 0 < \alpha < \Psi_{n-1}(n) \\ P_{\mu,\sigma}(n\hat{\sigma}^2 < n), & \text{otherwise} \end{cases} \\ &= \begin{cases} P_{\mu,\sigma}(n\hat{\sigma}^2 < \chi^2_{n-1,1-\alpha}), & \text{if } 0 < \alpha < \Psi_{n-1}(n) \\ P_{\mu,\sigma}(n\hat{\sigma}^2 < n), & \text{otherwise} \end{cases} \\ &= P_{\mu,\sigma}(n\hat{\sigma}^2 < \min\{\chi^2_{n-1,1-\alpha}, n\}) \\ &= \Psi_{n-1}(\sigma^{-2}\min\{\chi^2_{n-1,1-\alpha}, n\}) \end{aligned}$$

and

$$\begin{aligned} & P_{\mu,\sigma}(\text{Reject } H_0,\, \hat{\sigma}^2 \geq 1) \\ &= P_{\mu,\sigma}(\text{Reject } H_0,\, \hat{\sigma}^2\chi^2_{n-1,1-\alpha} \geq n,\, \hat{\sigma}^2 \geq 1) \\ &= E_{\mu,\sigma}\left[\frac{\alpha - \Psi_{n-1}(n\hat{\sigma}^{-2})}{1 - \Psi_{n-1}(n\hat{\sigma}^{-2})}\, I(\hat{\sigma}^2\chi^2_{n-1,1-\alpha} \geq n,\, n\hat{\sigma}^2 \geq n)\right] \\ &= E\left[\frac{\alpha - \Psi_{n-1}(n^2\sigma^{-2}/\chi^2_{n-1})}{1 - \Psi_{n-1}(n^2\sigma^{-2}/\chi^2_{n-1})}\, I(\sigma^2\chi^2_{n-1} \geq \max\{n, n^2/\chi^2_{n-1,1-\alpha}\})\right]. \end{aligned}$$

Therefore

$$\begin{aligned}
&\sup_{H_0} P_{\mu,\sigma}(\text{Reject } H_0) \\
&\quad \geq \quad \Psi_{n-1}(\min\{\chi^2_{n-1,1-\alpha}, n\}) \\
&\qquad\quad + E\left[\frac{\alpha - \Psi_{n-1}(n^2/\chi^2_{n-1})}{1 - \Psi_{n-1}(n^2/\chi^2_{n-1})}\, I(\chi^2_{n-1} \geq \max\{n, n^2/\chi^2_{n-1,1-\alpha}\})\right] \\
&\quad = \quad \min\{\alpha, \Psi_{n-1}(n)\} \\
&\qquad\quad + E\left[\frac{\alpha - \Psi_{n-1}(n^2/\chi^2_{n-1})}{1 - \Psi_{n-1}(n^2/\chi^2_{n-1})}\, I(\chi^2_{n-1} \geq \max\{n, n^2/\chi^2_{n-1,1-\alpha}\})\right].
\end{aligned}$$

Figure 3 shows graphs of the lower bound (3.2) for $n = 5, 10, 100$, and 500.

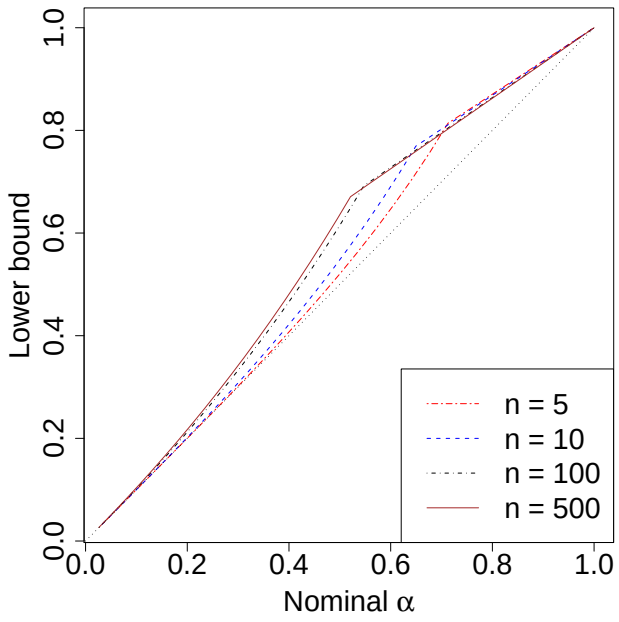


FIG 3. *Lower bounds (3.2) on the size of the bootstrap test of* $H_0 : \sigma^2 \geq 1$ *vs.* $H_1 : \sigma^2 < 1$ *based on the standard likelihood ratio* $M_n^{(1)}$. *The 45-degree line is the identity function.*

3.2.2. Cox Likelihood Ratio Statistic

Theorem 3.4. *The bootstrap test of* $H_0 : \sigma^2 \geq 1$ *vs.* $H_1 : \sigma^2 < 1$ *based on the Cox likelihood ratio has size* α *and is UMP if* $\chi^2_{n-1,\alpha} > n$. *Otherwise, it rejects* H_0 *w.p.1.*

Proof. Since

$$\hat\sigma_0^2\, \hat\sigma_1^{-2} = \begin{cases} \hat\sigma^{-2}, & \text{if } \hat\sigma^2 < 1 \\ \hat\sigma^2, & \text{if } \hat\sigma^2 \geq 1 \end{cases}$$

and

$$\hat\sigma_0^{-2} - \hat\sigma_1^{-2} = \begin{cases} 1 - \hat\sigma^{-2}, & \text{if } \hat\sigma^2 < 1 \\ \hat\sigma^{-2} - 1, & \text{if } \hat\sigma^2 \geq 1 \end{cases}$$

we have

$$\begin{aligned} 2n^{-1}M(\hat{\mu},\hat{\mu},\hat{\sigma}_0,\hat{\sigma}_1,\mathbf{X}_n) &= \log(\hat{\sigma}_0^2\hat{\sigma}_1^{-2}) + \hat{\sigma}^2(\hat{\sigma}_0^{-2} - \hat{\sigma}_1^{-2}) \\ &= \begin{cases} -\log(\hat{\sigma}^2) + \hat{\sigma}^2 - 1, & \text{if } \hat{\sigma}^2 < 1 \\ \log(\hat{\sigma}^2) - \hat{\sigma}^2 + 1, & \text{if } \hat{\sigma}^2 \geq 1 \end{cases} \end{aligned}$$

a strictly decreasing function of $\hat{\sigma}^2$. Thus the test statistic rejects H_0 for small values of $\hat{\sigma}^2$. The bootstrap null distribution of $\hat{\sigma}^2$ is $n^{-1}\hat{\sigma}_0^2\chi^2_{n-1}$, with critical value $n^{-1}\hat{\sigma}_0^2\chi^2_{n-1,1-\alpha}$. Hence the bootstrap test rejects H_0 if $\hat{\sigma}^2\,\hat{\sigma}_0^{-2} < n^{-1}\chi^2_{n-1,1-\alpha}$. But the left side of the inequality is never greater than 1, because

$$\hat{\sigma}^2\,\hat{\sigma}_0^{-2} = \begin{cases} \hat{\sigma}^2, & \text{if } \hat{\sigma}^2 < 1 \\ 1, & \text{otherwise.} \end{cases}$$

Therefore, if α is so large that $n^{-1}\chi^2_{n-1,1-\alpha} \geq 1$, the bootstrap test rejects H_0 w.p.1. Otherwise, if $n^{-1}\chi^2_{n-1,1-\alpha} < 1$, the test rejects H_0 if $\hat{\sigma}^2 < n^{-1}\chi^2_{n-1,1-\alpha}$, which coincides with the classical UMP unbiased test [10, pp. 154].

4. Testing Difference of Two Normal Means

Let $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ be independent random samples from $N(\mu,\sigma^2)$ and $N(\eta,\tau^2)$, respectively, and $N = m+n > 2$. We want to test

$$H_0 : \eta \leq \mu \quad \text{vs.} \quad H_1 : \eta > \mu. \tag{4.1}$$

The likelihood function for this case is

$$\begin{aligned} &L(\mu,\tau) \\ &= (2\pi)^{-(m+n)/2}\sigma^{-m}\tau^{-n}\exp\left\{-(2\sigma^2)^{-1}\sum(X_i-\mu)^2 - (2\tau^2)^{-1}\sum(Y_j-\eta)^2\right\} \\ &= (2\pi)^{-(m+n)/2}\sigma^{-m}\tau^{-n}\exp\{-(2\sigma^2)^{-1}\sum(X_i-\bar{X}_m)^2 - (2\tau^2)^{-1}\sum(Y_j-\bar{Y}_n)^2 \\ &\quad - m(2\sigma^2)^{-1}(\mu-\bar{X}_m)^2 - n(2\tau^2)^{-1}(\eta-\bar{Y}_n)^2\} \end{aligned}$$

and the unrestricted MLE of (μ,η) is $(\hat{\mu},\hat{\eta}) = (\bar{X}_m,\bar{Y}_n)$.

4.1. Known Variances

Let $V = (m\tau^2\bar{X}_m + n\sigma^2\bar{Y}_n)/(m\tau^2+n\sigma^2)$. The MLE of (μ,η) under H_0 is

$$(\hat{\mu}_0,\hat{\eta}_0) = \begin{cases} (\bar{X}_m,\bar{Y}_n), & \bar{Y}_n \leq \bar{X}_m \\ (V,V), & \bar{Y}_n > \bar{X}_m \end{cases}$$

and that under H_1 is

$$(\hat{\mu}_1,\hat{\eta}_1) = \begin{cases} (V,V), & \bar{Y}_n > \bar{X}_m \\ (\bar{X}_m,\bar{Y}_n), & \bar{Y}_n \leq \bar{X}_m. \end{cases}$$

4.1.1. Difference of Means Statistic

Theorem 4.1. *The size of the bootstrap test of (4.1) based on $\bar{Y}_n - \bar{X}_m$ is α if $\alpha < 1/2$ and is 1 if $\alpha \geq 1/2$.*

Proof. Let $S = \bar{Y}_n - \bar{X}_m$. The bootstrap test statistic $S^* = \bar{Y}_n^* - \bar{X}_m^*$ has a normal distribution with mean $\hat{\eta}_0 - \hat{\mu}_0 = S\,I(S < 0)$ and variance $m^{-1}\sigma^2 + n^{-1}\tau^2$. Thus its nominal level-α bootstrap critical value is $S\,I(S < 0) + z_\alpha\{m^{-1}\sigma^2 + n^{-1}\tau^2\}^{1/2}$ and the rejection region is $\max(S, 0) > z_\alpha\{m^{-1}\sigma^2 + n^{-1}\tau^2\}^{1/2}$. Clearly, the size of the test is attained at the boundary $\mu = \eta$. If $\alpha < 1/2$, the probability of rejecting H_0 when $\mu = \eta$ is exactly α. On the other hand, if $\alpha \geq 1/2$, then $z_\alpha \leq 0$ and the test rejects H_0 w.p.1.

4.1.2. Standard Likelihood Ratio Statistic

Theorem 4.2. *The size of the bootstrap test of (4.1) based on the standard likelihood ratio is* $\min(2\alpha, 1) + (1 - \alpha)\log\{1 - \min(\alpha, 1/2)\} > \alpha$.

Proof. The log-likelihood ratio statistic is

$$\begin{aligned}
&\log\{L(\hat{\mu}, \hat{\tau})/L(\hat{\mu}_0, \hat{\tau}_0)\} \\
&\quad = \quad \{m(2\sigma^2)^{-1}(V - \bar{X}_m)^2 + n(2\tau^2)^{-1}(V - \bar{Y}_n)^2\} I(\bar{Y}_n > \bar{X}_m) \\
&\quad = \quad mn(m\tau^2 + n\sigma^2)^{-2}(\bar{Y}_n - \bar{X}_m)^2\, I(\bar{Y}_n > \bar{X}_m).
\end{aligned}$$

Thus the test statistic is equivalent to $S = (\bar{Y}_n - \bar{X}_m)\, I(\bar{Y}_n > \bar{X}_m)$. The bootstrap distribution of S^* is normal with mean $\hat{\eta}_0 - \hat{\mu}_0$ and variance $n^{-1}\tau^2 + m^{-1}\sigma^2$, left-truncated at 0 with $P(S^* = 0) = \Phi\{(\hat{\mu}_0 - \hat{\eta}_0)(n^{-1}\tau^2 + m^{-1}\sigma^2)^{-1/2}\}$. Let $\delta = (\mu - \eta)(n^{-1}\tau^2 + m^{-1}\sigma^2)^{-1/2}$ and $W = (\bar{X}_m - \bar{Y}_n)(n^{-1}\tau^2 + m^{-1}\sigma^2)^{-1/2} \sim N(\delta, 1)$. We consider two cases.

1. $\bar{Y}_n \leq \bar{X}_m$. Then $S = 0$, $\hat{\eta}_0 - \hat{\mu}_0 = \bar{Y}_n - \bar{X}_m$, and $\Phi(W) \geq 1/2$. If $1 - \Phi(W) < \alpha$, the test is randomized and rejects H_0 with probability $\{\alpha - 1 + \Phi(W)\}/\Phi(W)$. Otherwise, if $1 - \Phi(W) \geq \alpha$, the test does not reject H_0.
2. $\bar{Y}_n > \bar{X}_m$. Then $S = \bar{Y}_n - \bar{X}_m > 0$, $\hat{\eta}_0 - \hat{\mu}_0 = 0$, and $P(S^* = 0) = 1/2$. If $\alpha < 1/2$, then the test rejects H_0 if $\bar{Y}_n - \bar{X}_m > z_\alpha(n^{-1}\tau^2 + m^{-1}\sigma^2)^{-1/2}$, i.e., $W < -z_\alpha$. Otherwise, if $\alpha \geq 1/2$, then the critical value is 0 and the test rejects w.p.1.

Therefore

$$\begin{aligned}
&P(\text{Reject } H_0) \\
&\quad = \quad P(\text{Reject } H_0,\, \bar{Y}_n \leq \bar{X}_m,\, 1 - \Phi(W) < \alpha) \\
&\qquad + P(\text{Reject } H_0,\, \bar{Y}_n > \bar{X}_m)\, I(\alpha < 1/2) + P(\text{Reject } H_0,\, \bar{Y}_n > \bar{X}_m)\, I(\alpha \geq 1/2) \\
&\quad = \quad E[\Phi(W)^{-1}\{\alpha - 1 + \Phi(W)\}\, I(W > z_\alpha^+)] \\
&\qquad + P(W < -z_\alpha)\, I(\alpha < 1/2) + P(W < 0)\, I(\alpha > 1/2) \\
&\quad = \quad E[\Phi(W)^{-1}\{\alpha - 1 + \Phi(W)\}\, I(W > z_\alpha^+)] + P(W < -z_\alpha^+)
\end{aligned}$$

and the result follows from Lemma 2.1.

4.1.3. Cox Likelihood Ratio Statistic

Theorem 4.3. *The bootstrap test of (4.1) based on the Cox likelihood ratio statistic is the same as that based on the difference of sample means; its size is* α *if* $\alpha < 1/2$ *and is 1 if* $\alpha \geq 1/2$.

Proof. The Cox log-likelihood ratio statistic is

$$\log\left\{\frac{L(\hat{\mu}_1,\hat{\tau}_1)}{L(\hat{\mu}_0,\hat{\tau}_0)}\right\} = \frac{mn(\bar{Y}_n-\bar{X}_m)^2}{2(m\tau^2+n\sigma^2)}\{I(\bar{Y}_n>\bar{X}_m)-I(\bar{Y}_n\le\bar{X}_m)\}.$$

Thus the test statistic is equivalent to $S=\bar{Y}_n-\bar{X}_m$ and the result follows from Theorem 4.1.

4.2. Unknown but Equal Variances

Suppose that $\tau^2=\sigma^2$ but their value is unknown. Then the likelihood function is

$$L(\mu,\tau,\sigma) = (2\pi\sigma^2)^{-N/2}\exp\left[-(2\sigma^2)^{-1}\left\{\sum_i(X_i-\mu)^2+\sum_j(Y_j-\eta)^2\right\}\right]$$

giving the unrestricted MLE

$$(\hat{\mu},\hat{\eta},\hat{\sigma}^2) = \left(\bar{X}_m,\bar{Y}_n,N^{-1}\left\{\sum_i(X_i-\bar{X}_m)^2+\sum_j(Y_j-\bar{Y}_n)^2\right\}\right).$$

Let $V=N^{-1}(m\bar{X}_m+n\bar{Y}_n)$ and

$$\begin{aligned}\widetilde{\sigma}^2 &= N^{-1}\left\{\sum_i(X_i-V)^2+\sum_j(Y_j-V)^2\right\}\\ &= \hat{\sigma}^2+mnN^{-2}(\bar{Y}_n-\bar{X}_m)^2.\end{aligned}$$

The corresponding MLEs under H_0 and H_1 are, respectively,

$$\begin{aligned}(\hat{\mu}_0,\hat{\eta}_0,\hat{\sigma}_0^2) &= \begin{cases}(\bar{X}_m,\bar{Y}_n,\hat{\sigma}^2), & \text{if } \bar{Y}_n<\bar{X}_m\\ (V,V,\widetilde{\sigma}^2), & \text{if } \bar{Y}_n\ge\bar{X}_m\end{cases}\\ (\hat{\mu}_1,\hat{\eta}_1,\hat{\sigma}_1^2) &= \begin{cases}(V,V,\widetilde{\sigma}^2), & \text{if } \bar{Y}_n<\bar{X}_m\\ (\bar{X}_m,\bar{Y}_n,\hat{\sigma}^2), & \text{if } \bar{Y}_n\ge\bar{X}_m.\end{cases}\end{aligned}$$

4.2.1. Difference of Means Statistic

Suppose $S=\bar{Y}_n-\bar{X}_m$. Then $S^*=\bar{Y}_n^*-\bar{X}_m^*$ has a normal distribution with mean $\hat{\eta}_0-\hat{\mu}_0$ and variance $N(mn)^{-1}\hat{\sigma}_0^2$. Let Υ_ν denote the t distribution function with ν degrees of freedom and let $s^2=\hat{\sigma}^2N(N-2)^{-1}$ be the usual unbiased estimate of σ^2.

Theorem 4.4. *The size of the bootstrap test of (4.1) based on $\bar{Y}_n-\bar{X}_m$ is*

$$\begin{aligned}&\sup_{H_0}P(\textit{Reject } H_0)\\ &= \begin{cases}0, & \textit{if } \alpha\le 1-\Phi(\sqrt{N})\\ 1-\Upsilon_{N-2}\left(z_\alpha\sqrt{\frac{N-2}{N-z_\alpha^2}}\right), & \textit{if } 1-\Phi(\sqrt{N})<\alpha<1/2\\ 1, & \textit{if } \alpha\ge 1/2\end{cases} \qquad (4.2)\\ &\to \begin{cases}\alpha, & \textit{if } \alpha<1/2\\ 1, & \textit{if } \alpha\ge 1/2\end{cases}\end{aligned}$$

as $N\to\infty$.

Proof. The hypothesis H_0 is rejected if

$$\begin{aligned} \bar{Y}_n - \bar{X}_m &> \hat{\eta}_0 - \hat{\mu}_0 + z_\alpha \hat{\sigma}_0 \sqrt{N/(mn)} \\ &= \begin{cases} \bar{Y}_n - \bar{X}_m + z_\alpha \hat{\sigma} \sqrt{N/(mn)}, & \text{if } \bar{Y}_n < \bar{X}_m \\ z_\alpha \widetilde{\sigma} \sqrt{N/(mn)}, & \text{if } \bar{Y}_n \geq \bar{X}_m. \end{cases} \end{aligned}$$

1. $\alpha < 1/2$. If $\bar{Y}_n < \bar{X}_m$, the test does not reject H_0. Otherwise, if $\bar{Y}_n \geq \bar{X}_m$, the test rejects H_0 if

$$\begin{aligned} & (\bar{Y}_n - \bar{X}_m)^2 > z_\alpha^2 \widetilde{\sigma}^2 N/(mn) \\ \iff \quad & (\bar{Y}_n - \bar{X}_m)^2 (1 - N^{-1} z_\alpha^2) > N z_\alpha^2 \hat{\sigma}^2/(mn) \\ \iff \quad & \alpha > 1 - \Phi(\sqrt{N}) \quad \text{and} \quad \sqrt{\frac{mn}{N}} \frac{\bar{Y}_n - \bar{X}_m}{s} > \sqrt{\frac{N-2}{N - z_\alpha^2}}. \end{aligned}$$

   Therefore if $\alpha \leq 1 - \Phi(\sqrt{N})$, the test does not reject H_0. Otherwise, if $1 - \Phi(\sqrt{N}) < \alpha < 1/2$, the rejection probability is maximized when $\eta = \mu$ at $1 - \Upsilon_{N-2}\left(z_\alpha \sqrt{(N-2)/(N - z_\alpha^2)}\right)$.
2. $\alpha \geq 1/2$. The test rejects H_0 w.p.1 because $z_\alpha < 0$.

Hence the result (4.2). The limit is due to $\Upsilon_\nu(x) \to \Phi(x)$ as $\nu \to \infty$, for every x. Figure 4 plots the size function (4.2) for $N = 3, 5, 10$, and 100.

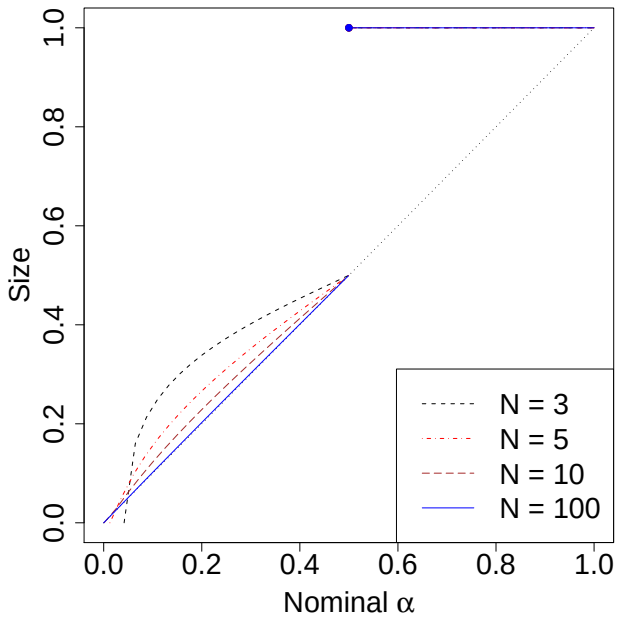


FIG 4. *Size of bootstrap test for a difference of normal means based on the difference of sample means, for equal but unknown variances (4.2). The 45-degree line is the identity function.*

4.2.2. Standard Likelihood Ratio Statistic

The log-likelihood ratio statistic is

$$(N/2)\log(\tilde{\sigma}^2/\hat{\sigma}^2)\, I(\bar{Y}_n > \bar{X}_m) = (N/2)\log\{1+mnN^{-2}(\bar{Y}_n-\bar{X}_m)^2\hat{\sigma}^{-2}\}\, I(\bar{Y}_n > \bar{X}_m)$$

which is equivalent to the positive part of the t-statistic:

$$S = \sqrt{mn/N}s^{-1}(\bar{Y}_n - \bar{X}_m)^+.$$

Theorem 4.5. *The size of the bootstrap test of (4.1) based on the standard likelihood ratio is bounded below by*

$$\min(\alpha, 1/2)+E\left[\frac{\alpha-1+\Phi\left(\sqrt{N/(N-2)}\,t_{N-2}\right)}{\Phi\left(\sqrt{N/(N-2)}\,t_{N-2}\right)}\, I\left(\sqrt{N/(N-2)}\,t_{N-2} \geq z_\alpha^+\right)\right]$$
$$\to \min(2\alpha, 1) + (1-\alpha)\log\{1-\min(\alpha,1/2)\}, \quad N\to\infty.$$

Proof. We again consider two situations.

1. $\bar{Y}_n > \bar{X}_m$. The bootstrap distribution of S^* is a t_{N-2} distribution left-truncated at 0 with probability 1/2. If $\alpha < 1/2$, then the test rejects H_0 if $S > t_{N-2,\alpha}$. Otherwise, if $\alpha \geq 1/2$, the test rejects H_0 w.p.1.
2. $\bar{Y}_n \leq \bar{X}_m$. The bootstrap distribution of S^* consists of the positive part of a noncentral t with $N-2$ degrees of freedom and noncentrality parameter $\delta = \sqrt{mn/N}\,\hat{\sigma}^{-1}(\bar{Y}_n - \bar{X}_m) = S\sqrt{N/(N-2)}$ and probability at 0 equal to $\Phi(-\delta)$. Since $S=0$, the test does not reject H_0 if $\alpha < 1-\Phi(-\delta) \iff -\delta < z_\alpha$. Otherwise, if $-\delta \geq z_\alpha$, then the test is randomized, rejecting H_0 with probability $\{\alpha-1+\Phi(-\delta)\}/\Phi(-\delta)$.

Thus

$$\begin{aligned}
&P(\text{Reject } H_0)\\
&= P(S > t_{N-2,\alpha},\, \bar{Y}_n > \bar{X}_m)\, I(\alpha < 1/2) + P(\bar{Y}_n > \bar{X}_m)\, I(\alpha \geq 1/2)\\
&\quad + P(\text{Reject } H_0,\, -\delta \geq z_\alpha,\, \bar{Y}_n \leq \bar{X}_m)\\
&= P(S > t_{N-2,\alpha})\, I(\alpha < 1/2) + (1/2) I(\alpha \geq 1/2)\\
&\quad + E\left\{\frac{\alpha-1+\Phi(-\delta)}{\Phi(-\delta)}\, I(-\delta \geq z_\alpha,\, \bar{Y}_n \leq \bar{X}_m)\right\}\\
&= \min(\alpha, 1/2) + E\left\{\frac{\alpha-1+\Phi(-\delta)}{\Phi(-\delta)}\, I(-\delta \geq z_\alpha,\, \bar{Y}_n \leq \bar{X}_m)\right\}\\
&= \min(\alpha, 1/2) + E\left\{\frac{\alpha-1+\Phi(-\delta)}{\Phi(-\delta)}\, I(-\delta \geq z_\alpha^+)\right\}.
\end{aligned}$$

Evaluating this probability at $\mu = \eta$ yields

$$\begin{aligned}
&\sup_{H_0} P(\text{Reject } H_0)\\
&\geq \min(\alpha, 1/2)\\
&\quad + E\left[\frac{\alpha-1+\Phi\left(\sqrt{N/(N-2)}\,t_{N-2}\right)}{\Phi\left(\sqrt{N/(N-2)}\,t_{N-2}\right)}\, I\left(\sqrt{N/(N-2)}\,t_{N-2} \geq z_\alpha^+\right)\right]\\
&\to \min(2\alpha, 1) + (1-\alpha)\log\{1-\min(\alpha,1/2)\}
\end{aligned}$$

as $N\to\infty$ by Lemma 2.1.

4.2.3. Cox Likelihood Ratio Statistic

Theorem 4.6. *The size of the bootstrap test of (4.1) based on the Cox likelihood ratio or the ordinary t-statistic is*

$$P(t_{N-2,t_{N-2}\sqrt{N/(N-2)},\,\alpha} < t_{N-2} < 0) + P(t_{N-2} > t^{+}_{N-2,\alpha}) \geq \alpha.$$

Proof. The Cox log-likelihood ratio simplifies to

$$(N/2)\log\{1 + mn(\bar{Y}_n - \bar{X}_m)^2 N^{-2}\hat{\sigma}^{-2}\}\{(I(\bar{Y}_n \geq \bar{X}_m) - I(\bar{Y}_n < \bar{X}_m)\}$$

which is an increasing function of the Student t statistic $S = \sqrt{mn/N}s^{-1}(\bar{Y}_n - \bar{X}_m)$. The bootstrap distribution of S is a noncentral $t_{N-2,\delta}$ with $N-2$ degrees of freedom and noncentrality parameter

$$\begin{aligned}
\delta &= \sqrt{mn/N}\hat{\sigma}_0^{-1}(\hat{\eta}_0 - \hat{\mu}_0) \\
&= \begin{cases} \sqrt{mn/N}\hat{\sigma}^{-1}(\bar{Y}_n - \bar{X}_m), & \text{if } \bar{Y}_n < \bar{X}_m \\ 0, & \text{if } \bar{Y}_n \geq \bar{X}_m \end{cases} \\
&= \begin{cases} \sqrt{N/(N-2)}S, & \text{if } \bar{Y}_n < \bar{X}_m \\ 0, & \text{if } \bar{Y}_n \geq \bar{X}_m. \end{cases}
\end{aligned}$$

Therefore

$$\begin{aligned}
P(\text{Reject } H_0) &= P(S > t_{N-2,\delta,\alpha},\ \bar{Y}_n < \bar{X}_m) + P(S > t_{N-2,\alpha},\ \bar{Y}_n \geq \bar{X}_m) \\
&= P(t_{N-2,\delta,\alpha} < S < 0) + P(S > t^{+}_{N-2,\alpha}).
\end{aligned}$$

Evaluating the probabilities at $\mu = \eta$ yields

$$\begin{aligned}
\sup_{H_0} P(\text{Reject } H_0) &\geq P(t_{N-2,t_{N-2}\sqrt{N/(N-2)},\alpha} < t_{N-2} < 0) + P(t_{N-2} > t^{+}_{N-2,\alpha}) \\
&\geq P(t_{N-2,0,\alpha} < t_{N-2} < 0) + \min(\alpha, 1/2) \\
&= (\alpha - 1/2)\, I(\alpha > 1/2) + \min(\alpha, 1/2) \\
&= \alpha.
\end{aligned}$$

5. Testing an Exponential Location Parameter

Let Exp(θ, τ) denote the distribution with density $\tau^{-1}\exp\{-\tau^{-1}(x-\theta)\}$, $x \geq \theta$. We consider testing hypotheses about θ with $\tau = 1$. The likelihood for a sample $X_1, \ldots, X_n$ from an Exp(θ,1) distribution is $\prod \exp\{-(x_i - \theta)\}I(x_{(1)} \geq \theta)$, where $x_{(1)}$ is the smallest order statistic. The unconstrained MLE is $\hat{\theta} = X_{(1)}$.

5.1. $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$

The MLE of θ is $\hat{\theta}_0 = \min(X_{(1)}, 0)$ and $\hat{\theta}_1 = \max(X_{(1)}, 0)$ under H_0 and H_1, respectively. Given $X_{(1)}$, the bootstrap data are independent observations from an Exp($\hat{\theta}_0$, 1) distribution.

5.1.1. Standard Likelihood Ratio Statistic

The standard log-likelihood ratio statistic is

$$\begin{aligned} S &= \sum_{i=1}^{n} \log\left[\frac{\exp\{-(X_i-\hat\theta)\}I(X_{(1)}\ge\hat\theta)}{\exp\{-(X_i-\hat\theta_0)\}I(X_{(1)}\ge\hat\theta_0)}\right] \\ &= n(\hat\theta-\hat\theta_0) \\ &= \begin{cases} 0, & X_{(1)}\le 0 \\ nX_{(1)}, & X_{(1)}\ge 0. \end{cases} \end{aligned}$$

Given $\hat\theta_0$, the bootstrap distribution of S is $\mathrm{Exp}(n\hat\theta_0,1)$, left-truncated at 0 with probability mass $1-\exp(n\hat\theta_0)$ there.

1. $X_{(1)}\ge 0$. Then $S=nX_{(1)}$, $\hat\theta_0=0$, the distribution of S^* is $\mathrm{Exp}(0,1)$ with upper-α critical point $\log(1/\alpha)$, and the test rejects H_0 if $nX_{(1)}>-\log\alpha$.
2. $X_{(1)}\le 0$. Then $S=0$, $\hat\theta_0=X_{(1)}$, and the distribution of S^* is $\mathrm{Exp}(nX_{(1)},1)$, left-truncated at 0 with probability $1-\exp(nX_{(1)})$ there. If $\alpha<\exp(nX_{(1)})$, the test never rejects H_0. Otherwise, the test rejects H_0 with probability $\{\alpha-\exp(nX_{(1)})\}/\{1-\exp(nX_{(1)})\}$.

Since $nX_{(1)}$ has an $\mathrm{Exp}(n\theta,1)$ distribution,

$$\begin{aligned} &P_\theta\{\text{Reject } H_0\} \\ &= P_\theta\{\text{Reject } H_0,\, X_{(1)}\ge 0\} + P_\theta\{\text{Reject } H_0,\, X_{(1)}<0\} \\ &= P_\theta(nX_{(1)}>-\log\alpha,\, X_{(1)}\ge 0) \\ &\quad + P_\theta\{\text{Reject } H_0,\, X_{(1)}<0,\, \exp(n\theta)<\exp(nX_{(1)})\le\alpha\} \\ &= P_\theta(nX_{(1)}>-\log\alpha) + E_\theta\left[\frac{\alpha-\exp(nX_{(1)})}{1-\exp(nX_{(1)})}\,I(n\theta<nX_{(1)}\le\log\alpha)\right] \\ &= \begin{cases} \alpha\exp(n\theta), & n\theta\ge\log\alpha \\ \alpha\exp(n\theta)+\int_{n\theta}^{\log\alpha}\{\alpha-\exp(y)\}\exp(n\theta-y)/\{1-\exp(y)\}\,dy, & n\theta\le\log\alpha. \end{cases} \end{aligned}$$

Now for $n\theta<\log\alpha$,

$$\begin{aligned} &\int_{n\theta}^{\log\alpha}\frac{\alpha-\exp(y)}{1-\exp(y)}\exp\{-(y-n\theta)\}\,dy \\ &= \exp(n\theta)\int_{\exp(n\theta)}^{\alpha}\frac{\alpha-z}{z^2(1-z)}\,dz \\ &= \exp(n\theta)\int_{\exp(n\theta)}^{\alpha}[\alpha z^{-2}-(1-\alpha)\{z^{-1}+(1-z)^{-1}\}]\,dz \\ &= \exp(n\theta)\left[-\alpha z^{-1}+(1-\alpha)\{\log(1-z)-\log z\}\right]_{\exp(n\theta)}^{\alpha} \\ &= \exp(n\theta)[(1-\alpha)\log(\alpha^{-1}-1)-1+\alpha\exp(-n\theta)-(1-\alpha)\log\{\exp(-n\theta)-1\}]. \end{aligned}$$

Therefore

$$P_\theta\{\text{Reject } H_0\} = \begin{cases} \alpha\exp(n\theta), & n\theta\ge\log\alpha \\ g_\alpha(\exp(n\theta)), & n\theta\le\log\alpha \end{cases}$$

where

$$g_\alpha(z)=\alpha+z(1-\alpha)[\log(\alpha^{-1}-1)-\log(z^{-1}-1)-1],\quad 0<z<\alpha.$$

Since $\lim_{z\to 0} g_\alpha(z)=\alpha$, $\lim_{z\to\alpha} g_\alpha(z)=\alpha^2$, and $g_\alpha''(z)>0$ for $0<z<\alpha$, we conclude that $\sup_{H_0} P_\theta\{\text{Reject } H_0\}=\lim_{\theta\to-\infty} g_\alpha(\exp(n\theta))=\alpha$.

5.1.2. Cox Likelihood Ratio Statistic

The Cox log-likelihood ratio statistic is

$$\begin{aligned} S &= \sum_{i=1}^{n} \log \left[\frac{\exp\{-(X_i - \hat{\theta}_1)\} I(X_{(1)} \geq \hat{\theta}_1)}{\exp\{-(X_i - \hat{\theta}_0)\} I(X_{(1)} \geq \hat{\theta}_0)} \right] \\ &= \begin{cases} -\infty, & X_{(1)} < 0 \\ nX_{(1)}, & X_{(1)} \geq 0. \end{cases} \end{aligned}$$

It follows that the bootstrap test behaves the same as that based on the standard likelihood ratio. We therefore have the following theorem.

Theorem 5.1. *For testing $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$ for a sample from an Exp(θ, 1) distribution, the bootstrap tests based on the standard and Cox likelihood ratios have size α.*

5.2. $H_0 : \theta \geq 0$ *vs.* $H_1 : \theta < 0$

The MLEs under H_0 and H_1 are $\hat{\theta}_0 = \max(X_{(1)}, 0)$ and $\hat{\theta}_1 = \min(X_{(1)}, 0)$, respectively.

5.2.1. Standard Likelihood Ratio Statistic

Theorem 5.2. *The bootstrap test of $H_0 : \theta \geq 0$ vs. $H_1 : \theta < 0$ based on the standard likelihood ratio test is completely randomized.*

Proof. The standard log-likelihood ratio statistic is

$$\begin{aligned} S &= \sum_{i=1}^{n} \log \left[\frac{\exp\{-(X_i - \hat{\theta})\} I(X_{(1)} \geq \hat{\theta})}{\exp\{-(X_i - \hat{\theta}_0)\} I(X_{(1)} \geq \hat{\theta}_0)} \right] \\ &= \begin{cases} \infty, & X_{(1)} < 0 \\ 0, & X_{(1)} \geq 0. \end{cases} \end{aligned}$$

Since $\hat{\theta}_0 \geq 0$, the distribution of S^* is degenerate at 0. On the other hand, $S = 0$ w.p.1 under H_0. Therefore the bootstrap test based on S is completely randomized.

5.2.2. Cox Likelihood Ratio Statistic

Theorem 5.3. *The bootstrap test of $H_0 : \theta \geq 0$ vs. $H_1 : \theta < 0$ based on the Cox likelihood ratio test rejects H_0 w.p.1 for any $0 < \alpha < 1$.*

Proof. The Cox log-likelihood ratio statistic is

$$\begin{aligned} S &= \sum_{i=1}^{n} \log \left[\frac{\exp\{-(X_i - \hat{\theta}_1)\} I(X_{(1)} \geq \hat{\theta}_1)}{\exp\{-(X_i - \hat{\theta}_0)\} I(X_{(1)} \geq \hat{\theta}_0)} \right] \\ &= \begin{cases} \infty, & X_{(1)} < 0 \\ n(\hat{\theta}_1 - \hat{\theta}_0), & X_{(1)} \geq 0 \end{cases} \\ &= \begin{cases} \infty, & X_{(1)} < 0 \\ -nX_{(1)}, & X_{(1)} \geq 0. \end{cases} \end{aligned}$$

1. $X_{(1)} < 0$. Then $\hat{\theta}_0 = 0$, the bootstrap data have an $\text{Exp}(0,1)$ distribution, and the distribution of S^* is the negative of an $\text{Exp}(0,1)$ distribution. Since $S = \infty$, the test rejects H_0 w.p.1 for any $0 < \alpha < 1$.
2. $X_{(1)} \geq 0$. Then $\hat{\theta}_0 > 0$, and the bootstrap data have an $\text{Exp}(X_{(1)}, 1)$ distribution. The distribution of S^* is the negative of an $\text{Exp}(nX_{(1)}, 1)$ distribution, with support $(-\infty, -nX_{(1)})$. Since $S = -nX_{(1)}$, the test rejects H_0 w.p.1 for any $0 < \alpha < 1$.

6. Conclusion

The results show that the size of the bootstrap test of hypotheses is unpredictable. It depends on the problem as well as the choice of test statistic. For example, in the case of testing a normal mean with known variance, the test based on the sample mean or the Cox likelihood ratio is UMP for $0 < \alpha \leq 1/2$, but it is sub-optimal when it is based on the standard likelihood ratio. On the other hand, if $\alpha > 1/2$, the test often rejects H_0 w.p.1. The overall conclusion is that the size of the test is typically larger than its nominal level. This may explain the high power that the test is found to possess in simulation experiments.

Appendix

Proof of Lemma 2.1.

First note that

$$\phi(x) - \phi(x-\theta) \begin{cases} > 0, & \text{if } x < \theta/2 \\ < 0, & \text{if } x > \theta/2. \end{cases} \tag{6.1}$$

Let $f(\theta)$ denote the function (2.2). We consider two cases.

1. $\alpha \geq 1/2$. Since $z_\alpha^+ = 0$, we have $f(\theta) = 1 - (1-\alpha)\int_0^\infty \phi(x-\theta)/\Phi(x)\,dx$ and

$$\begin{aligned} \frac{f(0) - f(\theta)}{1-\alpha} &= \int_0^\infty \frac{\phi(x-\theta) - \phi(x)}{\Phi(x)}\,dx \\ &= \int_0^{\theta/2} \frac{\phi(x-\theta) - \phi(x)}{\Phi(x)}\,dx + \int_{\theta/2}^\infty \frac{\phi(x-\theta) - \phi(x)}{\Phi(x)}\,dx \\ &\geq 2\int_0^{\theta/2} \{\phi(x-\theta) - \phi(x)\}\,dx + \int_{\theta/2}^\infty \{\phi(x-\theta) - \phi(x)\}\,dx \\ &= 2\{\Phi(-\theta/2) - \Phi(-\theta) - \Phi(\theta/2) + 1/2\} - \Phi(-\theta/2) + \Phi(\theta/2) \\ &= 2\{\Phi(\theta) - \Phi(\theta/2)\} \\ &\geq 0 \end{aligned}$$

where we use (6.1) in the first inequality. Hence

$$\begin{aligned} f(\theta) &\leq f(0) \\ &= 1 - (1-\alpha)\int_0^\infty \phi(x)/\Phi(x)\,dx \\ &= 1 + (1-\alpha)\log(1/2). \end{aligned}$$

2. $\alpha < 1/2$. Write $f(\theta) = J_1(\theta) + J_2(\theta)$, where

$$\begin{aligned} J_1(\theta) &= \alpha(1-\alpha)^{-1}\Phi(\theta - z_\alpha) + \Phi(-z_\alpha - \theta) \\ J_2(\theta) &= E\left\{ \frac{(1-2\alpha)\Phi(Z+\theta) - (1-\alpha)^2}{(1-\alpha)\Phi(Z+\theta)} I(Z+\theta \geq z_\alpha) \right\}. \end{aligned}$$

Since

$$\begin{aligned} \partial J_1(\theta)/\partial\theta &= \alpha(1-\alpha)^{-1}\phi(\theta - z_\alpha) - \phi(\theta + z_\alpha) \\ &= \phi(\theta + z_\alpha)\{\alpha(1-\alpha)^{-1}\exp(2\theta z_\alpha) - 1\}, \end{aligned}$$

$J_1(\theta)$ is decreasing-increasing, with minimum at $\theta_0 = (2z_\alpha)^{-1}\log\{(1-\alpha)/\alpha\} > 0$. Further, $J_1(0) = \lim_{\theta\to\infty} J_1(\theta) = \alpha(1-\alpha)^{-1}$. Therefore $J_1(\theta) < J_1(0)$ for $\theta > 0$.

To obtain a similar result for J_2, let

$$g(x) = \{(1-2\alpha)\Phi(x) - (1-\alpha)^2\}/\{(1-\alpha)\Phi(x)\}$$

which is increasing in x with $g(z_\alpha) = -\alpha/(1-\alpha)$ and $g(x) \to -\alpha^2/(1-\alpha)$ as $x \to \infty$. Hence $g(x) < 0$ for $x \geq z_\alpha$.

(a) $0 < \theta \leq 2z_\alpha$. Since $\phi(x) \leq \phi(x-\theta)$ for $x \geq z_\alpha$,

$$J_2(0) - J_2(\theta) = \int_{z_\alpha}^{\infty} g(x)\phi(x)\,dx - \int_{z_\alpha}^{\infty} g(x)\phi(x-\theta)\,dx \geq 0.$$

(b) $\theta > 2z_\alpha$. From (6.1),

$$\begin{aligned} & J_2(0) - J_2(\theta) \\ &= \int_{z_\alpha}^{\infty} g(x)\phi(x)\,dx - \int_{z_\alpha}^{\infty} g(x)\phi(x-\theta)\,dx \\ &= \int_{z_\alpha}^{\theta/2} g(x)[\phi(x) - \phi(x-\theta)]\,dx + \int_{\theta/2}^{\infty} g(x)[\phi(x) - \phi(x-\theta)]\,dx \\ &> -\frac{\alpha}{1-\alpha}\int_{z_\alpha}^{\theta/2}[\phi(x) - \phi(x-\theta)]\,dx - \frac{\alpha^2}{1-\alpha}\int_{\theta/2}^{\infty}[\phi(x) - \phi(x-\theta)]\,dx \\ &= \alpha(1-\alpha)^{-1}[-\{\Phi(\theta/2) - (1-\alpha) - \Phi(-\theta/2) + \Phi(z_\alpha - \theta)\} \\ &\quad + \alpha\{\Phi(\theta/2) - \Phi(-\theta/2)\}] \\ &= \alpha(1-\alpha)^{-1}\{K_1(\theta) + K_2(\theta)\} \end{aligned}$$

where

$$\begin{aligned} K_1(\theta) &= \Phi(-\theta/2) - \Phi(z_\alpha - \theta) \\ K_2(\theta) &= 1 - \alpha - \Phi(\theta/2) + \alpha\Phi(\theta/2) - \alpha\Phi(-\theta/2). \end{aligned}$$

Now $K_1(\theta) > 0$ for $\theta > 2z_\alpha$, $K_2'(\theta) = (\alpha - 1/2)\phi(\theta/2) < 0$, and $K_2(\theta) \to 0$ as $\theta \to \infty$. Thus $J_2(0) - J_2(\theta) \geq 0$.

Therefore $f(\theta) \leq f(0) = 2\Phi(-z_\alpha) - (1-\alpha)\int_{z_\alpha}^{\infty}\phi(x)/\Phi(x)\,dx = 2\alpha + (1-\alpha)\log(1-\alpha)$.

It remains to show that $f(0) > \alpha$ for all $0 < \alpha < 1$. Let $h(\alpha) = f(0) - \alpha$. Then

$$h(\alpha) = \begin{cases} \alpha + (1-\alpha)\log(1-\alpha), & \text{if } 0 < \alpha \leq 1/2 \\ 1 - \alpha - (1-\alpha)\log 2, & \text{if } 1/2 \leq \alpha < 1 \end{cases}$$

and h is continuous with $h(0) = h(1) = 0$, $h(1/2) = (1 - \log 2)/2 > 0$, and

$$h'(\alpha) = \begin{cases} -\log(1-\alpha) > 0, & \text{if } 0 < \alpha \leq 1/2 \\ -1 + \log 2 < 0, & \text{if } 1/2 \leq \alpha < 1. \end{cases}$$

Therefore $h(\alpha) > 0$ for $0 < \alpha < 1$, concluding the proof.

References

[1] Ané, C., Burleigh, J. G., McMahon, M. M. and Sanderson, M. J. (2005). Covarion structure in plastid genome evolution: A new statistical test. *Molecular Biology and Evolution*, **22**, 914–924.

[2] Carstens B. C., Bankhead III, A. and Joyce P. and Sullivan J. (2005). Testing population genetic structure using parametric bootstrapping and MIGRATE-N. *Genetica*, **124**, 71–75.

[3] Chung P. J. and Moura J. (2004). A GLRT and bootstrap approach to detection in magnetic resonance force microscopy. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[4] Cox, D. R. (1961). Tests of separate families of hypotheses. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**. Berkeley, CA, University of California Press. 105–123.

[5] Dalla V. and Hidalgo J. (2004). A parametric bootstrap test for cycles. *Journal of Econometrics*, **129**, 219–261.

[6] Dameusa A., Richter, F. G. C., Brorsen, B. W. and Sukhdial K. P. (2002). AIDS versus the Rotterdam Demand System: a Cox Test with Parametric Bootstrap. *Journal of Agricultural and Resource Economics*, **27**, 335–347.

[7] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1–26.

[8] Hunsberger S., Albert, P. S., Follmann, D. A. and Suh E. (2002). Parametric and semiparametric approaches to testing for seasonal trend in serial count data. *Biostatistics*, **3**, 289–298.

[9] Kaitibie, S., Nganje, W. E., Wade Brorsen, B. and Epplin, F. M. (2007). A Cox Parametric Bootstrap Test of the von Liebig Hypotheses. *Canadian Journal of Agricultural Economics*, **55**, 15–25.

[10] Lehmann E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York.

[11] Solow A. R., Costello, C. J. and Ward, M. (2003). Testing the Power Law Model for Discrete Size Data. *The American Naturalist*, **162**, 685–689.

[12] Walters S. J. and Campbell, M. J. (2004). The use of bootstrap methods for analysing health-related quality of life outcomes (particularly the SF-36). *Health Qual Life Outcomes*, **2**, 70.

[13] Wilson, E. B. and Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences*, **17**, 684–688.

IMS Collections
Vol. 0 (2009) 117–146
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000004

Nonparametric Estimation for Lévy Models Based on Discrete-Sampling

José E. Figueroa-López[1,*]

Purdue University

Abstract: A Lévy model combines a Brownian motion with drift and a pure-jump homogeneous process such as a compound Poisson process. The estimation of the Lévy density, the infinite-dimensional parameter controlling the jump dynamics of the process, is studied under a discrete-sampling scheme. In that case, the jumps are latent variables whose statistical properties can in principle be assessed when the frequency of observations increase to infinity. We propose nonparametric estimators for the Lévy density following *Grenander's method of sieves*. The associated problem of selecting a suitable approximating sieve is subsequently investigated using regular piece-wise polynomials as sieves and assuming standard smoothness conditions on the Lévy density. By sampling the process at a high enough frequency relative to the time horizon T, we show that it is feasible to choose the dimension of the sieve so that the rate of convergence of the *risk of estimation off the origin* is the best possible from a minimax point of view, and even if the estimation were based on the whole sample path of the process. The sampling frequency necessary to attain the optimal minimax rate is explicitly identified. The proposed method is illustrated by simulation experiments in the case of variance Gamma processes.

Contents

1 Introduction 118
1.1 Motivation and Some Background 118
1.2 The Statistical Problems and Methodology 120
1.3 An Overview of the Estimators and the Results 121
1.4 Outline 123
2 The Estimators and Central Limit Theorems 124
3 The Model Selection Problem 126
3.1 Analysis of the Variance Term 127
3.2 The Approximation Error for Besov Type Smooth Functions 129
3.3 Rate of Convergence for Smooth Functions Via Splines 130
3.4 About the Critical Mesh 131
4 Minimax Risk of Estimation for Smooth Lévy Densities 132
5 A Data-Driven Selection Method and Adaptability 133
6 An Example: Estimation of Variance Gamma Processes. 134
6.1 The Model 134

[1]Department of Statistics, Purdue University, West Lafayette, IN 47906, email: figueroa@stat.purdue.edu

*It is a pleasure to thank the Department of Statistics and Applied Probability at the University of California at Santa Barbara for their support while part of this research was conducted. The author is indebted to two anonymous referees for their careful reading of the paper and for their many suggestions that improve it considerably.

AMS 2000 subject classifications: Primary 62G05; secondary 62G20

Keywords and phrases: Lévy processes, nonparametric estimation, model selection

6.2 The Simulation Procedure 135
6.3 The Numerical Results 136
Appendix A: Technical Proofs 137
Appendix B: Figures 142
References 145

1. Introduction

1.1. Motivation and Some Background

In the last decade, Lévy processes have received a great deal of attention, fueled by numerous applications in the area of mathematical finance, to the extend that Lévy processes have become a fundamental building block in the modeling of asset prices with jumps (see e.g. [11] and [30]). The simplest of these models postulates that the price of a commodity (say a stock) at time t is determined by

$$S_t := S_0 e^{X_t}, \tag{1.1}$$

where $X := \{X_t\}_{t\geq 0}$ is a Lévy process. Even this simple extension of the classical Black-Scholes model, in which X is simply a Brownian motion with drift, is able to account for some fundamental empirical features commonly observed in time series of asset returns such as heavy tails, high-kurtosis, and asymmetry. More recently, other Lévy based models have been proposed to account for more stylized features of stock prices. These models include exponential time-changed Lévy processes (cf. [7]-[9]), and stochastic differential equations driven by multivariate Lévy processes (cf. [1], [31]). Lévy processes, as models capturing some of the most important features of returns and as "first-order approximations" to other more accurate models, should be considered first in developing and testing a successful statistical methodology. However, even in such parsimonious models, there are several issues in performing statistical inference by standard likelihood-based methods.

A Lévy process is the "discontinuous sibling" of a Brownian motion. Concretely, $X = \{X_t\}_{t\geq 0}$ is a Lévy process if X has independent and stationary increments, its paths are right-continuous with left limits, and it has no fixed jump times. The later condition means that, for any $t > 0$,

$$\mathbb{P}\left[\Delta X_t \neq 0\right] = 0,$$

where $\Delta X_t := X(t) - \lim_{s\nearrow t} X_s$ is the magnitude of the "jump" of X at time t. It can be proved that the only Lévy process with continuous paths is essentially the Brownian motion $W := \{W_t\}_{t\geq 0}$ up to a drift term bt (hence, the well-known Gaussian distribution of the increments of W is a byproduct of the stationarity and independence of its increments). The only deterministic Lévy process is of the form $X_t := bt$, for a constant b. Another distinguished type of Lévy process is a compound Poisson process defined as

$$Y_t := \sum_{i=1}^{N_t} \xi_i, \tag{1.2}$$

where N is a homogeneous Poisson process and the random variables ξ_i, $i \geq 1$, are mutually independent from one another, independent from N, and with common

distribution ρ. The process N dictates the jump times, which can occur "homogeneously" across time with an (average) intensity of λ jumps per unit time, while the sequence $\{\xi_i\}_{i\geq 1}$ determines the sizes of the jumps.

It turns out that the most general Lévy process is the superposition of a Brownian motion with drift, $\sigma W_t + bt$, a compound Poisson process, and the limit process resulting from making the jump intensity of a compensated compound Poisson process, $Y_t - \mathbb{E}\, Y_t$, to go to infinity while simultaneously allowing jumps of smaller sizes. The latter limiting process is governed by a measure ν such that the intensity of jumps is $\lambda_\varepsilon := \nu(\varepsilon \leq |x| < 1)$, the common distribution of the jump sizes is $\rho_\varepsilon(dx) := \mathbf{1}_{\{|x|\geq\varepsilon\}}\nu(dx)/\lambda_\varepsilon$, and the limit is when $\varepsilon \searrow 0$. For such a limit to converge to a "steady" process it must hold that

$$\int_{\{|x|<1\}} x^2 \nu(dx) < \infty.$$

The previous fundamental decomposition of a Lévy process is called the Lévy-Itô decomposition (see Section 19 in [29] for the details).

In summary, Lévy processes are determined by three "parameters": a non-negative real σ^2, a real b, and a measure ν on $\mathbb{R}\backslash\{0\}$ such that $\int (x^2 \wedge 1)\nu(dx) < \infty$. The measure ν controls the jump dynamics of the process X in that for any $A \in \mathcal{B}(\mathbb{R})$ whose indicator χ_A vanishes in a neighborhood of the origin,

$$\nu(A) = \frac{1}{t}\,\mathbb{E}\left[\sum_{s\leq t} \chi_A\left(\Delta X(s)\right)\right],$$

for any $t > 0$ (see Section 19 of [29]). Thus, $\nu(A)$ gives the average number of jumps (per unit time) whose magnitudes fall in the set A. A common assumption in Lévy-based financial models is that ν is determined by a function $s : \mathbb{R}\backslash\{0\} \to [0, \infty)$, called the *Lévy density*, as follows

$$\nu(A) = \int_A s(x)dx, \quad \forall A \in \mathcal{B}(\mathbb{R}\backslash\{0\}).$$

Intuitively, the value of s at x_0 provides information on the frequency of jumps with sizes "close" to x_0. In the case of the compound Poisson process (1.2), the Lévy measure is $\nu(dx) = \lambda\rho(dx)$.

By allowing a general Lévy process X in (1.1), instead of just a Brownian motion with drift as in the Black-Scholes model, one can incorporate two very appealing features: sudden changes in the price dynamics and some freedom in the distribution for the log return $\log\{S_t/S_s\} = X_t - X_s$. The possible distributions belong to the so-called class of *infinitely-divisible distributions*, a very rich class which include most known parametric families of distributions. We recall that an infinitely divisible distribution μ is characterized by the so-called *Lévy-Khinchin representation* of its characteristic function.

There are two key properties of a Lévy process that are exploited in this work. The first property relates ν with the short-term moments of X_t. Concretely, if φ is *ν-continuous, bounded, and vanishing in a neighborhood of the origin*, then

$$\lim_{\Delta\to 0} \frac{1}{\Delta}\,\mathbb{E}\,\varphi(X_\Delta) = \int \varphi(x)\nu(dx) = \int \varphi(x)s(x)dx. \tag{1.3}$$

The limiting relation (1.3) is straightforward when X is a compound Poisson process. A proof of (1.3) for a general Lévy process can be found in [29] (see his

Corollary 8.9). Let us remark that (1.3) is also valid for certain unbounded functions φ, which does not necessarily vanish in a neighborhood of the origin, but rather converge to 0 at a proper rate (see [15] for more details). The second key property is related to the decomposition of X into two *independent* processes: one accounting for the "small" jumps and a compound Poisson process collecting the "big" jumps. Concretely, let

$$\widetilde{X}^{\varepsilon}_t := \sum_{s\leq t} \Delta X_s \mathbf{1}_{\{|\Delta X_s|>\varepsilon\}},$$

be the piece-wise constant process associated with those jumps of X with sizes larger than ε. Then, $\widetilde{X}^{\varepsilon}$ is a compound Poisson process independent of $X - \widetilde{X}^{\varepsilon}$.

1.2. The Statistical Problems and Methodology

We are interested in estimating the Lévy density s on a window of estimation $D := [a,b] \subset \mathbb{R}\backslash\{0\}$, based on discrete observations of the process on a finite interval $[0,T]$. *We remark that the domain D is "separated" from the origin; that is to say, the estimation window D lies outside of a neighborhood of the origin.* If the whole path of the process were available (and hence, the jumps of the process would be available), the problem would be identical to the estimation of the intensity of a non-homogeneous Poisson process on a fixed time interval, say $[0,1]$, based on T independent copies of the process. However, under discrete-sampling, the times and sizes of jumps are latent (unobservable) variables, whose statistical properties can be assessed when the frequency of observations increase to infinity at a certain speed relative to the time horizon. Hence, we will aim at determining the performance of our estimation method as both frequency and time horizon increase.

We adopt the so-called *method of sieves* originally proposed by [18] and implemented by Birgé, Massart, and others (see e.g. [3] & [5]) in several classical nonparametric problems such as density estimation and regression. This approach consists of the following general steps. First, choose a family of finite-dimensional *linear models* of functions, called *sieves*, with good approximation properties. Common sieves are splines, trigonometric polynomials, or wavelets. Second, specify a distance between functions relative to which the best approximation to s, in a given linear model, is going to be defined and characterized. Finally, devise an estimator, called the *projection estimator*, for the best approximation of s in the given linear model. It is important to point out that in principle there is no guarantee that the projection estimator will be nonnegative. In practice, one barely faces this problem when working with a large sample size, which is exactly the situation when nonparametric methods are recommended.

A linear model has the generic form

$$\mathcal{S} := \{\beta_1\varphi_1 + \cdots + \beta_d\varphi_d : \beta_1,\ldots,\beta_d \in \mathbb{R}\}, \tag{1.4}$$

where $\varphi_1,\ldots,\varphi_d$ are given functions, typically taken to be orthonormal with respect to the inner product $\langle p,q\rangle := \int_D p(x)q(x)dx$. In the sequel, $\|\cdot\|$ stands for the associated norm $\langle\cdot,\cdot\rangle^{1/2}$ on $\mathbb{L}^2(D,dx)$. Relative to the distance induced by $\|\cdot\|$, the element of $\mathcal{S}$ closest to s, i.e. the *orthogonal projection* of s on $\mathcal{S}$, is

$$s^{\perp}(x) := \sum_{j=1}^{d} \nu(\varphi_j)\varphi_j(x), \tag{1.5}$$

where $\nu(\varphi_j) := \langle \varphi_j, s\rangle = \int \varphi_j(x)s(x)dx$. Then, the method of sieves boils down to estimate the *orthogonal projection* (1.5) on an "adequate" sieve $\mathcal{S}$. The core problem in this paper is to determine what a good sieve is. A very large linear model $\mathcal{S}$ will allow to attain a close approximation to s, but will entail necessarily a high estimation variance as the result of the large number of coefficients β_i to be estimated. Therefore, an essential task, called *model selection*, consists of selecting a linear model $\mathcal{S}$ accomplishing a good tradeoff between the error of approximation (or mis-specification error) and the standard error of the estimation. Concretely, one wishes to minimize the risk of the estimator $\hat{s}$, which in turn can be decomposed into two antagonist terms as follows:

$$\mathbb{E}\,\|s-\hat{s}\|^2 = \|s-s^{\perp}\|^2 + \mathbb{E}\,\|s^{\perp}-\hat{s}\|^2. \tag{1.6}$$

The first term, called the *bias term*, accounts for the error of the approximation, while the second, called the *variance term*, accounts for the standard error of the estimation.

1.3. An Overview of the Estimators and the Results

We assume that the Lévy process $\{X_t\}_{t\geq 0}$ is being sampled over a time horizon $[0,T]$ at discrete times $0<t_1^n<\cdots<t_n^n=T$. In the sequel, $t_0^n:=0$, $\pi^n:=\{t_k^n\}_{k=0}^n$, and $\bar{\pi}^n := \max_k\{t_k^n - t_{k-1}^n\}$, the so-called mesh of the partition. We shall sometimes drop the superscript n in π^n and t_i^n. The following statistics are the main building blocks of our estimators:

$$\hat{\beta}^{\pi^n}(\varphi) := \frac{1}{t_n}\sum_{k=1}^{n}\varphi\left(X_{t_k^n}-X_{t_{k-1}^n}\right). \tag{1.7}$$

In the case of a quadratic function $\varphi(x)=x^2$, $\sum_{k=1}^n \varphi\left(X_{t_k^n}-X_{t_{k-1}^n}\right)$ is called the realized quadratic variation (or variance) of the process. Thus, the statistics (1.7) can be interpreted as the average realized $\varphi-$variation of the process per unit time based on the observations $X_{t_1^n},\ldots,X_{t_n^n}$.

To explain the motivation behind the estimator in (1.7), let us assume for now that sampling observations are equally-spaced in time so that $\Delta_n := t_i^n - t_{i-1}^n = T/n$ for all i, and hence,

$$\begin{aligned}\mathbb{E}\left\{\hat{\beta}^{\pi^n}(\varphi)\right\} &= \frac{1}{\Delta_n}\mathbb{E}\,\varphi\left(X_{\Delta_n}\right),\\ \mathrm{Var}\left\{\hat{\beta}^{\pi^n}(\varphi)\right\} &= \frac{1}{T}\left(\frac{1}{\Delta_n}\mathbb{E}\,\varphi^2\left(X_{\Delta_n}\right)\right) - \frac{1}{n}\left(\frac{1}{\Delta_n}\mathbb{E}\,\varphi\left(X_{\Delta_n}\right)\right)^2.\end{aligned}$$

In view of (1.3), it is now evident that

$$\lim_{n\to\infty}\mathbb{E}\left\{\hat{\beta}^{\pi^n}(\varphi)\right\} = \int\varphi(x)s(x)dx, \quad \text{and} \quad \lim_{T\to\infty}\sup_n \mathrm{Var}\left(\hat{\beta}^{\pi^n}(\varphi)\right)=0, \tag{1.8}$$

if φ is *ν-continuous, bounded, and vanishing in a neighborhood of the origin.* In statistical terms, (1.8) means that the statistic $\hat{\beta}^{\pi^n}(\varphi)$ is an asymptotically unbiased estimator of $\int\varphi(x)s(x)dx$ with associated risk vanishing uniformly when time horizon T increases. The previous argument leads us to propose

$$\hat{s}^{\pi^n}(x) := \sum_{j=1}^{d}\hat{\beta}^{\pi^n}(\varphi_j)\varphi_j(x), \tag{1.9}$$

as a natural estimator for the orthogonal projection $s^{\perp}$ defined in (1.5). In view of (1.8), $\hat{s}^{\pi^n}$ is a "consistent" estimator for $s^{\perp}$, in the integrated mean-square sense, as both the time horizon $T = t_n^n$ and the sampling frequency n/t_n^n go to ∞. The general sampling case will be considered in Section 2 as well as other statistical properties. It is worth pointing out that $\hat{s}^{\pi^n}$ is independent of the specific orthonormal basis of $\mathcal{S}$ as it can be proved that $\hat{s}^{\pi^n}$ is the unique solution of the minimization problem

$$\min_{f\in\mathcal{S}} \gamma_D^{\pi}(f),$$

where $\gamma_D^{\pi^n} : L^2(D, dx) \to \mathbb{R}$ is given by

$$\gamma_D^{\pi^n}(f) \equiv -\frac{2}{t_n^n}\sum_{k=1}^{n} f(X_{t_k^n} - X_{t_{k-1}^n}) + \int f^2(x)dx. \tag{1.10}$$

In the literature of model selection (see e.g. [4] and [25]), $\gamma_D^{\pi^n}$ is called the *contrast function*.

Finding the best sieve $\mathcal{S}$ to estimate s, even if we stick with using the class of projection estimators in (1.9), is impossible because s is unknown. However, it is possible to select a reasonably good model under certain qualitative assumptions on the parameter s, typically expressed by requiring s to be a member of a certain class Θ of smooth functions. Concretely, suppose we are interested in selecting a good model out of a family of linear models $\{\mathcal{S}_m\}_{m\in\mathcal{M}}$ (here, $\mathcal{M}$ is a suitable set of labels). Let $m^* := m^*(\pi)$ be the *optimal minimax element of* $\{\hat{s}_m\}_{m\in\mathcal{M}}$ *on* Θ, defined as

$$m^* := \text{arginf}_{m\in\mathcal{M}}\{\sup_{s\in\Theta} \mathbb{E}\,\|s - \hat{s}_m\|^2\}.$$

By requiring certain conditions on Θ and by choosing a suitable family of sieves $\{\mathcal{S}_m\}_{m\in\mathcal{M}}$, we can ensure that

$$\mathbb{E}\,\|s - \hat{s}_{m^*(\pi)}\|^2 \to 0 \tag{1.11}$$

as the mesh of the partition $\pi = \{t_k\}_{k\geq 1}$ vanishes and the time horizon $T := t_n$ goes to infinity. Our goal will be to select a linear model $\hat{m}(\pi) \in \mathcal{M}$ so that the projection estimator on this model, $\hat{s}_{\hat{m}(\pi)}$, "attains" the minimax rate of convergence in (1.11), in the sense that

$$\limsup \frac{\mathbb{E}\,\|s - \hat{s}_{\hat{m}(\pi)}\|^2}{\mathbb{E}\,\|s - \hat{s}_{m^*(\pi)}\|^2} < \infty, \tag{1.12}$$

where the limit is taken as $\bar{\pi} \to 0$ and $T \to \infty$. In order to be able to determine in a "simple" way the rate of convergence of $\hat{s}_{\hat{m}(\pi)}$, we shall control the sampling frequency, measured by $\bar{\pi}$, in function of the time horizon T. In the case of a finite-jump activity process with a jump intensity of λ jumps per unit time, we can expect that it suffices to sample at a faster rate than $1/T$ (that is, $\bar{\pi}\,T \to 0$). It is intuitive that in general the sampling frequency will depends on how close the window of estimation D is to the origin (see Section 3.4). The limit result (1.12) and the rate of convergence of projection estimators for a certain class of smooth Lévy densities are addressed in Section 3.

In this paper, we will show that the rate of convergence that can be attained using projection estimation on sieves is actually the best possible among all feasible estimators, given the information available on s (namely, that s belongs to a certain

class Θ of smooth functions), and even if the estimators were based on continuous-time sampling of the process. Concretely, define $\hat{s}_T^*$ be the minimax estimator,

$$\hat{s}_T^* := \arg\inf_{\hat{s}} \sup_{s\in\Theta} \mathbb{E}\,\|s-\hat{s}\|^2 < \infty,$$

where the infimum is over *all the estimators* $\hat{s}$ of s based on $\{X(t)\}_{0\le t\le T}$. Then, by sampling at a high enough frequency (relative to T), we can accomplish that

$$\limsup_{T\to\infty} \frac{\mathbb{E}\,\|s-\hat{s}_{\hat{m}}\|^2}{\mathbb{E}\,\|s-\hat{s}_T^*\|^2} < \infty.$$

The rate of convergence of the minimax estimator will be provided in Section 4.

Let us finish by pointing out that the model selection problem was already analyzed in Figueroa-López & Houdré (2006) using the statistics

$$\hat{\beta}^c(\varphi) := \frac{1}{t_n}\sum_{t\le t_n} \varphi\left(\Delta X_t\right), \tag{1.13}$$

which intrinsically required continuous-time sampling of the process to determine the jumps ΔX_t. In the cited paper, the statistics (1.7) were proposed as good proxies of (1.13). Indeed, convergence in distribution is not hard to check, but moreover, recently [20] prove that (1.7) converges in probability to (1.13) when $n\to\infty$ (for fixed T). To the best of our knowledge, an analysis of the model selection problem for Lévy densities, under discrete sampling schemes, has not been considered before the present work.

1.4. Outline

The paper is structured as follows. In Section 2, we introduce the estimators proposed in this paper and study some basic statistical properties. In particular, we prove a CLT for the estimator $\hat{\beta}^\pi(\varphi)$ of (1.7) centered at the inner product $\beta(\varphi)=\int\varphi(x)s(x)dx$. In Section 3, we describe how to control the risk of the projection estimators by imposing three conditions. First, the time horizon T should be large enough (compared to the complexity of the sieves). Second, the time span between consecutive observations should be small enough compared to the time horizon. Finally, the sieves should have good approximating properties in general classes of smooth functions. We show that by ensuring the three previous conditions and by suitably choosing the dimension of the sieve (in terms of the presumed smoothness of the function s), the rate of convergence of the risk is of order $O(T^{-2\alpha/(2\alpha+1)})$ as $T\to\infty$ provided that the parameter s has "degree of smoothness" α.

In Section 4, the minimax risk of estimation, defined by

$$\inf_{\hat{s}_T}\sup_{s\in\Theta} \mathbb{E}_s\|s-\hat{s}_T\|^2,$$

is studied. Here, the infimum is over all estimators $\hat{s}_T$ which can be computed from the *whole sample paths* of X on the interval $[0,T]$ and the supremum is over all Lévy densities in a class Θ of functions that are smooth in $D=[a,b]$. We found that the minimax risk converges at an order of $O(T^{-2\alpha/(2\alpha+1)})$, where α is a parameter that measures the smoothness of the functions on Θ. For instance, if s has d continuous derivatives in D, then $\alpha\ge d$. The rate of convergence of the estimation is faster when α increases. Sections 3 and 4 justify the claim of the abstract: "...we show that

it is feasible to choose the dimension of the sieve so that the rate of convergence of the *risk of estimation off the origin* is the best possible from a minimax point of view, and even if the estimation were based on the whole sample path of the process".

In Section 5, we propose a data-driven selection method for the sieve. Instead of deciding the dimension of the sieve from a presumed degree of smoothness of s (as it was suggested in Section 3), we propose to choose the sieve that minimizes an unbiased estimator of the risk of the projection estimator corresponding to that sieve. Since the proposed estimator of the risk will require the knowledge of all jumps of X up to time T, we replace it by a natural discrete-based proxy, where the jumps ΔX_t are replaced by the increments $X_{t_k} - X_{t_{k-1}}$. Section 6 illustrates the statistical methods using simulation experiments in the case of a variance gamma Lévy model. We finish with an Appendix where some technical proofs are given.

2. The Estimators and Central Limit Theorems

In this section, our goal is to survey some statistical properties of the estimators (1.7) and (1.9). We already mentioned a few of these in the case of *regular sampling*[1] and of bounded $\nu-$continuous *test functions* φ vanishing in a neighborhood of the origin. In the framework of this paper, this kind of test functions indeed suffices to recover and estimate the Lévy density off the origin. Our first result is a simple application of the Central Limit Theorem (CLT) for independent random variables (cf. [20] [Theorem 3.2] for the case of regular sampling). In the following results, Z stands for a standard Normal random variable.

Proposition 2.1. *Let φ be ν-continuous, bounded, and such that $\varphi(x) = o(|x|)$, as $x \to 0$. Then,*

$$\sqrt{t_n}\left(\hat{\beta}^\pi(\varphi) - \mathbb{E}\,\hat{\beta}^\pi(\varphi)\right) \xrightarrow{\mathfrak{D}} \nu(\varphi^2)^{\frac{1}{2}}\, Z, \tag{2.1}$$

as $t_n \to \infty$ and $\bar{\pi} \to 0$.

Proof. Let $\Gamma_t(\varphi) := \mathbb{E}\,\varphi^2(X_t) - \{\mathbb{E}\,\varphi(X_t)\}^2$ and $\Delta_k := t_k - t_{k-1}$. We can write

$$\sqrt{t_n}\left(\hat{\beta}^\pi(\varphi) - \mathbb{E}\,\hat{\beta}^\pi(\varphi)\right) = \sum_{k=1}^n \xi_k^\pi,$$

where $\xi_k^\pi = \frac{1}{\sqrt{t_n}}\left\{\varphi\left(X_{t_k} - X_{t_{k-1}}\right) - \mathbb{E}\,\varphi\left(X_{t_k - t_{k-1}}\right)\right\}$. Under the assumption of this Proposition, it turns out that $\lim_{t\to 0}\frac{1}{t}\Gamma_t(\varphi) = \nu(\varphi^2)$ (see Lemma 5.5 in Jacod (2007)), and thus,

$$\bar{\sigma}^2_{n,\pi} := \mathrm{Var}\sum_{k=1}^n \xi_k^\pi = \frac{1}{t_n}\sum_{k=1}^n \Gamma_{\Delta_k}(\varphi) \longrightarrow \nu(\varphi^2),$$

as the mesh $\bar{\pi} := \max_k\{t_k - t_{k-1}\} \to 0$. Due to the boundedness of φ, we have that, for $\bar{\pi}$ small enough,

$$\frac{|\xi_k^\pi|}{\bar{\sigma}_{n,\pi}} \le C\frac{1}{\sqrt{t_n}} \to 0,$$

as $t_n \to \infty$. Then, (2.1) follows from the Central Limit Theorem for independent random variables (see e.g. the Corollary following Theorem 7.1.2 in [10]).

[1] Sampling equally spaced in time.

In order to provide an explicit centering in (2.1), we need to estimate the rate of convergence of the bias $\mathbb{E}\,\hat{\beta}^{\pi}(\varphi) - \nu(\varphi)$. Since

$$\mathbb{E}\,\hat{\beta}^{\pi}(\varphi) - \nu(\varphi) = \frac{1}{t_n}\sum_{k=1}^{n}\Delta_k\left\{\frac{1}{\Delta_k}\,\mathbb{E}\,\varphi(X_{\Delta_k}) - \nu(\varphi)\right\},$$

the problem is equivalent to analyzing the rate of convergence in (1.3). To achieve this goal, we need to impose some regularity on either the Lévy process or the moment functions φ. Following the second approach, [15] shed light on this problem for functions $\varphi \in C_b^2(\mathbb{R})$; namely, twice-continuously differentiable functions φ such that $\limsup_{|x|\to\infty}|\varphi^{(i)}(x)| < \infty$, for $i = 0, 1, 2$. Below, $*$ denotes the *convolution operator* $\nu_1 * \nu_2(\varphi) := \iint \varphi(x_1 + x_2)\nu(dx_1)\nu(dx_2)$, and L denotes the *infinitesimal generator* of the process X (see e.g. Sato (1999)), which is known to be given by

$$(L\varphi)(x) := \frac{\sigma^2}{2}\varphi''(x) + b\varphi'(x) + \int\left(\varphi(y+x) - \varphi(x) - y\varphi'(x)\mathbf{1}_{\{|y|\le 1\}}\right)\nu(dy); \tag{2.2}$$

see Theorem 31.5 in [29] and also in Proposition 2.3 in [15]. The following result can be found in [15] (see Proposition 3.1), where a proof is provided for a certain class of unbounded functions φ:

Lemma 2.2. *If $\varphi \in C_b^2(\mathbb{R})$ vanishes in a neighborhood of the origin, then*

$$\lim_{t\to 0}\frac{1}{t}\left\{\frac{1}{t}\;\mathbb{E}\,\varphi(X_t) - \nu(\varphi)\right\} = \nu_\varepsilon(L\varphi) - \frac{1}{2}\,\nu_\varepsilon * \nu_\varepsilon(\varphi), \tag{2.3}$$

where $\nu_\varepsilon(dx) := \mathbf{1}_{\{|x|>\varepsilon\}}\nu(dx)$.

The following is an easy consequence of the previous two results.

Theorem 2.3. *Under the assumptions of Lemma 2.2,*

$$\sqrt{t_n}\left(\hat{\beta}^{\pi}(\varphi) - \nu(\varphi)\right) \xrightarrow{\mathfrak{D}} \nu(\varphi^2)^{\frac{1}{2}}\,Z \tag{2.4}$$

as $t_n \to \infty$ and $\bar{\pi} \to 0$ so that $\bar{\pi}\sqrt{t_n} \to 0$.

Proof. It suffices to prove that

$$D_n := \sqrt{t_n}\left(\;\mathbb{E}\;\hat{\beta}^{\pi}(\varphi) - \int\varphi(x)\nu(dx)\right) \longrightarrow 0.$$

Writing $\Delta_k = t_k - t_{k-1}$ and using (2.3), for $\bar{\pi} = \max_k \Delta_k$ small enough, there exists a constant C such that

$$|D_n| \le \frac{1}{\sqrt{t_n}}\sum_{k=1}^{n}\Delta_k\left|\frac{1}{\Delta_k}\;\mathbb{E}\,\varphi(X_{\Delta_k}) - \nu(\varphi)\right| \le C\frac{1}{\sqrt{t_n}}\sum_{k=1}^{n}\Delta_k^2 \le C\bar{\pi}\sqrt{t_n} \to 0,$$

by assumption.

Remark 2.4. *As a direct consequence, it follows that $\hat{\beta}^{\pi}(\varphi)$ is a consistent estimator for $\nu(\varphi)$ as $t_n \to \infty$ and $\bar{\pi} \to 0$ so that $\bar{\pi}\sqrt{t_n} \to 0$. As a matter of fact, it suffices that $t_n \to \infty$ and $\bar{\pi} \to 0$, provided that e.g. f is ν-continuous, bounded, and $f(x) = o(|x|^2)$, as $x \to 0$. A proof of this statement is outlined in [32] for regular sampling observations, while the general case is considered in [14].*

In view of the linearity of $\hat{\beta}^{\pi}(\cdot)$ and $\nu(\cdot)$, we conclude that

Corollary 2.5. *Let Ξ be the class of functions $\varphi \in C_0^2(\mathbb{R})$ that vanish in a neighborhood of the origin. Suppose that the linear model $\mathcal{S}$ in (1.4) is such that $\{\varphi_j\}_{j=1}^d \subseteq \Xi$. Then, the projection estimator $\hat{s}^{\pi}(x)$ in (1.9) satisfies the limiting relation*

$$\sqrt{t_n}\left(\hat{s}^{\pi}(x) - s^{\perp}(x)\right) \xrightarrow{\mathfrak{D}} V^{1/2}(x)\, Z \tag{2.5}$$

as $t_n \to \infty$ and $\bar{\pi} \to 0$ so that $\bar{\pi}\sqrt{t_n} \to 0$, where $V(x) := \int f^2(y)\nu(dy)$ with $f(y) := \sum_{j=1}^d \varphi_j(x)\varphi_j(y)$.

Remark 2.6. *Notice that we have the following bound for the variance*

$$V(x) \leq \|s\|_{\infty,D} \sum_{j=1}^{d} \varphi_j^2(x),$$

where $\|s\|_{\infty,D} := \sup_{y\in D} s(y)$.

We can relax the regularity conditions on the moment functions φ by using a simple integration by parts formula (see Remark 3.3 below). A different approach could be to impose additional regularity conditions on the Lévy process itself. In this direction, [28] studies series expansions for the transition density $p_t(x)$ of X_t as powers of t. For instance, one of their results states that if p_t is monotonically decreasing for $x > b$ and $x < -c$, for some $b, c > 0$, then for any $\eta > 0$, there exists $\varepsilon' > 0$ and $t_0 > 0$, such that

$$\frac{1}{t}p_t(x) = e^{-\int_{\{|y|>\varepsilon\}} s(y)dy} s(x) + O_{\varepsilon,\eta}(t), \tag{2.6}$$

for $|x| > \eta$. Such a result will allow us to estimate the rate of convergence in (1.3) if φ vanishes around the origin, since

$$\frac{1}{\Delta}\ \mathbb{E}\,\varphi(X_\Delta) - \nu(\varphi) = \int \varphi(x)\left\{\frac{1}{\Delta}p_\Delta(x) - s(x)\right\}dx.$$

However, we should warn that the derivation of (2.6) in [28] is not completely formal[2], and hence, we avoid to use such an approach in the sequel. See [17] for more insight on the small-time polynomial expansions of the transition distributions of the Lévy process.

3. The Model Selection Problem

In this part we describe how to control the risk (1.6) of the projection estimators by imposing two conditions. First, the time horizon T should be large enough (compared to the complexity of the sieves), while the sampling frequency is kept small compared to the time horizon. These conditions will ensure that the variance term of (1.6) is of order $O(T^{-1})$. Second, the sieves should have good approximating properties in general classes of smooth functions so that when the Lévy density is

[2] The main problem arises from the application of Lemma 1 in [28]. The value of t_0 actually depends on δ. Later on in their proof, δ is taken arbitrarily small, which is likely to result in $t_0 \to 0$ (unless otherwise proved).

presumed to have "degree of smoothness α", the bias term of (1.6) is of order $O(m^{-\alpha})$, where m is the dimension of the sieve (see Section 3.2 for the details). We prove that under the above conditions, we can tune up the dimension of the sieve to the presumed smoothness of s so that the rate of convergence of the risk is of order $O(T^{-2\alpha/(2\alpha+1)})$.

3.1. Analysis of the Variance Term

Consider the setting and notation of the introduction. For simplicity, we focus on estimation windows D in the positive reals (that is, $D := [a, b]$, for some $0 < a < b \leq \infty$). By making the sampling frequency per unit time high enough relative to the sampling horizon T, we can estimate the rate at which the *variance term* of the risk (1.6) decreases in the time horizon T. In the subsequent sections, we will see that this estimate actually leads to a rate of convergence for the risk which is optimal, even if our estimation were based on the whole sample path $\{X_t\}_{t\leq T}$. We shall need the following technical lemma, which we prove in the appendix for the sake of completeness.

Lemma 3.1. *For any $T > 0$, there exist $\delta_T > 0$ and $k > 0$ (independent of T) such that*

$$\sup_{y\in D}\left|\frac{1}{\Delta}\,\mathbb{P}\left[X_\Delta \geq y\right] - \nu([y,\infty))\right| < k\,\frac{1}{T}$$

for all $\Delta < \delta_T$.

The mesh size δ_T will play a very important role below as the asymptotic results in the sequel will hold true as far as the sampling frequency, measured by $\bar{\pi} := \max\{t_k - t_{k-1}\}$, is such that $\bar{\pi} < \delta_T$. Thus, from a practical point of view, estimating δ_T is crucial. We will discuss this point in more detail in the Section 3.4.

The following easy estimate will be useful in the sequel.

Lemma 3.2. *Suppose that φ has support $[c, d] \subset \mathbb{R}_+\backslash\{0\}$, where φ is continuous with continuous derivative. Then,*

$$\left|\frac{\mathbb{E}\,\varphi\left(X_\Delta\right)}{\Delta} - \nu(\varphi)\right| \leq \left(|\varphi(c)| + \int_c^d |\varphi'(u)|\,du\right) M_\Delta([c,d]),$$

where $M_\Delta([c,d]) := \sup_{y\in[c,d]}\left|\frac{1}{\Delta}\,\mathbb{P}\left[X_\Delta \geq y\right] - \nu([y,\infty))\right|$.

Proof. The result follows from the following identities

$$\mathbb{E}\,\varphi(X_\Delta) = \varphi(c)\,\mathbb{P}\left[X_\Delta \geq c\right] + \int_c^\infty \varphi'(u)\,\mathbb{P}\left[X_\Delta \geq u\right]du,$$

$$\int \varphi(x)\nu(dx) = \varphi(c)\nu\left([c,\infty)\right) + \int_c^\infty \varphi'(u)\nu\left([u,\infty)\right)du.$$

These are standard consequences of Fubini's Theorem.

Remark 3.3. *We can apply the previous two lemmas to obtain CLTs for $\hat{\beta}^\pi$ and $\hat{s}^\pi$. Indeed, if φ is as in Lemma 3.2 and, for each T, the partition π_T has mesh smaller than δ_T, the critical value in Lemma 3.1, then (2.3) holds true. The projection estimator $\hat{s}^\pi$ will satisfy (2.5) provided that the basis functions φ are as in Lemma 3.2.*

We are now ready to estimate the variance term. We shall impose conditions on the approximating linear models so that the estimates of the above Lemmas are applicable.

Standing assumption 1. *The linear model $\mathcal{S}$ of (1.4) is generated by an orthonormal basis $\mathcal{G} := \{\varphi_j\}_{j=1}^d$ such that each φ_j is bounded with continuous derivative on the interior of its support, which is assumed to be of the form $[x_{j-1}, x_j] \subset D$.*

In the sequel, we will need the following notation:

$$D_1(\mathcal{S}) := \inf_{\mathcal{G}} \max\left\{\|\varphi\|_\infty^2 : \varphi \in \mathcal{G}\right\}, \tag{3.1}$$

$$D_2(\mathcal{S}) := \inf_{\mathcal{G}} \max\left\{\|\varphi'\|_1^2 : \varphi \in \mathcal{G}\right\}, \tag{3.2}$$

where the infimums are over all orthonormal bases $\mathcal{G}$ of $\mathcal{S}$.

Proposition 3.4. *There exists a constant $K > 0$ such that*

$$\mathbb{E}\,\|s^\perp - \hat{s}^\pi\|^2 \le K\,\frac{\dim(\mathcal{S})}{T}, \tag{3.3}$$

for any linear model $\mathcal{S}$ satisfying the Standing Assumption 1, and for any partition $\pi : 0 = t_0 < \cdots < t_n = T$ such that $T > \max\{D_1(\mathcal{S}), D_2(\mathcal{S})\}$ and $\bar{\pi} < \delta_T$, where δ_T is the "critical" mesh size introduced in Lemma 3.1.

Proof. Fix an orthonormal basis $\mathcal{G} := \{\varphi_j\}_{j=1}^d$ of $\mathcal{S}$. Let $D_t(\varphi) := \frac{1}{t}\,\mathbb{E}\,\varphi(X_t) - \nu(\varphi)$. For any $\varphi_j \in \mathcal{G}$, we have

$$\mathbb{E}\left\{\hat{\beta}^\pi(\varphi_j) - \nu(\varphi_j)\right\}^2 \le \frac{1}{t_n}\int \varphi_j^2(x)\nu(dx) + \\ + \frac{1}{t_n^2}\sum_{k=1}^n \left|D_{\Delta_k}(\varphi_j^2)\right|\Delta_k + \left\{\frac{1}{t_n}\sum_{k=1}^n \left|D_{\Delta_k}(\varphi_j)\right|\Delta_k\right\}^2,$$

where $\Delta_k := t_k - t_{k-1}$. Then, from the previous two lemmas, when $\bar{\pi} < \delta_T$,

$$\mathbb{E}\left\{\hat{\beta}^\pi(\varphi_j) - \nu(\varphi_j)\right\}^2 \le \frac{1}{T}\int \varphi_j^2(x)\nu(dx) \\ + \frac{k}{T^2}\left(|\varphi_j^2(x_{j-1})| + \int_{x_{j-1}}^{x_j}\left|2\varphi_j(u)\varphi_j'(u)\right|du\right) \\ + \frac{k^2}{T^2}\left(|\varphi_j(x_{j-1})| + \int_{x_{j-1}}^{x_j}\left|\varphi_j'(u)\right|du\right)^2,$$

which can be simplified further as follows

$$\mathbb{E}\left\{\hat{\beta}^\pi(\varphi_j) - \nu(\varphi_j)\right\}^2 \le \frac{1}{T}\int \varphi_j^2(x)\nu(dx) + \frac{2k^2}{T^2}\left(\|\varphi_j\|_\infty + \|\varphi_j'\|_1\right)^2 \\ \le \frac{\|s\cdot\chi_{_D}\|_\infty}{T} + 8k^2\frac{\max_{j'}\|\varphi_{j'}\|_\infty^2 + \|\varphi_{j'}'\|_1^2}{T^2}.$$

Then,

$$\mathbb{E}\,\|s^\perp - \hat{s}^\pi\|^2 \le \frac{\dim(\mathcal{S})}{T}\left\{\|s\cdot\chi_{_D}\|_\infty + 8k^2\frac{\max_{j'}\|\varphi_{j'}\|_\infty^2 + \|\varphi_{j'}'\|_1^2}{T}\right\}.$$

Now, it is evident that (3.3) holds whenever $T > \max\{D_1(\mathcal{S}), D_2(\mathcal{S})\}$.

3.2. The Approximation Error for Besov Type Smooth Functions

As it is customary, the bias term in (1.6) will be estimated by imposing certain degree of smoothness on the function s. Concretely, the restriction of the Lévy density s to $D := [a,b]$ is assumed to belong to the *Besov space* $\mathcal{B}^{\alpha}_{\infty}\left(L^p([a,b])\right)$ for some $p \in [2,\infty]$ and $\alpha > 0$ (see for instance [12] and references therein for background on these spaces). The space $\mathcal{B}^{\alpha}_{\infty}\left(L^p([a,b])\right)$ consists of those functions f belonging to $L^p([a,b])$ if $0 < p < \infty$ (or being uniformly continuous if $p = \infty$) such that

$$|f|_{\mathcal{B}^{\alpha}_{\infty}(L^p)} \equiv \sup_{\delta>0} \frac{1}{\delta^{\alpha}} \sup_{0<h\leq\delta} \|\Delta^r_h(f,\cdot)\|_p < \infty,$$

with $r := [\alpha]+1$. Here, $\Delta_h(f,x) \equiv f(x+h) - f(x)$ and $\Delta^r_h(f,x)$ is the r^{th}-order difference of f defined recursively by

$$\Delta^r_h(f,x) \equiv \Delta_h(\Delta^{r-1}_h(f,\cdot),x),$$

for x's such that $x + rh \in D$ and $r \in \mathbb{N}$.

The Besov class is closely related to the so-called class of *Lipschitz functions*. For constants $k \in \mathbb{N}$ and $\beta \in (0,1]$, f is said to belong to $\mathrm{Lip}(k+\beta, L^p([a,b]))$ if $f, \ldots, f^{(k-1)}$ are absolutely continuous (on $[a,b]$) and $f^{(k)}$ belongs to $L^p((a,b))$ and satisfies

$$\sup_{h>0} \frac{1}{h^{\beta}} \|\Delta_h(f^{(k)},\cdot)\|_p < \infty. \tag{3.4}$$

It is known that if $\beta < 1$ and $1 \leq p \leq \infty$, then $f \in \mathrm{Lip}(k+\beta, L^p([a,b]))$ if and only if f is a.e. equal to a function in $\mathcal{B}^{\alpha}_{\infty}\left(L^p([a,b])\right)$ with $\alpha := k+\beta$. In general, $\mathrm{Lip}(k+\beta, L^p([a,b])) \subset \mathcal{B}^{k+\beta}_{\infty}\left(L^p([a,b])\right)$, for any $0 < p \leq \infty$ (see e.g. [12]). Notice that when $p = \infty$, the condition (3.4) takes the form:

$$|f^{(k)}(x) - f^{(k)}(y)| \leq L|x-y|^{\beta}, \tag{3.5}$$

for all $x, y \in (a,b)$ and some $L < \infty$.

An important reason for working with the Besov-type smooth functions is the availability of estimates of the approximation error by splines, trigonometric polynomials, and wavelets (see [12] and [3] for more details). For instance, if $\mathcal{S}_{k,m}$ denotes the space of piecewise polynomials of degree at most k, based on a regular partition of $[a,b]$ with m classes, and $s \in \mathcal{B}^{\alpha}_{\infty}\left(L^p([a,b])\right)$ with $\alpha < k+1$, then there exists a constant $c_{[\alpha]} < \infty$ such that

$$\inf_{f\in\mathcal{S}_{k,m}} \|s-f\|_p \leq c_{[\alpha]}|s|_{\mathcal{B}^{\alpha}_{\infty}(L^p)}(b-a)^{\alpha}m^{-\alpha}. \tag{3.6}$$

Thus, when $p \geq 2$, the orthogonal projection of s on $\mathcal{S}_{k,m}$, denoted by $s^{\perp}_m$, is such that

$$\|s - s^{\perp}_m\| \leq c_{[\alpha]}(b-a)^{\frac{1}{2}-\frac{1}{p}+\alpha}|s|_{\mathcal{B}^{\alpha}_{\infty}(L^p)}m^{-\alpha}. \tag{3.7}$$

Notice that the elements of $\mathcal{S}_{k,m}$ are not necessarily smooth (not even continuous) and hence, they are not "splines" in the standard sense of the literature, where a spline is understood as a smooth piece-wise polynomial. The upper bound (3.6) is actually true if we restrict to certain splines of $\mathcal{S}_{k,m}$ (say B-splines) (see (10.1) in Chapter 2 of [12]). For the sake of completeness let us describe in detail the

space $S_{k,m}$ as well as give estimates for the constants (3.1-3.2). Let Q_j be the Legendre polynomials of order j on $\mathbb{L}^2([-1,1],dx)$. The space $\mathcal{S}_{k,m}$ is generated by the orthonormal functions

$$\hat{\varphi}_{i,j}(x) := \sqrt{\frac{2j+1}{x_i - x_{i-1}}}\, Q_j\left(\frac{2x-(x_i+x_{i-1})}{x_i - x_{i-1}}\right)\mathbf{1}_{(x_{i-1},x_i)}(x),$$

for $i = 1,\dots,m$ and $j = 0,\dots,k$, and where $a = x_0 < \cdots < x_m = b$ are equally-spaced points. It is well-known that $|Q_j(x)| \le 1$ and $|Q_j'(x)| \le Q_j'(1) = \frac{j(j+1)}{2}$. Then, fixing $\Delta_x := x_i - x_{i-1} = \frac{b-a}{m}$, we have that

$$\hat{\varphi}_{i,j}'(x) = 2\sqrt{2j+1}\,\Delta_x^{-3/2}\, Q_j'\left(\frac{2x-(x_i+x_{i-1})}{x_i - x_{i-1}}\right)\mathbf{1}_{(x_{i-1},x_i)}(x),$$

$$\|\hat{\varphi}_{i,j}'\|_1 \le 2\sqrt{2j+1}\Delta_x^{-3/2}\int_{x_{i-1}}^{x_i}\sup_u |Q_j'(u)|dx \le \sqrt{2j+1}\Delta_x^{-1/2}(j)(j+1).$$

It is now clear that

$$D_2(\mathcal{S}_{k,m}) \le \max_{i,j}\left\{\|\varphi_{i,j}'\|_1^2\right\} \le \frac{(k+1)^2k^2(2k+1)}{b-a}\,m.$$

In a similar manner one can check that

$$D_1(\mathcal{S}_{k,m}) \le \frac{(k+1)^2(2k+1)}{b-a}\,m.$$

3.3. Rate of Convergence for Smooth Functions Via Splines

As a consequence of the variance and bias term estimates given in the previous two parts, we now estimate the rate of convergence on D of the projection estimators (1.9), on the regular piece-wise polynomials $\{\mathcal{S}_{k,m}\}_{m\ge1}$, assuming that the Lévy density s is in the Besov class $\mathcal{B}_\infty^\alpha\left(L^p([a,b])\right)$ with $p \ge 2$ and $\alpha < k+1$. It turns out that under the stated conditions, projection estimators converge at a rate at least as good as $T^{-2\alpha/(2\alpha+1)}$. The following result is valid provided that, for each time horizon T, the mesh of the sampling times π_T is smaller than the critical mesh δ_T introduced in Lemma 3.1. In Section 4, we will see that this rate is actually the best possible even under continuous sampling.

Proposition 3.5. *Let $\hat{m}_T := \left\lceil T^{1/(2\alpha+1)}\right\rceil$ and let $\Theta(R,L)$ be the class of Lévy densities s such that $\|s\cdot\chi_D\|_\infty < R$, and such that the restriction of s to $D := [a,b]$ is a member of $\mathcal{B}_\infty^\alpha\left(L^p([a,b])\right)$ with $|s|_{\mathcal{B}_\infty^\alpha(L^p)} < L$ and $p \ge 2$. Then,*

$$\limsup_{T\to\infty} T^{2\alpha/(2\alpha+1)} \sup_{s\in\Theta(R,L)} \mathbb{E}\left[\|s-\hat{s}_T\|^2\right] < \infty, \tag{3.8}$$

where for each T, the estimator $\hat{s}_T$ is given by (1.7) and (1.9) with $\mathcal{S} = \mathcal{S}_{k,\hat{m}_T}$, $k > \alpha - 1$, and a mesh $\bar{\pi}_T$ smaller than δ_T.

Proof. From the two previous parts, there exists a constant K (depending on k,a,b,α,R,p,L) such that

$$\|s - s_m^\perp\| \le K\, m^{-\alpha} \quad \text{and} \quad \mathbb{E}\,\|s^\perp - \hat{s}_m^\pi\|^2 \le K\,\frac{m}{T},$$

for $m \in \mathcal{M}_T := \{m' : T > Km'\}$ and $\bar{\pi} < \delta_T$. Then for a constant M and for large enough T,

$$\sup_{s\in\Theta(R,L)} \mathbb{E}\left[\|s - \hat{s}_T\|^2\right] \leq M\left\{\left[T^{1/(2\alpha+1)}\right]^{-2\alpha} + \left[T^{1/(2\alpha+1)}\right]T^{-1}\right\}.$$

The limit (3.8) is now clear.

Example 3.6. *If s has continuous bounded derivative on $D := [a,b] \subset \mathbb{R}\backslash\{0\}$ (hence, $s \in \mathcal{B}^{\alpha}_{\infty}\left(L^{\infty}([a,b])\right)$, for any $\alpha < 1$), then one can construct regular histogram estimators converging to s on D at a rate faster than $T^{-1/2}$ if one selects the number of classes approximately equal to $T^{1/2}$ and the mesh of the partition π smaller than δ_T.*

3.4. About the Critical Mesh

The critical mesh, introduced in Lemma 3.1, gives a bound on the mesh of the sampling frequency needed to estimate in a simple way the rate of convergence of the variance term (see Proposition 3.4). Of course, any hope for a feasible implementation of this estimation scheme will require an explicit estimate of this critical mesh. In the compound Poisson case (when $\nu(\mathbb{R}\backslash\{0\}) < \infty$), it turns out that $\delta_T = o\left(\frac{1}{T}\right)$ suffices. In the general case, we have the following result, which tell us, in particular, that the sampling frequency needs to be higher when one wishes to estimate the Lévy density closer to the origin.

Proposition 3.7. *Let $\rho > 0$ such that $a\rho > 1$. Then, there exists $T_0(\rho) > 0$ and $k > 0$ such that*

$$\sup_{y\in D}\left|\frac{1}{\Delta}\mathbb{P}\left[X_\Delta \geq y\right] - \nu([y,\infty))\right| < k\,\frac{1}{T}$$

for all $T > T_0$ and $\Delta < T^{-\frac{1}{\rho}T}$.

Proof. As in the proof of Lemma 3.1, we can obtain

$$\sup_{y\in D}\left|\frac{1}{t}\mathbb{P}\left[X_t \geq y\right] - \nu([y,\infty))\right| \leq \frac{1}{t}\mathbb{P}\left[X^{\varepsilon}_t \geq a\right] + 2c\frac{1}{T} + \nu([a,\infty))\,\mathbb{P}\left[|X^{\varepsilon}_t| \geq \eta\right] + \lambda_\varepsilon\,\mathbb{P}\left[|X^{\varepsilon}_t| \geq \eta\right] + \nu([a,\infty))\,\lambda_\varepsilon t + \lambda^2_\varepsilon t,$$

valid for $T > 1/a$, $\eta = \frac{1}{T}$, and $0 < \varepsilon < a - \eta$ (here $c := \sup_{a-\eta\leq x\leq b+\eta} s(x)$). Fix $\varepsilon > 0$ sufficiently small so that $\rho < \frac{1}{\varepsilon}$. Let us recall that there exists $y_0 := y_0(\rho)$ such that

$$\mathbb{P}\left[|X^{\varepsilon}_t| \geq y\right] \leq \exp\{\rho y_0 \log y_0\}\exp\{\rho y - \rho y \log y\}\, t^{y\rho}$$

for all $t < \frac{y}{y_0(\rho)}$ (see e.g. [28]). In particular, when $y = \eta = \frac{1}{T}$ and $t < T^{-\frac{1}{\rho}T}$, for T sufficiently large that $T^{-\frac{1}{\rho}T+1} < \frac{1}{y_0(\rho)}$,

$$\mathbb{P}\left[|X^{\varepsilon}_t| \geq \eta\right] \leq kT^{-1}.$$

Similarly, when $y = a$ and $t < T^{-\frac{1}{\rho}T}$,

$$\frac{1}{t}\mathbb{P}\left[X^{\varepsilon}_t \geq a\right] \leq kt^{a\rho-1} < kT^{-\frac{1}{\rho}T(a\rho-1)} < kT^{-1},$$

if $T > \frac{\rho}{a\rho-1}$. This proves the result since ε is fixed.

Remark 3.8. *The estimate of the critical mesh given in Proposition 3.7 can be improved substantially. Indeed, in a forthcoming paper, we will show that it suffices that* $\Delta = o(T^{-1})$.

4. Minimax Risk of Estimation for Smooth Lévy Densities

In this section, we show that the rate of convergence $O(T^{-2\alpha/(2\alpha+1)})$ attained by projection estimators is the best possible, in the sense that there is no estimator $\hat{s}^*_T$ that can converge to s faster than $T^{-2\alpha/(2\alpha+1)}$, for any $s \in \Theta$, even assuming continuous-time sampling. In order to prove this, we will assess the long-run behavior of the *minimax risk* on Θ, roughly defined as

$$\inf_{\hat{s}} \sup_{s\in\Theta} \mathbb{E}_s \left[d\left(s, \hat{s}\right) \right],$$

where the infimum is taken over *all possible* estimators $\hat{s}$, and $d(s, \hat{s})$ measures the distance between $\hat{s}$ and s.

Traditionally, the performance of nonparametric estimators is gauged by comparing the rate of convergence of the estimator in question to the rate of convergence of the minimax risk when the available data increases. The rates of convergence of minimax risks are available in most of the traditional nonparametric problems. For instance, Ibragimov and Has'minskii [19] and Barron et. al. [2] provided this kind of asymptotics for the problem of density estimation based on i.i.d. random variables, while Kutoyants [22] and Reynaud-Bouret [25] considered the problem of intensity estimation of a *finite* Poisson point processes. This last set-up is relevant for our problem since the jumps of a Lévy process can be associated with a (possibly infinite) Poisson point process on $\mathbb{R}_+ \times \mathbb{R}\backslash\{0\}$ (see e.g. Theorem 19.2 in [29]). Using this connection, we adapt below a result from [22] to obtain the long-run asymptotics of the minimax risk of estimation of the Lévy density off the origin. The idea of the proof, due to Ibragimov and Has'minskii [19], is based on the statistical toolbox for distributions satisfying the *Local Asymptotic Normality* (LAN) property (see Chapters II and Section IV.5 of [19]).

Let us introduce some notation. Here, $\ell : \mathbb{R} \to \mathbb{R}$ stands for a *loss function* satisfying the following:

(i) $\ell(\cdot)$ is nonnegative, $\ell(0) = 0$ but not identically 0, and ℓ continuous at 0;
(ii) ℓ is symmetric: $\ell(u) = \ell(-u)$ for all u;
(iii) $\{u : \ell(u) < c\}$ is a convex set for any $c > 0$, ;
(iv) $\ell(u)\exp\{\varepsilon|u|^2\} \to 0$ as $|u| \to \infty$ for any $\varepsilon > 0$.

We consider Lévy densities $s : \mathbb{R}\backslash\{0\} \to \mathbb{R}_+$ that are k times differentiable on an interval $[a, b] \subset \mathbb{R}\backslash\{0\}$ and satisfy (3.5) for all $x, y \in [a, b]$. For given $k \in \mathbb{N}$ and $\beta \in (0, 1]$, we denote such a class of functions by $\Theta_{k+\beta}(L; [a, b])$. The proof of the result below is presented in the Appendix 6.3.

Theorem 4.1. *If* x_0 *is an interior point of the interval* $[a, b]$*, then*

$$\liminf_{T\to\infty} \left\{ \inf_{\hat{s}_T} \sup_{s\in\Theta} \mathbb{E}_s \left[\ell\left(T^{\alpha/(2\alpha+1)} \left(\hat{s}_T(x_0) - s(x_0) \right) \right) \right] \right\} > 0, \tag{4.1}$$

where $\alpha := k + \beta$, $\Theta := \Theta_\alpha(L; [a, b])$ *and the infimum is over all the estimators* $\hat{s}_T$ *of* s *based on* $\{X(t)\}_{0\le t\le T}$.

The previous result can be strengthen to be uniform in $x_0 \in (a,b)$ and as a consequence, the long-run behavior of the minimax risk under the integrated mean-square distance can be assessed. The proof of the next result is given in Appendix 6.3.

Corollary 4.2. *Under the notation and conditions of Theorem 4.1, the following two limits hold:*

$$\liminf_{T\to\infty} \left\{ \inf_{\hat{s}_T} \inf_{x\in(a,b)} \sup_{s\in\Theta} \mathbb{E}_s \left[\ell \left(T^{\alpha/(2\alpha+1)} \left(\hat{s}_T(x) - s(x) \right) \right) \right] \right\} > 0, \tag{4.2}$$

$$\liminf_{T\to\infty} T^{2\alpha/(2\alpha+1)} \left\{ \inf_{\hat{s}_T} \sup_{s\in\Theta} \mathbb{E}_s \left[\int_a^b \left(\hat{s}_T(x) - s(x) \right)^2 dx \right] \right\} > 0. \tag{4.3}$$

Remark 4.3. *The previous result is also valid for classes slightly smaller than* $\Theta_\alpha(L;[a,b])$ *such as*

$$\Theta = \Theta_\alpha(L;[a,b]) \cap \{ s : \|s\|_{\mathbb{L}^\infty([a,b])} < R \},$$

which is closely related to the Besov class $\Theta(R,L)$ *of (3.8). Indeed,* $\Theta_\alpha(L;[a,b])$ *is contained in* $\mathcal{B}^\alpha_\infty\left(\mathbb{L}^\infty([a,b])\right)$ *(see Section 2.9 of [12]), and thus,*

$$\liminf_{T\to\infty} T^{2\alpha/(2\alpha+1)} \left\{ \inf_{\hat{s}_T} \sup_{s\in\Theta(R,L)} \mathbb{E}_s \left[\int_a^b \left(\hat{s}_T(x) - s(x) \right)^2 dx \right] \right\} > 0. \tag{4.4}$$

We conclude that there is no reasonable *estimator* $\hat{s}_T$ *of* s *capable of outperforming the rate* $T^{-2\alpha/(2\alpha+1)}$ uniformly on Θ*: there is always an* $s \in \Theta$ *for which*

$$T^{2\alpha/(2\alpha+1)}\, \mathbb{E}_s \left[\|\hat{s}_T - s\|^2 \right] > B,$$

for some $B > 0$ *and for large enough* T*. Therefore, the estimator described in Proposition 3.5 achieves the optimum rate of convergence on* $\Theta(R,L)$ *from a minimax point of view.*

5. A Data-Driven Selection Method and Adaptability

The model selection criterion described in Section 3.3, where one tunes up the number of classes m to the "smoothness" of s, has the obvious drawback of requiring (or at least presuming) the smoothness parameter α. In the literature of nonparametric statistics, one wishes to devise data-driven selection methods that can *adapt* to arbitrary degree of smoothness (see e.g. Birgé and Massart [5] for an extensive exposition of the topic).

A typical approach for adaptive model selection schemes consists of minimizing an unbiased estimator of the risk of estimation. This approach was developed in [13] in the context of Lévy density estimation. Let us briefly discuss the findings there. The key idea comes from the following refinement of (1.6):

$$\mathbb{E}\left[\|s - \hat{s}^c\|^2 \right] = \|s\|^2 + \mathbb{E}\left[-\|\hat{s}^c\|^2 + \mathrm{pen}^c(\mathcal{S}) \right], \tag{5.1}$$

where $\hat{s}^c$ is as in (1.9) substituting $\hat{\beta}^\pi(\varphi)$ by the statistics $\hat{\beta}^c(\varphi)$ of (1.13), $s^\perp$ is the orthogonal projection in (1.5), and $\mathrm{pen}^c(\mathcal{S})$ is defined in terms of an orthonormal basis $\mathcal{G} := \{\varphi_1, \ldots, \varphi_d\}$ of $\mathcal{S}$ by the formula:

$$\mathrm{pen}^c(\mathcal{S}) \equiv \frac{2}{T^2} \sum_{t\le T} \sum_{\varphi\in\mathcal{G}} \varphi^2(\Delta X_t). \tag{5.2}$$

Equation (5.1) shows that the risk of $\hat{s}^c$ moves "parallel" to the expectation of the *observable statistics* $-\|\hat{s}^c\|^2 + \text{pen}^c(\mathcal{S})$, suggesting the selection of the model that minimizes such statistics. Concretely, given a collection of sieves $\{\mathcal{S}_m, m \in \mathcal{M}\}$, we should choose the projection estimator $\tilde{s}^c \equiv \hat{s}^c_{\hat{m}}$, where

$$\hat{m} \equiv \text{argmin}_{m\in\mathcal{M}} \left\{-\|\hat{s}^c_m\|^2 + \text{pen}^c(\mathcal{S}_m)\right\}.$$

Such an estimator $\tilde{s}^c$ is called a *penalized projection estimator* (p.p.e.) since the role of $\text{pen}^c(\mathcal{S})$ is to penalize large linear models.

In [16], it is shown that the p.p.e. $\tilde{s}^c$ is adaptive in the class of Besov Lévy densities of Section 3.2 in the sense that $\tilde{s}^c$ attains the optimal rate of convergence $O(T^{-2\alpha/(2\alpha+1)})$ without using the knowledge of α. Unfortunately, the previous approach intrinsically requires continuous-time sampling of the process to determine the jumps ΔX_t. However, the analysis could still be useful if one uses the natural discrete-based proxies of $\hat{\beta}^c$ and pen^c, where the jumps ΔX_t are replaced by the increments $X_{t_k} - X_{t_{k-1}}$. This idea leads to the estimators $\hat{s}^\pi$ in (1.9) and to the statistic

$$\text{pen}^\pi(\mathcal{S}) = \frac{2}{T^2}\sum_{k=1}^{n}\sum_{\varphi\in\mathcal{G}}\varphi^2(X_{t_k} - X_{t_{k-1}}) \tag{5.3}$$

as the penalization term. In the light of the previous arguments, we proposed a discrete-based model selection criterion as follows

$$\begin{aligned}\hat{m}^\pi &\equiv \text{argmin}_{m\in\mathcal{M}} \left\{-\|\hat{s}^\pi_m\|^2 + \text{pen}^\pi(\mathcal{S}_m)\right\} \\ &= \text{argmin}_{m\in\mathcal{M}} \left\{ -\sum_{\varphi\in\mathcal{G}_m}\{\hat{\beta}^\pi(\varphi)\}^2 + \text{pen}^\pi(\mathcal{S}_m)\right\}\end{aligned} \tag{5.4}$$

where $\mathcal{G}_m$ is an orthonormal basis of $\mathcal{S}_m$, $\hat{\beta}^\pi$ is given by (1.7), and pen^π is given by (5.3). The resulting estimator

$$\widetilde{s} := s^\pi_{\hat{m}^\pi} \tag{5.5}$$

will be called (discrete-based) *penalized projection estimator*.

We hope to extend in a future work the adaptability result in [16] for this discrete-based p.p.e. In the sequel, we illustrate the performance of these estimators for an infinite-jump activity Lévy process of relevance in the area of mathematical finance.

6. An Example: Estimation of Variance Gamma Processes.

6.1. The Model

Variance Gamma processes were proposed in [23] (see also [8]) as substitutes to the Brownian Motion in the Black-Scholes model. Since their introduction, this kind of processes have received a great dealt of attention, even in the financial industry. For an introduction to many basic properties of variance Gamma processes and other related processes, the reader is referred to Knotz et al. [21].

There are two useful representations for this type of processes. A variance Gamma process $X = \{X(t)\}_{t\geq 0}$ is a Brownian motion with drift, time changed by a Gamma Lévy process. Concretely,

$$X(t) = \theta U(t) + \sigma W(U(t)), \tag{6.1}$$

where $\{W(t)\}_{t\geq 0}$ is a standard Brownian motion, $\theta \in \mathbb{R}$, $\sigma > 0$, and $U = \{U(t)\}_{t\geq 0}$ is an independent Gamma Lévy process with density at time t given by

$$f_t(x) = \frac{x^{t/\nu - 1} \exp\left(-\frac{x}{\nu}\right)}{\nu^{t/\nu}\Gamma\left(\frac{t}{\nu}\right)}. \tag{6.2}$$

Notice that $\mathrm{E}[U(t)] = t$ and $\mathrm{Var}[U(t)] = \nu t$; therefore, the random clock U has a "mean rate" of one and a "variance rate" of ν. There is no loss of generality in restricting the mean rate of the Gamma process U to one since, as a matter of fact, any process of the form

$$\theta_1 V(t) + \sigma_1 W(V(t)),$$

where $V(t)$ is an arbitrary Gamma Lévy process, $\theta_1 \in \mathbb{R}$, and $\sigma_1 > 0$, has the same law as a process of the form (6.1) with suitably chosen θ, σ, and ν. This a consequence of the *self-similarity*[3] property of Brownian motion and the fact that ν in (6.2) is a scale parameter.

The process X is itself a Lévy process since Gamma processes are *subordinators* (see Theorem 30.1 of [29]). Moreover, it is not hard to check that "statistically" X is the difference of two Gamma Lévy processes (see e.g. (2.1) of [6]):

$$\{X(t)\}_{t\geq 0} \stackrel{\mathfrak{D}}{=} \{X_+(t) - X_-(t)\}_{t\geq 0}, \tag{6.3}$$

where $\{X_+(t)\}_{t\geq 0}$ and $\{X_-(t)\}_{t\geq 0}$ are Gamma Lévy processes with respective Lévy measures

$$\nu_\pm(dx) = \alpha \exp\left(-\frac{x}{\beta^\pm}\right) dx, \quad \text{for } x > 0.$$

Here, $\alpha = 1/\nu$ and

$$\beta^\pm = \sqrt{\frac{\theta^2\nu^2}{4} + \frac{\sigma^2\nu}{2}} \pm \frac{\theta\nu}{2}.$$

As a consequence of this decomposition, the Lévy density of X takes the form

$$s(x) = \begin{cases} \frac{\alpha}{|x|} \exp\left(-\frac{|x|}{\beta^-}\right) & \text{if } x < 0, \\ \frac{\alpha}{x} \exp\left(-\frac{x}{\beta^+}\right) & \text{if } x > 0, \end{cases} \tag{6.4}$$

where $\alpha > 0$, $\beta^- \geq 0$, and $\beta^+ \geq 0$ (of course, $|\beta^-| + |\beta^+| > 0$). As in the case of Gamma Lévy processes, α controls the overall jump activity, while β^+ and β^- take respectively charge of the intensity of large positive and negative jumps. In particular, the difference between $1/\beta^+$ and $1/\beta^-$ determines the frequency of drops relative to rises, while their sum measures the frequency of large moves relative to small ones.

6.2. The Simulation Procedure

The above two representations provide straightforward methods to simulate a variance Gamma model. One way will be to simulate the Gamma Lévy processes $\{X_+(t)\}_{0\leq t\leq T}$ and $\{X_-(t)\}_{0\leq t\leq T}$ of (6.3) using the series representation method introduced in Rosiński [26]. The other approach is to generate the random time change $\{U(t)\}_{0\leq t\leq T}$ of (6.1), and then construct a discrete skeleton from the increments $X(i\Delta t) - X((i-1)\Delta t)$, $i \geq 1$. The increments of X are simply simulated using normal random variables with mean and variances determined by the increments of U.

[3] namely, $\{W(ct)\}_{t\geq 0} \stackrel{\mathfrak{D}}{=} \{c^{1/2}W(t)\}_{t\geq 0}$, for any $c > 0$.

6.3. The Numerical Results

In this part we illustrate the performance of the projection estimators (1.9) and the model selection criterion described in Section 5 using simulation experiments. The approximating linear models $\mathcal{S}_m$ considered here are the span of the indicator functions $\chi_{[x_0,x_1]},\dots,\chi_{(x_{m-1},x_m]}$, where $x_0<\cdots<x_m$ is a regular partition of an interval $D\equiv[a,b]$, with $0<a$ or $b<0$. We perform the following numerical experiment. First, we simulate the variance gamma Lévy process with specified (known) parameter settings. Then, we apply the penalized projection estimator defined by (5.4)-(5.5). Finally, to assess the accuracy of the nonparametric estimator, the true parametric model of s is subsequently fit to the nonparametric estimator using a least-square errors method. Concretely, if $\widetilde{s}:=\hat{s}^{\pi}_{\hat{m}^{\pi}}$ is the discrete-based p.p.e. and s^{θ} is the function (6.4), where we set $\theta:=(\alpha,\beta^-,\beta^+)$, then we find

$$\hat{\theta}_{_{NP}}:=\operatorname{argmin}_{\theta}\sum_{i=0}^{\hat{m}^{\pi}-1}(\widetilde{s}(\bar{x}_i)-s^{\theta}(\bar{x}_i)^2 \tag{6.5}$$

where $\bar{x}_i$ is the midpoint of the interval $[x_i,x_{i+1}]$. This approach provides a nonparametric based estimators for the parameters of the variance Gamma process.

Notice that, from an algorithmic point of view, the estimation for the variance Gamma model using penalized projection is not different from the estimation for the Gamma process. We can simply estimate both tails of the variance Gamma process separately. However, from the point of view of maximum likelihood estimation (MLE), the problem is numerically challenging. Even though the marginal density functions have "closed" form expressions[4] (see [8]), there are well-documented issues with MLE (see for instance [24]). The likelihood function is highly flat for a wide range of parameters and good starting values as well as convergence are critical. Also, the separation of parameters and the identification of the variance Gamma process from other classes of the generalized hyperbolic Lévy processes is difficult. In fact, difference between subclasses in terms of likelihood is small. It is important to mention that these issues worsen when dealing with "high-frequency" data.

Let us consider a numerical example motivated by the empirical findings of [8] based on daily returns on the S&P stock index from January 1992 to September 1994 (see their Table I). Using maximum likelihood methods, the annualized estimates of the parameters for the variance Gamma model were reported to be $\hat{\theta}_{_{ML}}=-0.00056256$, $\hat{\sigma}^2_{_{ML}}=0.01373584$, and $\hat{\nu}_{_{ML}}=0.002$, from where we obtain $\hat{\alpha}_{_{ML}}=500$, $\hat{\beta}^+_{_{ML}}=0.0037056$, and $\hat{\beta}^-_{_{ML}}=0.0037067$.

Figures 1 and 2 show respectively the left- and right- tails of the true Lévy density and the (discrete-based) penalized projection estimator as well as their corresponding best-fit variance Gamma Lévy densities using (6.5), and their marginal probability density functions (pdf) scaled by $1/\Delta t$ (the reciprocal of the time span between observations). The estimation was based on 5000 simulated increments with Δt equal to one-eight of a day. The figures seem quite comforting. To get a better idea of the performance of the method, Figures 3 and 4 show the sampling distributions of the estimates of α^- and β^+ obtained from applying the least-square method to the penalized projection estimators. The histograms are based on 1000 samples of size 5000 with $\Delta t=1/8$ of a day. This experiment shows clear, though

[4] More concretely, the density is terms of Bessel special functions of third kind. For more information, see also Section 4.1 in Knotz et al. [21].

not critical, underestimation of the parameter α and overestimation of the parameters β's. A simple method of moments (based on the first four moments) yields better results (see Figures 5 and 6). Nonparametric methods are not free-lunches and usually the gain in robustness is paid by a loss in efficiency.

To illustrate the seriousness of applying an efficient estimation method to a misspecified model let us consider a close relative of the variance Gamma process: the CGMY model in [6]. This is defined as a pure-jump Lévy process with Lévy density of the form

$$s_m(x) = \begin{cases} \frac{\alpha^-}{|x|^{\upsilon+1}} \exp\left(-\frac{|x|}{\beta^-}\right) & \text{if } x < 0, \\ \frac{\alpha^+}{x^{\upsilon+1}} \exp\left(-\frac{x}{\beta^+}\right) & \text{if } x > 0, \end{cases} \tag{6.6}$$

where $\upsilon > 0$. In the case when $\alpha^- = 0$ and $\nu = 0$, we recover a Gamma Lévy process, for which MLE are widely available. Let us take $\alpha^+ = \beta^+ = 1$ and $\upsilon = .1$. We can estimate the parameter υ using a Zolotarev type estimator. This can be done so since the CGMY Lévy process is a *tempered stable Lévy process*, whose short-term increments behave like stable processes (see Rosínski [27] for details).

The following table shows the sampling average and standard deviations of the estimators of α^+, β^+, and υ by two methods based on 100 simulation runs. The first method estimates υ using the Zolotarev's estimator $\hat{\upsilon}$, then computes the piece-wise constant p.p.e. $\widetilde{s}$ of (5.5), and finally, estimate α^+ and β^+ via the LSE method (6.5) replacing s^θ by the Lévy density s_m of (6.6) with $\theta = (\alpha^+, \beta^+)$ and fixing $\alpha_- = 0$ and $\upsilon = \hat{\upsilon}$. The second method assumes (erroneously) that the underlying model is a Lévy gamma process and performs maximum likelihood estimation.

Δt	Penalized Projection Estimators/Least-Squares Fit			Misspecified Gamma MLE	
	$\hat{\alpha}^+_{NP}$	$\hat{\beta}^+_{NP}$	$\hat{\upsilon}_{Zolotarev}$	$\hat{\alpha}^+_{MLE}$	$\hat{\beta}^+_{MLE}$
.01	**1.03** (0.15)	**0.97** (0.14)	0.09 (0.0002)	**1.2** (0.08)	**0.89** (0.079)

TABLE 1

Sampling mean in **bold** *and standard errors in parenthesis of the estimators of α^+, β^+, and υ in the CGMY model with theoretical values $\alpha^- = 0$, $\beta^+ = 1$, $\alpha^+ = 1$, and $\upsilon = .1$. Sample size is 100 paths.*

The results above shows that sometime a modestly efficient robust nonparametric method is preferably to a very efficient estimation method.

Appendix A: Technical Proofs

Proof of Lemma 3.1. The idea is to exploit the well-known decomposition of the Lévy process as a compound Poisson process $\widetilde{X}^\varepsilon$ plus an independent Lévy process $X^\varepsilon := X - \widetilde{X}^\varepsilon$ with compactly supported Lévy measure $\nu_\varepsilon(dx) := \mathbf{1}_{\{|x| \le \varepsilon\}}\, \nu(dx)$, for a suitable chosen $\varepsilon > 0$. Concretely, here

$$\widetilde{X}^\varepsilon_t = \sum_{i=1}^{N_t} \xi_i,$$

for a homogeneous Poisson process $\{N_t\}_{t \ge 0}$ with intensity $\lambda_\varepsilon := \nu(\{|x| > \varepsilon\})$ and for independent random variables $\{\xi_i\}$ with distribution $\frac{1}{\lambda_\varepsilon}\mathbf{1}_{\{|x|>\varepsilon\}}\nu(dx)$. Clearly,

$$\begin{aligned} \frac{1}{t}\,\mathbb{P}\left[X_t \ge y\right] = \frac{1}{t}\,\mathbb{P}\left[X^\varepsilon_t \ge y\right] e^{-\lambda_\varepsilon t} + \mathbb{P}\left[X^\varepsilon_t + \xi_1 \ge y\right] e^{-\lambda_\varepsilon t}(\lambda_\varepsilon) \\ + \sum_{n=2}^{\infty} \mathbb{P}\left[X^\varepsilon_t + \sum_{i=1}^{n} \xi_i \ge y\right] e^{-\lambda_\varepsilon t} \lambda_\varepsilon^n t^{n-1}. \end{aligned}$$

Then, we have

$$\left|\frac{1}{t}\,\mathbb{P}\,[X_t \geq y] - \nu([y,\infty))\right| \leq \frac{1}{t}\,\mathbb{P}\,[X^\varepsilon_t \geq y] + |\lambda_\varepsilon\,\mathbb{P}\,[X^\varepsilon_t + \xi_1 \geq y] - \nu([y,\infty))| + \nu([y,\infty))\,\lambda_\varepsilon t + \lambda^2_\varepsilon t.$$

The second term on the right hand side of this inequality can itself be decomposed as follows:

$$|\lambda_\varepsilon\,\mathbb{P}\,[X^\varepsilon_t + \xi_1 \geq y] - \nu([y,\infty))| \leq \int_{y-\eta}^{y+\eta} s(x)\,dx + \lambda_\varepsilon\,\mathbb{P}\,[|X^\varepsilon_t| \geq \eta] + \nu([y,\infty))\,\mathbb{P}\,[|X^\varepsilon_t| \geq \eta]$$

for each $\eta > 0$ such that $a - \eta > \varepsilon$. Since s is bounded off the origin, there exists a $k > 0$ such that

$$\int_{y-\eta}^{y+\eta} s(x)dx \leq k\,\eta$$

for all $y \in D$. Fix $0 < \eta < \frac{1}{T} \wedge a$ and $0 < \varepsilon < a - \eta$. Then,

$$\sup_{y\in D}\left|\frac{1}{t}\,\mathbb{P}\,[X_t \geq y] - \nu([y,\infty))\right| \leq \frac{1}{t}\,\mathbb{P}\,[X^\varepsilon_t \geq a] + \frac{k}{T} + \nu([a,\infty))\,\mathbb{P}\,[|X^\varepsilon_t| \geq \eta] + \lambda_\varepsilon\,\mathbb{P}\,[|X^\varepsilon_t| \geq \eta] + \nu([a,\infty))\,\lambda_\varepsilon t + \lambda^2_\varepsilon t.$$

Finally, since $\lim_{t\to 0}\frac{1}{t}\,\mathbb{P}\,[X^\varepsilon_t \geq a] = 0$ and $\lim_{t\to 0}\,\mathbb{P}\,[|X^\varepsilon_t| \geq \eta] = 0$, we can choose $\delta_T > 0$ sufficiently small to make each of the terms smaller than $1/T$ when $t < \delta_T$.

Proof of Theorem 4.1.

(i) Fix a Lévy density $s_0 \in \Theta_\alpha\,(L/2;[a,b])$ such that $s_0(x) > 0$, for all $x \in \mathbb{R}_0 := \mathbb{R}\backslash\{0\}$, and a constant $\kappa > 0$. Also, let $g : \mathbb{R} \to \mathbb{R}_+$ be a symmetric function with compact support $\mathbb{K}_g$, satisfying (3.5) with $L/2$ (instead of L). Moreover, the support of $x \to g\,(\kappa(x - x_0))$, denoted by $\mathbb{K}$, does not contain the origin and also,

$$s_0(x) - \kappa^{-\alpha} g\,(\kappa(x - x_0)) > 0, \quad \forall x \in \mathbb{R}_0.$$

Let

$$s_\theta(x) := s_0(x) + \theta T^{-\frac{\alpha}{2\alpha+1}}\, g\left(\kappa T^{\frac{1}{2\alpha+1}}(x - x_0)\right), \quad x \in \mathbb{R}_0,$$

and notice that $s_\theta \in \Theta$ whenever $|\theta| < \kappa^{-\alpha}$.

(ii) Without loss of generality we assume that $\mathbb{K} \cap [-1,1] = \emptyset$. We follow the notation in [29] (Section 33). Let $\mathbb{P}^{(T)}_\theta$ be the distribution (on $\mathbb{D}[0,T]$) of a Lévy process $\{X(t)\}_{0\leq t\leq T}$ with Lévy density s_θ (the other two parameters of the generating triplet remain constant). We proceed to prove that $\left\{\mathbb{P}^{(T)}_\theta : \theta \in (-\kappa^{-\alpha}, \kappa^{-\alpha})\right\}$ is LAN at $\theta = 0$ (see e.g. Definition II.2.1 in [19]. By Theorems 33.1 and 33.2 in [29], $\mathbb{P}^{(T)}_\theta \approx \mathbb{P}^{(T)}_0$ and the likelihood function, $L_\theta(\omega) := \frac{d\,\mathbb{P}^{(T)}_\theta}{d\,\mathbb{P}^{(T)}_0}(\omega)$ is given by

$$L_\theta(\omega) := \exp\left\{\int_0^T\int_{\mathbb{K}} \ln\left[1 + \frac{\theta T^{-\frac{\alpha}{2\alpha+1}}}{s_0(x)}\, g\left(\kappa T^{\frac{1}{2\alpha+1}}(x - x_0)\right)\right]\xi(dt,dx;\omega) - \theta T^{1-\frac{\alpha}{2\alpha+1}}\int_{\mathbb{K}} g\left(\kappa T^{\frac{1}{2\alpha+1}}(x - x_0)\right)dx\right\},$$

where $\xi(dt,dx;\omega)$ is the random measure on $\mathbb{R}_+ \times \mathbb{R}_0$ associated with the jumps of $\omega \in \mathbb{D}[0,T]$; that is,

$$\xi(A;\omega) := \#\{(t,x) : \Delta w_t := w_t - w_{t^-} = x\}, \quad A \subset \mathbb{R}_+ \times \mathbb{R}_0.$$

Under $\mathbb{P}_0^{(T)}$, ξ is a Poisson random measure with mean measure $s_0(x)dxdt$. We denote $\bar{\xi}(dt,dx;\omega) := \xi(dt,dx;\omega) - s_0(x)dxdt$. The likelihood $L_\theta(\omega)$ can be written as follows:

$$L_\theta(\omega) = \exp\left\{\theta\Delta_{_T} - \frac{\theta^2}{2}\sigma_{_T}^2 + r_{_T}(\theta)\right\},$$

where

$$\begin{aligned}
\Delta_{_T} &= T^{-\frac{\alpha}{2\alpha+1}} \int_0^T\!\!\int_{\mathbb{K}} s_0^{-1}(x) g\left(\kappa T^{\frac{1}{2\alpha+1}}(x-x_0)\right)\bar{\xi}(dt,dx),\\
\sigma_{_T}^2 &= T^{1-\frac{2\alpha}{2\alpha+1}} \int_{\mathbb{K}} s_0^{-1}(x) g^2\left(\kappa T^{\frac{1}{2\alpha+1}}(x-x_0)\right)dx,\\
r_{_T}(\theta) &= -\frac{\theta^2}{2} T^{-\frac{2\alpha}{2\alpha+1}} \int_0^T\!\!\int_{\mathbb{K}} s_0^{-2}(x) g^2\left(\kappa T^{\frac{1}{2\alpha+1}}(x-x_0)\right)\bar{\xi}(dt,dx)\\
&\quad + \int_0^T\!\!\int_{\mathbb{K}} R\left(\theta T^{-\frac{\alpha}{2\alpha+1}} s_0^{-1}(x) g\left(\kappa T^{\frac{1}{2\alpha+1}}(x-x_0)\right)\right)\xi(dt,dx),
\end{aligned}$$

and $R(u) := \ln(1+u) - u + \frac{u^2}{2}$. We want to prove that there are nomalizing constants $\varphi_T > 0$ such that

$$\mathcal{L}_{\mathbb{P}_0^{(T)}}(\varphi_T\Delta_{_T}) \xrightarrow{\mathfrak{D}} \mathcal{N}(0,1), \;\; \varphi_T^2\sigma_{_T}^2 \to 1, \;\; \text{and} \;\; r_{_T}(\theta) \overset{\mathbb{P}_0^{(T)}}{\to} 0$$

as $T \to \infty$. To prove the first limit, we invoke the CLT for Poisson integrals by verifying the Liapunov condition (see Theorem 1.1 and Remark 1.2 of [22]). Indeed, for $T > 1$, we have that

$$\begin{aligned}
&T^{-\frac{\alpha(2+\delta)}{2\alpha+1}} \int_0^T\!\!\int_{\mathbb{K}} s_0^{-2-\delta}(x) g^{2+\delta}\left(\kappa T^{\frac{1}{2\alpha+1}}(x-x_0)\right)(s_0(x))\,dxdt =\\
&\quad \kappa^{-1}T^{1-\frac{\alpha(2+\delta)}{2\alpha+1}-\frac{1}{2\alpha+1}} \int_{\mathbb{K}_g} s_0^{-1-\delta}(\kappa^{-1}T^{-\frac{1}{2\alpha+1}}u + x_0) g^{2+\delta}(u)\,du \overset{T\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Similarly, for large enough T,

$$\begin{aligned}
\mathrm{Var}(\Delta_{_T}) &= T^{-\frac{2\alpha}{2\alpha+1}} \int_0^T\!\!\int_{\mathbb{K}} s_0^{-2}(x) g^2\left(\kappa T^{\frac{1}{2\alpha+1}}(x-x_0)\right)(s_0(x))\,dxdt =\\
&\overset{T\to\infty}{\longrightarrow} \kappa^{-1}s_0^{-1}(x_0)\int_{\mathbb{K}_g} g^2(u)du.
\end{aligned}$$

Then, $\mathcal{L}_{\mathbb{P}_0^{(T)}}(\Delta_{_T}) \xrightarrow{\mathfrak{D}} \mathcal{N}(0,I_0^2)$ with $I_0^2 := \kappa^{-1}s_0^{-1}(x_0)\int_{\mathbb{K}_g} g^2(u)du$, and $\sigma_{_T}^2 \to I_0^2$. We now verify that $r_{_T}(\theta)$ vanishes in probability. Notice that the first term of $r_{_T}$ converges to 0 since its mean is 0 and its variance vanishes.
Similarly, the second term of $r_T(\theta)$ converges to 0 in probability because its mean and variance both goes to 0. Indeed, using that $|R(u)| \le |u|^3/3$, the

absolute value of its expectation satisfies

$$\left|\int_0^T\int_{\mathbb{K}} R\left(\theta T^{-\frac{\alpha}{2\alpha+1}}s_0^{-1}(x)g\left(\kappa T^{\frac{1}{2\alpha+1}}(x-x_0)\right)\right)(s_0(x))dxdt\right|$$
$$\leq \frac{|\theta|^3}{3}T^{1-\frac{3\alpha}{2\alpha+1}}\int_{\mathbb{K}} s_0^{-2}(x)g^3\left(\kappa T^{\frac{1}{2\alpha+1}}(x-x_0)\right)dx \overset{T\to\infty}{\longrightarrow} 0.$$

A similar reasoning applies to the variance. Therefore, $\{\mathbb{P}_\theta^{(T)}\}_{\theta\in(-\kappa^{-\alpha},\kappa^{-\alpha})}$ is Locally Asymptotically Normal (LAN) at $\theta=0$ (with the normalizing constants $\varphi_T := I_0^{-1}$).

(iii) By Theorem II.12.1 and Remark II.12.2 in [19], if $\hat\theta_T$ is any estimator of θ based on $\{X(t)\}_{0\leq t\leq T}$, then

$$\liminf_{T\to\infty}\sup_{|\theta|<\kappa^{-\alpha}}\mathbb{E}_\theta\left[\ell_0\left(I_0\left(\hat\theta_T-\theta\right)\right)\right]\geq B, \tag{A.1}$$

where $B := \mathbb{E}\left[\ell_0(Z)\chi_{[|Z|<I_0\kappa^{-\alpha}/2]}\right]$ and $Z\sim\mathcal{N}(0,1)$. Now, let $\hat s_T(\cdot)$ be an arbitrary estimator based on $\{X(t)\}_{0\leq t\leq T}$ and let

$$\hat\theta_T := T^{\frac{\alpha}{2\alpha+1}}g^{-1}(0)\left(\hat s_T(x_0)-s_0(x_0)\right).$$

Since $\theta = T^{\frac{\alpha}{2\alpha+1}}g^{-1}(0)\left(s_\theta(x_0)-s_0(x_0)\right)$, we can write

$$g(0)\left(\hat\theta_T-\theta\right) = T^{\frac{\alpha}{2\alpha+1}}\left(\hat s_T(x_0)-s_\theta(x_0)\right).$$

If we take $\ell_0(u) := \ell\left(g(0)I_0^{-1}u\right)$, (A.1) becomes:

$$\begin{aligned}B &\leq \liminf_{T\to\infty}\sup_{|\theta|<\kappa^{-\alpha}}\mathbb{E}_\theta\left[\ell_0\left(I_0\left(\hat\theta_T-\theta\right)\right)\right]\\ &= \liminf_{T\to\infty}\sup_{|\theta|<\kappa^{-\alpha}}\mathbb{E}_\theta\left[\ell\left(T^{\frac{\alpha}{2\alpha+1}}\left(\hat s_T(x_0)-s_\theta(x_0)\right)\right)\right].\end{aligned}$$

Since $\{s_\theta:\theta\in(-k^{-\alpha},k^{-\alpha})\}\subset\Theta$,

$$\liminf_{T\to\infty}\sup_{s\in\Theta}\mathbb{E}_s\left[\ell\left(T^{\frac{\alpha}{2\alpha+1}}\left(\hat s_T(x_0)-s(x_0)\right)\right)\right]\geq B, \tag{A.2}$$

where

$$B := 2^{-3/2}\pi^{-1/2}\int_{|z|<I_0\kappa^{-\alpha}/2}\ell(g(0)I_0^{-1}z)e^{-z^2/2}dz. \tag{A.3}$$

This implies (4.1) because the lower bound B does not depend on the family of estimators $\hat s_T$. Indeed, for each $\varepsilon>0$, let $\hat s_T^{(\varepsilon)}$ be such that

$$\begin{aligned}&\sup_{s\in\Theta}\mathbb{E}_s\left[\ell\left(T^{\frac{\alpha}{2\alpha+1}}\left(\hat s_T^{(\varepsilon)}(x_0)-s(x_0)\right)\right)\right]\\ &< \inf_{\hat s_T}\sup_{s\in\Theta}\mathbb{E}_s\left[\ell\left(T^{\frac{\alpha}{2\alpha+1}}\left(\hat s_T(x_0)-s(x_0)\right)\right)\right]+\varepsilon.\end{aligned}$$

Taking the lim inf as $T\to\infty$ on both sides, we obtain (4.1) since ε is arbitrary.

Proof of Corollary 4.2.

(i) Following the same reasoning as in Theorem 4.1, we first prove that for any family of estimators $\{\hat{s}_T\}_{T>0}$ and arbitrary points $\{x_T\}_{T>0} \subset (a,b)$,

$$\liminf_{T\to\infty} \sup_{s\in\Theta} \mathbb{E}_s \left[\ell\left(T^{\frac{\alpha}{2\alpha+1}}\left(\hat{s}_T(x_T) - s(x_T)\right)\right)\right] \geq C \tag{A.4}$$

for some constant $C > 0$, which is independent of the family of estimators and of the points. Fix a Lévy density $s_0 \in \Theta_\alpha\left(L/2; [a,b]\right)$ such that $s_0(x) > 0$ for all $x \in \mathbb{R}_0 := \mathbb{R}\backslash\{0\}$, and a constant $\kappa > 0$. Again, let $g : \mathbb{R} \to \mathbb{R}_+$ be a symmetric function with compact support $\mathbb{K}_g$, satisfying (3.5) with $L/2$ (instead of L). Moreover, for any $y \in (a,b)$, the support of $x \to g\left(\kappa(x-y)\right)$ does not contain the origin and

$$s_0(x) - \kappa^{-\alpha} g\left(\kappa(x-y)\right) > 0, \quad \forall x \in \mathbb{R}_0.$$

Let

$$s_{\theta,T}(x) := s_0(x) + \theta T^{-\frac{\alpha}{2\alpha+1}} g\left(\kappa T^{\frac{1}{2\alpha+1}}(x - x_T)\right), \quad x \in \mathbb{R}_0,$$

for $|\theta| < \kappa^{-\alpha}$. Let $\mathbb{P}_\theta^{(T)}$ be the distribution (on $\mathbb{D}[0,T]$) of a Lévy process $\{X(t)\}_{0\leq t\leq T}$ with Lévy density $s_{\theta,T}$. Following the proof of Theorem 4.1, $\{\mathbb{P}_\theta^{(T)} : \theta \in (-\kappa^{-\alpha}, \kappa^{-\alpha})\}$ is LAN at $\theta = 0$ with the normalizing constants

$$\varphi_T := \kappa^2 \left(\int_{\mathbb{K}_g} s_0^{-1}(\kappa^{-1}T^{-\frac{1}{2\alpha+1}}u + x_T) g^2(u)\, du\right)^{-2},$$

where $\mathbb{K}_g$ denotes the support of g and it is being assumed that $[-1,1] \cap \cup_{y\in[a,b]}\{y + \kappa^{-1}\mathbb{K}_g\} = \emptyset$. Observe that there is an $m > 0$ for which $\inf_{T\geq 1} \varphi_T \geq m$.

(ii) By Theorem II.12.1 and Remark II.12.2 in Ibragimov & Has'minskii (1981), for any $\delta > 0$,

$$\liminf_{T\to\infty} \sup_{|\theta|<\delta\varphi_T} \mathbb{E}_\theta \left[\ell_0\left(\varphi_T^{-1}\left(\hat{\theta}_T - \theta\right)\right)\right] \geq C, \tag{A.5}$$

where $C := \mathbb{E}\left[\ell_0(Z)\chi_{[|Z|<\delta/2]}\right]$ and $Z \sim \mathcal{N}(0,1)$. Since $\ell_0(|y|)$ is increasing in y,

$$\liminf_{T\to\infty} \sup_{|\theta|<\delta\varphi_T} \mathbb{E}_\theta \left[\ell_0\left(m^{-1}\left(\hat{\theta}_T - \theta\right)\right)\right] \geq C. \tag{A.6}$$

Now, setting,

$$\hat{\theta}_T := T^{\frac{\alpha}{2\alpha+1}} g^{-1}(0)\left(\hat{s}_T(x_T) - s_0(x_T)\right),$$

it follows that

$$\sup_{s\in\Theta} \mathbb{E}_s \left[\ell\left(T^{\frac{\alpha}{2\alpha+1}}\left(\hat{s}_T(x_T) - s(x_T)\right)\right)\right] \geq \sup_{|\theta|<\delta\varphi_T} \mathbb{E}_\theta \left[\ell\left(g(0)\left(\hat{\theta}_T - \theta\right)\right)\right].$$

Taking lim inf as $T \to \infty$, (A.4) is obtained with

$$C = 2^{-3/2}\pi^{-1/2} \int_{|z|<\delta/2} \ell(g(0)\, m\, z) e^{-z^2/2} dz. \tag{A.7}$$

(iii) To obtain (4.2), for each $\varepsilon > 0$, let $\hat{s}_T^{(\varepsilon)} \in \Theta$ and $x_T^{(\varepsilon)} \in (a,b)$ be such that

$$\sup_{s\in\Theta} \mathbb{E}_s\left[\ell\left(T^{\frac{\alpha}{2\alpha+1}}\left(\hat{s}_T^{(\varepsilon)}\left(x_T^{(\varepsilon)}\right) - s\left(x_T^{(\varepsilon)}\right)\right)\right)\right] \leq \inf_{x\in(a,b)} \inf_{\hat{s}_T} \sup_{s\in\Theta} \mathbb{E}_s\left[\ell\left(T^{\frac{\alpha}{2\alpha+1}}\left(\hat{s}_T(x) - s(x)\right)\right)\right] + \varepsilon.$$

Next, take the lim inf as $T \to \infty$ on both sides above and apply (A.4).

(iv) We now prove (4.4). Fix a measurable estimator $\hat{s}_T$ and a $s \in \Theta$. By Fubini's Theorem,

$$\mathbb{E}_s\left[\int_a^b \left(\hat{s}_T(x) - s(x)\right)^2 dx\right] = \int_a^b \mathbb{E}_s\left[\left(\hat{s}_T(x) - s(x)\right)^2\right] dx.$$

Now, for each $\varepsilon > 0$, there exists an $x_0^{(\varepsilon)} \in (a,b)$ satisfying

$$\frac{1}{b-a}\int_a^b \mathbb{E}_s\left[\left(\hat{s}_T(x) - s(x)\right)^2\right] dx \geq \mathbb{E}_s\left[\left(\hat{s}_T\left(x_0^{(\varepsilon)}\right) - s\left(x_0^{(\varepsilon)}\right)\right)^2\right] - \varepsilon.$$

Then,

$$\begin{aligned}\frac{1}{b-a}\sup_{s\in\Theta} \mathbb{E}_s\left[\int_a^b \left(\hat{s}_T(x) - s(x)\right)^2 dx\right] &\geq \sup_{s\in\Theta} \mathbb{E}_s\left[\left(\hat{s}_T(x_0^{(\varepsilon)}) - s(x_0^{(\varepsilon)})\right)^2\right] - \varepsilon \\ &\geq \inf_{x\in(a,b)} \sup_{s\in\Theta} \mathbb{E}_s\left[\left(\hat{s}_T(x) - s(x)\right)^2\right] - \varepsilon.\end{aligned}$$

Letting $\varepsilon \to 0$, (4.4) becomes a consequence of (4.2) with $\ell(u) = u^2$.

Appendix B: Figures

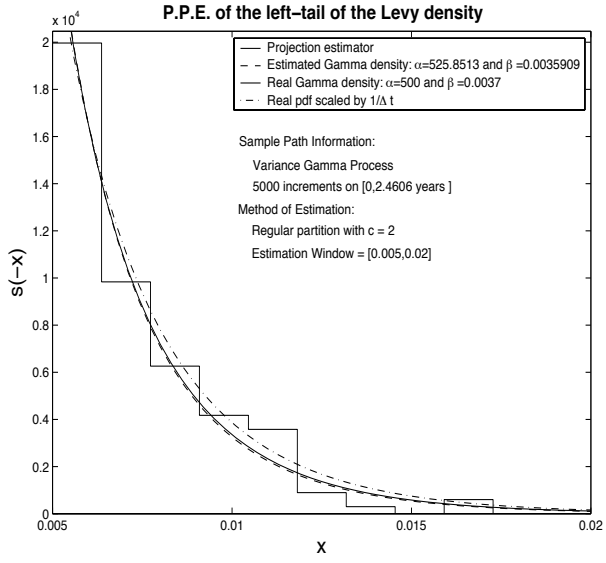


FIG 1. *Penalized projection estimation of the left-tail of the variance gamma Levy density.*

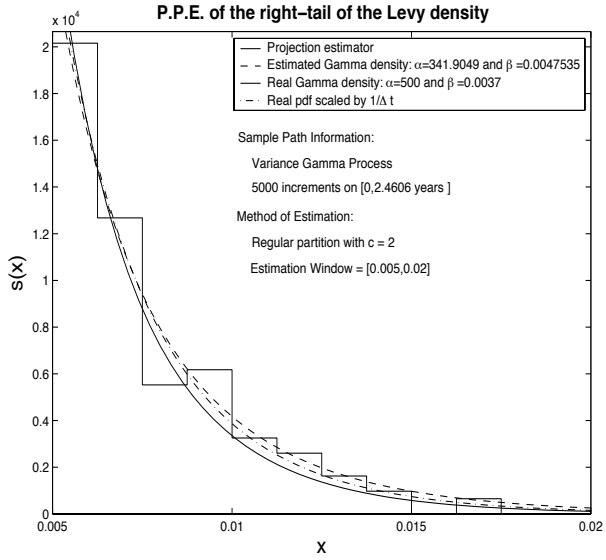


FIG 2. *Penalized projection estimation of the right-tail of the variance Gamma Levy density.*

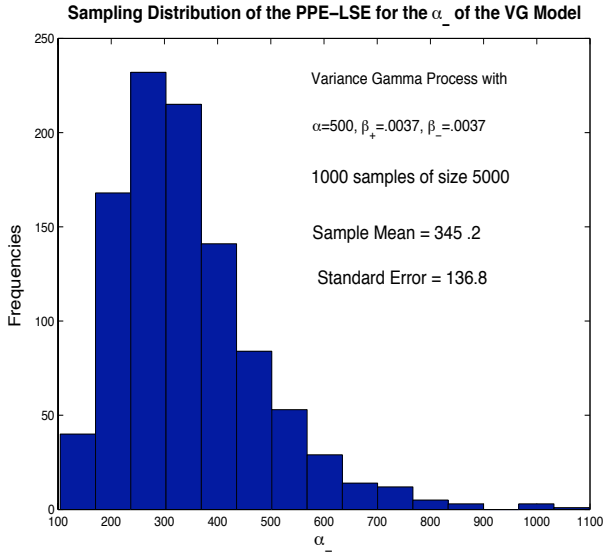


FIG 3. *Sampling Distribution for the Estimates of α^- obtained from the PPE and the LSE method.*

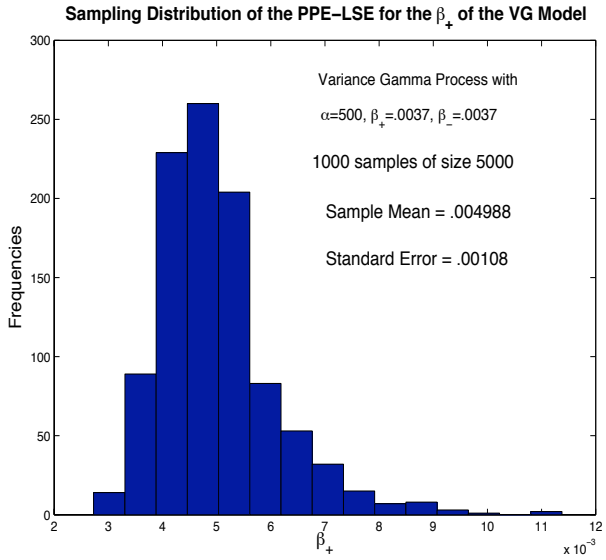


FIG 4. *Sampling Distribution for the Estimates of β^+ obtained from the PPE and the LSE method.*

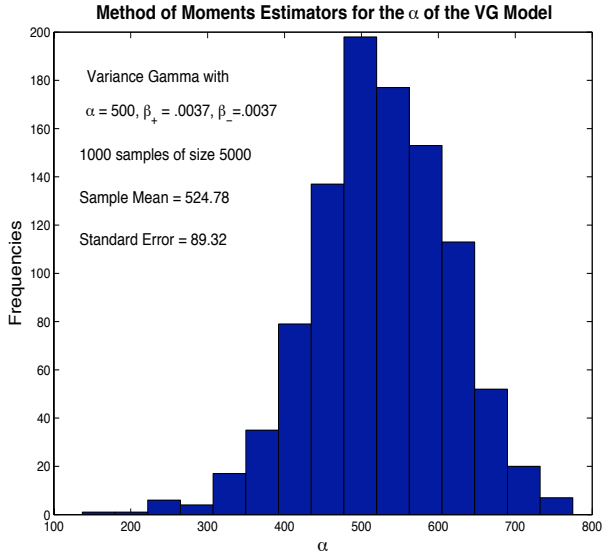


FIG 5. *Sampling Distribution for the Estimator of α obtained by the Method of Moments.*

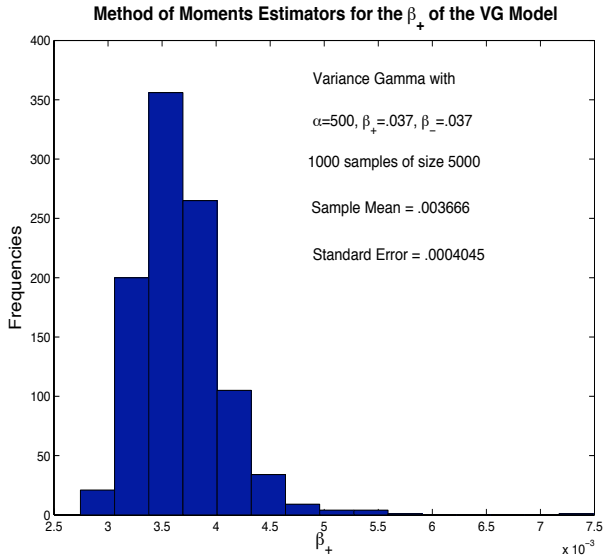


FIG 6. *Sampling Distribution for the Estimator of β^+ obtained by the Method of Moments.*

References

[1] BARNDORFF-NIELSEN O. E. and SHEPHARD N. (2001). Modelling by Lévy processess for financial economics. *Lévy Processes. Theory and Applications* by Barndorff-Nielsen, O.E., Mikosch, T. and Resnick, S.I. 283–318

[2] BARRON A., BIRGÉ L., and MASSART P. (1995). Model selection via penalization. *Technical Report*, **54**, Université Paris-Sud.

[3] BARRON A., BIRGÉ L., and MASSART P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, **113**, 301–413.

[4] BIRGÉ L., and MASSART P. (1994). Minimum contrast estimation on sieves. *Technical Report*, **34**, Université Paris-Sud.

[5] BIRGÉ L., and MASSART P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, 55–87.

[6] CARR P., GEMAN, H., MADAN, D., and YOR M. (2002). The fine structure of asset returns: An empirical investigation. *Journal of Business*, 305–332.

[7] CARR P., GEMAN, H., MADAN, D., and YOR M. (2003). Stochastic volatility for Lévy processes. In *Mathematical Finance*, **13**, 345–382.

[8] CARR P., MADAN, D., and CHANG E. (1998). The variance Gamma process and option pricing. *European Finance Review*, **2**, 79–105.

[9] CARR P., and WU, L. (2005). Time-changed Levy processes and option pricing. *Journal of Financial Economics*, **71.1**, 113–141.

[10] CHUNG K. L. (2001). *A course in Probability Theory*. Academic Press.

[11] CONT R., and TANKOV P. (2003). *Financial modelling with Jump Processes*. Chapman & Hall.

[12] DEVORE R. A., and LORENTZ G. G. (1993). *Constructive Approximation*. Springer-Verlag.

[13] FIGUEROA-LÓPEZ J. E. (2004). Nonparametric estimation of Lévy processes with a view towards mathematical finance. *PhD Thesis*. Georgia Institute of Technology, Atlanta, GA 30332, April 2004. http://etd.gatech.edu.No.etd-04072004-122020.

[14] FIGUEROA-LÓPEZ J. E. (2007). On the non-parametric estimation of some time-changed Lévy models. In preparation.

[15] FIGUEROA-LÓPEZ J. E. (2008). Small-time moment asymptotics for Lévy processes. *Statistics and Probability Letters*. DOI:10.1016/j.spl.2008.07.012.

[16] FIGUEROA-LÓPEZ J. E. and HOUDRÉ C. (2006). Risk bounds for the non-parametric estimation of Lévy processes. *IMS Lecture Notes - Monograph Series. High Dimensional Probability*, **51**, 96–116.

[17] FIGUEROA-LÓPEZ J. E. and HOUDRÉ C. (2008). Small-time expansions for the transition distributions of Lévy processes. Available at *arXiv:0809.0849v1 [math.PR]*.

[18] GRENANDER U. (1981). *Abstract Inference*. John Wiley & Sons.

[19] IBRAGINOV I. A. and HAS'MINSKII R. Z. (1981). *Statistical Estimation. Asymptotic Theory*. Springer-Verlag, Berlin, New York, Heidelberg.

[20] Jacod J. (2007). Asymptotic properties of power variations of Lévy processes. *ESAIM:P&S*, **11**, 173–196.
[21] Knotz S., Kozubowski T. J., and Podgórski K. (2001). *The Laplace Distribution and Generalizations: A revisit with Applications to Communications, Economics, Engineering, and Finance*. Birkhauser, Boston.
[22] Kutoyants Y. A. (1998). *Statistical Inference for Spatial Poisson Processes*. Springer.
[23] Madan D. B. and Seneta E. (1990). The variance Gamma model for share market returns. *Journal of Business*, **63**, 511–524.
[24] Prause K. (1999). The generalized hyperbolic model: estimation, financial derivatives, and risk measures. *Ph.D. Thesis*. University of Freiburg.
[25] Reynaud-Bouret P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory & Related Fields*, **126.1**, 103–153.
[26] Rosiński J. (2001). Series representations of Lévy processes from the perspective of point processes. In *Lévy processes-Theory and Applications*, 401–415.
[27] Rosiński J. (2007). Tempering stable processes. *Stochastic processes and their applications*, **117**, 677–707.
[28] Rüschendorf L. and Woerner J. (2002). Expansion of transition distributions of Lévy processes in small time. *Bernoulli*, **8**, 81–96.
[29] Sato K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press.
[30] Schoutens W. (2003). *Lévy Processes: Pricing Financial Derivatives*. John Wiley.
[31] Todorov V. (2005). Econometric analysis of jump-driven stochastic volatility models. *Duke University*, Working Paper.
[32] Woerner J. (2003). Variational sums and power variation: a unifying approach to model selection and estimation in semimartingale models. *Statistics and Decisions*, **21**, 47–68.

IMS Collections
Volume Title
Vol. 0 (0000) 147–173
© Institute of Mathematical Statistics, 0000
arXiv: math.PR/0000000

On the Estimation of Symmetric Distributions under Peakedness Order Constraints

Javier Rojo[1,*] **and José Batún-Cutz**

Rice University and Universidad Autónoma de Yucatán, Mérida, México

Abstract: Consider distribution functions F and G and suppose that F is *more peaked* about a than G is about b. The problem of estimating F or G, or both, when F and G are symmetric, arises quite naturally in applications. The empirical distribution functions F_n and G_m will not necessarily satisfy the order constraint imposed by the experimental conditions. Rojo and Batun-Cutz (2007) proposed some estimators that are strongly uniformly consistent when both m and n tend to infinity. However the estimators fail to be consistent when only either m or n tend to infinity. Here estimators are proposed that circumvent these problems and the asymptotic distribution of the estimators is delineated. A simulation study compares these estimators in terms of Mean Squared Error and Bias behavior with their competitors.

Contents

1 Introduction 147
2 New Estimators and Their Finite Sample Properties 151
2.1 Definition of the New Estimators 152
2.2 Bias Functions 153
2.3 Estimators as Projections onto Appropriate Convex Spaces 155
2.4 Peakedness Order of New and Previous Estimators 155
3 Asymptotics 156
3.1 Strong Uniform Convergence 157
3.2 Weak Convergence 158
4 Example with Sib-pair Data: An Illustration 162
5 Simulation Work 162
6 Conclusions 163
Acknowledgements 172
References 172

1. Introduction

The concept of stochastic order was pioneered by Lehmann (1955), and applications to hypotheses testing were discussed in Lehmann (1959), henceforth referred to as TSH-1. Lehmann and Rojo (1992) provided characterizations of stochastic

*Corresponding author
[1]Department of Statistics, MS-138; Rice University; 6100 Main Street; Houston, TX 77005; and School of Mathematics; Universidad Autónoma de Yucatán Mérida México e-mail: jrojo@rice.edu; jbatun@uady.mx

AMS 2000 subject classifications: Primary 62G30, 62G20; secondary 60F05
Keywords and phrases: Partial orders, Nonparametric statistics, Stochastic order, Empirical process, Weak convergence

ordering in terms of the maximal invariant with respect to the group of monotone transformations, and connections with other partial orderings were provided. Since the publication of TSH-1, there has been a large number of papers discussing various types of stochastic orders and their properties. Thus, one finds a large literature on stochastic orders in Economics (e.g. first-, second-, third-order stochastic dominance), reliability (e.g. IFR, IFRA, NBU, etc.), and applied probability (e.g. Laplace transform and dispersive orders). Marshall and Olkin (2007) and Shaked and Shantikumar (2007) are excellent references to the literature on stochastic orders.

The attention to this area of statistics and applied probability is well deserved. These concepts arise naturally in many applications in engineering, survival analysis, biology, economics, etc.

In corrosion engineering, for example, the times until pitting of metals immersed in a corrosive environment are measured under different solution corrosivities to discern the impact of the solution acidity on the pitting corrosion times. Shibata and Takeyama (1977) present data which strongly supports the belief that the times until pitting should be shorter in some sense, for the more corrosive environment. In toxicity studies, cells are grown in environments containing different levels of toxic materials (e.g. Arenaz *et al* (1992)). Invariably, the data supports the intuitive notion that the stronger the toxic solution is, the shorter the lifetimes of the organisms.

Another set of examples arises from clinical trials. This is illustrated by a clinical trial run to evaluate the efficiency of maintenance chemotherapy for acute myelongeneous leukemia (AML). The trial was conducted at Stanford University (Embury *et al* (1977)). After reaching a state of remission through treatment by chemotherapy, the patients who entered the study were randomized into two groups. The first group received maintenance chemotherapy; the second group did not. One would then expect that in this case, the survival times in the control group would be stochastically smaller than those in the first group.

Stochastic ordering, together with failure rate ordering, and monotone likelihood ratio ordering, are examples of *location* orderings. There are situations, however, when the interest lies in comparing distributions based on their *spread* rather than on their location.

Various concepts of spread, concentration, or dispersion have appeared in the literature. For example, Brown and Tukey (1946), Fraser (1957), Bickel and Lehmann (1979), Lehmann (1988), Doksum (1969), and Shaked (1980), define F to be more dispersive than G , denoted as $F >_d G$, if, for every $u > v$,

$$F^{-1}(u) - F^{-1}(v) \geq G^{-1}(u) - G^{-1}(v). \tag{1.1}$$

Shaked (1982), Bartoszewicz (1985a, 1985b, 1986), Oja (1981), and Rojo and He (1991), among others, have discussed various characterizations and properties of the dispersive order. Doksum (1969) utilized this concept to study power properties of rank tests, and showed that the power of certain rank tests is isotonic with respect to this order. Rojo (1995b, 1999) considered the problem of estimating the quantile function F^{-1} and the distribution function F when $F <_d G$, and the asymptotic theory of the resulting estimators was delineated. Rojo and Wang (1994) also showed that the power of tests based on L-statistics is isotonic with respect to the dispersive order. For other properties of the dispersive order, and connections with other partial orderings, see Bickel and Lehmann (1979), Proschan (1965), Karlin (1968), Shaked (1980, 1982), and Schweder (1982). When F and G

are assumed symmetric, (1.1) can be seen to be equivalent to

$$F^{-1}(u) - F^{-1}(1/2) \geq (\leq)\ G^{-1}(u) - G^{-1}(1/2)$$

depending on whether $u \geq (\leq)\ 1/2$.

Birnbaum (1948) proposed a different concept of dispersion based on the distribution functions rather than on the quantile functions. According to Brinbaum, the distribution function F is more peaked about the point a than the distribution function G is about the point b if, for all $x \geq 0$,

$$F((x+a)^-) - F(-x+a) \geq G((x+b)^-) - G(-x+b), \tag{1.2}$$

where $h(x^-) = \lim_{\epsilon \downarrow 0} h(x-\epsilon)$. We will write $F >_p G$ whenever (1.2) holds. It is easy to see that the condition (1.2) is equivalent to

$$\begin{array}{ll} F(x^-) \geq G(x^-) & \text{for } x \geq 0 \\ F(x) \leq G(x) & \text{for } x < 0. \end{array} \tag{1.3}$$

whenever F and G are symmetric about the point 0.

When F and G are continuous, it is easy to see that (1.2) is equivalent to requiring that $|X-a|$ be stochastically smaller than $|Y-b|$, and, although in general $F <_d G \not\Rightarrow F >_p G$ and $F >_p G \not\Rightarrow F <_d G$, it is easy to verify that $F <_d G \Rightarrow F >_p G$, when F and G are symmetric and continuous. When a and b in (1.2) are, respectively, the means of F and G, the condition (1.2) implies the obvious order on the variances of F and G.

An interesting example from statistical genetics, discussed in Rojo *et al* (2007), illustrates the importance of this concept in applications. Haseman-Elston (1972) proposed a regression model to assess the effect of a candidate gene on a phenotype when using sib-paired data. There have been some modifications of the initial proposal (see e.g. Elston *et al.* (2000)). The original model, Haseman-Elston (1972), represents the expected value of the squared phenotypic differences as a linear function of the proportion of alleles shared identical by descent (IBD) at the locus of interest. Let λ_i represent the proportion of alleles shared identical by descent ($\lambda_i = 0, \frac{1}{2}$, or 1). The Haseman and Elston (1972) regression model may then be written as follows: $E(X_i|\lambda_i) = \alpha + \beta\lambda_i$, where X_i represents the squared sib-pair difference for the i^{th} sib-pair conditional on λ_i. Writing $Z_{1i} = \theta + g_{1i} + \varepsilon_{1i}$ and $Z_{2i} = \theta + g_{2i} + \varepsilon_{2i}$ where Z_{1i} and Z_{2i} represent, respectively, the phenotype values for siblings one and two, and where θ is the population mean, and g_{ij} and ε_{ij} are the genetic and the residual effects, respectively, the model is then represented as

$$E(X_j|\lambda_j) \quad = \quad \eta_\varepsilon^2 + 2(1-\lambda_j)\eta_g^2$$

where, $\eta_\varepsilon^2 = E((\varepsilon_{1i} - \varepsilon_{2i})^2)$ and η_g^2 represents the variance in the trait due to allelic variation at the locus of interest. As a consequence of linkage between the candidate gene and the phenotype, siblings sharing two alleles IBD at the locus of interest will tend to be more similar than siblings sharing one allele IBD, and siblings sharing one allele IBD will in turn be more similar than siblings sharing no alleles IBD. It is then clear that phenotypical similarity of sibs within the same pair is being measured in terms of the spread of the distribution of the differences of the siblings' phenotypical measurements.

Existing sib-paired data illustrates very clearly that the distribution functions of sib-pair differences are symmetrically distributed. This will happen, for example, if

$(X - \mu_X, Y - \mu_Y)$ has the same distribution as $(\mu_X - X, \mu_Y - Y)$, as it happens under the assumption of a bivariate normal distribution, and if the means μ_X and μ_Y are equal, then the sib-pair differences are symmetrically distributed. When the candidate gene is linked to the phenotype of interest, the cumulative distributions of the differences within sib-pairs are ordered by peakedness. This is illustrated by sib-paired data on plasma Lipoprotein (a) data. Figure 1 shows the empirical

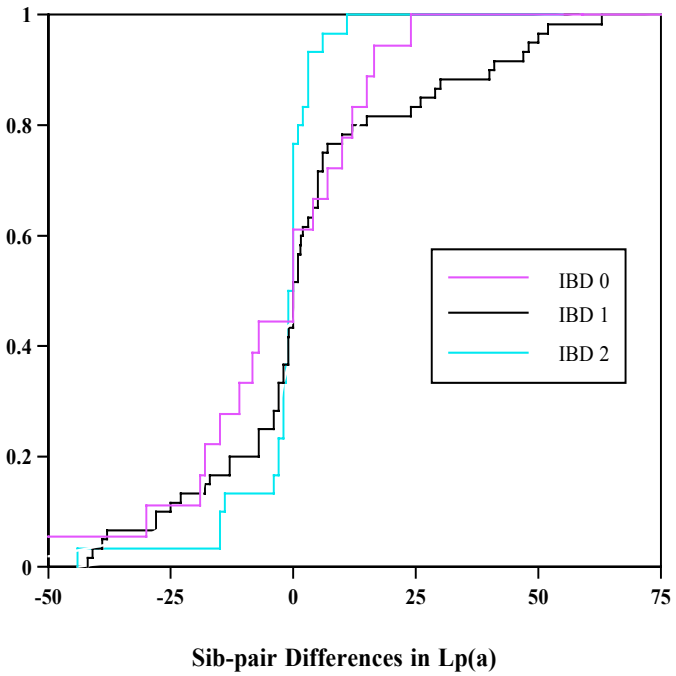


FIG 1. *Empirical distribution functions of phenotypic differences for the sib-pair data.*

distribution functions for plasma Lipoprotein (a) differences within sib-pairs for a sample of Caucasian individuals from the Dallas metroplex area. The pairs of siblings were classified into groups according to the number of shared alleles identical by descent.

Note that the assumptions of symmetry and peakedness are close to being satisfied, but the plots also show areas where these characteristics do not hold. We will illustrate our estimators later in section 4, by computing them for this example.

The points a and b about which peakedness of F and G will obtain, will be assumed known throughout this work. In the linkage example to be considered in section 4, the assumption of known a and b can be justified under the assumption of bivariate normality of the siblings' phenotypes with equal means. This is a common assumption in the literature. Thus, irrespective of whether a and b are known or unknown, the difference of the phenotypes is always symmetric about zero. Dropping the assumption of bivariate normality of the sib-pairs phenotypes, existing models, see *e.g.* Liu (1988) Table 15.7, yield a zero mean for the phenotypic differences. We, therefore, will assume that a and b are zero.

The goals of this paper are to develop estimators for symmetric F and G, which satisfy (1.2), and to delineate their asymptotic theory.

Under the assumption that F and G are discrete distributions satisfying (1.2), El Barmi and Rojo (1997) provided the nonparametric maximum likelihood estimators

of F and G and tests were given to test the hypothesis of homogeneity of F and G against the alternative that F and G satisfy (1.2). Rojo, Batun, and Durazo (2007) proposed estimators for continuous F and G, when (1.2) holds and the case of censored data was also considered, but without the symmetry assumption. Rojo and Batun-Cutz (2007) proposed estimators for symmetric F and G when (1.2) holds using results from Schuster (1975), and the asymptotic theory was delineated for the case when both n and $m \to \infty$. El Barmi and Mukerjee (2008), following the ideas in Rojo (2004) and Rojo and Batun (2007), proposed estimators which are consistent for F (G) and their asymptotic theory was developed. Unfortunately, the proofs of their asymptotic results for the estimators of F and G depend on letting **both** n and m increase to infinity. The purpose of this paper is to consider modifications of the estimators proposed by Rojo and Batun-Cutz (2007) that yield consistent estimators for F (G) when only n (m) $\to \infty$. The asymptotic distribution theory is considered and a simulation study compares the estimators to the estimator of El Barmi and Mukerjee (2008).

The organization of this paper is as follows: Sections 2 proposes the estimators and finite sample properties are discussed. Section 3 delineates the asymptotic theory showing that the estimators are strongly and uniformly consistent and their asymptotic theory is developed. Section 4 illustrates the new estimators using the sib-pair data, and section 5 discusses the results of computer simulations which compare the bias and mean squared error of the new estimators with the bias and mean squared error of the estimators of Rojo and Batun-Cutz (2007) and El Barmi and Mukerjee (2008).

Although the estimators proposed in Rojo and Batun-Cutz (2007) have larger absolute bias than the estimators proposed here, the selection of the better estimators based on Mean Squared Error (MSE) behavior is not as clear. Whereas the new estimators have smaller MSE in a neighborhood of zero, the estimators of Rojo and Batun-Cutz have smaller MSE in the tails of the distributions, and the region of the support of the distribution where the latter estimators behave better seems to increase as the tail-heaviness of the distributions increase.

2. New Estimators and Their Finite Sample Properties

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent random samples from the symmetric distributions (about 0) F and G respectively, and let F_n and G_m be the empirical distribution functions based on $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$. Suppose than $F >_p G$. Rojo and Batun-Cutz (2007) considered the problem of the estimation of F and G under the peakedness restriction and proposed the following strongly uniformly consistent estimators

$$F^1_{n,m} = \Phi_1(\Phi_2(F_n, \Phi_1(G_m))) \tag{2.1}$$

$$F^2_{n,m} = \Phi_2(\Phi_1(F_n), \Phi_1(G_m)), \tag{2.2}$$

where Φ_1 and Φ_2 are operators defined by

$$\Phi_1(f)(x) = \frac{1}{2}(f(x) + 1 - f(-x^-)), \text{ and}$$

$$\Phi_2(f,g)(x) = \begin{cases} \max\{f(x), g(x)\} & \text{if} \quad x \geq 0 \\ \min\{f(x), g(x)\} & \text{if} \quad x < 0. \end{cases}$$

Note that the operator Φ_1 symmetrizes the function f, Schuster (1975), and the operator Φ_2 imposes the "stochastic order" restriction (see, e.g., Lo (1987), Rojo and Ma (1996), and Rojo (2004)). Unfortunately the estimators $F^i_{n,m}$, for $i = 1, 2$ do not converge to F when only $n \to \infty$. This follows since, for example, for $F^2_{n,m}$ when $x > 0$ and $\varepsilon > 0$,

$$\lim_{n\to\infty} P[F^2_{n,m}(x) - F(x) > \varepsilon] \quad \geq \quad P[\Phi_1(G_m(x)) - F(x) > \varepsilon] > 0.$$

This is a drawback of $F^2_{n,m}$ that is also shared by $F^1_{n,m}$ and $G^i_{n,m}$ for $i = 1, 2$, and the strong uniform consistency of these estimators requires that both m and n tend to infinity. To circumvent this problem, new estimators are proposed here.

2.1. Definition of the New Estimators

Let $\widehat{F}_n = \Phi_1(F_n)$ and $\widehat{G}_m = \Phi_1(G_m)$ be the symmetrized empirical distribution functions (Schuster, 1975). Then the empirical distribution function, and the symmetrized empirical distribution function of the combined samples are defined as follows:

$$C_{n,m} = \frac{n}{m+n}F_n + \frac{m}{n+m}G_m \text{ and}$$
$$\widehat{C}_{n,m} = \Phi_1(C_{n,m}) = \frac{n}{m+n}\widehat{F}_n + \frac{m}{n+m}\widehat{G}_m \tag{2.3}$$

respectively. Then our new estimators for F and G are

$$\widehat{F}^1_{n,m} = \Phi_1(\Phi_2(F_n, C_{n,m})), \tag{2.4}$$

$$\widehat{G}^1_{n,m} = \Phi_1(\Phi^*_2(G_m, C_{n,m})), \tag{2.5}$$

$$\widehat{F}^2_{n,m} = \Phi_2(\Phi_1(F_n), \Phi_1(C_{n,m})), \text{ and} \tag{2.6}$$

$$\widehat{G}^2_{n,m} = \Phi^*_2(\widehat{G}_m, \widehat{C}_{n,m}), \tag{2.7}$$

where

$$\Phi^*_2(f, g)(x) = \begin{cases} \min\{f(x), g(x)\} & \text{if} \quad x \geq 0 \\ \max\{f(x), g(x)\} & \text{if} \quad x < 0. \end{cases}$$

Note that the estimators $\widehat{F}^1_{n,m}$ and $\widehat{G}^1_{n,m}$ first impose the constraint of "stochastic order" by requiring that the estimator of F (G) be larger (smaller) than $C_{n,m}$ for $x \geq 0$ and smaller (larger) than $C_{n,m}$ for $x < 0$. The second requirement of symmetry is then imposed by the operator Φ_1. By contrast, the estimators $\widehat{F}^2_{n,m}$ and $\widehat{G}^2_{n,m}$, first impose the constraint of symmetry and then, through the operator Φ_2, the constraint of "stochastic order" is imposed.

El Barmi and Mukerjee (2008) proposed estimators for F and G when $F <_p G$. In our notation and making the appropriate change for the case $F >_p G$, their estimator for F is given, for $x \geq 0$, by

$$F^*_{nm}(x) \quad = \quad \frac{1}{2}(1 + \max\left\{F_n(x) - F_n(-x^-), C_{nm}(x) - C_{nm}(-x^-)\right\}).$$

This estimator is the same as our estimator $\widehat{F}^2_{n,m}$ since for $x \geq 0$,

$$\begin{aligned}\widehat{F}^2_{n,m}(x) &= \max\left\{\frac{1}{2}(1+F_n(x)-F_n(-x^-)), \frac{1}{2}(1+C_{nm}(x)-C_{nm}(-x^-))\right\} \\ &= \frac{1}{2}+\frac{1}{2}\max\left\{F_n(x)-F_n(-x^-), C_{nm}(x)-C_{nm}(-x^-)\right\}) \\ &= F^*_{nm}(x).\end{aligned}$$

Therefore, by symmetry, $\widehat{F}^2_{n,m} = F^*_{nm}$.

2.2. Bias Functions

The operator Φ_1 does not introduce any bias in the "symmetrization" procedure. In fact, it is well known that $\widehat{F}_n$ and $\widehat{G}_m$ are unbiased estimators for F and G, and have smaller variance than F_n and G_m respectively. However, the operators Φ_2 and Φ_2^* do introduce bias when estimating F and G. The bias function of the estimators are discussed next and compared to the estimator provided by El Barmi and Mukerjee (2008).

For $x \geq 0$ define $F^+_n(x) = \frac{1}{n}\sum_{i=1}^n I_{[-x\leq X_i\leq x]}$, $F^{+*}_{nm} = \max\{F^+_n, \frac{nF^+_n+mG^+_m}{n+m}\}$ and finally, let $F^*_{nm} = \frac{1}{2}(1+F^{+*}_{nm})$; G^+_m, $G^{+*}_{n,m}$ and $G^*_{n,m}$ are defined similarly. The estimator F^*_{nm} is the estimator for F studied by El Barmi and Mukerjee (2008) following ideas of Rojo (2004). Note that for $x \geq 0$,

$$\begin{aligned}E(F^*_{nm}(x)) &= \frac{1}{2}+\frac{1}{2}E(F^{+*}_{nm}(x)) \\ &= \frac{1}{2}+\frac{1}{2}E\{F^+_n(x)+\max\{0, \frac{m}{m+n}(G^+_m(x)-F^+_n(x))\}\} \\ &= \frac{1}{2}+\frac{1}{2}E(F^+_n(x))+\frac{m}{2(m+n)}E\{\max(0, G^+_m(x)-F^+_n(x))\}\end{aligned}$$

and since $\frac{1}{2}+\frac{1}{2}E(F^+_n(x)) = F(x)$,

$$Bias(F^*_{nm}(x)) = \frac{m}{2(m+n)}E\{\max(0, G^+_m(x)-F^+_n(x))\}. \tag{2.8}$$

Note that $Bias(F^*_{nm}(x)) \to 0$ as $\frac{n}{m} \to \infty$. Since our estimator $\widehat{F}^2_{nm}$ defined by (2.6) turns out to be equal F^*_{nm}, then its bias function is also given by (2.8).

Now consider the estimator $\widehat{F}^1_{nm}$ given by (2.4). For $x \geq 0$,

$$\begin{aligned}\widehat{F}^1_{n,m}(x) &= \Phi_1(\max(F_n(x), C_{nm}(x))) \\ &= \frac{1}{2}\{1+\max(F_n(x), C_{nm}(x))-\min(F_n(-x^-), C_{nm}(-x^-))\} \\ &= \frac{1}{2}\left\{1+F_n(x)-F_n(-x^-)+\max(0, C_{nm}(x)-F_n(x))\right. \\ &\quad \left.+\max(0, F_n(-x^-)-C_{nm}(-x^-))\right\}.\end{aligned}$$

Thus, $E(\widehat{F}^1_{n,m}(x)) = F(x)+\frac{1}{2}E(\max(0, C_{nm}(x)-F_n(x)))+\frac{1}{2}E(\max(0, F_n(-x^-)-$

$C_{nm}(-x^-)))$ and then, for $x \geq 0$

$$\begin{aligned} Bias(\widehat{F}^1_{n,m}(x)) &= \frac{1}{2}E(\max(0, \frac{m}{n+m}(G_m(x) - F_n(x)))) \\ &\quad + \frac{1}{2}E(\frac{m}{n+m}\max(0, F_n(-x^-) - G_m(-x^-))) \\ &= \frac{m}{2(m+n)}\{E(\max(0, G_m(x) - F_n(x))) \\ &\quad + E(\max(0, F_n(-x^-) - G_m(-x^-)))\} \\ &\geq \frac{m}{2(m+n)}E(\max(0, G^+_m(x) - F^+_n(x))) = Bias(F^*_{nm}). \end{aligned}$$

This result will also follow from the fact that $\widehat{F}^1_{nm} >_p \widehat{F}^2_{nm} = F^*_{nm}$.

Next consider the estimator F^2_{nm} defined in equation (2.2) and in Rojo and Batun-Cutz (2007):

$$F^2_{nm}(x) = \max\left\{\frac{1}{2}(1 + F_n(x) - F_n((-x)^-)), \frac{1}{2}(1 + G_m(x) - G_m((-x)^-))\right\}.$$

It follows easily that $E(F^2_{nm}(x)) = F(x) + \frac{1}{2}E(\max(0, G^+_m(x) - F^+_n(x)))$ and hence $Bias(F^2_{nm}(x)) = \frac{1}{2}E(\max(0, G^+_m(x) - F^+_n(x))) > Bias(F^*_{nm})$, for $x \geq 0$.

Finally, consider the estimator F^1_{nm} given in Rojo and Batun-Cutz (2007). For $x \geq 0$

$$\begin{aligned} F^1_{nm}(x) &= \frac{1}{2}\left\{1 + \max(F_n(x), \frac{1}{2}(1 + G_m(x) - G_m((-x)^-))) \right. \\ &\quad \left. - \min(F_n(-x), \frac{1}{2}(1 + G_m((-x)^-) - G_m((x))))\right\} \\ &= \frac{1}{2}(1 + F_n(x) - F_n((-x)^-)) + \frac{1}{2}\max(0, \frac{1}{2}(1 - 2F_n(x) + G^+_m(x))) \\ &\quad + \frac{1}{2}\max(0, \frac{1}{2}(-1 + 2F_n((-x)^-) + G^+_m(x))). \end{aligned}$$

Therefore,

$$\begin{aligned} E(F^1_{nm}(x)) = F(x) &+ \frac{1}{4}E(\max(0, (1 - 2F_n(x) + G^+_m(x)))) \\ &+ \frac{1}{4}E(\max(0, (-1 + 2F_n((-x)^-) + G^+_m(x)))). \end{aligned}$$

Then, for $x \geq 0$,

$$\begin{aligned} Bias(F^1_{nm}(x)) &= \frac{1}{4}E\left(\max\{0, G^+_m(x) - F^+_n(x) - F_n((-x)^-) - F_n(x) + 1\}\right) \\ &\quad + \frac{1}{4}E\left(\max\{0, G^+_m(x) - F^+_n(x) - 1 + F_n(x) + F_n((-x)^-)\}\right). \end{aligned}$$

The last expression is then seen to be equal to

$$\begin{aligned} &\frac{1}{4}E(\max\{\max(0, G^+_m(x) - F^+_n(x) - F_n((-x)^-) - F_n(x) + 1), \\ &\qquad \max(G^+_m(x) - F^+_n(x) + F_n((-x)^-) + F_n(x) - 1, 2(G^+_m - F^+_n)\}) \\ &\geq \frac{1}{4}E(\max(0, 2(G^+_m(x) - F^+_n(x)))) = Bias(F^*_{n,m}). \end{aligned}$$

The corresponding inequalities for the case of $x < 0$ follow by symmetry. Similar results may be obtained for the estimators $G^1_{n,m} = \Phi_1(\Phi^*_2(\Phi_1(F_n), G_m))$, $G^2_{n,m} = \Phi^*_2(\Phi_1(F_n), \Phi_1(G_m))$, and $\widehat{G}^1_{n,m}$ and $\widehat{G}^2_{n,m}$. It is easy to see that all the estimators for F have positive (negative) bias for $x > 0$ ($x < 0$), while the estimators for G have negative (positive) bias for $x > 0$ ($x < 0$). The following proposition summarizes the results about the bias functions.

Proposition 1. *Let $F >_p G$ be symmetric distribution functions, and let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent random samples from F and G respectively. The bias functions of the estimators for F and G given by (2.1), (2.2), (2.4), (2.5), (2.6), and (2.7), satisfy the following properties. For all x,*

$$\begin{aligned}(i)|Bias(\widehat{F}^1_{n,m}(x))| &\geq |Bias(\widehat{F}^2_{n,m}(x))| \\ &= \frac{m}{2(m+n)} E\{\max(0, G^+_m(|x|) - F^+_n(|x|)\} = |Bias(F^*_{n,m}(x))|\end{aligned}$$

(ii) $|Bias(F^1_{n,m}(x))| \geq |Bias(F^2_{n,m}(x))| \geq |Bias(\widehat{F}^2_{n,m}(x))|$
(iii) $|Bias(\widehat{G}^1_{n,m}(x))| \geq |Bias(\widehat{G}^2_{n,m}(x))| = -\frac{m}{2(m+n)} E\{\min(0, F^+_n(|x|) - G^+_m(|x|))\}$
(iv) $|Bias(G^1_{n,m}(x))| \geq |Bias(G^2_{n,m}(x))| \geq |Bias(\widehat{G}^2_{n,m}(x))|$.

2.3. Estimators as Projections onto Appropriate Convex Spaces

Recall the definitions of the new estimators given by (2.4) - (2.7). Schuster (1975) showed that the operator Φ_1 projects its argument to its closest symmetric distribution. That is, letting $\mathcal{F}$ be the convex set of symmetric distributions about zero, then for an arbitrary distribution H, $\| \Phi_1(H) - H \|_\infty = \inf_{F \in \mathcal{F}} \| H - F \|_\infty$. Rojo and Ma (1996), and Rojo and Batun-Cutz (2007) have shown that the operator Φ_2 has the property that for arbitrary distributions H and G, $|\Phi_2(H(x), G(x)) - H(x)| = \inf_{F \in \mathcal{F}^*} |F(x) - G(x)|$, where $\mathcal{F}^*$ is the convex set of distributions F satisfying (1.3). Thus, for F and G distribution functions let

$$\mathcal{F}_1 = \{\text{distribution functions } F \text{ satisfying (1.3) with } G \text{ replaced by } C_{n,m}\}$$

$$\mathcal{F}^*_1 = \{\text{symmetric distributions } F \text{ satisfying (1.3) with } G \text{ replaced by } \Phi_1(C_{n,m})\}$$

$$\text{and} \qquad \mathcal{F}^*_2 = \{\text{all symmetric at 0 distribution functions}\}.$$

Thus the estimator $\widehat{F}^2_{n,m}$ first projects F_n onto $\mathcal{F}^*_2$ and then projects $\Phi_1(F_n)$ onto $\mathcal{F}^*_1$. By contrast, the estimator $\widehat{F}^1_{n,m}$ first projects F_n onto $\mathcal{F}_1$ to obtain $\Phi_2(F_n, C_{n,m})$ and then projects the latter onto $\mathcal{F}^*_1$. With appropriate changes in the above notation, similar comments hold for the estimators $\widehat{G}^i_{n,m}$ for $i = 1, 2$.

2.4. Peakedness Order of New and Previous Estimators

By construction, the estimators $F^i_{n,m}$ and $\widehat{F}^i_{n,m}$, for $i = 1, 2$ are more peaked than the estimators $G^i_{n,m}$ and $\widehat{G}^i_{n,m}$, respectively. Rojo and Batun-Cutz (2007) showed that $F^1_{n,m} >_p F^2_{n,m}$. The next theorem provides comparisons in terms of peakedness for several of the estimators and provides a simple relationship between $F^2_{n,m}$ and $\widehat{F}^2_{n,m}$.

Lemma 1. *Let $F >_p G$ be symmetric distribution functions, and let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent random samples from F and G respectively. Consider the estimators for F and G given by (2.1), (2.2), (2.4), (2.5), (2.6), (2.7). Then*

(i) $\widehat{F}^2_{n,m} = \frac{n}{n+m}\widehat{F}_n + \frac{m}{n+m}F^2_{n,m}$

(ii) $\widehat{F}^1_{n,m} >_p \widehat{F}^2_{n,m} >_p \widehat{G}^2_{n,m} >_p \widehat{G}^1_{n,m}$

(iii) $F^1_{n,m} >_p F^2_{n,m} >_p \widehat{F}^2_{n,m}$, *and* $G^1_{n,m} <_p G^2_{n,m} <_p \widehat{G}^2_{n,m}$.

Proof. (i) For $x \geq 0$,

$$\begin{aligned}\widehat{F}^2_{n,m}(x) = \max\{\widehat{F}_n(x), \widehat{C}_{n,m}(x)\} &= \frac{n}{n+m}\widehat{F}_n(x) + \frac{m}{n+m}\max\{\widehat{F}_n(x), \widehat{G}_m(x)\} \\ &= \frac{n}{n+m}\widehat{F}_n(x) + \frac{m}{n+m}F^2_{n,m}(x).\end{aligned}$$

The result then follows by symmetry.

(ii) First we prove that $\widehat{F}^1_{n,m} >_p \widehat{F}^2_{n,m}$. Let $x \geq 0$, then

$$\begin{aligned}\widehat{F}^1_{n,m}(x) &= \frac{1}{2}\left[\max\{F_n(x), C_{n,m}(x)\} + 1 - \min\{F_n((-x)^-), C_{n,m}((-x)^-)\}\right] \\ &\geq \frac{1}{2}\left[C_{n,m}(x) + 1 - C_{n,m}((-x)^-)\right] = \widehat{C}_{n,m}(x). \qquad (2.9)\end{aligned}$$

Using similar arguments it can be shown that $\widehat{F}^1_{n,m}(x) \geq \widehat{F}_n(x)$. Therefore, combining the last inequality and (2.9) we obtain $\widehat{F}^1_{n,m}(x) \geq \widehat{F}^2_{n,m}(x)$. The result follows from symmetry.

We now prove that $\widehat{F}^2_{n,m} >_p \widehat{G}^2_{n,m}$. For $x \geq 0$, $\widehat{F}^2_{n,m}(x) = \max\{\widehat{F}_n(x), \widehat{C}_{n,m}(x)\} \geq \widehat{C}_{n,m}(x) \geq \widehat{G}^2_{n,m}(x)$. The result follows by symmetry.

Since for $x \geq 0$, $\widehat{G}^1_{n,m}(x) \leq \widehat{C}_{n,m}(x)$ and $\widehat{G}^1_{n,m}(x) \leq \widehat{G}_m(x)$. Then $\widehat{G}^2_{n,m} >_p \widehat{G}^1_{n,m}$ by symmetry.

Finally consider (iii). The result that $F^1_{n,m} >_p F^2_{n,m}$ follows from Rojo and Batun-Cutz (2007). The result that $F^2_{n,m} >_p \widehat{F}^2_{n,m}$ follows from the arguments used to prove that Bias$(F^2_{n,m})$ $\geq$Bias$(\widehat{F}^2_{n,m})$.

Note that (i) implies that for $x \geq 0$, $Bias(F^2_{n,m}(x)) = \frac{m+n}{m}Bias(\widehat{F}^2_{n,m}(x))$, so that $|Bias(F^2_{n,m}(x))| = \frac{m+n}{m}|Bias(\widehat{F}^2_{n,m}(x))|$ for all x, thus providing a more accurate description of the result about bias given in proposition 1.

3. Asymptotics

This section discusses the strong uniform convergence of the estimators and their asymptotic distribution theory. One important aspect of the asymptotic results for the estimators $\hat{F}^i_{n,m}$ $(\hat{G}^i_{n,m})$, $i = 1, 2$ discussed here is that they hold even when only n (m) tends to infinity. This is in sharp contrast with the results of Rojo and Batun-Cutz (2007) and those of El Barmi and Mukerjee (2008). We discuss the strong uniform convergence first.

3.1. Strong Uniform Convergence

The following theorem provides the strong uniform convergence of the estimators $\widehat{F}^i_{n,m}$ ($\widehat{G}^i_{n,m}$), $i = 1, 2$. The results use the strong uniform convergence of the symmetrized $\hat{F}_n$ ($\hat{G}_m$) to F (G) as $n \to \infty$ ($m \to \infty$), Schuster (1975).

Theorem 3.1. *Let F and G be symmetric distribution functions with $F >_p G$, and let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent random samples from F and G respectively. Then,*

(i) *$\widehat{F}^i_{n,m}$, for $i = 1, 2$, converge uniformly with probability one to F as $n \to \infty$.*
(ii) *$\widehat{G}^i_{n,m}$ for $i = 1, 2$ converge uniformly with probability one to G as $m \to \infty$.*

Proof. **(i)** Consider $\widehat{F}^2_{n,m}$ first. Then, for $x \geq 0$,

$$\widehat{F}^2_{n,m}(x) - F(x) = \widehat{F}_n(x) - F(x) + \frac{m}{n+m} \max\{0, \widehat{G}_m(x) - \widehat{F}_n(x)\}. \tag{3.1}$$

But, since $F(x) \geq G(x)$,

$$\widehat{G}_m(x) - \widehat{F}_n(x) \leq \widehat{G}_m(x) - G(x) + F(x) - \widehat{F}_n(x).$$

Hence

$$\begin{aligned} \max\{0, \widehat{G}_m(x) - \widehat{F}_n(x)\} &\leq \max\{0, \widehat{G}_m(x) - G(x) + F(x) - \widehat{F}_n(x)\} \\ &\leq |\widehat{G}_m(x) - G(x)| + |\widehat{F}_n(x) - F(x)| \end{aligned} \tag{3.2}$$

and therefore, the left side of (3.1) is bounded above by

$$|\widehat{F}_n(x) - F(x)| + (\frac{m}{m+n})\{|\widehat{G}_m(x) - G(x)| + |\widehat{F}_n(x) - F(x)|.\}$$

Since $\widehat{F}_n$, and $\widehat{G}_m$ are strongly and uniformly consistent for F and G, then as $n \to \infty$, with probability one,

$$\sup_{x \geq 0} |\widehat{F}^2_{n,m}(x) - F(x)| \to 0,$$

regardless of whether $m \to \infty$ or not. When $x < 0$ the result follows by symmetry.

Let us now consider the case of $\widehat{F}^1_{n,m}$. For $x \geq 0$

$$\begin{aligned} \widehat{F}^1_{n,m}(x) - F(x) &= \widehat{F}_n(x) - F(x) + \frac{1}{2}\frac{m}{n+m} [\max\{0, G_m(x) - F_n(x)\} \\ &\quad - \min\{0, G_m(-x^-) - F_n(-x^-)\}]. \end{aligned} \tag{3.3}$$

Since $F(x) \geq G(x)$ and $F(-x) \leq G(-x)$, then it follows that

$$\max\{0, G_m(x) - F_n(x)\} - \min\{0, G_m(-x^-) - F_n(-x^-)\}$$

is bounded above by

$$\max\{0, G_m(x) - G(x) + F(x) - F_n(x)\} - \min\{0, G_m(-x^-) - G(-x) + F(-x) - F_n(-x^-)\}$$

and, therefore, the left side of (3.2) is bounded above by

$$
\begin{aligned}
|\widehat{F}_n(x) - F(x)| &+ \frac{1}{2}\frac{m}{m+n}(|G_m(x) - G(x)| + |F(x) - F_n(x)| \\
&+ |G_m(-x^-) - G(-x)| + |F(-x) - F_n(-x^-)|). \qquad (3.4)
\end{aligned}
$$

Taking the supremum over x in (3.4), and then letting $n \to \infty$, the result follows, whether $m \to \infty$ or not, from the strong uniform convergence of $\widehat{F}_n$, G_m, and F_n to F, G, and F respectively. The result for $x < 0$ follows by symmetry.

(iii) The proof for the strong uniform convergence of $\widehat{G}^2_{n,m}$ to G, when only $m \to \infty$ is similar. We sketch the proof. For $x < 0$

$$
\widehat{G}^2_{n,m}(x) - G(x) = \widehat{G}_m(x) - G(x)) + \frac{n}{n+m}\max\{0, \widehat{F}_n(x) - \widehat{G}_m(x)\}.
$$

Therefore, since $F(x) < G(x)$ for $x < 0$, $\max\{0, \widehat{F}_n(x) - \widehat{G}_m(x)$ is bounded above by

$$
\max\{0, \widehat{F}_n(x) - F(x) + G(x) - \widehat{G}_m(x)\} \leq |\widehat{F}_n(x) - F(x)| + |G(x) - \widehat{G}_m(x)|.
$$

When $m \to \infty$, the result follows, regardless of whether $n \to \infty$ or not, from the strong uniform convergence of $\widehat{F}_n$ and $\widehat{G}_m$ and using a symmetry argument to handle the case of $x > 0$.

(iv) This case is omitted as it follows from similar arguments.

3.2. Weak Convergence

Consider first the point-wise asymptotic distribution for $\widehat{F}^i_{n,m}$, $i = 1, 2$. Recall that

$$
\sqrt{n}(\widehat{F}_n(x) - F(x)) \xrightarrow{W} N\left(0, \frac{F(-|x|)(2F(|x|) - 1)}{2}\right).
$$

Therefore, when $n/m \to \infty$, and using (3.1)-(3.4), Slutsky's theorem and the central limit theorem for $\widehat{F}_n$, we get the following result:

$$
\sqrt{n}(\widehat{F}^i_{nm}(x) - F(x)) \xrightarrow{W} N\left(0, \frac{F(-|x|)(2F(|x|) - 1)}{2}\right). \qquad (3.5)
$$

Thus, under these conditions, $\widehat{F}^i_{n,m}$, $i = 1, 2$, are $\sqrt{n}-$equivalent and have the same asymptotic distribution as the symmetrized $\widehat{F}_n$ which happens to have the same asymptotic limit as in the case when G is completely known. Note that this result assumes only that $n/m \to \infty$ and hence the result holds if m is fixed and $n \to \infty$. This is in sharp contrast with the results of El Barmi and Mukerjee (2008) that require that both n and m tend to infinity. Similar results hold for the estimators $\widehat{G}^i_{n,m}$, $i = 1, 2$. These are summarized in the following theorem.

Theorem 3.2. *Suppose that $F >_p G$ and let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be random samples from F and G respectively. Then for $i = 1, 2$,*

(i) If $n/m \to \infty$ then

$$
\sqrt{n}(\widehat{F}^i_{nm}(x) - F(x)) \xrightarrow{\mathcal{D}} N\left(0, \frac{F(-|x|)(2F(|x|) - 1)}{2}\right).
$$

(ii) If $m/n \to \infty$ then

$$\sqrt{n}(\widehat{G}^i_{nm}(x) - G(x)) \quad \xrightarrow{\mathcal{D}} \quad N\left(0, \frac{G(-|x|)(2G(|x|)-1)}{2}\right).$$

We now turn our attention to the weak convergence of the processes

$$\left\{\sqrt{n}\left(\widehat{F}^i_{nm}(x) - F(x)\right) : -\infty < x < \infty\right\},$$

and

$$\left\{\sqrt{n}\left(\widehat{G}^i_{nm}(x) - G(x)\right) : -\infty < x < \infty\right\},$$

for $i = 1, 2$. Only the results for $\widehat{F}^i_{n,m}$, $i = 1, 2$ will be discussed in detail as the results for $\widehat{G}^i_{n,m}$, $i = 1, 2$ can be obtained by similar arguments. Although the processes $\left\{\sqrt{n}\left(\widehat{F}^i_{nm}(x) - F(x)\right) : -\infty < x < \infty\right\}$ for $i = 1, 2$ are correlated, we are only interested in their marginal behavior. For that purpose let $\{W_1(x) : -\infty < x < \infty\}$ denote a mean zero Gaussian process with covariance function

$$(3.6) \quad E(W_1(x)W_1(y)) \quad = \quad \begin{cases} \frac{1}{2}(1 - F(y))(F(x) - F(-x)) & \text{if } |y| > |x| \\ \frac{1}{2}F(x)(F(-y) - F(y)) & \text{if } |y| < |x|, \end{cases}$$

and let $\{W_2(x) : -\infty < x < \infty\}$ denote a mean zero Gaussian process with covariance function

$$(3.7) \quad E(W_2(x)W_2(y)) \quad = \quad \begin{cases} \frac{1}{2}(1 - G(y))(G(x) - G(-x)) & \text{if } |y| > |x| \\ \frac{1}{2}G(x)(G(-y) - G(y)) & \text{if } |y| < |x|. \end{cases}$$

We have the following result:

Theorem 3.3. *Under the conditions of the previous Theorem,*

(i) If $n/m \to \infty$, then

$$\{\sqrt{n}(\widehat{F}^i_{nm}(x) - F(x)), -\infty < x < \infty\} \quad \xrightarrow{W} \quad \{W_1(x) : -\infty < x < \infty\}, \text{ and}$$

(ii) If $m/n \to \infty$, then

$$\{\sqrt{n}(\widehat{G}^i_{nm}(x) - G(x)), -\infty < x < \infty\} \quad \xrightarrow{W} \quad \{W_2(x) : -\infty < x < \infty\}.$$

Proof. The proof follows easily by the continuous mapping Theorem after observing that the weak limit of $\{\sqrt{n}(\widehat{F}_n(x) - F(x))), -\infty < x < \infty\}$ is the process $\{W_1(x) : -\infty < x < \infty\}$, together with the fact that, using (3.1),

$$(3.8) \quad \widehat{F}^2_{n,m}(x) - F(x) = \widehat{F}_n(x) - F(x) + \frac{m}{n+m}\max\{0, \widehat{G}_m(x) - \widehat{F}_n(x)\},$$

with $\| \sqrt{n}\frac{m}{n+m}\{\max 0, \widehat{G}_m - \widehat{F}_n\} \|_\infty \to 0$ with probability one, where $\| \cdot \|_\infty$ denotes the sup norm. Similar arguments yield the results for the other processes.

The asymptotic theory for $\widehat{F}^2_{n,m}$ was discussed by El Barmi *et al* (2008) for the case that both n and m go to infinity and hence their result does not include our result here when m is bounded and $n \to \infty$. When $n/m \to c$ with $0 \le c < \infty$, and $F(x) > G(x)$ for all $x > 0$ the weak limit of $\{\sqrt{n}(\widehat{F}^i_{nm}(x) - F(x)), -\infty < x < \infty\}$ is $\{W_1(x) : -\infty < x < \infty\}$, for $i = 1, 2$, which is the weak limit of the

process $\{\sqrt{n}(F_{n,2}(x) - F(x)), -\infty < x < \infty\}$ discussed in Rojo and Batun-Cutz (2007). Let $\{Z(x), -\infty < x < \infty\}$ represent the weak limit of the empirical process $\{\sqrt{n}(F_n(x) - F(x)), -\infty < x < \infty\}$. That is $\{Z(x), -\infty < x < \infty\}$ is a mean zero Gaussian process with covariance function $E(Z(t)Z(s)) = F(s)(1-F(t))$ for $s \leq t$. When $n/m \to c$ with $0 \leq c < \infty$, and $F(x) = G(x)$ for all x the weak limits of $\{\sqrt{n}(\widehat{F}^i_{nm}(x) - F(x)), -\infty < x < \infty\}$ for $i = 1, 2$ follow from the results in Rojo (2004) as follows:

Theorem 3.4. *Let $F(x) = G(x)$ for all x and let $n/m \to c$ for $0 \leq c < \infty$. Let $\{W_i(x), -\infty < x < \infty\}$, for $i = 1, 2$ be the mean zero Gaussian processes with covariance functions given by (3.6) and (3.7), respectively. Let $W_i^*(x) = W_i(|x|)sgn(x)$, for $i = 1, 2$. Then*

(i) The process $\sqrt{n}(\widehat{F}^2_{n,m} - F(x)), -\infty < x < \infty\}$ converges weakly to the process $\{\max(W_1^*(x), \frac{\sqrt{c}W_2^*(x)+cW_1^*(x)}{1+c}), -\infty < x < \infty\}$ with $W_1^* \stackrel{\mathcal{D}}{=} W_2^*$ and independent.

(ii) The process $\sqrt{n}(\widehat{F}^1_{n,m} - F(x)), -\infty < x < \infty\}$ converges weakly to the process $\{H(|x|)sgn(x), -\infty < x < \infty]\}$ where $H(x) = \frac{1}{2}\{\max\{Z_1(x), \frac{c}{1+c}Z_1(x) + \frac{\sqrt{c}}{1+c}Z_2(x)\} - \min\{Z_1(-x), \frac{c}{1+c}Z_1(-x) + \frac{\sqrt{c}}{1+c}Z_2(-x)\}$, and $\{Z_i(x), -\infty < x < \infty\}$, $i = 1, 2$ are independent copies of the process $\{Z(x), -\infty < x < \infty\}$.

Proof. (i) Consider $\widehat{F}^2_{n,m}$ first. When $F(x) = G(x)$ for all x, it follows from (3.8) that, for $x \geq 0$,

$$\sqrt{n}(\widehat{F}^2_{n,m}(x) - F(x)) = \max\{\sqrt{n}(\widehat{F}_n(x) - F(x)), \sqrt{n/m}\frac{m}{n+m}\sqrt{m}(\widehat{G}_m(x) - G(x)) + \frac{n}{n+m}\sqrt{n}(F(x) - \widehat{F}_n(x))\}. \tag{3.9}$$

By the independence of $\widehat{F}_n$ and $\widehat{G}_m$ and their weak convergence to W_1 and W_2, it follows that the bivariate process

$$\{\sqrt{n/m}\frac{m}{n+m}\sqrt{m}(\widehat{G}_m(x) - G(x)), \frac{n}{n+m}\sqrt{n}(F(x) - \widehat{F}_n(x)), -\infty < x < \infty\}$$

converges weakly to the process $\{\frac{\sqrt{c}}{1+c}W_2(x), \frac{c}{1+c}W_1(x), -\infty < x < \infty\}$. Since for $x < 0$, $\widehat{F}^2_{n,m}(x) - F(x) \stackrel{\mathcal{D}}{=} F(-x) - \widehat{F}^2_{n,m}(-x)$, the result then follows for $0 < c < \infty$ from the continuous mapping theorem after observing that the mapping $h(x, y) = (\frac{1+c}{c}y, x+y)$ is continuous, and then applying it to (3.9) to get the result. The case of $c = 0$ follows immediately since it then follows that the second term on the right side of (3.9) converges to zero in probability.

(ii) Note that for $x > 0$

$$\begin{aligned}\widehat{F}^1_{nm}(x) - F(x) &= \frac{1}{2}\max\{F_n(x) - F(x), \\ &\quad \frac{n}{m+n}(F_n(x) - F(x)) + \frac{m}{n+m}(G_m(x) - F(x))\} \\ &+ \frac{1}{2}\min\{F_n(-x) - F(-x), \\ &\quad \frac{n}{m+n}(F_n(-x) - F(-x)) + \frac{m}{n+m}(G_m(-x) - F(-x))\}.\end{aligned}$$

Since the function $h(x,y,z,w) = \frac{1}{2}\left[\max\{x, x+z\} - min\{y, y+w\}\right]$ is continuous, by the continuous mapping theorem we obtain

$$\sqrt{n}(\widehat{F}^1_{nm}(x) - F(x)) \xrightarrow{W} \frac{1}{2}\Big[\max\{Z_1(x), \frac{c}{1+c}Z_1(x) + \frac{\sqrt{c}}{1+c}Z_2(x)\} - \min\{Z_1(-x), \frac{c}{1+c}Z_1(x) + \frac{\sqrt{c}}{1+c}Z_2(-x)\}\Big] = H(x). \tag{3.10}$$

The result then follows after considering the case $x < 0$ and following a similar argument.

It has been observed, e.g. Rojo (1995a), Rojo (2004), and Rojo and Batun-Cutz (2007), that weak convergence of the processes of interest fails to hold when the underlying distributions F and G coincide at some point x_0 and are unequal in some neighborhood to the right of x_0. That is the case here as well. Suppose that $F(x_0) = G(x_0)$ for $x_0 > 0$ and $F(x) > G(x)$ for $x \in (x_0, x_0 + \nu)$, $\nu > 0$. If $\frac{m}{n} \to c$, $0 < c \leq \infty$, as $m, n \to \infty$, it follows that

$$\sqrt{n}(\widehat{F}^1_{nm}(x_0) - F(x_0)) \xrightarrow{\mathcal{D}} H(|x_0|)sgn(x_0), \tag{3.11}$$

with $H(x)$ defined as in (ii) of theorem 3.4 with $(Z_1(x_0), Z_2(x_0))$ a zero-mean bivariate normal distribution vector with covariance $(1 - F(x_0))F(x_0)$.

However, for $x \in (x_0, x_0 + \nu)$ the sequence $\sqrt{n}(\widehat{F}^1_{nm}(x) - F(x))$ converges in distribution to the distribution given in (3.5). Then it can be seen, using arguments as in Rojo (1995a), that the process $\{\sqrt{n}(\widehat{F}^1_{nm}(x) - F(x)) : -\infty < x < \infty\}$ is not tight and hence cannot converge weakly.

We finish this section with results that provide the weak convergence of the processes $\{\sqrt{n}(F^i_{n,m}(x) - F(x)), -\infty < x < \infty\}$ for $i = 1, 2$, in the case that $F(x) = G(x)$ for all x.

Theorem 3.5. *Let $n/m \to c$ with $0 \leq c < \infty$, and $F(x) = G(x)$ for all x.*

(i) The process $\{\sqrt{n}(F^2_{n,m}(x) - F(x)), -\infty < x < \infty\}$ converges weakly to

$$\{sgn(x)\max\{sgn(x)W_1(x), sgn(x)\sqrt{c}W_2(x), -\infty < x < \infty\},$$

where W_1 and W_2 are independent copies of the mean zero Gaussian process with covariance function defined by (3.6).

(ii) The process $\{\sqrt{n}(F^1_{n,m}(x) - F(x)), -\infty < x < \infty\}$ converges weakly to

$$\frac{1}{2}\{\max\{Z(xsgn(x)), \sqrt{c}W(xsgn(x)\} - sgn(x)\min\{Z(-xsgn(x)), \sqrt{c}W(-xsgn(x))\}; -\infty < x < \infty\},$$

where Z and W are independent mean zero Gaussian process with covariance functions defined by $E(Z(s)Z(t)) = F(s)(1 - F(t))$ for $s < t$, and (3.6) respectively.

Proof. (i) The result follows from the independence of $\{\sqrt{n}(\widehat{F}^*_n(x) - F(x)), -\infty < x < \infty\}$ and $\{\sqrt{m}(\widehat{G}^*_m(x) - G(x)), -\infty < x < \infty\}$, their weak convergence to W_1 and W_2, and the continuous mapping theorem after observing that

$$\sqrt{n}(F^2_{n,m}(x) - F(x)) = sgn(x)\max\{sgn(x)\sqrt{n}(\widehat{F}^*_n(x) - F(x)), sgn(x)(\sqrt{\frac{n}{m}}\sqrt{m}(\widehat{G}^*_m(x) - G(x))\}.$$

(ii) Consider the case of $x > 0$ and write

$$\begin{aligned}\sqrt{n}(F^1_{n,m}(x)-F(x)) &= \frac{\sqrt{n}}{2}\{1-2F(x)+\max(F_n(x),\widehat{G}_m(x))-\min(F_n(-x),\widehat{G}_m(-x))\}\\ &= \frac{1}{2}\{\max\{\sqrt{n}(F_n(x)-F(x)),\sqrt{\frac{n}{m}}\sqrt{m}(\widehat{G}_m(x)-G(x))\}\\ &\quad -\min\{\sqrt{n}(F_n(-x)-F(-x)),\sqrt{\frac{n}{m}}\sqrt{m}(\widehat{G}_m(-x)-G(-x))\}\}.\end{aligned}$$

For $x < 0$, a similar argument leads to

$$\begin{aligned}\sqrt{n}(F^1_{n,m}(x)-F(x)) &= \frac{1}{2}\{\min\{\sqrt{n}(F_n(x)-F(x)),\sqrt{\frac{n}{m}}\sqrt{m}(\widehat{G}_m(x)-G(x))\}\\ &\quad -\max\{\sqrt{n}(F_n(-x)-F(-x)),\sqrt{\frac{n}{m}}\sqrt{m}(\widehat{G}_m(-x)-G(-x))\}\},\end{aligned}$$

The result then follows by the continuous mapping theorem after letting $n/m \to c$ with $\sqrt{n}(F_n(x)-F(x))$ and $\sqrt{m}(\widehat{G}_m(x)-G(x))$ independent and converging weakly to Z and W respectively.

4. Example with Sib-pair Data: An Illustration

In this section the estimator $\widehat{F}^2_{n,m}$ is illustrated by using the sib-paired data for the Caucasian population in the Dallas metroplex area. As can be observed from Figure 2, the new estimated distribution functions now satisfy both the constraint of symmetry and the constraint of peakedness. Thus, since siblings with two alleles identical by descent are more similar than those siblings sharing only one allele identical by descent, the distribution function denoted by IBD2 is more peaked about zero than the other two distribution functions. Similar comments apply to the other comparisons.

5. Simulation Work

Monte Carlo simulations were performed to study the finite-sample properties of the estimators $\widehat{F}^1_{nm}$ and $\widehat{F}^2_{nm}$ defined by (2.4) and (2.6) respectively. We consider various examples of underlying distributions (Normal, Cauchy, Laplace, mixtures of normals, and T), and sample sizes ($n = 10, 20, 30$ for F and $m = 10, 20, 30$ for G). Each simulation consisted of 10,000 replications.

Figures 3 and 4 show the bias functions for the four estimators considered here. Figure 3 considers $F \sim Cauchy(0, 1)$ and $G \sim Cauchy(0, 2)$, and Figure 4 considers the case with $F \sim Laplace(0, 1)$ and $G \sim Laplace(0, 1.5)$. As shown in Proposition 1, the estimator $\widehat{F}^2_{n,m}$ has uniformly the smallest absolute bias. These Figures are representative of the results that we obtained. One result that holds in all of our simulations is that $|Bias(F^1_{n,m}(x))| \geq |Bias(\widehat{F}^1_{n,m}(x))|$ for all x. Unfortunately, we are unable to prove this conjecture.

Turning our attention to comparing the estimators in terms of the Mean Squared Error (MSE) Figures 5 - 10 show the ratio of the MSE of the empirical distribution to the MSE of each of the four estimators considered here. These plots are representative of all the examples considered. As it can be seen from the plots, the

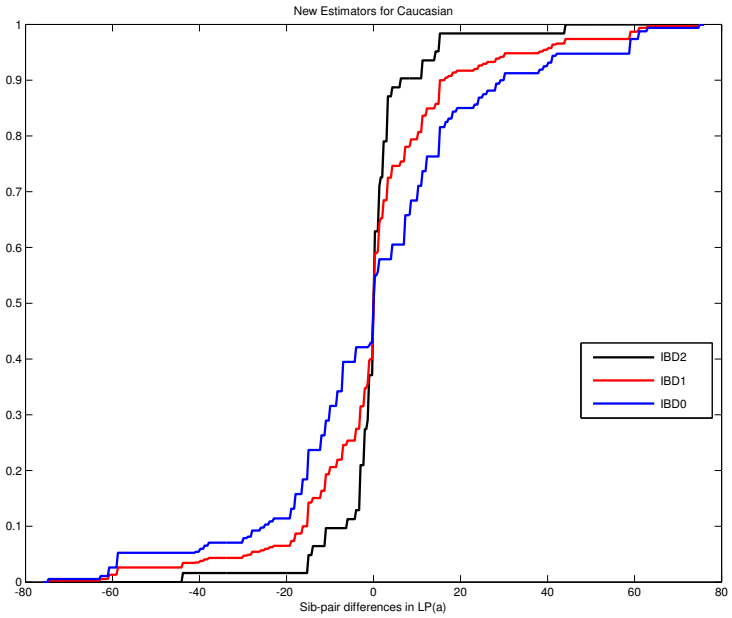


FIG 2. *Order restricted estimators for the sib-pair data incorporating peakedness.*

empirical distribution function is dominated by the estimators in every case and for all x. Whereas the estimators $\widehat{F}^i_{n,m}$ behave better than the estimators $F^i_{n,m}$, $i = 1, 2$ in a neighborhood of zero, the roles are reversed on the tails of the underlying distribution. What is observed is that the region of the support of F where $\widehat{F}^i_{n,m}$ dominate $F^i_{n,m}$, $i = 1, 2$ shrinks as the tails of the distributions get heavier, and when the distribution G is far from F. Thus, there is no clear choice among the four estimators, unless the tail is of special interest, in which case the estimator $F^2_{n,m}$ seems to be the clear choice.

6. Conclusions

Estimators were proposed for the distribution functions F and G when it is known that $F >_p G$, and F and G are symmetric about zero. The estimator for F (G) was seen to be strongly uniformly consistent when only n (m) goes to infinity and the asymptotic theory of the estimators was delineated without requiring that both n and m go to infinity. Finite sample properties of the estimators were considered and it was shown that the estimator $\widehat{F}^2_{n,m}$ has the uniformly smaller absolute bias of the four estimators considered here. The choice of which estimator is best in terms of mean squared error (mse), however, is not clear. Although the estimators $\widehat{F}^i_{n,m}$ for $i = 1, 2$ have smaller mse than the estimators $F^i_{n,m}$, $i = 1, 2$ in a neighborhood of zero, the tails are problematic for $\widehat{F}^i_{n,m}$ and the estimators $F^i_{n,m}$ tend to have smaller mse as demonstrated by the simulation study.

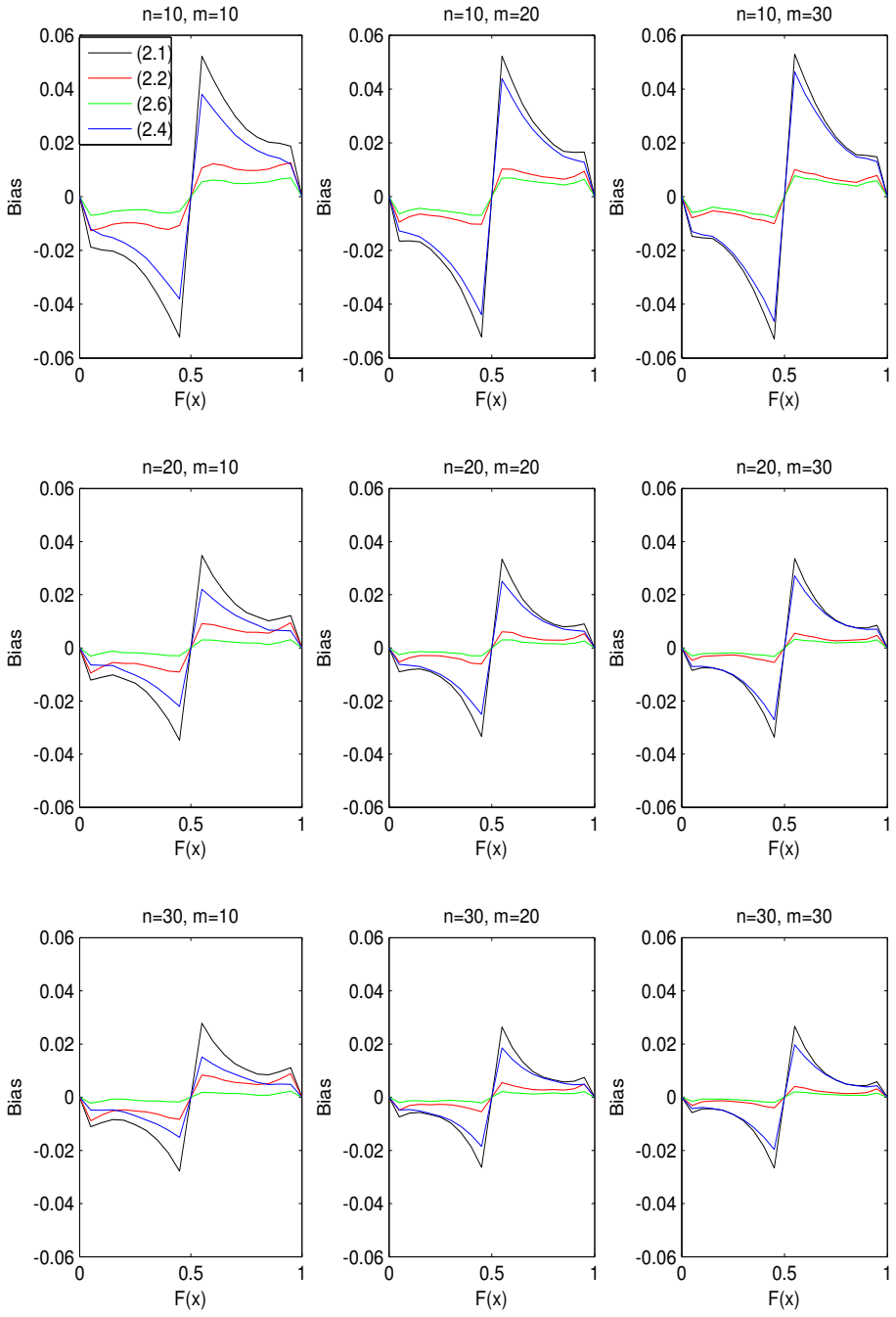


FIG 3. *Bias of the estimators when estimating F∼Cauchy(0,1) with G∼Cauchy(0,2)*

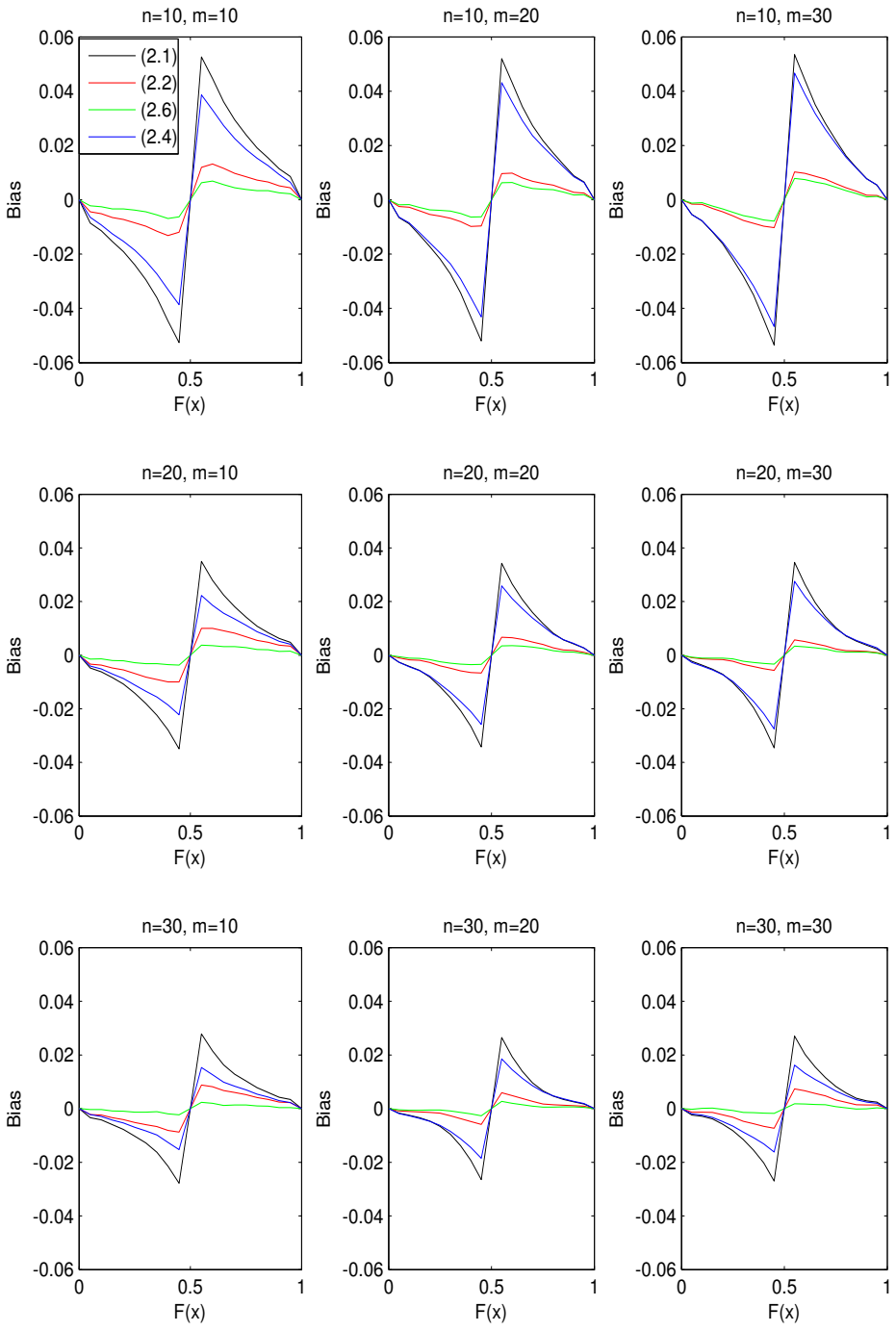


FIG 4. *Bias of the estimators when estimating F∼Laplace(0,1) with G∼Laplace(0,2).*

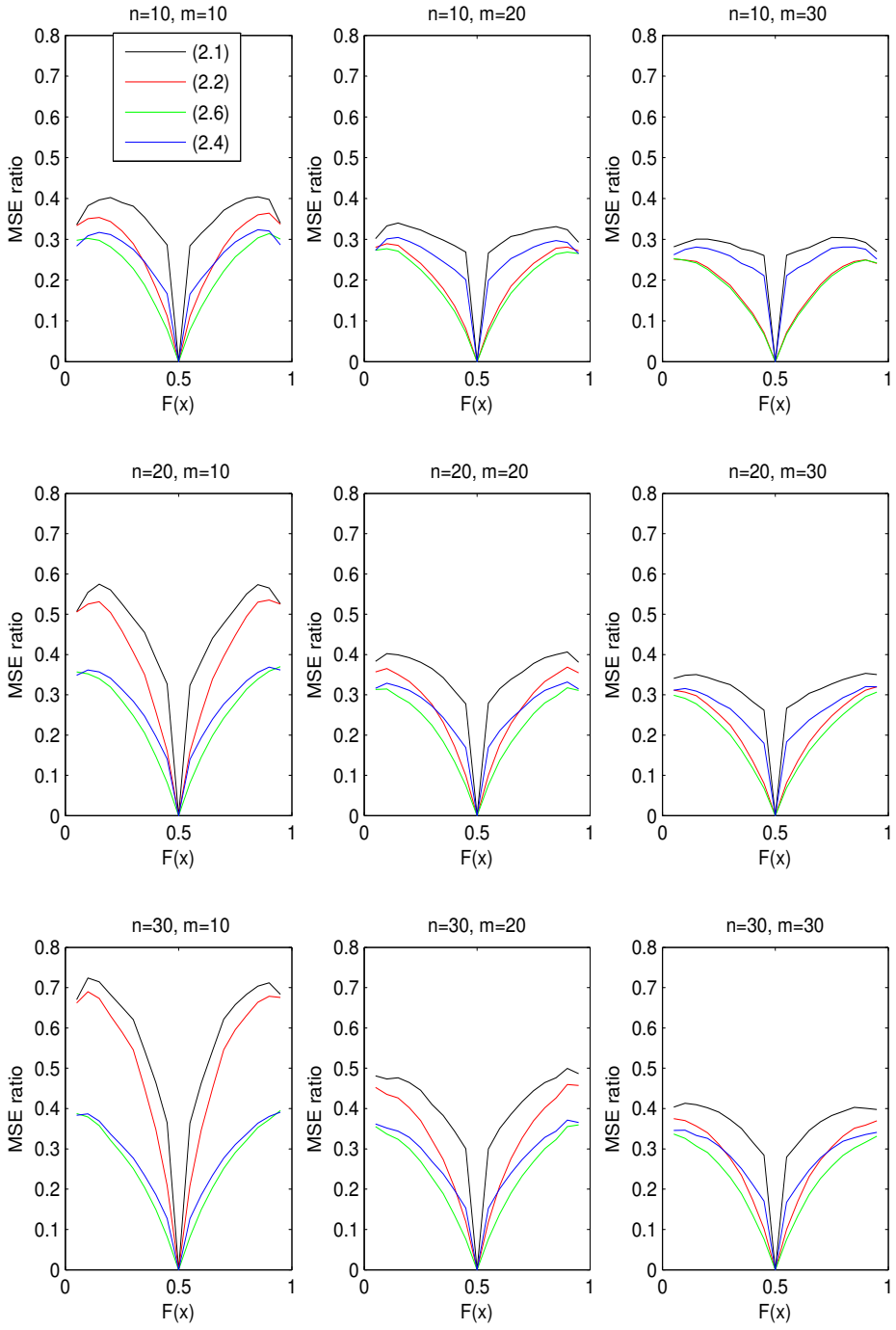


FIG 5. *Mean Squared Error of the estimators when estimating F∼Normal(0,1) with G∼Normal(0,1.1).*

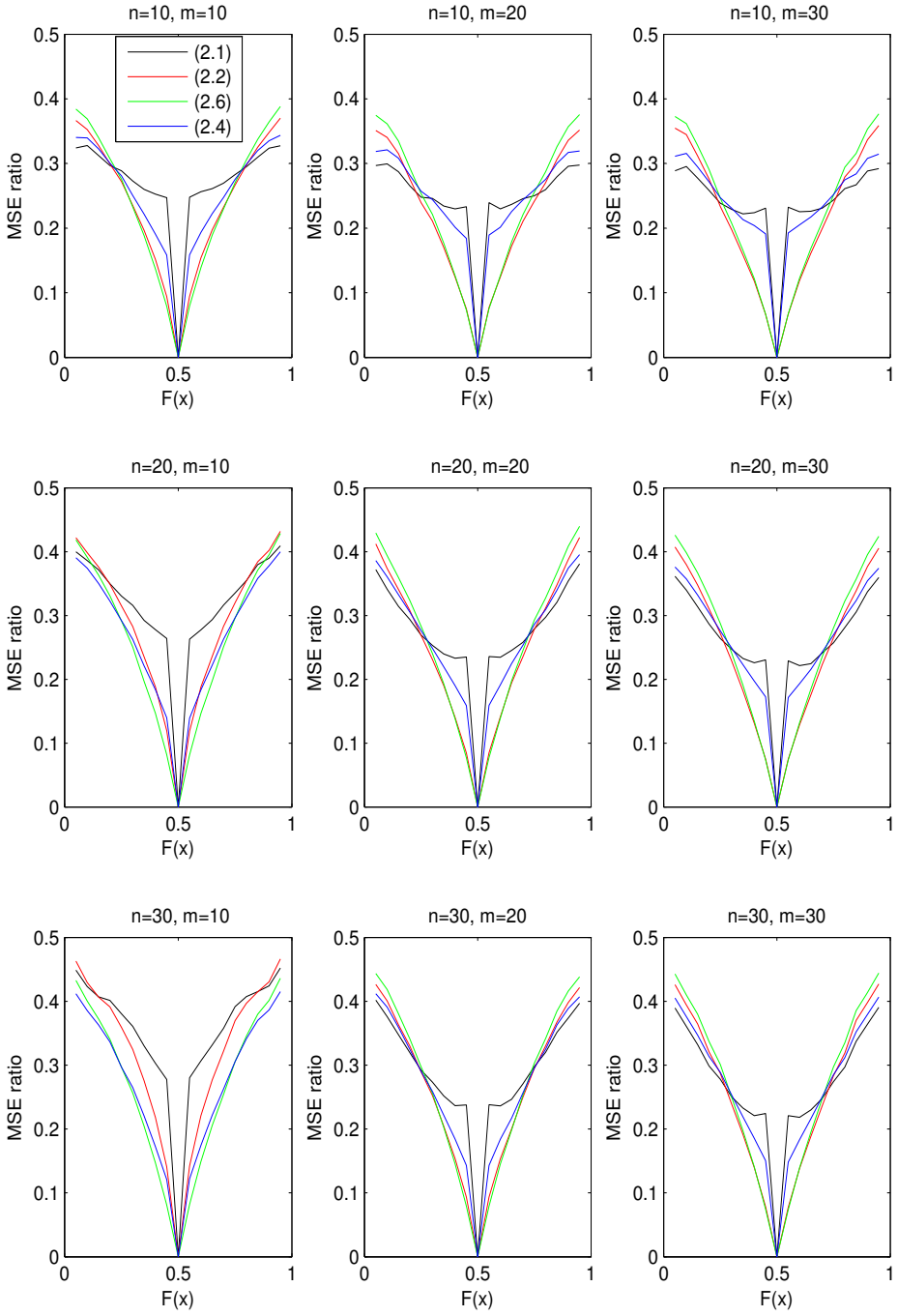


FIG 6. *Mean Squared Error of the estimators when estimating F∼Normal(0,1) with G∼Normal(0,2).*

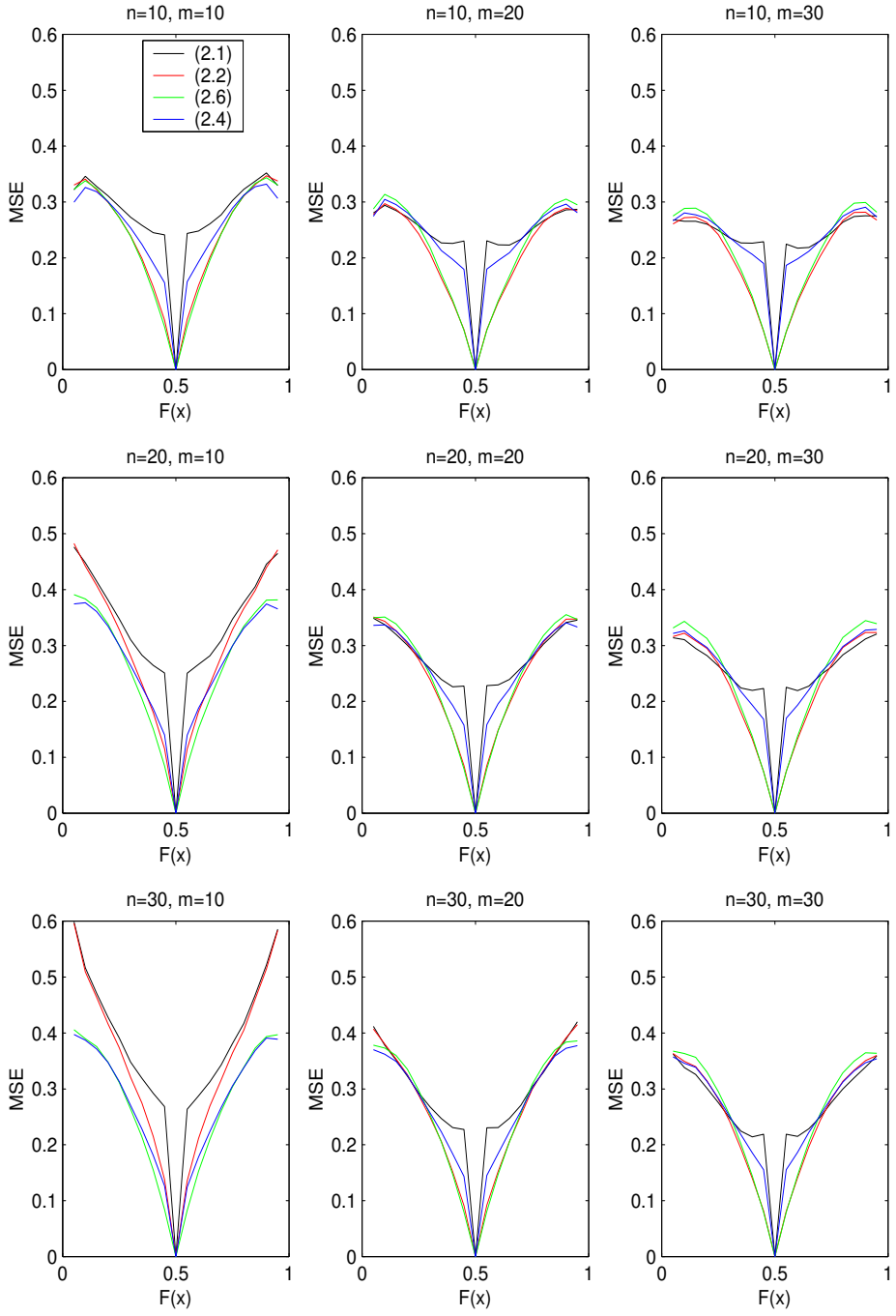


FIG 7. *Mean Squared Error of the estimators when estimating F∼Cauchy(0,1) with G∼Cauchy(0,1.5).*

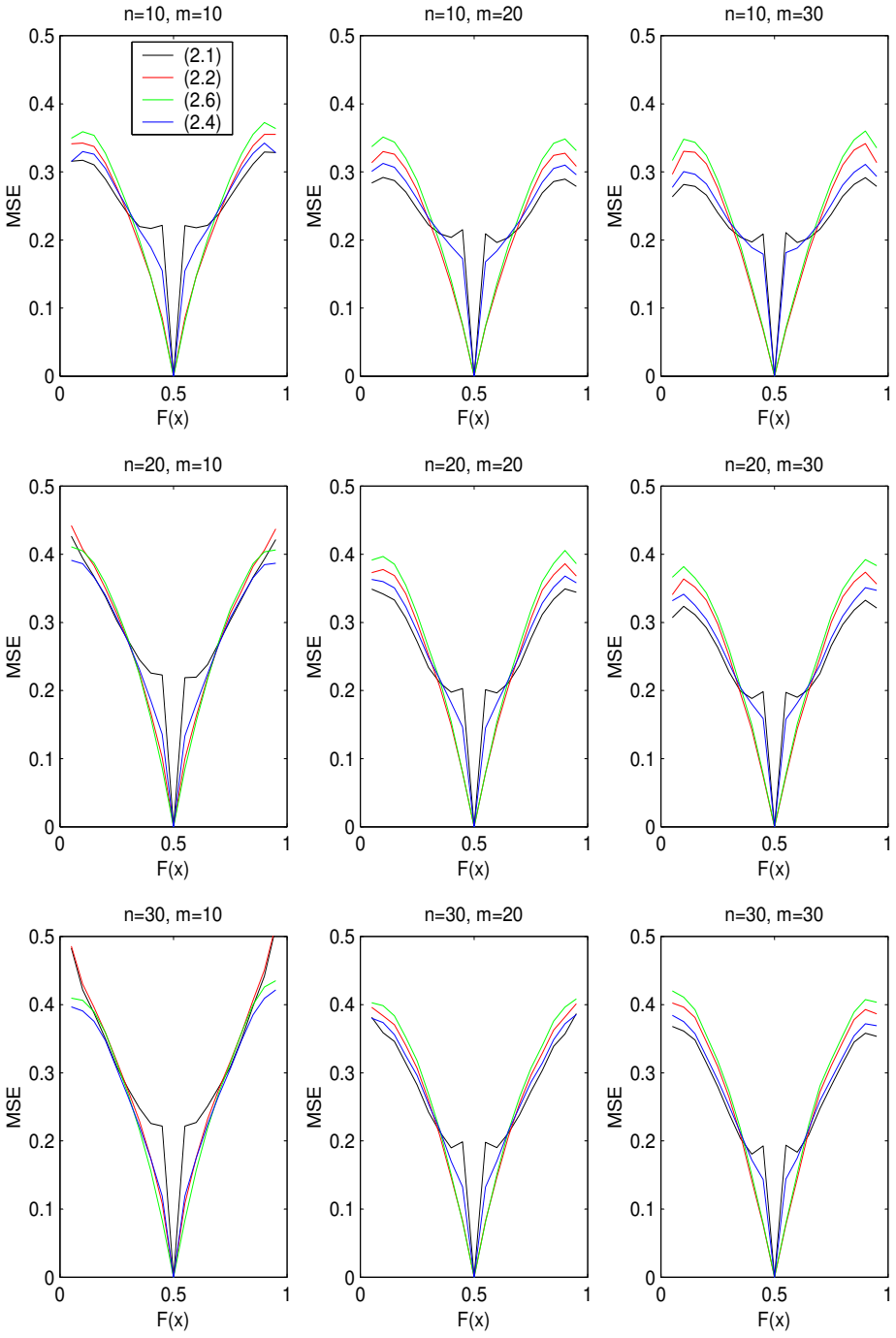


FIG 8. *Mean Squared Error of the estimators when estimating F∼Cauchy(0,1) with G∼Cauchy(0,2).*

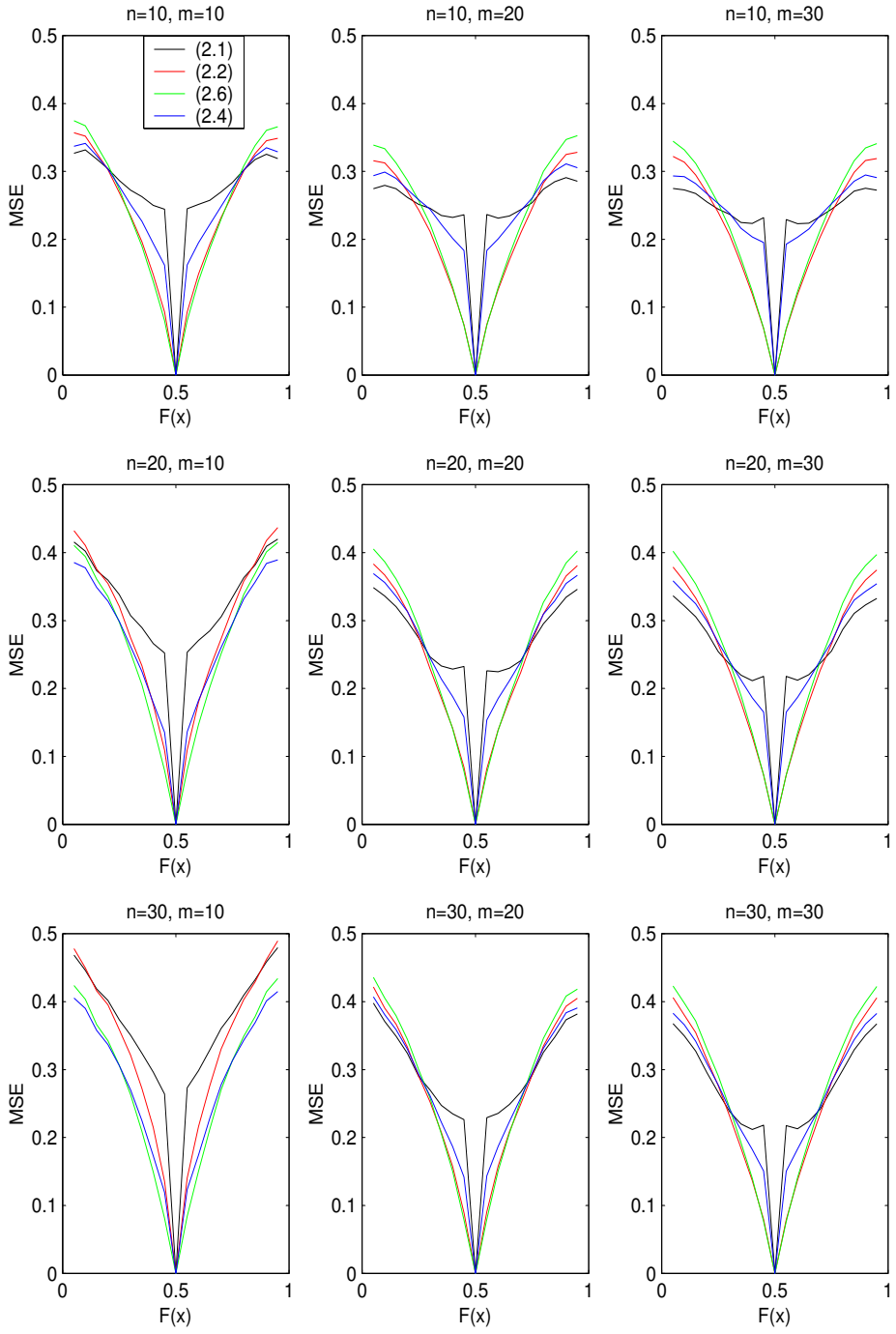


FIG 9. *Mean Squared Error of the estimators when estimating F∼Laplace(0,1) with G∼Laplace(0,1.5).*

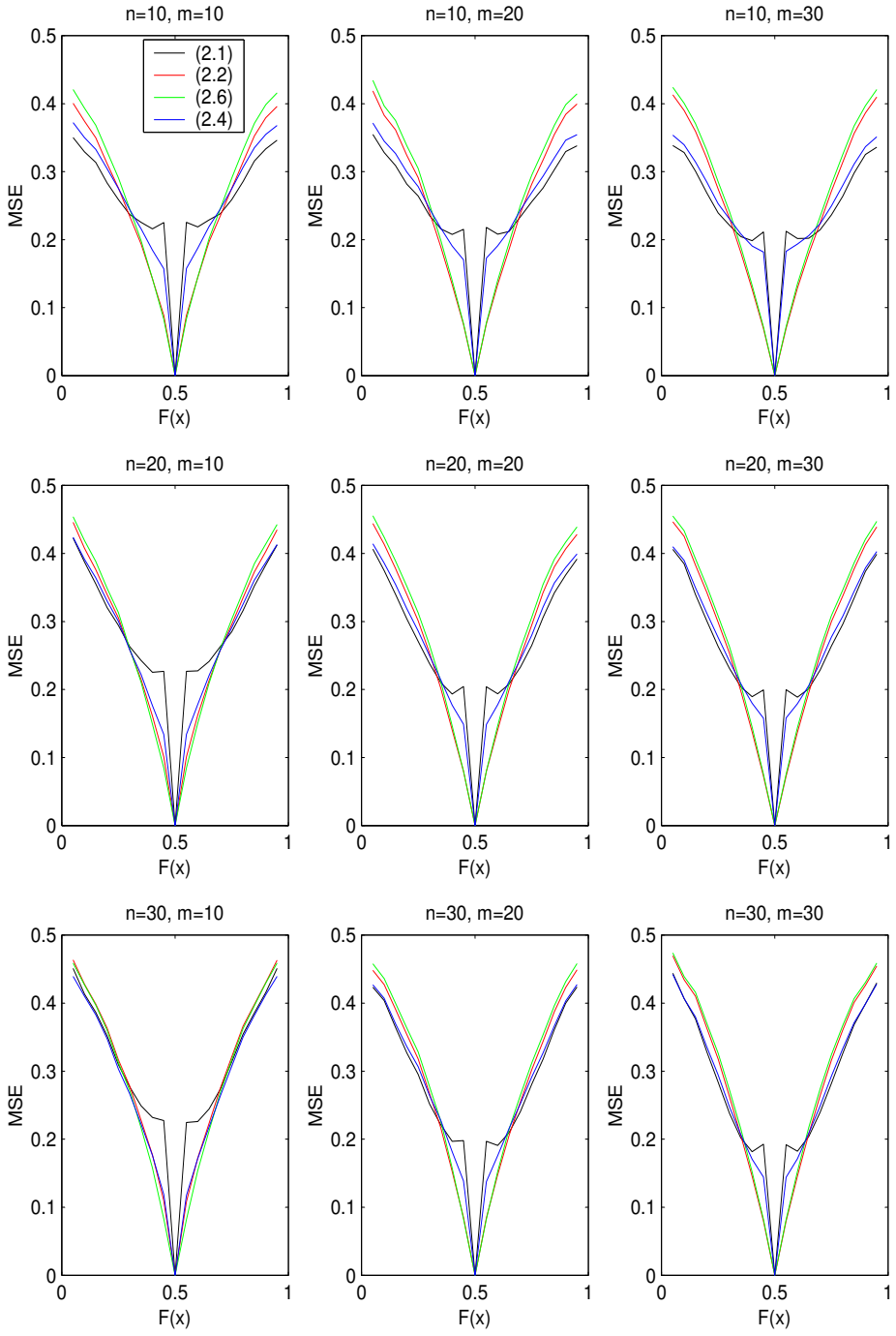


FIG 10. *Mean Squared Error of the estimators when estimating F∼Laplace(0,1) with G∼Laplace(0,2).*

Acknowledgements

The work of the first author was partially supported by NSF Grant DMS-053246, NSA Grant H98230-06-1-0099, and NSF REU Grant DMS-0552590. The second author acknowledges support by project PROMEP/103.5/08/2988. The authors are grateful to Drs Rudy Guerra of Rice University and Jonathan Cohen of University of Texas Southwestern Medical Center who made their data accessible. Their work was supported by NIH grant RO1 HL-53917.

References

[1] Arenaz, P., Bitticks L., Panell, K. H. and García, S. (1992). Genotoxic potential of crown ethers in mammalian cells; induction of sister chromatid exchanges. *Mutation Research*, **280**, 109–115.
[2] Bartoszewicz, J. (1985a). Moment inequalities for order statistics from ordered families of distributions. *Metrika*, **32**, 383–389.
[3] Bartoszewicz, J. (1985b). Dispersive ordering and monotone failure rate distributions. *Adv. Appl. Prob.*, **17**, 472–47.
[4] Bartoszewicz, J. (1986). Dispersive ordering and the total time on test transformation. *Statist. Prob. Lett.*, **4**, 285–288.
[5] Bickel, P. J. and Lehmann, E. L. (1979). Descriptive statistics for nonparametric models. IV. Spread. In *Contributions to Statistics, Jaroslav Hájek Memorial Volume*, ed. J. Jureckova. Dordrecht, Riedel, 33–40.
[6] Birnbaum, Z. W. (1948). On random variables with comparable peakedness. *Ann. Math. Statist.*, **19**, 6–81.
[7] Brown, G. and Tukey, J. W. (1946). Some distributions of sample means. *Ann. Math. Statist.*, **7**, 1–12.
[8] Doksum, K. A. (1969). Starshaped transformations and the power of rank tests. *The Annals of Mathematical Statistics*, **40**, 1167–1176.
[9] El Barmi, H. and Rojo, J. (1997). Likelihood ratio test for peakedness in multinomial populations. *J. Nonparam. Statist.*, **7**, 221–237.
[10] El Barmi, H. and Mukerjee, H. (2008). Peakedness and peakedness ordering in symmetric distributions. *Journal of Multivariate Analysis*, doi:10.1016/j.jmva2008.06.011 (In Press)
[11] Elston, R. C., Boxbaum, S. and Olson, M. (2000). Haseman and Elston Revisited. *Genetic Epidem.*, **19**, 1–17.
[12] Embury, S. H., Elias, L., Heller, P. H., Hood, C. E., Greenberg, P. L. and Schrier, S. L. (1977). Remission maintenance therapy in acute myelogenous lukemia. *Western Journal of Medicine*, **126**, 267–272.
[13] Fraser, D. A. S. (1957). *Nonparametric Methods in Statistics*. Wiley, New York.
[14] Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, **2**, 2–19.
[15] Karlin, S. (1968). *Total Positivity*. Stanford University Press, CA.
[16] Lehmann, E. L. (1955). Ordered families of distributions. *Ann Statist.*, **26**, 399–419.
[17] Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.
[18] Lehmann, E. L. (1988). Comparing location experiments. *Ann. Statist.*, **16**, 521–533.
[19] Lehmann, E. L. and Rojo, J. (1992). Invariant directional orderings. *Ann Statist.*, **20**, 2100–2110.
[20] Liu, B. H. (1988). *Statistical Genomics, Linkage, Mappings, and QTL Analyisis*. CRC Press, New York.
[21] Lo, S. H. (1987). Estimation of distribution functions under order restrictions. *Statistics and Decisions*, **5**, 251–262.
[22] Marshall, A. W. and Olkin, I. (2007). *Life Distributions: Structure of Nonparametric, Semiparametric, and Parametric Families*. Springer Science+Bussines Media, LCC, New York.
[23] Oh, M. (2004). Inference for peakedness ordering between two distributions. *J. Kor. Statist. Soc.*, **33**, 303–312.
[24] Oja, H. (1981). On location, scale, skewness, and kurtosis of univariate distributions. *Scand. J. Statist.*, **8**, 154–168.
[25] Proschan, F. (1965). Peakedness of distributions of convex combinations. *Ann. Math. Statist.*, **36**, 1703–1706.
[26] Rojo, J. and He, G. Z. (1991). New properties and characterizations of the dispersive ordering. *Statist. & Prob. Lett.*, **11**, 365–372.
[27] Rojo, J. and Wang, J. (1994). Test based on L-statistics to test the equality in dispersion of two probability distributions. *Statistics and Probability Letters*. **21**, 107–113.
[28] Rojo, J. (1995a). On the weak convergence of certain estimators of stochastically ordered survival functions. *J. Nonparam. Statist.*, **4**, 349–363.

[29] Rojo, J. (1995b). Nonparametric quantile estimation under order constraints, *J. Nonparam. Statist.*, **5**, 185-200.
[30] Rojo, J. and Ma, Z. (1996). On the estimation of stochastically ordered survival functions. *J. Statist. Compu. and Simu.*, **55**, 1–21.
[31] Rojo, J. (1998). Estimation of the quantile function of an IFRA distribution. *Scand. J. Statist.*, **25.2**, 293–310.
[32] Rojo, J. (1999). On the estimation of a survival function under a dispersive order constraint. *J. Nonparam. Statist.*, **11**, 107–135.
[33] Rojo, J. (2004) On the estimation of survival functions under a stochastic order constraint. *The First Erich L. Lehmann Symposium - Optimality*, (J. Rojo, *ed*), IMS LNMS Vol 44, 37–61.
[34] Rojo, J. and Batun-Cutz, J. (2007). Estimation of symmetric distributions subjects to peakedness order. *Series in Biostatistics Vol 3, Advances in Statistical Modeling and Inference* Chapter 13, 649–670.
[35] Rojo, J., Batun-Cutz, J. and Durazo, R. (2007). Inference under peakedness restrictions. *Statistica Sinica* **17.3**, 1165–1189.
[36] Schuster, E. (1975). Estimating the distribution function of a symmetric distribution. *Biometrika,* 62, 3, 631–635.
[37] Schweder, T. (1982). On the dispersion of mixtures. *Scand. J. Statist.*, **9**, 165–169.
[38] Shaked, M. (1980). On mixtures from exponential families. *J. R. Statist. Soc. B.*, **42**, 192–198.
[39] Shaked, M. (1982). Dispersive orderings of distributions. *J. Appl. Prob.*, **19**, 310–320.
[40] Shaked, M. and Shantikumar, J. G. (2007) *Stochastic Orders*. Springer Science+Bussines Media, LCC, New York.
[41] Shibata, T. and Takeyama, T. (1977). Stochastic theory of pitting corrosion. *Corrosion*, **33.7**, 243.

IMS Collections
Vol. 0 (2009) 174–190
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000005

A Functional Generalized Linear Model with Curve Selection in Cervical Pre-cancer Diagnosis using Fluorescence Spectroscopy

Hongxiao Zhu[1] **and Dennis D. Cox**[2]

Rice University

Abstract: A functional generalized linear model is applied to spectroscopic data to discriminate disease from non-disease in the diagnosis of cervical pre-cancer. For each observation, multiple functional covariates are available, and it is of interest to select a few of them for efficient classification. In addition to multiple functional covariates, some non-functional covariates are also used to account for systematic differences caused by these covariates. Functional principal components are used to reduce the model to multivariate logistic regression and a grouped Lasso penalty is applied to the reduced model to select useful functional covariates among multiple curves.

Contents

1 Introduction 174
2 Functional Generalized Linear Model with Curve Selection 176
3 Simulation Study 179
4 Real Data Application–Fluorescence Spectral Curve Selection and Cervical Pre-Cancer Diagnosis 181
5 Determining the Related Parameters 183
6 Discussion 186
Appendix: Proof of Proposition 1 187
Acknowledgements 189
References 189

1. Introduction

Classification with functional data is a challenging problem due to the high dimensionality of the observation space. One solution is to reduce the dimension and use the reduced features for classification, such as the work of Hall et al. [6], Zhao et al. [19] and Ferré and Villa [5]). Another way is to use generalized linear regression by treating the class labels as responses and functional observations as predictors, which was proposed by James [8] and Müller and Stadtmüller [11]. Ratcliffe et al.

[1]Department of Statistics, Rice University, 6100 Main St. MS-138, Houston, Texas 77005, U.S.A., email: hxzhu@stat.rice.edu
[2]Department of Statistics, Rice University, 6100 Main St. MS-138, Houston, Texas 77005, U.S.A., email: dcox@stat.rice.edu
AMS 2000 subject classifications: 60K35; secondary 60K37
Keywords and phrases: Functional generalized linear model, curve selection, grouped lasso, fluorescence spectroscopy, cervical cancer

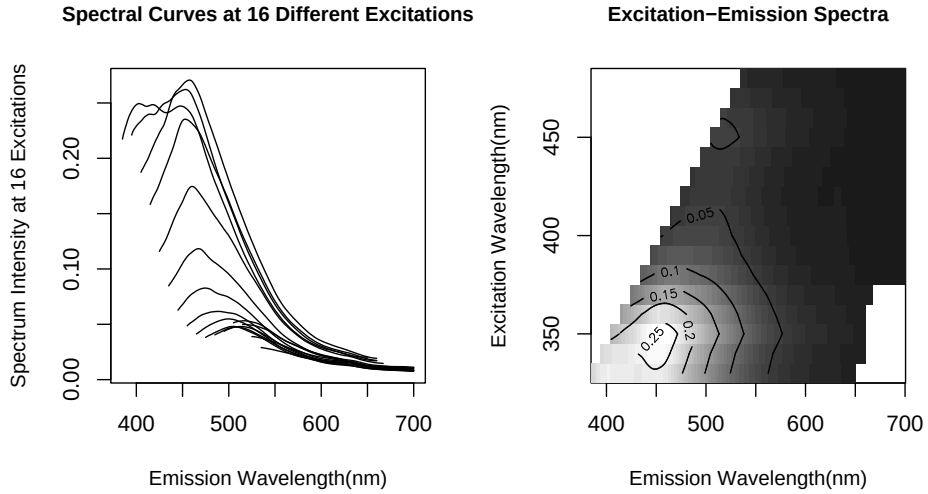


FIG 1. *Left Panel: Fluorescence spectral curves at different excitation wavelengths. Right Panel: The image plot of fluorescence spectroscopy data (EEM).*

[14] and Leng and Müller [9] applied this type of modeling to medical and gene expression data, respectively. Our basic concern in this study is the case when there are multiple functions per observation in a classification problem, and we wish to perform a curve selection to select few important curves and perform classification based on the selected curves.

The example that motivated our work is fluorescence spectroscopy data being investigated for cervical pre-cancer diagnosis. Fluorescence spectroscopy is an optical technique proposed for cervical pre-cancer screening. As a non-invasive, low-cost diagnosis tool, it provides a promising alternative to the existing methods for early-stage cancer diagnosis. One important step in this type of diagnosis is to discriminate the diseased observations from normal based on the high dimensional functional data — the fluorescence spectral measurements. In many clinical studies, several different spectra can be produced and used simultaneously for diagnosis ([12]), which makes the classification difficult since introducing more spectra not only provides more information but also more noise. Among these multiple spectral curves, it is suspected that some spectral curves contain more disease related information and hence are more "important" than others (see [3]). Furthermore, in order to produce an inexpensive commercial device, we would like to measure as few spectra as is necessary. This makes it beneficial to use statistical analysis to find out those curves that are good enough for diagnosis and remove the unnecessary ones, which can improve the diagnostic accuracy and reduce the cost.

The data studied in this paper are from a clinical study in which multiple fluorescence spectra were measured at the same sites where biopsies were taken for pathological diagnosis. Each observation consists of several spectral curves measured in the following way: an excitation light at a certain fixed excitation wavelength is produced to illuminate the cervical tissue. The excitation light is absorbed by various endogenous fluorescent molecules in tissue, resulting in emission of fluorescent light. The emitted fluorescent light is measured by an optical detector and the spectrum is obtained as one smooth curve. The excitation light is varied at several different wavelengths and gives multiple spectral curves for each measurement. The

left panel of Figure 1 shows the plot of all spectral curves from one measurement. Each measurement contains 16 spectral curves measured at excitation wavelengths ranging from 330 nm to 480 nm with increments of 10 nm. Each spectral curve contains fluorescence intensities recorded on a range of emission wavelengths between 385nm and 700nm. If we use a color plot to represent the intensities, we can stack all the 16 spectra and obtain an image as shown in the right panel of Figure 1. We call such fluorescence spectroscopy measurements excitation-emission matrices (EEMs).

This study aims to select a subset of spectral curves from the 16 available curves for the purpose of classification. We will look at the problem from the functional data analysis ([13]) point of view and propose a functional generalized linear model, which will select among multiple functional predictors and perform binary classification. The proposed model allows both functional predictors and non-functional predictors. The non-functional predictors are variables associated with the measurements which may cause systematic difference in spectra, such as tissue type of the measurement site, or the menopausal status of patients.

The structure of this paper is as follows: Section 2 introduces the functional generalized linear model with curve selection and Section 3 provides a simulation study. The real data application to the fluorescence spectroscopy data is presented in Section 4, and details on determining related parameters are discussed in Section 5. A more general discussion is given in Section 6.

2. Functional Generalized Linear Model with Curve Selection

Consider n i.i.d. observations where each observation contains J functions. For $i = 1, \ldots, n$ and $j = 1, \ldots, J$, let $x_{ij}(t)$ denote the jth function observed from the ith observation, where $E[x_{ij}(t)] = \mu_j(t)$. Note that the J functions within each observation can be rather arbitrary hence we assume different mean function $\mu_j(t)$ for each $x_{ij}(t)$. In addition to functional data, we assume there is a non-functional vector z_i associated with each observation. Suppose the responses we observed are binary variables y_i. Similarly to James [8] and Müller and Stadtmüller [11], we propose a functional generalized linear model to connect the binary responses with the predictors. Let $p_i = Pr\,(y_i = 1|z_i, x_{ij}(t), j = 1, \ldots, J)$ and

$$p_i = g^{-1}(\eta_i) \tag{2.1}$$

$$\eta_i = \alpha_0 + z_i^T \alpha + \sum_{j=1}^{J} \int_{T_j} \beta_j(t)(x_{ij}(t) - \mu_j(t))dt \tag{2.2}$$

where T_j is the domain of $x_{ij}(t)$, α_0 is the univariate intercept, α is a vector of coefficients for the non-functional predictors, and the $\beta_j(t)$'s are the functional regression coefficients. For convenience, we center $x_{ij}(t)$ at its mean in the integrand. Here the link function $g(\cdot)$ is a one-to-one continuous function. To perform curve selection, we propose the following constraint on the functional regression coefficients:

$$\sum_{j=1}^{J} ||\beta_j||_{L^2} < s \tag{2.3}$$

where $||f||_{L^2} = (\int f^2(t)dt)^{1/2}$, s is a pre-defined constant. Note that (2.3) is a combined constraint of L^2 norm and l^1 norm. This is an extension of the group-wise variable selection in the multivariate setting proposed by Yuan and Li [18].

Because of the properties of this combined constraint, we expect $\beta_j \equiv 0$ for a number of j's, depending on the value of s.

Due to the infinite dimensionality of functional data, multivariate methods can not be used directly for solving the above proposed model. One can discretize $x_{ij}(t)$ on a finite grid and transform the problem to a multivariate regression model, but the number of grid points is an issue and there will be high correlation between contiguous grid points because of the "functional" properties of $x_{ij}(t)$. A natural choice is to apply standard functional dimension reduction methods to reduce the dimension first and solve the problem on the reduced space. If we assume $\forall j$, $x_{ij}(t) \in \mathcal{H}_j$ for some separable Hilbert space $\mathcal{H}_j$, and $E[x_{ij}(t)] = \mu_j(t)$, we can expand $x_{ij}(t) - \mu_j(t)$ on a set of orthonormal basis $\{\phi_k^j\}_{k=1}^{\infty}$

$$x_{ij}(t) - \mu_j(t) = \sum_{k=1}^{\infty} c_{ijk}\phi_k^j(t) \tag{2.4}$$

and a truncated version of (2.4) can be used to approximate $x_{ij}(t)$ since $\sum_{k=1}^{\infty} |c_{ijk}|^2 < \infty$. And similarly, we assume $\beta_j(t) \in \mathcal{H}_j, \forall j$, and this gives

$$\beta_j(t) = \sum_{k=1}^{\infty} b_{jk}\phi_k^j(t) \tag{2.5}$$

Note that the orthonormal basis $\{\phi_k^j\}_{k=1}^{\infty}$ can be chosen to be a known basis such as a Fourier basis or a wavelet basis. If in addition, we assume $x_{ij}(t) \in L_2[\Omega \times T_j]$ for the domain T_j and the underlying sample space Ω, i.e., $\int_{T_j} E[x_{ij}(t)^2] < \infty, \forall j$, Mercer's theorem and Karhunen-Loève theorem ([2]) suggest taking the orthonormal basis to be the eigenfunctions of the covariance operator K, where K is defined by

$$Kx(t) = \int x(s)k(s,t)ds, \qquad k(s,t) = Cov(x(s), x(t)). \tag{2.6}$$

In this case, the coefficients $\{c_{ijk}, k = 1, \ldots, \infty\}$ are called functional principal component scores of the functional data. Using the functional principal component method is different from using a known basis in that the eigenbasis functions need to be estimated. Various estimating methods are proposed as in Ramsay and Silverman [13], and in Hall, Müller and Wang [7].

Once the functional principal component scores or the orthonormal basis coefficients have been estimated, we can reduce equation (2.2) to

$$\eta_i = \alpha_0 + z_i^T\alpha + \sum_{j=1}^{J}\sum_{k=1}^{\delta_j} c_{ijk}b_{jk} \tag{2.7}$$

where δ_j is the truncation parameter for the jth functional predictor. We thus transfer the functional regression to multivariate regression. The constraint condition (2.3) will be reduced to

$$\sum_{j=1}^{J} ||b_j||_2 < t \tag{2.8}$$

where $b_j = (b_{j1}, \ldots, b_{j\delta_j})$ and $||\cdot||_2$ stands for the Euclidean norm. Curve selection can thus be performed through selecting variables in (2.7) using the grouped

Lasso type constraint (2.8), i.e., if one curve $x_j(t)$ is selected, then the coefficients $b_{jk}, k = 1, \ldots, \delta_j$, will all be non-zero. The Lasso (Least Absolute Shrinkage and Selection Operator) was first proposed by Tibshirani [16] for model selection in linear regression models. The basic idea was to find a subset of the covariates with non-zero coefficients by applying an l_1 constraint to the regression coefficients based on the ordinary least square estimation. Yuan and Lin [18] extended the regular Lasso to cases where the covariates can be grouped, such as multi-factor ANOVA. They combine the l_1 and l_2 constraints so that the resulting model selects variables at the group level and is invariant under group-wise orthogonal transformation. To solve our problem based on the reduced model (2.7) and (2.8), we borrow the algorithm proposed by Meier et al. [10], where they extend the group-wise lasso regression of Yuan and Lin [18] to a logistic regression setup. Assume the link function in (2.1) is a logit link, i.e.,

$$\log(\frac{p_i}{1-p_i}) = \eta_i \tag{2.9}$$

The estimate can be obtained by minimizing the convex function

$$Q_\lambda(\theta) = -l(\theta) + \lambda \sum_{j=1}^{J} s(\delta_j)||b_j||_2 \tag{2.10}$$

where $\theta = \{\alpha_0, \alpha, b_j, j = 1, \ldots, J\}$, and $l(\cdot)$ is the log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} \{y_i \eta_i - \log(1 + \exp(\eta_i))\} \tag{2.11}$$

Here $s(\delta_j)$ is used to rescale the penalty with respect to the dimensionality of b_j, usually taken to be $\sqrt{\delta_j}$, and $\lambda > 0$ is the tuning parameter to control the amount of penalty. Note that in the model of Meier et al. [10], they only allow one unpenalized term, i.e., only the intercept term is unpenalized. In our proposed model, in addition to the intercept α_0, we allow the coefficients α of nonfunctional predictors to be unpenalized. Meier et al. stated the attainability of the minimum of the optimization problem in their paper and provided a proof. Actually, some conditions must be satisfied for the attainability to hold. Here we provide a general sufficient condition for the minimum of Equation (2.10) to be attained.

Proposition 1. *For* $0 < \sum_{i=1}^{n} y_i < n, \lambda > 0, s(\delta_j) > 0, \forall j$, *assume the design matrix* X *formed by*

$$X = \begin{pmatrix} 1 & z_1^T & c_{111} & \ldots & c_{11\delta_1} & \ldots & \ldots & c_{1J1} & \ldots & c_{1J\delta_J} \\ \vdots & & & & & & & & & \\ 1 & z_n^T & c_{n11} & \ldots & c_{n1\delta_1} & \ldots & \ldots & c_{nJ1} & \ldots & c_{nJ\delta_J} \end{pmatrix}$$

is an n *by* m *matrix of rank* m, $n \geq m$. *Assume the maximum likelihood estimator for the logistic regression (with log-likelihood in Equation (2.11)) exists. Then the Equation (2.10) has an unique minimizer* θ^*.

The proof for Proposition 1 is in the Appendix. Meier et al. [10] proposed a Block Coordinate Gradient Descent algorithm to solve the group lasso logistic regression and provided an R package called *grplasso*. We will use this package to perform curve selection based on reduced model in equations (2.7) and (2.8). The initiation of the algorithm is the same as in *grplasso*.

3. Simulation Study

To verify the performance of the proposed method in classification problems with multiple functional covariates, we generate $n = 1000$ i.i.d. observations. Each observation contains one non-functional covariate and three functional covariates. The non-functional covariate is generated from uniform $(0, 1)$ distribution. And the three functional covariates are generated using the first 4 cosine basis functions on the domain $[0, 1]$, i.e., using basis $\phi_0(t) = 1, \phi_k(t) = \sqrt{2}\cos(k\pi t), k = 1, \ldots, 3$. For each functional covariate, the 4 coefficients of the cosine basis are generated independently from a normal distribution with some fixed mean and variance 0.5. We set the coefficient functions for the first and third functional covariates to be zero and set the coefficient function for the second to be non-zero. Figure 2 shows the plot of both non-functional covariates and functional covariates for the first 50 observations. The binary responses y_i are generated by sampling from a Bernoulli distribution with success probability $p_i = (1 + \exp(-\eta_i))^{-1}$, where η_i is computed from equation (2.2) using numerical integration. The proportion of 1's among the binary y_i's is 57.3%. The data are randomly split into a training set and a test set with 800 observations in the training set and 200 observations in the test set.

To apply the proposed model to these data, one can choose different types of orthonormal basis for dimension reduction. Since the data are generated using cosine basis, we will show the results of using the cosine basis so that the estimated coefficients can be compared with their known true values. We have also tried using functional principal components, and the curve selection and prediction results are very similar to that of using cosine basis.

For the choice of cosine basis, we reduce the dimension of the functional predictors using the first 4 cosine basis functions. The group-wise lasso regression algorithm of Meier et al. [10] is then applied to the reduced scores. Figure 3 shows the estimation paths for the regression coefficients as a function of λ. Note that for the estimated coefficient function $\hat{\beta}_j$, we plotted their L^2 norm, i.e., $||\hat{\beta}_j|| = \sqrt{\int_{T_j} \hat{\beta}_j(t)^2 dt}$, where the function $\hat{\beta}_j$ are obtained through inverse transform of the estimated coefficients $\hat{b}_j$. From Figure 3, we see that for a large range of λ, i.e., $15.7 < \lambda < 115$, the method correctly picked out the non-zero coefficient function $\hat{\beta}_2$. The values of $\hat{\beta}_2(t)$ at 6 selected λ's is plotted in Figure 4 in comparison with the true $\beta_2(t)$. Table 1 shows the estimated coefficients(in form of the cosine basis scores $\hat{b}_j$) compared with the true values under the 6 selected λ's. From Table 1, we see that as the penalty parameter λ increases, the magnitudes of the estimated coefficients shrink toward 0. When $\lambda = 0$, the estimates are equal to the maximum likelihood estimates, which gives nonzero estimates to all coefficients. When λ ranges from 22.4 to 89.6, the coefficients corresponding to the first and third curve are exactly 0, and the coefficients corresponding to the second curve are nonzero. For $\lambda > 14.1$, the estimates are (almost all) closer to 0 than the true values. We believe that these shrinkage effects are caused by the continuous-shrinkage property of Ridge and Lasso penalty (see Tibshirani [16]). It has been suggested that there may be large bias in the estimators related to the inconsistency of the original Lasso under certain conditions, i.e., that the Lasso does not satisfy the "oracle properties" (Fan and Li [4], Zhao and Yu [20]). Some modifications have been proposed to overcome the drawbacks of Lasso and make the estimators satisfy the oracle properties(see Zou [22]). In this study, we only focus on the curve selection and prediction, but more research can be done on the consistency of the grouped-Lasso regression under the functional data setup.

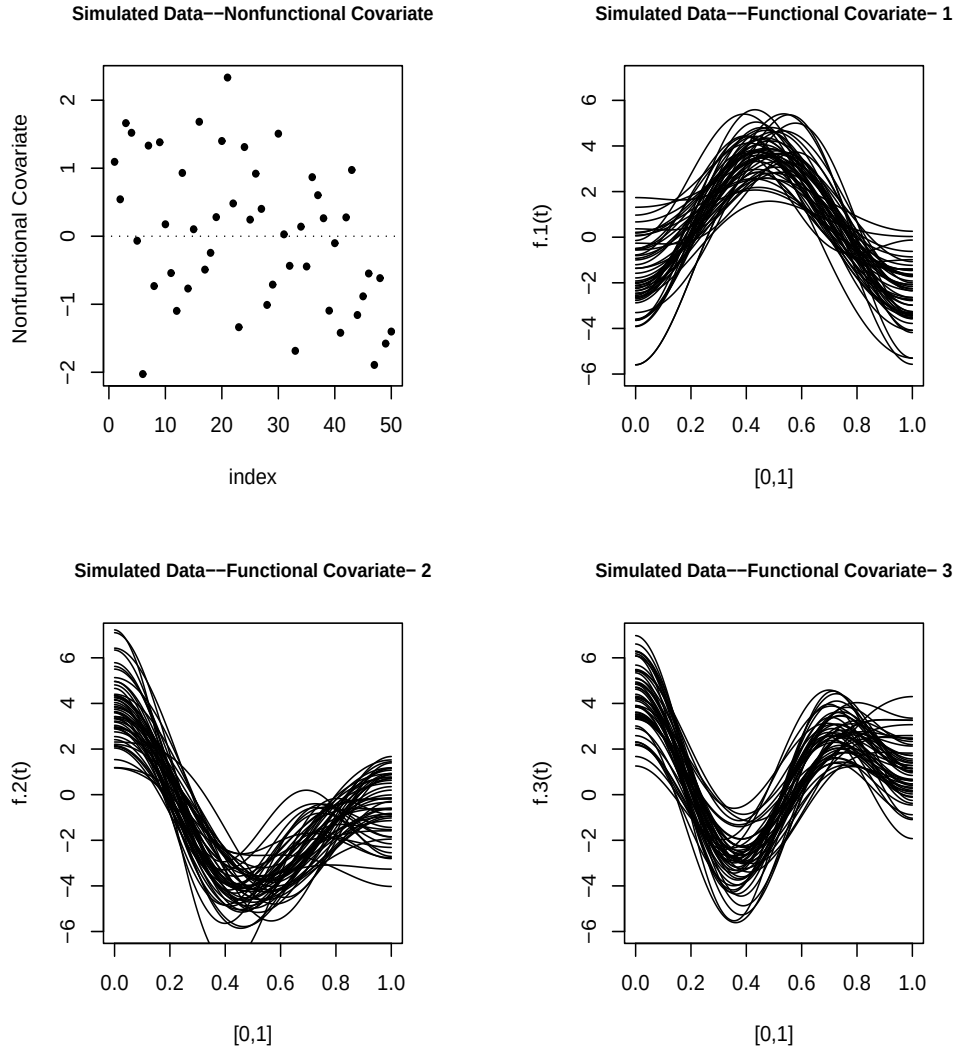


FIG 2. *Data plot of both non-functional covariates and functional covariates for the first* 50 *observations used in simulation.*

To perform prediction on test set, the estimated coefficient function $\hat{\beta}_j(t), j = 1, 2, 3$ are plugged into the test set using (2.2) and the estimated success probability $\hat{p}_i$ are computed for each observation, from which we can plot a ROC curve(see Zweig & Campbell [23]) for each λ. From each ROC curve, we pick a point that maximizes the sum of sensitivity and specificity, and this point will be used as the optimal classification point. The misclassification rate at the optimal point and the corresponding area under the ROC curves are computed at different values of λ and plotted in Figure 5. From Figure 5, we find that when λ is around 22.4, the prediction on the test set gives the best sensitivity(93%) and specificity(73%) and an fairly large area under ROC curve (0.88), and the corresponding misclassification rate is 16%.

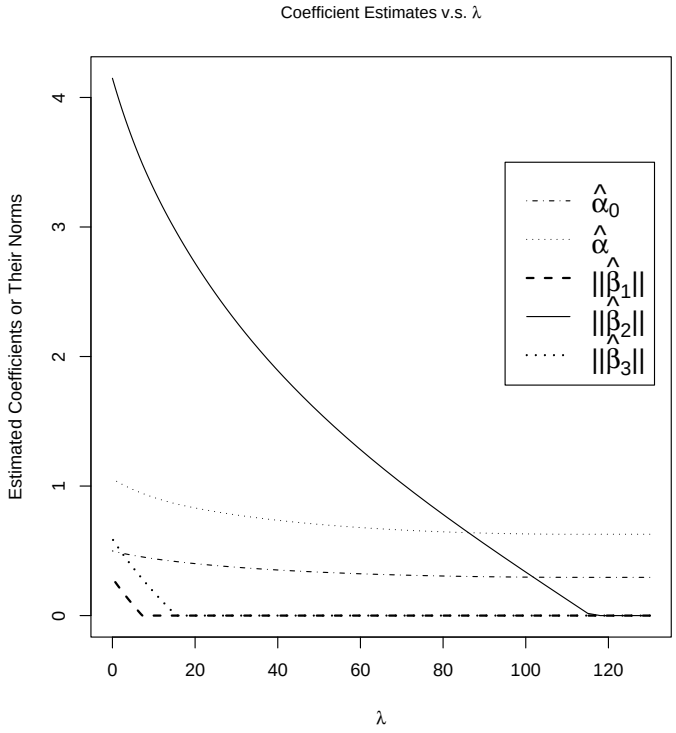


FIG 3. *Estimated paths of coefficient vector at different λ values*

4. Real Data Application–Fluorescence Spectral Curve Selection and Cervical Pre-Cancer Diagnosis

Totally 717 EEM measurements were made on 306 patients, and each measurement contains 16 spectral curves. Measurements were taken from different sites on the cervix and may include repeated measurements at the same site. All the measurements were made using the same instrument (called FastEEM3) in the same clinic (British Columbia Cancer Agency, Vancouver, CA). Data were split into a training set and a test set with 396 measurements in the training set and 321 in the test set. The proportions of diseased cases within each set are 0.21, 0.20, respectively. Two non-functional covariates are considered in this study: the colposcopic tissue type of the measurements, and the menopausal status. Colposcopic tissue type is a binary variable indicating two types of tissue — squamous and columnar, which is obtained prior the fluorescence spectroscopy measurements. Menopausal status of a patient is a categorical variable which has three levels: pre-, peri- and post-menopause. The first 5 functional principal components are chosen as the scores extracted from each functional predictor, which reduce the data to a total of 80 scores. To reduce bias, the test set scores (the scores of orthonormal basis) are computed based on only information obtained from the training set. For example, the eigenfunctions used for computing functional principal components scores of the test set are estimated from the training set.

The grouped lasso logistic regression is used to pick the excitation wavelengths as λ decreases from 6 to 0. Due to the large number of curves, the plot of coefficient path is hard to visualize. In Figure 6, we summarize the excitation spectral curves

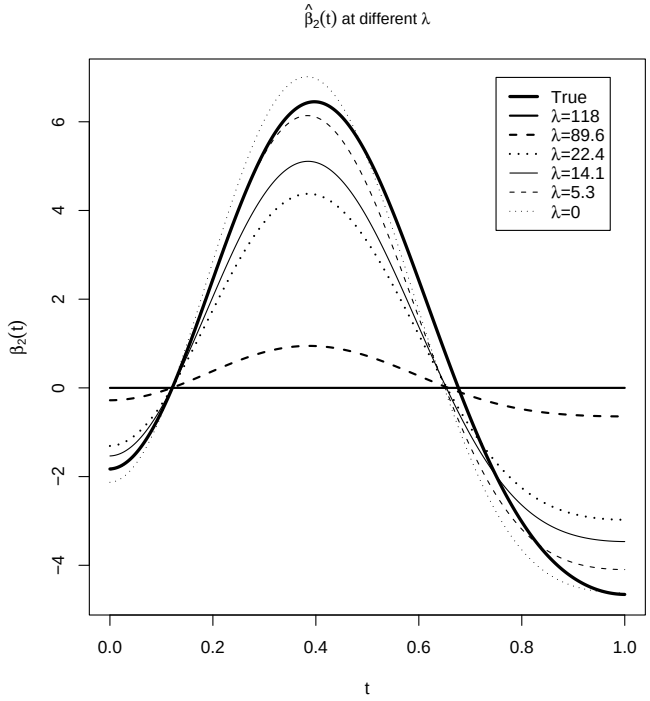


FIG 4. *Estimated coefficient function* $\hat{\beta}_2(t)$ *at* 6 *selected* λ *values and the true* $\beta_2(t)$.

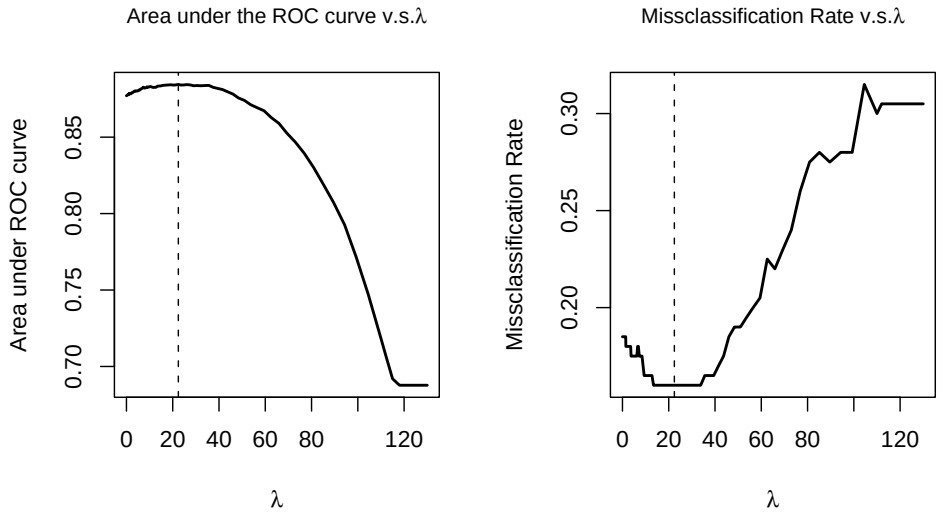


FIG 5. *Prediction results at different* λ *values.*

selected at different λ values. For example, from Figure 6 we find that when $3.73 < \lambda \leq 5.08$, the curves at excitations 340,410,420,460 are selected. At larger values of λ, the penalty is heavier, and fewer curves are selected. When $\lambda = 0$, there is no penalty, and all curves are selected. The order of selection from larger λ values to smaller λ values suggests the importance of curves in the regression. For example,

TABLE 1
The estimated coefficient values compared with the true values at different λs

		Estimated coefficients at different λ values					
Coef	*True Values*	λ=118	λ=89.6	λ=22.4	λ=14.1	λ=5.3	λ=0
α_0	0.5	0.3	0.3	0.39	0.42	0.46	0.5
α	1	0.63	0.64	0.82	0.87	0.97	1.06
b_{11}	0	0	0	0	0	0.03	0.15
b_{12}	0	0	0	0	0	-0.04	-0.17
b_{13}	0	0	0	0	0	0.04	0.18
b_{14}	0	0	0	0	0	0	-0.01
b_{21}	1	0	0.13	0.58	0.67	0.79	0.9
b_{22}	2	0	0.31	1.43	1.67	2.01	2.29
b_{23}	-3	0	-0.42	-1.92	-2.24	-2.66	-3.02
b_{24}	-1	0	-0.18	-0.84	-0.99	-1.21	-1.41
b_{31}	0	0	0	0	0	0.02	0.03
b_{32}	0	0	0	0	0.01	0.07	0.13
b_{33}	0	0	0	0	0.04	0.34	0.56
b_{34}	0	0	0	0	0.01	0.09	0.14

the excitation curves are ordered by $340 > 460 > 420 > 410$ according to the order of being selected. The estimated coefficients at different values of λ are used to predict in the test set, from where we can evaluate the performance of different λ values. Due to the fact that the total proportion of diseased cases is small, the misclassification rate is not an ideal criterion for evaluating the prediction result (see [21], page 22 for details). To reduce the risk of false negatives, we wish to keep the sensitivity high enough and sacrifice some specificity. Hence for each fixed λ, we pick a point from the empirical ROC curve using the criterion that the sum of the sensitivity and specificity is maximized. The Figure 7 shows the area under the curves and the optimal sum of the sensitivity and specificity at different values of λ. At $\lambda = 1.64$, the sum is maximized at 1.44 with sensitivity 87% and specificity 57%, and the corresponding area under ROC curve is 0.77, and misclassification rate is 38%.

Since the main purpose of the above analysis is for curve selection rather than classification, once the functional covariates are selected, different classifiers can be applied to perform classification based on the selected subset of curves. In addition to logistic regression, we also performed classification with 3 other classifiers using the selected curves. By choosing $\lambda = 1.64$, we selected function predictor curves at excitations: 330, 340, 360, 370, 410, 420, 460 and 480, and used the first 5 functional principal components to reduce the dimension. We refitted the logistic model without penalty and compared the prediction results on the test set with 3 other classifiers in Table 2. The corresponding ROC curves are plotted in Figure 8. From Figure 8, we find that logistic regression, k-nearest neighbor(KNN) and linear discriminant analysis(LDA) provide similar ROC curves. The highest sum of sensitivity and specificity is 1.43, obtained by KNN, which is only slightly smaller than the grouped lasso results at $\lambda = 1.64$. The LDA method provides the same specificity with logistic regression but higher sensitivity.

5. Determining the Related Parameters

In our proposed model, two types of parameters need to be determined: the tuning parameter λ and the truncation parameters δ_j, $j = 1, \ldots, J$. In this section, we discuss how to determine these parameters.

The choice of tuning parameter λ is important for prediction. In Meier et al. [10]

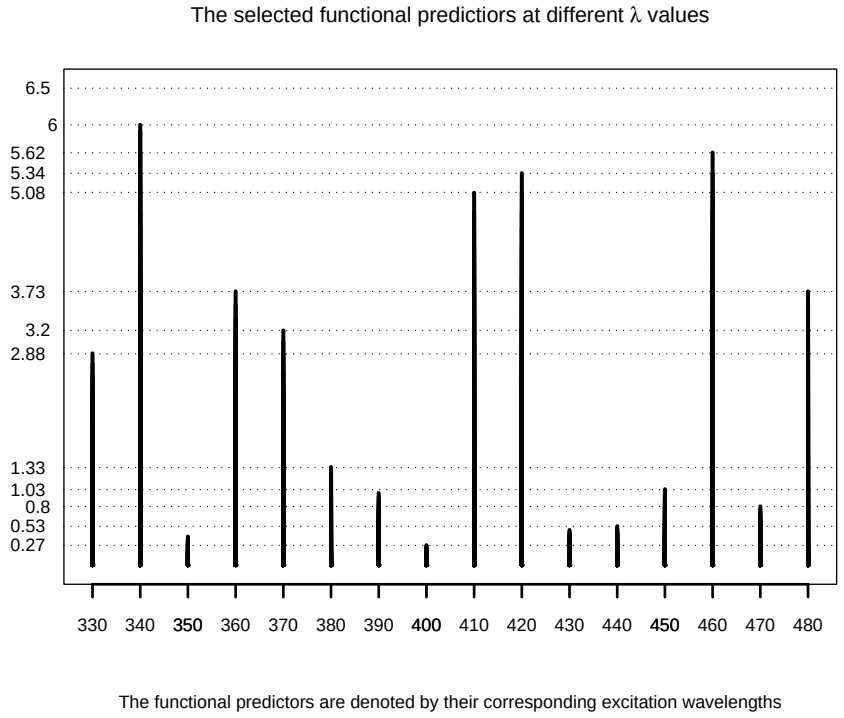


FIG 6. *The selected functional predictors (fluorescence spectral curves denoted by excitation wavelengths) at different λ values.*

and in our paper, a test set is used to choose the λ with the best prediction performance. However, there are also cases where only a small number of observations are available and splitting out a test set is not possible. In this case, we can adopt model selection criteria such as AIC, practical C_p or BIC. AIC tends to select a model with optimal prediction, while BIC tends to identify the true sparse model if the true model is included in the candidate set (see Yang [17]). In the grouped Lasso linear regression model, Yuan and Lin [18] proposed an approximation to the degrees of freedom and used a C_p criterion for selecting the tuning parameter λ. Whether this criterion can be extended to logistic regression case for selecting λ is an open question.

In addition to the tuning parameter λ, the truncation parameter δ_j in equation (2.7) is also one concern of the study. In the real application of Section 4, we let

TABLE 2

The classification results using 4 different methods on the selected curves. Auc: Area under ROC curve. MisR: Misclassification rate. Sens: Sesitivity. Speci: Specificity. Sum: The sum of sensitivity and specificity. Logistic: logistic regression. KNN: k-nearest neighbor. LDA: linear discriminant analysis. SVM: support vector machine.

Method	Auc	MisR	Sens	Speci	Sum
Logistic	0.76	31%	71%	68%	1.39
KNN	0.68	27%	68%	74%	1.43
LDA	0.75	31%	75%	68%	1.42
SVM	0.64	28%	48%	79%	1.26

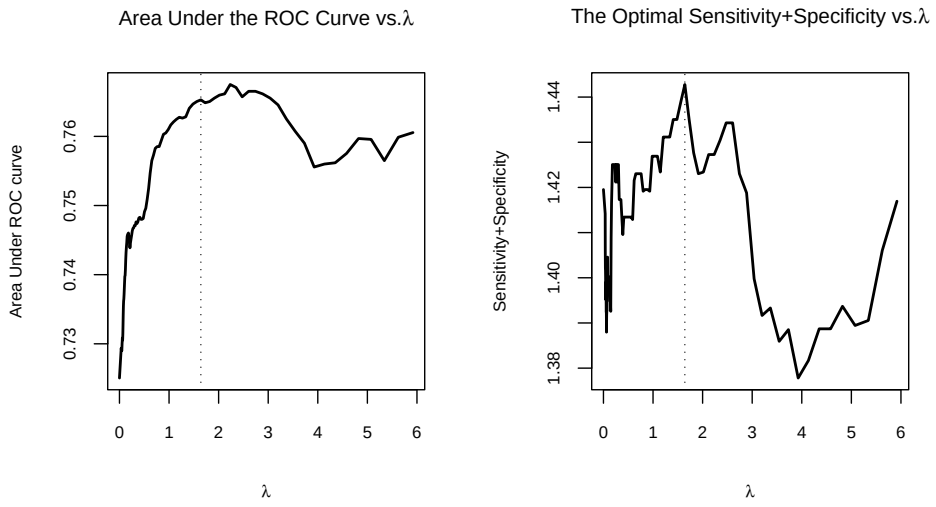


FIG 7. *Prediction results at different λ values.*

$\delta_j \equiv \delta$ and reported the curve selection and prediction results with $\delta = 5$. To find out whether other choices of δ are better for prediction, we compute the prediction results for the test set at different number of δ but fixing $\lambda = 1.64$. The quantitative prediction results are plotted in Figure 9. From Figure 9, we can see that using 11 functional principal components, the area under the ROC curve are maximized at 0.780, and the sum of sensitivity and specificity are maximized at 1.47, with a relatively small misclassification rate 31%. The sensitivity and specificity reach 81%, 66%, respectively.

It is also suspected that the optimal λ may interact with δ so determining one by fixing the other may be suboptimal. In our study, we also have tried to determine both the parameters by training the model under different combinations of them, and predicting on the test set. It turns out that at around $\lambda = 1.64$ the prediction results of the model is better than other choices of λ, and this is quite stable across different choice of δ, especially for δ greater than 3. In Figure 10, We plot the area under the ROC curve for 11 different δ and for appropriately selected λ values across a meaningful range, i.e., $\lambda = (5, 3, 1.64, 1.5, 1, 0.27)$. It shows that the line with $\lambda = 1.64$ stays on the top for δ's larger than 3. The reason for the small interaction between λ's and δ's can be the following: the orthogonal basis approximation tends to be accurate with only a few components. For example, in functional principle components, over 97% of the variability will be counted in the first priciple component score for all excitation curves. Later components only add details to the model but does not change the likelihood dramatically. Therefore the miminum of Equation (2.10) as a function of λ does not change much when δ changes. But this is not true for non-orthogonal basis approximation methods such as B-spline.

Note that choosing $\delta_j \equiv \delta$ is just a convenient choice, which has the advantage that it leaves only two parameters to determine and cross validation is feasible for determining these parameters. However, it also brings in the risk of loosing information. In general, one may use different truncation parameters if there are large differences on the properties of the curves such as smoothness. If all curves are obtained through similar sources and are similar in shape and other above

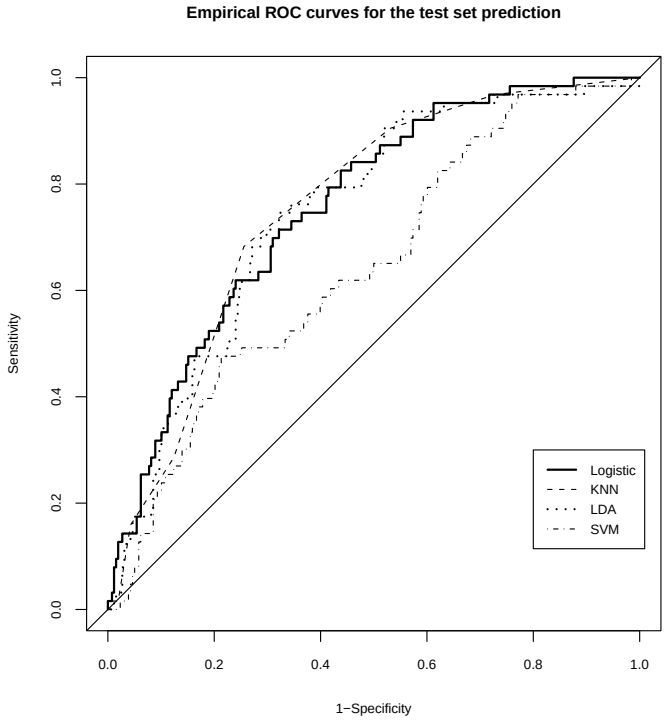


FIG 8. *ROC curves obtained when training 4 different classifiers based on selected curves and predicting on the test set.*

mentioned properties, it would be safe to choose a common δ. As an alternative, since the step of estimating $\{c_{ijk}, k = 1, \ldots, \delta_j, j = 1, \ldots, J\}$ can be independent of the group-wise Lasso step, one can use approximation criteria such as error sum of squares(SSE) to determine the truncation parameters for each curve. For example, if using functional principal component, we can choose a level of approximation(e.g., let the percentage of variabilities explained to be greater than 99%) and select the number of eigenfunctions to achieve this. However, better approximation does not necessarily give better prediction.

6. Discussion

We have proposed a functional logistic regression model to perform classification and curve selection. This model automatically selects among the functional covariates through the grouped Lasso variable selection. The proposed model gives information about which curves will be selected if we are willing to use a subset of the functional covariates for classification. For example, under penalty $\lambda = 5$, the best four functional predictors selected in our real data application are curves at excitations 340, 410, 420 and 460. The selected functional covariates can then be used with different classifiers for accurate classification.

There are several aspects that can be studied in more detail. Firstly, the basis expansion step can be combined more tightly with the grouped Lasso regression step using techniques similar to Müller and Stadtmüller [11]. It is necessary to investigate the consistency properties of the estimated coefficient function $\beta_j(t)$'s,

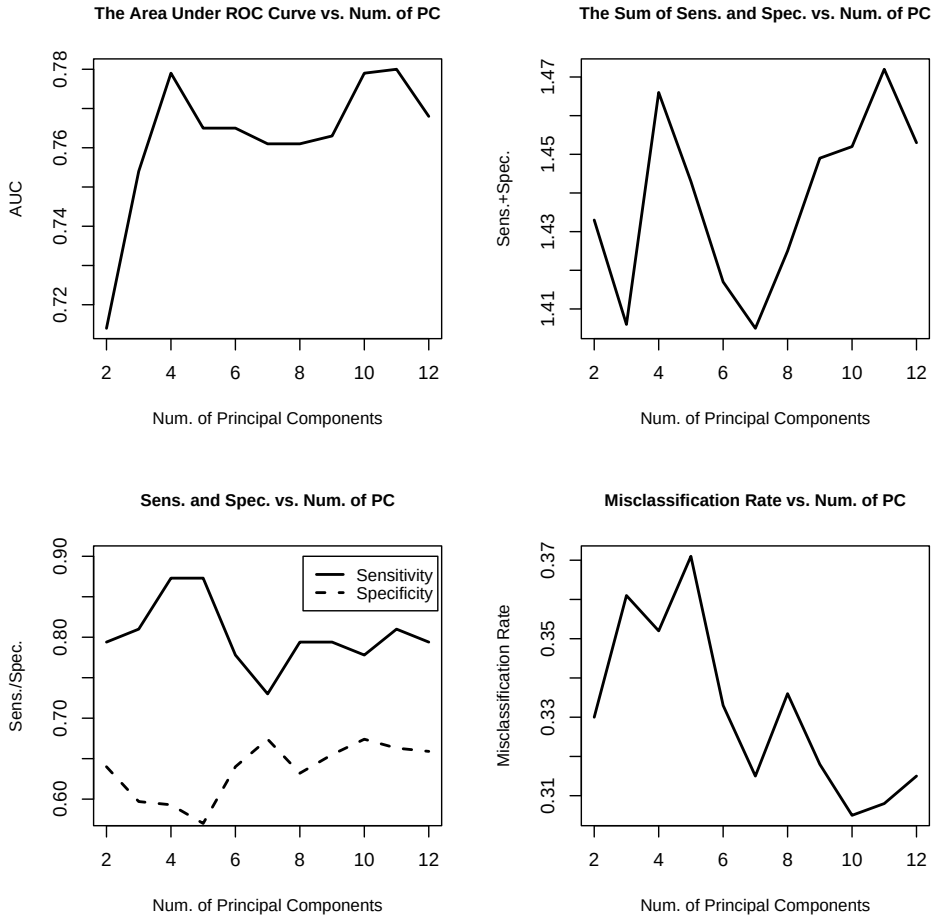


FIG 9. *Prediction results using different number of basis.*

such as the oracle property. The algorithm of Meier et al. [10] requires that the tuning parameter λ to be predefined on a grid of values, where they proposed a way to find the range of λ of interest. This method, although faster, makes it difficult to find a precise λ value that is optimal for prediction purposes. Efficient algorithms for searching λ is of great importance especially when functional data is involved.

Alternative methods for curve selection can be formulated through the Bayesian paradigm. Bayesian variable selection models can be derived for selecting variables at a group level and thus can be used for curve selection as well.

Appendix: Proof of Proposition 1

The proof of Proposition 1 uses a result stated in the following lemma.

Lemma 1. *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a strictly convex function with a minimizer $\tilde{x}$, and let $g : \mathbb{R}^n \mapsto [0, \infty)$ be a convex function. Then $f + g$ has a unique minimizer x^* in $\mathbb{R}^n$.*

Proof. Let $h(x) = f(x) + g(x)$. It is easy to show that $h(x)$ is strictly convex from

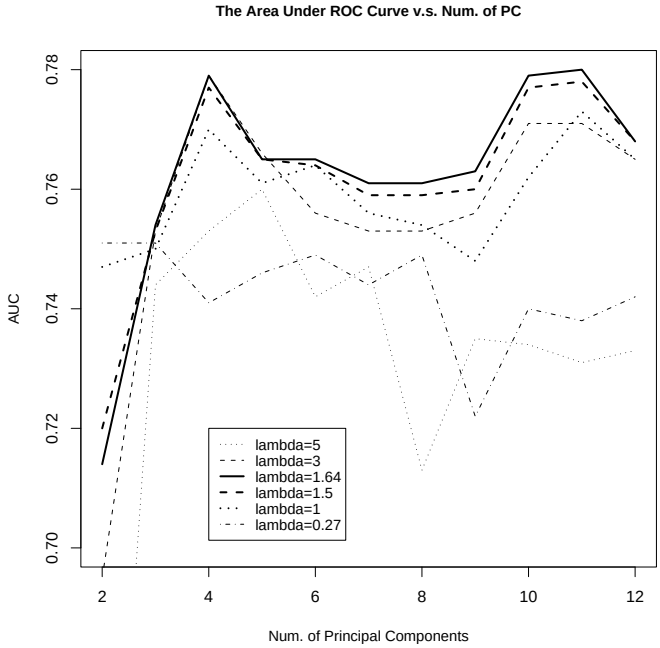


FIG 10. *The area under the ROC curves for 6 different* λ *values and* 11 *different choice of FPC's.*

the definition. We claim that the existence of a minimizer $\tilde{x}$ of f implies that h is coercive, which means $h(x) \to \infty$ as $||x|| \to \infty$. The coerciveness and strict convexity of h implies the existence of a unique minimizer x^*.

To show that h is coercive, it is sufficient to show that f is coercive (since $g \geq 0$). The minimizer $\tilde{x}$ of f is the unique minimizer of f by strict convexity. Also, f is convex hence is continuous on $\mathbb{R}^n$ (see [15], page 82). Thus $\forall\ r > 0, \forall\ x$ such that $||x - \tilde{x}|| > r$, we claim

$$f(x) > \frac{b}{r}||x - \tilde{x}|| + f(\tilde{x})$$

where $b = \inf\{f(x) : ||x - \tilde{x}|| = r\} - f(\tilde{x})$. Note that b exists and $b > 0$ by continuity of f. To show this inequality, let $x_0 = r(x - \tilde{x})/(||x - \tilde{x}||) + \tilde{x}$, so that x_0 lies on the line formed by x and $\tilde{x}$, with $||x_0 - \tilde{x}|| = r$ and $||x - x_0|| = ||x - \tilde{x}|| - r$. Thus $f(x_0) - f(\tilde{x}) \geq b$ by the definition of b. Now let $\alpha = r/||x - \tilde{x}||$. We see that $x_0 = \alpha x + (1 - \alpha)\tilde{x}$. By strict convexity of f,

$$f(x_0) < \alpha f(x) + (1 - \alpha) f(\tilde{x})$$

Thus

$$\begin{aligned} \frac{b}{r}||x - \tilde{x}|| + f(\tilde{x}) &\leq (f(x_0) - f(\tilde{x}))\frac{||x - \tilde{x}||}{r} + f(\tilde{x}) \\ &< (\alpha f(x) + (1 - \alpha) f(\tilde{x}) - f(\tilde{x}))\frac{||x - \tilde{x}||}{r} + f(\tilde{x}) \\ &= f(x) \end{aligned}$$

Since $||x - \tilde{x}|| \geq ||x|| - ||\tilde{x}||$, $||x|| \to \infty$ implies $||x - \tilde{x}|| \to \infty$, which implies $f(x) \to \infty$ by the above inequality and the facts that $b > 0, r > 0$, $f(\tilde{x})$ finite. Therefore, f is coercive, and so is h.

Since h is coercive, we have $h(x) \to \infty$ as $||x|| \to \infty$. Therefore, if we pick an arbitrary point $x_1 \in \mathbb{R}^n$, there exists a constant $\delta > 0$ such that $h(x) > h(x_1)$ for all $||x - x_1|| > \delta$. Since the domain $||x - x_1|| \leq \delta$ is compact and $h(x)$ is strictly convex on it, $h(x)$ has a unique minimizer in $||x - x_1|| \leq \delta$, which we denote as x^*. (A strictly convex real valued function defined on a compact domain has a unique minimum on its domain.) This x^* is also the global minimizer since $h(x) > h(x_1) \geq h(x^*)$ on $||x - x_1|| > \delta$.

Proof of Proposition 1: Based on results in Lemma 1, we let f to be $-l(\theta)$ and g to be $\lambda \sum_{j=1}^{J} s(\delta_j)||b_j||_2$, therefore our objective function in Equation (2.10) is the sum of f and g, where $\theta = \{\alpha_0, \alpha, b_j, j = 1, \ldots, J\}$, and $l(\theta) = \sum_{i=1}^{n} y_i \eta_i - \log(1 + \exp(\eta_i))$ with $\eta_i = \alpha_0 + z_i^T \alpha + \sum_{j=1}^{J} \sum_{k=1}^{\delta_j} c_{ijk} b_{jk}$.

Firstly, we show that $-l(\theta)$ is strictly convex. It is sufficient to show that its Hessian is positive definite. Since the Hessian takes the form

$$\bigtriangledown_\theta^2(-l(\theta)) = X^T D X$$

where $D = \text{diag}\{\exp(\eta_i)/(1 + \exp(\eta_i))^2, i = 1, \ldots, n\}$. It is positive definite since X is of rank m (full rank). Secondly, since the maximum likelihood estimator exists, $-l(\theta)$ has an unique minimizer. The existence of maximum likelihood estimator for logistic regression requires some conditions for the design matrix X. Basically, the n rows of X can not be completely separated or quasi-completely separated in $\mathbb{R}^m$. See [1] for details. In practice, as long as we can find a numerical solution for the MLE at $\lambda = 0$, we would believe that the maximum likelihood estimator exists. Finally, let $g(b) = \lambda \sum_{j=1}^{J} s(\delta_j)||b_j||_2$, $b^T = (b_1^T, \ldots, b_J^T)$. It is easy to see that $g(b)$ is convex by the triangle inequality. Therefore by Lemma 1, $Q_\lambda(\theta)$ has a unique minimizer θ^*.

Acknowledgements

This research was supported by the National Cancer Institute grant PO1-CA82710 and by the National Science Foundation grant DMS0505584. We thank the referees for constructive comments, and thank Dr. Wotao Yin for helpful discussions on the convex optimization.

References

[1] Albert, A and Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**, 1–10.
[2] Ash, R. B. (1975). *Topics in Stochastic Processes*. Academic Press, New York.
[3] Chang, S. K., Follen, M., Malpica, A., Utzinger, U., Staerkel, G., Cox, D., Atkinson, E. N., MacAulay, C. and Richards-Kortum, R. (2002). Optimal excitation wavelengths for discrimination of cervical neoplasia. *IEEE Transactions on Biomedical Engineering*, **49**, 1102–1110.
[4] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348–1360.
[5] Ferré, L. and Villa, N. (2006). Multilayer perceptron with functional inputs: an Inverse Regression Approach. *the Scandinavian Journal of Statistics*, **33**, 807–823.
[6] Hall, P., Poskitt, D. S. and Presnell, B. (2001). A Functional data-analytic approach to signal discrimination. *Technometrics*, **43**, 157–214.
[7] Hall, P., Müller, H. and Wang, J. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist*, **34**, 1493–1517.
[8] James, G. M. (2002). Generalized linear models with functional predictors. *J. R. Stat. Soc. Ser. B*, **64**, 411–432.
[9] Leng, X. and Müller, H. (2005). Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, **22**, 68–76.

[10] Meier, L., Geer, S. and Bühlmann, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B*, **70**, 53–71.
[11] Müller, H. and Stadtmüller, U. (2005). Generalized functional linear models. *Ann. Statist*, **33**, 774–805.
[12] Ramanujam, N., Mitchell, M. F., Mahadevan, A., Thomsen,S., Malpica, A., Wright,T., Atkinson, N. and Richards-Kortum, R. (1996). Spectorscopic diagnosis of cervical intraepithelial neoplasia(CIN) in vivo using laser induced fluorescence spectra at multiple excitation wavelengths. *Lasers Surg. Med.*, **19**, 63–67.
[13] Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. Springer-Verlag, New York.
[14] Ratcliffe, S. J., Heller, G. Z. and Leader, L. R. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression. *Statistics in Medicine*, **21**, 1115–1127.
[15] Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.
[16] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288.
[17] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, **92**, 937–950.
[18] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B*, **68**, 49–67.
[19] Zhao, X., Marron, J. S. and Wells, M. T. (2004). The functional data analysis view of longitudinal data. *Statistica Sinica*, **4**, 789–808.
[20] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Machine Learning Research*, **7**, 2541–2563.
[21] Zhou, X. and Obuchowski, N. A. (2002). *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, Inc., New York.
[22] Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101**, 1418–1429.
[23] Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39**, 561–577.

IMS Collections
Vol. 0 (2009) 191–216
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000018

Nonparametric Estimation of Hemodynamic Response Function: A Frequency Domain Approach

Ping Bai[1], **Young Truong**[2] **and Xuemei Huang**[3]

University of North Carolina at Chapel Hill

Abstract: Hemodynamic response function (HRF) has played an important role in many recent functional magnetic resonance imaging (fMRI) based brain studies where the main focus is to investigate the relationship between stimuli and the neural activity. Standard statistical analysis of fMRI data usually calls for a "canonical" model of HRF, but it is uncertain how well this fits the actual HRF. The objective of this paper is to exploit the experimental designs by modeling the stimulus sequences using stochastic point processes. The identification of the stimulus-response relationship will be conducted in the frequency domain, which will be facilitated by fast Fourier transforms (FFT). The usefulness of this approach will be illustrated using both simulated and real human brain data. Under regularity conditions, it is shown that the estimated HRF possesses an asymptotic normal distribution.

Contents

1 Introduction 192
2 A Frequency Domain Method for Estimating HRF 193
3 A Brief Survey of HRF Modeling 195
4 Simulated Numerical Results for HRF estimation 197
5 A Real Data Analysis 197
5.1 Experiment Paradigm and Data Description 197
5.2 Analysis and Results 198
5.3 Implications 201
5.4 Discussions 202
6 Concluding Remarks 203
7 Sampling Properties of the Estimates 205
7.1 Point Process 205
7.2 Stationary Time Series 207
7.3 Cumulants and Spectra 207
7.4 Fast Fourier Transforms 208
7.5 Complex Normal 210
7.6 Asymptotics for Periodograms 210

[1]Department of Statistics and Operational Research, University of North Carolina, Chapel Hill, NC 27599-3260, email: pingbai@gmail.com
[2]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-3260, email: truong@bios.unc.edu
[3]Department of Neurology, University of North Carolina, Chapel Hill, NC 27599-3260, and Department of Neurology, Milton S. Hershey Medical Center H037, 500 University Drive, Penn State University, email: xhuang3@hmc.psu.edu

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35
Keywords and phrases: fMRI, hemodynamic response function, transfer function model, point process, spectral analysis

7.7 Window Estimates — The Smoothed Periodograms 211
7.8 Estimating the Transfer Function 212
7.9 Estimating the Hemodynamic Response Function 213
Acknowledgements 214
References 214

1. Introduction

Consider a statistical problem in which data are acquired by applying stimuli at times $\tau_1 < \tau_2 < \cdots$ and simultaneously a varying response $Y(t)$ is recorded. Suppose it is desired to conduct the associated statistical inference based on the model:

$$Y(t) = \sum_j h(t - \tau_j) + \epsilon(t), \tag{1.1}$$

where $h(\cdot)$ is an unknown function and $\epsilon(t)$ is a stationary, zero mean, noise series with power spectrum given by $s_{\epsilon\epsilon}(\cdot)$. It is assumed that the function $h(t) = 0$ for $t < 0$ and will have finite duration.

This type of problems has played an important role in the fields of psychology, neurobiology, neurology, radiology, biomedical engineering, and many others, where data acquisition is carried out in functional magnetic resonance imaging (fMRI) experiments. As a noninvasive technique, fMRI allows us to study dynamic physiological processes at a time scale of seconds. The basis of fMRI is the Blood Oxygenation Level Dependent (BOLD) effect [40]. Due to differential magnetic susceptibility of oxygenated (oxygen-rich) hemoglobin and deoxygenated hemoglobin, the BOLD effect reflects the changes in hemodynamics which in turn yields greater MRI intensity when brain activity increases (see [30]). It is this hemodynamic response to the underlying neuronal activity that makes the fMRI signal in brain areas of activation a blurred and delayed version of the stimuli. Figure 1 shows the recorded BOLD signals (solid line) triggered by a single event (dashed line on the left panel) and a sequence of consecutive of stimuli (dashed line on the right panel) respectively. Both of them show the blur-and-delay effect caused by the hemodynamic response.

In practice, the BOLD effect is modeled through the convolution of the stimulus sequence $X(\cdot)$ and a hemodynamic response function (HRF) $h(\cdot)$ given by

$$\text{BOLD}(t) = h \otimes X(t) = \int h(t-u)X(u)du, \tag{1.2}$$

whose discrete time version is represented by model (1.1). Typically, an fMRI dataset consists of a 3D grid of voxels, each containing a time series of measurements that reflect brain activity. For each of roughly 200,000 voxels that lie inside the brain images, we wish to carry out the estimation of the HRF which will subsequently be applied to infer those voxels that were activated under certain experimental conditions.

The rest of this paper is organized as follows. Our estimate is described in Section 2, which is based on frequency domain methods applied to the point processes and ordinary time series. A brief survey of HRF modeling is provided in Section 3. Section 4 illustrates the performance of our proposed method through a simulated data analysis. For the purpose of comparison, a real fMRI data set is analyzed using the proposed method and a popular fMRI tool in Section 5. Discussions and concluding remarks are given in Section 6. Proofs are given in the last section of the paper.

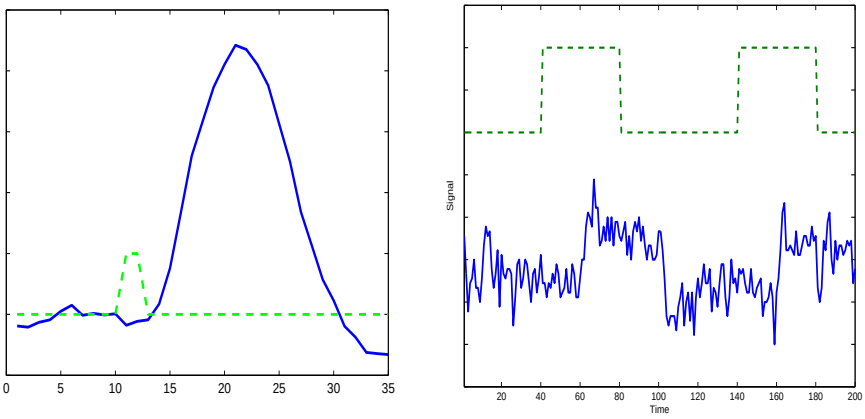


FIG 1. *Left Panel: The recorded BOLD signal (solid line) triggered by a single event (dashed line). Right Panel: The recorded BOLD signal (solid line) triggered by a typical block-design sequence (dashed line).*

2. A Frequency Domain Method for Estimating HRF

Model (1.1) has the structure of linear time invariant system carrying the stimuli $X(t)$ onto an response time series $Y(t)$. These models are generally studied by the frequency domain methods based on cross-spectral analysis (see [10]).

Define the *system transfer function* by

$$H(f) = \sum_u h(u) \exp(-iuf), \qquad f \in \mathbb{R}.$$

Define the finite Fourier transform of $Y(t)$ by

$$\varphi_0^T(f) \equiv \varphi_Y^T(f) = \sum_{t=0}^{T-1} \exp(-ift) Y(t)$$

with a similar definition for $\varphi_\epsilon^T(f)$, $f \in \mathbb{R}$. Also, define

$$\varphi_1^T(f) \equiv \varphi_X^T(f) = \sum_{t=0}^{T-1} \exp(-ift) X(t) = \sum_j \exp(-if\tau_j), \qquad f \in \mathbb{R},$$

the last sum is over the available stimuli before time $T-1$. It follows from (1.1) that

$$\varphi_0^T(f) = H(f)\varphi_1^T(f) + \varphi_\epsilon^T(f), \qquad f \in \mathbb{R}. \tag{2.1}$$

Now let m_f denote the integer $m \in \{0, 1, \ldots, T-1\}$ such that $2\pi m/T$ is closest to the (angular) frequency $f \in (0, \pi/2)$. Let K denote a positive integer. Then, for smooth $H(\cdot)$,

$$\begin{aligned} \varphi_0^T\left(\frac{2\pi}{T}(m_f + k)\right) &\approx H(f)\varphi_1^T\left(\frac{2\pi}{T}(m_f + k)\right) \\ &\quad + \varphi_\epsilon^T\left(\frac{2\pi}{T}(m_f + k)\right), \quad k = 0, \pm 1, \ldots, \pm K, \end{aligned} \tag{2.2}$$

hold for the $2K+1$ nearest frequencies around f of the form $2\pi(m_f+k)/T$. Thus a reasonable estimate of $H(f)$ can be obtained by regressing $\varphi_0^T(2\pi(m_f+k)/T)$ on $\varphi_1^T(2\pi(m_f+k)/T)$ for $k=0,\pm1,\ldots,\pm K$, which is given by

$$\hat{H}(f)=\hat{s}_{01}(f)/\hat{s}_{11}(f), \qquad f\in\mathbb{R}, \tag{2.3}$$

where

$$\hat{s}_{jj'}(f)=(2K+1)^{-1}\sum_{k=-K}^{K}\tilde{s}_{jj'}\left(\frac{2\pi}{T}(m_f+k)\right), \tag{2.4}$$

$$\tilde{s}_{jj'}(f)=(2\pi T)^{-1}\varphi_j^T(f)\overline{\varphi_{j'}^T(f)}, \qquad f\in\mathbb{R}, \quad j,j'\in\{0,1\}. \tag{2.5}$$

Here $\overline{a}$ is the conjugation of $a\in\mathbb{C}$. This is similar to the linear regression setting and therefore the residual sum of squares (RSS) is given by

$$\hat{s}_{\epsilon\epsilon}(f)=\frac{2K+1}{2K+1-1}\left(\hat{s}_{00}(f)-\hat{s}_{01}(f)\hat{s}_{11}^{-1}(f)\hat{s}_{10}(f)\right). \tag{2.6}$$

Note that

$$\hat{s}_{\epsilon\epsilon}(f)\propto\hat{s}_{00}(f)\Big(1-\frac{|\hat{s}_{01}(f)|^2}{\hat{s}_{11}(f)\hat{s}_{00}(f)}\Big)=\hat{s}_{00}(f)(1-R_{01}^2(f)),$$

where

$$|\hat{R}_{01}(f)|^2=\frac{|\hat{s}_{01}(f)|^2}{\hat{s}_{00}(f)\hat{s}_{11}(f)}, \qquad f\in\mathbb{R},$$

is the squared coherence, which lies between 0 and 1, the closer it is to 1 the stronger is the linear relationship between the two series.

Let $s_{\epsilon\epsilon}(\cdot)$ denote the power spectrum of the noise series. It can be shown that (see Section 7) the estimate $\hat{H}(f)$ is asymptotically complex normal with mean $H(f)$ and variance $s_{\epsilon\epsilon}(f)/K\hat{s}_{11}(f)$. And $\hat{H}(f_1)$, $\hat{H}(f_2)$, …, $\hat{H}(f_M)$ are asymptotically independent normal for distinct f_1, f_2, …, f_M [7].

In practice, we use a smoother estimate known as window estimate by observing that (2.4) can be written more generally as

$$\hat{s}_{jj'}(f)=\sum_{k\neq0}b^{-1}W\left(b^{-1}\left(f-\frac{2\pi k}{T}\right)\right)\tilde{s}_{jj'}\left(\frac{2\pi k}{T}\right), \qquad f\in\mathbb{R}, \tag{2.7}$$

where $W(\cdot)$ is a non-negative function called the weight or window function, and $b\equiv b_T\searrow0$ is the smoothing parameter. It has been shown that (2.7) has better sampling properties than (2.4) as an estimate of the cross-spectrum of the bivariate time series. See [10] and Section 7. From now on, our estimate of $H(f)$ will be based on (2.3) and (2.7).

We remark that when $j=j'$, then (2.5) becomes

$$\tilde{s}_{jj}(f)=(2\pi T)^{-1}\varphi_j^T(f)\overline{\varphi_j^T(f)}=(2\pi T)^{-1}|\varphi_j^T(f)|^2, \qquad f\in\mathbb{R}, \tag{2.8}$$

which is the *periodogram* of the series $Y(t)$ when $j=0$, or of the series $X(t)$ when $j=1$. The periodogram is an important statistic in spectral time series analysis.

Under certain conditions, $\hat{R}_{01}(f)$ is asymptotically normal with mean $R_{01}(f)$ and variance proportional to constant $(1-R_{01}^2(f))/(Tb)$. Moreover, if $R_{01}(f)=0$, then

$$F(f)=\frac{c|\hat{R}_{01}(f)|^2}{1-|\hat{R}_{01}(f)|^2}\sim F_{2,2c}, \tag{2.9}$$

where $c = (bT/\gamma) - 1$ and $\gamma = \int \lambda^2$ with λ being the lag-window generator [39].

This result can be used to test for a response to the stimulus by computing a test statistic for significant activation $F(f_a)$ at the fundamental frequency of activation f_a. Under the null hypothesis of no activation, the F-statistic at the fundamental frequency of activation, $F(f_a)$, has a F distribution with 2 and $2c$ degrees of freedom. Large values of $F(f_a)$ indicate a large effect at the fundamental frequency.

The estimate of the impulse response function $h(\cdot)$ is then given by

$$\hat{h}(u) = \frac{1}{Q}\sum_{q=0}^{Q} \hat{H}\left(\frac{2\pi q}{Q}\right) \exp\left(i\frac{2\pi uq}{Q}\right),$$

where $Q \equiv Q_T$ denote a sequence of positive integers tending to ∞ with T. Under certain conditions, $(\hat{h}(u_1), \ldots, \hat{h}(u_J))$ is asymptotically normal with mean $(h(u_1), \ldots, h(u_J))$ and covariance matrix

$$\frac{2\pi}{bT}\int W(\lambda)^2 d\lambda \cdot \frac{1}{Q^2}\int \exp\left(i(u_j - u_k)\lambda\right)\frac{s_{\epsilon\epsilon}(\lambda)}{s_{11}(\lambda)}d\lambda, \qquad j,k = 1,2,\ldots J.$$

See Section 7 for more details.

3. A Brief Survey of HRF Modeling

The basis of model (1.1) is the linearity of BOLD fMRI responses when multiple stimuli are presented in succession. This was first studied by Boynton, et al. [6]. The linearity arises from the fact that a stimulus induces the neural activity in a specific region of the brain. This then brings blood flow changes (hemodynamics) in that region, while BOLD fMRI responses are measured from these blood flow changes. In addition to giving this clear picture of how BOLD fMRI works, the linear transform model is important in two respects. Firstly, the assumption of linearity of the fMRI response and neural activity makes it possible to determine changes in neural activity by the amplitude changes in hemodynamic response. Secondly, this linear transform model also shows that when multiple stimuli are presented in succession, the hemodynamic response would be the summation of the individual responses generated by the single stimulus respectively.

Modeling the relationship between the fMRI response and stimuli is a key step towards detecting fMRI activity. Standard statistical analysis is carried out based on the following model:

$$Y(t) = \beta\sum_j h(t - \tau_j) + \epsilon(t), \tag{3.1}$$

where the HRF $h(\cdot)$ is pre-specified and β is a voxel specific parameter, to be utilized for testing fMRI activity [15, 21, 23, 24, 31, 33, 43]. The assumptions made about the shape of the HRF vary among different methods. Some of them are very stringent, while others are relatively more flexible. Typically, a "canonical" HRF is employed to process fMRI data. Some studies have reported variation in the shape of HRF across subjects [2, 4], and within the same subject across regions [11, 32, 38].

Detecting fMRI activity has also been evolved from using block-designs (where the stimulus times τ_i occur consecutively to form a block) to event-related fMRI

(ER-fMRI) designs [41]. In the latter case, stimuli (or events) are applied for short bursts in a stochastic manner. The recorded BOLD fMRI signals measure transient changes in brain activity associated with discrete events. This feature makes ER-fMRI a useful tool to estimate the change in the MR signal triggered by neuronal activity.

As an early study of ER-fMRI, Dale and Buckner [16] correlated the selective averaging (or time averaging) data and the fixed HRF induced data in a voxelwise manner. Friston et al. [24] employed a Poisson function with a fixed parameter for the HRF. In the general linear model (GLM) framework, Friston et al. [22] estimated the HRF by two given temporal basis functions. To enhance its flexibility, this idea was extended by Josephs et al. [31] to include more basis functions. These are very important contributions since the basis sets allow one to estimate an HRF of arbitrary shape for different events at different voxels of the brain, and at the same time the inferences can be easily made. Many studies on modeling HRF have since focused on the refinement and improvement of the basis sets idea. For example, Woolrich et al. [44] introduced a technique by applying some constraints to avoid nonsensical HRF, which is a big problem when using simple basis functions. More recently, Lindquist and Wager [34] proposed another method, using three superimposed inverse logistic functions, to model the HRF. This paper also described some most popular HRF modeling techniques, such as smooth finite impulse response (FIR) filter [29], canonical HRF with time and dispersion derivatives [14] and the canonical SPM HRF [25]. A flexible method based on splines has been considered by Zhang et al. [47].

From a Bayesian perspective, Genovese [26] and Gössl et al. [28] proposed to model the HRF by a number of parameters and prior distributions are given to each parameter. See also Woolrich et al. [44] and Lindquist and Wager [34]. Inferences of the parameters were then made at each voxel using Markov Chain Monte Carlo (MCMC) technique. The disadvantage of these methods is the slow performance of general MCMC techniques for the inferences.

The above methods are referred to as the time-domain methods. We now consider the frequency-domain approach. Assuming a periodic stimulus design, fMRI time series analysis can be greatly simplified in the frequency domain, which is more natural as the problem of modeling the relationship between the response and the stimuli is reduced to a few parameters related to the stimulus frequency information. One of the first frequency domain approaches is given by Lang and Zeger [33], who used model (3.1) along with a two-parameter gamma function to model the HRF. The two parameters vary at different voxels and hence the estimated HRF varies from voxel to voxel. It was reported that this approach has an identifiability problem of the parameters. The issue was addressed in Marchini and Ripley [37] using a fixed HRF approach.

We remark that a common theme among the time-domain methods for testing activity is the two-step procedure: the extra step is required for modeling parametrically the temporal noise series. This will affect the power of the test. While the above frequency approaches avoided the noise modeling part, they lack the ability to address: (1) the varying HRF issue, and (2) different types of stimulus designs. Moreover, the linear transformed model of fMRI response has not been tested and some studies [16, 30] reported the presence of non-linearity.

In Section 2, we described a regression approach based on model (1.1) for addressing these problems. The application of point processes to model the stimuli is novel in the HRF modeling literature. The procedure is also greatly simplified. Namely, it enables us to estimate the HRF directly and simultaneously test the lin-

earity assumption in a single step. The idea of using point processes can be traced back to the work of Brillinger [8, 9] who applied this to identification problems emerged in neurophysiological neural spike train analysis.

4. Simulated Numerical Results for HRF estimation

We illustrate the usefulness of our procedure in a simulated study. Here we use one of the HRF's from the literature [16, 27] to generate the response. The stimulus sequence contains a brief trial (1 second) intermixed with events "on" or "off" (Figure 2a). In this experiment, each trial lasts for 18 seconds and there are sixteen runs. So the "average" (because of the random "on" or "off") frequency of the event is $18/288 = 0.0625$. The estimated power spectrum (Figure 2b) and the frequency of the event are precisely estimated. The second peak corresponds to the frequency of every other event. In the first experiment (Figure 3a-b), the response is given by $Y(t) = a \int h(t-u)X(u)du + \epsilon(t)$ with $a = 0.5$ and the noise ϵ is generated from an AR(1) with coefficient 0.7: $\epsilon(t) = 0.7\epsilon(t-1) + z(t)$, $z(t) \sim N(0, .3^2)$. In the second experiment (Figure 3c-d), the noise is generated from an ARMA(2,2): $\epsilon(t) - 0.8897\epsilon(t-1) + 0.4858\epsilon(t-2) = z(t) - 0.2279z(t-1) + 0.2488z(t-2)$, $z(t) \sim N(0, .3^2)$. The AR(1) time series model was chosen to represent the default settings in Statistical Paramtric Mapping (SPM) [21] and FMRIB Software Library (FSL) [43], while the ARMA case was mainly for testing teh strengths of our method under other types of correlated structures. The coefficients were selected to illustrate the performance of the procedure under moderate serially correlated noise. Large coherency at the stimulus frequency of 0.0625 indicates that the activation is strong, and there is some linearity in the response and the stimulus series. This is also confirmed by the highly significant F-statistics (Figures 4a-d). The significant level is a dashed line that sits near the bottom of the graph. The variability of the proposed estimated is illustrated in Figures 5a-c using various noise level with SD=0.3, 0.5. We remark that the number of runs (=16) used in these simulations is based on recently published articles.

We further apply the procedure to examine the main application of fMRI to detect regions of activation. These are illustrated in Figures 6, 7. In these experiments, the responses are generated from

$$Y(t) = a \int h(t-u)X(u)\,du + \epsilon(t),$$

with varying a to show contrast of the regions so that the sub-region has a higher value of a. The noise component is ARMA(2,2), the same as in the previous experiment with SD=0.3. The regions of activation are clearly captured (Figure 6) when the contrast ratio is high. The effect of the contrast ratio on the detection of region of activation is depicted in Figures 7. It is evident that the level of detection depends on the contrast ratio.

5. A Real Data Analysis

5.1. Experiment Paradigm and Data Description

In this study, an fMRI data set was obtained from one human subject performing a predefined event sequence as visually instructed. The stimulus sequence includes two different events: right-hand and left-hand finger tapping. Each finger tapping

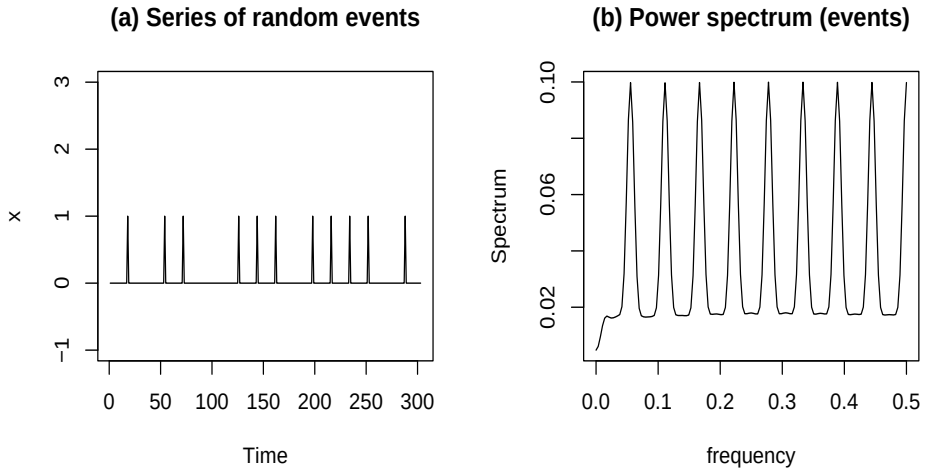


FIG 2. Plot of the stimulus series (16 trials) with on-off pattern stimuli (every 18 seconds). The whole duration lasts about $16 \times 18 = 288$ seconds. (b) The estimated power spectrum provides frequency of the occurred events. The frequency associated with the first peak is about $18/288 = 0.0625$. The second peak gives frequency of every other event, etc.

movement lasted around 1 second. The order of the sequence was predefined in a random way. To avoid the overlapping of consecutive events, the time interval between two successive events was randomly selected from Uniform$[18, 22]$. A typical sequence of stimuli is $\{R, L, R, R, L, L\}$.

During the experiment, 47 MR scans were acquired on a modified 3T Siemens MAGNETOM Vision system. Each acquisition consisted of 49 contiguous slices. Each slice contained 64×64 voxels. Hence there were $64 \times 64 \times 49$ voxels from each scan. The size of each voxel is 3mm $\times$ 3mm $\times$ 3mm. Each acquisition took 2.9388 seconds, with the scan to scan repetition time (TR) set to be 3 seconds.

5.2. Analysis and Results

The data set was preprocessed using SPM5 [21]. The preprocessing included realignment, slice timing correction, coregistration and spatial smoothing. We then analyzed the processed data set using both our proposed method and SPM5. When using SPM5, we used a canonical HRF with time and dispersion derivatives to model the hemodynamic response [25] and its functional form is shown in Figure 5. A t-statistic map was generated to show the activations triggered by the stimuli and part of them is shown on the first row of Figure 8.

When using the proposed method to detect which regions of the brain were activated by the finger tapping movements, we generated a spatial color map of the p-value for each voxel. The p-values were calculated based on the test defined by (2.9). Thus the activation regions are identified by the F statistics that are significant. The p-map generated this way is shown on the second row in Figure 8.

The four image slices represent the spatial maps of the right-hand activation. The red areas illustrate activated brain regions. Brighter color indicates higher intensity. Our p-maps demonstrate the classic brain activation patterns during hand movement as described above. However, the t-maps of the same four slices generated using SPM5 do not show any activation, as seen from the first row of Figure 8.

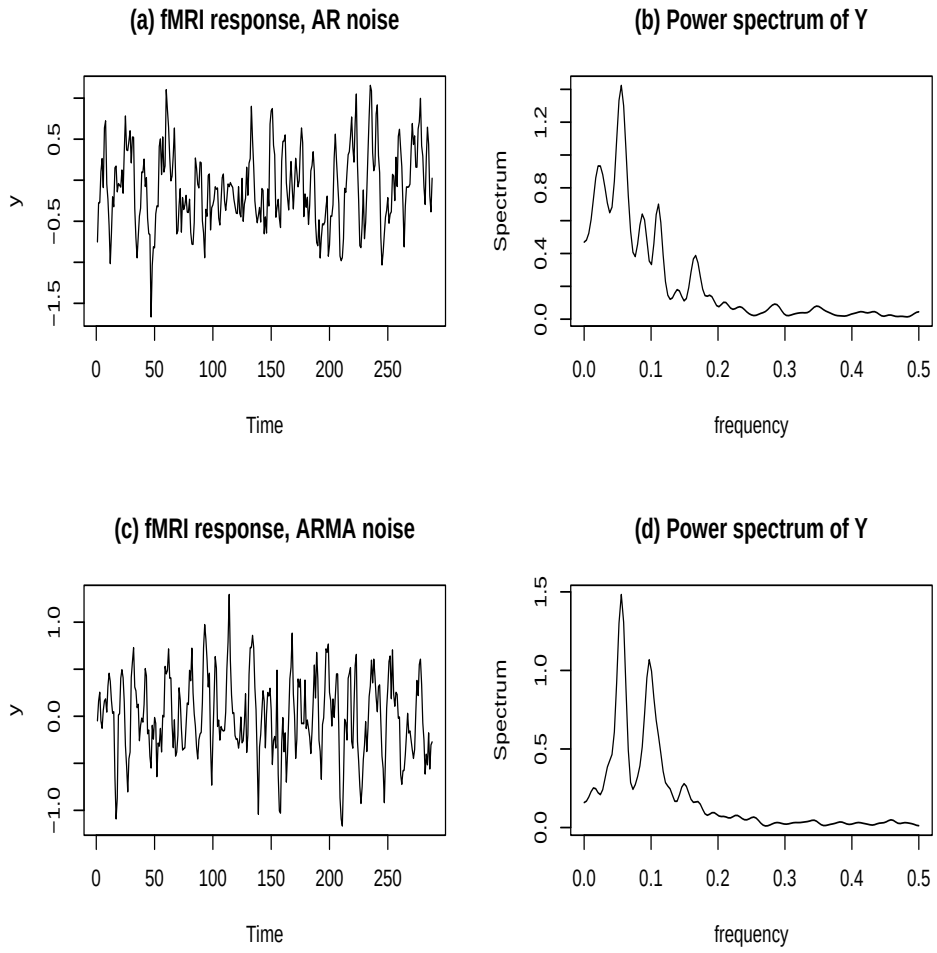


FIG 3. The fMRI responses. In these experiments, the responses are generated from $Y(t) = (0.5)\int h(t-u)X(u)\,du + \epsilon(t)$ with $\epsilon(t) = 0.7\epsilon(t-1) + z(t)$, $z \sim N(0, .3^2)$ in (a) and (b); $\epsilon(t) - 0.8897\epsilon(t-1) + 0.4858\epsilon(t-2) = z(t) - 0.2279z(t-1) + 0.2488z(t-2)$, $z \sim N(0, .3^2)$ in (c) and (d). The stimuli are the same as in Fig. 2.

Next we plot the estimated HRFs at voxels which are shown to be activated according to Figure 8. Figure 9 displays the estimated HRFs for voxels (with $F > 20$) selected from primary motor cortex (PMC). Figure 10 displays the estimated HRFs in cerebellum, and Figure 11 shows the estimated HRFs in the supplementary motor area (SMA). These figures were obtained by first computing the F-statistics (2.9) followed with the selection of those voxels with the F values greater than 20. This threshold was chosen to adjust for the multiple comparison effect and was carried out by computing the F statistics over those voxels that are known to be not activated by the stimuli. For example, we used the WFU PickAtlas software [36] to generate region of interest (ROI) mask in the cerebro spinal fluid area of the brain. Then computed the F statistics over this area, followed with a density estimate (eg, kernel method, or simply histogram) to select the thresholding value. There are about 20,000 voxels in the cerebro spinal fluid area which can be used to calibrate the null distribution for detecting fMRI activity.

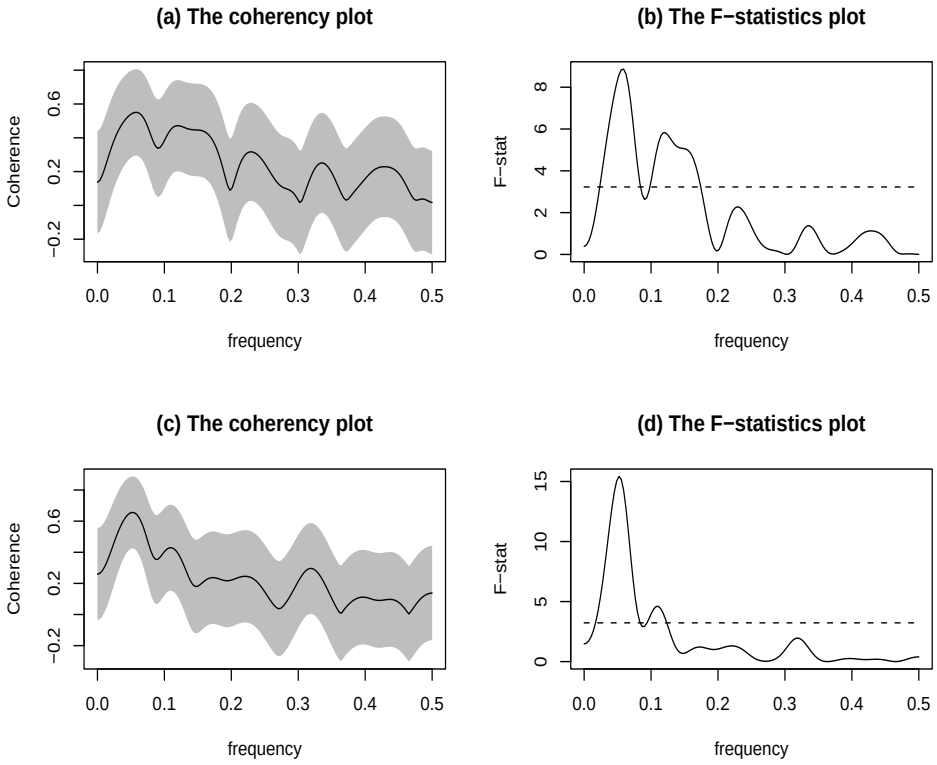


FIG 4. (a) and (c): The estimated coherency function with pointwise confidence intervals. Large coherence values at the event frequencies indicate perfect linear time invariant system used in this simulation. (b) and (d): The F-test for coherency with the dashed-line showing the significant level. (a) and (b) have the same conditions as (a) and (b) of Fig. 3; similarly for (c) and (d).

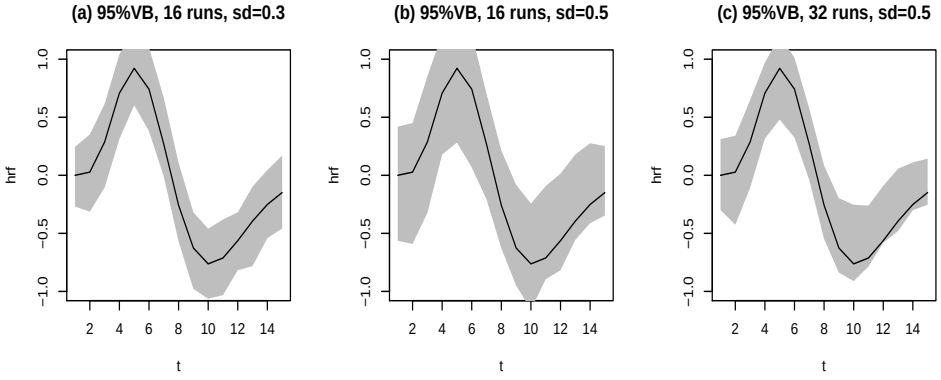


FIG 5. (a) 95% variation bands (VB) are obtained from estimates of the HRF using 100 random samples simulated from the model by trimming off the 2.5% both ends. (b) Same as in (a) except the noise SD has increased to 0.5. This clearly indicates the variance of the proposed estimate depends on the variance of the noise. (c) Same as in (b) by doubling the number of runs. This illustrates the variance of the estimate is inversely proportional to the duration of the experiment. In these experiments, the responses are generated from $Y(t) = \int h(t-u)X(u)\,du + \epsilon(t)$ with $\epsilon \sim \text{ARMA}(2,2)$. The parameters are the same as in Fig. 3.

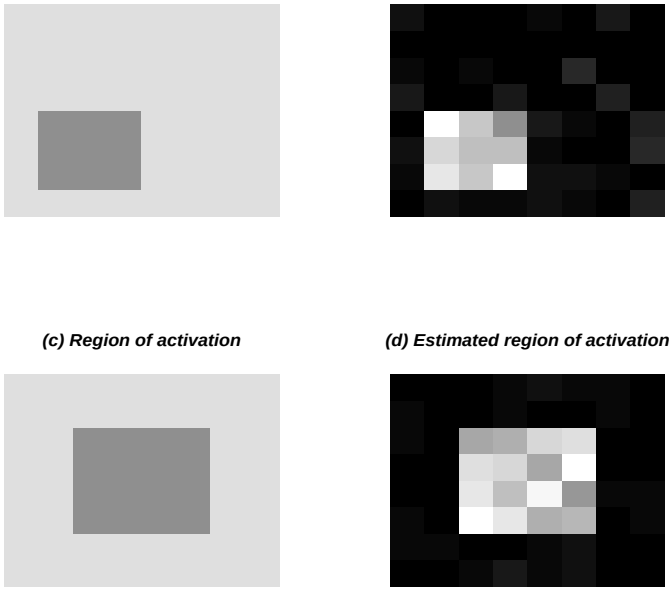


FIG 6. Estimated region of activation. In these experiments, the responses are generated from $Y(t) = a \int h(t-u)X(u)\,du + \epsilon(t)$ with $a = 0.1$ outside and $a = 1.0$ inside the sub-region, respectively. There are two slices with different region sizes and locations. Each slice is 8×8. (a) and (c) are the true regions, (b) and (d) are the estimated regions. The noise ϵ is generated from an ARMA(2,2). The stimuli are the same as in Fig. 2.

All of them have the form that agrees with empirical experience. It is well established that the contralateral cerebral hemisphere motor areas such as primary motor cortex (PMC), and ipsilateral cerebellar areas play dominant role in motor functions in normal human subjects [1, 3, 17, 19]. Our new methods validate unequivocally this known motor activation pattern with single finger movement in a single subject, whereas traditional SPM5 failed to do. Adequate imaging analysis techniques to demonstrate the involvement of those structures during motor function is superior important. PMC is the primary brain region directly control human movement, basal ganglia and cerebellar modulate its functions through a number of cortical motor associated areas of the brain (such as SMA). Dysfunctions of these structures have known to cause a variety of movement disorders such as Parkinson's disease and cerebellar ataxia [1]. Our methods might provide "higher resolution" statistical analysis methods for clinical and neuroscientist to define the roles of these structures in disease stages using fMRI.

5.3. Implications

(1) We demonstrated that our method handles statistical issues related to event-related experiments well. (2) It is nonparametric in the sense that the functional form of HRF is not specified a priori. Hence it is an useful diagnostic tool for other approaches that may be biased because of misspecification of HRF. (3) Variation of HRF in the brain has been under active research [18], and the nonparametric

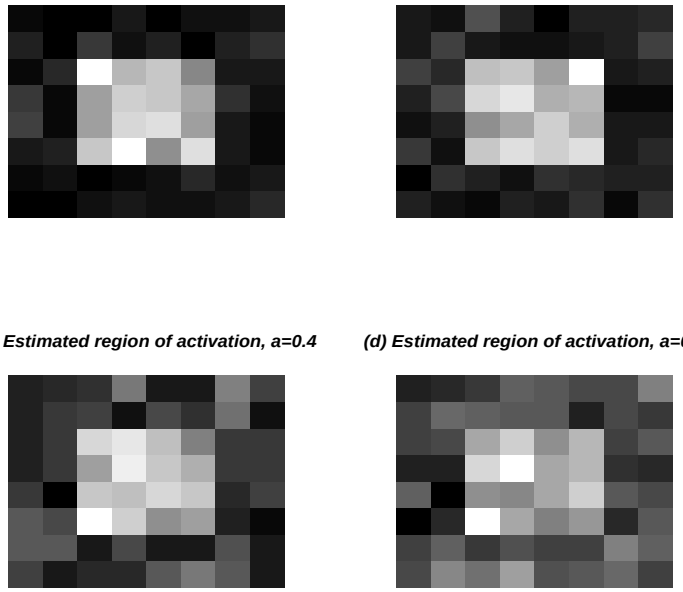


FIG 7. Estimated region of activation with varying contrast ratios. In these experiments, the responses are generated from $Y(t) = a\int h(t-u)X(u)\,du + \epsilon(t)$ with $a = 0.2, 0.3, 0.4, 0.5$ outside the sub-region and $a = 1.0$ inside the sub-region. The noise ϵ is generated from an ARMA(2,2). These results illustrate that accuracy of the estimates depends on the signal-to-noise (or contrast) ratio: The contrast ratio is proportional to $1/a$. (a) Here $a = 0.2$ implies that the signal is weaker than that in Fig. 6, but the contrast is still high and so the estimated region can still be clearly identified. (b) The contrast here is weaker with $a = 0.3$. (c) Weaker contrast with $a = 0.4$, and (d) fuzzy region due to the weakest contrast used in this experiment.

approach offers a systematic way to study the variation without requiring HRF to have the same shape over all voxels. (4) The linear relationship specified through the BOLD signal can be examined statistically by carrying out a formal test of hypothesis. This is important in verifying the linearity assumption employed in SPM [20, 23, 24, 45] in the process of constructing the T-map. (5) It is relatively easy to interpret the results using our approach as no prior specification of HRF is required (as is done in SPM [21]/FSL [43]/AFNI [15]).

5.4. Discussions

There are many ways to detect fMRI activity. The critical problem is to estimate the statistical significance, which depends on the estimation of both the magnitude of the response to the stimulus and the serial dependence of the time series and especially on the assumptions made in that estimation. Nonparametric spectral density estimation is shown to be self-calibrating and accurate when compared to several other time-domain approaches [12, 13], SPM: [20, 23, 24, 45, 46]. In particular, spectral technique to detect periodic and event-related activations has a distribution theory with significance levels down to 1 in 100,000, levels which are needed when a whole brain image is under consideration. The technique is especially resistant to high frequency artifacts that are found in some datasets and

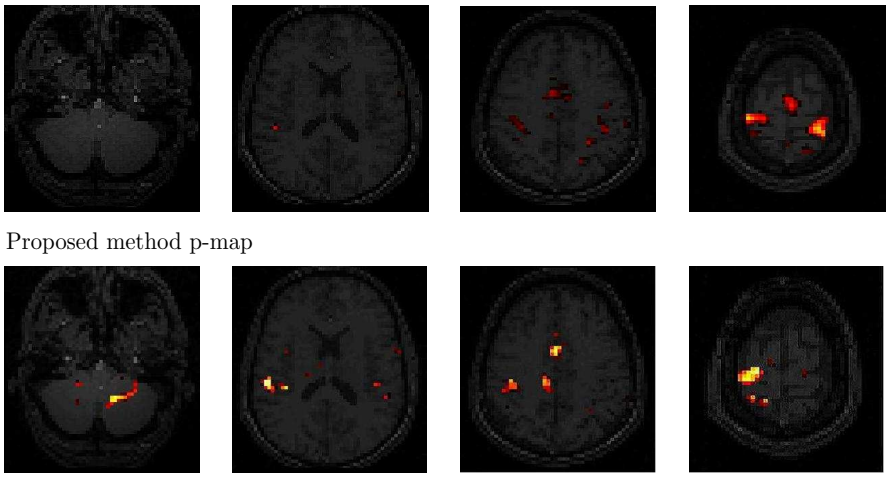


FIG 8. *The four related slices that contain the areas activated by right-hand finger tapping. The first row consists of the t-maps generated by SPM5 and they do not show any activation. The second row contains the p-maps generated by the proposed method. The first slice indicates the activated areas in cerebellum. The second slice contains basal ganglia. The third slice contains supplementary motor area (SMA) and the fourth slice shows primary motor cortex (PMC).*

it was demonstrated that time-domain approaches may be sufficiently susceptible to these effects to give misleading results. Also, these techniques are capable for detecting activations in clumps of a few (even one) voxel in periodic designs, yet produce essentially no false positive detections at any voxels in null datasets [37].

6. Concluding Remarks

It is now widely accepted that fMRI modeling requires flexible HRF modeling, with the HRF varying spatially and between subjects. Flexibility in linear modeling has been introduced with the use of basis functions [22]. However, basis functions suffer from a number of limitations. They impose a hard constraint on the allowed HRF shape and often the extent of the constraint is difficult to control and/or interpret. To overcome these problems, we formulated a procedure based on model (1.1) and FFT. The usefulness has been demonstrated empirically.

We remark that time-domain methods such as SPM [21], FSL [43], FIR [34, 35] and local adaptive spline estimate [47] in modeling the HRF are generally very sensitive to the linearity assumption and the error structures they employ. Any approach proposed within the time-domain may have difficulty providing resistant estimates. There is also no guarantee that the parametric noise model chosen will be sufficiently flexible to capture the true form of the correlation structure even if artifacts are removed and a model selection procedure is employed [18, 37]. Therefore significant loss in statistical efficiency can occur if these assumptions are invalidated. In contrast, if these assumptions are valid then the use of a frequency approach will result in a comparatively small loss in efficiency [10]. When considering voxel time series from fMRI datasets there can be no guarantees that the correct time domain

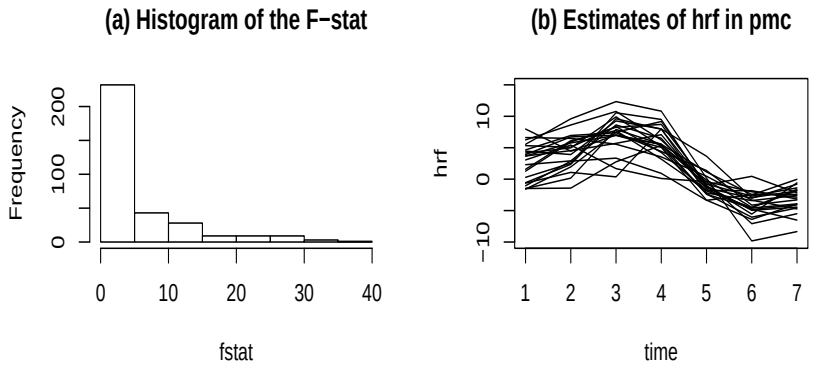


FIG 9. *The histogram for the F-stat and the HRF estimates in the primary motor cortex (PMC) area with $F > 20$. Each finger-tapping task lasted around 1 second. The order of the sequence was predefined in a random way. The time interval between two successive events was randomly selected from Uniform (18,22). Each acquisition took 2.9388 seconds, with the scan to scan repetition time (TR) set to 3 seconds.*

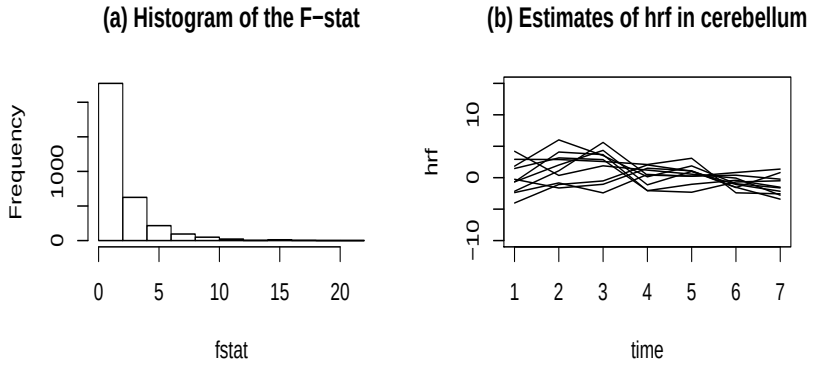


FIG 10. *The histogram for the F-stat and the HRF estimates in the cerebellum area with $F > 15$.*

approach has been chosen and a frequency approach seems the most prudent in this context. It is often demonstrated that the assumptions of commonly proposed time-domain models are not resistant to high frequency artifacts.

It is generally believed that the direct analysis of nonperiodic designs will not be as simple as that of the periodic designs, since the response due to the stimulus will be spread over a range of frequencies. Marchini and Ripley [37] suggested that this may be addressed by combining their method with the iteratively reweighted least squares [33] in the spectral domain and the basis functions [22]. However, this method will not be easily extended to model the HRF discussed in this paper. By formulating the problem using point processes, the frequency method advocated by [37] can be easily generalized to handle event-related designs. We also observe that our method is applicable to block designs since the stimuli can be put next to each other to form a block. Thus this unified approach significantly improves the estimation of the HRF described in [33, 37].

The flexible frequency approach proposed here acts as an insurance policy against the results being badly affected by artifacts, and is guaranteed to be near-optimal under all realistic operational conditions. It also offers a quick and accurate way to

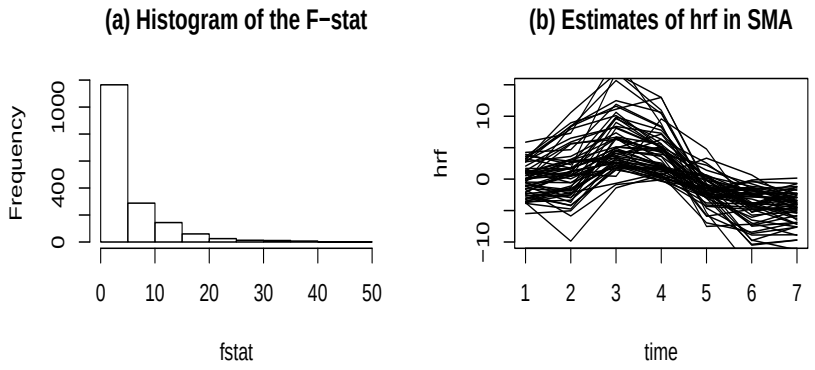


FIG 11. *The histogram for the F-stat and the HRF estimates in the cerebellum area with* $F > 15$.

check the calibration of the procedure. Further investigations will be carried out for an extensive comparative study on these maps. A concern about our procedure is the choice of weight function $W(\cdot)$ and bandwidth b given in (2.7). The former is less crucial and it can be addressed by choosing one of commonly used weight functions described in Newton [39]. Bandwidth selection appears to be more serious and it would seem to require adaptive methods such as cross-validation. Based on our experience and the fact that the HRF (blood flow) is relatively smooth, the choice of bandwidth therefore plays a less significant role. Nevertheless, we do observe that the spectral properties of the stimuli can be closely examined by the designs of the experimental protocols, which to some extent can help determine the smoothness of the estimate of HRF. This project is currently underway along with the use of splines for estimating the spectra.

7. Sampling Properties of the Estimates

The sampling properties of the HRF estimate will rely on the spectral properties of the stimulus $X(t)$, which is a point process. They also depend on the spectral properties of the response $Y(t)$ and the noise series $\epsilon(t)$, which are real-valued ordinary stationary time series. Thus this section starts with a brief summary of the spectral properties of stationary time series which will be denoted by $X(t)$. This is followed by a discussion of the cumulants, which are essential for establishing asymptotic distributions of the estimates. Subsequent sections describe sampling properties of various statistics involved in establishing the properties of HRF estimate.

7.1. Point Process

Consider a point process $X(t)$ with points occurring at times $0 \le \tau_1 \le \tau_2 \le \cdots$ with $X(t)$ denoting the number of points in the interval $(0, t]$. When it exists, the rate of the process at time t is given by

$$p_X(t) = \lim_{v \downarrow 0} \frac{1}{v} \mathsf{E}\big(X(t+v) - X(t)\big).$$

The expected number of points in the small interval $(t, t+v]$ is given by $p_X(t)v + o(v)$. Suppose orderliness, that is, the points of its realizations are isolated, multiple

points do not occur. Then $dX(t) = 0$ or 1 and one must have

$$\mathsf{P}\{dX(t) = 1\} = p_X(t)\,dt.$$

The rate function $p_X(t)$ is seen to have an interpretation as a probability.

In the second-order case one defines the second-order product density as

$$p_{XX}(t_1, t_2) = \lim_{v_1, v_2 \downarrow 0} \frac{1}{v_1 v_2} \mathsf{E}\big(X(t_1 + v_1) - X(t_1)\big)\big(X(t_2 + v_2) - X(t_2)\big), \quad t_1 \neq t_2.$$

In view of the orderliness of the process, $\mathsf{P}\{dX(t) = 1 \text{ and } dX(t) = 1\} = \mathsf{P}\{dX(t) = 1\}$, the case $t_1 = t_2$ can be included via

$$\begin{aligned}\mathsf{P}\{dX(t_1) = 1 \text{ and } dX(t_2) = 1\} &= \mathsf{E}\big(dX(t_1)dX(t_2)\big) \\ &= \big(p_{XX}(t_1, t_2) + \delta(t_1 - t_2)p_X(t_2)\big)\,dt_1\,dt_2,\end{aligned}$$

where $\delta(\cdot)$ is the Dirac delta function: $\delta(\cdot) \geq 0$ and $\int \delta(t)\varphi(t)\,dt = \varphi(0)$ for infinitely differentiable real-valued function φ with compact support. It is useful to recall here that δ is the (generalized) derivative of the Heaviside function: $H(\cdot) = 1_{(0,\infty)}(\cdot)$. See [42].

The covariance density of the process is defined by

$$q_{XX}(t_1, t_2) = p_{XX}(t_1, t_2) - p_X(t_1)p_X(t_2),$$

with the interpretation

$$\operatorname{cov}\{dX(t_1), dX(t_2)\} = \big(q_{XX}(t_1, t_2) + \delta(t_2 - t_1)p_X(t_2)\big)\,dt_1\,dt_2.$$

The conditional intensity of the process is defined by $p_{XX}(t)/p_X(t)$ with the interpretation

$$\mathsf{P}\{dX(t_2) = 1 \mid dX(t_1) = 1\} = \big(p_{XX}(t_1, t_2)/p_X(t_1)\big)\,dt_2.$$

A point process is said to be stationary when its probability properties are unaffected by simple shifts of time. In this case one has

$$\mathsf{P}\{dX(t) = 1\} = p_X\,dt,$$

$$\mathsf{P}\{dX(t_1) = 1 \text{ and } dX(t_2) = 1\} = \big(p_{XX}(t_2 - t_1) + \delta(t_2 - t_1)p_X\big)\,dt_1\,dt_2$$

$$\operatorname{cov}\{dX(t_1), dX(t_2)\} = \big(q_{XX}(t_2 - t_1) + \delta(t_2 - t_1)p_X\big)\,dt_1\,dt_2.$$

By analogy with what is done in the ordinary time series case one may define the power spectrum at frequency f by

$$s_{XX}(f) = \frac{1}{2\pi} \int e^{-ifu}\big(\operatorname{cov}\{dX(t + u), dX(t)\}/dt\big)\,du.$$

For multivariate process $X(t) = \{X_1(t), \ldots, X_m(t)\}$, it may be convenient to consider

$$\mathsf{P}\{dX_j(t) = 1\} = C_j\,dt, \qquad j = 1, \ldots, m,$$

and

$$\operatorname{cov}\{dX_j(t + u), dX_k(t)\} = C_{jk}(du)\,dt, \qquad j, k = 1, \ldots, m.$$

The power spectrum at frequency f is defined by

$$s_{jk}(f) = \frac{1}{2\pi} \int e^{-ifu} C_{jk}(du), \qquad j, k = 1, \ldots, m.$$

7.2. Stationary Time Series

Let $X(t) = (X_1(t), \ldots, X_r(t))$, $t = 0, 1, 2, \ldots$, denote a vector-valued stationary time series. Set

$$C_{jk}(u) = \operatorname{cov}\{X_j(t+u), X_k(t)\}, \qquad j, k = 1, \ldots, r.$$

The power spectrum at frequency f is defined by

$$s_{jk}(f) = \frac{1}{2\pi} \sum_u e^{-ifu} C_{jk}(u), \qquad f \in \mathbb{R}, \quad j, k = 1, \ldots, r.$$

7.3. Cumulants and Spectra

Definition 1. *Let $X_1, X_2, \ldots, X_r$ denote random variables with finite rth moment. The rth order joint cumulant of $X_1, X_2, \ldots, X_r$ is defined by*

$$\operatorname{cum}(X_1, X_2, \ldots, X_r) = \sum (-1)^{p-1} (p-1)! \left(\prod_{j \in \nu_1} X_j \right) \ldots \left(\prod_{j \in \nu_p} X_j \right),$$

where the summation extends over all partitions $\nu_1, \ldots, \nu_p$, $p = 1, \ldots, r$ of $\{1, 2, \ldots, r\}$.

Remarks

1. When $X_1 = X_2 = \cdots = X_r$, the definition gives the cumulant of order r of a univariate random variable.
2. $\operatorname{cum}(X_1, X_2, \ldots, X_r)$ is also given by the coefficient of $(i)^r t_1 \ldots t_r$ in the Taylor series expansion of $\log(E \exp i \sum_1^r X_j t_j)$.

Given r time series $X_1(t), X_2(t), \ldots, X_r(t)$ with each having finite rth moment, we define

$$C_{1,\ldots,r}(t_1, t_2, \ldots, t_r) = \operatorname{cum}(X_1(t_1), X_2(t_2), \ldots, X_r(t_r)).$$

For stationary time series,

$$C_{1,\ldots,r}(t_1, t_2, \ldots, t_r) = C_{1,\ldots,r}(t_1 - t_r, t_2 - t_r, \ldots, t_{r-1} - t_r, 0),$$

which is a function of $r-1$ variables. In this case, the rth order cumulant spectrum, $s_{1,\ldots,r}(f_1, f_2, \ldots, f_{r-1})$, is defined by

$$s_{1,\ldots,r}(f_1, f_2, \ldots, f_{r-1}) =$$

$$(2\pi)^{-k+1} \sum_{u_1, u_2, \ldots, u_{r-1}} C_{1,\ldots,r}(u_1, u_2, \ldots, u_{r-1}) \exp\left(-i \sum_{j=1}^{r-1} u_j f_j \right),$$

$$f_1, f_2, \ldots, f_{r-1} \in \mathbb{R}, \quad r \geq 2.$$

For a more detailed discussion of cumulants and their spectra, see [10].

7.4. Fast Fourier Transforms

Let $a_j(\cdot) : \mathbb{R} \to \mathbb{R}$, $j = 1, 2$, denote **tapering functions**. The discrete Fourier transform for the univariate series X_j is defined by

$$\varphi_j^T(f) \equiv \varphi_{X_j}^T(f) = \sum_t a_j(t/T) X_j(t) \exp(-ift), \qquad f \in \mathbb{R}, \quad j = 1, 2.$$

For vector-valued series $\mathbf{X}$, it is given by

$$\boldsymbol{\varphi}^T(f) \equiv \boldsymbol{\varphi}_{\mathbf{X}}^T(f) = \begin{pmatrix} \varphi_1^T(f) \\ \varphi_2^T(f) \end{pmatrix}, \qquad d \in \mathbb{R}.$$

Set $a_j^T(t) = a_j(t/T)$, $j = 1, 2$. For $j_m \in \{1, 2\}$, $m = 1, \ldots, M$,

$$A_{j_1,\ldots,j_M}^T(f) = \sum_t \left(\prod_{m=1}^M a_{j_m}^T(t) \right) \exp(-ift), \quad f \in \mathbb{R}.$$

Condition 1. *The tapering function $a(\cdot) : \mathbb{R} \to \mathbb{R}$ has a compact support with bounded first derivative. Furthermore,*

$$\int a(u)\, du = 1 \qquad \textit{and} \qquad \int |a(u)|\, du < \infty.$$

Condition 2. *The covariance function satisfies*

$$\sum_u C_{jk}(u) < \infty,$$

and

$$\sum_{u_1,\ldots,u_{M-1}} C_{j_1 \ldots j_M}(u_1, \ldots, u_{M-1}) < \infty, \qquad j_1, \ldots, j_M = 1, 2.$$

The second part of the above condition is necessary for establishing the asymptotic properties of the estimates to be considered in this section.

Lemma 1. *Suppose Conditions 1 and 2 hold. Then*

$$\sup_{f_1,\ldots,f_M} \Big| \mathrm{cum}(\varphi_{j_1}^T(f_1), \ldots, \varphi_{j_M}^T(f_M)) - (2\pi)^{M-1} A_{j_1,\ldots,j_M}^T(f_1 + \cdots + f_M)\, s_{j_1,\ldots,j_M}(f_1, \ldots, f_M) \Big| = o(T).$$

Condition 3. *The covariance function satisfies*

$$\sum_u |u| c_{jk}(u) < \infty,$$

and

$$\sum_{u_1,\ldots,u_{M-1}} |u_1 \cdots u_{M-1}| C_{j_1 \ldots j_M}(u_1, \ldots, u_{M-1}) < \infty, \qquad j_1, \ldots, j_M = 1, 2.$$

Lemma 2. *Under Conditions 1 and 3,*

$$\sup_{f_1,\ldots,f_M} \Big| \mathrm{cum}(\varphi_{j_1}^T(f_1), \ldots, \varphi_{j_M}^T(f_M)) - (2\pi)^{M-1} A_{j_1,\ldots,j_M}^T(f_1 + \cdots + f_M)\, s_{j_1,\ldots,j_M}(f_1, \ldots, f_M) \Big| = O(1).$$

Proof. We now prove Lemmas 1 and 2. If follows from

$$|a_j(t+u)a_k(t+v) - a_j(t)a_k(t)| \le |a_j(t+u)a_k(t+v) - a_j(t+u)a_k(t)| + |a_j(t+u)a_k(t) - a_j(t)a_k(t)|$$

and Condition 1 that there is a constant K_1 such that

$$|\sum_t a^T_{j_1}(t+u_1)\cdots a^T_{j_{M-1}}(t+u_{M-1})a^T_{j_M}(t)\exp(-ift) - A^T_{j_1\ldots j_M}(f)| \le K_1(|u_1|+\cdots+|u_{M-1}|).$$

By the cumulant property,

$$\begin{aligned}
&\operatorname{cum}(\varphi^T_{j_1}(f_1),\ldots,\varphi^T_{j_M}(f_M))\\
&= \sum_{t_1}\cdots\sum_{t_M} a^T_{j_1}(t_1)\cdots a^T_{j_M}(t_M)\exp\left(-i\sum_{m=1}^{M} f_m t_m\right)\\
&\qquad\times C_{j_1,\ldots,j_M}(t_1-t_M,\ldots,t_{M-1}-t_M)\\
&= \sum_{u_1=-2(T-1)}^{2(T-1)}\cdots\sum_{u_{M-1}=-2(T-1)}^{2(T-1)}\exp\left(-i\sum_{m=1}^{M-1} f_m t_m\right)C_{j_1,\ldots,j_M}(u_1,\ldots,u_{M-1})\\
&\qquad\times\sum_t a^T_{j_1}(t+u_1)\cdots a^T_{j_{M-1}}(t+u_{M-1})a^T_{j_M}(t)\exp\left(-i\sum_{m=1}^{M} f_m t\right)\\
&= \epsilon_T + \sum_{u_1=-2(T-1)}^{2(T-1)}\cdots\sum_{u_{M-1}=-2(T-1)}^{2(T-1)}\exp\left(-i\sum_{m=1}^{M-1} f_m t_m\right)\\
&\qquad C_{j_1,\ldots,j_M}(u_1,\ldots,u_{M-1})A^T_{j_1\ldots j_M}(f_1+\cdots+f_M),
\end{aligned}$$

where

$$|\epsilon_T| \le K_2\sum_{u_1=-2(T-1)}^{2(T-1)}\cdots\sum_{u_{M-1}=-2(T-1)}^{2(T-1)}(|u_1|+\cdots+|u_{M-1}|)C_{j_1,\ldots,j_M}(u_1,\ldots,u_{M-1}).$$

It now follows from Condition 3,

$$T^{-1}|\epsilon_T| \le K_2\sum_{u_1=-2(T-1)}^{2(T-1)}\cdots\sum_{u_{M-1}=-2(T-1)}^{2(T-1)}T^{-1}(|u_1|+\cdots+|u_{M-1}|)C_{j_1,\ldots,j_M}(u_1,\ldots,u_{M-1}),$$

$T^{-1}(|u_1|+\cdots+|u_{M-1}|)\to 0$ and the dominated convergence theorem that

$$|\epsilon_T| = o(T). \tag{7.1}$$

Lemmas 1 and 2 follow from this and

$$\begin{aligned}
&s_{j_1,\ldots,j_M}(f_1,\ldots,f_{M-1})\\
&\quad= (2\pi)^{M-1}\sum\cdots\sum\exp\left(-i\sum_{1}^{M-1} f_m u_m\right)C_{j_1,\ldots,j_M}(u_1,\ldots,u_{M-1}) + o(1).
\end{aligned}$$

7.5. Complex Normal

Let X denote an k-dimensional random vector whose components are complex-valued random variables. If, for some $\mu \in \mathbb{C}^k$ and $k \times k$ Hermitian non-negative definite matrix Σ (that is, $\Sigma = \bar{\Sigma}^\top$),

$$\begin{pmatrix} \operatorname{Re} X \\ \operatorname{Im} X \end{pmatrix} \sim N_{2k}\left(\begin{pmatrix} \operatorname{Re}\mu \\ \operatorname{Im}\mu \end{pmatrix}, \tfrac{1}{2} \begin{pmatrix} \operatorname{Re}\Sigma & -\operatorname{Im}\Sigma \\ \operatorname{Im}\Sigma & \operatorname{Re}\Sigma \end{pmatrix} \right),$$

we say X has a complex normal distribution with mean μ and covariance matrix Σ, and is abbreviated by $X \sim N_k^c(\mu, \Sigma)$.

The FFT is asymptotically normal with mean specified according to the frequency f as described below.

Theorem 7.1. *Under Conditions 1 and 2, $\varphi_j^T(f)$ is asymptotically*

1. $N_1^c(0, 2\pi T s_{jj}(f) A_{jj}(0))$ *if* $f \neq 0 \mod \pi$,
2. $N_1(T c_j A_{jj}(0), 2\pi T s_{jj}(0) A_{jj}(0))$ *if* $f = 0, \pm 2\pi, \ldots$,
3. $N_1(0, 2\pi T s_{jj}(\pi) A_{jj}(0))$ *if* $f = \pm\pi, \ldots$.

Note that $A_{jj}(0) = \int a_j^2$. The above result implies that the real and the imaginary part of $\varphi_j^T(f)$ are approximately independent. Each is approximately normal with mean and variance $\pi T s_{jj}(f) \int a_j^2$.

Proof. To prove Theorem 7.1, we note that by Condition 1, $A^T_{j_1,\ldots,j_M}(f) = O(T)$. Recall that the Gaussian distribution has cumulants of order greater than 2 vanishes. The desired result now follows from Lemmas 1, 2 and that fact that

$$\begin{aligned} & T^{-M/2} \operatorname{cum}(\varphi^T_{j_1}(f_1), \ldots, \varphi^T_{j_M}(f_M)) \\ & \quad = T^{-M/2}(2\pi)^{M-1} A^T_{j_1,\ldots,j_M}(f_1 + \cdots + f_M)\, s_{j_1,\ldots,j_M}(f_1, \ldots, f_M) + o(T^{1-M/2}) \\ & \quad \to 0 \qquad \text{for } M > 2 \text{ as } T \to \infty. \end{aligned}$$

7.6. Asymptotics for Periodograms

The distributions of the FFT suggests the following statistic:

$$\tilde{s}_{jj}(f) = |\varphi_j^T(f)|^2 \Big/ \left(2\pi \sum_t [a_j(t/T)]^2 \right), \qquad f \in \mathbb{R}.$$

This is called **periodogram** and is an estimate of the spectral density function s_{jj}. For more historical remarks, see [10]. Note that if there is no tapering function, the periodogram is given by

$$\tilde{s}_{jj}(f) = (2\pi T)^{-1} |\varphi_j^T(f)|^2, \qquad f \in \mathbb{R}.$$

Let $f_m = 2\pi m/T$, $m = 0, \pm 1, \pm 2, \ldots, \pm T/2$ denote the Fourier frequencies. The result below describes the asymptotic distribution of the periodograms.

Theorem 7.2. *Under Conditions 1–3, $\tilde{s}_{jj}(f_m)$, $m = 1, \ldots, M = T/2$, are asymptotically independent $s_{jj}(f_m)\chi_2^2/2$. Also $\tilde{s}_{jj}(f)$ is asymptotically $s_{jj}(f)\chi_1^2$ for $f = \pm\pi, \pm 3\pi, \ldots$, independent of the $\tilde{s}_{jj}(f_m)$, $m = 1, \ldots, T/2$.*

Proof. The proof follows from Theorem 7.1 and the definition of the chi-square distribution.

The above result shows that the asymptotic variance of the periodogram is approximately $s_{jj}(f)^2$, which is usually positive. Thus the periodogram is not a consistent estimate of the spectral density function. The following section will present a class of consistent estimates obtained by smoothing the periodograms.

7.7. *Window Estimates — The Smoothed Periodograms*

A class of consistent estimates can be obtained by using a running mean or local average of the periodograms. Specifically, set

$$\hat{s}(f_m) = (2K+1)^{-1} \sum_{k=-K}^{K} \tilde{s}_{jj}\left(\frac{2\pi}{T}(m+k)\right).$$

It follows from the asymptotic distributional properties of the periodograms (Theorem 7.2) that $\hat{s}(f_m)$, $m = 1, \ldots, T/2$, are asymptotically independent with $\hat{s}(f) \sim s(f)\chi^2_{4K+2}/(4K+2)$ if $f \neq 0$, and $\hat{s}(0) \sim s(0)\chi^2_{2K}/(2K)$. An important implication of the above result is that consistency can be achieved by letting $K \to \infty$ and $K/T \to 0$ as $T \to \infty$.

More generally, let $W(\cdot)$ denote a weight function. Set

$$\hat{s}_{jj'}(f) = \sum_{k \neq 0} b_T^{-1} W\left(b_T^{-1}\left(f - \frac{2\pi k}{T}\right)\right) \tilde{s}_{jj'}\left(\frac{2\pi k}{T}\right), \tag{7.2}$$

where

$$\tilde{s}_{jj'}(f) = (2\pi T)^{-1} \varphi_j^T(f) \overline{\varphi_{j'}^T(f)}, \qquad f \in \mathbb{R},$$

and b_T is referred to as the **bandwidth** or **window width** that will be specified more clearly later. Certain properties of the weight function $W(\cdot)$ will be required in order to assure that the above estimate is consistent.

Condition 4. *The weight function $W(\cdot) : \mathbb{R} \to \mathbb{R}$ is a symmetric probability density function with a compact support $[-\pi, \pi]$.*

Under this condition, the bias of the window estimate is given by

$$E(\hat{s}_{jj}(f)) = \int W(\lambda) s_{jj}(f - b_T\lambda)\, d\lambda + O(T^{-1}b_T^{-1}).$$

In fact, more properties can be obtained and are stated in the following result.

Theorem 7.3. *Under Conditions 1–3 and suppose that the spectral density function s_{jj} does not vanish. Let $b_T \to 0$ and $b_T T \to \infty$ as $T \to \infty$. Then, $\hat{s}(f_m)$, $m = 1, \ldots, M$, are asymptotically normal with mean zero and covariance structure given by*

$$\lim_{T\to\infty} b_T T \operatorname{cov}(\hat{s}(f_1), \hat{s}(f_2)) = \begin{cases} 0 & \text{if } f_1 \neq f_2, \\ 2\pi s(f)^2 \int W^2 & \text{otherwise.} \end{cases} \tag{7.3}$$

Proof. Direct computation shows that

$$\begin{aligned}&\operatorname{cov}(\tilde{s}_{jj}(f_1), \tilde{s}_{jj}(f_2))\\ &\quad = s_{jj}(f_1)\left\{\left(\frac{\sin T(f_1+f_2)/2}{T\sin(f_1+f_2)/2}\right)^2 + \left(\frac{\sin T(f_1-f_2)/2}{T\sin(f_1-f_2)/2}\right)^2\right\} + O(1/T).\end{aligned}$$

Moreover,

$$\begin{aligned}\operatorname{cov}(\hat{s}_{jj}(f_1), \hat{s}_{jj}(f_2)) &= 2\pi T^{-1}\int W^T(f_1-\lambda)W^T(f_2-\lambda)s_{jj}(\lambda)^2\,d\lambda\\ &\quad + 2\pi T^{-1}\int W^T(f_1-\lambda)W^T(f_2+\lambda)s_{jj}(\lambda)^2\,d\lambda\\ &\quad + O(b_T^{-2}T^{-2}) + O(T^{-1}),\end{aligned}$$

where

$$W^T(\lambda) = b_T^{-1}\sum_{k=-\infty}^{\infty} W(b_T^{-1}(\lambda+2\pi k)).$$

The indicated covariance structure (7.3) is an easy consequence of these results.

To obtain the asymptotic normality, we need to show that all cumulants of order higher than 2 tend to zero as $T\to\infty$. This is carried out by directly computing the cumulants of the window estimates in a manner similar to the proof of Lemma 1.

7.8. Estimating the Transfer Function

Theorem 7.4. *Suppose that*

1. $\epsilon(t)$, $t = 0, 1, \ldots$ *satisfy Condition 2 and have mean zero,*
2. $X(t)$ *is uniformly bounded and* $s_{11} \neq 0$,
3. $\sum_u |u|h(u) < \infty$,
4. W *in Condition 4 is a uniform kernel.*

Let $b_T \to 0$, $b_T T \to \infty$, $b_T^5 T \to 0$ *as* $T\to\infty$. *Then* $\hat{H}(f_1), \ldots, \hat{H}(f_M))$ *is complex normal with mean* $(E\hat{H}(f_1), \ldots, E\hat{H}(f_M))$ *and covariance matrix whose entries are given by*

$$\operatorname{cov}(\hat{H}(f_1), \hat{H}(f_2)) = \eta(f_1-f_2)\frac{2\pi s_{\epsilon\epsilon}(f_1)}{b_T T s_{11}(f_1)}\int W^2,$$

where $\eta(0) = 1$ *and* $\eta(f) = 0$ *for* $f \neq 0$.

The weight function W is assumed to be uniform on $[-\pi, \pi]$ in order to simplify the presentation of the above asymptotic properties of the estimate of the transfer function. A more general approach can be found in [10].

Proof. We begin the proof of Theorem 7.4 with two lemmas.

Lemma 3. *Let* $(\mathbf{V}_n)$ *denote a sequence of random vectors converging in distribution to* $\mathbf{V}$. *Then there exists a probability space such that* $\mathbf{V}_n$ *converges to* $\mathbf{V}$ *almost surely.*

Proof. The proof can be found in [5].

Lemma 4. *Let* $(\mathbf{V}_n)$ *denote a sequence of random vectors in* $\mathbb{R}^p$ *converging in distribution to* $N_p^c(\mathbf{0}, \mathbf{I}_p)$ *and* $(\mathbf{U}_n)$ *a sequence of* $p\times p$ *unitary matrices. Then* $\mathbf{U}_n\mathbf{V}_n$ *converges to* $N_p^c(\mathbf{0}, \mathbf{I}_p)$ *as* $n\to\infty$.

Proof. This follows from Lemma 3.

Before proceeding to the proof, we remark that the following argument is simplified by assuming the series X to be non-random. The result nevertheless holds for general random X. Let φ_j^T, $j = 0, 1$, be the Fourier transform of Y and X, respectively. Let $2\pi k/T$ denote the Fourier frequency that is nearest to λ. Then

$$\begin{aligned}&\varphi_0^T(2\pi(k+l)/T)\\ &\quad = H(2\pi(k+l)/T)\varphi_1^T(2\pi(k+l)/T) + \varphi_\epsilon^T(2\pi(k+l)/T) + O(1)\\ &\quad = H(\lambda)\varphi_1^T(2\pi(k+l)/T) + \varphi_\epsilon^T(2\pi(k+l)/T) + O(1), \quad l = 0, \pm1, \ldots, \pm m,\end{aligned}$$

where $O(1)$ is uniformly in l. Now let $\mathbf{D}_0$ denote the $1 \times (2m+1)$ matrix given by

$$\mathbf{D}_0 = \frac{1}{\sqrt{2\pi T}}\left[\ \varphi_0^T(2\pi(k-m)/T) \quad \cdots \quad \varphi_0^T(2\pi k/T) \quad \cdots \quad \varphi_0^T(2\pi(k+m)/T)\ \right].$$

Define $\mathbf{D}_1$ and $\mathbf{D}_\epsilon$ similarly. Then

$$\mathbf{D}_0 = H(f)\mathbf{D}_1 + \mathbf{D}_\epsilon + O(T^{-1/2}).$$

Let $\mathbf{U} \equiv \mathbf{U}^T = [\mathbf{U}_1, \mathbf{U}_0]$ be a $(2m+1)\times(2m+1)$ unitary matrix whose first column is $\mathbf{U}_1 = \mathbf{D}_1^H(\mathbf{D}_1\mathbf{D}_1^H)^{-1/2}$, where $\mathbf{D}^H = \overline{\mathbf{D}}^\top$ is the conjugate transpose of $\mathbf{D}$. Then

$$\mathbf{D}_0\mathbf{U} = H(f)\mathbf{D}_1\mathbf{U} + \mathbf{D}_\epsilon\mathbf{U} + O(T^{-1/2}).$$

The first and the remaining columns of these matrices yield

$$[\hat{H}(f) - H(f)]\hat{s}_{11}(f)^{1/2}(2m+1)^{1/2} = \mathbf{D}_\epsilon\mathbf{U}_1 + O(T^{-1/2}), \tag{7.4}$$

$$\mathbf{D}_0\mathbf{U}_0 = \mathbf{D}_\epsilon\mathbf{U}_0 + O(T^{-1/2}). \tag{7.5}$$

By the property of the unitary matrix,

$$\begin{aligned}(2m+1)\hat{s}_{00} = \mathbf{D}_0\mathbf{D}_0^H &= \mathbf{D}_0\mathbf{U}_1\mathbf{U}_1^H\mathbf{D}_0^H + \mathbf{D}_0\mathbf{U}_0\mathbf{U}_0^H\mathbf{D}_0^H\\ &= \mathbf{D}_0\mathbf{D}_1^H(\mathbf{D}_1\mathbf{D}_1^H)^{-1}\mathbf{D}_1\mathbf{D}_0^H + \mathbf{D}_0\mathbf{U}_0\mathbf{U}_0^H\mathbf{D}_0^H.\end{aligned}$$

Thus

$$\hat{s}_{\epsilon\epsilon} = \mathbf{D}_0\mathbf{U}_0\mathbf{U}_0^H\mathbf{D}_0^H = \mathbf{D}_\epsilon\mathbf{U}_0\mathbf{U}_0^H\mathbf{D}_\epsilon^H + O_p(T^{-1/2}). \tag{7.6}$$

Now, according to Theorem 7.1, $\mathbf{D}_\epsilon \to_d N^c_{2m+1}N(\mathbf{0}, s_{\epsilon,\epsilon}(f)\mathbf{I})$ and therefore $s_\epsilon(f)^{-1/2}\mathbf{D}_\epsilon \to_d N^c_{2m+1}N(\mathbf{0}, \mathbf{I})$. By Lemma 4, $s_\epsilon(f)^{-1/2}\mathbf{D}_\epsilon\mathbf{U} \to_d N^c_{2m+1}N(\mathbf{0}, \mathbf{I})$, or $\mathbf{D}_\epsilon\mathbf{U} \to_d N^c_{2m+1}N(\mathbf{0}, s_\epsilon(f)\mathbf{I})$. This, together with (7.4) and (7.6) yield the desired result. This completes the proof of the theorem.

7.9. Estimating the Hemodynamic Response Function

From

$$H(f) = \sum_u h(u)\exp(-iuf),$$

we see that the hemodynamic response function is given by

$$h(u) = \frac{1}{2\pi}\int_0^{2\pi} H(f)\exp(iuf)\,df, \qquad u = 0, \pm1, \pm2, \ldots.$$

Let $\hat{H}(f)$ denote an estimate of $H(f)$ given by the last section, and let $Q \equiv Q_T$ denote a sequence of positive integers tending to ∞ with T. As an estimate of $h(u)$ by approximating the integral using finite sums, we define

$$\hat{h}(u) = \frac{1}{Q}\sum_{q=0}^{Q} \hat{H}\left(\frac{2\pi q}{Q}\right) \exp\left(i\frac{2\pi uq}{Q}\right), \qquad u = 0, \pm 1, \pm 2, \ldots .$$

Theorem 7.5. *Suppose that*

1. $\epsilon(t)$, $t = 0, 1, \ldots$ *satisfy Condition 2 and have mean zero,*
2. $X(t)$ *is uniformly bounded and* $s_{11} \neq 0$,
3. $\sum_u |u|h(u) < \infty$,
4. W *in Condition 4 is a uniform kernel.*

Let $Qb \to 0$ *as* $T \to \infty$. *Then*

$$E\hat{h}(u) = h(u) + \sum_{q \neq 0} h(u + qQ) + O(b) + O(T^{-1/2}).$$

In particular, $\hat{h}(u)$ *is asymptotically unbiased. Furthermore,* $\hat{h}(u_1), \ldots, \hat{h}(u_M)$ *are asymptotically normal with mean* $h(u_1), \ldots, h(u_M)$ *and covariance structure*

$$\operatorname{cov}(\hat{h}(u), \hat{h}(v)) = \frac{2\pi}{QbT}\Lambda^T(u, v)\int W^2 + O\big(T^{-1}\big), \quad u, v = 0, \pm 1, \pm 2, \ldots,$$

where

$$\Lambda^T(u, v) = \frac{1}{Q}\sum_{q=0}^{Q} \exp\left(i\frac{2\pi(u - v)q}{Q}\right) s_{\epsilon\epsilon}(2\pi q/Q)/s_{11}(2\pi q/Q).$$

Proof. The proof of Theorem 7.5 is tedious and very computational. We outline the argument here. The proof starts by assuming the X series to be non-random. The asymptotic normality then follows from the computation of the joint cumulants, it is shown that cumulants of order greater than 2 of $\hat{h}(u_1), \ldots, \hat{h}(u_M)$ tend to zero as $T \to \infty$. The desired result for the random X follows by invoking a standard technique in nonparametric regression for handling ratio of two random variates.

Acknowledgements

We like to thank Mechelle Lewis, Andrew Smith, Suman Sen and Roxanne Poole for their comments and helps to make the data available to us. We are grateful to three referees for their critical and constructive comments that improve the readability of the paper significantly. Finally, we thank Allen Song for the physics behind MRI.

References

[1] Afifi, A. K. and Bergman, R. A. (1998). *Functional Neuroanatomy: Text and Altas*. McGraw-Hill, USA.
[2] Aguirre, G. K., Zarahn, E. and D'Esposito, M. (1998). The Variability of Human, BOLD Hemodynamic Responses. *NeuroImage*, **8**, 360–369.
[3] Alexander, G. E., DeLong, M. R. and Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci.*, **9**, 357–381.
[4] Bénar, C. G., Gross, D. W., Wang, Y., Petre, V., Pike, B., Dubeau, F. and Gotman, J. (2002). The BOLD Response to Interictal Epileptiform Discharges. *NeuroImage*, **17**, 1182–1192.
[5] Billingsley, P. (1995). *Probability and Measures, 3rd ed.*. Wiley, New York, USA.

[6] Boynton, G. M., Engel, S. A., Glover, G. H. and Heeger, D. J. (1996). Linear Systems Analysis of Functional Magnetic Resonance Imaging in Human V1. *The Journal of Neuroscience*, **16**, 4207–4221.

[7] Brillinger, D. R. (1974). Cross-Spectral Analysis of Processes with Stationary Invrements Including the Stationary G/G/∞ Queue. *AOP*, **2**, 815–827.

[8] Brillinger, D. R. (1975). The Identification of Point Process Systems. *AOP*, **3**, 909–929.

[9] Brillinger, D. R. (1978). A note on the estimation of evoked response. *Biological Cybernetics*, **31:3**, 141–144.

[10] Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*. Holden-Day, San Francisco, CA.

[11] Buckner, R. L. and Koutstaal, W. and Schacter, D. L. and Dale, A. M. and Rotte, M. and Rosen, B. R. (1998). Functional-Anatomic Study of Episodic Retrieval: II. Selective Averaging of Event-Related fMRI Trials to Test the Retrieval Success Hypothesis. *NeuroImage*, **7**, 163–175.

[12] Bullmore, E., Brammer, M., Williams, S. C. R., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R. and Sham, P. (1996). Statistical methods of estimation and inference for functional MR image analysis. *IEEE Trans. Med. Img.*, **35**, 261–277.

[13] Bullmore, E. T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E. and Brammer, M. J. (1999). Statistical methods of estimation and inference for functional MR image analysis. *IEEE Trans. Med. Img.*, **18**, 32–42.

[14] Caulhoun, V. D., Stevens, M. C., Pearlson, G. D. and Kiehl, K. A. (2004). fMRI analysis with the general linear model: removal of latency-induced amplitude bias by incorporation of hemodynamic derivative terms. *NeuroImage*, **22**, 252–257.

[15] Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, **29**, 162–173. http://afni.nimh.nih.gov/afni.

[16] Dale, A. M. and Buckner, R. L. (1997). Selective Averaging of Rapidly Presented Individual Trials Using fMRI. *Human Brain Mapping*, **5**, 329–340.

[17] DeLong, M. R., Alexander, G. E., Georgopoulos, A. P., Crutcher, M. D., Mitchell, S. J. and Richardson, R. T. (1984). Role of basal ganglia in limb movements. *Hum. Neurobiol.*, **2**, 235–244.

[18] Duann, J. R., Jung, T. P., Kuo, W. J., Yeh, T. C., Makeig, S., Hsieh, J. C. and Sejnowski, T. J. (1998). Single-Trial Variability in Event-Related BOLD Signals. *NeuroImage*, 823–835.

[19] Fried, I., Katz, A., McCarthy, G., Sass, K. J., Williamson, P. and Spencer, S. S. and Spencer, D. D. (1991). Functional organization of human supplementary motor cortex studied by electrical stimulation. *J. Neurosci.*, **11**, 3656-3666.

[20] Friston, K., Holmes, A., Worsley, K., Poline, J. B., Grasby, P., Williams, S., Frackowiak, R. and Turner, R. (1995). Analysis of fMRI time series revisited. *NeuroImage*, **2**, 45–53.

[21] Friston, K. J., Ashburner, J., Kiebel, S. J., Nichols, T. E. and Penny, W. D. (2007). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, Academic Press. http://www.fil.ion.ucl.ac.uk/spm/.

[22] Friston, K. J., Frith, C. D., Turner, R. and Frackowiak, R. S. J. (1995). Characterizing Evoked Hemodynamics with fMRI. *NeuroImage*, **2**, 157–165.

[23] Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D. and Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, **2**, 189–210.

[24] Friston, K. J., Jezzard, P. and Turner, R. (1994). Analysis of Functional MRI Time-Series. *Human Brain Mapping*, **1**, 153–171.

[25] Friston, K. J., Josephs, O., Rees, G. and Turner, R. (1998). Nonlinear event-related responses in fMRI. *Magnetic Resonance in Medicine*, **39**, 41–52.

[26] Genovese, C. R. (2000). A Bayesian Time-Course Model for Functional Magnetic Resonance Imaging Data (with discussion). *JASA*, **95:451**, 691–703.

[27] Glover, G. H. (1999). Deconvolution of impulse response in event-related fmri. *NeuroImage*, **9**, 416–429.

[28] Gössl, C., Fahrmeir, L. and Auer, D. P. (2001). Bayesian Modeling of the hemodynamic response function in BOLD fMRI. *NeuroImage*, **14**, 140–148.

[29] Goutte, C., Nielsen, F. A. and Hansen, L. K. (2000). Modeling the haemodynamic response in fMRI using smooth FIR filters. *IEEE Transactions on Medical Imaging*, **19**, 1188–1201.

[30] Huettel, S. A., Song, A. W. and McCarthy, G. (2004). *Functional Magnetic Resonance Imaging*. Sinauer Associates, Inc..

[31] Josephs, O., Turner, R. and Friston, K. (1995). Event-Related fMRI. *Human Brain Mapping*, **5**, 243–248.

[32] Kang, J. K., Bénar, C. G., Al-Asmi, A., Khani, Y. A., Pike, G. B., Dubeau, F. and Gotman, J. (2003). Using patient-specific hemodynamic response functions in combined EEG-fMRI studies in epilepsy. *NeuroImage*, **20**, 1162-1170.

[33] Lange, N. and Zeger, S. L. (1997). Non-linear Fourier Time Series Analysis for Human Brain Mapping by Functional Magnetic Resonance Imaging (with discussion). *JRSS*, **46:1**, 1–29.

[34] Lindquist, M. and Wager, T. D. (2007). Validity and power in hemodynamic response modeling: A comparison study and a new approach. *Human Brain Mapping*, Available online.

[35] Lindquist, M. and Wager, T. D. (2006). Validity and Power in Hemodynamic Response Modeling:

A comparison study and a new approach. *Human Brain Mapping*, DOI: 10.1002/hbm.20310.

[36] Maldjian, J. A., Laurienti, P. J. and Burdette, J. H. (2004). Precentral Gyrus Discrepancy in Electronic Versions of the Talairach Atlas. *NeuroImage*, **21:1**, 450–455.

[37] Marchini, J. L. and Ripley, B. D. (2000). A new statistical approach to detecting significant activation in functional MRI. *NeuroImage*, **12**, 366–380.

[38] Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E. and Buckner, R. L. (2000). Characterizing the Hemodynamic Response: Effects of Presentation Rate, Sampling Procedure, and the Possibility of Ordering Brain Activity Based on Relative Timing. *NeuroImage*, **11**, 735-759.

[39] Newton, H. J. (1988). *Timeslab: A Time Series Analysis Laboratory*. Wadsworth & Brooks/Cole.

[40] S. Ogawa, D. W. Tank, R. Menon, J. M. Ellermann, S. G. Kim, H. Merkle and K. Uğurbil (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *PNAS*, **89**, 5951–5955.

[41] Rosen, B. R., Buckner, R. L. and Dale, A. M. (1998). Event-related functional MRI: past, present, and future. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 773–780.

[42] Schwartz, L. (1966). *Mathematics for the Physical Sciences*, Hermann.

[43] Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I. and Flitney, D. E., Niazy, R., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M. and Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, **23(S1)**, 208–219. http://www.fmrib.ox.ac.uk/fsl/.

[44] Woolrich, M. W., Behrens, T. E. and Smith, S. M. (2004). Constrained linear basis sets for HRF modelling using Variational Bayes. *NeuroImage*, **21**, 1748–1761.

[45] Worsley, K. and Friston, K. (1995). Analysis of fMRI time series revisited-again. *NeuroImage*, **2**, 173–181.

[46] Zarahn, E., Aguirre, G. K. and D'Esposito, M. (1997). Empirical analyses of BOLD fMRI statistics. *NeuroImage*, **5**, 179–197.

[47] Zhang, C. M., Jiang, Y. and Yu, T. (2007). A comparative study of one-level and two-level semiparametric estimation of hemodynamic response function for fMRI data. *Statistics in Medicine*, **26**, 3845–3861.

IMS Collections
Vol. 0 (2009) 217–236
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000012

Mixed Models, Posterior Means and Penalized Least-Squares

Yolanda Muñoz Maldonado[1]

Michigan Technological University

Abstract: This paper reviews the connections between estimators that derive from three different modeling methodologies: Mixed-effects models, Bayesian models and Penalized Least-squares. Extension of classical results on the equivalence for smoothing spline estimators and best linear unbiased prediction and/or posterior analysis of certain Gaussian signal-plus-noise models is examined in a more general setting. These connections allow for the application of an efficient, linear time algorithm, to estimate parameters, compute random effects predictions and evaluate likelihoods in a large class of model scenarios. We also show that the methods of generalized cross-validation, restricted maximum likelihood and unbiased risk prediction can be used to estimate the variance components or adaptively select the smoothing parameters in any of the three settings.

Contents

1 Introduction 217
2 Equivalence Theorem 220
3 Examples 224
3.1 Varying Coefficient Models 225
3.2 Ridge Regression and Penalized Spline Regression 229
3.3 Mixed-Effects Model 230
4 Summary 232
A State-Space Forms 233
References 235

1. Introduction

Mixed-effects model methodology, penalized least-squares and Bayesian random-effects models are widely used statistical tools. However, due to the dissimilar nature of the settings in which they are typically formulated, connections between these three techniques as well as the fundamental reasons for the connections, have often been overlooked. In this paper, we review some of the well known results that connect smoothing spline estimators, Gaussian signal-plus-noise models and best linear unbiased prediction of mixed-effects models and show that they are but one aspect of a general framework that allows for "cross-platform" development in mixed-effects models, using frequentist or Bayesian approaches, and/or penalized

[1]Department of Mathematical Sciences, Michigan Technological University, Houghton, MI USA, email: ymunoz@mtu.edu
AMS 2000 subject classifications: Primary 62J05; Secondary 65C20
Keywords and phrases: Smoothing splines, p-splines, ridge regression, varying coefficient models, Bayesian prediction, Kalman fiter, adaptive selection

least-squares (PLS) criteria.

The relationship between particular cases of frequentist and Bayesian mixed-effects models and PLS has been exposed before. For example, Lindley and Smith [29] proposed the use of prior information on the parameters of a fixed effects linear model under the assumption of the parameters having exchangeable distributions. In the early development of the Bayesian theory for smoothing splines, Wahba [43] noticed the intimate connection between estimators resulting from spline smoothing and a Gaussian model with diffuse initial conditions. Robinson [35] remarked on applications of Best Linear Unbiased Predictors (BLUP's) for estimation of variance parameters, randomized block designs and their link to empirical Bayes methods and Kriging. Speed [40] pointed out, in a comment to Robinson's article, that smoothing spline estimators were in fact BLUP's of a certain mixed effects model. In the PLS framework, it is well known that smoothing splines estimators are a special case of penalized splines estimators (P-splines) [37]. Wahba [47] and Cressie [7] discussed the links between splines and kriging estimates and Nychka used the representation of smoothing splines as a type of ridge estimator to further relate smoothing spline estimation and kriging [32].

More recently, researchers have been using the connection between smoothing spline estimators and particular mixed-effects models to compute smoothing spline estimators [see 4, 19, 48]. Ruppert et al. [36] mentioned the correspondence between penalized spline smoothers and prediction in the mixed-effects model and remarked on the advantages of using existing mixed-effects model techniques and software in a semi-parametric regression setting. Eubank et al. [11] took advantage of the relationship between smoothing splines and the Gaussian model of [43] to provide a general development that includes the efficient computation of estimators in a varying coefficient model context.

Using connections that have been established for various special cases, we synthesize them and present a formal result that details precisely when penalized least-squares estimation, BLUP for a mixed-effects model and posterior mean analysis of a mixed-effects model with diffuse priors on some of the random effects (hereafter referred to as simply the Bayesian model) produce identical estimators. We then describe how this can be exploited in many cases of interest to provide a computationally efficient algorithm for evaluation of estimators and likelihoods, computation of predictions, and construction of Bayesian prediction intervals. The implemented algorithm reduces the computational effort of calculating the aforementioned quantities by two orders of magnitude over what would normally be the case for a direct mixed-effects model approach. We also establish a result showing that the methods of Generalized Cross-validation (GCV), Restricted Maximum Likelihood (REML) or the equivalent technique of Generalized Maximum Likelihood (GML) and Unbiased Risk Prediction (UBR) can be used in any of the three settings to adaptively estimate the smoothing parameters or variance components.

The following three examples will be used throughout the paper to illustrate the utility of our approach.

Example 1: Varying Coefficient Models. Varying coefficient models generalize ordinary linear regression models by allowing for regression coefficients that change dynamically as a function of independent variables. The simplest example of this are the so-called time varying coefficient models where there is only one effect modifying covariate. In that setting, we have response variables y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, that depend on some predictor variables $x_{1ij}, \ldots, x_{Kij}$ through a relationship of the

form

$$y_{ij} = \sum_{k=1}^{K} \beta_k(t_{ij}) x_{kij} + e_{ij}, \tag{1}$$

where the $\beta_k(\cdot)$'s are unknown coefficient functions of a covariate t and the e_{ij} represent random error terms. Models like (1) were first introduced by [21] who proposed obtaining estimators through minimization of the PLS criterion

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ y_{ij} - \sum_{k=1}^{K} f_k(t_{ij}) x_{kij}(t_{ij}) \right\}^2 + \sum_{k=1}^{K} \lambda_k \int_0^1 [f_k^{(r)}(t)]^2 dt \tag{2}$$

over functions $f_1, \ldots, f_K$ having r square integrable derivatives, and $g^{(s)}(t)$ being the s^{th}derivative of the function g. The parameters $\lambda_k \geq 0$ control the smoothness of the coefficient functions and the minimizers can be shown to be natural splines of degree $2r-1$ with knots at the unique elements of the set $\{t_{ij}\}$.

Example 2: Ridge Regression. Consider the linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{3}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of responses, $\mathbf{X}$ is a known $n \times p$ matrix of predictor variables of rank p, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients and $\boldsymbol{e}$ is a normally distributed vector of errors with $\mathrm{E}(\boldsymbol{e}) = \mathbf{0}$ and $\mathrm{E}(\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}) = \sigma_e^2 \mathbf{I}$, with "T" denoting the transpose of a matrix and $\mathbf{I}$ an identity matrix of suitable dimension. The generalized ridge regression estimator of $\boldsymbol{\beta}$ is then given by $\hat{\boldsymbol{\beta}} = [\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{K}]^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$. This estimator can be obtained by minimizing the PLS criterion

$$(\boldsymbol{y} - \mathbf{X}\boldsymbol{a})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{X}\boldsymbol{a}) + \boldsymbol{a}^{\mathrm{T}}\mathbf{K}\boldsymbol{a}, \tag{4}$$

over $\{\boldsymbol{a} : \boldsymbol{a} \in I\!R^p\}$, with $\mathbf{K}$ a diagonal matrix having elements $\lambda_i \geq 0$, for $i = 1, \ldots, p$. A special instance of (4) is given by ordinary ridge regression in which case the predictor variables are usually standardized and $\mathbf{K}$ has the form $\lambda\mathbf{I}$, for $\lambda > 0$. Other variations of generalized ridge regression are the P-splines estimators of [9] and of [36]. We will now describe the later approach in more detail.

Suppose that we have a collection of points on the plane, (t_i, y_i), $i = 1, \ldots, n$, and want to fit them using scatter-plot smoothing methodology. P-splines provide one popular approach for accomplishing this that arise from using a spline basis to construct the $\mathbf{X}$ matrix in (3). That is, for some integer $m \geq 0$ and a fixed set of knots $\xi_1 < \xi_2 < \cdots < \xi_p$, we take $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m+p}]$ with $\boldsymbol{x}_1$ a n-vector of all ones, $\boldsymbol{x}_j = [t_1^{j-1}, \ldots, t_n^{j-1}]^{\mathrm{T}}$, $j = 2, \ldots, m$ and $\boldsymbol{x}_{m+j} = [(t_1 - \xi_j)_+^{m-1}, \ldots, (t_n - \xi_j)_+^{m-1}]^{\mathrm{T}}$, for $j = 1, \ldots, p$ with $(x)_+^r$ being x^r for $x \geq 0$ and zero otherwise. A P-spline smoother is then found by minimizing (4), with the matrix $\mathbf{K}$ having the form

$$\mathbf{K} = \begin{bmatrix} \mathbf{0}_{m\times m} & \mathbf{0}_{m\times p} \\ \mathbf{0}_{p\times m} & \lambda\mathbf{I} \end{bmatrix}, \tag{5}$$

with $\mathbf{0}_{r\times s}$ being an r by s matrix of all zeros.

Example 3: Randomized Block Design. Linear mixed-effects models have been applied for analysis of data arising from situations involving repeated measures and experimental designs with factors that can be seen as a combination of fixed and

random effects. Some types of randomized block designs fall in the last category, for example, when the experimental units are randomly selected and each has repeated measurements. For this particular type of design, the experimental units are assumed to be the factor (or blocking criterion), that makes them relatively homogeneous with respect to a measured response. One way of modeling this type of problems is

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{b} + \boldsymbol{e}, \tag{6}$$

where $\mathbf{X}$ is the design matrix for the fixed-effects, $\boldsymbol{\theta}$ is the parameter vector for the fixed-effects and $\boldsymbol{b}$ is a random vector of blocking factors. This is not the only model that can be used with this type of design, but it will serve the purpose of this paper.

The remainder of the paper is organized as follows. In section 2 we present a result that connects estimators/predictions that are obtained from mixed-effects models, penalized least-squares estimation and Bayesian formulations. We also address the issue of estimation of the variance components and smoothing parameters that arise from their respective contexts. In this latter respect, we establish that GCV, REML/GML and UBR can all be used to obtain the above mentioned estimators. Section 3 illustrates the implementation of our main result using the three examples mentioned in this section. Section 4 concludes with some comments about the use of the theorems in section 2 and the employment of the Kalman filter algorithm.

2. Equivalence Theorem

To begin, we will give a detailed description of the three modeling scenarios that are the focus of this section.

- **Mixed-effects model :** Consider first the linear mixed-effects model

$$\boldsymbol{y} = \mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b} + \boldsymbol{e}, \tag{7}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of responses and $\mathbf{T}$ and $\mathbf{U}$ are design matrices for the fixed and random effects of dimensions $n \times m$ and $n \times q$, respectively. Here, we take $\boldsymbol{\theta}$ to be a $m \times 1$ vector of fixed effects and $\boldsymbol{b}$ to be a $q \times 1$ normally distributed random vector with zero mean and variance-covariance matrix $\mathrm{Var}(\boldsymbol{b}) = \sigma_b^2\mathbf{R}$. The random effects $\boldsymbol{b}$ are assumed to be independent of the $n \times 1$ vector of random errors, $\boldsymbol{e}$, which in turn, is assumed to be normally distributed with zero mean and variance-covariance matrix $\sigma_e^2\mathbf{I}$. For this model, as well as for the Bayesian model below, the parameters σ_e^2 and σ_b^2 are the so called variance components. It is often convenient to reparameterized the variance components as $\lambda = \sigma_b^2/\sigma_e^2$ so that $\mathrm{Var}(\boldsymbol{y}) = \sigma_e^2(\lambda\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} + \mathbf{I})$. The value of $\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}$ can be predicted using its BLUP.

- **Bayesian Model:** Similar to the previous case, in this setting we assume that

$$\boldsymbol{y} = \mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b} + \boldsymbol{e}, \tag{8}$$

with $\mathbf{T}$ and $\mathbf{U}$ fixed matrices. However, we now also take $\boldsymbol{\theta}$ to be random and model it as being independent of $\boldsymbol{b}$ and $\boldsymbol{e}$, with a zero mean, normal prior distribution having variance-covariance matrix $\mathrm{Var}(\boldsymbol{\theta}) = \nu\mathbf{I}$. The vector of random effects, $\boldsymbol{b}$, is also assumed to be normally distributed with zero mean and $\mathrm{Var}(\boldsymbol{b}) = \sigma_b^2\mathbf{R}$. Prediction of $\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}$ can be accomplished via the use of its posterior mean. In the absence of an informative prior for $\boldsymbol{\theta}$ a diffuse formulation can be employed

wherein ν is allowed to diverge. *Note: notice that this is not truly a Bayesian model since there are no priors on the variance components. It is named Bayesian model for the sake of identification.*

- **Penalized Least-Squares:** In this case we have $\boldsymbol{y} = \mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b} + \boldsymbol{e}$ with $\boldsymbol{\theta}$ and $\boldsymbol{b}$ being non random and $\boldsymbol{e}$ is a vector of zero mean, normally distributed random errors with variance-covariance matrix $\text{Var}(\boldsymbol{e}) = \sigma_e^2\mathbf{I}$. The parameters are to be estimated by minimizing the PLS criterion

$$\text{PLS}(\boldsymbol{a}, \boldsymbol{c}) = (\boldsymbol{y} - \mathbf{T}\boldsymbol{a} - \mathbf{U}\boldsymbol{c})^T(\boldsymbol{y} - \mathbf{T}\boldsymbol{a} - \mathbf{U}\boldsymbol{c}) + \lambda \boldsymbol{c}^T\mathbf{R}^{-1}\boldsymbol{c}, \tag{9}$$

with respect to $\boldsymbol{a}$ and $\boldsymbol{c}$. Here, $\mathbf{R}^{-1}$ is a penalty matrix and λ is the parameter that controls how heavily we penalize the coefficients $\boldsymbol{c}$.

Having these three scenarios in mind, we now state the following theorem.

Theorem 2.1. *The Best Linear Unbiased Predictor (BLUP) of* $\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}$ *in (7) is given explicitly by*

$$\hat{\boldsymbol{y}} = \mathbf{A}_\lambda \boldsymbol{y}, \tag{10}$$

where

$$\mathbf{A}_\lambda = \{\mathbf{I} - \mathbf{Q}^{-1}[\mathbf{I} - \mathbf{T}(\mathbf{T}^T\mathbf{Q}^{-1}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{Q}^{-1}]\}, \tag{11}$$

and

$$\mathbf{Q} = (\lambda \mathbf{U}\mathbf{R}\mathbf{U}^T + \mathbf{I}). \tag{12}$$

This result is numerically the same as the limiting value (as $\nu \to \infty$*) of* $E[\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y}]$ *in (8) and the minimizer of (9).*

Proof. To simplify the proof let us assume that the design matrices $\mathbf{U}$ and $\mathbf{T}$, as well as $\mathbf{R}$, are all full rank matrices (we will later relax this assumption).

Under model (7), the first two moments of $\boldsymbol{y}$ are given by

$$\text{E}(\boldsymbol{y}) = \mathbf{T}\boldsymbol{\theta} \quad \text{and} \quad \text{Var}(\boldsymbol{y}) = \sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\text{T}} + \sigma_e^2\mathbf{I}.$$

Using the distribution of $\boldsymbol{y}$ given $\boldsymbol{b}$ and the distribution of $\boldsymbol{b}$, we can then find the joint density of $\boldsymbol{y}$ and $\boldsymbol{b}$ and obtain the normal equations of [23]:

$$\begin{aligned} \mathbf{T}^{\text{T}}\mathbf{T}\boldsymbol{\theta} + \mathbf{T}^{\text{T}}\mathbf{U}\boldsymbol{b} &= \mathbf{T}^{\text{T}}\boldsymbol{y} \\ \mathbf{U}^{\text{T}}\mathbf{T}\boldsymbol{\theta} + (\mathbf{U}^{\text{T}}\mathbf{U} + \mathbf{R}_\lambda^{-1})\boldsymbol{b} &= \mathbf{U}^{\text{T}}\boldsymbol{y}, \end{aligned}$$

for $\mathbf{R}_\lambda = \lambda\mathbf{R}$.

After some algebra and using the Sherman-Morrison-Woodbury formula in [24] we have

$$\mathbf{Q}^{-1} = \mathbf{I} - \mathbf{U}(\mathbf{U}^{\text{T}}\mathbf{U} + \mathbf{R}_\lambda)^{-1}\mathbf{U}^{\text{T}}, \tag{13}$$

$$\hat{\boldsymbol{\theta}} = (\mathbf{T}^{\text{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1}\mathbf{T}^{\text{T}}\mathbf{Q}^{-1}\boldsymbol{y}$$

and

$$\hat{\boldsymbol{b}} = (\mathbf{U}^{\text{T}}\mathbf{U} + \mathbf{R}_\lambda^{-1})^{-1}\mathbf{U}^{\text{T}}[\mathbf{I} - \mathbf{T}(\mathbf{T}^{\text{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1}\mathbf{T}^{\text{T}}\mathbf{Q}^{-1}]\boldsymbol{y}.$$

In this way, the predicted values of $\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}$ are given by

$$\hat{\boldsymbol{y}} = \{\mathbf{I} - \mathbf{Q}^{-1}[\mathbf{I} - \mathbf{T}(\mathbf{T}^{\text{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1}\mathbf{T}^{\text{T}}\mathbf{Q}^{-1}]\}\boldsymbol{y}. \tag{14}$$

To show that minimization of the PLS criterion produces the same numerical answer as the BLUP of (7), we differentiate PLS($\boldsymbol{a}, \boldsymbol{c}$) with respect to $\boldsymbol{a}$ and $\boldsymbol{c}$ to obtain normal equations which together with (13) give us the same answer as in (14).

It remains to show that under the Bayesian model with diffuse prior, $\lim_{\eta\to\infty} \mathrm{E}(\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y})$ also agrees with (14). In this case, the joint distribution of $\mathbf{T}\boldsymbol{\theta}+\mathbf{U}\boldsymbol{b}$ and $\boldsymbol{y}$ is found to be normal with zero mean vector and variance-covariance matrix given by

$$\begin{pmatrix} \nu\mathbf{T}\mathbf{T}^{\mathrm{T}} + n^{-1}\sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} & \nu\mathbf{T}\mathbf{T}^{\mathrm{T}} + n^{-1}\sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} \\ (\nu\mathbf{T}\mathbf{T}^{\mathrm{T}} + n^{-1}\sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}})^{\mathrm{T}} & \nu\mathbf{T}\mathbf{T}^{\mathrm{T}} + n^{-1}\sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} + \sigma_e^2\mathbf{I} \end{pmatrix}.$$

Standard multivariate analysis results then produce

$$\begin{aligned} \mathrm{E}(\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y}) &= \mathrm{Cov}(\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}, \boldsymbol{y})[\mathrm{Var}(\boldsymbol{y})]^{-1}\boldsymbol{y} \\ &= (\nu\mathbf{T}\mathbf{T}^{\mathrm{T}} + n^{-1}\sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}}) \\ &\quad \times(\nu\mathbf{T}\mathbf{T}^{\mathrm{T}} + n^{-1}\sigma_b^2\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} + \sigma_e^2\mathbf{I})^{-1}\boldsymbol{y}. \end{aligned}$$

Letting λ be as in (12), $\eta = \nu/\sigma_e^2$ and recalling equation (12) we obtain

$$\mathrm{E}(\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y}) = (\eta\mathbf{T}\mathbf{T}^{\mathrm{T}} + \mathbf{U}\mathbf{R}_\lambda\mathbf{U}^{\mathrm{T}})(\eta\mathbf{T}\mathbf{T}^{\mathrm{T}} + \mathbf{Q})^{-1}\boldsymbol{y}. \tag{15}$$

Applying the Sherman-Morrison-Woodbury formula [24] on $(\eta\mathbf{T}\mathbf{T}^{\mathrm{T}} + \mathbf{Q})^{-1}$ and using a little algebra we get

$$\begin{aligned} (\eta\mathbf{T}\mathbf{T}^{\mathrm{T}} + \mathbf{Q})^{-1} &= \mathbf{Q}^{-1} - \\ &\quad \mathbf{Q}^{-1}\mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1}[\eta^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1} + \mathbf{I}]^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}. \end{aligned}$$

For η sufficiently large, the eigenvalues of $(\eta^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1}$ are all less than one. So, applying a power series expansion on $(\eta\mathbf{T}\mathbf{T}^{\mathrm{T}} + \mathbf{Q})^{-1}$ [16], substituting this expansion in (15), and with the aid of some straight forward calculus we have that $\lim_{\eta\to\infty}\mathrm{E}(\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y})$ is exactly the same expression as in (14).

Now, let us go back to our assumption of $\mathbf{U}$, $\mathbf{T}$ and $\mathbf{R}$ being full rank matrices. This may not be always the case. For example, if we approach estimation from the PLS criterion perspective, there are cases (such as spline smoothing), where $\mathbf{R}$ has less than full rank. To deal with this instance, suppose that the matrix $\mathbf{U}\mathbf{R}\mathbf{U}^T$ is not invertible. In this situation, the matrix $\mathbf{Q} = (\lambda\mathbf{U}\mathbf{R}\mathbf{U}^T + \mathbf{I})$ will still be invertible and our only concern is that the matrix $\mathbf{T}$ is less than full rank. In that case, we can employ conditional inverses [e.g., 18, pp. 31] and the theorem will still hold.

A result such as theorem 2.1 is important because, as pointed out by [4, 19, 48] and [36], one can take advantage of existing methodology and software to facilitate and enhance our analyses. The difference here is that theorem 2.1 is not restricted to the smoothing spline case of [43]; the BLUP result by [40] and referenced by [4, 19, 48]; or to the Bayesian mixed model of [29]. Instead we see that, quite generally, methodology from any particular one of the three frameworks can be potentially applied to obtain useful developments for the other two.

In each of the scenarios described by theorem 2.1, it will generally be necessary to estimate the parameter λ. The following result is a generalization of theorem 5.6 in [12] that allows us to apply three standard methods to the problem of adaptively selecting this parameter. The methods considered here are GCV, UBR and GML

which respectively produce estimators of λ via minimization of

$$\text{GCV}(\lambda) = \frac{n^{-1}\text{RSS}(\lambda)}{[n^{-1}\text{tr}(\mathbf{I}-\mathbf{A}_\lambda)]^2}, \tag{16}$$

$$\text{UBR}(\lambda) = n^{-1}\text{RSS}(\lambda) + 2n^{-1}\sigma_e^2\text{tr}(\mathbf{A}_\lambda), \tag{17}$$

and

$$\text{GML}(\lambda) = \frac{\boldsymbol{y}^T(\mathbf{I}-\mathbf{A}_\lambda)\boldsymbol{y}}{|\mathbf{I}-\mathbf{A}_\lambda|_+^{1/(n-m)}}. \tag{18}$$

Here, tr denotes the trace of a matrix, $\text{RSS}(\lambda) = (\boldsymbol{y}-\hat{\boldsymbol{y}})^T(\boldsymbol{y}-\hat{\boldsymbol{y}})$and $|\mathbf{I}-\mathbf{A}_\lambda|_+$ is the product of the nonzero eigenvalues of $\mathbf{I}-\mathbf{A}_\lambda$.

We note in passing that GML is equivalent to the method of REMS that is a popular approach to variance component estimation. [See, e.g. 40]. In terms of the relationship between criteria (16)-(18) we can establish the following result.

Theorem 2.2. E[GCV(λ)], E[UBR(λ)] *and* E[REMS/GML(λ)] *are all minimized at* $\lambda = \sigma_b^2/\sigma_e^2$.

Proof. To establish theorem 2.2 first note that the arguments in [12, pp. 244–247] can be easily modified to account for either the GCV or UBR part of the theorem. The main difference is that here we are not working with diffuse priors. Thus, we will concentrate on sketching the part of the proof that pertains to equation (18).

Let $\lambda_o = \sigma_b^2/\sigma_e^2$ and write $\mathbf{I}-\mathbf{A}_\lambda = \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{Q}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$, for a $\mathbf{B}$ such that $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}$, $\mathbf{B}\mathbf{B}^{\mathbf{T}} = \mathbf{I} - \mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1}$ and $\mathbf{B}^{\mathrm{T}}\mathbf{T} = 0$. Then,

$$\begin{aligned}\mathbf{B}^{\mathrm{T}}\mathbf{Q}\mathbf{B} &= \mathbf{B}^{\mathrm{T}}(n\lambda\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}} + \mathbf{I})\mathbf{B} \\ &= n\lambda\mathbf{B}^{\mathrm{T}}\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}}\mathbf{B} + \mathbf{I}.\end{aligned}$$

Define the matrix of eigenvalues for $\mathbf{B}^{\mathrm{T}}\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}}\mathbf{B}$ with corresponding matrix of eigenvectors $\mathbf{V}$ as $\mathbf{\Lambda} = \text{diag}\{d_1, \ldots, d_{n-m}\}$. Then, we can write

$$\mathbf{B}^{\mathrm{T}}\mathbf{Q}\mathbf{B} = \mathbf{V}(\lambda\mathbf{\Lambda} + \mathbf{I})\mathbf{V}^{\mathrm{T}}.$$

Now, taking expectation with respect to $\boldsymbol{e}$ and $\boldsymbol{b}$ we can show that

$$\begin{aligned}\text{E[REMS/GML}(\lambda)] &= \frac{\sigma_e^2\text{tr}[(\mathbf{I}-\mathbf{A}_\lambda)] + \lambda_o\text{tr}[(\mathbf{I}-\mathbf{A}_\lambda)(\mathbf{Q}-\mathbf{I})]}{[\prod_{i=1}^{n-m}(\lambda d_i+1)^{-1/(n-m)}]}, \\ &= \frac{\sigma_e^2}{\prod_{i=1}^{n-m}(\lambda d_i+1)^{-1/(n-m)}}\sum_{i=1}^{n-m}\frac{(\lambda_o d_i+1)}{(\lambda d_i+1)}.\end{aligned}$$

Now, take the difference of the logarithms of the expectations E[REMS/GML(λ)] and E[REMS/GML(λ_o)]. A sufficient condition for minimization of the REMS/GML criterion at λ_o is then seen to be

$$\log\left[\frac{1}{n-m}\sum_{i=1}^{n-m}\frac{(\lambda_o d_i+1)}{(\lambda d_i+1)}\right] - \frac{1}{(n-m)}\sum_{i=1}^{n-m}\log\left[\frac{(\lambda_o d_i+1)}{\lambda d_i+1}\right] \geq 0.$$

However, this is an immediate consequence of Jensen's inequality.

Criteria (16)–(18) have long been used for the selection of smoothing or penalty parameters. Golub et al. [17] proposed (16) as a method to choose the ridge regression parameter in a standard regression model like (3) and Craven and Wahba

[6] introduced GCV as a method for choosing the smoothing parameter in non-parametric regression. Wahba [46], Kohn et al. [27] and Stein [41] compared the performance of GCV and REML/GML for the smoothing spline case.

Unlike the methods of REML/GML in the PLS framework, GCV and UBR have not been applied in the context of mixed-effects models. Theorem 2.2 suggests that GCV may be another suitable method for estimation of variance components in this context. The fact that the GCV estimator of the variance components shares the REML/GML estimator attribute of minimizing the expectation of the risk, seems to indicate that both estimators will have similar properties and behavior under the mixed-effects model (as it has been shown for the PLS and the Bayesian models [see 27, 46]). However, this needs to be confirmed by studying the distributional and consistency properties of the GCV estimator of σ_e^2 and σ_b^2 under the mixed-effects model and this is a topic for future research.

3. Examples

In this section we focus on the examples introduced in section 1 and exemplify the advantages of using existing methodology for one particular framework (the Bayesian model) to the other two. In particular, we will use a Kalman filter algorithm to compute estimators and predictions that arise in the three scenarios considered in theorem 2.1. Perhaps the most common application of the Kalman filter has been in a Bayesian context [see 3, 28]. Specifically, Kohn and Ansley [25], using Wahba's Gaussian model (a particular case of our Bayesian model), reformulated the model into a state-space representation and thereby obtained an efficient $O(n)$ algorithm for computing smoothing spline estimators. Theorem 2.1 allows us to extend this approach to non spline smoothing situations and obtain an efficient, Kalman filter based, computational algorithm provided that the random components in theorem 2.2 admit a state-space representation. This algorithm also permits the evaluation of likelihood functions, making it possible to obtain REMS/GML estimators for variance components or smoothing parameters.

Description of the Kalman filter is beyond the scope of this paper. Instead, we will focus on establishing a state-space representation for the three examples and refer the reader to [11, 13] and [14] for a more complete development. To accomplish this, it suffices to give only a brief discussion concerning the form of a state-space model.

Any response y_i can be represented using a state-space model if the observation at time i can be expressed as a function of the observation at time $i-1$. More formally, a state-space model is composed of a set of response equations

$$y_i = \mathbf{h}^{\mathrm{T}}(t_i)\boldsymbol{x}(t_i) + e_i, \tag{19}$$

and a system of state equations

$$\boldsymbol{x}(t_{i+1}) = \mathbf{F}(t_i)\boldsymbol{x}(t_i) + \boldsymbol{u}(t_i). \tag{20}$$

with $t_i \in [0,1]$ and $0 = t_0 \leq t_1 < \cdots < t_n$. The y_i are observed quantities and the e_i, $\boldsymbol{u}(t_i)$, $\boldsymbol{x}(t_i)$, are all unobservable with $\boldsymbol{u}(t_0)\ldots,\boldsymbol{u}(t_{n-1})$, $e_1,\ldots,e_n$ and the initial state, $\boldsymbol{x}(t_0)$, all being zero mean, uncorrelated normal random variables. In general, the $\boldsymbol{x}(t_i)$ and $\boldsymbol{u}(t_i)$ may be vector valued with $\boldsymbol{u}(t_i)$ having variance-covariance matrix $\mathbf{R}_{u(t_i)}$. For our purposes we will treat the vectors $\mathbf{h}(t_i)$ and the transition matrix $\mathbf{F}(t_i)$ in (19)– (20) as being known.

We will proceed now to demonstrate the application of the equivalence theorem in the context of our three examples.

3.1. Varying Coefficient Models

To illustrate the varying coefficient case, we will examine the progesterone profiles data (figure 1) of [4]. The data consists of metabolite progesterone profiles, measured daily in urine over the course of a menstrual cycle in a group of 51 women.

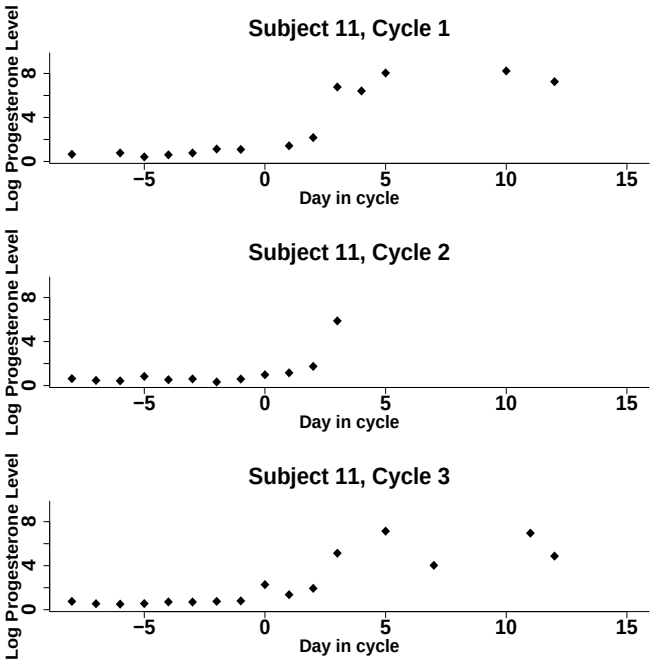


FIG 1. *Observed progesterone measurements for subject 11 in the non-conceptive group. The plots correspond to three of the four cycles for subject 11 and show the log progesterone concentration versus day in the cycle. All cycles have missing observations. Days corresponding to the menses were excluded.*

The women in the study were divided into two groups: 29 in the non-conceptive group and 22 in the conceptive group. Each woman contributed a different number of cycles, ranging from 1 to 5 cycles and some of the cycles have missing values.

The goal of the analysis is to detect differences between the conceptive and non-conceptive group profiles. To do this we will express the varying coefficient model (1) with the formulation in (9), apply theorem 2.1 and find the equivalent formulation (8) in the Bayesian framework in order to use the efficient Kalman filter algorithm of [13].

For simplicity, assume that we have complete data and the same number of cycles per woman (later we will relax these assumptions). Let the log progesterone level of the c^{th}cycle for the w^{th}woman at time t_i be denoted by y_{wci} and model this response as

$$y_{wci} \quad = \quad \beta_1(t_i)X_{1wci} + \beta_2(t_i)X_{2wci} + e_{wci}$$

where $i = 1, \ldots, 24$, and $t_1 = -8, t_2 = -7, \ldots, t_{24} = 15$ are the days in a menstrual cycle. The cycles c range from 1 to 5 and $w = 1, \ldots, 29$ correspond to women in the non-conceptive group and the rest belong to the conceptive group.

Assume that the $\beta_k(\cdot)$'s, $k = 1, 2$, are smooth functions of t. Usually, this translates into assuming that the functions belong to a Hilbert space of order m [see

22]. To find the estimated profiles we minimize a particular PLS criterion, where the penalty is applied to the integral of the square of the second derivative of the $\beta_k(\cdot)$'s. The minimizers of $\beta_1(\cdot)$, $\beta_2(\cdot)$ are natural splines of order m, with $m = 3$, that can be represented by a linear combinations of basis functions

$$\sum_{q=0}^{m-1} \theta_{kq} t_i^q + \sum_{r=1}^{24} b_{kr} \xi_r(t_i),$$

with knots at each of the design points t_i, and

$$\xi_r(t_i) = \int_0^{\min\{t_r,t_i\}} \frac{(t_i-u)^{m-1}(t_r-u)^{m-1}du}{[(m-1)!]^2}. \tag{21}$$

Equation (21) is one of the usual reproducing kernels of a Hilbert space of order m [22].

Let $\boldsymbol{y}_{wc} = [y_{wc}(t_1), \ldots, y_{wc}(t_{24})]^{\mathrm{T}}$ be the vector of responses for women w that contains all the daily observations in the c cycle, and $\boldsymbol{\xi}_i = [\xi_1(t_i), \xi_2(t_i), \ \ldots, \xi_{24}(t_i)]^{\mathrm{T}}$ and $\boldsymbol{t}_i = [t_i^0, t_i^1, \ldots, t_i^{m-1}]^{\mathrm{T}}$ be the vectors of basis functions evaluated at the times t_i's.

Denote the vector of coefficients for β_1 and β_2 as $\boldsymbol{\theta}_1 = [\theta_{10}, \theta_{11}, \ldots, \theta_{1,(m-1)}]^{\mathrm{T}}$, $\boldsymbol{\theta}_2 = [\theta_{20}, \theta_{21}, \ldots, \theta_{2,(m-1)}]^{\mathrm{T}}$, $\boldsymbol{b}_1 = [b_{10}, b_{11}, \ldots, b_{1,24}]^{\mathrm{T}}$, $\boldsymbol{b}_2 = [b_{21}, b_{22}, \ldots, b_{2,24}]^{\mathrm{T}}$, respectively. Construct $\boldsymbol{t}$, $\boldsymbol{\xi}$ and $\boldsymbol{X}$ such that

$$\boldsymbol{t} = \begin{bmatrix} \boldsymbol{t}_1^{\mathrm{T}} \\ \boldsymbol{t}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{t}_{24}^{\mathrm{T}} \end{bmatrix}, \boldsymbol{\xi} = \begin{bmatrix} \boldsymbol{\xi}_1^{\mathrm{T}} \\ \boldsymbol{\xi}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{\xi}_{24}^{\mathrm{T}} \end{bmatrix} \text{ and } X = \begin{bmatrix} X_{1wc1} & X_{2wc1} \\ X_{1wc2} & X_{2wc2} \\ \vdots & \vdots \\ X_{1wc24} & X_{2wc24} \end{bmatrix}.$$

Let $\mathbf{T}_{wc} = \boldsymbol{t} \bigotimes \mathbf{X}$ and $\mathbf{U}_{wc} = \boldsymbol{\xi} \bigotimes \mathbf{X}$, where $\boldsymbol{A} \bigotimes \mathbf{B}$ denotes the Kronecker product of the matrices $\boldsymbol{A}$ and $\mathbf{B}$ and it is equal to $a_{ij}\mathbf{B}$.

For each woman's cycle we have the model $\mathbf{T}_{wc}\boldsymbol{\theta}^\star + \mathbf{U}_{wc}\boldsymbol{b}^\star + \boldsymbol{e}_{ew}$, where $\boldsymbol{\theta}^\star = [\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}}]^{\mathrm{T}}$, $\boldsymbol{b}^\star = [\boldsymbol{b}_1^{\mathrm{T}}, \boldsymbol{b}_2^{\mathrm{T}}]^{\mathrm{T}}$ and $\boldsymbol{e}_{wc}$ is the corresponding vector of errors. Denote by $\boldsymbol{y}$ and $\boldsymbol{e}$ the vectors resulting from stacking the vectors $\boldsymbol{y}_{wc}$ and $\boldsymbol{e}_{wc}$, (i.e., $\boldsymbol{y} = [\boldsymbol{y}_{1,1}^{\mathrm{T}}, \boldsymbol{y}_{1,2}^{\mathrm{T}}, \ldots, \boldsymbol{y}_{1,5}^{\mathrm{T}}, \boldsymbol{y}_{2,1}^{\mathrm{T}}, \ldots, \boldsymbol{y}_{51,5}^{\mathrm{T}}]^{\mathrm{T}}$), and let $\mathbf{T} = \mathrm{diag}\{\mathbf{T}_{wc}\}_{\substack{w=1,51 \\ c=1,5}}$ and $\mathbf{U} = \mathrm{diag}\{\mathbf{U}_{wc}\}_{\substack{w=1,51 \\ c=1,5}}$. Then, we can construct the model $\mathbf{T}\boldsymbol{\theta}+\mathbf{U}\boldsymbol{b}+\boldsymbol{e}$, where $\boldsymbol{\theta} = \mathbf{1} \bigotimes \boldsymbol{\theta}^\star$, $\boldsymbol{b} = \mathbf{1} \bigotimes \boldsymbol{b}^\star$, and minimize criteria (9), where $\mathbf{R}^{-1} = \mathbf{U}$.

By theorem 2.1, this is equivalent to find $\lim_{\nu \to \infty} \mathrm{E}[\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y}]$, where $\boldsymbol{\theta}$, $\boldsymbol{b}$ and $\boldsymbol{e}$ are independent of each other and normally distributed with zero mean and variance-covariance matrices $\nu\mathbf{I}$, $\sigma_b^2\mathbf{U}^{-1}$, and $\sigma_e^2\mathbf{I}$, respectively. In this case, the smoothing parameter λ in the PLS model can be found using the variance components, σ_e^2 and σ_b^2 since $n\lambda = \sigma_b^2/\sigma_e^2$, where n is the total number of observations in the data.

The equivalent Bayesian representation of the varying coefficient model will allow us to make use of the Bayesian theory and apply it to our PLS setting. Specifically, we can follow [13] and transform the Bayesian model into a state-space model, as they indicate, and apply their efficient algorithm to compute the varying coefficients and respective confidence bands. Their approach also shows how to reformulate the matrices in the Bayesian model so the unbalanced design does not represent a problem in the computation of the estimators. For details on how to find the state-space model form, or on how to apply this efficient algorithm, we refer the readers to the

appendix and to the above mentioned authors, respectively.

To see what are the advantages of using this equivalence representation of the PLS, let us first explore the extent that the Kalman filter can speed up computations. To investigate this issue we carried out a run time comparison between our Kalman filter approach, the "standard way" of estimation assuming a mixed-effects model approach (both in SAS), and, only as a reference, we provide the time used in the method developed by Brumback and Rice [4]. We need to point out that these are the reported times in their 1998 paper and that there has been great improvement in computational speed since the publication of this paper. Table 1 shows the required times for computing the estimated conceptive and non-conceptive functions (see figure 2).

TABLE 1

Run time comparisons between the Kalman filter algorithm of [11] implemented in SAS, SAS proc mixed and Brumback and Rice's (1998) approach. Both Kalman filter and Brumback and Rice's approach include the time it took to calculate the smoothing parameter. The proc mixed time does not include this computation.

Method	**Real Time**
Kalman Filter	**14.60 secs.**
PROC MIXED	4 hrs. 12 mins. 15 secs.
Brumback and Rice	1 hr. 50 mins.

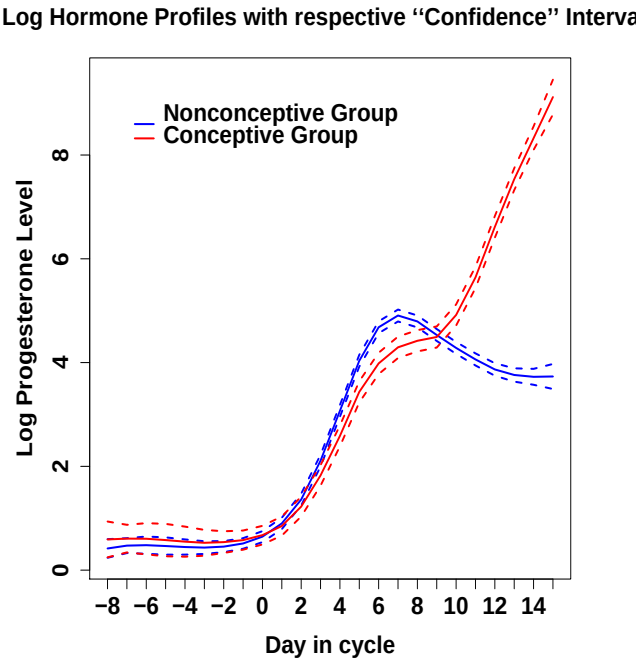


FIG 2. *Smooth estimates for non conceptive and conceptive mean groups with respective 95% pointwise confidence intervals. The corresponding smoothing parameters were computed using the GML method implemented through a Kalman filter algorithm.*

The first time in table 1 corresponds to the time employed by the Kalman filter algorithm of [11] implemented in SAS and using a computer with a 3.2GHz processor and 1G RAM. This algorithm used 2004 observations (missing values were

omitted) and calculated the estimated coefficient functions and corresponding 95% confidence intervals. The second time is the result of using a mixed-effects model representation and taking advantage of SAS proc mixed (the same equipment was used). The last time is the one reported by [4]. They implemented an eigenvalue-eigenvector decomposition on a mixed-effects model representation of the profiles, separately for each group, and combined the times for both groups and the estimation of the variance components. We calculated the smoothing parameters via REML/GML using the Kalman filter algorithm and it took approximately 10.5 seconds in SAS (this time is included in the computation of the Kalman filter in table 1). These parameters were used in both the SAS and Kalman filter calculation of the varying coefficient functions (we didn't want to calculate the smoothing parameters with SAS Proc Mixed given that it already took 4 hrs. to calculate the profiles without estimating the variance components). The reason why SAS takes so long to estimate the functions is due to the complex covariance structure of the model and the number of observations. The convenient SAS built-in covariance structures were not an option [see comment by 4], and the inversion of a general $n \times n$ matrix requires $O(n^3)$ operations versus the $O(n)$ used by the Kalman filter.

Another advantage of using the Bayesian interpretation in our PLS model is that the Kalman filter allows us to easily obtain confidence intervals as well as point estimators. In this respect, we use the relationship between PLS and the Bayesian model to provide Bayesian $100(1-\alpha)\%$ confidence (or prediction) intervals which parallel those developed by [44] and [31]. Specifically, we "estimate" the i^{th}component of $\beta_k(t_i)$ via the interval $\beta_k(t_i) \pm z_{1-\alpha/2}\sqrt{\hat{\sigma}_e^2 \times a_{ii}}$, where $\hat{\sigma}_e^2 = [(\boldsymbol{y}-\mathbf{A}_\lambda \boldsymbol{y})^T(\boldsymbol{y}-\mathbf{A}_\lambda \boldsymbol{y})]/(n-m)$, a_{ii} is the i^{th}diagonal element of the corresponding hat matrix $\mathbf{A}_\lambda$ for β_k and $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ standard normal percentile.

Wahba's "Bayesian Confidence Intervals" have been often used in the nonparametric community. Wahba [44] showed that the average of the coverage probability across points of these pointwise intervals is very close to nominal level for large n. She also commented that even if the confidence intervals are derived from a Bayesian perspective, they perform well in the frequentist realm. Nychka [32] offers an excellent discussion on why this is true.

In their paper, Brumback and Rice [4] utilized a hierarchical bootstrap method to assess the variability of the fitted functions instead of using the variance components estimators (it is well know that these estimators often underestimate the true parameters). For each bootstrapped sample 1.5 hours was required to obtain the estimated sample profiles (as reported by [4]). As a result, a partially parametric version of the method was implemented [see 4, for more details]). They computed 35 bootstrap samples and this took approximately 45 mins. In contrast, the confidence intervals computed in this paper for the progesterone profiles were obtained with the same computational effort involved in the estimation of the profiles.

Our estimated function profiles seem to agree with the ones obtained by Brumback and Rice. In addition, the "confidence" intervals also allow us to see that, on average, the production of progesterone in the conceptive group drops significantly from day 4 until around day 8 (when ovulation occurs) as compared to the hormone production of the non conceptive group. This result differs from the findings by [4]. Their bootstrap sample suggested that the decrease in progesterone for the conceptive group was not significant. The discrepancy between our findings and those of Brumback and Rice may be due to the small bootstrap sample they employed in their analysis or with our interpretation of the confidence intervals. Nychka [32]

pointed out that these intervals may not be reliable at specific points, even more if those points are part of a sharp peak or deep valley in the function. However, he also mentioned that it provides "a reasonable measure of the spline estimate's accuracy provided that the point for evaluation is chosen independently of the shape" of the function. It is known that the women are more fertile around day 3 to day 9, making it an interval targeted for studying before the start of the analysis. Also, we do not consider that the bump that forms in the interval of interest is that sharp. Hence, we believe that the confidence intervals provide reasonable evidence that the profiles are different.

3.2. Ridge Regression and Penalized Spline Regression

To exemplify the use of theorem 2.1 in the ridge regression setting, we have selected a data set that is freely distributed by *StatLib* at http://lib.stat.cmu.edu. The data set consists of 1150 heights measured at 1 micron intervals along the drum of a roller (i.e. parallel to the axis of the roller). The units of height are not given and the zero reference height is arbitrary.

To fit this data we used a model of the form (3) with $\mathbf{X} = [\mathbf{T}, \mathbf{U}]$ and corresponding vector of coefficients $\boldsymbol{\beta} = [\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{b}^{\mathrm{T}}]^{\mathrm{T}}$, where

$$(22) \qquad \mathbf{T} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_{1150} \end{bmatrix} \quad \text{and} \quad \mathbf{U} = \begin{bmatrix} (t_1-\xi_1)_+ & \cdots & (t_1-\xi_k)_+ \\ (t_2-\xi_1)_+ & \cdots & (t_2-\xi_k)_+ \\ \vdots & \cdots & \vdots \\ (t_{1150}-\xi_1)_+ & \cdots & (t_{1150}-\xi_k)_+ \end{bmatrix}.$$

The generalized ridge regression estimator of $\boldsymbol{\beta}$ is then obtained by minimizing the PLS criterion (4), with $\mathbf{K}$ as in (5) and $m = 2$.

Applying the results of theorem 2.1 we can write a parallel mixed-effects model representation for this ridge regression problem. This particular framework was considered by [36] who describe in section 4.9 of their book how to represent p-splines as BLUP's and illustrated how to use available software packages, like SAS proc mixed or the S-PLUS function lme, to obtain a fitted curve for the data. In view of the equivalence theorem, an alternative approach would be to use the connection between PLS and the Bayesian model so the Kalman filter can be implemented for purposes of computing estimators and "confidence" intervals. Another comprehensive description of the use of P-splines in the semi-parametric regression setting using Bayesian techniques is given in [5]. In this paper, we will use the Bayesian connection.

Assume that the vectors $\boldsymbol{\theta}$, $\boldsymbol{b}$ and $\boldsymbol{e}$ are independently normally distributed with zero mean and respective variance-covariance matrices $\nu\mathbf{I}$, $\sigma_b^2\mathbf{I}$ and $\sigma_e^2\mathbf{I}$. Then, by the equivalence theorem, the minimizer of (4) is the same as the limit, when ν is allowed to go to infinity, of the posterior mean of $\mathbf{T}\boldsymbol{\theta} + \mathbf{U}\boldsymbol{b}|\boldsymbol{y}$.

Again, this Bayesian model representation of the ridge regression example will permit the use of the Kalman filter algorithm for the computation of the estimated function and its respective "confidence intervals". For the explicit form of the state-space model see the appendix.

For this particular example, we considered two different model versions, one using $k = 150$ knots and the other with $k = 1150$ knots. This because we wanted to contrast the computational performance of the P-splines versus the computational effort required using smoothing splines and the Kalman filter. We should remark

here that, when $k = 1150$, basically we have a smoothing spline estimator which basis functions are the polynomials and the truncated power functions $(t_i - \xi_k)_+$, for $i = 1, \ldots, 1150$.

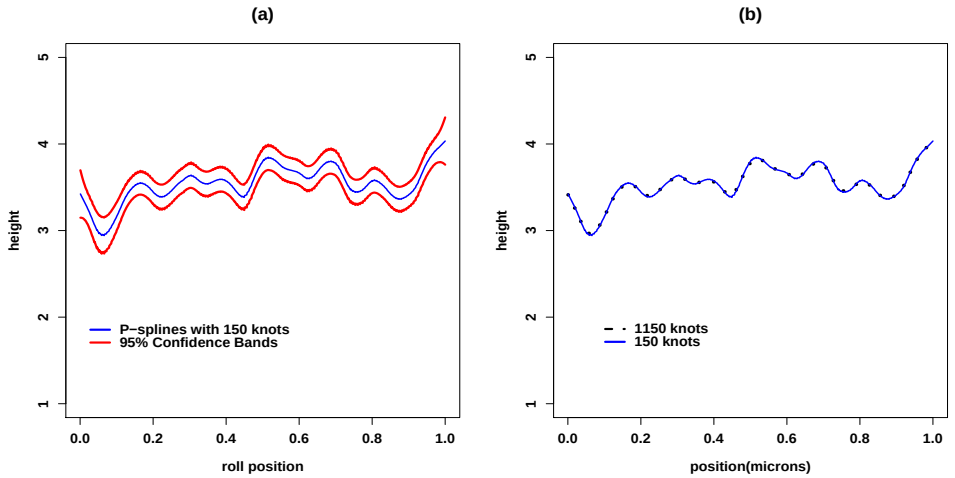


FIG 3. *Figure (a) shows the P-spline estimator with 150 knots for the roller height with its respective 95% confidence bands. The corresponding figure for the P-spline estimator with 1150 knots looks exactly the same and has been omitted. Figure (b) shows the comparison between the p-spline estimator with 150 knots and the P-spline with 1150 knots. There is no visual difference and both procedures yielded an estimated error variance of 0.36.*

Figure 3 shows the smooth estimated curve for the roller height and corresponding 95% "confidence intervals". The smoothing parameters were, respectively, $\lambda_{150} = 0.043$ and $\lambda_{1150} = 0.095$. They were selected via GCV and as we can see the GCV methods adjusts the smoothing parameters according to the number of knots used.

One of the main arguments in favor of using P-splines in lieu of smoothing splines is that, by reducing the number of knots involved in the model, we increase the computational efficiency involved in calculating the spline estimator. This is true when using brute force methods, i.e., direct inversion of the matrix (12). However, when using the proposed Kalman filter algorithm, the computational advantage of the P-splines over the smoothing splines disappear as we can see in table 2.

TABLE 2

Run time comparisons between a p-splines estimator with 150 knots, a P-spline estimator with 1150 knots and a smoothing spline estimator (technically 1150 knots but using as basis functions the polynomials and equation (21)). Both estimators were computed using code in R and the time does not include computation of the smoothing parameter.

Knots	**Real Time**
P-spline 150	48.34 secs.
P-spline 1150	54.55 secs.
Smoothing Spline	48.26 secs.

3.3. Mixed-Effects Model

In this last example, we want to illustrate the application of the equivalence theorem in the mixed-effects model setting. By finding the equivalent Bayesian model

of a mixed-effects model representation we will demonstrate the use of Kalman filtering for estimating parameters in a setting that it has seldom being used and that it can benefit from the reduced computational burden of estimating parameters and variance components. It is true that, if we have a "reasonable" number of observations and a specific covariance structure, like the ones provided by existing software, it will be advisable to use these procedures in lieu of the Kalman filter. However, there are occasions where the number of observations is really large. Then we can take advantage of the computational efficiency of the Kalman filter.

Our example deals with a randomized block design, where the data consists of 37 patients, which represent the random blocks, and a set of consecutive Hamilton depression scores measured over the course of 6 weeks (see figure 4). The data set is part of a study conducted by [34] and it is available at http://tigger.uic.edu/ hedeker/.

We model the data as

$$y_{ij} \quad = \quad \beta_0 + \beta_1 \text{week} + b_i + e_{ij},$$

where the y_{ij}'s are the depression scores, for $i = 1, \ldots, 37$, β_0 and β_1 are fixed parameters and week $= 0, 1 \ldots, 5$, is the week number where the score was measured. The random effects due to each patient are denoted by the b_i's and they are independent of the errors e_{ij}'s which are generated by an autoregressive process of order 1, i.e., $e_{ij} = \phi e_{i,j-1} + a_j(t_i)$, with ϕ a constant and $a_j(t_i)$ independent, identically distributed zero mean errors with variance σ_e^2.

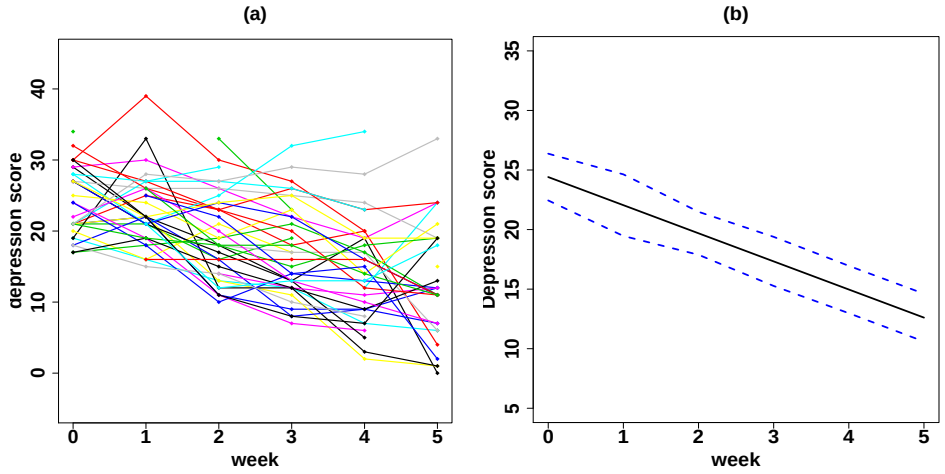


FIG 4. *(a) Hamilton depression scores for 37 patients measured over the period of 6 weeks. (b) Estimated regression line,* $y = 24.41 - 2.36$week, *with respective 95% confidence bands.*

Let $\boldsymbol{y}$ be the vector of depression scores such that $\boldsymbol{y} = [\boldsymbol{y}_1^{\mathrm{T}}, \ldots, \boldsymbol{y}_{37}^{\mathrm{T}}]^{\mathrm{T}}$, for $\boldsymbol{y}_i = [y_{i0}, y_{i1}, \ldots, y_{i5}]^{\mathrm{T}}$. Denote by $\mathbf{1}_n$, the vector of all ones of dimension $n \times 1$ and $\mathbf{week} = [0, 1, \ldots, 5]^{\mathrm{T}}$. In matrix form our model becomes

$$\boldsymbol{y} \quad = \quad \mathbf{T}\boldsymbol{\theta} + \boldsymbol{b} + \boldsymbol{e},$$

with $\boldsymbol{\theta} = [\beta_0, \beta_1]^{\mathrm{T}}$, $\mathbf{T} = [\mathbf{1}_{37} \bigotimes \mathbf{1}_5, \mathbf{1}_{37} \bigotimes \mathbf{week}]$, and $\boldsymbol{b} = \mathbf{1}_5 \bigotimes [b_1, b_2, \ldots, b_{37}]^{\mathrm{T}}$ and $\boldsymbol{e} = [\boldsymbol{e}_1^{\mathrm{T}}, \boldsymbol{e}_2^{\mathrm{T}}, \ldots, \boldsymbol{e}_{37}^{\mathrm{T}}]^{\mathrm{T}}$, for $\boldsymbol{e}_i = [e_{i0}, e_{i1}, \ldots, e_{i5}]^{\mathrm{T}}$. Here, we model the $\boldsymbol{b}$ as normally

distributed with zero mean and variance-covariance matrix $\mathbf{R} = \{\xi_r(t_1)\}_{\substack{r=1,5\\i=1,5}}$ and

$$\mathbf{W} = \begin{bmatrix} \frac{\sigma_e^2}{1-\phi^2} & \frac{\phi}{1-\phi^2} & \frac{\phi^2}{1-\phi^2} & \cdots & \frac{\phi^n}{1-\phi^2} \\ \frac{\phi}{1-\phi^2} & \frac{\sigma_e^2}{1-\phi^2} & \frac{\phi}{1-\phi^2} & \cdots & \frac{\phi^{n-1}}{1-\phi^2} \\ \vdots & \cdots & \ddots & \cdots & \vdots \\ \frac{\phi^n}{1-\phi^2} & \frac{\phi^{n-1}}{1-\phi^2} & \cdots & \frac{\phi}{1-\phi^2} & \frac{\sigma_e^2}{1-\phi^2} \end{bmatrix},$$

where W is the variance-covariance matrix of the AR(1) errors.

To find the corresponding Bayesian model, let $\boldsymbol{b}$ and $\boldsymbol{e}$ keep their distributions and assume that $\boldsymbol{\theta}$ is normally distributed with zero mean and variance-covariance matrix $\nu\mathbf{I}$. Once in the Bayesian form, we check that our observations Y_{ij} can be represented using the state-space equations (19-20). The equivalence theorem hold regardless of the state-space structure but, if we have that structure, then we can apply the efficient Kalman filter algorithm of [13] and estimate all our parameters with linear computational efficiency.

Figure 4 shows the estimated regression line for the Hamilton depression scores over the 6 week period. The variance components for this example are estimated via REML/GML and are $\hat{\phi} = 0.97$, $\hat{\sigma}_e^2 = 1.214$ and $\hat{\sigma}_b^2 = 0.00132$. The corresponding estimated values for the regression coefficients are $\hat{\theta}_0 = 24.41$ and $\hat{\theta}_1 = -2.36$.

4. Summary

In this paper, we have reviewed known results concerning the numerical equivalence of 1) a smoothing spline estimator and a particular mixed-effects model and 2) a smoothing spline estimator and the posterior mean of Wahba's Gaussian model and focus on the more general framework of frequentist and Bayesian mixed-effects models and penalized least-squares estimation as seen in Theorem 2.1. This result broadens the number of methodological resources available for computing BLUPs, posterior means, likelihoods and minimizers of penalized least squares criteria and facilitates the use of existing methodological tools, as exemplified by theorem 2.2 and our examples.

The link between the Bayesian mixed-effects model and the two other model settings allowed us to obtain Bayesian "confidence" intervals for the profile groups (instead of the computationally demanding bootstrap method of Brumback and Rice) and facilitated the analysis of the profile differences during the fertile days. Example 2 showed us that the Kalman filter implementation is not restricted to Wahba's Bayesian model. More generally, the idea carries over to settings involving p-splines, Kernel estimators, differences, etc. Lastly, this link allows for the implementation of a computationally efficient Kalman filter algorithm in many cases of interest. Kalman filter algorithms have been used to compute smoothing splines type estimators [19, 25, 26, 48]. But, they have been sparsely used in mixed-effects model settings. To this author knowledge, only [38] and, more recently, [30] have applied the Kalman filter to mixed-effects models. In the mixed-effects framework, the techniques employed for the analysis of large data sets require the use of computer intensive methods like the EM or MCMC algorithms [1, 39], conjugate gradient iterative methods [42], or the use of high performance computing environments. Some of the methods mentioned in these references assume that observations are generated by Brownian motion or ARMA processes and, whenever we have this type of processes, we have a state-space structure that can be exploited, as demonstrated

in our examples, to reduce the computational burden. Observations generated by longitudinal analysis (as in example 3), repeated measurements or any process that depends on an ordering variable can also frequently be assumed to have a state-space representation and can, as a result, benefit from the computational efficiency of the Kalman filter.

Appendix A: State-Space Forms

In this section we will explicitly describe the state-space forms used for the application of the Kalman filter in each of our examples. Since the form of the errors $\boldsymbol{u}(t_i)$ in equation (20) is assumed to be the same for the varying-coefficient case and the mixed-effects model, we will show the derivation for the varying coefficient case and detail the small changes needed for the mixed-effects case. We will leave for the last the ridge regression example.

To employ the Kalman filter for computation of the varying coefficient example we need to show that the varying coefficients have a state-space representation. That is, we need to be able to write equation (1) using equations (19)– (20). Since the $\beta_k(\cdot)$ are assumed to be smooth functions of t, we model them as

$$\beta_k(t_i) = \sum_{q=0}^{m-1} \theta_{kq} t^q + \sigma_b^2 Z_k(t_i), \tag{A.1}$$

for $k = 1, 2$ and $m = 2$, where (without loss of generality) we can take t in $[0, 1]$ and

$$Z_k(t) = \int_0^1 \frac{(t-u)_+^{m-1}}{(m-1)!} dW_k(u),$$

with $W_k(\cdot)$ standard Wiener processes. To simplify matters, first assume that $\beta_k(t_i) = \sigma_b^2 Z_k(t_i)$. Then, $\beta_k(t_{i+1})$ can be written as σ_b^2 times

$$\int_0^{t_i} \frac{(t_{i+1}-s)_+^{m-1}}{(m-1)!} dW_k(s) + \int_{t_i}^{t_{i+1}} \frac{(t_{i+1}-s)_+^{m-1}}{(m-1)!} dW_k(s).$$

Taking

$$u_k(t_i) = \int_{t_i}^{t_{i+1}} \frac{(t_{i+1}-u)_+^{m-1}}{(m-1)!} dW_k(u),$$

for $t_i < t_j$, the covariance between $u_k(t_i)$ and $u_k(t_j)$ is found to be equal to

$$\int_0^{t_i} \frac{(t_i-u)^{m-1}(t_j-u)^{m-1}}{[(m-1)!]^2} du.$$

For the remaining integral, add and subtract t_i inside $(t_{i+1}-u)^{m-1}$ and apply the Binomial theorem. Upon doing this, a state-space representation results with $\mathbf{F}(t_i)$ equal to

$$\mathbf{F}(t_i) = \begin{bmatrix} 1 & (t_{i+1}-t_i) & \frac{(t_{i+1}-t_i)^2}{2!} & \dots & \frac{(t_{i+1}-t_i)^{m-1}}{(m-1)!} \\ 0 & 1 & (t_{i+1}-t_i) & \dots & \frac{(t_{i+1}-t_i)^{m-2}}{(m-2)!} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & . & . & . & 1 \end{bmatrix}, \tag{A.2}$$

$\mathbf{Z}_k(t_i) = [Z_k(t_i), Z_k^{(1)}(t_i), \ldots, Z_k^{(m-1)}(t_i)]^T$, $\boldsymbol{u}_k(t_i) = [u_k(t_i), u_k^{(1)}(t_i), \ldots, u_k^{(m-1)}(t_i)]^T$ and $\boldsymbol{Z}_k(t_{i+1}) = \mathbf{F}(t_i)\boldsymbol{Z}_k(t_i) + \boldsymbol{u}_k(t_i)$.

Now, rearranging the observations with respect to the time t_i define

$$\boldsymbol{y}_{iw}^{\mathrm{T}} = [y_{iw1}, \ldots, y_{iwc_w}]^{\mathrm{T}}$$

with $\boldsymbol{y}_{iw}^{\mathrm{T}}$ the responses for woman w at time t_i observed at cycles $1, \ldots, c_w$, with corresponding vector of random errors $\boldsymbol{e}_{iw}$. Let $\boldsymbol{x}(t_i) = [\boldsymbol{Z}_1(t_i), \boldsymbol{Z}_2(t_i)]^T$, $\boldsymbol{u}(t_i) = [\boldsymbol{u}_1(t_i), \boldsymbol{u}_2(t_i)]^T$ and $\boldsymbol{X}_{kwi} = [X_{kw1i}, \ldots, X_{kwc_wi}]^T$. Then, taking

$$\boldsymbol{h}(t_i) = [\boldsymbol{X}_{1wi}^T, \mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{X}_{2wi}^T, \mathbf{0}, \ldots, \mathbf{0}]^T,$$

we arrive at the state-space model

$$\boldsymbol{y}_{iw} = \boldsymbol{h}(t_i)\boldsymbol{x}(t_i) + \boldsymbol{e}_{iw},$$

$$\boldsymbol{x}(t_{i+1}) = \mathbf{F}^{\star}(t_i)\boldsymbol{x}(t_i) + \boldsymbol{u}(t_i),$$

where $\mathbf{F}^{\star}(t_i)$ is the block diagonal matrix of size $2m \times 2m$ with diagonal blocks $\mathbf{F}(t_i)$, $i = 1, \ldots, n$.

Application of the standard Kalman filter to the vector of observations $\boldsymbol{y}_{iw}$ will yield coefficient functions estimates that disregard the polynomial term in (A.1). To account for that, we must employ the diffuse Kalman filter as in [13]. This entails a slight modification of our approach wherein the Kalman filter is applied to the vector of observations $\boldsymbol{y}_{iw}$ and each of the vectors $\mathbf{1}_{\boldsymbol{n}}$, $\boldsymbol{t}$, $\boldsymbol{t}^2 \ldots, \boldsymbol{t}^{(m-1)}$, where $\boldsymbol{t}^r = [t_1^r, t_2^r, \ldots, t_n^r]^{\mathrm{T}}$ [see 11, for a detailed derivation].

For our mixed-effects example we need to show that $\boldsymbol{e}$ can be represented in a state-space form and stack the respective state vectors, errors and matrices. We will proceed as follows: since the errors $e(t_i)$ are generated by an AR(1) process, they can be written as $e(t_{i+1}) = \phi e(t_i) + a_j(t_i)$, with ϕ a non random coefficient. This entails that the transition matrix $\mathbf{F}^{\star}(t_i) = diag\{\mathbf{F}(t_i), \phi\}$, with $\mathbf{F}(t_i)$ as in (A.2) and $\boldsymbol{h}(t_i) = [1, 0, 1]$. Take the state vector, $\boldsymbol{x}(t_i)$, to be equal to $[\boldsymbol{Z}_k(t_i)^{\mathrm{T}}, e(t_i)]^{\mathrm{T}}$, $\boldsymbol{u}(t_i) = [\boldsymbol{u}_k(t_i)^{\mathrm{T}}, a_j(t_i)]^{\mathrm{T}}$ with $m = 2$, where $\boldsymbol{Z}_k(t_i)$ and $\boldsymbol{u}_k(t_i)$ are as in the varying coefficient case. Specific details about the form of the state vector and the vector $\boldsymbol{u}(t_i)$ of the state equation (20), as well as a more general form for an ARMA model, can be found in [14].

Lastly, the state-space representation for the ridge regression example is found by taking the state vector to be $\boldsymbol{x}(t_i) = [x(t_i), x^{(1)}(t_i), \ldots, x^{(m-1)}(t_i)]^T$, with $x(t_i) = \sum_{k=1}^{j} \beta_k (t_i - \xi_k)^{m-1}$ for $t_i \in [\xi_j, \xi_{j+1})$ (using the definition of the truncated power function), and $x^{(r)}(t_i)$ the r^{th}derivative of $x(t_i)$, $r = 1, \ldots, (m-1)$. Then,

$$\boldsymbol{x}(t_{i+1}) = \mathbf{F}(t_i)\boldsymbol{x}(t_i) + \boldsymbol{u}(t_i),$$

with $\mathbf{F}(t_i)$ as in (A.2) and $\boldsymbol{u}(t_i) = [u(t_i), u^{(1)}(t_i), \ldots, u^{(m-1)}(t_i)]^{\mathrm{T}}$, where

$$u(t_i) = \begin{cases} 0 & \text{if} \quad t_{i+1} \in [\xi_j, \xi_{j+1}) \\ \beta_{j+1}(t_{i+1} - \xi_{j+1})^{m-1} & \text{if} \quad t_{i+1} \in [\xi_{j+1}, \xi_{j+2}) \end{cases}.$$

To complete the state-space formulation, take the vector $\mathbf{h}(t_i)$ to have dimension $m \times 1$ with one in the first position and the rest of its elements equal to zero.

References

[1] Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Statistical Methods in Medical Research*, **55**, 117–128.

[2] Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering*. Prentice Hall, Englewood Cliffs, New Jersey.

[3] Ansley, C. F. and Kohn, R. (1985). Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions. *Annals of Statistics*, **13**, 1286–1316.

[4] Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93**, 961–993.

[5] Crainiceanu, C. M., Ruppert, D. and Wand, M. P. (2005). Bayesian analysis for penalized spline regression using WinBugs. *Journal of Statistical Software*, **14**, 1–24.

[6] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerical Mathematic*, **31**, 377–403.

[7] Cressie, N. (1990). Reply: letters to the editor. *The American Statistician*, **44**, 256–258.

[8] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–22.

[9] Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–102.

[10] Eppright, E., Fox, H., Fryer, B., Lamkin, G., Vivian, V. and Fuller, E. (1972). Nutrition of infants and preschool children in the North Central Region of the United States of America. *World Review of Nutrition and Dietetics*, **14**, 269–332.

[11] Eubank, R. L., Huang, C., Muñoz Maldonado, Y., Wang, N., Wang, S. and Buchanan, R. J. (2004). Smoothing spline estimation in varying coefficient models. *Journal of the Royal Statistical Society, Series B*, **66**, 653–667.

[12] Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*, 1*st* ed. Marcel Dekker Inc., New York.

[13] Eubank, R. L., Huang, C. and Wang, S. (2003). Adaptive order selection for spline smoothing. *Journal of Computational and Graphical Statistics*, **12**, 2546–2559.

[14] Eubank, R. L. (2006). *A Kalman Filter Primer*. Chapman & Hall/CRC, Boca Raton, FL.

[15] de Finetti, B. (1964). Foresight: its logical laws, its subjective sources in *Studies in Subjective Probability*. Wiley, New York.

[16] Gantmakher, F. R. (1959). *The Theory of Matrices*. Chelsea Pub. Co., New York.

[17] Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized-cross validation as a method for choosing a good ridge parameter. *Technometrics*, **58**, 215–223.

[18] Graybill, F. A. (1976). *The theory and application of the linear model*. Duxbury, North Scituate, Massachusetts.

[19] Guo, W. (2002). Functional mixed effects models. *Biometrics*, **58**, 121–128.

[20] Guo, W. (2003). Functional data analysis in longitudinal settings using smoothing splines. *Statistical Methods in Medical Research*, **13**, 1–24.

[21] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, **4**, 757–796.

[22] Heckman, N. (1997). The theory and application of penalized least squares methods or reproducing kernel Hilbert spaces made easy. *Technical Report* **216**. The University of British Columbia.

[23] Henderson, C. R., Kempthorne, O., Searle, S. R. and Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **15**, 192–218.

[24] Householder, A. (1964). *The Theory aof Matrices in Numerical Analysis*. Dover, New York.

[25] Kohn, R. and Ansley, C. F. (1987). A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM Journal of Scientific and Statistical Computing*, **8**, 33–48.

[26] Kohn, R. and Ansley, C. F. (1989). A fast algorithm for signal extraction, influence and cross-validation in state-space models. *Biometrika*, **76**, 65–79.

[27] Kohn, R., Ansley, C. F. and Tharm, D. (1993). Performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the American Statistical Association*, **86**, 1042–1050.

[28] Koopman, S. J. and Durbin, J. (1998). Fast filtering and smoothing for multivariate state space models. *Journal of Times Series Analysis*, **21**, 281–296.

[29] Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, **34**, 1–41.

[30] Piehpo, H. P. and Ogutu, J. O. (2007). Simple state-space models in a mixed-model framework. *The American Statistician*, **61**, 224–232.

[31] Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association*, **83**, 1134-1143.

[32] Nychka, D. (2000). Spatial process estimates as smoothers. *Smoothing and Regression. Approaches, Computation and Application*. Wiley, New York.

[33] Patterson, H. D. and Thompson, R. (1971). Recovery of interblock information when cell sizes are unequal. *Biometrika*, **58**, 545–554.

[34] Riesby, N., Gram, L.F., Bech, P., Nagy, A., Petersen, G. O., Ortmann, J., Ibsen, I., Dencker, S.

J., Jacobsen, O., Krautwald, O., Søndergaard, I. and Chirstiansen, J. (1977). Imipramine: clinical effects and pharmacokinetic variability. *Psychopharmacology*, **54**, 263–272.

[35] Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, **6**, 15–32.

[36] Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

[37] Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735–757.

[38] Sallas, W. M. and Harville, D. A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, **76**, 860–869.

[39] Schafer, J. L. and Recai, M. Y. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, **11**, 437–457.

[40] Speed, T. (1991). That BLUP is a good Thing: the estimation of random effects: comment. *Statistical Science*, **6**, 42–44.

[41] Stein, M. L. (2000). A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *Annals of Statistics*, **18**, 1139–1157.

[42] Strandén, I. and Lidauer, M. (1999). Solving large mixed linear models using preconditioned conjugate gradient iteration. *Journal of Diary Science*, **82**, 2779-2787.

[43] Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, Series B*, **40**, 364–372.

[44] Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Series B*, **40**, 364–372.

[45] Wahba, G. (1990). Spline models for observational data. *SIAM*, **59**. Philadelphia, Pennsylvania.

[46] Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics*, **13**, 1378–1402.

[47] Wahba, G. (1990). Comment on Cressie: letters to the editor. *The American Statistician*, **44**, 255–256.

[48] Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association*, **93**, 341–348.

[49] Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B*, **60**, 159–174.

IMS Collections
Vol. 0 (2009) 237–251
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000007

From Charged Polymers to Random Walk in Random Scenery

Xia Chen[1,*] **and Davar Khoshnevisan**[2,†]

University of Tennessee and University of Utah

Abstract: We prove that two seemingly-different models of random walk in random environment are generically quite close to one another. One model comes from statistical physics, and describes the behavior of a randomly-charged random polymer. The other model comes from probability theory, and was originally designed to describe a large family of asymptotically self-similar processes that have stationary increments.

Contents

1 Introduction and the Main Results 237
2 Preliminary Estimates 239
3 Proofs of the Main Results 245
Acknowledgement. 250
References 251

1. Introduction and the Main Results

The principal goal of this article is to show that two apparently-disparate models—one from statistical physics of disorder media (Kantor and Kardar (1991), Derrida et al (1992), Derrida and Higgs (1994)) and one from probability theory (Kesten and Spitzer (1979), Bolthausen (1989))—are very close to one another.

In order to describe the model from statistical physics, let us suppose that $q := \{q_i\}_{i=1}^{\infty}$ is a collection of i.i.d. mean-zero random variables with finite variance $\sigma^2 > 0$. For technical reasons, we assume here and throughout that

$$\mu_6 := \mathrm{E}(q_1^6) < \infty. \tag{1.1}$$

In addition, we let $S := \{S_i\}_{i=0}^{\infty}$ denote a random walk on $\mathbf{Z}^d$ with $S_0 = 0$ that is independent from the collection q. We also rule out the trivial case that S_1 has only one possible value.

The object of interest to us is the random quantity

$$H_n := \sum_{1\le i<j\le n} q_i q_j \mathbf{1}_{\{S_i=S_j\}}. \tag{1.2}$$

[1]Department of Mathematics, University of Tennessee, Knoxville, TN 37996-1300, email: xchen@math.utk.edu
*Research supported in part by NSF grant DMS-0704024.
[2]Department of Mathematics, University of Utah, Salt Lake City, UT 84112-0090, email: davar@math.utah.edu
†Research supported in part by NSF grant DMS-0706728.
AMS 2000 subject classifications: primary 60K35; secondary 60K37
Keywords and phrases: Polymer measures, random walk in random scenery

In statistical physics, H_n denotes a random Hamiltonian of spin-glass type that is used to build Gibbsian polymer measures. The q_i's are random charges, and each realization of S corresponds to a possible polymer path; see the paper by Kantor and Kardar (1991), its subsequent variations by Derrida et al (1992, 1994) and Wittmer et al (1993), and its predecessos by Garel and Orland (1988) and Obukhov (1986). The resulting Gibbs measure then corresponds to a model for "random walk in random environment." Although we do not consider continuous processes here, the continuum-limit analogue of H_n has also been studied in the literature (Buffet and Pulé (1997), Martìnez and Petritis (1996)).

Kesten and Spitzer (1979) introduced a different model for "random walk in random environment," which they call *random walk in random scenery*.[1] We can describe that model as follows: Let $Z := \{Z(x)\}_{x\in\mathbf{Z}^d}$ denote a collection of i.i.d. random variables, with the same common distribution as q_1, and independent of S. Define

$$W_n := \sum_{i=1}^{n} Z(S_i). \tag{1.3}$$

The process $W := \{W_n\}_{n=0}^{\infty}$ is called *random walk in random scenery*, and can be thought of as follows: We fix a realization of the d-dimensional random field Z—the "scenery"—and then run an independent walk S on $\mathbf{Z}^d$. At time j, the walk is at S_j; we sample the scenery at that point. This yields $Z(S_j)$, which is then used as the increment of the process W at time j.

Our goal is to make precise the assertion that if n is large, then

$$H_n \approx \gamma^{1/2}\cdot W_n \quad \text{in distribution}, \tag{1.4}$$

where

$$\gamma := \begin{cases} 1 & \text{if } S \text{ is recurrent},\\ \sum_{k=1}^{\infty} \mathrm{P}\{S_k = 0\} & \text{if } S \text{ is transient}. \end{cases} \tag{1.5}$$

Our derivation is based on a classification of recurrence vs. transience for random walks that appears to be new. This classification [Theorem 2.4] might be of independent interest.

We can better understand (1.4) by considering separately the cases that S is transient versus recurrent. The former case is simpler to describe, and appears next.

Theorem 1.1. *If S is transient, then*

$$\frac{W_n}{n^{1/2}} \xrightarrow{\mathcal{D}} N(0\,,\sigma^2) \quad \textit{and} \quad \frac{H_n}{n^{1/2}} \xrightarrow{\mathcal{D}} N(0\,,\gamma\sigma^2). \tag{1.6}$$

Kesten and Spitzer (1979) proved the assertion about W_n under more restrictive conditions on S. Similarly, Chen (2008) proved the statement about H_n under more hypotheses.

Before we can describe the remaining [and more interesting] recurrent case, we define

$$a_n := \left(n\sum_{k=0}^{n}\mathrm{P}\{S_k = 0\}\right)^{1/2}. \tag{1.7}$$

[1] Kesten and Spitzer ascribe the terminology to Paul Shields.

It is well known (Polya (1921), Chung and Fuchs (1951)) that S is recurrent if and only if $a_n/n^{1/2} \to \infty$ as $n \to \infty$.

Theorem 1.2. *If S is recurrent, then for all bounded continuous functions $f : \mathbf{R}^d \to \mathbf{R}$,*

$$\mathrm{E}\left[f\left(\frac{W_n}{a_n}\right)\right] = \mathrm{E}\left[f\left(\frac{H_n}{a_n}\right)\right] + o(1), \tag{1.8}$$

where $o(1)$ converges to zero as $n \to \infty$. Moreover, both $\{W_n/a_n\}_{n\geq 1}$ and $\{H_n/a_n\}_{n\geq 1}$ are tight.

We demonstrate Theorems 1.1 and 1.2 by using a variant of the replacement method of Liapounov (1900) [pp. 362–364]; this method was rediscovered later by Lindeberg (1922), who used it to prove his famous central limit theorem for triangular arrays of random variables.

It can be proved that when S is in the domain of attraction of a stable law, W_n/a_n converges in distribution to an explicit law (Kesten and Spitzer (1979), Bolthausen (1989)). Consequently, H_n/a_n converges in distribution to the same law in that case. This fact was proved earlier by Chen (2008) under further [mild] conditions on S and q_1.

We conclude the introduction by describing the growth of a_n under natural conditions on S.

Remark 1.3. Suppose S is strongly aperiodic, mean zero, and finite second moments, with a nonsingular covariance matrix. Then, S is transient iff $d \geq 3$, and by the local central limit theorem, as $n \to \infty$,

$$\sum_{k=1}^{n} \mathrm{P}\{S_k = 0\} \sim \text{const} \times \begin{cases} n^{1/2} & \text{if } d = 1, \\ \log n & \text{if } d = 2. \end{cases} \tag{1.9}$$

See, for example (Spitzer (1976) [**P9** on p. 75]). Consequently,

$$a_n \sim \text{const} \times \begin{cases} n^{3/4} & \text{if } d = 1, \\ (n \log n)^{1/2} & \text{if } d = 2. \end{cases} \tag{1.10}$$

This agrees with the normalization of Kesten and Spitzer (1979) when $d = 1$, and Bolthausen (1989) when $d = 2$.

2. Preliminary Estimates

Consider the local times of S defined by

$$L_n^x := \sum_{i=1}^{n} \mathbf{1}_{\{S_i = x\}}. \tag{2.1}$$

A little thought shows that the random walk in random scenery can be represented compactly as

$$W_n = \sum_{x \in \mathbf{Z}^d} Z(x) L_n^x. \tag{2.2}$$

There is also a nice way to write the random Hamiltonian H_n in local-time terms. Consider the "level sets,"

$$\mathcal{L}_n^x := \{i \in \{1,\dots,n\} :\ S_i = x\}. \tag{2.3}$$

It is manifest that if $j \in \{2,\dots,n\}$, then $L_j^x > L_{j-1}^x$ if and only if $j \in \mathcal{L}_n^x$. Thus, we can write

$$\begin{aligned} H_n &= \frac{1}{2}\left(\sum_{x\in\mathbf{Z}^d}\left|\sum_{i=1}^n q_i \mathbf{1}_{\{S_i=x\}}\right|^2 - \sum_{i=1}^n q_i^2\right) \\ &= \sum_{x\in\mathbf{Z}^d} h_n^x, \end{aligned} \tag{2.4}$$

where

$$h_n^x := \frac{1}{2}\left(\left|\sum_{i\in\mathcal{L}_n^x} q_i\right|^2 - \sum_{i\in\mathcal{L}_n^x} q_i^2\right). \tag{2.5}$$

We denote by $\widehat{\mathrm{P}}$ the conditional measure, given the entire process S; $\widehat{\mathrm{E}}$ denotes the corresponding expectation operator. The following is borrowed from Chen (2008) [Lemma 2.1].

Lemma 2.1. *Choose and fix some integer $n \geq 1$. Then, $\{h_n^x\}_{x\in\mathbf{Z}^d}$ is a collection of independent random variables under $\widehat{\mathrm{P}}$, and*

$$\widehat{\mathrm{E}} h_n^x = 0 \quad \textit{and} \quad \widehat{\mathrm{E}}\left(|h_n^x|^2\right) = \frac{\sigma^2}{2} L_n^x\left(L_n^x - 1\right) \qquad \mathrm{P}\textit{-a.s.} \tag{2.6}$$

Moreover, there exists a nonrandom positive and finite constant $C = C(\sigma)$ such that for all $n \geq 1$ and $x \in \mathbf{Z}^d$,

$$\widehat{\mathrm{E}}\left(|h_n^x|^3\right) \leq C\mu_6 \left|L_n^x\left(L_n^x - 1\right)\right|^{3/2} \qquad \mathrm{P}\textit{-a.s.} \tag{2.7}$$

Next we develop some local-time computations.

Lemma 2.2. *For all $n \geq 1$,*

$$\sum_{x\in\mathbf{Z}^d} \mathrm{E} L_n^x = n \quad \textit{and} \quad \sum_{x\in\mathbf{Z}^d} \mathrm{E}\left(|L_n^x|^2\right) = n + 2\sum_{k=1}^{n-1}(n-k)\mathrm{P}\{S_k = 0\}. \tag{2.8}$$

Moreover, for all integers $k \geq 1$,

$$\sum_{x\in\mathbf{Z}^d} \mathrm{E}\left(|L_n^x|^k\right) \leq k!\, n \left|\sum_{j=0}^n \mathrm{P}\{S_j = 0\}\right|^{k-1}. \tag{2.9}$$

Proof. Since $\mathrm{E} L_n^x = \sum_{j=1}^n \mathrm{P}\{S_j = x\}$ and $\sum_{x\in\mathbf{Z}^d} \mathrm{P}\{S_j = x\} = 1$, we have $\sum_x \mathrm{E} L_n^x = n$. For the second-moment formula we write

$$\begin{aligned} \mathrm{E}\left(|L_n^x|^2\right) &= \sum_{1\leq i\leq n} \mathrm{P}\{S_i = x\} + 2\sum_{1\leq i<j\leq n}\sum \mathrm{P}\{S_i = S_j = x\} \\ &= \sum_{1\leq i\leq n} \mathrm{P}\{S_i = x\} + 2\sum_{1\leq i<j\leq n}\sum \mathrm{P}\{S_i = x\}\mathrm{P}\{S_{j-i} = 0\}. \end{aligned} \tag{2.10}$$

We can sum this expression over all $x \in \mathbf{Z}^d$ to find that

$$\sum_{x\in\mathbf{Z}^d} \mathrm{E}\left(|L_n^x|^2\right) = n + 2\sum_{1\le i<j\le n}\sum \mathrm{P}\{S_{j-i}=0\}. \tag{2.11}$$

This readily implies the second-moment formula. Similarly, we write

$$\begin{aligned}
&\mathrm{E}\left(|L_n^x|^k\right)\\
&\le k!\sum_{1\le i_1\le\cdots\le i_k\le n}\cdots\sum \mathrm{P}\{S_{i_1}=\cdots=S_{i_k}=x\}\\
&= k!\sum_{1\le i_1\le\cdots\le i_k\le n}\cdots\sum \mathrm{P}\{S_{i_1}=x\}\mathrm{P}\{S_{i_2-i_1}=0\}\cdots\mathrm{P}\{S_{i_k-i_{k-1}}=0\} \qquad (2.12)\\
&\le k!\sum_{i=1}^n \mathrm{P}\{S_i=x\}\cdot\left|\sum_{j=1}^n \mathrm{P}\{S_j=0\}\right|^{k-1}.
\end{aligned}$$

Add over all $x \in \mathbf{Z}^d$ to finish.

Our next lemma provides the first step in a classification of recurrence [versus transience] for random walks.

Lemma 2.3. *It is always the case that*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{x\in\mathbf{Z}^d}\mathrm{E}\left(|L_n^x|^2\right) = 1 + 2\sum_{k=1}^\infty \mathrm{P}\{S_k=0\}. \tag{2.13}$$

Proof. Thanks to Lemma 2.2, for all $n \ge 1$,

$$\frac{1}{n}\sum_{x\in\mathbf{Z}^d}\mathrm{E}\left(|L_n^x|^2\right) = 1 + 2\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\mathrm{P}\{S_k=0\}. \tag{2.14}$$

If S is transient, then the monotone convergence theorem ensures that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{x\in\mathbf{Z}^d}\mathrm{E}\left(|L_n^x|^2\right) = 1 + 2\sum_{k=1}^\infty \mathrm{P}\{S_k=0\}. \tag{2.15}$$

This proves the lemma in the transient case.

When S is recurrent, we note that (2.14) readily implies that for all integers $m \ge 2$,

$$\begin{aligned}
\liminf_{n\to\infty}\frac{1}{n}\sum_{x\in\mathbf{Z}^d}\mathrm{E}\left(|L_n^x|^2\right) &\ge 1 + 2\sum_{k=1}^{m-1}\left(1-\frac{k}{m}\right)\mathrm{P}\{S_k=0\}\\
&\ge 1 + \sum_{1\le k\le m/2}\mathrm{P}\{S_k=0\}.
\end{aligned} \tag{2.16}$$

Let $m \uparrow \infty$ to deduce the lemma.

Next we "remove the expectation" from the statement of Lemma 2.3.

Theorem 2.4. *As $n \to \infty$,*

$$\frac{1}{n}\sum_{x\in\mathbf{Z}^d}(L_n^x)^2 \to 1 + 2\sum_{k=1}^{\infty}\mathrm{P}\{S_k = 0\} \quad \textit{in probability.} \tag{2.17}$$

Remark 2.5. The quantity $I_n := \sum_{x\in\mathbf{Z}^d}(L_n^x)^2$ is the socalled *self-intersection local time* of the walk S. This terminology stems from the following elementary calculation: For all integers $n \geq 1$,

$$I_n = \sum_{1\leq i,j\leq n}\sum \mathbf{1}_{\{S_j=S_i\}}. \tag{2.18}$$

Consequently, Theorem 2.4 implies that a random walk S on $\mathbf{Z}^d$ is recurrent if and only if its self-intersection local time satisfies $I_n/n \to \infty$ in probability.

Remark 2.6. Nadine Guillotin–Plantard has kindly pointed out to us that the mode of convergence in Theorem 2.4 can be strengthened to almost-sure convergence. This requires a direct subadditivity argument (Guillotin–Plantard (2004)). It follows also from the estimates that follow, together with a classical blocking argument, which we skip.

Proof. First we study the case that $\{S_i\}_{i=0}^{\infty}$ is transient.

Define

$$Q_n := \sum_{1\leq i<j\leq n}\sum \mathbf{1}_{\{S_i=S_j\}}. \tag{2.19}$$

Then it is not too difficult to see that

$$\sum_{x\in\mathbf{Z}^d}(L_n^x)^2 = 2Q_n + n \qquad \text{for all } n \geq 1. \tag{2.20}$$

This follows immediately from (2.18), for example. Therefore, it suffices to prove that, under the assumption of transience,

$$\frac{Q_k}{k} \to \sum_{j=1}^{\infty}\mathrm{P}\{S_j = 0\} \quad \text{in probability as } k \to \infty. \tag{2.21}$$

Lemma 2.3 and (2.20) together imply that

$$\lim_{k\to\infty}\frac{\mathrm{E}Q_k}{k} = \sum_{j=1}^{\infty}\mathrm{P}\{S_j = 0\}. \tag{2.22}$$

Hence, it suffices to prove that $\operatorname{Var} Q_n = o(n^2)$ as $n \to \infty$. In some cases, this can be done by making an explicit [though hard] estimate for $\operatorname{Var} Q_n$; see, for instance, (Chen (2008)[Lemma 5.1]), and also the technique employed in the proof of Lemma 2.4 of Bolthausen (1989). Here, we opt for a more general approach that is simpler, though it is a little more circuitous. Namely, in rough terms, we write Q_n as $Q_n^{(1)} + Q_n^{(2)}$, where $\mathrm{E}Q_n^{(1)} = o(n)$, and $\operatorname{Var} Q_n^{(2)} = o(n^2)$. Moreover, we will soon see that $Q_n^{(1)}, Q_n^{(2)} \geq 0$, and this suffices to complete the proof.

For all $m := m_n \in \{1, \ldots, n-1\}$ we write

$$Q_n = Q_n^{1,m} + Q_n^{2,m}, \tag{2.23}$$

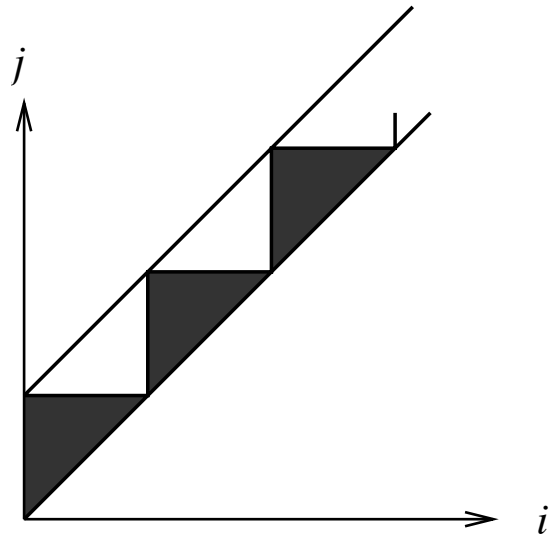


FIG 1. *A decomposition of* Q_n

where

$$Q_n^{1,m} := \sum_{\substack{1\le i<j\le n:\\ j\ge i+m}}\sum \mathbf{1}_{\{S_i=S_j\}} \quad \text{and} \quad Q_n^{2,m} := \sum_{\substack{1\le i<j\le n:\\ j<i+m}}\sum \mathbf{1}_{\{S_i=S_j\}}. \tag{2.24}$$

Because $n > m$, we have

$$\mathrm{E}Q_n^{1,m} \le n\sum_{k=m}^{\infty} \mathrm{P}\{S_k = 0\}. \tag{2.25}$$

We estimate the variance of $Q_n^{2,m}$ next. We do this by first making an observation.

Throughout the remainder of this proof, define for all subsets Γ of $\mathbf{N}^2$,

$$\Upsilon(\Gamma) := \sum_{(i,j)\in\Gamma}\sum \mathbf{1}_{\{S_i=S_j\}}. \tag{2.26}$$

Supppose $\Gamma_1, \Gamma_2, \ldots, \Gamma_\nu$ are finite disjoint sets in $\mathbf{N}^2$, with common cardinality, and the added property that whenever $1 \le a < b \le \nu$, we have $\Gamma_a < \Gamma_b$ in the sense that $i < k$ and $j < l$ for all $(i\,,j) \in \Gamma_a$ and $(k\,,l) \in \Gamma_b$. Then, it follows that

$$\{\Upsilon(\Gamma_\nu)\}_{\mu=1}^{\nu} \quad \text{is an i.i.d. sequence.} \tag{2.27}$$

For all integers $p \ge 0$ define

$$\begin{aligned} B_p^m &:= \left\{(i\,,j) \in \mathbf{N}^2 : (p-1)m < i < j \le pm\right\},\\ W_p^m &:= \left\{(i\,,j) \in \mathbf{N}^2 : (p-1)m < i \le pm < j \le (p+1)m\right\}. \end{aligned} \tag{2.28}$$

In Figure 1, $\{B_p^m\}_{p=1}^\infty$ denotes the collection black and $\{W_p^m\}_{p=1}^\infty$ the white triangles that are inside the slanted strip.

We may write

$$Q_{(n-1)m}^{2,m} = \sum_{p=1}^{n-1}\Upsilon(B_p^m) + \sum_{p=1}^{n-1}\Upsilon(W_p^m). \tag{2.29}$$

Consequently,

$$\operatorname{Var} Q^{2,m}_{(n-1)m} \le 2\operatorname{Var}\sum_{p=1}^{n-1}\Upsilon(B^m_p) + 2\operatorname{Var}\sum_{p=1}^{n-1}\Upsilon(W^m_p). \tag{2.30}$$

If $1 \le a < b \le m-1$, then $B^m_a < B^m_b$ and $W^m_a < W^m_b$. Consequently, (2.27) implies that

$$\operatorname{Var} Q^{2,m}_{(n-1)m} \le 2(n-1)\left[\operatorname{Var}\Upsilon(B^m_1) + \operatorname{Var}\Upsilon(W^m_1)\right]. \tag{2.31}$$

Because $\Upsilon(B^m_1)$ and $\Upsilon(W^m_1)$ are individually sums of not more than $\binom{m}{2}$-many ones,

$$\operatorname{Var} Q^{2,m}_{(n-1)m} \le 2(n-1)m^2. \tag{2.32}$$

Let $Q^{(1)}_n := Q^{1,m}_n$ and $Q^{(2)}_n := Q^{2,m}_n$, where $m = m_n := n^{1/4}$ [say]. Then, $Q_n = Q^{(1)}_n + Q^{(2)}_n$, and (2.25) and (2.32) together imply that $\mathrm{E}Q^{(1)}_{(n-1)m} = o((n-1)m)$. Moreover, $\operatorname{Var} Q^{(2)}_{(n-1)m} = o((nm)^2)$. This gives us the desired decomposition of $Q_{(n-1)m}$. Now we complete the proof: Thanks to (2.22),

$$\mathrm{E}Q^{(2)}_{(n-1)m} \sim nm\cdot\sum_{j=1}^{\infty}\mathrm{P}\{S_j = 0\} \quad \text{as } n\to\infty. \tag{2.33}$$

Therefore, the variance of $Q^{(2)}_{(n-1)m}$ is little-o of the square of its mean. This and the Chebyshev inequality together imply that $Q^{(2)}_{(n-1)m}/(nm)$ converges in probability to $\sum_{j=1}^{\infty}\mathrm{P}\{S_j = 0\}$. On the other hand, we know also that $Q^{(1)}_{(n-1)m}/(nm)$ converges to zero in $L^1(\mathrm{P})$ and hence in probability. Consequently, we can change variables and note that as $n\to\infty$,

$$\frac{Q_{nm}}{nm} \to \sum_{j=1}^{\infty}\mathrm{P}\{S_j = 0\} \quad \text{in probability.} \tag{2.34}$$

If k is between $(n-1)m$ and nm, then

$$\frac{Q_{(n-1)m}}{nm} \le \frac{Q_k}{k} \le \frac{Q_{nm}}{(n-1)m}. \tag{2.35}$$

This proves (2.21), and hence the theorem, in the transient case.

In order to derive the recurrent case, it suffices to prove that $Q_n/n \to \infty$ in probability as $n\to\infty$.

Let us choose and hold an integer $m \ge 1$—so that it does *not* grow with n—and observe that $Q_n \ge Q^{2,m}_n$ as long as n is sufficiently large. Evidently,

$$\begin{aligned}\mathrm{E}Q^{2,m}_n &= \sum_{\substack{1\le i<j\le n:\\ j<i+m}}\sum\mathrm{P}\{S_j = S_i\}\\ &= (n-1)\sum_{k=1}^{m-1}\mathrm{P}\{S_k = 0\}.\end{aligned} \tag{2.36}$$

We may also observe that (2.32) continues to hold in the present recurrent setting. Together with the Chebyshev inequality, these computations imply that as $n \to \infty$,

$$\frac{Q^{2,m}_{n(m-1)}}{n} \to \sum_{k=1}^{m-1} \mathrm{P}\{S_k = 0\} \quad \text{in probability.} \tag{2.37}$$

Because $Q_{n(m-1)} \geq Q^{2,m}_{n(m-1)}$, the preceding implies that

$$\lim_{n\to\infty} \mathrm{P}\left\{ \frac{Q_{n(m-1)}}{n} \geq \frac{1}{2} \sum_{k=1}^{m} \mathrm{P}\{S_k = 0\} \right\} = 1. \tag{2.38}$$

A monotonicity argument shows that $Q_{n(m-1)}$ can be replaced by Q_n without altering the end-result; see (2.35). By recurrence, if $\lambda > 0$ is any predescribed positive number, then we can choose [and fix] our integer m such that $\sum_{k=1}^{m} \mathrm{P}\{S_k = 0\} \geq 2\lambda$. This proves that $\lim_{n\to\infty} \mathrm{P}\{Q_n/n \geq \lambda\} = 1$ for all $\lambda > 0$, and hence follows the theorem in the recurrent case.

3. Proofs of the Main Results

Now we introduce a sequence $\{\xi_x\}_{x\in\mathbf{Z}^d}$ of random variables, independent [under P] of $\{q_i\}_{i=1}^{\infty}$ and the random walk $\{S_i\}_{i=0}^{\infty}$, such that

$$\mathrm{E}\xi_0 = 0, \quad \mathrm{E}\left(\xi_0^2\right) = \sigma^2, \quad \text{and} \quad \widehat{\mu}_3 := \mathrm{E}\left(|\xi_0|^3\right) < \infty. \tag{3.1}$$

Define

$$\widehat{h}^x_n := \left| \frac{L^x_n \left(L^x_n - 1\right)}{2} \right|^{1/2} \xi_x \qquad \text{for all } n \geq 1 \text{ and } x \in \mathbf{Z}^d. \tag{3.2}$$

Evidently, $\{\widehat{h}^x_n\}_{x\in\mathbf{Z}^d}$ is a sequence of [conditionally] independent random variables, under $\widehat{\mathrm{P}}$, and has the same [conditional] mean and variance as $\{h^x_n\}_{x\in\mathbf{Z}^d}$.

Lemma 3.1. *There exists a positive and finite constant $C_* = C_*(\sigma)$ such that if $f : \mathbf{R}^d \to \mathbf{R}$ is three time continuously differentiable, then for all $n \geq 1$,*

$$\left| \mathrm{E}f\left(\sum_{x\in\mathbf{Z}^d} \widehat{h}^x_n \right) - \mathrm{E}f(H_n) \right| \leq C_* M_f (\widehat{\mu}_3 + \mu_6) n \left| \sum_{j=0}^{n} \mathrm{P}\{S_j = 0\} \right|^2, \tag{3.3}$$

with $M_f := \sup_{x\in\mathbf{R}^d} |f'''(x)|$.

Proof. Temporarily choose and fix some $y \in \mathbf{Z}^d$, and notice that

$$\begin{aligned} &f(H_n) \\ &= f\left(\sum_{x\in\mathbf{Z}^d\setminus\{y\}} h^x_n \right) + f'\left(\sum_{x\in\mathbf{Z}^d\setminus\{y\}} h^x_n \right) h^y_n + \frac{1}{2} f''\left(\sum_{x\in\mathbf{Z}^d\setminus\{y\}} h^x_n \right) |h^y_n|^2 \\ &\quad + R_n, \end{aligned} \tag{3.4}$$

where $|R_n| \leq \frac{1}{6}\|f'''\|_\infty |h^y_n|^3$. It follows from this and Lemma 2.1 that

$$\begin{aligned} &\widehat{\mathrm{E}} f(H_n) \\ &= \widehat{\mathrm{E}} f\left(\sum_{x\in\mathbf{Z}^d\setminus\{y\}} h^x_n \right) + \frac{\sigma^2}{2} L^y_n \left(L^y_n - 1\right) \widehat{\mathrm{E}} f''\left(\sum_{x\in\mathbf{Z}^d\setminus\{y\}} h^x_n \right) + R^{(1)}_n, \end{aligned} \tag{3.5}$$

where

$$\begin{aligned}\left|R_n^{(1)}\right| &\le \frac{CM_f\mu_6}{12}\left|L_n^x\left(L_n^x-1\right)\right|^{3/2} \qquad \text{P-a.s.}\\ &\le \frac{CM_f\mu_6}{12}\left|L_n^y\right|^3.\end{aligned} \tag{3.6}$$

We proceed as in (3.4) and write

$$\begin{aligned}&f\left(\widehat{h}_n^y+\sum_{x\in\mathbf{Z}^d\setminus\{y\}}h_n^x\right)\\ &=f\left(\sum_{x\in\mathbf{Z}^d\setminus\{y\}}h_n^x\right)+f'\left(\sum_{x\in\mathbf{Z}^d\setminus\{y\}}h_n^x\right)\widehat{h}_n^y+\frac{1}{2}f''\left(\sum_{x\in\mathbf{Z}^d\setminus\{y\}}h_n^x\right)\left|\widehat{h}_n^y\right|^2\\ &\quad+\widehat{R}_n,\end{aligned} \tag{3.7}$$

where $|\widehat{R}_n|\le\frac{1}{6}M_f|\widehat{h}_n^y|^3\le\frac{1}{12\sqrt{2}}M_f|L_n^y|^3\,|\xi_y|^3$. It follows from this and Lemma 2.1 that

$$\begin{aligned}&\widehat{\mathrm{E}}f\left(\widehat{h}_n^y+\sum_{x\in\mathbf{Z}^d\setminus\{y\}}h_n^x\right)\\ &=\widehat{\mathrm{E}}f\left(\sum_{x\in\mathbf{Z}^d\setminus\{y\}}h_n^x\right)+\frac{\sigma^2}{2}L_n^y\left(L_n^y-1\right)\widehat{\mathrm{E}}f''\left(\sum_{x\in\mathbf{Z}^d\setminus\{y\}}h_n^x\right)+R_n^{(2)},\end{aligned} \tag{3.8}$$

where $|R_n^{(2)}|\le\frac{1}{12\sqrt{2}}\widehat{\mu}_3M_f|L_n^y|^3$. Define $C_*:=(C+1)/2$ to deduce from the preceding and (3.5) that P-a.s.,

$$\left|\widehat{\mathrm{E}}f\left(\widehat{h}_n^y+\sum_{x\in\mathbf{Z}^d\setminus\{y\}}h_n^x\right)-\widehat{\mathrm{E}}f\left(\sum_{x\in\mathbf{Z}^d}h_n^x\right)\right|\le\frac{A}{6}|L_n^y|^3, \tag{3.9}$$

where $A:=C_*M_f(\widehat{\mu}_3+\mu_6)$. The preceding computes the effect of replacing the contribution of h_n^x to H_n but the independent quantity $\widehat{h}_n^y$, for each fixed y, and uses only the fact that the $\hat{h}$'s are a conditionally independent sequence with the same means and variances as their corresponding h's. Therefore, if we choose and fix another point $y\in\mathbf{Z}^d\setminus\{y\}$, then the very same constant A satisfies the following: Almost surely [P],

$$\left|\widehat{\mathrm{E}}f\left(\widehat{h}_n^z+\widehat{h}_n^y+\sum_{x\in\mathbf{Z}^d\setminus\{y,z\}}h_n^x\right)-\widehat{\mathrm{E}}f\left(\widehat{h}_n^y+\sum_{x\in\mathbf{Z}^d\setminus\{y\}}h_n^x\right)\right|\le\frac{A}{6}|L_n^z|^3. \tag{3.10}$$

And hence, the triangle inequality yields the following: P-a.s.,

$$\begin{aligned}&\left|\widehat{\mathrm{E}}f\left(\widehat{h}_n^z+\widehat{h}_n^y+\sum_{x\in\mathbf{Z}^d\setminus\{y,z\}}h_n^x\right)-\widehat{\mathrm{E}}f\left(\sum_{x\in\mathbf{Z}^d}h_n^x\right)\right|\\ &\qquad\le\frac{A}{6}\left(|L_n^y|^3+|L_n^z|^3\right).\end{aligned} \tag{3.11}$$

Because $\sum_{x\in\mathbf{Z}^d} h_n^x = H_n$, it is now possible to see how we can iterate the previous inequality to find that P-a.s.,

$$\left|\widehat{\mathrm{E}} f\left(\sum_{x\in\mathbf{Z}^d} \widehat{h}_n^x\right) - \widehat{\mathrm{E}} f(H_n)\right| \le \frac{A}{6}\sum_{y\in\mathbf{Z}^d} |L_n^y|^3. \tag{3.12}$$

We take expectations and appeal to Lemma 2.2 to finish.

Next, we prove Theorem 1.1.

Proof of Theorem 1.1. We choose, in Lemma 3.1, the collection $\{\xi_x\}_{x\in\mathbf{Z}^d}$ to be i.i.d. mean-zero normals with variance σ^2. Then, we apply Lemma 3.1 with $f(x) := g(x/n^{1/2})$ for a smooth bounded function g with bounded derivatives. This yields,

$$\left|\mathrm{E} g(H_n/n^{1/2}) - \mathrm{E} g\left(\frac{1}{n^{1/2}}\sum_{x\in\mathbf{Z}^d}\widehat{h}_n^x\right)\right| \le \frac{\text{const}}{n^{1/2}}. \tag{3.13}$$

In this way,

$$\begin{aligned}\sum_{x\in\mathbf{Z}^d}\widehat{h}_n^x &\overset{\mathcal{D}}{=} \frac{\sigma}{\sqrt{2}}\left|\sum_{x\in\mathbf{Z}^d} L_n^x\,(L_n^x-1)\right|^{1/2} N(0\,,1) \qquad \text{under } \widehat{\mathrm{P}}\\ &= \frac{\sigma}{\sqrt{2}}\left|-n+\sum_{x\in\mathbf{Z}^d}(L_n^x)^2\right|^{1/2} N(0\,,1),\end{aligned} \tag{3.14}$$

where $\overset{\mathcal{D}}{=}$ denotes equality in distribution, and $N(0\,,1)$ is a standard normal random variable under $\widehat{\mathrm{P}}$ as well as P. Therefore, in accord with Theorem 2.4,

$$\begin{aligned}\frac{1}{n^{1/2}}\sum_{x\in\mathbf{Z}^d}\widehat{h}_n^x &\overset{\mathcal{D}}{=} \frac{\sigma}{\sqrt{2}}\left|-1+\frac{1}{n}\sum_{x\in\mathbf{Z}^d}(L_n^x)^2\right|^{1/2} N(0\,,1)\\ &= o_{\widehat{\mathrm{P}}}(1) + \gamma^{1/2}\cdot N(0\,,\sigma^2),\end{aligned} \tag{3.15}$$

where $o_{\widehat{\mathrm{P}}}(1)$ is a term that converges to zero as $n\to\infty$ in $\widehat{\mathrm{P}}$-probability a.s. [P]. Equation (3.13) then completes the proof in the transient case.

Theorem 1.2 relies on the following "coupled moderate deviation" result.

Proposition 3.2. *Suppose that S is recurrent. Consider a sequence $\{\epsilon_j\}_{j=1}^\infty$ of nonnegative numbers that satisfy the following:*

$$\lim_{n\to\infty}\epsilon_n^3 n\left|\sum_{k=1}^n \mathrm{P}\{S_k=0\}\right|^2 = 0. \tag{3.16}$$

Then for all compactly supported functions $f:\mathbf{R}^d\to\mathbf{R}$ that are infinitely differentiable,

$$\lim_{n\to\infty}\left|\mathrm{E}\left[f\left(\epsilon_n W_n\right)\right] - \mathrm{E}\left[f\left(\epsilon_n H_n\right)\right]\right| = 0, \tag{3.17}$$

Proof. We apply Lemma 3.1 with the ξ_x's having the same common distribution as q_1, and with $f(x) := g(\epsilon_n x)$ for a smooth and bounded function g with bounded derivatives. This yields,

$$\begin{aligned}\left|\mathrm{E}\left[g\left(\epsilon_n\sum_{x\in\mathbf{Z}^d}\left|L_n^x\left(L_n^x-1\right)\right|^{1/2}Z(x)\right)\right]-\mathrm{E}\left[g\left(\epsilon_n H_n\right)\right]\right| \\ \le 2C_*M_g\mu_6 n\epsilon_n^3\left|\sum_{k=0}^n\mathrm{P}\{S_k=0\}\right|^2 \\ = o(1),\end{aligned} \tag{3.18}$$

owing to Lemma (3.4).

According to Taylor's formula,

$$\begin{aligned}&g\left(\epsilon_n\sum_{x\in\mathbf{Z}^d}\left|L_n^x\left(L_n^x-1\right)\right|^{1/2}Z(x)\right)\\ &\quad=g\left(\epsilon_n\sum_{x\in\mathbf{Z}^d}Z(x)L_n^x\right)+\epsilon_n\sum_{x\in\mathbf{Z}^d}\left(\left|L_n^x\left(L_n^x-1\right)\right|^{1/2}-L_n^x\right)Z(x)\cdot R,\end{aligned} \tag{3.19}$$

where $|R|\le\sup_{x\in\mathbf{R}^d}|g'(x)|$. Thanks to (2.2), we can write the preceding as follows:

$$\begin{aligned}&g\left(\epsilon_n\sum_{x\in\mathbf{Z}^d}\left|L_n^x\left(L_n^x-1\right)\right|^{1/2}Z(x)\right)-g\left(\epsilon_n W_n\right)\\ &\qquad=\epsilon_n\sum_{x\in\mathbf{Z}^d}\left(\left|L_n^x\left(L_n^x-1\right)\right|^{1/2}-L_n^x\right)Z(x)\cdot R.\end{aligned} \tag{3.20}$$

Consequently, P-almost surely,

$$\begin{aligned}&\left|\widehat{\mathrm{E}}\left[g\left(\epsilon_n\sum_{x\in\mathbf{Z}^d}\left|L_n^x\left(L_n^x-1\right)\right|^{1/2}Z(x)\right)\right]-\widehat{\mathrm{E}}\left[g\left(\epsilon_n W_n\right)\right]\right|\\ &\le\sup_{x\in\mathbf{R}^d}|g'(x)|\sigma\cdot\epsilon_n\left\{\widehat{\mathrm{E}}\left(\sum_{x\in\mathbf{Z}^d}\left(\left|L_n^x\left(L_n^x-1\right)\right|^{1/2}-L_n^x\right)^2\right)\right\}^{1/2}.\end{aligned} \tag{3.21}$$

We apply the elementary inequality $(a^{1/2}-b^{1/2})^2\le|a-b|$—valid for all $a,b\ge0$—to deduce that P-almost surely,

$$\begin{aligned}&\left|\widehat{\mathrm{E}}\left[g\left(\epsilon_n\sum_{x\in\mathbf{Z}^d}\left|L_n^x\left(L_n^x-1\right)\right|^{1/2}Z(x)\right)\right]-\widehat{\mathrm{E}}\left[g\left(\epsilon_n W_n\right)\right]\right|\\ &\le\sup_{x\in\mathbf{R}^d}|g'(x)|\sigma\cdot\epsilon_n\left\{\widehat{\mathrm{E}}\left(\sum_{x\in\mathbf{Z}^d}L_n^x\right)\right\}^{1/2}\\ &=\sup_{x\in\mathbf{R}^d}|g'(x)|\sigma\cdot\epsilon_n n^{1/2}.\end{aligned} \tag{3.22}$$

We take E-expectations and apply Lemma (3.4) to deduce from this and (3.18) that

$$|\mathrm{E}\left[g\left(\epsilon_n W_n\right)\right] - \mathrm{E}\left[g\left(\epsilon_n H_n\right)\right]| = o(1). \tag{3.23}$$

This completes the proof.

Our proof of Theorem 1.2 hinges on two more basic lemmas. The first is an elementary lemma from integration theory.

Lemma 3.3. *Suppose $X := \{X_n\}_{n=1}^{\infty}$ and $Y := \{Y_n\}_{n=1}^{\infty}$ are $\mathbf{R}^d$-valued random variables such that: (i) X and Y each form a tight sequence; and (ii) for all bounded infinitely-differentiable functions $g : \mathbf{R}^d \to \mathbf{R}$,*

$$\lim_{n\to\infty} |\mathrm{E}g(X_n) - \mathrm{E}g(Y_n)| = 0. \tag{3.24}$$

Then, the preceding holds for all bounded continuous functions $g : \mathbf{R}^d \to \mathbf{R}$.

Proof. The proof uses standard arguments, but we repeat it for the sake of completeness.

Let $K_m := [-m\,, m]^d$, where m takes values in $\mathbf{N}$. Given a bounded continuous function $g : \mathbf{R}^d \to \mathbf{R}$, we can find a bounded infinitely-differentiable function $h_m : \mathbf{R}^d \to \mathbf{R}$ such that $|h_m - g| < 1/m$ on K_m. It follows that

$$\begin{aligned} |\mathrm{E}g(X_n) - \mathrm{E}g(Y_n)| &\le 2/m + |\mathrm{E}h_m(X_n) - \mathrm{E}h_m(Y_n)| \\ &\quad + 2\sup_{x\in\mathbf{R}^d} |g(x)|\big(\mathrm{P}\{X_n \notin K_m\} + \mathrm{P}\{Y_n \notin K_m\}\big). \end{aligned} \tag{3.25}$$

Consequently,

$$\begin{aligned} &\limsup_{n\to\infty} |\mathrm{E}g(X_n) - \mathrm{E}g(Y_n)| \\ &\qquad \le 2/m + 2\sup_{x\in\mathbf{R}^d} |g(x)| \sup_{j\ge 1}\left(\mathrm{P}\{X_j \notin K_m\} + \mathrm{P}\{Y_j \notin K_m\}\right). \end{aligned} \tag{3.26}$$

Let m diverge and appeal to tightness to conclude that the left-had side vanishes.

The final ingredient in the proof of Theorem 1.1 is the following harmonic-analytic result.

Lemma 3.4. *If $\epsilon_n := 1/a_n$, then (3.16) holds.*

Proof. Let ϕ denote the characteristic function of S_1. Our immediate goal is to prove that $|\phi(t)| < 1$ for all but a countable number of $t \in \mathbf{R}^d$. We present an argument, due to Firas Rassoul-Agha, that is simpler and more elegant than our original proof.

Suppose S_1' is an independent copy of S_1, and note that whenever $t \in \mathbf{R}^d$ is such that $|\phi(t)| = 1$, $D := \exp\{it\cdot(S_1 - S_1')\}$ has expectation one. Consequently, $\mathrm{E}(|D-1|^2) = \mathrm{E}(|D|^2) - 1 = 0$, whence $D = 1$ a.s. Because S_1 is assumed to have at least two possible values, $S_1 \ne S_1'$ with positive probability, and this proves that $t \in 2\pi\mathbf{Z}^d$. It follows readily from this that

$$\left\{t \in \mathbf{R}^d :\ |\phi(t)| = 1\right\} = 2\pi\mathbf{Z}^d, \tag{3.27}$$

and in particular, $|\phi(t)| < 1$ for almost all $t \in \mathbf{R}^d$.

By the inversion theorem (Spitzer (1976)[**P3**(b), p. 57]), for all $n \geq 0$,

$$\mathrm{P}\{S_n = 0\} = \frac{1}{(2\pi)^d} \int_{(-\pi,\pi)^d} \{\phi(t)\}^n \, dt. \tag{3.28}$$

This and the dominated convergence theorem together tell us that $\mathrm{P}\{S_n = 0\} = o(1)$ as $n \to \infty$, whence it follows that

$$\sum_{k=1}^{n} \mathrm{P}\{S_k = 0\} = o(n) \quad \text{as } n \to \infty. \tag{3.29}$$

For our particular choice of ϵ_n we find that

$$\epsilon_n^3 n \left| \sum_{k=1}^{n} \mathrm{P}\{S_k = 0\} \right|^2 = \left(\frac{1}{n} \sum_{k=1}^{n} \mathrm{P}\{S_k = 0\} \right)^{1/2}, \tag{3.30}$$

and this quantity vanishes as $n \to \infty$ by (3.29). This proves the lemma.

Proof of Theorem 1.2. Let $\epsilon_n := 1/a_n$. In light of Proposition 3.2, and Lemmas 3.3 and 3.4, it suffices to prove that the sequences $n \mapsto \epsilon_n W_n$ and $n \mapsto \epsilon_n H_n$ are tight.

Lemma 2.2, (2.2), and recurrence together imply that for all n large,

$$\begin{aligned} \mathrm{E}\left(|\epsilon_n W_n|^2 \right) &= \sigma^2 \epsilon_n^2 \sum_{x \in \mathbf{Z}^d} \mathrm{E}\left(|L_n^x|^2 \right) \\ &\leq \text{const} \cdot \epsilon_n^2 n \sum_{k=1}^{n} \mathrm{P}\{S_k = 0\} \\ &= \text{const}. \end{aligned} \tag{3.31}$$

Thus, $n \mapsto \epsilon_n W_n$ is bounded in $L^2(\mathrm{P})$, and hence is tight.

We conclude the proof by verifying that $n \mapsto \epsilon_n H_n$ is tight. Thanks to (2.4) and recurrence, for all n large,

$$\begin{aligned} \mathrm{E}\left(|\epsilon_n H_n|^2 \right) &\leq \text{const} \cdot \epsilon_n^2 \mathrm{E} \sum_{x \in \mathbf{Z}^d} (L_n^x)^2 \\ &\leq \text{const} \cdot \epsilon_n^2 n \sum_{k=1}^{n} \mathrm{P}\{S_k = 0\} \\ &= \text{const}. \end{aligned} \tag{3.32}$$

Confer with Lemma 2.2 for the penultimate line. Thus, $n \mapsto \epsilon_n H_n$ is bounded in $L^2(\mathrm{P})$ and hence is tight, as was announced.

Acknowledgement.

We wish to thank Siegfried Hörmann, Richard Nickl, Jon Peterson, and Firas Rassoul-Agha for many enjoyable discussions, particularly on the first portion of Lemma 3.4. We are grateful to Nadine Guillotin–Plantard for her valuable suggestions and typographical corrections, as well as providing us with references to her paper (Guillotine–Plantard (2004)). Special thanks are extended to Firas Rassoul-Agha for providing us with his elegant argument that replaced our clumsier proof of the first part of Lemma 3.4. Finally, we thank two anonymous referees who made valuable suggestions and pointed out misprints.

References

[1] Bolthausen, E. (1989). A central limit theorem for two-dimensional random walks in random sceneries. *Ann. Probab.*, **17.1**, 108–115.

[2] Buffet, E. and Pulé, J. V. (1997). A model of continuous polymers with random charges. *J. Math. Phys.* **38**, 5143–5152.

[3] Chen, X. (2008). Limit laws for the energy of a charged polymer. *Annales de l'Institut Henri Poincaré: Probab. et Statist.*, **44**, 638–672.

[4] Chung, K. L. and Fuchs, W. H. J. (1951). On the distribution of values of sums of random variables. *Mem. Amer. Math. Soc.*, **6**.

[5] Derrida, B., Griffiths, R. B. and Higgs, R. G. (1992). A model of directed walks with random interactions. *Europhys. Lett.*, **18**, 361–366.

[6] Derrida, B. and Higgs, R. G. (1994). Low-temperature properties of directed walks with random self interactions. *J. Phys. A*, **27**, 5485–5493.

[7] Garel, T. and Orland, H. (1988). Mean-field model for protein folding. *Europhys. Lett.*, **6**, 307–310.

[8] Guillotin–Plantard, N. (2004). Sur la convergence faible des systèmes dynamiques échantillonnés. *Annales de l'Institut Fourier*, **54.1**, 255–278.

[9] Kantor, Y. and Kardar, M. (1991). Polymers with self-interactions. *Europhys. Lett.*, **14**, 421–426.

[10] Kesten, H. and Spitzer, F. (1979). A limit theorem related to a new class of self-similar processes. *Z. Wahrsch. Verw. Gebiete*, **50.1**, 5–25.

[11] Liapounov, A. M. (1900). Sur une proposition de la théorie des probabilités. *Bull. de'l Acad. Impériale des Sci. St. Petérsbourg* **13.4**, 359–386.

[12] Lindeberg, W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechtnung. *Math. Z.*, **15**, 211–225.

[13] Martìnez, S. and Petritis, D. (1996). Thermodynamics of a Brownian bridge polymer model in a random environment. *J. Phys. A*, **29**, 1267–1279.

[14] Kesten, H. and Spitzer, F. (1978). A limit theorem related to a new class of self-similar processes. *Z. Wahrsch. Verw. Gebiete*, **50.1**, 5–25.

[15] Obukhov, S. P. (1986). Configurational statistics of a disordered polymer chain. *J. Phys. A*, **19**, 3655–3664.

[16] Pólya, G. (1921). Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt in Straßennetz. *Math. Ann.*, **84**, 149–160.

[17] Spitzer, F. (1976). *Principles of Random Walks*, second edition. Springer–Verlag, New York–Heidelberg.

[18] Wittmer, J., Johner, A. and Joanny, J. F. (1993). Random and alternating polyampholytes. *Europhys. Lett.*, **24**, 263–268.

IMS Collections
Vol. 0 (2009) 252–265
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000011

Recovery of Distributions via Moments

Robert M. Mnatsakanov[1,*] **and Artak S. Hakobyan**[2,*]

West Virginia University

Abstract: The problem of recovering a cumulative distribution function (cdf) and corresponding density function from its moments is studied. This problem is a special case of the classical moment problem. The results obtained within the moment problem can be applied in many indirect models, e.g., those based on convolutions, mixtures, multiplicative censoring, and right-censoring, where the moments of unobserved distribution of actual interest can be easily estimated from the transformed moments of the observed distributions. Nonparametric estimation of a quantile function via moments of a target distribution represents another very interesting area where the moment problem arises. In all such models one can apply the present results to recover a function via its moments. In this article some properties of the proposed constructions are derived. The uniform rates of convergence of the approximation of cdf, its density function, quantile and quantile density function are obtained as well.

Contents

1 Introduction 252
2 Notation and Assumptions 254
3 Asymptotic Properties of $F_{\alpha,\nu}$ and $f_{\alpha,\nu}$ 255
4 Some Applications and Examples 258
Acknowledgments 264
References 264

1. Introduction

The probabilistic Stielties moment problem can be described as follows: let a sequence $\nu = \{\mu_j, j = 0, 1, \dots\}$ of real numbers be given. Find a probability distribution on the non-negative real line $\mathbb{R}_+ = [0, \infty)$, such that $\mu_j = \int t^j dF(t)$ for $j \in \mathbb{N} = \{0, 1, \dots\}$. The classical Stieltjes moment problem was introduced first by Stieltjes [21]. When the support of the distribution F is compact, say, $supp\{F\} = [0, T]$ with $T < \infty$, then the corresponding problem is known as a Hausdorff moment problem.

Consider two important questions related to the Stieltjes (or Hausdorff) moment problem:

(*i*) If the distribution F exists, is it uniquely determined by the moments $\{\mu_j\}$?

(*ii*) How is this uniquely defined distribution F reconstructed?

[1]Department of Statistics, West Virginia University, Morgantown, WV 26506, email: rmnatsak@stat.wvu.edu

*Work supported by the National Institute for Occupational Safety and Health, contract no. 212-2005-M-12857.

[2]Department of Industrial and Management Systems Engineering, West Virginia University, Morgantown, WV 26506, email: artakhak@yahoo.com

AMS 2000 subject classifications: Primary 62G05; secondary 62G20

Keywords and phrases: Probabilistic moment problem, M-determinate distribution, Moment-recovered distribution, Uniform rate of convergence

If there is a positive answer to question (*i*) we say that a distribution F is moment-determinate (M-determinate), otherwise it is M-indeterminate.

In this paper we mainly address the question of recovering the M-determinate distribution (density and quantile functions) via its moments in the Hausdorff moment problem, i.e., we study question (*ii*). Another question we focus on here is the estimation of an unknown distribution and its quantile function, given the estimated moments of the target distribution.

It is known from the probabilistic moment problem that under suitable conditions an M-determinate distribution is uniquely defined by its moments. There are many articles that investigated the conditions (for example, the Carlemann's and the Krein's conditions), under which the distributions are either M-determinate or M-indeterminate. See, e.g., Akhiezer [2], Feller [6], Lin [10-11], and Stoyanov [22-24] among others. However, there are very few works dealing with the reconstruction of distributions via their moments. Several inversion formulas were obtained by inverting the moment generating function and Laplace transform (Shohat and Tamarkin [20], Widder [27], Feller [6], Chauveau *et al.* [4], and Tagliani and Velasquez [25]). These methods are too restrictive, since there are many distributions for which the moment generating function does not exist even though all the moments are finite.

The reconstruction of an M-determinate cdf by means of mixtures having the same assigned moments as the target distribution have been proposed in Lindsay *et al.* [12]. Note that this procedure requires calculations of high-order Hankel determinants, and due to ill-conditioning of the Hankel matrices this method is not useful when the number of assigned moments is large. The reconstruction of an unknown density function using the Maximum Entropy principle with the specified ordinary and fractional moments has been studied in Kevasan and Kapur [9] and Novi Inverardi *et al.* [18], among others.

In Mnatsakanov and Ruymgaart [17] the constructions (2.2) and (3.13) (see Sections 2 and 3 below) have been introduced, and only their convergence has been established.

Different types of convergence of maximum entropy approximation have been studied by Borwein and Lewis [3], Frontini and Tagliani [7], and Novi Inverardi *et al.* [18], but the rates of approximations have not been established yet. Our construction enables us to derive the uniform rate of convergence for moment-recovered cdfs $F_{\alpha,\nu}$, corresponding quantile function Q_α, and the uniform convergence of the moment-recovered density approximation $f_{\alpha,\nu}$, as the parameter $\alpha \to \infty$. Other constructions of moment-recovered cdfs and pdfs (see, (3.13) and (3.14) in Remark 3.2) were proposed in Mnatsakanov [13-14], where the uniform and L_1-rates of the approximations were established.

The paper is organized as follows: in Section 2 we introduce the notation and assumptions, while in Section 3 we study the properties of $F_{\alpha,\nu}$ and $f_{\alpha,\nu}$. Note that our construction also gives a possibility to recover different distributions through the simple transformations of moment sequences of given distributions (see Theorem 3.1 in Section 3 and similar properties derived in Mnatsakanov [13]: Theorem 1 and Corollary 1). In Theorem 3.2 we state the uniform rate of convergence for moment-recovered cdfs. In Theorem 3.3 as well as in Corollaries 3.1 and 3.2 we apply the constructions (2.2) and (3.11) to recover the pdf f, the quantile function Q, and the corresponding quantile density function q of F given the moments of F. In Section 4 some other applications of the constructions (2.2) and (3.11) are discussed: the uniform convergence of the empirical counterpart of (2.2), the rate of approximation of moment-recovered quantile function (see (4.4) in Section 4) along with the demixing and deconvolution problems in several particular models.

Note that our approach is particularly applicable in situations where other estimators cannot be used, e.g., in situations where only moments (empirical) are available. The results obtained in this paper will not be compared with similar results derived by other methods. We only carry out the calculations of moment-recovered cdfs, pdfs, and quantile functions, and compare them with the target distributions via graphs in several simple examples. We also compare the performances of $F_{\alpha,\nu}$ and $f_{\alpha,\nu}$ with the similar constructions studied in Mnatsakanov [13-14] (see, Figures 1 (b) and 3 (b)). The moment-estimated quantile function $\hat{Q}_\alpha$ and well known Harrell-Davis quantile function estimator $\hat{Q}_{HD}$ (Sheather and Marron [19]) defined in (4.6) and (4.7), respectively, are compared as well (see, Figure 2 (b)).

2. Notation and Assumptions

Suppose that the M-determinate cdf F is absolute continuous with respect to the Lebesgue measure and has support $[0, T]$, $T < \infty$. Denote the corresponding density function by f. Our method of recovering the cdf $F(x)$, $0 \leqslant x \leqslant T$, is based on an inverse transformation that yields a solution of the Hausdorff moment problem.

Let us denote the moments of F by

$$\mu_{j,F} = \int t^j d\,F(t) = (\mathcal{K}F)(j), j \in \mathbb{N}, \tag{2.1}$$

and assume that the moment sequence $\nu = (\mu_{0,F}, \mu_{1,F}, \dots)$ determines F uniquely.

An approximate inverse of the operator $\mathcal{K}$ from (2.1) constructed according to

$$\left(\mathcal{K}_\alpha^{-1}\nu\right)(x) = \sum_{k=0}^{[\alpha x]} \sum_{j=k}^{\infty} \frac{(-\alpha)^{j-k}}{(j-k)!} \frac{\alpha^k}{k!}\, \mu_{j,F},\ 0 \leqslant x \leqslant T\,,\ \alpha \in \mathbb{R}_+, \tag{2.2}$$

is such that $\mathcal{K}_\alpha^{-1}\mathcal{K}F \to_w F$, as $\alpha \to \infty$ (see, Mnatsakanov and Ruymgaart [17]). Here $\to_w$ denotes the weak convergence of cdfs, i.e. convergence at each continuity point of the limiting cdf. The success of the inversion formula (2.2) hinges on the convergence

$$P_\alpha(t,x) = \sum_{k=0}^{[\alpha x]} \frac{(\alpha t)^k}{k!}\, e^{-\alpha t} \to \begin{cases} 1, & t < x \\ 0, & t > x \end{cases}, \tag{2.3}$$

as $\alpha \to \infty$. This result is immediate from a suitable interpretation of the left hand side as a sum of Poisson probabilities.

For any moment sequence $\nu = \{\nu_j, j \in \mathbb{N}\}$, let us denote by F_ν the cdf recovered via $F_{\alpha,\nu} = \mathcal{K}_\alpha^{-1}\nu$ according to (2.2), when $\alpha \to \infty$, i.e.

$$F_{\alpha,\nu} \to_w F_\nu\,,\ as\ \alpha \to \infty\,. \tag{2.4}$$

Note that if $\nu = \{\mu_{j,F}, j \in \mathbb{N}\}$ is the moment sequence of F, the statement (2.4) with $F_\nu = F$ is proved in Mnatsakanov and Ruymgaart [17].

To recover a pdf f via its moment sequence $\{\mu_{j,F}, j \in \mathbb{N}\}$, consider the ratio:

$$f_{\alpha,\nu}(x) = \frac{\Delta F_{\alpha,\nu}(x)}{\Delta},\ \ \Delta = \frac{1}{\alpha}\,, \tag{2.5}$$

where $\Delta F_{\alpha,\nu}(x) = F_{\alpha,\nu}(x+\Delta) - F_{\alpha,\nu}(x)$ and $\alpha \to \infty$.

In the sequel the uniform convergence on any bounded interval in $\mathbb{R}_+$ will be denoted by $\longrightarrow_u$, while the sup-norm between two functions f_1 and f_2 by $||\ f_1 - f_2\ ||$. Note also that the statements from Sections 3 and 4 are valid for distributions defined on any compact $[0,T], T < \infty$. Without loss of generality we assume that F has support $[0,1]$.

3. Asymptotic Properties of $F_{\alpha,\nu}$ and $f_{\alpha,\nu}$

In this Section we present asymptotic properties of the moment-recovered cdf $F_{\alpha,\nu}$ and pdf $f_{\alpha,\nu}$ functions based on the transformation $\mathcal{K}_\alpha^{-1}\nu$ (2.2). The uniform approximation rate of $F_{\alpha,\nu}$ and the uniform convergence of $f_{\alpha,\nu}$ are derived as well.

Denote the family of all cdfs defined on $[0,1]$ by $\mathbb{F}$. The construction (2.2) gives us the possibility to recover also two non-linear operators $\mathcal{A}_k : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$, $k = 1,2$, defined as follows: denote the convolution with respect to the multiplication operation on $\mathbb{R}_+$ by

$$F_1 \otimes F_2(x) = \int F_1(x/\tau)\, dF_2(\tau) := \mathcal{A}_1(F_1,F_2)(x), \quad 0 \leqslant x \leqslant 1, \tag{3.1}$$

while the convolution with respect to the addition operation is denoted by

$$F_1 \star F_2(x) = \int F_1(x-\tau)\, dF_2(\tau) := \mathcal{A}_2(F_1,F_2)(x), \quad 0 \leqslant x \leqslant 2\,.$$

For any two moment sequences $\nu_1 = \{\mu_{j,F_1}, j \in \mathbb{N}\}$ and $\nu_2 = \{\mu_{j,F_2}, j \in \mathbb{N}\}$, define $\nu_1 \odot \nu_2 = \{\mu_{j,F_1} \times \mu_{j,F_2}, j \in \mathbb{N}\}$ and $\nu_1 \oplus \nu_2 = \{\bar{\nu}_j, j \in \mathbb{N}\}$, where

$$\bar{\nu}_j = \sum_{m=0}^{j} \binom{j}{m} \mu_{m,F_1} \times \mu_{j-m,F_2}\,, \tag{3.2}$$

while $\mu_F^{\odot k} = \{\mu_{j,F}^k, j \in \mathbb{N}\}$ and $F^{\otimes k} = F \otimes \cdots \otimes F$ for the corresponding k-fold convolution (cf. (3.1)). Also denote by $F \circ \phi^{-1}$ the composition $F(\phi^{-1}(x)), x \in [0,1]$, with ϕ - continuous and increasing function $\phi : [0,1] \to [0,1]$.

Since cdfs $\mathcal{A}_1(F_1,F_2) = F_1 \otimes F_2$, $\mathcal{A}_2(F_1,F_2) = F_1 \star F_2$, and $F \circ \phi^{-1}$ have compact support, they all are M-determinate and have the moment sequences $\nu_1 \odot \nu_2$, $\nu_1 \oplus \nu_2$, and $\nu = \{\bar{\mu}_j\,, j \in \mathbb{N}\}$, with

$$\bar{\mu}_j = \int [\phi(t)]^j\, dF(t), \tag{3.3}$$

respectively. Hence, applying Theorem 3.1 from Mnatsakanov and Ruymgaart [17] a statement similar to the one in Mnatsakanov [13] (see Theorem 1 and Corollary 1, where $T = T' = 1$) is obtained. Besides, the following statement is true:

THEOREM 3.1. *If $\nu = \sum_{k=1}^m \beta_k\, \mu_F^{\odot k}$, where $\sum_{k=1}^m \beta_k = 1, \beta_k > 0$, then (2.4) holds with*

$$F_\nu = \sum_{k=1}^{m} \beta_k F^{\otimes k}\,. \tag{3.4}$$

Proof. The equation (3.4) follows from Theorem 1 (i) (Mnatsakanov [13]) and the linearity of $\mathcal{K}_\alpha^{-1}\nu$.

The construction (2.2) is also useful when recovering the quantile function $Q(t) = \inf\{x : F(x) \geqslant t\}$ via moments (see (4.5) in Section 4). Define $Q_\alpha = F_{\alpha,\nu_Q}$, where

$$\nu_Q = \{\int_0^1 [F(u)]^j \, du, j \in \mathbb{N}\}. \tag{3.5}$$

The following statement is true:

COROLLARY 3.1. *If F is continuous, then $Q_\alpha \to_w Q$, as $\alpha \to \infty$.*

Proof. Replacing the functions ϕ and F in (3.3) by F and the uniform cdf on $[0,1]$, respectively, we obtain from Theorem 1 (iv) (Mnatsakanov [13]) that $Q_\alpha = F_{\alpha,\nu_Q} \to_w F_\nu = F^{-1}$ as $\alpha \to \infty$.

Under additional conditions on the smoothness of F one can obtain the uniform rate of convergence in (2.4) and, hence, in Theorem 3.1 too. Consider the following condition

$$F'' = f' \text{ is bounded on } [0,1]. \tag{3.6}$$

THEOREM 3.2. *If $\nu = \{\mu_{j,F}, j \in \mathbb{N}\}$, and (3.6) holds, we have*

$$\sup_{0 \leqslant x \leqslant 1} \left|F_{\alpha,\nu}(x) - F(x)\right| = O\left(\frac{1}{\alpha}\right), \quad \text{as } \alpha \to \infty. \tag{3.7}$$

Proof. Let us use the following representation

$$P_\alpha(t,x) = \mathbf{P}\{N_{\alpha t} \leqslant \alpha x\} = \mathbf{P}\{S_{[\alpha x]} \geqslant \alpha t\}.$$

Here $\{N_{\alpha t}, t \in [0,1]\}$ is a Poisson process with intensity αt, $S_m = \sum_{k=0}^m \xi_k$, $S_0 = 0$, with ξ_k being *iid* $Exp(1)$ random variables. Integration by parts gives

$$\begin{aligned} F_{\alpha,\nu}(x) = (\mathcal{K}_\alpha^{-1}\,\nu)(x) &= \int_0^1 \sum_{k=0}^{[\alpha x]} \frac{(\alpha t)^k}{k!} \sum_{j=k}^{\infty} \frac{(-\alpha t)^{j-k}}{(j-k)!}\, dF(t) \\ &= \int_0^1 P_\alpha(t,x) d\,F(t) = \int_0^1 \mathbf{P}\{S_{[\alpha x]} \geqslant \alpha t\} dF(t) \\ &= F(t)\,\mathbf{P}\{S_{[\alpha x]} \geqslant \alpha t\}\Big|_0^1 - \int_0^1 F(t)\, d\,\mathbf{P}\{S_{[\alpha x]} \geqslant \alpha t\} \\ &= \mathbf{P}\{S_{[\alpha x]} \geqslant \alpha\} + \int_0^1 F(t)\, d\,\mathbf{P}\{S_{[\alpha x]} \leqslant \alpha t\} = \int_0^\infty F(t)\, d\,\mathbf{P}\{S_{[\alpha x]} \leqslant \alpha t\}. \end{aligned} \tag{3.8}$$

Thus, (3.6) and the argument used in Adell and de la Cal [1] yield (3.7).

REMARK 3.1. When $supp\{F\} = \mathbb{R}_+$, $F_{\alpha,\nu}(x) = \int_0^\infty P_\alpha(t,x) d\,F(t)$ (cf. with (3.8)). According to Mnatsakanov and Klaassen [16] (see the proof of Theorem 3.1), one can derive the exact rate of approximation of $F_{\alpha,\nu}$ in the space $L_2(\mathbb{R}_+, dF)$. Namely, if the pdf f is bounded, say by $C > 0$, then

$$\int_0^\infty \Big(F_{\alpha,\nu}(x) - F(x)\Big)^2 dF(x) \leq \frac{2\,C}{\alpha}.$$

Now let us consider the moment-recovered density function $f_{\alpha,\nu}$ defined in (2.5) and denote by $\Delta(f,\delta) = \sup_{|t-s|\leq\delta} |f(t) - f(s)|$ the modulus of continuity of f, where $0 < \delta < 1$.

THEOREM 3.3. *If the pdf f is continuous on* $[0,1]$, *then* $f_{\alpha,\nu} \longrightarrow_u f$ *and*

$$|| f_{\alpha,\nu} - f ||\leq \Delta(f,\delta) + \frac{2\,||\,f\,||}{\alpha\delta^2} + o\Big(\frac{1}{\alpha}\Big), \quad as \;\; \alpha \to \infty. \tag{3.9}$$

Proof. Since $[\alpha(x + 1/\alpha)] = [\alpha x] + 1$, for any $x \in [0,1]$, we have

$$f_{\alpha,\nu}(x) = \alpha\,\Big[\sum_{k=0}^{[\alpha x]+1}\sum_{j=k}^{\infty}\frac{(-\alpha)^{j-k}}{(j-k)!}\,\frac{\alpha^k}{k!}\,\mu_{j,F} - \sum_{k=0}^{[\alpha x]}\sum_{j=k}^{\infty}\frac{(-\alpha)^{j-k}}{(j-k)!}\,\frac{\alpha^k}{k!}\,\mu_{j,F}\Big], \tag{3.10}$$

and, after some algebra (3.10) yields

$$f_{\alpha,\nu}(x) = \frac{\alpha^{[\alpha x]+2}}{\Gamma([\alpha x]+2)}\cdot\sum_{m=0}^{\infty}\frac{(-\alpha)^m}{m!}\,\mu_{m+[\alpha x]+1,F}\,. \tag{3.11}$$

Let $g(t,a,b)$ denote a gamma pdf with shape and scale parameters a and b, respectively. Substitution of (2.1) into the right hand side of (3.11) gives

$$\begin{aligned} f_{\alpha,\nu}(x) &= \frac{\alpha^{[\alpha x]+2}}{\Gamma([\alpha x]+2)}\int_0^1\sum_{m=0}^{\infty}\frac{(-\alpha t)^m}{m!}t^{[\alpha x]+1}\,dF(t) \\ &= \int_0^1 g(t,[\alpha x]+2,\alpha)\,f(t)dt. \end{aligned} \tag{3.12}$$

To show (3.9), note that the pdf g in (3.12) has mean $([\alpha x]+2)/\alpha$ and variance $([\alpha x]+2)/\alpha^2$, respectively. The rest of the proof is similar to the lines of Theorem 1 (i) (Mnatsakanov [14]).

REMARK 3.2. In Mnatsakanov [13-14] the uniform and L_1-rates of moment-recovered approximations of F and f defined by

$$F^*_{\alpha,\nu}(x) = \sum_{k=0}^{[\alpha x]}\sum_{j=k}^{\alpha}\binom{\alpha}{j}\binom{j}{k}(-1)^{j-k}\,\mu_{j,F} \tag{3.13}$$

and

$$f^*{}_{\alpha,\nu}(x) = \frac{\Gamma(\alpha+2)}{\Gamma([\alpha x]+1)}\sum_{m=0}^{\alpha-[\alpha x]}\frac{(-1)^m\,\mu_{m+[\alpha x],F}}{m!\;(\alpha-[\alpha x]-m)!}, \;\; x\in[0,1]\,,\;\alpha\in\mathbb{N}, \tag{3.14}$$

are established. In Section 4, see Example 4.2, the cdf $F(x) = x^3 - 3\,x^3\ln x$ and its density function $f(x) = -9x^2\ln x, 0 \leqslant t \leqslant 1$, are recovered using $F_{\alpha,\nu}$ and $F^*{}_{\alpha,\nu}$, and $f_{\alpha,\nu}$ and $f^*{}_{\alpha,\nu}$ constructions, (see Figures 1 (b) and 3 (b), respectively).

The formulas (3.11) and (3.14) with $\nu = \nu_Q$ defined according to (3.5) can be used to recover a quantile density function

$$q(x) = Q'(x) = \frac{1}{f(F^{-1}(x))}\,, \quad x\in[0,1].$$

For example, consider $f_{\alpha,\nu_Q} := q_\alpha$: the application of the first line in (3.12) with F^{-1} instead of F yields

$$q_\alpha(x) = \int_0^1 g(F(u), [\alpha x] + 2, \alpha)\, du \tag{3.15}$$

and corresponding moment-recovered quantile density function

$$q_{\alpha,\beta}(x) = \int_0^1 g(F_{\beta,\nu}(u), [\alpha x] + 2, \alpha)\, du\,, \quad \alpha\,, \beta \in \mathbb{N}.$$

Here $F_{\beta,\nu}$ is a moment-recovered cdf of F. As a consequence of Theorem 3.3 we have the following

COROLLARY 3.2. *If $q_\alpha = f_{\alpha,\nu_Q}$, with ν_Q defined in (3.5), and f is continuous on* [0, 1] *with* $\inf_{0\leqslant x\leqslant 1} f(x) > \gamma > 0$, *then* $q_\alpha \longrightarrow_u q$ *and*

$$|| \, q_\alpha - q \, || \leq \frac{\Delta(f,\delta)}{\gamma^2} + \frac{2\,||\, f\,||}{\alpha\,\delta^2\,\gamma^2} + o\Big(\frac{1}{\alpha}\Big)\,, \quad as \quad \alpha \to \infty.$$

Finally, note that taking $\nu = \nu_Q$ in (3.14), we derive another approximation $q^*_\alpha = f^*_{\alpha,\nu_Q}$ of q based on Beta densities $\beta(\cdot, a, b)$ with the shape parameters $a = [\alpha x] + 1$ and $b = \alpha - [\alpha x] + 1$:

$$q^*_\alpha(x) = \int_0^1 \beta(F(u), [\alpha x] + 1, \alpha - [\alpha x] + 1)\, du\,. \tag{3.16}$$

4. Some Applications and Examples

In this Section the construction of the moment-recovered cdf $F_{\alpha,\nu}$ is applied to the problem of nonparametric estimation of a cdf, its density and a quantile functions as well as to the problem of demixing in exponential, binomial and negative binomial mixtures, and deconvolution in error-in-variable model. In Theorems 4.1 we derive the uniform rate of convergence for the empirical counterpart of $F_{\alpha,\nu}$ denoted by $\widetilde{F}_\alpha$, i.e. for $\widetilde{F}_\alpha = F_{\alpha,\hat{\nu}}$, where $\hat{\nu}$ is the sequence of all empirical moments of the sample from F. In Theorem 4.2 the uniform rate of approximation for moment-recovered quantile function of F is obtained. Finally, the graphs of moment-recovered cdfs, pdfs, and quantile functions are presented in Figures 1-3.

Direct model. Let $X_1, \ldots, X_n$ be a random sample from F defined on $[0, 1]$. Denote by $\hat{F}_n$ the empirical cdf (ecdf) of the sample $X_1, \ldots, X_n$:

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n I_{[0,t]}(X_i)\,, \quad 0 \leqslant t \leqslant 1\,.$$

Substitution of the empirical moments

$$\hat{\nu}_j = \frac{1}{n}\sum_{i=1}^n X_i^j\,,\ j \in \mathbb{N}\,,$$

instead of $\mu_{j,F}$ into (2.2) yields

$$\widetilde{F}_\alpha(x) = F_{\alpha,\hat{\nu}}(x) = \int_0^1 P_\alpha(t,x) d\,\hat{F}_n(t) = \int_0^1 \mathbf{P}\{S_{[\alpha x]} \geqslant \alpha t\} d\,\hat{F}_n(t)\,.$$

Furthermore, the empirical analogue of (3.8) admits a similar representation

$$\widetilde{F}_\alpha(x) = \int_0^\infty \hat{F}_n(t)\, d\,\mathbf{P}\{S_{[\alpha x]} \leqslant \alpha t\}.$$

The application of the Theorem 3.2 and the asymptotic properties of $\hat{F}_n$ yield

THEOREM 4.1. *If $\nu = \{\mu_{j,F}, j \in \mathbb{N}\}$, then under condition (3.6) we have*

$$\sup_{0\leqslant x\leqslant 1} \left|\widetilde{F}_\alpha(x) - F(x)\right| = O\left(\frac{1}{\sqrt{n}}\right) + O\left(\frac{1}{\alpha}\right) \quad \text{a.s.,} \quad \text{as} \quad \alpha\,, n \to \infty. \tag{4.1}$$

REMARK 4.1. In Mnatsakanov and Ruymgaart [17] the weak convergence of the moment-empirical processes $\{\sqrt{n}\{\widetilde{F}_n(t) - F(t)\}, t \in [0,1]\}$ to the Brownian bridge is obtained.

Of course, when the sample is directly drawn from the cdf F of actual interest, one might use the ecdf $\hat{F}_n$ and empirical process $U_n = \sqrt{n}\,(\hat{F}_n - F)$. The result mentioned in Remark 4.1 yields, that even if the only information available is the empirical moments, we still can construct different test statistics based on the moment-empirical processes $\widetilde{U}_n = \sqrt{n}\,(\widetilde{F}_n - F)$.

On the other hand, using the construction (3.11), one can estimate the density function f given only the estimated or empirical moments in:

$$f_{\alpha,\hat{\nu}}(x) = \frac{\alpha^{[\alpha x]+2}}{\Gamma([\alpha x]+2)} \sum_{m=0}^{\infty} \frac{(-\alpha)^m}{m!}\, \hat{\nu}_{m+[\alpha x]+1}\,,\ x \in [0,1]\,. \tag{4.2}$$

REMARK 4.2. In practice, the parameter α as well as the number of summands in (4.2) (and the number of summands in the inner summation of $F_{\alpha,\hat{\nu}}$) can be chosen as the functions of n: $\alpha = \alpha(n) \to \infty$ and $M = M(n) \to \infty$ as $n \to \infty$, that optimize the accuracy of corresponding estimates. Further analysis is required to derive the asymptotic forms of $\alpha(n)$ and $M(n)$ as $n \to \infty$. This question is currently under investigation and is beyond the scope of the present article.

Note that the construction (4.2) yields the estimator $\hat{f}_\alpha(x) = f_{\alpha,\hat{\nu}}$ with $\hat{\nu} = \{\hat{\nu}_j, j \in \mathbb{N}\}$:

$$\hat{f}_\alpha(x) = \frac{\alpha}{n} \sum_{i=1}^{n} \frac{(\alpha X_i)^{[\alpha x]+1}}{([\alpha x]+1)!}\, e^{-\alpha X_i} = \frac{1}{n}\sum_{i=1}^{n} g(X_i, [\alpha x]+2, \alpha),\ x \in [0,1]\,.$$

Here $g(\cdot, [\alpha x]+2, \alpha)$ is defined in (3.12). The estimator $\hat{f}_\alpha$ does not represent a traditional kernel density estimator of f. It is defined by a δ-sequence, which consists of the gamma density functions of varying shapes (the shape and the rate parameters are equal to $[\alpha x]+2$ and α, respectively). It is natural to use this estimator when $supp\{F\} = [0,\infty)$, since, in this case, the supports of f and gamma kernel densities coincide and one avoids the boundary effect of $\hat{f}_\alpha$ (cf. Chen [5]).

Some asymptotic properties such as the convergence in probability of $\hat{f}_\alpha$ uniformly on any bounded interval and the Integrated Mean Squared Error (*IMSE*) of $\hat{f}_\alpha$ have been studied in Mnatsakanov and Ruymgaart [17] and Chen [5], respectively.

Applying the results from Mnatsakanov and Khmaladze [15], where the necessary and sufficient conditions for L_1-consistency of general kernel density estimates

are established, one can prove in a similar way (cf. Mnatsakanov [14], Theorem 3) that if f is continuous on $[0,1]$, then $E \mid\mid \hat{f}_\alpha - f \mid\mid_{L_1} \to 0$, as $\sqrt{\alpha}/n \to 0$ and $\alpha\,, n \to \infty$.

Exponential mixture model. Let $Y_1, \ldots, Y_n$ be a random sample from the mixture of exponentials

$$G(x) = \int_0^T (1 - e^{-x/\tau})\, dF(\tau)\,,\ x \geqslant 0\,.$$

The unknown cdf F can be recovered according to the construction $F_{\alpha,\nu} = \mathcal{K}_\alpha^{-1}\,\nu$ with $\nu = \{\mu_{j,G}/j!, j \in \mathbb{N}\}$. Similarly, given the sample $Y_1, \ldots, Y_n$ from G and taking $F_{\alpha,\hat{\nu}} = \mathcal{K}_\alpha^{-1}\,\hat{\nu}$, where $\hat{\nu} = \{\hat{\mu}_{j,G}/j!, j \in \mathbb{N}\}$, we obtain the estimate of F. Here $\{\hat{\mu}_{j,G}, j \in \mathbb{N}\}$ are the empirical moments of the sample $Y_1, \ldots, Y_n$. The regularized inversion of the noisy Laplace transform and the L_2-rate of convergence were obtained in Chauveau *et al.* [4].

Binomial and negative binomial mixture models. When $Y_1, \ldots, Y_n$ is a random sample from the binomial or negative binomial mixture distributions, respectively:

$$p(x) := P(Y = x) = \int_0^1 \binom{m}{x} \tau^x\,(1-\tau)^{m-x} dF(\tau)\,,\ x = 0, \ldots m\,,$$

$$p(x) := P(Y = x) = \int_0^1 \frac{\Gamma(r+x)}{\Gamma(r)\,x!} \left(\frac{1}{1+\tau}\right)^r \left(\frac{\tau}{1+\tau}\right)^x dG(\tau)\,,\ x = 0, 1, \ldots\,,$$

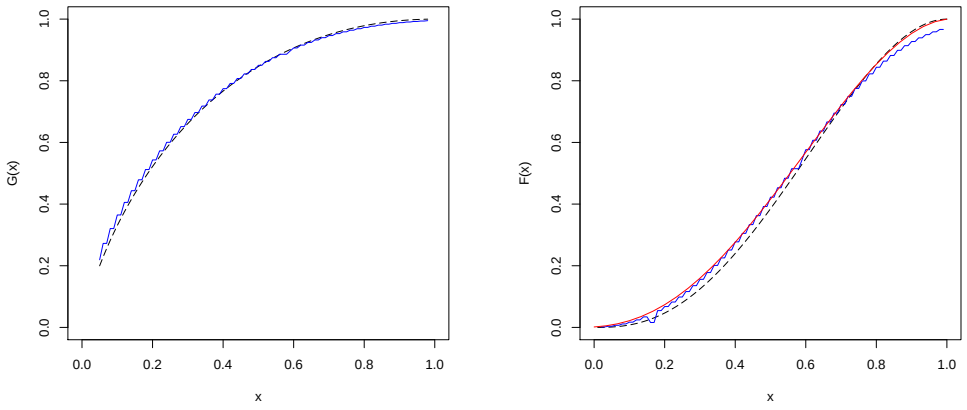
where m and r are given positive integers. Assume that the unknown mixing cdfs F and G are such that F has at most $\frac{m+1}{2}$ support points in $(0,1)$, while G is a right continuous cdf on $(0,1)$. In both models the mixing distributions are identifiable (see, for example, Teicher [26] for binomial mixture model). Note also that the jth moments of F and G are related to the jth factorial moments of corresponding Y_i's in the following ways:

$$\mu_{j,F} = \frac{1}{m^{[j]}}\, E(Y_1^{[j]}) \ \text{ and } \ \mu_{j,G} = \frac{1}{r_{(j)}}\, E(Y_1^{[j]})\,.$$

Here $y^{[j]} = y(y-1)\cdots(y-j+1)$ and $r_{(j)} = r(r+1)\cdots(r+j-1)$. To estimate F and G one can use the moment-recovered formulas (2.2) or (3.13) with $\mu_{j,F}$ and $\mu_{j,G}$ defined in previous two equations where the theoretical factorial moments are replaced by corresponding empirical counterparts. The asymptotic properties of the derived estimators of F and G will be studied in a separate work.

Deconvolution problem: error-in-variable model. Consider the random variable $Y = X + U$, with cdf G, where U (the error) has some known symmetric distribution F_2, X has cdf F_1 with a support $[0,T]$, and U and X are independent. This model, known as an error-in-variable model, corresponds to the convolution $G = F_1 \star F_2$. Assuming that all moments of X and U exist, the moments $\{\bar{\nu}_j, j \in \mathbb{N}\}$ of Y are described by (3.2). Hence, given the moments of U (with $E(U) = 0$), we can recalculate the moments of F_1 as follows: $\mu_{1,F_1} = \bar{\nu}_1$, $\mu_{2,F_1} = \bar{\nu}_2 - \mu_{2,F_2}$, and so on. So that, assuming that we already calculated μ_{k,F_1}, or estimated them by μ^*_{k,F_1} for $1 \leqslant k \leqslant j-2$, we will have, for any $j \geqslant 1$:

$$\mu_{j,F_1} = \bar{\nu}_j - \sum_{m=2}^{j} \binom{j}{m} \mu_{m,F_2} \times \mu_{j-m,F_1}$$



(a) (b)

FIG 1. *(a) Approximation of* $G(x) = x - x\ln x$ *by* $F_{\alpha,\nu}$ *and (b) Approximation of* $G(x^3)$ *by* $F_{\alpha,\nu}$ *and by* $F^*_{\alpha,\nu}$

or, respectively,

$$\mu^*_{j,F_1} = \hat{\mu}_{j,G} - \sum_{m=2}^{j} \binom{j}{m} \mu_{m,F_2} \times \mu^*_{j-m,F_1}$$

given the sample $Y_1, \ldots, Y_n$ from cdf G. Now the moment-recovered estimate of F_1 will have the form $F_{\alpha,\hat{\nu}} = \mathcal{K}_\alpha^{-1}\hat{\nu}$, where $\hat{\nu} = \{\mu^*_{j,F_1}, j \in \mathbb{N}\}$. The alternative construction of the kernel type estimate of F_1 based on the Fourier transforms is studied in Hall and Lahiri [8], where the $\sqrt{n}$-consistency and other properties of the estimated moments $\mu^*_{j,F_1}, j \in \mathbb{N}$, are derived as well.

Example 4.1. Consider the moment sequence $\mu = \{1/(j+1), j \in \mathbb{N}\}$. The corresponding moment-recovered distribution $F_{\alpha,\mu} = \mathcal{K}_\alpha^{-1}\mu$ is a good approximation of $F(x) = x$ already with $\alpha = 50$ and $M = 100$.

Assume now that we want to recover the distribution G with corresponding moments $\nu_{j,G} = 1/(j+1)^2, j \in \mathbb{N}$. Since we can represent $\nu_G = \mu \odot \mu$, we conclude from Theorem 1 (i) in Mnatsakanov [13], that $G = F \otimes F$, with $F(x) = x$, and hence $G(x) = x - x\ln x, 0 \leqslant x \leqslant 1$. We plotted the curves of F_{α,ν_G} (the solid line) and G (the dashed line) on Figure 1 (a). We took $\alpha = 50$ and $M = 200$, the number of terms in the inner summation of the formula (2.2). From Figure 1 (a) we can see that the approximation of G by F_{α,ν_G} at $x = 0$ is not as good as inside of the interval $[0, 1]$. This happened because the condition (3.6) from Theorem 3.2 is not valid for $g'(x) = G''(x) = -1/x$.

Example 4.2. To recover the distribution F via moments $\nu_j = 9/(j+3)^2, j \in \mathbb{N}$, note that $\nu_j = \nu_{aj,G}$, with $a = 1/3$. Hence, $F(x) = G(x^3) = x^3 - x^3\ln(x^3), 0 \leqslant x \leqslant 1$ (Theorem 1 (iii), Mnatsakanov [13]). We conducted computations of moment-recovered cdf $F_{\alpha,\nu}$ when $\alpha = 50$ and the number of terms in the inner summation of the formula (2.2) is equal to 200. Also, we calculated $F^*_{\alpha,\nu}$ defined in (3.13) with $\alpha = 32$. See Figure 1 (b), where we plotted $F_{\alpha,\nu}$ (the solid blue line), $F^*_{\alpha,\nu}$ (the

solid red line), and F (the dashed line), respectively. These two approximations of cdf F justify a good fit already with $\alpha = 50$ and $M = 200$ for the first one and with $\alpha = 32$ for the second one. From Figure 1 (b) we can see that the performance of $F^*_{\alpha,\nu}$ is slightly better compared to $F_{\alpha,\nu}$: $F^*_{\alpha,\nu}$ does not have the "boundary" effect around $x = 1$.

Estimation of a quantile function Q and quantile density function q. Assume that a random variable X has a continuous cdf F defined on $[0, 1]$. To approximate (estimate) the quantile function Q given only the moments (estimated moments) of F, one can use Corollary 3.1. Indeed, after some algebra, we have

$$Q_\alpha(x) = F_{\alpha,\nu_Q}(x) = \int_0^1 P_\alpha\big(F(u), x\big)\, du\,, \quad 0 \leqslant x \leqslant 1\,, \tag{4.3}$$

where ν_Q and $P_\alpha(\cdot,\cdot)$ are defined in (3.5) and in (2.3), respectively. Comparing (4.3) and (3.8) we can prove in a similar way (see, the proof of Theorem 3.2) the following

THEOREM 4.2. *If f' is bounded and $\inf_{0\leqslant x\leqslant 1} f(x) > \gamma > 0$, then*

$$\sup_{0\leqslant x\leqslant 1} \big|Q_\alpha(x) - Q(x)\big| = O\left(\frac{1}{\alpha}\right), \quad \text{as } \alpha \to \infty. \tag{4.4}$$

Now, given only the moment sequence ν of F, one can construct the approximation $Q_{\alpha,\beta}$ of Q by substituting the moment-recovered cdf $F_{\beta,\nu}$ (instead of F) in the right hand side of (4.3). Let us denote the corresponding approximation of Q by

$$Q_{\alpha,\beta}(x) = \int_0^1 P_\alpha\big(F_{\beta,\nu}(u), x\big)\, du\,, \quad \alpha\,, \beta \in \mathbb{N}. \tag{4.5}$$

Figure 2 (a) shows the cdf $F(x) = x^3 - x^3\ln(x^3)$ (the dashed line), introduced in Example 4.2, and its quantile approximation $Q_{\alpha,\beta}$ (the solid line), when $\nu = \{9/(j+3)^2, j \in \mathbb{N}\}$, $\alpha = \beta = 100$, and $M = 200$.

Replacing F by the empirical $\hat{F}_n$ in (4.3), (3.15), and in (3.16) yields the following estimators, respectively, based on the spacings $\Delta X_{(i)} = X_{(i)} - X_{(i-1)}, i = 1, \ldots, n+1$:

$$\hat{Q}_\alpha(x) = F_{\alpha,\hat{\nu}_Q}(x) = \int_0^1 P_\alpha\big(\hat{F}_n(u), x\big)\, du = \sum_{i=1}^{n+1} \Delta X_{(i)}\, P_\alpha\big(\frac{i-1}{n}, x\big)\,, \tag{4.6}$$

$$\hat{q}_\alpha(x) = \int_0^1 g(\hat{F}_n(u), [\alpha x]+2, \alpha)\, du = \sum_{i=1}^{n+1} \Delta X_{(i)}\, g\big(\frac{i-1}{n}, [\alpha x]+2, \alpha\big)\,,$$

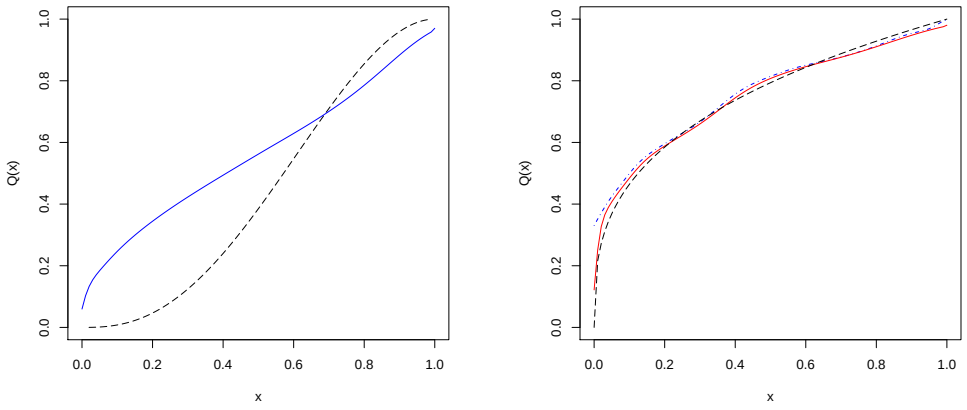
and

$$\hat{q}^*_\alpha(x) = \sum_{i=1}^{n+1} \Delta X_{(i)}\, \beta\big(\frac{i-1}{n}, [\alpha x]+1, \alpha - [\alpha x]+1\big)\,.$$

Here $\hat{\nu}_Q = \{\int_0^1 [\hat{F}_n(u)]^j du, j \in \mathbb{N}\}$, while $X_{(i)}, i = 1, \ldots, n, X_{(0)} = 0, X_{(n+1)} = 1$, are the order statistics of the sample $X_1, \ldots, X_n$.

Now, let us compare the curves of $\hat{Q}_\alpha$ and the well known Harrell-Davis estimator

$$\hat{Q}_{HD}(x) = \sum_{i=1}^{n} X_{(i)}\, \Delta Beta\Big(\frac{i}{n},\, (n+1)x,\, (n+1)(1-x)\Big)\,, \tag{4.7}$$



(a) (b)

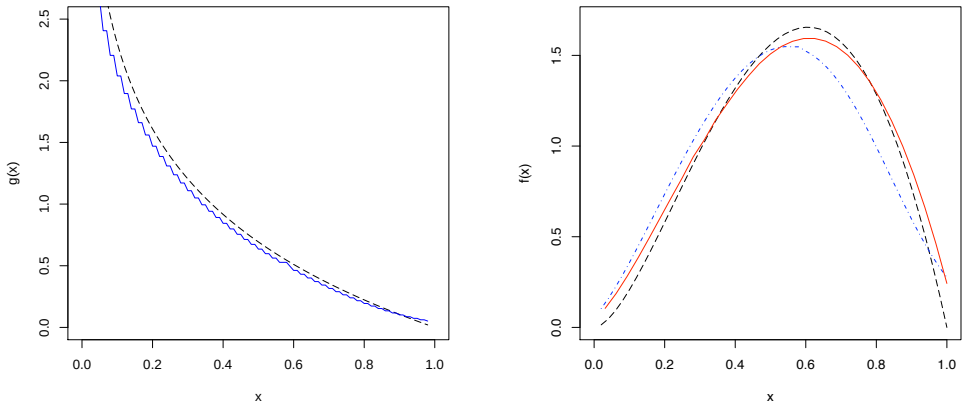
FIG 2. *(a) Approximation of Q by $Q_{\alpha,\beta}$ and (b) Estimation of $Q(x) = x^{1/3}$ by $\hat{Q}_\alpha$ and by $\hat{Q}_{HD}$*

where $Beta\left(\cdot, a, b\right)$ denotes the cdf of a $Beta$ distribution with the shape parameters $a > 0$ and $b > 0$. For asymptotic expressions of MSE and the bias term of $\hat{Q}_{HD}$ we refer the reader to Sheather and Marron [19]. Let us generate $n = 100$ independent random variables $X_1, \ldots, X_n$ from $F(x) = x^3, 0 \leqslant x \leqslant 1$. Taking $\alpha = 100$, we estimate (see, Figure 2 (b)) the corresponding quantile function $Q(x) = x^{1/3}, 0 \leqslant x \leqslant 1$, (the dashed line) by means of $\hat{Q}_\alpha$ (the solid line) and by $\hat{Q}_{HD}$ (the dashed-dotted line), defined in (4.6) and (4.7), accordingly. Through simulations we conclude that the asymptotic behavior of the moment-recovered estimator $\hat{Q}_\alpha$ and the Harrell-Davis estimator $\hat{Q}_{HD}$ are similar. The MSE and other properties of $\hat{Q}_\alpha, \hat{q}_\alpha$, and $\hat{q}^*_\alpha$ will be presented in a separate article.

Example 4.1 (continued). Assume now that we want to recover pdf of the distribution G studied in the Example 4.1 via the moments $\nu_{j,G} = 1/(j+1)^2, j \in \mathbb{N}$. On the Figure 3 (a) we plotted the curves of the moment-recovered density $f_{\alpha,\nu}$ (the solid line) defined by (3.11) and $g(x) = G'(x) = -\ln x,\ 0 \leqslant x \leqslant 1$ (the dashed line), respectively. Here we took $\alpha = 50$ and $M = 200$.

Example 4.2 (continued). Now let us recover the pdf $f(x) = -9x^2 \ln x, 0 \leqslant x \leqslant 1$, of distribution F defined in Example 4.2 where $\nu_{j,F} = 9/(j+3)^2, j \in \mathbb{N}$. We applied the approximations $f_{\alpha,\nu}$ and ${f^*}_{\alpha,\nu}$ defined in (3.11) and (3.14), respectively, by calculating the values of $f_{\alpha,\nu}$ and ${f^*}_{\alpha,\nu}$ at the points $x = k/\alpha$, $k = 1, 2, \ldots, \alpha$. Figure 3 (b) shows the curves of $f_{\alpha,\nu}$ (the blue dashed-dotted line), and $f^*_{\alpha,\nu}$ (the red solid line), and f (the black dashed line). Here, we took $\alpha = 50$ and $M = 200$ when calculating $f_{\alpha,\nu}$ and $\alpha = 32$ in $f^*_{\alpha,\nu}$. One can see that the performance of $f^*_{\alpha,\nu}$ with $\alpha = 32$ is better than the performance of $f_{\alpha,\nu}$ with $\alpha = 50$ and $M = 200$.

After conducting many calculations of moment-recovered approximants for several models we conclude that the accuracy of the formulas (2.2) and (3.11) are not as good as the ones defined in (3.13) and (3.14) in the Hausdorff case. On the other hand, the constructions (2.2) and (3.11) could be useful in the Sieltjes moment problem as well.



(a) (b)

FIG 3. *(a) Approximation of* $g(x) = -\ln x$ *by* $f_{\alpha,\nu}$ *and (b) Approximation of* $f(x) = -9x^2 \ln x$ *by* $f_{\alpha,\nu}$ *and* $f^*_{\alpha,\nu}$

Acknowledgments

The authors are thankful to Estate Khmaladze, E. James Harner, and Cecil Burchfiel for helpful discussions. The work of the first author was supported by the National Institute for Occupational Safety and Health, contract no. 212-2005-M-12857. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

References

[1] ADELL, J. A. and DE LA CAL, J. (1993). On the uniform convergence of normalized Poisson mixtures to their mixing distribution. *J. Statist. Prob. Letters*, **18**, 227–232.
[2] AKHIEZER, N. I. (1965). *The Classical Moment Problem and Some Related Questions in Analysis*. Oliver & Boyd, Edinburgh.
[3] BORWEIN, J. M. and LEWIS, A. S. (1993). A survey of convergence results for maximum entropy methods. In Djafari, M. and Demoments, G., eds. *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, 39–48.
[4] CHAUVEAU, D. E., VAN ROOIJ, A. C. M. and RUYMGAART, F. H. (1994). Regularized inversion of noisy Laplace transforms. *Advances in Appl. Math.*, **15**, 186–201.
[5] CHEN, S. X. (2000). Probability density function estimation using gamma kernels. *Ann. Inst. Statist. Math.*, **52**, 471–480.
[6] FELLER, W. (1971). *An Introduction to Probability Theory and its Applications*. vol. II. Wiley, New York.
[7] FRONTINI, M. and TAGLIANI, A. (1997). Entropy-convergence in Stieltejs and Hamburger moment problem. *Applied Math. and Computation*, **88**, 39–51.
[8] HALL, P. and LAHIRI, S. N. (2008). Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Statist.*, **36**, 2110–2134.
[9] KEVASAN, H. K. and KAPUR, J. N. (1992). *Entropy Optimization Principles with Applications*. Academic Press. New York.
[10] LIN, G. D. (1992). Characterizations of distributions via moments. *Sankhya*, **54**, Series A, 128–132.
[11] LIN, G. D. (1997). On the moment problem. *J. Statist. Prob. Letters*, **35**, 85–90.
[12] LINDSAY, B. G., PILLA, R. S. and BASAK, P. (2000). Moment-based approximations of distributions using mixtures: theory and applications. *Ann. Inst. Statist. Math.*, **52**, 215–230.
[13] MNATSAKANOV, R. M. (2008). Hausdorff moment problem: reconstruction of distributions. *J. Statist. Prob. Letters*, **78**, 1612–1618.

[14] Mnatsakanov, R. M. (2008). Hausdorff moment problem: Reconstruction of probability density functions. *J. Statist. Prob. Letters*, **78**, 1869–1877.
[15] Mnatsakanov, R. M. and Khmaladze, E. V. (1981). On L_1-convergence of statistical kernel estimators of distribution densities. *Soviet Math. Dokl.*, **23**, 633–636.
[16] Mnatsakanov, R. M. and Klaassen, C. A. J. (2003). Estimation of the mixing distribution in the Poisson mixture models. Uncensored and censored samples. *Proc. Hawaii International conference on Statistics and Related Fields*, Honolulu, Hawaii, June 4-8, 2003, 1-18, CD: ISSN# 1539–7211.
[17] Mnatsakanov, R. M. and Ruymgaart, F. H. (2003). Some properties of moment-empirical cdf's with application to some inverse estimation problems. *Math. Meth. Statist.*, **12**, 478–495.
[18] Novi Inverardi, P. L., Petri, A., Pontuale, G. and Tagliani, A. (2003). Hausdorff moment problem via fractional moments. *Applied Mathematics and Computation*, **144**, 61–74.
[19] Sheather, S. J. and Marron, J. S. (1990). Kernel quantile estimators. *JASA*, **85**, No 410, 410–416.
[20] Shohat, J. A. and Tamarkin, J. D. (1943). *The Problem of Moments*. Amer. Math. Soc., New York.
[21] Stieltjes, T. (1894). Recherches sur les fractions continues. *Anns Sci. Fac. Univ. Toulouse (1894-1895)*, **8** J1–J122; A5–A47.
[22] Stoyanov, J. (1997). *Counterexamples in Probability*. 2nd Ed. Wiley, Chichester.
[23] Stoyanov, J. (2000). Krein condition in probabilistic moment problems. *Bernoulli*, **6**, 939–949.
[24] Stoyanov, J. (2004). Stieltjes classes for moment indeterminate probability distributions. *J. Appl. Probab.*, **41A**, 281–294.
[25] Tagliani, A. and Velasquez, Y. (2003). Numerical inversion of the Laplace transform via fractional moments. *Applied Mathematics and Computation*, **143**, 99–107.
[26] Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.*, **34**, 1265–1269.
[27] Widder, D. V. (1934). The inversion of the Laplace integral and the related moment problem. *Transactions of the Amer. Math. Soc.*, **36**, 107–200.

IMS Collections
Vol. 0 (2009) 266–275
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000013

Asymptotic Efficiency of Simple Decisions for the Compound Decision Problem

Eitan Greenshtein[1,*] **and Ya'acov Ritov**[2,*]

Duke University and Jerusalem, Israel

Abstract: We consider the compound decision problem of estimating a vector of n parameters, known up to a permutation, corresponding to n independent observations, and discuss the difference between two symmetric classes of estimators. The first and larger class is restricted to the set of all permutation invariant estimators. The second class is restricted further to simple symmetric procedures. That is, estimators such that each parameter is estimated by a function of the corresponding observation alone. We show that under mild conditions, the minimal total squared error risks over these two classes are asymptotically equivalent up to essentially $O(1)$ difference.

Contents

1 Introduction 266
2 Basic Lemma and Notation 269
3 Two Valued Parameter 269
4 Dense μ's 270
5 Main Result 272
References 274

1. Introduction

Let $\mathcal{F} = \{F_\mu : \mu \in \mathcal{M}\}$ be a parameterized family of distributions. Let $Y_1, Y_2 \dots$ be a sequence of independent random variables, where Y_i takes value in some space $\mathcal{Y}$, and $Y_i \sim F_{\mu_i}$, $i = 1, 2, \dots$. For each n, we suppose that the sequence $\mu_{1:n}$ is known up to a permutation, where for any sequence $x = (x_1, x_2, \dots)$ we denote the subsequence $x_s, \dots, x_t$ by $x_{s:t}$. We denote by $\boldsymbol{\mu} = \boldsymbol{\mu}_n$ the set $\{\mu_1, \dots, \mu_n\}$, i.e., $\boldsymbol{\mu}$ is $\mu_{1:n}$ without any order information. We consider in this note the problem of estimating $\mu_{1:n}$ by $\hat{\mu}_{1:n}$ under the loss $\sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2$, where $\hat{\mu}_{1:n} = \Delta(Y_{1:n})$. We assume that the family $\mathcal{F}$ is dominated by a measure ν, and denote the corresponding densities simply by $f_i = f_{\mu_i}$, $i = 1, \dots, n$. The important example is, as usual, $F_{\mu_i} = N(\mu_i, 1)$.

Let $\mathcal{D}^S = \mathcal{D}^S_n$ be the set of all *simple symmetric decision functions* Δ, that is, all Δ such that $\Delta(Y_{1:n}) = (\delta(Y_1), \dots, \delta(Y_n))$, for some function $\delta : \mathcal{Y} \to \mathcal{M}$. In particular, the best simple symmetric function is denoted by $\Delta^S_{\boldsymbol{\mu}} = (\delta^S_{\boldsymbol{\mu}}(Y_1), \dots, \delta^S_{\boldsymbol{\mu}}(Y_n))$:

$$\Delta^S_{\boldsymbol{\mu}} = \arg\min_{\Delta \in \mathcal{D}^S_n} \mathrm{E}\, ||\Delta - \mu_{1:n}||^2,$$

[1]Department of Statistical Sciences, Duke University, Durham, NC 27708-0251, USA, email: eitan.greenshtein@gmail.com
[2]Jerusalem, Israel, email: yaacov.ritov@gmail.com
*Partially supported by NSF grant DMS-0605236, and an ISF Grant.
AMS 2000 subject classifications: Primary 62C25; secondary 62C12, 62C07
Keywords and phrases: Compound decision, Simple decision rules, Permutation invariant rules

and denote

$$r_n^S = \mathrm{E}\,||\Delta_{\boldsymbol{\mu}}^S(Y_{1:n}) - \mu_{1:n}||^2,$$

where, as usual, $\|a_{1:n}\|^2 = \sum_{i=1}^n a_i^2$.

The class of simple rules may be considered too restrictive. Since the μs are known up to a permutation, the problem seems to be of matching the Ys to the μs. Thus, if $Y_i \sim N(\mu_i, 1)$, and $n = 2$, a reasonable decision would make $\hat{\mu}_1$ closer to $\mu_1 \wedge \mu_2$ as Y_2 gets larger. The simple rule clearly remains inefficient if the μs are well separated, and generally speaking, a bigger class of decision rules may be needed to obtain efficiency. However, given the natural invariance of the problem, it makes sense to be restricted to the class $\mathcal{D}^{PI} = \mathcal{D}_n^{PI}$ of all permutation invariant decision functions, i.e, functions Δ that satisfy for any permutation π and any $(Y_1, \ldots, Y_n)$:

$$\Delta(Y_1, \ldots, Y_n) = (\hat{\mu}_1, \ldots, \hat{\mu}_n) \quad \Longleftrightarrow \quad \Delta(Y_{\pi(1)}, \ldots, Y_{\pi(n)}) = (\hat{\mu}_{\pi(1)}, \ldots, \hat{\mu}_{\pi(n)}).$$

Let

$$\Delta_{\boldsymbol{\mu}}^{PI} = \arg\min_{\Delta \in \mathcal{D}^{PI}} \mathrm{E}\,||\Delta(Y^n) - \mu_{1:n}||^2$$

be the optimal permutation invariant rule under $\boldsymbol{\mu}$, and denote its risk by

$$r_n^{PI} = E||\Delta_{\boldsymbol{\mu}}^{PI}(Y_{1:n}) - \mu_{1:n}||^2.$$

Obviously $\mathcal{D}^S \subset \mathcal{D}^{PI}$, and whence $r_n^S \geq r_n^{PI}$. Still, 'folklore', theorems in the spirit of De Finetti, and results like Hannan and Robbins (1955), imply that asymptotically (as $n \to \infty$) $\Delta_{\mu^n}^{PI}$ and $\Delta_{\mu^n}^S$ will have 'similar' mean risks: $r_n^S - r_n^{PI} = o(n)$. Our main result establishes conditions that imply the stronger claim, $r_n^S - r_n^{PI} = O(1)$.

To repeat, $\boldsymbol{\mu}$ is assumed known in this note. In the general decision theory framework the unknown parameter is the order of its member to correspond with $Y_{1:n}$, and the parameter space, therefore, corresponds to the set of all the permutations of $1, \ldots, n$.

An asymptotic equivalence as above implies, that when we confine ourselves to the class of permutation invariant procedures, we may further restrict ourselves to the class of simple symmetric procedures, as is usually done in the standard analysis of compound decision problems. The later class is smaller and simpler.

The motivation for this paper stems from the way the notion of oracle is used in some sparse estimation problems. Consider two oracles, both *know* the value of $\boldsymbol{\mu}$. Oracle I is restricted to use only a procedure from the class $\mathcal{D}^{PI}$, while Oracle II is further restricted to use procedures from $\mathcal{D}^S$. Obviously Oracle I has an advantage, our results quantify this advantage and show that it is asymptotically negligible. Furthermore, starting with Robbins (1951) various oracle-inequalities were obtained showing that one can achieve nearly the risk of Oracle II, by a 'legitimate' statistical procedure. See, e.g., the survey Zhang (2003), for oracle-inequalities regarding the difference in risks. See also Brown and Greenshtein (2007), and Wenuha and Zhang (2007) for oracle inequalities regarding the ratio of the risks. However, Oracle II is limited, and hence, these claims may seem to be too weak. Our equivalence results, extend many of those oracle inequalities to be valid also with respect to Oracle I. We needed a stronger result than the usual objective that the mean risks are equal up to $o(1)$ difference. Many of the above mentioned recent applications of the compound decision notion are about sparse situations when most of the μs are in fact 0, the mean risk is $o(1)$, and the only interest is in total risk.

Let $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$ be some arbitrary ordering of $\boldsymbol{\mu}$. Consider now the Bayesian model under which $(\pi, Y_{1:n})$, π a random permutation, have a distribution given by

$$\begin{aligned} &\pi \text{ is uniformly distributed over } \mathcal{P}(1:n); \\ &\text{Given } \pi,\ Y_{1:n} \text{ are independent, } Y_i \sim F_{\boldsymbol{\mu}_{\pi(i)}},\ i = 1, \ldots, n, \end{aligned} \tag{1.1}$$

where for every $s < t$, $\mathcal{P}(s:t)$ is the set of all permutations of $s, \ldots, t$. The above description induces a joint distribution of $(M_1, \ldots, M_n, Y_1, \ldots, Y_n)$, where $M_i \equiv \boldsymbol{\mu}_{\pi(i)}$, for a random permutation π.

The first part of the following proposition is a simple special case of general theorems representing the best invariant procedure under certain groups as the Bayesian decision with respect to the appropriate Haar measure; for background see, e.g., Berger (1985), Chapter 6. The second part of the proposition was derived in various papers starting with Robbins (1951).

In the following proposition and proof, $\mathrm{E}_{\mu_{1:n}}$ is the expectation under the model in which the observations are independent, $Y_i \sim F_{\mu_i}$, and $\mathrm{E}_{\boldsymbol{\mu}}$ is the expectation under the above joint distribution of $Y_{1:n}$ and $M_{1:n}$. Note that under the latter model, for any $i = 1, \ldots, n$, marginally $M_i \sim \mathbb{G}_n$, the empirical measure defined by the vector $\boldsymbol{\mu}$, and conditional on $M_i = m$, $Y_i \sim F_m$.

Proposition 1.1. *The best simple and permutation invariant rules are given by*

(i) $\Delta^{PI}_{\boldsymbol{\mu}}(Y_{1:n}) = E_{\boldsymbol{\mu}}\big(M_{1:n}\big|Y_{1:n}\big)$.
(ii) $\Delta^{S}_{\boldsymbol{\mu}}(Y_{1:n}) = (E_{\boldsymbol{\mu}}(M_1|Y_1), \ldots, E_{\boldsymbol{\mu}}(M_n|Y_n)$.
(iii) $r^S_n = r^{PI}_n + \mathrm{E}_{\boldsymbol{\mu}} \|\Delta^S_{\boldsymbol{\mu}} - \Delta^{PI}_{\boldsymbol{\mu}}\|^2$.

Proof. We need only to give the standard proof of the third part. First, note that by invariance $\Delta^{PI}_{\boldsymbol{\mu}}$ is an equalizer (over all the permutations of $\boldsymbol{\mu}$), and hence $\mathrm{E}_{\mu_{1:n}}(\Delta^{PI}_{\boldsymbol{\mu}} - \mu_{1:n})^2 = \mathrm{E}_{\boldsymbol{\mu}}(\Delta^{PI}_{\boldsymbol{\mu}} - M_{1:n})^2$. Also $\mathrm{E}_{\mu_{1:n}}(\Delta^S_{\boldsymbol{\mu}} - \mu_{1:n})^2 = \mathrm{E}_{\boldsymbol{\mu}}(\Delta^S_{\boldsymbol{\mu}} - M_{1:n})^2$. Then, given the above joint distribution,

$$\begin{aligned} r^S_n &= \mathrm{E}_{\boldsymbol{\mu}} \|\Delta^S_{\boldsymbol{\mu}} - M_{1:n}\|^2 \\ &= \mathrm{E}_{\boldsymbol{\mu}}\, \mathrm{E}_{\boldsymbol{\mu}}\{\|\Delta^S_{\boldsymbol{\mu}} - M_{1:n}\|^2 | Y_{1:n}\} \\ &= \mathrm{E}_{\boldsymbol{\mu}}\, \mathrm{E}_{\boldsymbol{\mu}}\{\|\Delta^S_{\boldsymbol{\mu}} - \Delta^{PI}_{\boldsymbol{\mu}}\|^2 + \|\Delta^{PI}_{\boldsymbol{\mu}} - M_{1:n}\|^2 | Y_{1:n}\} \\ &= r^{PI}_n + \mathrm{E}_{\boldsymbol{\mu}} \|\Delta^S_{\boldsymbol{\mu}} - \Delta^{PI}_{\boldsymbol{\mu}}\|^2. \end{aligned}$$

We now briefly review some related literature and problems. On simple symmetric functions, compound decision and its relation to empirical Bayes, see Samuel (1965), Copas (1969), Robbins (1983), Zhang (2003), among many other papers.

Hannan and Robbins (1955) formulated essentially the same equivalence problem in testing problems, see their Section 6. They show for a special case an equivalence up to $o(n)$ difference in the 'total risk' (i.e., non-averaged risk). Our results for estimation under squared loss are stated in terms of the total risk and we obtain $O(1)$ difference.

Our results have a strong conceptual connection to De Finetti's Theorem. The exchangeability induced on $M_1, \ldots, M_n$, by the Haar measure, implies 'asymptotic independence' as in De Finetti's theorem, and consequently asymptotic independence of $Y_1, \ldots, Y_n$. Thus we expect $E(M_1|Y_1)$ to be asymptotically similar to $E(M_1|Y_1, \ldots, Y_n)$. Quantifying this similarity as n grows, has to do with the rate of convergence in De Finetti's theorem. Such rates were established by Diaconis and Freedman (1980), but are not directly applicable to obtain our results.

After quoting a simple result in the following section, we consider in Section 3 the special important, but simple, case of two-valued parameter. In Section 4 we obtain a strong result under strong conditions. Finally, the main result is given in Section 5, it covers the two preceding cases, but with some price to pay for the generality.

2. Basic Lemma and Notation

The following lemma is standard in comparison of experiments theory; for background on comparison of experiments in testing see Lehmann (1986), p-86. The proof follows a simple application of Jensen's inequality.

Lemma 2.1. *Consider two pairs of distributions,* $\{G_0, G_1\}$ *and* $\{\tilde{G}_0, \tilde{G}_1\}$*, such that the first pair represents a weaker experiment in the sense that there is a Markov kernel* $\mathbb{K}$*, and* $G_i(\cdot) = \int \mathbb{K}(y, \cdot)\, d\tilde{G}_i(y)$*,* $i = 1, 2$*. Then*

$$\mathrm{E}_{G_0}\, \psi\big(\frac{dG_1}{dG_0}\big) \le \mathrm{E}_{\tilde{G}_0}\, \psi\big(\frac{d\tilde{G}_1}{d\tilde{G}_0}\big)$$

for any convex function ψ

For simplicity denote $f_i(\cdot) = f_{\mu_i}(\cdot)$, and for any random variable X, we may write $X \sim g$ if g is its density with respect to a certain dominating measure. Finally, for simplicity we use the notation y_{-i} to denote the sequence $y_1, \ldots, y_n$ without its i member, and similarly $\boldsymbol{\mu}_{-i} = \{\mu_1, \ldots, \mu_n\} \setminus \{\mu_i\}$. Finally $f_{-i}(Y_{-j})$ is the marginal density of Y_{-j} under the model (1.1) conditional on $M_j = \mu_i$.

3. Two Valued Parameter

We suppose in this section that μ can get one of two values which we denote by $\{0, 1\}$. To simplify notation we denote the two densities by f_0 and f_1.

Theorem 3.1. *Suppose that either of the following two conditions holds:*

(i) $f_{1-\mu}(Y_1)/f_\mu(Y_1)$ *has a finite variance under both* $\mu \in \{0, 1\}$*.*
(ii) $\sum_{i=1}^n \mu_i/n \to \gamma \in (0, 1)$*, and* $f_{1-\mu}(Y_1)/f_\mu(Y_1)$ *has a finite variance under one of* $\mu \in \{0, 1\}$*.*

Then $\mathrm{E}_{\boldsymbol{\mu}} \|\hat{\mu}^S - \hat{\mu}^{PI}\|^2 = O(1)$*.*

Proof. Suppose condition (i) holds. Let $K = \sum_{i=1}^n \mu_i$, and suppose, WLOG, that $K \le n/2$. Consider the Bayes model of (1.1). By Bayes Theorem

$$P(M_1 = 1|Y_1) = \frac{K f_1(Y_1)}{K f_1(Y_1) + (n-K) f_0(Y_1)}.$$

On the other hand

$$\begin{aligned}
&P(M_1 = 1|Y_{1:n}) \\
&= \frac{K f_1(Y_1) f_{K-1}(Y_{2:n})}{K f_1(Y_1) f_{K-1}(Y_{2:n}) + (n-K) f_0(Y_1) f_K(Y_{2:n})} \\
&= \frac{K f_1(Y_1)}{K f_1(Y_1) + (n-K) f_0(Y_1)} \Big(1 + \frac{(n-K) f_0(Y_1)}{K f_1(Y_1) + (n-K) f_0(Y_1)} \big(\frac{f_K}{f_{K-1}}(Y_{2:n}) - 1\big)\Big)^{-1} \\
&= P(M_1 = 1|Y_1) \Big(1 + \gamma \big(\frac{f_K}{f_{K-1}}(Y_{2:n}) - 1\big)\Big)^{-1},
\end{aligned}$$

where, with some abuse of notation $f_k(Y_{2:n})$ is the joint density of $Y_{2:n}$ conditional on $\sum_{j=2}^{n}\mu_j = k$, and the random variable γ is in $[0,1]$. We prove now that $f_K/f_{K-1}(Y_{2:n})$ converges to 1 in the mean square.

We use Lemma 2.1 (with ψ the square) to compare the testing of $f_K(Y_{2:k})$ vs. $f_{K-1}(Y_{2:k})$ to an easier problem, from which the original problem can be obtained by adding a random permutation. Suppose for simplicity and WLOG that in fact $Y_{2:K}$ are i.i.d. under f_1, while $Y_{K+1:n}$ are i.i.d. under f_0. Then we compare

$$g_{K-1}(Y_{2:n}) = \prod_{j=2}^{K} f_1(Y_j) \prod_{j=K+1}^{n} f_0(Y_j),$$

the true distribution, to the mixture

$$g_K(Y_{2:n}) = g_{K-1}(Y_{2:n})\frac{1}{n-K}\sum_{j=K+1}^{n}\frac{f_1}{f_0}(Y_j).$$

However, the likelihood ratio between g_K and g_{K-1} is a sum of $n-K$ terms, each with mean 1 (under g_{K-1}) and finite variance. The ratio between the gs is, therefore, $1+O_p(n^{-1/2})$ in the mean square. By Lemma 2.1, this applies also to the fs' ratio.

Consider now the second condition. By assumption, K is of the same order as n, and we can assume, WLOG, that the f_1/f_0 has a finite variance under f_0. With this understanding, the above proof holds for the second condition.

The condition of the theorem is clearly satisfied in the normal shift model: $F_i = N(\mu_i, 1)$, $i = 1, 2$. It is satisfied for the normal scale model, $F_i = N(0, \sigma_i^2)$, $i = 1, 2$, if K is of the same order as n, or if $\sigma_0^2/2 < \sigma_1^2 < 2\sigma_0^2$.

4. Dense $\boldsymbol{\mu}$'s

We consider now another simple case in which $\boldsymbol{\mu}$ can be ordered $\mu_{(1)}, \ldots, \mu_{(n)}$ such that the difference $\mu_{(i+1)} - \mu_{(i)}$ is uniformly small. This will happen if, for example, $\boldsymbol{\mu}$ is in fact a random sample from a distribution with density with respect to Lebesgue measure, which is bounded away from 0 on its support, or more generally, if it is sampled from a distribution with short tails. Denote by $Y_{(1)}, \ldots, Y_{(n)}$ and $f_{(1)}, \ldots, f_{(n)}$ the Ys and fs ordered according to the μs.

We assume in this section

(B1) For some constants A_n and V_n which are bounded by a slowly converging to infinite sequence:

$$\max_{i,j}|\mu_i - \mu_j| = A_n,$$
$$\mathrm{Var}\Big(\frac{f_{(j+1)}}{f_{(j)}}(Y_{(j)})\Big) \le \frac{V_n}{n^2}.$$

Note that condition **(B1)** holds for both the normal shift model and the normal scale model, if $\boldsymbol{\mu}$ behaves like a sample from a distribution with a density as above.

Theorem 4.1. *If Assumption **(B1)** holds then*

$$\sum_{i=1}^{n}|\hat{\mu}_i^{PI} - \hat{\mu}_i^{S}|^2 = O_p(A_n^2 V_n^2/n).$$

Proof. By definition

$$\hat{\mu}_1^S = \frac{\sum_{i=1}^n \mu_i f_i(Y_i)}{\sum_{i=1}^n f_i(Y_i)},$$
$$\hat{\mu}_1^{PI} = \frac{\sum_{i=1}^n \mu_i f_i(Y_1) f_{-i}(Y_{2:n})}{\sum_{i=1}^n f_i(Y_1) f_{-i}(Y_{2:n})},$$

where f_{-i} is the density of $Y_{2:n}$ under $\boldsymbol{\mu}_{-i}$:

$$f_{-i}(y_{2:m}) = \frac{1}{(n-1)!} \sum_{\pi \in \mathcal{P}(2:n)} \prod_{j=2}^n f_{\pi(j)}(y_j) \frac{f_1}{f_i}(y_i).$$

The result will follow if we argue that

$$|\mu_1^{PI} - \mu_1^S| \le \max_{i,j} |\mu_i - \mu_j| \big(\max_{i,j} \frac{f_{-i}}{f_{-j}}(Y_{2:n}) - 1\big) = O_p(A_n V_n/n). \tag{4.1}$$

That is, $\max_i |f_{-i}(Y_{2:n})/f_{-1}(Y_{2:n}) - 1| = O_p(V_n/n)$. In fact we will establish a slightly stronger claim that $\|f_{-i} - f_{-1}\|_{TV} = O_p(V_n/n)$, where $\|\cdot\|_{TV}$ denotes the total variation norm.

We will bound this distance by the distance between two other densities. Let $g_{-1}(y_{2:n}) = \prod_{j=2}^n f_j(y_j)$, the true distribution of $Y_{2:n}$. We define now a similar analog of f_{-i}. Let r_j and $y_{(r_j)}$ be defined by $f_j = f_{(r_j)}$ and $y_{(r_j)} = y_j$, $j = 1, \dots, n$. Suppose, for simplicity, that $r_i < r_1$. Let

$$g_{-i}(y_{2:n}) = g_{-1}(y_{2:n}) \prod_{j=r_i}^{r_1 - 1} \frac{f_{(j+1)}}{f_{(j)}}(y_{(j)}).$$

The case $r_1 < r_i$ is defined similarly. Note that g_{-i} depends only on $\boldsymbol{\mu}_{-i}$. Moreover, if $\tilde{Y}_{2:n} \sim g_{-j}$, then one can obtain $Y_{2:n} \sim f_{-j}$ by the Markov kernel that takes $\tilde{Y}_{2:n}$ to a random permutation of itself. It follows from Lemma 2.1

$$\begin{aligned} \|f_{-i} - f_{-1}\|_{TV} &\le \|g_{-i} - g_{-1}\|_{TV} \\ &= \mathrm{E}_{\mu_{2:n}} \Big| \frac{g_{-i}}{g_{-1}}(Y_{2:n}) - 1 \Big| \\ &= \mathrm{E}_{\mu_{2:n}} \Big| \prod_{j=k}^{r_1 - 1} \frac{f_{(j+1)}}{f_{(j)}}(Y_{(j)}) - 1 \Big| \end{aligned}$$

But, by assumption

$$R_k = \prod_{j=k}^{r_1 - 1} \frac{f_{(j+1)}}{f_{(j)}}(Y_{(j)})$$

is a reversed L_2 martingale, and it follows from Assumption **(B1)** that

$$\max_{k < r_1} |R_k - 1| = O_p(A_n V_n/n).$$

Similar argument applies to i, $r_i > r_1$, yielding

$$\max_i \|f_{-i} - f_{-1}\|_{TV} = O_p(A_n V_n/n)$$

We established (4.1). The theorem follows.

5. Main Result

We assume:

(G1) For some $C < \infty$: $\max_{i\in\{1,\ldots,n\}} |\mu_i| < C$, and $\max_{i,j\in 1,\ldots,n} \mathrm{E}_{\mu_i}(f_{\mu_j}(Y_1)/f_{\mu_i}(Y_1))^2 < C$. Also, there is $\gamma > 0$ such that $\min_{i,j\in 1,\ldots,n} P_{\mu_i}(f_{\mu_j}(Y_1)/f_{\mu_i}(Y_1) > \gamma) \geq 1/2$.

(G2) The random variables

$$p_j(Y_i) = \frac{f_j(Y_i)}{\sum_{k=1}^n f_k(Y_i)}, \quad i,j = 1,\ldots,n.$$

are bounded in expectation by

$$\mathrm{E}\sum_{i=1}^n \sum_{j=1}^n \big(p_j(Y_i)\big)^2 < C$$

$$\sum_{i=1}^n \mathrm{E}\,\frac{1}{n\min_j p_j(Y_i)} < Cn$$

$$\mathrm{E}\sum_{i=1}^n \frac{\sum_{j=1}^n \big(p_j(Y_i)\big)^2}{n\min_j p_j(Y_i)} < C.$$

Both assumptions describe a situation where the μs do not "separate". They cannot be too far one from another, geometrically or statistically (Assumption **(G1)**), and they are dense in the sense that each Y can be explained by many of the μs (Assumption **(G2)**). The conditions hold for the normal shift model if $\boldsymbol{\mu}_n$ are uniformly bounded: Suppose the common variance is 1 and $|\mu_j| < A_n$. Then

$$\begin{aligned}
\mathrm{E}\sum_{j=1}^n \Big(\frac{f_j(Y_1)}{\sum_{k=1}^n f_k(Y_1)}\Big)^2 &= \mathrm{E}\,\frac{\sum_{j=1}^n f_j^2(Y_1)}{(\sum_{k=1}^n f_k(Y_1))^2} \\
&\leq \mathrm{E}\,\frac{ne^{-Y_1^2+2A_n|Y_1|-A_n^2}}{(ne^{-(Y_1^2-2A_n|Y_1|+A_n^2)/2})^2} \\
&= \frac{1}{n}\mathrm{E}\,e^{4A_n|Y_1|} \\
&= \frac{1}{n}\Big(e^{8A_n^2+4A_n\mu_1} + e^{8A_n^2-4A_n\mu_1}\Big) \leq \frac{2}{n}e^{12A_n^2}.
\end{aligned}$$

and the first part of **(G2)** hold. The other parts follow a similar calculations.

Theorem 5.1. *Assume that **(G1)** and **(G2)** hold. Then*

(i) $$\mathrm{E}\,\|\Delta_{\boldsymbol{\mu}}^S - \Delta_{\boldsymbol{\mu}}^{PI}\|^2 = O(1)$$

(ii) $$r_n^S - r_n^{PI} = O(1).$$

Corollary 5.2. *Suppose $\mathcal{F} = \{N(\mu,1) : |\mu| < c\}$ for some $c < \infty$, then the conclusions of the theorem follow.*

Proof. It was mentioned already in the introduction that when we are restricted to permutation invariant procedure we can consider the Bayesian model under which $(\pi, Y_{1:n})$, π a random permutation, have a distribution given by (1.1). Fix now $i \in \{1,\ldots,n\}$. Under this model we want to compare

$$\mu_i^S = E(\mu_{\pi(i)}|Y_i), \quad i = 1,\ldots,n$$

to

$$\mu_i^{PI} = E(\mu_{\pi(i)}|Y_{1:n}), \quad i = 1, \ldots, n.$$

More explicitly:

$$\begin{aligned} \mu_i^S &= \frac{\sum_{j=1}^n \mu_j f_j(Y_i)}{\sum_{j=1}^n f_j(Y_i)} \\ &= \sum_{j=1}^n \mu_j p_j(Y_i), \quad i = 1, \ldots, n \\ \mu_i^{PI} &= \frac{\sum_{j=1}^n \mu_j f_j(Y_i) f_{-j}(Y_{-i})}{\sum_{j=1}^n f_j(Y_i) f_{-j}(Y_{-i})} \\ &= \sum_{j=1}^n \mu_j p_j(Y_i) W_j(Y_{-i}, Y_i), \quad i = 1, \ldots, n, \end{aligned} \tag{5.1}$$

where for all $i, j = 1, \ldots, n$, $f_j(Y_i)$ was defined in Section 2, and

$$\begin{aligned} p_j(Y_i) &= \frac{f_j(Y_i)}{\sum_{k=1}^n f_k(Y_i)}, \\ W_j(Y_{-i}, Y_i) &= \frac{f_{-j}(Y_{-i})}{\sum_{k=1}^n p_k(Y_i) f_{-k}(Y_{-i})} \end{aligned}$$

Note that $\sum_{k=1}^n p_k(Y_i) = 1$, and $W_j(Y_{-i}, Y_i)$ is the likelihood ratio between two (conditional on Y_i) densities of Y_{-i}, say g_{j0} and g_1. Consider two other densities (again, conditional on Y_i):

$$\begin{aligned} \tilde{g}_{j0}(Y_{-i}|Y_i) &= f_i(Y_j) \prod_{m \neq i,j} f_m(Y_m), \\ \tilde{g}_{j1}(Y_{-i}|Y_i) &= \tilde{g}_{j0}(Y_{-i}|Y_i) \Big(\sum_{k \neq i,j} p_k(Y_i) \frac{f_j}{f_k}(Y_k) + p_i(Y_i) \frac{f_j}{f_i}(Y_j) + p_j(Y_i) \Big) \end{aligned}$$

Note that $g_{j0} = \tilde{g}_{j0} \circ \mathbb{K}$ and $g_1 = \tilde{g}_{j1} \circ \mathbb{K}$, where $\mathbb{K}$ is the Markov kernel that takes Y_{-i} to a random permutation of itself. It follows from Lemma 2.1 that

$$\begin{aligned} \mathrm{E}\big(|W_j(Y_{-i}, Y_i) - 1|^2 \big| Y_i\big) &\leq \mathrm{E}_{\tilde{g}_{j1}} \Big(\frac{\tilde{g}_{j0}}{\tilde{g}_{j1}} - 1 \Big)^2 \\ &= \mathrm{E}_{\tilde{g}_{j0}} \Big(\frac{\tilde{g}_{j0}}{\tilde{g}_{j1}} - 2 + \frac{\tilde{g}_{j1}}{\tilde{g}_{j0}} \Big). \end{aligned} \tag{5.2}$$

This expectation does not depend on i except for the value of Y_i. Hence, to simplify notation, we take WLOG $i = j$. Denote

$$\begin{aligned} L &= \frac{\tilde{g}_{j1}}{\tilde{g}_{j0}} = p_j(Y_j) + \sum_{k \neq i} p_k(Y_j) \frac{f_j}{f_k}(Y_k) \\ V &= \frac{n}{4} \gamma \min_k p_k(Y_j), \end{aligned}$$

where γ is as in **(G1)**. Then by (5.2)

$$\begin{aligned}\mathrm{E}\big(|W_j(Y_{-j},Y_j)-1|^2\big|Y_j\big) &\le \mathrm{E}_{\tilde{g}_{j0}}\Big(\frac{1}{L}-2+L\Big)\\ &= \mathrm{E}_{\tilde{g}_{j0}}\frac{(L-1)^2}{L}\\ &\le \frac{1}{V}\mathrm{E}_{\tilde{g}_{j0}}(L-1)^2\mathbb{1}(L>V)+\mathrm{E}_{\tilde{g}_{j0}}\frac{\mathbb{1}(L\le V)}{L}\\ &\le \mathrm{E}_{\tilde{g}_{j0}}\frac{\mathbb{1}(L\le V)}{L}+\frac{1}{V}\sum_{k=1}^{n}p_k^2(Y_j),\end{aligned} \tag{5.3}$$

by **G1**. Bound

$$L\ge\gamma\min_k p_k(Y_j)\sum_{k=1}^{n}\mathbb{1}(\frac{f_j}{f_k}(Y_k)>\gamma)\ge\gamma\min_k p_k(Y_j)(1+U),$$

where $U\sim B(n-1,1/2)$ (the 1 is for the ith summand). Hence

$$\begin{aligned}\mathrm{E}_{\tilde{g}_{j0}}\frac{\mathbb{1}(L\le V)}{L} &\le \frac{1}{\gamma\min_k p_k(Y_j)}\sum_{k=0}^{\lceil n/4\rceil}\frac{1}{k+1}\binom{n-1}{k}2^{-n+1}\\ &= \frac{1}{\gamma n\min_k p_k(Y_j)}\sum_{k=0}^{\lceil n/4\rceil}\binom{n}{k+1}2^{-n+1}\\ &= O(e^{-n})\frac{1}{\gamma n\min_k p_k(Y_j)}\end{aligned} \tag{5.4}$$

by large deviation.

From **(G1)**, **(G2)**, (5.1), (5.3), and (5.4):

$$\begin{aligned}\mathrm{E}\,\mathrm{E}\big((\mu_i^S-\mu_i^{PI})^2\big|Y_i\big) &= \mathrm{E}\,\mathrm{E}\Big(\Big(\sum_{j=1}^{n}\mu_j p_j(Y_i)\big(W_j(Y_{-i},Y_i)-1\big)\Big)^2\Big|Y_i\Big)\\ &\le \max_j|\mu_j|^2\mathrm{E}\sum_{j=1}^{n}p_j(Y_i)\mathrm{E}\big(W_j(Y_{-i},Y_i)-1\big)^2\Big)\Big|Y_i\Big)\\ &\le \kappa C^3/n,\end{aligned}$$

for some κ large enough. Claim (i) of the theorem follows. Claim (ii) follows (i) by Proposition 1.1.

References

[1] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2^{nd} edition. Springer-Verlag, New York.

[2] Brown, L. D. and Greenshtein, E. (2007). Non parametric empirical Bayes and compound decision approaches to estimation of a high dimensional vector of normal means. Manuscript.

[3] Copas, J. B. (1969). Compound decisions and empirical Bayes (with discussion). *JRSSB* **31** 397–425.

[4] Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *Ann.Prob.*, **8.4**, 745–764.

[5] Hannan, J. F. and Robbins, H. (1955). Asymptotic solutions of the compound decision problem for two completely specified distributions. *Ann.Math.Stat.*, **26.1**, 37–51.

[6] Lehmann, E. L. (1986). *Testing Statistical Hypothesis*, 2^{nd} edition. Wiley & Sons, New York.

[7] Robbins, H. (1951). Asymptotically subminimax solutions of compound decision problems. *Proc. Third Berkeley Symp.*, 157–164.
[8] Robbins, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Stat.*, **11**, 713–723.
[9] Samuel, E. (1965). On simple rules for the compound decision problem. *JRSSB*, **27**, 238–244.
[10] Zhang, C. H. (2003). Compound decision theory and empirical Bayes methods. (Invited paper. *Ann. Stat.*, **31**, 379–390.
[11] Wenuha, J. and Zhang, C. H. (2007) General maximum likelihood empirical Bayes estimation of normal means. Manuscript.

IMS Collections
Vol. 0 (2009) 276–303
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000002

Large Sample Statistical Inference for Skew-Symmetric Families on the Real Line

Rolando Cavazos–Cadena[1,*]
and Graciela González–Farías[2,*]

Universidad Autónoma Agraria Antonio Narro
and Centro de Investigación en Matemáticas A. C.

Abstract: For a general family of one-dimensional skew-symmetric probability densities, the application of the maximum likelihood method to the estimation of the asymmetry parameter λ is studied. Under mild conditions, the existence and consistency of a sequence $\{\hat{\lambda}_n\}$ of maximum likelihood estimators is established, and the limit distributions of $\{\hat{\lambda}_n\}$ and the sequence of likelihood ratios are determined under the null hypothesis $\mathcal{H}_0 : \lambda = 0$. These latter conclusions, which hold under differential singularity of the likelihood function at $\lambda = 0$, extend to the present framework results recently obtained for general statistical models with null Fisher information.

Dedicated to Professor O. Hernández-Lerma, on the occasion of his sixtieth birthday

Contents

1 Introduction 276
2 Identifiability 279
3 Existence of Maximum Likelihood Estimators 281
4 Consistency 285
5 Asymptotic Distribution 289
6 Technical Preliminaries 292
7 Proof of Theorem 5.1 299
References 302

1. Introduction

This work concerns likelihood inference for the general one-dimensional skew-symmetric family, which is constructed as follows: Given two symmetric densities

[1]Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro, Buenavista, Saltillo COAH, 25315, **MÉXICO**, email: rcavazos@uaaan.mx
[2]Centro de Investigación en Matemáticas A. C., Apartado Postal 402, Guanajuato, GTO, 36240, **MÉXICO**, email: farias@cimat.mx
*This research was supported by the PSF Organization under Grant 2007-3, and by CONACYT under Grants 25357 and 45974-F.
AMS 2000 subject classifications: Primary 62F10; secondary 62F12
Keywords and phrases: Boundedness of maximum likelihood estimators, Kullback's inequality, Lateral Taylor series, Strong law of large numbers, Central limit theorem, Asymptotic normality

f and g on the real line—that is, $f(x) = f(-x)$ and $g(x) = g(-x)$ for all $x \in \mathbb{R}$—let

$$G(x) := \int_{-\infty}^{x} g(z)\, dz \tag{1.1}$$

be the cumulative distribution function of density g. Notice that $\int_{\mathbb{R}}[1 - G(\lambda w)]f(w)\, dw = \int_{\mathbb{R}}[1 - G(-\lambda w)]f(-w)\, dw = \int_{\mathbb{R}} G(\lambda w)f(w)\, dw$ for every $\lambda \in \mathbb{R}$, so that $\int_{\mathbb{R}} 2f(w)G(\lambda w)\, dw = 1$. This argument, due to Azzalini (1985), shows that for each $\lambda \in \mathbb{R}$

$$\rho(x; \lambda) := 2f(x)G(\lambda x), \quad x \in \mathbb{R} \tag{1.2}$$

is a genuine density, and the collection

$$S(f, g) := \{\rho(\cdot; \lambda) \,|\, \lambda \in \mathbb{R}\}, \tag{1.3}$$

is the skew-symmetric family determined by f and g. When the parametrization $\lambda \mapsto \rho(\cdot; \lambda)$ is one-to-one, $f(\cdot) = \rho(\cdot; 0)$ is the unique symmetric density in $S(f, g)$, and λ can be considered as a measure of the asymmetry (or skewness) of density $\rho(\cdot; \lambda)$. The first systematic treatment of a skew-symmetric family was presented in Azzalini (1985, 1986) for the case in which f and g coincide with the standard normal density φ, and location-scale and regression models based on $S(\varphi, \varphi)$ as well as multivariate extensions have been intensively studied during the last twenty years; see Azzalini and Dalla Valle (1996), Azzalini and Capitanio (1999), Pewsey (2000), Genton (2004) and the references therein. The analysis of the maximum likelihood method for the location-scale model based on $S(\varphi, \varphi)$ has proved most challenging since, in that context, the Fisher information matrix has incomplete rank at $\lambda = 0$, a problem that also arises for the skew exponential family, which includes the skew normal location-scale model as a particular case (Azzalini, 1986, DiCiccio and Monti, 2004). The singularity of the information matrix has been analyzed via the centered parametrization introduced in Azzalini (1985) and asymptotic results are based on the recent work by Rotzintzky *et al.* (2000). Using the conclusions in this latter paper, rates of convergence for maximum likelihood estimators are derived in Chiogna (2005), and Sartori (2006) studied the finiteness of the estimator of the asymmetry parameter.

As suggested by the previous comments, family $S(\varphi, \varphi)$ has been intensively studied, and a great effort has been done on generalizations of that model (Genton, 2004), so that looking for inference results applicable to a broad class of skew families is, certainly, an interesting problem. This work is *a first step* in this direction since, although no scale or location parameters will be introduced, the maximum likelihood method applied to the estimation of the asymmetry parameter λ will be studied under rather *minimal conditions* on densities f and g, making the likelihood inference problem a very interesting one. *The first objective* of this note is

(i) To establish the existence of a sequence $\{\hat{\lambda}_n\}$ of maximum likelihood estimators of λ and to prove its consistency.

To see the interest behind this problem, denote the (kernel log-) likelihood corresponding to a single observation x by

$$\ell(\lambda; x) := \log(G(\lambda x)), \tag{1.4}$$

and observe that if g is the standard normal density, then $\ell(\cdot; x)$ is strictly concave for $x \neq 0$, a property that yields the existence and consistency of maximum likelihood estimators (Newey and McFadden, 1993). However, strict concavity of $\ell(\cdot; x)$

is far from being a general property, and may fail in common cases, for instance, if g is the Laplace density. In this work, problem (i) above will be studied under the minimal assumption that the parametrization $\lambda \mapsto \rho(\cdot;\lambda)$ is identifiable and, since the parameter space is not compact for the $S(f,g)$ family, to the best of the authors' knowledge, in this context the existence and consistency of maximum likelihood estimators can not be directly obtained form general available results. The second problem studied in this work concerns the asymptotic distribution of $\{\hat{\lambda}_n\}$ and, as usual, the analysis below requires differentiability assumptions on $\ell(\cdot;x)$, and then, on density g. This problem will be studied under conditions allowing g to be non smooth at $x = 0$, which can be roughly described as follows:

A1: g is continuous on $\mathbb{R}$, is 'smooth' outside 0, and has lateral derivatives at zero.

Under this requirement, it is not difficult to see that if the true parameter value, say ν, is non-null, then $\ell(\cdot;x)$ is smooth on a neighborhood of ν if $\ell(\nu;x) < \infty$. Thus, when the information number at ν, given by $\mathcal{I}(\nu) = \int_{\mathbb{R}} (\partial_\lambda \ell(\nu;x))^2 \rho(\nu;x)\,dx (> 0)$, is finite, under standard regularity conditions Wald's classical results yield that $\sqrt{n\mathcal{I}(\nu)}(\hat{\lambda}_n - \nu)$ has a standard normal distribution at the limit (Lehmann and Casella, 1998, Section 6.3, Shao, 1999, Section 4.4). However, under the null hypothesis $\mathcal{H}_0 : \lambda = 0$ such a direct conclusion is not possible, since $g(\cdot)$ is not necessarily differentiable in a neighborhood of zero under condition A1 above. Also, observe that $\partial_\lambda \ell(0;x) = 2g(0)x$, so that $\mathcal{I}(0) = 4g(0)^2 \int_R x^2 f(x)\,dx$; thus, $\mathcal{I}(0)$ is null if $g(0) = 0$ and, again, in this case the asymptotic distribution of $\{\hat{\lambda}_n\}$ can not be obtained from the results in the aforementioned references. The case of null information at $\lambda = 0$ was recently studied in a general context by Rotnitzky *et al.* (2000) under several assumptions, including (a) compactness of the parameter space, and (b) the existence of higher order derivatives $\partial_\lambda^k \ell(\lambda;x)$ in a neighborhood of zero. Since in the present context neither the parameter space is compact, nor higher order derivatives $\partial_\lambda^k \ell(\lambda;x)$ exist around zero, if $g(0) = 0$ then the limiting distribution can not be solved by direct application of available results under condition A1. Therefore, *the second problem* considered in this note is

(ii) to determine both the limit distribution of the (appropriately normalized) sequence $\{\hat{\lambda}_n\}$ under the hypothesis $\mathcal{H}_0 : \lambda = 0$, and the asymptotic null distribution of the likelihood ratio statistic.

The approach used below to study problems (i) and (ii) can be described as follows: The monotonicity of the mapping $\lambda \mapsto \ell(\lambda;x)$ is used to establish the existence of maximizers $\hat{\lambda}_n$ of the observed likelihood when the sample size n is large enough, whereas the proof for the consistency of $\{\hat{\lambda}_n\}$ follows the ideas in Cavazos-Cadena and Gonzalez-Farías (2007) where, under mild conditions, it is shown that sequence $\{\hat{\lambda}_n\}$ is consistent if and only if it is bounded with probability 1; here, after establishing the boundedness property, a direct proof of consistency is given via a simple consequence of Kullback's inequality. Concerning problem (ii), as in the results presented in Lehmann and Casella (1998, Section 6.3), Shao (1999, Section 4.4), or in Rotnitzky *et al.* (2000), the analysis is based on Taylor series expansions for the observed likelihood and its derivative around zero, which in the present context are *lateral expansions*; they are used to show that the maximum likelihood estimator is no null with probability increasing to 1, and the asymptotic distributions are obtained via the central limit theorem and the strong law of large numbers.

The organization of the paper is as follows: Problem (i) is analyzed in the following

three sections. Thus, in Section 2 the identification property of the parametrization $\lambda \mapsto \rho(\lambda;\cdot)$ is discussed, the existence of a sequence $\{\hat{\lambda}_n\}$ of maximum likelihood estimators is established in Section 3, and the consistency of $\{\hat{\lambda}_n\}$ is proved in Section 4. After this point, the reminder of the paper concerns problem (ii). In Section 5 the basic dominance and smoothness assumptions are formally introduced, and the main asymptotic result is stated as Theorem 5.1. Next, in Section 6 the necessary technical tools concerning *lateral Taylor series* for the likelihood function and its first derivative are established and, finally, the exposition concludes in Section 7 with a proof of Theorem 5.1.

NOTATION. Throughout the remainder $X_1, X_2, X_3, \ldots$ stands for a sequence of independent and identically distributed random variables with common density belonging to family $S(f,g)$ in (1.3): For each $n = 1, 2, , \ldots$, set

$$X_1^n := (X_1, \ldots, X_n). \tag{1.5}$$

The distribution of the sequence $(X_1, X_2, \ldots)$ when ν is the true parameter value is denoted by $P_\nu[\,\cdot\,]$, whereas $E_\nu[\,\cdot\,]$ stands for the corresponding expectation operator. On the other hand, if $H(\cdot)$ is a function defined around zero, $H(0+) := \lim_{x \searrow 0} H(x)$, and $H(0-) := \lim_{x \nearrow 0} H(x)$, whereas given $A \subset \mathbb{R}$, I_A is the indicator function of set A, that is, $I_A(x) = 1$ if $x \in A$, whereas $I_A(x) = 0$ when $x \notin A$. Finally, the following convention is enforced: $\sum_{i=a}^{b} C_i = 0$ when $b < a$.

2. Identifiability

In this section the identifiability of the parametrization $\lambda \mapsto \rho(\cdot;\lambda)$ is briefly discussed. This condition establishes that different parameters correspond to different densities, and plays a fundamental role in parametric estimation (Newey and McFadden, 1993).

Assumption 2.1. *The mapping $\lambda \mapsto \rho(\cdot;\lambda)$ is one-to-one, that is,*

$$\int_{\mathbb{R}} |\rho(x;\lambda) - \rho(x;\nu)|\,dx = \int_{\mathbb{R}} 2f(x)|G(\lambda x) - G(\nu x)|\,dx \neq 0, \quad \text{if } \lambda \neq \nu;$$

see (1.2).

Some primitive conditions ensuring this requirement are now given.

Lemma 2.1. *Assumption 2.1 holds under either of the following conditions (i)–(iii).*
(i) For every nonempty open interval $J \subset \mathbb{R}$, $\int_J f(x)\,dx > 0$;
(ii) $\int_J g(x)\,dx > 0$ for each nonempty open interval $J \subset \mathbb{R}$;
(iii) There exists $\delta > 0$ such that

$$\int_J g(x)\,dx > 0 \quad \text{and} \quad \int_J f(x)\,dx > 0,$$

for each nonempty open interval $J \subset (0,\delta)$.

Proof. Let λ and ν be fixed and *different* real numbers, and suppose that

$$\int_{\mathbb{R}} f(x)|G(\lambda x) - G(\nu x)|\,dx = 0. \tag{2.1}$$

A contradiction will be obtained under each of the three conditions in the lemma. *Assume that condition (i) holds.* Given an open interval $J \subset \mathbb{R}$ with positive length, (2.1) and $\int_J f(x)\,dx > 0$ together yield that $|G(\lambda x) - G(\nu x)| = 0$ for some $x \in J$, so that $\{x \,|\, G(\lambda x) - G(\nu x) = 0\}$ is dense in $\mathbb{R}$. Since $G(\cdot)$ is continuous, this set is also closed, so that

$$G(\lambda x) = G(\nu x) \quad \text{for all } x \in \mathbb{R}. \tag{2.2}$$

This fact implies that $\lambda \neq 0$ and $\nu \neq 0$. Indeed, if $\nu = 0$, it follows that $\lambda \neq 0$ and the right-hand side equals $1/2$ for all x, whereas the values of left-hand side cover the whole interval $(0,1)$ as x moves on $\mathbb{R}$. Thus, $\nu \neq 0$ and, similarly, $\lambda \neq 0$. Moreover, (2.2) also yields that λ and ν have the same sign since, otherwise, as $x \to \infty$ one side of the equality converges to 1 and the other converges to 0. Therefore, recalling that $\lambda \neq \nu$, it follows that $|\lambda| \neq |\nu|$, and without loss of generality it can be assumed that $\beta = |\lambda|/|\nu| = \lambda/\nu \in (0,1)$. Replacing x by y/ν, (2.2) yields that, for all $y \in \mathbb{R}$, $G(y) = G(\beta y)$, and then

$$G(y) = G(\beta^n y), \quad y \in \mathbb{R}, \quad n = 1, 2, 3, \ldots.$$

Letting n go to ∞, the continuity of $G(\cdot)$ and the inclusion $\beta \in (0,1)$ yield that $G(y) = G(0) = 1/2$ for all $y \in \mathbb{R}$, in contradiction with the basic properties of a distribution function.
Under condition (ii), $G(b) - G(a) = \int_a^b g(x)\,dx > 0$ for $a < b$, so that $G(\cdot)$ is strictly increasing. Thus, since $\lambda \neq \nu$, $|G(\lambda x) - G(\nu x)| > 0$ for all $x \neq 0$, and it follows that (2.1) is equivalent to $\int_{\mathbb{R}} f(x)\,dx = 0$, which is not possible, since $f(\cdot)$ is a density.
Suppose that condition (iii) occurs. In this context, $G(\cdot)$ is strictly increasing on the interval $(0,\delta)$ and then on $(-\delta,\delta)$, by symmetry. Next, define $\delta_1 := \delta/(|\lambda| + |\nu| + 1)$ and, recalling that $\lambda \neq \nu$, notice that if $x \in [\delta_1/2, \delta_1]$ then λx and νx are different points in $(-\delta,\delta)$, so that $|G(\lambda x) - G(\nu x)| > 0$. Thus, by continuity of $G(\cdot)$, $\min_{x \in [\delta_1/2, \delta_1]} |G(\lambda x) - G(\nu x)| =: \varepsilon > 0$. Consequently,

$$\int_{[\delta_1/2, \delta_1]} f(x)|G(\lambda x) - G(\nu x)|\,dx \geq \varepsilon \int_{[\delta_1/2, \delta_1]} f(x)\,dx > 0,$$

where the inclusion $[\delta_1/2, \delta_1] \subset (0,\delta)$ was used to set the last inequality. Therefore, (2.1) can not occur under condition (iii).

According to the previous result, Assumption 2.1 is valid under mild requirements on densities f and g, and it is interesting to observe that if conditions (i)–(iii) in Lemma 2.1 do not hold, then identifiability may fail.

Example 2.1. *Let the symmetric densities f and g be such that $f(x) = 0$ for $x \in [-1,1]$, whereas $g(x) = 0$ for $|x| > 1$. In this case, it is not difficult to see that the general density $\rho(\cdot;\lambda) \in S(f,g)$ satisfies $\rho(\cdot;\lambda) = \rho(\cdot;1)$ for $\lambda \geq 1$, and $\rho(\cdot;\lambda) = \rho(\cdot;-1)$ when $\lambda \leq -1$, so that Assumption 2.1 does not hold.*

The basic consequence of Assumption 2.1, which plays a central role in the subsequent development, is established in the following lemma. Firstly, recall that $G(\cdot)$ is increasing and notice that if $\lambda_2 > \lambda_1$ then

$$\int_0^\infty f(x)|G(\lambda_2 x) - G(\lambda_1 x)|\,dx = \int_0^\infty f(x)(G(\lambda_2 x) - G(\lambda_1 x))\,dx,$$

and

$$\int_{-\infty}^{0} f(x)|G(\lambda_2 x) - G(\lambda_1 x)|\,dx = \int_{-\infty}^{0} f(x)(G(\lambda_1 x) - G(\lambda_2 x))\,dx = \int_{0}^{\infty} f(x)(G(\lambda_2 x) - G(\lambda_1 x))\,dx,$$

where the second equality comes from $\int_{\mathbb{R}} f(x)G(\lambda_i x)\,dx = 1/2$ for $i = 1, 2$, so that

$$\int_R f(x)|G(\lambda_2 x) - G(\lambda_1 x)|\,dx = 2\int_0^{\infty} f(x)(G(\lambda_2 x) - G(\lambda_1 x))\,dx = 2\int_{-\infty}^{0} f(x)(G(\lambda_1 x) - G(\lambda_2 x))\,dx, \quad \lambda_1 < \lambda_2. \tag{2.3}$$

Lemma 2.2. *Under the identifiability Assumption 2.1, the following assertions (i) and (ii) hold:*
(i) For each $\lambda \in \mathbb{R}$, $\int_0^{\infty} f(x)G(\lambda x)\,dx > 0$ *and* $\int_{-\infty}^{0} f(x)G(\lambda x)\,dx > 0$.
(ii) There exists a function $c : \mathbb{R} \to (0, \infty)$ *such that*

$$P_\lambda[X_1 < -c(\lambda)] > 0 \quad \textit{and} \quad P_\lambda[X_1 > c(\lambda)] > 0, \quad \lambda \in \mathbb{R}.$$

Proof. (i) Notice that

$$0 \le \int_0^{\infty} f(x)G(\lambda_1 x)\,dx \le \int_0^{\infty} f(x)G(\lambda_2 x)\,dx, \quad \lambda_1 < \lambda_2,$$

since $G(\cdot)$ is increasing. Now, suppose that $\int_0^{\infty} f(x)G(\lambda_2 x)\,dx = 0$ for some $\lambda_2 \in \mathbb{R}$. In this case the above display yields $\int_0^{\infty} f(x)G(\lambda_1 x)\,dx = 0$ for every $\lambda_1 \le \lambda_2$, and then,

$$\int_R f(x)|G(\lambda_2 x) - G(\lambda_1 x)|\,dx = 0, \quad \lambda_1 \le \lambda_2,$$

by the first equality in (2.3), contradicting Assumption 2.1; consequently, $\int_0^{\infty} f(x)G(\lambda x)\,dx > 0$ for all $\lambda \in \mathbb{R}$, whereas the other part of the conclusion can be obtained along similar lines.
(ii) By the monotone convergence theorem, as $\varepsilon \searrow 0$, $\int_{\varepsilon}^{\infty} f(x)G(\lambda x)\,dx \nearrow \int_0^{\infty} f(x)G(\lambda x)\,dx$ and $\int_{-\infty}^{-\varepsilon} f(x)G(\lambda x)\,dx \nearrow \int_{-\infty}^{0} f(x)G(\lambda x)\,dx$ for each λ. Therefore, by part (i), there exists $c(\lambda) > 0$ such that $\int_{c(\lambda)}^{\infty} f(x)G(\lambda x)\,dx > 0$ and $\int_{-\infty}^{-c(\lambda)} f(x)G(\lambda x)\,dx > 0$, and the conclusion follows.

3. Existence of Maximum Likelihood Estimators

The objective of this section is to establish the existence of a sequence of maximum likelihood estimators, an idea that is formally stated below. To begin with, given a fixed sample size $n > 0$, for each possible sample $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ define the average (kernel log-)likelihood function $L_n(\cdot; \mathbf{x})$ by

$$L_n(\lambda; \mathbf{x}) := \frac{1}{n}\sum_{k=1}^{n} \ell(\lambda; x_k) = \frac{1}{n}\sum_{k=1}^{n} \log(G(\lambda x_k)), \quad \lambda \in \mathbb{R}, \tag{3.1}$$

(see (1.1) and (1.4)), where the usual convention $\log(0) := -\infty$ is enforced. Since $G(\cdot)$ is continuous and takes values in $[0,1]$, $L_n(\cdot;\cdot)$ is continuous function from $\mathbb{R}\times\mathbb{R}^n$ into $[-\infty,0]$; moreover, for every $Q\subset\{1,2,\ldots,n\}$

$$L_n(\lambda;\mathbf{x}) \le \frac{1}{n}\sum_{i\,:i\in Q}\log(G(\lambda x_i)) \le 0, \quad \lambda\in\mathbb{R}, \quad \emptyset\neq Q. \tag{3.2}$$

Definition 3.1. *Let $\{\lambda_n : \mathbb{R}^n \to \mathbb{R}\}$ be a sequence of (Borel) measurable functions and set*

$$\hat{\lambda}_n := \lambda_n(X_1^n).$$

In this case, $\{\hat{\lambda}_n\}$ is a sequence of maximum likelihood of estimators of λ if

$$P_{\lambda_0}\left[\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty}[L_k(\hat{\lambda}_k;X_1^k)\ge L_k(\lambda;X_1^k) \text{ for all } \lambda]\right] = 1, \quad \lambda_0\in\mathbb{R}. \tag{3.3}$$

Remark 3.1. *(i) By continuity, $L_n(\hat{\lambda}_n;X_1^n)\ge L_n(\lambda;X_1^n)$ occurs for every $\lambda\in\mathbb{R}$ if and only if it holds for each rational number, so that $[L_n(\hat{\lambda}_n;X_1^n)\ge L_n(\lambda;X_1^n),\ \lambda\in\mathbb{R}]$ is an event.*
(ii) In words, $\{\hat{\lambda}_n\}$ is a sequence of maximum likelihood estimators of λ if, with probability 1 and regardless of the true parameter value, $\hat{\lambda}_n$ maximizes the observed average likelihood function $L_n(\cdot;X_1^n)$ whenever n is large enough. The event within brackets in (3.3) is the inferior limit of the events $[L_k(\hat{\lambda}_k;X_1^k)\ge L_k(\lambda;X_1^k),\ \lambda\in\mathbb{R}]$, and when (3.3) holds then, as $k\to\infty$, $P_{\lambda_0}[L_k(\hat{\lambda}_k;X_1^k)\ge L_k(\lambda;X_1^k)$ for all $\lambda]\to 1$; see, for instance, Billingsley (1995, Section 4).

The main objective of this section is to prove the following result.

Theorem 3.1. *Under Assumption 2.1, there exists a sequence of maximum likelihood estimators of λ.*

The proof of this theorem has been divided into three simple lemmas involving the following notation: For each $\mathbf{x}\in\mathbb{R}^n$, set

$$m_n(\mathbf{x}) := \sup_{\lambda\in\mathbb{R}} L_n(\lambda;\mathbf{x}), \tag{3.4}$$

so that (3.1) and (3.2) lead to

$$0 \ge m_n(\mathbf{x}) \ge L_n(0;\mathbf{x}) = -\log(2), \quad \mathbf{x}\in\mathbb{R}^n, \tag{3.5}$$

since $G(0)=1/2$. Next, define

$$\mathcal{M}_n(\mathbf{x}) := \{\nu\in\mathbb{R}\,|\,L_n(\nu;\mathbf{x})=m_n(\mathbf{x})\}, \quad \mathbf{x}\in\mathbb{R}^n, \tag{3.6}$$

which is a closed subset of $\mathbb{R}$, by the continuity of $L_n(\cdot;\mathbf{x})$. As can be seen from the monotonicity of $\ell(\cdot;x)$, the set $\mathcal{M}_n(\mathbf{x})$ may be empty if the observed sample $\mathbf{x}\in\mathbb{R}^n$ does not contain observations of different sign. The first step to the proof of Theorem 3.1 is the following lemma, showing that $\mathcal{M}_n(\mathbf{x})$ is nonempty and compact if $\mathbf{x}\in\mathbb{R}^n$ contains components with opposite signs. For each integer $n\ge 2$, set

$$\mathcal{S}_n := \{\mathbf{x}\in\mathbb{R}^n\,|\,x_ix_j<0 \text{ for some } i \text{ and } j \text{ with } 1\le i\neq j\le n\} \tag{3.7}$$

and observe that $\mathcal{S}_n$ is an open subset of $\mathbb{R}^n$.

Lemma 3.1. *For each integer $n \geq 2$ and $\mathbf{x} \in \mathcal{S}_n$, the set of maximizers $\mathcal{M}_n(\mathbf{x})$ is nonempty and compact.*

Proof. Given $\mathbf{x} \in \mathcal{S}_n$, select indexes i^* and j^* such that $x_{i^*} < 0$ and $x_{j^*} > 0$, so that $\lim_{\lambda\to\infty} \log(G(\lambda x_{i^*})) = -\infty = \lim_{\lambda\to-\infty} \log(G(\lambda x_{j^*}))$. After setting $Q = \{i^*\}$ and $Q = \{j^*\}$ in (3.2), these convergences yield that

$$\lim_{|\lambda|\to\infty} L_n(\lambda; \mathbf{x}) = -\infty,$$

and it follows that the set $\mathcal{M}_n(\mathbf{x})$—consisting of the maximizers of the continuous function $L_n(\cdot; \mathbf{x})$—is nonempty and compact.

Next, for each $n \geq 2$, define $\lambda_n^+ : \mathcal{S}_n \to \mathbb{R}$ by

$$\lambda_n^+(\mathbf{x}) := \max \mathcal{M}_n(\mathbf{x}), \quad \mathbf{x} \in \mathcal{S}_n, . \tag{3.8}$$

so that $\lambda_n^+(\mathbf{x})$ is the largest element in $\mathcal{M}_n(\mathbf{x})$. Observe that $\lambda_n^+(\mathbf{x}) \in \mathcal{M}_n(\mathbf{x})$ is a well-defined finite number for each $\mathbf{x} \in \mathcal{S}_n$, by Lemma 3.1. As it is shown below, this function $\lambda_n^+(\cdot)$ is upper semi-continuous.

Lemma 3.2. *Let the integer $n \geq 2$ be arbitrary but fixed, and suppose that $\{\mathbf{x}_k\} \subset \mathcal{S}_n$ is such that $\lim_{k\to\infty} \mathbf{x}_k = \mathbf{y} \in \mathcal{S}_n$. In this context,*
(i) If $\nu_k \in \mathcal{M}_n(\mathbf{x}_k)$ for each k, then sequence $\{\nu_k\}$ is bounded, and
(ii) Every limit point of $\{\nu_k\}$ belongs to $\mathcal{M}_n(\mathbf{y})$.
Consequently,
(iii) $\lambda_n^+(\cdot)$ is upper semi-continuous.

Proof. To begin with, write $\mathbf{x}_k = (x_{k1}, \ldots, x_{kn})$ and notice that, since $\nu_k \in \mathcal{M}_n(\mathbf{x}_k)$,

$$\frac{1}{n}\sum_{i=1}^{n} \log(G(\lambda x_{ki})) = L_n(\lambda; \mathbf{x}_k) \leq L_n(\nu_k; \mathbf{x}_k) = \frac{1}{n}\sum_{i=1}^{n} \log(G(\nu_k x_{ki})), \tag{3.9}$$

for ever $\lambda \in \mathbb{R}$; in particular,

$$-\log(2) = L_n(0; \mathbf{x}_k) \leq \frac{1}{n}\sum_{i=1}^{n} \log(G(\nu_k x_{ki})). \tag{3.10}$$

(i) *Assume* now that $\limsup_{k\to\infty} \nu_k = \infty$ and, taking a subsequence if necessary, without loss of generality suppose that $\nu_k \to \infty$. In this context, select an index i^* such that $y_{i^*} < 0$, which is possible since $\mathbf{y} \in \mathcal{S}_n$ (see (3.7)) and observe that the convergence $x_{ki^*} \to y_{i^*} < 0$ leads to $\nu_k x_{ki^*} \to -\infty$ as $k \to \infty$, so that $\log(G(\nu_k x_{ki^*}) \to -\infty$, and then, via (3.2) with $Q = \{i^*\}$, this yields

$$\lim_{k\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log(G(\nu_k x_{ki}) = -\infty,$$

which contradicts (3.10); it follows that $\limsup_k \nu_k < \infty$, whereas the inequality $\liminf_k \nu_k > -\infty$ can be established along similar lines.
(ii) Let ν^* be an arbitrary limit point of $\{\nu_k\}$, and notice that, by part (i), ν^* is finite; selecting a subsequence, if necessary, assume that $\nu_k \to \nu^*$. In this case, taking the limit as k goes to ∞ in (3.9), it follows that

$$L_n(\lambda; \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} \log(G(\lambda y_k)) \leq \frac{1}{n}\sum_{i=1}^{n} \log(G(\nu^* y_k)) = L_n(\nu^*; \mathbf{y}), \quad \lambda \in \mathbb{R},$$

i.e., $\nu^* \in \mathcal{M}_n(\mathbf{y})$.
(iii) Let $\mathbf{y} \in \mathcal{S}_n$ be arbitrary. If $\{\mathbf{x}_k\} \subset \mathcal{S}_n$ is such that $\lim_k \mathbf{x}_k = \mathbf{y}$, recalling that $\lambda_n^+(\mathbf{x}_k) \in \mathcal{M}_n(\mathbf{x}_k)$, part (ii) with $\nu_k = \lambda_n^+(\mathbf{x}_k)$ yields that $\limsup_k \lambda_n^+(\mathbf{x}_k) \in \mathcal{M}_n(\mathbf{y})$, and then $\limsup_k \lambda_n^+(\mathbf{x}_k) \leq \lambda_n^+(\mathbf{y})$, by (3.8), so that $\lambda_n^+(\cdot)$ is upper semi-continuous.

The last step before the proof of Theorem 3.1 is the following consequence of Lemma 2.2(i).

Lemma 3.3. *Under Assumption 2.1,*

$$\lim_{n\to\infty} P_\nu \left[\bigcap_{k=n}^{\infty} [X_1^k \in \mathcal{S}_k] \right] = 1, \quad \nu \in \mathbb{R};$$

see (1.5) for notation.

Proof. Let $\nu \in \mathbb{R}$ be fixed and observe that for every $i = 1, 2, \ldots,$

$$P_\nu[X_i \leq 0] = 1 - P_\nu[X_i > 0] = 1 - 2\int_0^\infty f(x)G(\nu x)\,dx =: \rho_-(\nu) \in [0,1),$$

where the inclusion stems from Lemma 2.2(i), so that for each $k > 0$, $\rho_-(\nu)^k = P_\nu[X_i \leq 0,\ 1 \leq i \leq k]$; similarly, $P_\nu[X_i \geq 0,\ 1 \leq i \leq k] = \rho_+(\nu)^k$ for some $\rho_+(\nu) \in [0,1)$. Since

$$[X_1^k \in \mathcal{S}_k]^c = [X_i \leq 0,\ 1 \leq i \leq k] \cup [X_i \geq 0,\ 1 \leq i \leq k],$$

it follows that $P_\nu\left[[X_1^k \in \mathcal{S}_k]^c\right] \leq 2\rho(\nu)^k$, where $\rho(\nu) \in [0,1)$ is given by $\rho(\nu) := \max\{\rho_-(\nu), \rho_+(\nu)\}$. Observing that (1.5) and (3.7) together yield that $[X_1^n \in \mathcal{S}_n] \subset [X_1^k \in \mathcal{S}_k]$ for $k \geq n$, it follows that

$$P_\nu \left[\bigcap_{k=n}^{\infty} [X_1^k \in \mathcal{S}_k] \right] \geq P_\nu\left[X_1^n \in \mathcal{S}_n\right] \geq 1 - 2\rho(\nu)^n$$

and the conclusion is obtained taking the limit as $n \to \infty$.

Notice that the last display and Lemma 3.1 together shows that, with probability increasing to 1 at a geometric rate, the function $L_n(\cdot; X_1^n)$ has a maximizer when n is large enough.

Proof of Theorem 3.1. Select a point $\lambda^* \in \mathbb{R}$ and for each positive integer k define $\lambda_k : \mathbb{R}^k \to \mathbb{R}$ as follows: $\lambda_1(\cdot) \equiv \lambda^*$, whereas, for $k \geq 2$, $\lambda_k(\mathbf{x}) := \lambda^*$ if $\mathbf{x} \in \mathbb{R}^k \setminus \mathcal{S}_k$, and $\lambda_k(\mathbf{x}) := \lambda_k^+(\mathbf{x})$ if $\mathbf{x} \in \mathcal{S}_k$. Since the $\mathcal{S}_k$'s are open sets, Lemma 3.2(iii) implies that each function $\lambda_k(\cdot)$ is measurable, and then $\hat{\lambda}_k = \lambda_k(X_1^k)$ is a genuine statistic. Using (3.4)–(3.6) and (3.8), this specification yields $[X_1^k \in \mathcal{S}_k] \subset [L_k(\hat{\lambda}_k; X_1^k) \geq L_k(\lambda; X_1^k), \lambda \in \mathbb{R}]$ for $k \geq 2$. Therefore, for each integer $m \geq 2$,

$$\begin{aligned}\bigcap_{k=m}^{\infty} [X_1^k \in \mathcal{S}_k] &\subset \bigcap_{k=m}^{\infty} [L_k(\hat{\lambda}_k; X_1^k) \geq L_k(\lambda; X_1^k), \lambda \in \mathbb{R}] \\ &\subset \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} [L_k(\hat{\lambda}_k; X_1^k) \geq L_k(\lambda; X_1^k), \lambda \in \mathbb{R}],\end{aligned}$$

a relation that yields that, for every parameter ν,

$$P_\nu\left[\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty}[L_k(\hat{\lambda}_k;X_1^k)\geq L_k(\lambda;X_1^k),\lambda\in\mathbb{R}]\right]\geq P_\nu\left[\bigcap_{k=m}^{\infty}[X_1^k\in\mathcal{S}_k]\right].$$

After taking the limit as $m\to\infty$, an application of Lemma 3.3 leads to

$$P_\nu\left[\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty}[L_k(\hat{\lambda}_k;X_1^k)\geq L_k(\lambda;X_1^k),\lambda\in\mathbb{R}]\right]=1,\quad \nu\in\mathbb{R},$$

so that, by Definition 3.1, $\{\hat{\lambda}_n\}$ is a sequence of maximum likelihood estimators.

4. Consistency

The objective of this section is to show that a sequence $\{\hat{\lambda}_n\}$ of maximum likelihood estimators of λ is consistent, *i.e.*, that $\{\hat{\lambda}_n\}$ converges to the true parameter value with probability 1.

Theorem 4.1. *Suppose that Assumption 2.1 holds and let $\{\hat{\lambda}_n\}$ be a sequence of maximum likelihood estimators. In this context,*

$$P_\nu\left[\lim_{n\to\infty}\hat{\lambda}_n=\nu\right]=1,\quad \nu\in\mathbb{R}.$$

The proof of this result relies on the two lemmas stated below and involves the following notation: Throughout the remainder of the section $\{\hat{\lambda}_n\}$ is a given sequence of maximum likelihood estimators of λ and the event Ω^* is given by

$$\Omega^*:=\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty}[L_k(\hat{\lambda}_k;X_1^k)\geq L_k(\lambda;X_1^k),\ \lambda\in\mathbb{R}]. \tag{4.1}$$

Also, for each $\nu\in\mathbb{R}$,

$$\begin{aligned}\Omega_\nu^- &:= \left[\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}I[X_i\leq -c(\nu)]=P_\nu[X\leq -c(\nu)]\right],\\ \Omega_\nu^+ &:= \left[\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}I[X_i\geq c(\nu))]=P_\nu[X\geq c(\nu)]\right],\end{aligned} \tag{4.2}$$

where $c(\nu)>0$ is as in Lemma 2.2(ii). Notice that the strong law of large numbers and Definition 3.1 yield

$$P_\nu[\Omega^*]=P_\nu[\Omega_\nu^+]=1=P_\nu[\Omega_\nu^-]\quad \nu\in\mathbb{R}. \tag{4.3}$$

The core of the proof of Theorem 4.1 is the following boundedness property.

Lemma 4.1. *Under Assumption 2.1, if $\{\hat{\lambda}_n\}$ is a sequence of maximum likelihood estimators of λ, then*

$$P_\nu[\limsup_{n\to\infty}|\hat{\lambda}_n|<\infty]=1,\quad \nu\in\mathbb{R}.$$

Proof. It will be shown, by contradiction, that the event

$$[\limsup_{n\to\infty} \hat{\lambda}_n = \infty] \cap \Omega^* \cap \Omega_\nu^- \quad \text{is empty.} \tag{4.4}$$

To achieve this goal, *suppose* that the sample trajectory $X_1, X_2, \ldots$ is such that the above intersection occurs, and observe that *along this path* assertions (a)–(c) below hold:
(a) Since the event $[\limsup_n \hat{\lambda}_n = \infty]$ occurs when $X_1, X_2, \ldots$ is observed, there exist a (trajectory dependent) subsequence $\{n_k\}$ such that

$$n_k \geq k \quad \text{and} \quad \hat{\lambda}_{n_k} \geq k$$

for all positive integers k;
(b) Using that $X_1, X_2, \ldots$ is such that Ω^* occurs, it follows that $\hat{\lambda}_n$ maximizes $L_n(\cdot; X_1^n)$ for n large enough, so that there exists a positive integer M such that

$$L_n(\hat{\lambda}_n; X_1^n) \geq L_n(0; X_1^n) = -\log(2), \quad n \geq M;$$

(c) Since the observation of $X_1, X_2, \ldots$ implies that Ω_ν^- occurs,

$$\frac{1}{n}\sum_{i=1}^{n} I[X_i \leq -c(\nu)] \to P_\nu[X_1 \leq -c(\nu)] \quad \text{as } n \to \infty.$$

Notice now that for each integer $M_1 > M$ and $k > M_1$, (a) and (b) together yield

$$\begin{aligned} -\log(2) \leq L_{n_k}(\hat{\lambda}_{n_k}; X_1^{n_k}) &= \frac{1}{n_k}\sum_{i=1}^{n_k} \log(G(\hat{\lambda}_{n_k} X_i)) \\ &\leq \frac{1}{n_k}\sum_{i=1}^{n_k} \log(G(\hat{\lambda}_{n_k} X_i)) I[X_i \leq -c(\nu)] \\ &\leq \frac{1}{n_k}\sum_{i=1}^{n_k} \log[G(-M_1 c(\nu))] I[X_i \leq -c(\nu)] \end{aligned}$$

where (3.2) with $Q = \{i : i \leq n_k,\ X_i \leq -c(\nu)\}$ was used to set the second inequality, and the third one follows from the monotonicity of $\log(G(\cdot))$, since $\lambda_{n_k} > k > M_1$. From this point, letting k go to ∞, (c) leads to

$$-\log(2) \leq \log(G(-M_1 c(\nu)))P_\nu[X_1 \leq -c(\nu)];$$

and, recalling that $\lim_{x\to-\infty} \log(G(x)) = -\infty$ and that $P_\nu[X_1 \leq -c(\nu)]$ and $c(\nu)$ are both positive (by Lemma 2.2(ii)), taking the limit as $M_1 \to \infty$, it follows that $-\log(2) \leq -\infty$, which is a contradiction, establishing (4.4). Therefore, $[\limsup_{n\to\infty} \hat{\lambda}_n = \infty] \subset (\Omega^*)^c \cup (\Omega_\nu^-)^c$, inclusion that yields

$$P_\nu[\limsup_{n\to\infty} \hat{\lambda}_n = \infty] = 0, \quad \nu \in \mathbb{R},$$

by (4.3). Similarly, it can be established that $[\liminf_{n\to\infty} \hat{\lambda}_n = -\infty] \subset (\Omega^*)^c \cup (\Omega_\nu^+)^c$, so that $P_\nu[\liminf_{n\to\infty} \hat{\lambda}_n = -\infty] = 0$ for all $\nu \in \mathbb{R}$; the conclusion follows combining this fact with the above display.

To continue, observe that since $\log(G(\cdot)) \leq 0$, $E_\nu[\log(G(\lambda X_1))]$ is always a well-defined non positive number, where the expectation may assume the value $-\infty$. Also, since $x \mapsto x\log(x)$ is bounded on $(0,1)$, $E_\nu[\log(G(\nu x))] = \int_{\mathbb{R}} 2\log(G(\nu x))G(\nu x)f(x)\,dx$ is finite.

Lemma 4.2. *Suppose that Assumption 2.1 holds and let $\nu \in \mathbb{R}$ be arbitrary but fixed.*
(i) [Kullback's inequality.] *For each $\lambda \in \mathbb{R} \setminus \{\nu\}$,*

$$E_\nu[\log(G(\lambda X_1))] < E_\nu[\log(G(\nu X_1))].$$

(ii) Assume that $\{r_k\}$ and $\{s_k\}$ are two real sequences such that, for some $\nu^* \in \mathbb{R}$,

$$r_k \searrow \nu^* \quad \text{and} \quad s_k \nearrow \nu^* \quad \text{as } k \to \infty.$$

and suppose that, for every $k = 1, 2, 3, \ldots$, the following inequality holds:

$$E_\nu[\log(G(\nu X))] \leq E_\nu[\log(G(r_k X))I[X \geq 0]] + E_\nu[\log(G(s_k X))I[X < 0]]. \tag{4.5}$$

In this case, $\nu = \nu^*$.

Proof. (i) If $\lambda \neq \nu$, Assumption 2.1 and the *strict concavity* of the logarithmic function yield, via Jensen's inequality, that

$$\begin{aligned} \int_{\mathbb{R}} \log\left(\frac{\rho(x;\lambda)}{\rho(x;\nu)}\right)\rho(x;\nu)\,dx &< \log\left(\int_{\mathbb{R}} \frac{\rho(x;\lambda)}{\rho(x;\nu)}\rho(x;\nu)\,dx\right) \\ &= \log\left(\int_{\mathbb{R}} \rho(x;\lambda)\,dx\right) = 0. \end{aligned}$$

Observing that $\rho(x;\lambda)/\rho(x;\nu) = G(\lambda x)/G(\nu x)$ when $\rho(x;\nu)$ is positive, the above inequality can be written as $E_\nu[\log(G(\lambda X_1)) - \log(G(\nu X_1))] < 0$, which yields the desired conclusion since, as already noted, $E_\nu[\log(G(\nu X_1))]$ is finite.
(ii) Notice that $r_k \searrow \nu^*$ leads to $r_k X \searrow \nu^* X$ on $[X \geq 0]$, whereas $s_k \nearrow \nu^*$ implies that $s_k X \searrow \nu^* X$ on $[X < 0]$. Therefore, since $-\log(G(\cdot))$ is decreasing and nonnegative,

$$0 \leq -\log(G(r_k X))I[X \geq 0] \nearrow -\log(G(\nu^* X))I[X \geq 0]$$

and

$$0 \leq -\log(G(s_k X))I[X < 0] \nearrow -\log(G(\nu^* X))I[X < 0].$$

Now, an application of the monotone convergence theorem yields

$$E_\nu[\log(G(r_k X))I[X \geq 0]] \searrow E_\nu[\log(G(\nu^* X))I[X \geq 0]]$$

and

$$E_\nu[\log(G(s_k X))I[X < 0]] \searrow E_\nu[\log(G(\nu^* X))I[X < 0]],$$

convergences that, after taking the limit as k goes to ∞ in (4.5), lead to

$$E_\nu[\log(G(\nu X))] \leq E_\nu[\log(G(\nu^* X))],$$

and then $\nu = \nu^*$, by part (i).

After these preliminaries, the proof of Theorem 4.1 is presented below. The argument uses the following notation, where $\mathcal{Q}$ stands for the set of rational numbers: For each $\nu \in \mathbb{R}$, define

$$\Omega_\nu^0 := \left[\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log(G(\nu X_i)) = E_\nu[\log(G(\nu X_1))\right]; \tag{4.6}$$

$$\Omega_\nu^1 := \left[\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log(G(\lambda X_i))I[X_i\geq 0] = E_\nu[\log(G(\lambda X_1))I[X_1\geq 0]],\quad \lambda\in\mathcal{Q}\right], \tag{4.7}$$

and

$$\Omega_\nu^2 := \left[\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log(G(\lambda X_i))I[X_i< 0] = E_\nu[\log(G(\lambda X_1))I[X_1< 0]],\quad \lambda\in\mathcal{Q}\right]. \tag{4.8}$$

Since $\mathcal{Q}$ is denumerable, the strong law of large numbers yields $P_\nu[\Omega_\nu^i] = 1$ for $i = 0, 1, 2$, and then, setting

$$\Omega_\nu = \Omega_\nu^0\cap\Omega_\nu^1\cap\Omega_\nu^2, \tag{4.9}$$

it follows that

$$P_\nu[\Omega_\nu] = 1. \tag{4.10}$$

Proof of Theorem 4.1. Given $\nu\in\mathbb{R}$, it is sufficient to show that

$$\left[\limsup_n|\hat\lambda_n|<\infty\right]\cap\Omega^*\cap\Omega_\nu\subset\left[\lim_n\hat\lambda_n=\nu\right], \tag{4.11}$$

where Ω^* and Ω_ν are specified in (4.1) and (4.9), respectively. Indeed, if this inclusion is valid, then (4.3), (4.10) and Lemma 4.1 together imply that $P_\nu[\lim_n\hat\lambda_n = \nu] = 1$. To establish (4.11) let $X_1, X_2, X_3, \ldots$ be a fixed trajectory such that $\left[\limsup_n|\hat\lambda_n|<\infty\right]\cap\Omega^*\cap\Omega_\nu$ occurs, select an arbitrary limit point ν^* of the associated sequence $\{\hat\lambda_n\}$, and observe the following facts (a)–(c):

(a) ν^* is finite, since $\limsup_n|\hat\lambda_n|<\infty$ holds when $X_1, X_2, X_3, \ldots$ is observed. Let r and s be arbitrary rational numbers satisfying

$$s<\nu^*<r \tag{4.12}$$

and select a sequence $\{n_k\}$ of positive integers such that

$$n_k>k,\quad s<\hat\lambda_{n_k}<r,\quad k=1,2,\ldots. \tag{4.13}$$

(b) Since the path $X_1, X_2, X_3, \ldots$ is such that Ω^* occurs, $\hat\lambda_n$ is a maximizer of $L_n(\cdot; X_1^n)$ when n is large enough, say $n > M$; see (4.1). In particular, $L_n(\nu; X_1^n)\leq$

$L_n(\hat{\lambda}_n; X_1^n)$ when $n > M$, and then, replacing n by n_k in this inequality and using the monotonicity of $\log(G(\cdot))$ as well as (4.13), it follows that

$$\begin{aligned}
\frac{1}{n_k}\sum_{i=1}^{n_k}\log(G(\nu X_i)) &= L_{n_k}(\nu; X_1^{n_k}) \\
&\le L_{n_k}(\hat{\lambda}_{n_k}; X_1^{n_k}) \\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\log(G(\hat{\lambda}_{n_k} X_i)) \\
&\le \frac{1}{n_k}\sum_{i=1}^{n_k}\log(G(rX_i))I[X_i \ge 0] \\
&\quad + \frac{1}{n_k}\sum_{i=1}^{n_k}\log(G(sX_i))I[X_i < 0], \quad k > M. \qquad (4.14)
\end{aligned}$$

(c) Since the trajectory $X_1, X_2, \ldots$ is such that Ω_ν occurs, a glance to (4.6)–(4.9) immediately yields that, as $k \to \infty$,

$$\begin{aligned}
\frac{1}{n_k}\sum_{i=1}^{n_k}\log(G(\nu X_i)) &\to E_\nu[\log(G(\nu X))], \\
\frac{1}{n_k}\sum_{i=1}^{n_k}\log(G(rX_i))I[X_i \ge 0] &\to E_\nu[\log(G(rX)I[X \ge 0]], \text{ and} \\
\frac{1}{n_k}\sum_{i=1}^{n_k}\log(G(sX_i))I[X_i < 0] &\to E_\nu[\log(G(sX)I[X < 0]].
\end{aligned}$$

After taking the limit as $k \to \infty$ in (4.14), these convergences yield that

$$E_\nu[\log(G(\nu X))] \le E_\nu[\log(G(rX)I[X \ge 0]] + E_\nu[\log(G(sX)I[X < 0]],$$

and then, since r and s are arbitrary rational numbers satisfying (4.12), from Lemma 4.2(ii) it follows that $\nu^* = \nu$. In short, it has been proved that along an arbitrary path $X_1, X_2, \ldots$ for which the intersection $\left[\limsup_n |\hat{\lambda}_n| < \infty\right] \cap \Omega^* \cap \Omega_\nu$ occurs, the corresponding sequence $\{\hat{\lambda}_n\}$ has ν as its unique limit point, so that $\hat{\lambda}_n \to \nu$ as $n \to \infty$. This establishes (4.11) and, as already noted, completes the proof.

5. Asymptotic Distribution

The remainder of the paper concerns the asymptotic behavior of a consistent sequence $\{\hat{\lambda}_n\}$ of maximum likelihood estimators of λ, whose existence is guaranteed by Assumption 2.1. As already mentioned, the large sample properties of $\{\hat{\lambda}_n\}$ will be studied under the null hypothesis $\mathcal{H}_0 : \lambda = 0$, and the analysis below requires two properties on the densities g and f generating the family $S(f, g)$, namely, (i) smoothness of density g outside of $\{0\}$ and a 'good' behavior of its derivatives around $\lambda = 0$, and (ii) a moment-dominance condition involving both densities f and g. After a formal presentation of these assumptions, the main result is stated at the end of the section, and the corresponding proof is given after establishing the necessary technical preliminaries.

Assumption 5.1. *For some nonnegative integer r—hereafter referred to as the critical order— the following conditions hold:*
(i) The symmetric density $g(x)$ is continuous on $\mathbb{R}$ and has derivatives up to order $2r+2$ on the interval $(0,\infty)$;
(ii) $D^k g(0+) = \lim_{x\searrow 0} D^k g(x)$ exists and is finite for $k = 0, 1, 2, \ldots, 2r+1$, and
(iii) $D^r g(0+) \neq 0$, whereas $D^s g(0+) = 0$ for $0 \leq s < r$.

Remark 5.1. *(i) Under this Assumption, $g(\cdot)$ statisfies that $D^r g(0+) = \lim_{x\searrow 0} r!g(x)/x^r \geq 0$, by continuity of g if $r = 0$, or by L'Hopital's rule, if $r > 0$, so that $D^r g(0+) > 0$, since $D^r g(0+)$ is no null. It follows that a density g satisfying Assumption 5.1 can be expressed as $g(x) = |x|^r h(|x|)$ where r is a nonnegative integer (the critical order), $h : [0,\infty) \to [0,\infty)$ is continuous with $h(0) > 0$, and has derivatives of order up to $2r+2$ on $(0,\infty)$, which are 'well-behaved' near zero so that the required lateral limits of the derivatives of g exist at $x = 0$. Thus, besides the smoothness requirement on the whole interval $(0,\infty)$, the core of Assumption 5.1 essentially concerns the local behavior of g around the origin.*
(ii) Under Assumption 5.1, density $g(\cdot)$ is continuous, so that $G(\cdot)$ has continuous derivative, and then $\partial_\lambda \ell(\lambda; x) = \partial_\lambda \log(G(\lambda x))$ exists if $\ell(\lambda; x)$ is finite. Suppose now that $\hat{\lambda}_n$ maximizes $L_n(\cdot; X_1^n)$. In this case $-\log(2) = L_n(0; X_1^n) \leq L_n(\hat{\lambda}_n; X_1^n)$ implies that $\log(G(\hat{\lambda}_n X_i)$ is finite for every $i = 1, 2, \ldots, n$, by (3.1), so that $L_n(\cdot; X_1^n)$ is differentiable at $\hat{\lambda}_n$, and then the likelihood equation holds: $\partial_\lambda L_n(\hat{\lambda}_n; X_1^n) = 0$.

By symmetry, Assumption 5.1 yields that $g(\cdot)$ also has derivatives up to order $2r+2$ on the interval $(-\infty, 0)$; indeed, if $k \leq 2r+2$ then $D^k g(x) = (-1)^k D^k g(-x)$ for $x \neq 0$, so that

$$D^k g(0-) = \lim_{x\nearrow 0} D^k g(x) = (-1)^k D^k g(0+), \quad k = 0, 1, 2, \ldots, 2r+1, \tag{5.1}$$

and the nullity of $D^k g(0+)$ for $0 \leq k < r$ implies that g has null (bilateral) derivative at $x = 0$ of any order less that r. On the other hand, under Assumption 5.1 the cumulative distribution function G in (1.1) has derivatives up to order $2r+3$ on $\mathbb{R} \setminus \{0\}$, and using the relation $D^k G(x) = D^{k-1} g(x)$ for $x \neq 0$ and $k > 0$ it follows that

$$D^k G(0) = 0, \quad 1 \leq k < r+1 \tag{5.2}$$

and

$$D^k G(0+) = D^{k-1} g(0+), \quad D^k G(0-) = D^{k-1} g(0-), \quad r+1 \leq k \leq 2r+2. \tag{5.3}$$

Next, define

$$H(x) := \log(G(x)), \quad x \in \mathbb{R}, \tag{5.4}$$

and observe the equalities

$$\begin{aligned} \ell(\lambda; x) &= H(\lambda x), \quad x \in \mathbb{R}, \\ \partial_\lambda^k \ell(\lambda; x) &= D^k H(\lambda x) x^k, \quad x \neq 0, \quad 1 \leq k \leq 2r+3 \end{aligned} \tag{5.5}$$

see (1.4). It follows that the lateral limits of $\partial_\lambda^k \ell(\cdot; x)$ at zero are given by

$$\partial_\lambda^k \ell(0+; x) = \begin{cases} D^k H(0+) x^k, & \text{if } x > 0; \\ D^k H(0-) x^k, & \text{if } x < 0. \end{cases} \tag{5.6}$$

and

$$\partial_\lambda^k \ell(0-;x) = \begin{cases} D^k H(0-)x^k, & \text{if } x>0; \\ D^k H(0+)x^k, & \text{if } x<0. \end{cases} \tag{5.7}$$

The analysis below uses (lateral) Taylor expansions of order $2r+2$ for $L_n(\cdot;X_1^n)$ around zero, and it is necessary to have an integrable bound for the residual as well as finite second moments for the coefficients. For this reason, the following conditions will be enforced.

Assumption 5.2. *Conditions (i) and (ii) below hold, where r is the the critical order in Assumption 5.1:*
(i) $E_0[X_1|^{4r+2}] = \int_{\mathbb{R}} x^{4r+2} f(x)\,dx < \infty$;
(ii) There exists a function $W:\mathbb{R}\to[0,\infty)$ and $\delta>0$ such that

$$|\partial_\lambda^{2r+3}\ell(\lambda;\cdot)| \le W(\cdot), \quad 0<|\lambda|\le\delta, \quad \text{and} \quad \int_{\mathbb{R}} W(x)f(x)\,dx<\infty. \tag{5.8}$$

Remark 5.2. *The moment requirement in Assumption 5.2(i) concerns only density f, whereas the dominance condition in the second part involves a relation between g and f. Using (5.4), it can be shown by induction that, for $x\neq 0$, $D^kH(x)$ is a polynomial in $D^sg(x)/G(x), s=0,1,2,\dots,k-1$, and then, setting*

$$M_r := \max\{|D^k g(x)|/G(x) : 0\le k\le 2r+2,\ x\neq 0\},$$

via (5.5) it follows that there exists a constant B such that $|\partial_\lambda^{2r+3}\ell(\lambda,x)| \le BM_r|x|^{2r+3}$ so that, if $M_r<\infty$, then Assumption 5.2 holds entirely if the moment condition in part (i) is valid. It is not difficult to see that M_r is finite when there exists $x_0>0$ for which (a) or (b) below occur:
(a) $g(x)$ is a rational function for $x>x_0$, as it is the case if $g(\cdot)$ is a multiple of a t-density for x large enough;
(b) $g(x)=p(x)e^{-\beta x}$ on (x_0,∞), where β is a positive constant and $p(x)$ is a polynomial or, more generally, a linear combination of terms of the form x^s; this occurs, for instance, when $g(\cdot)$ is proportional to a mixture of gamma densities on (x_0,∞).

To state the result on the large sample distribution of maximum likelihood estimators, set

$$V_{r+1} := 4\left(\frac{D^r g(0+)}{(r+1)!}\right)^2 E_0\left[X_1^{2r+2}\right] > 0 \tag{5.9}$$

which, as it will be shown later, is the variance of $\partial_\lambda^{r+1}\ell(0+,X_i)/(r+1)!$; for the strict inequality, see Remark 5.1(i).

Theorem 5.1. *Let $\{\hat\lambda_n\}$ be a consistent sequence of maximum likelihood estimators of λ, and suppose that Assumptions 5.1 and 5.2 hold. In this context, under the hypothesis $\mathcal{H}_0:\lambda=0$, the following convergences (i) and (ii) occur as $n\to\infty$, where r is the critical order in Assumption 5.1, and Z is a random variable with standard normal distribution:*

$$(i) \qquad (nV_{r+1})^{1/(2(r+1))}\ \hat\lambda_n \xrightarrow{\text{d}} |Z|^{1/(r+1)}\text{sign}(Z),$$

and

$$(ii) \qquad 2n[L_n(\hat\lambda_n;X_1^n) - L_n(0;X_1^n)] \xrightarrow{\text{d}} Z^2.$$

Remark 5.3. *(i) Suppose that the critical index r is null. In this case Theorem 5.1 (i) yields that $(nV_1)^{1/2}\ \hat{\lambda}_n \xrightarrow{\mathrm{d}} |Z|\mathrm{sign}(Z) \stackrel{\mathrm{d}}{=} Z$. This conclusion coincides with that obtained from the general classical results presented, for instance, in Lehmann and Casella (1998, Section 6.3), or Shao (1999, Section 4.4), where derivatives up to order 2 are required for $g(\cdot)$ around zero; under Assumption 5.1, only the lateral limits of Dg and D^2g exist at zero.*
(ii) Suppose that the critical order r is positive and that $g(\cdot)$ has (bilateral) derivative of order r at zero, so that $D^r g(0+) = D^r g(0-)$. Since $D^r g(0+) \neq 0$ it follows from (5.1) that r is an even integer, and $g(\cdot)$ has derivatives up to order r on the real line. Thus, setting $s = r+1$, s is odd, $\ell(\cdot; x)$ has derivatives up to order s on $\mathbb{R}$, and $\partial_\lambda^k \ell(0;\cdot) = 0$ for $1 \leq k < s$, whereas $\partial_\lambda^s \ell(0, x) \neq 0$ for $x \neq 0$. If, among other conditions, $g(\cdot)$ has derivatives up to order $2s$ at zero, an application of Theorem 1 in Rotnitzky et al. (2000) yields the conclusions in Theorem 5.1; notice, however, that Assumption 5.1 only ensures the existence of the lateral limits $D^k g(0\pm)$ for $s < k \leq 2s$, so that Theorem 5.1 extends Theorem 1 in Rotnitzky et al. (2000) to the framework of this work.

The rather technical proof of Theorem 5.1, requiring some explicit computations for the lateral limits $\partial_\lambda^k \ell(0\pm; x)$ in terms of density $g(\cdot)$, will be given after the preliminaries established in the following section.

6. Technical Preliminaries

This section is dedicated to establish the basic tools that will be used to prove Theorem 5.1, namely, lateral Taylor expansions around the origin for the average likelihood $L_n(\cdot; X_1^n)$ and its first derivative; via the absolute value function, such expansions are stated below as single equations. The following notation will be used:

$$\Delta(x) := \partial_\lambda^{r+1}\ell(0-; x) - \partial_\lambda^{r+1}\ell(0+; x), \quad x \in \mathbb{R}. \tag{6.1}$$

Theorem 6.1. *Suppose that Assumptions 5.1 and 5.2 hold. In this case, the assertions (i)—(iii) below occur, where r is the critical order in Assumption 5.1, $\delta > 0$ and $W(\cdot)$ are as in Assumption 5.2, and*

$$\Delta_n := \frac{1}{n}\sum_{i=1}^{n} \Delta(X_i);$$

see (6.1).
(i) For each positive integer n and $\alpha \in (-\delta, \delta)$,

$$\begin{aligned} &L_n(\alpha; X_1^n) - L_n(0; X_1^n) \\ &= |\alpha|^r \alpha \left\{ \sum_{k=r+1}^{2r+1} \frac{\partial_\lambda^k L_n(0+; X_1^n)}{k!} |\alpha|^{k-r-1} \right. \\ &\left. + \left(\frac{\partial_\lambda^{2r+2} L_n(0+; X_1^n) + \Delta_n I_{(-\infty,0)}(\alpha)}{(2r+2)!} + \frac{W_n^*(\alpha)}{(2r+3)!}\alpha \right) |\alpha|^r \alpha \right\}, \end{aligned} \tag{6.2}$$

and

$$\begin{aligned}&\partial_\lambda L_n(\alpha, X_1^n)\\ &= |\alpha|^r \Bigg\{ \sum_{k=r}^{2r} \frac{\partial_\lambda^{k+1} L_n(0+; X_1^n)}{k!} |\alpha|^{k-r}\\ &+ \left(\frac{\partial_\lambda^{2r+2} L_n(0+; X_1^n) + \Delta_n I_{(-\infty,0)}(\alpha)}{(2r+1)!} + \frac{\tilde{W}_n(\alpha)}{(2r+2)!}\alpha \right) |\alpha|^r \alpha \Bigg\},\end{aligned} \tag{6.3}$$

where the random variables $W_n^*(\alpha)$ and $\tilde{W}_n(\alpha)$ satisfy

$$|W_n^*(\alpha)|,\ |\tilde{W}_n(\alpha)| \le W_n := \frac{1}{n}\sum_{i=1}^n W(X_i). \tag{6.4}$$

(ii) Under $\mathcal{H}_0 : \lambda = 0$ for each $k = r+1, \dots, 2r+1$, the following convergences hold as $n \to \infty$:

$$\sqrt{n}\ \partial_\lambda^k L_n(0+; X_1^n) \xrightarrow{\mathrm{d}} \mathcal{N}(0, v_k), \quad \text{where } v_k = E_0\left[\left(\partial_\lambda^k \ell(0+; X_1) \right)^2 \right], \tag{6.5}$$

whereas

$$\Delta_n \to 0 \quad \text{and} \quad 2\frac{\partial_\lambda^{2r+2} L_n(0+; X_1^n)}{(2r+2)!} \to -V_{r+1} \quad P_0\text{-a.s.}; \tag{6.6}$$

see (5.9).

The proof of this theorem relies on explicit formulas for $\partial_\lambda^k \ell(0\pm; x)$ in terms of density $g(\cdot)$ and, in this direction, the following lemma concerning the lateral limits at zero of the derivatives of function $H(\cdot)$ in (5.4) will be useful.

Lemma 6.1. *Suppose that Assumption 5.1 holds. In this case, the lateral limits at 0 of the derivatives of function $H(\cdot)$ in (5.4) satisfy the following relations (i)–(iii):*
(i) $D^k H(0+) = D^k H(0-) = 0$, $1 \le k < r+1$;
(ii) If $r+1 \le k < 2r+2$, then

$$D^k H(0+) = 2D^{k-1} g(0+) \quad \text{and} \quad D^k H(0-) = 2D^{k-1} g(0-).$$

$$D^{2r+2} H(0+) = 2D^{2r+1} g(0+) - \frac{1}{2}\binom{2r+2}{r+1} \left(D^{r+1} H(0+) \right)^2, \tag{iii}$$

and

$$D^{2r+2} H(0-) = 2D^{2r+1} g(0-) - \frac{1}{2}\binom{2r+2}{r+1} \left(D^{r+1} H(0-) \right)^2.$$

Proof. Recalling that the distribution function $G(x)$ is continuous and has derivatives up to order $2r+3$ on $\mathbb{R} \setminus \{0\}$; from (5.4) it follows that $G(x)DH(x) = DG(x)$ and, via Leibinitz' formula,

$$G(x)D^k H(x) + \sum_{i=1}^{k-1} \binom{k-1}{i} D^i G(x) D^{k-i} H(x) = D^k G(x)$$

for $x \neq 0$ and $2 \leq k \leq 2r+3$; since $G(\cdot)$ is continuous and $G(0) = 1/2$, taking lateral limit as x approaches to zero these equalities lead to

$$DH(0\pm) = 2DG(0)$$

and, for $2 \leq k \leq 2r+2$,

$$D^k H(0\pm) + 2\sum_{i=1}^{k-1} \binom{k-1}{i} D^i G(0\pm) D^{k-i} H(0\pm) = 2D^k G(0\pm).$$

Since $D^k G(0\pm) = 0$ when $1 \leq k \leq r$, by (5.2), these relations yield $D^k H(0\pm) = 0$ for $1 \leq k \leq r$, establishing part(i), as well as

$$D^{r+1} H(0\pm) = 2D^{r+1} G(0\pm), \tag{6.7}$$

and

$$\begin{aligned} &\text{for } r+1 < k \leq 2r+2, \\ &D^k H(0\pm) + 2\sum_{i=r+1}^{k-1} \binom{k-1}{i} D^i G(0\pm) D^{k-i} H(0\pm) = 2D^k G(0\pm). \end{aligned} \tag{6.8}$$

To prove part (ii), select an integer k such that $r+1 < k < 2r+2$. In this case, if $k > i \geq r+1$ then $1 \leq k-i < r+1$, and then $D^{k-i} H(0\pm) = 0$, by part (i), so that the summation in the above display is null. Therefore, $D^k H(0\pm) = 2D^k G(0\pm)$, and combining this with (6.7) it follows that

$$D^k H(0\pm) = 2D^k G(0\pm), \quad r+1 \leq k < 2r+2,$$

equalities that yield part (ii) via (5.3). To conclude, observe that if $k = 2r+2$ then $2r+1 \geq i > r+1$ implies that $1 \leq k-i < r+1$, and in this case $D^{k-i} H(0\pm) = 0$, by part (i), so that the terms in the summation in (6.8) with $k = 2r+2$ are null when $i > r+1$. Consequently,

$$D^{2r+2} H(0\pm) + 2\binom{2r+1}{r+1} D^{r+1} G(0\pm) D^{r+1} H(0\pm) = 2D^{2r+2} G(0\pm);$$

since

$$2D^{r+1} G(0\pm) = 2D^r g(0\pm) = D^{r+1} H(0\pm)$$

and $D^{2r+2} G(0\pm) = D^{2r+1} g(0\pm)$, by (5.3) and part (ii), respectively, the conclusion follows observing that $\binom{2r+1}{r+1} = 2^{-1}\binom{2r+2}{r+1}$.

The expressions in the previous lemma are used below to determine the lateral limits of $\partial_\lambda^k \ell(\cdot; x)$ at zero in terms of density $g(\cdot)$.

Lemma 6.2. *Under Assumption 5.1, assertions (i)–(v) below hold:*
(i) $\partial_\lambda^k \ell(0+; \cdot) = 0 = \partial_\lambda^k \ell(0-; \cdot)$ *for* $1 \leq k \leq r$.
(ii) For each $x \in \mathbb{R}$ *and* $r+1 \leq k < 2r+2$,

$$\begin{aligned} \partial_\lambda^k \ell(0+; x) &= 2D^{k-1} g(0+) |x|^{k-1} x, \\ \partial_\lambda^k \ell(0-; x) &= 2D^{k-1} g(0-) |x|^{k-1} x. \end{aligned}$$

(iii) $\partial_\lambda^k \ell(0-; x) = (-1)^{k-1} \partial_\lambda^k \ell(0+; x)$ *for* $r+1 \leq k < 2r+2$ *and* $x \in \mathbb{R}$.

(iv) For each $x \in \mathbb{R}$

$$\partial_\lambda^{2r+2}\ell(0+;x) = 2D^{2r+1}g(0+)|x|^{2r+1}x - \frac{1}{2}\binom{2r+2}{r+1}\left(\partial_\lambda^{r+1}\ell(0+;x)\right)^2, \tag{6.9}$$

and

$$\partial_\lambda^{2r+2}\ell(0-;x) = 2D^{2r+1}g(0-)|x|^{2r+1}x - \frac{1}{2}\binom{2r+2}{r+1}\left(\partial_\lambda^{r+1}\ell(0-;x)\right)^2. \tag{6.10}$$

Consequently,
(v) The difference between $\partial_\lambda^{2r+2}\ell(0-;x)$ *and* $\partial_\lambda^{2r+2}\ell(0+;x)$ *is given by*

$$\Delta(x) = -4D^{2r+1}g(0+)\ |x|^{2r+1}x, \quad x \in \mathbb{R}; \tag{6.11}$$

see (6.1).

Proof. From Lemma 6.1(i), part (i) follows via (5.6) and (5.7), whereas these latter equalities and Lemma 6.1(ii) together yield that, for $r+1 \le k \le 2r+1$, (a) and (b) below hold:
(a) For $x \ge 0$,

$$\partial_\lambda^k\ell(0+;x) = 2D^{k-1}g(0+)x^k \quad \text{and} \quad \partial_\lambda^k\ell(0-;x) = 2D^{k-1}g(0-)x^k;$$

(b) If $x < 0$,

$$\begin{aligned}\partial_\lambda^k\ell(0+;x) &= 2D^{k-1}g(0-)x^k \\ &= 2D^{k-1}g(0+)(-1)^{k-1}x^k = 2D^{k-1}g(0+)|x|^{k-1}x, \\ \partial_\lambda^k\ell(0-;x) &= 2D^{k-1}g(0+)x^k \\ &= 2D^{k-1}g(0-)(-1)^{k-1}x^k = 2D^{k-1}g(0-)|x|^{k-1}x,\end{aligned}$$

where (5.1) was used to set the second equalities. These facts (a) and (b) together lead to part (ii), which implies part (iii) via (5.1). To establish part (iv), notice that Lemma 6.1(iii) and (5.6) together imply that
For $x \ge 0$

$$\begin{aligned}\partial_\lambda^{2r+2}\ell(0+;x) &= 2D^{2r+1}g(0+)x^{2r+2} - \frac{1}{2}\binom{2r+2}{r+1}\left(D^{r+1}H(0+)x^{r+1}\right)^2 \\ &= 2D^{2r+1}g(0+)x^{2r+2} - \frac{1}{2}\binom{2r+2}{r+1}\left(\partial_\lambda^{r+1}\ell(0+;x)\right)^2,\end{aligned}$$

showing that (6.9) holds for $x \ge 0$, whereas combining Lemma 6.1(iii) with relations (5.6) and (5.1) it follows that if $x < 0$, then

$$\begin{aligned}&\partial_\lambda^{2r+2}\ell(0+;x) \\ &\quad = 2D^{2r+1}g(0-)x^{2r+2} - \frac{1}{2}\binom{2r+2}{r+1}\left(D^{r+1}H(0-)x^{r+1}\right)^2 \\ &\quad = 2D^{2r+1}g(0+)(-1)^{2r+1}x^{2r+1}x - \frac{1}{2}\binom{2r+2}{r+1}\left(\partial_\lambda^{r+1}\ell(0+;x)\right)^2 \\ &\quad = 2D^{2r+1}g(0+)|x|^{2r+1}x - \frac{1}{2}\binom{2r+2}{r+1}\left(\partial_\lambda^{r+1}\ell(0+;x)\right)^2,\end{aligned}$$

and then (6.9) also holds for $x < 0$. Equality (6.10) can be established along similar lines and, finally, observing that $\partial_\lambda^{r+1}\ell(0+;x)$ and $\partial_\lambda^{r+1}\ell(0-;x)$ have the same absolute value, by part (iii), via part (iv), (6.11) follows immediately from (5.1) and (6.1).

Next, the above expressions will be used to write lateral Taylor expansions for $\ell(\cdot;x)$ and $\partial_\lambda\ell(\cdot;x)$ around the origin.

Lemma 6.3. *Suppose that Assumptions 5.1 and 5.2 hold. In this case, the following expansions are valid for $x \in \mathbb{R}$ and $\alpha \in (-\delta,\delta)\setminus\{0\}$:*

$$\begin{aligned}&\ell(\alpha,x)-\ell(0;x)\\&=|\alpha|^r\alpha\left\{\sum_{k=r+1}^{2r+1}\frac{\partial_\lambda^k\ell(0+;x)}{k!}|\alpha|^{k-r-1}\right.\\&\qquad\left.+\left(\frac{\partial_\lambda^{2r+2}\ell(0+;x)+I_{(-\infty,0)}(\alpha)\Delta(x)}{(2r+2)!}+\frac{W^*(\alpha,x)}{(2r+3)!}\alpha\right)|\alpha|^r\alpha\right\},\end{aligned}$$

and

$$\begin{aligned}&\partial_\lambda\ell(\alpha,x)\\&\quad=|\alpha|^r\left\{\sum_{k=r}^{2r}\frac{\partial_\lambda^{k+1}\ell(0+;x)}{k!}|\alpha|^{k-r}\right.\\&\qquad\left.+\left(\frac{\partial_\lambda^{2r+2}\ell(0+;x)+I_{(-\infty,0)}(\alpha)\Delta(x)}{(2r+1)!}+\frac{\tilde W(\alpha,x)}{(2r+2)!}\alpha\right)|\alpha|^r\alpha\right\},\end{aligned}$$

where $\Delta(\cdot)$ is as in (6.1), and

$$|W^*(\alpha,x)|,\ |\tilde W(\alpha,x)|\le W(x). \tag{6.12}$$

Proof. Select $\alpha_0\neq 0$ with the same sign as α and $|\alpha_0|<|\alpha|$, so that the closed interval joining α_0 and α is contained in $(-\delta,\delta)\setminus\{0\}$. Since $\ell(\cdot;x)$ has derivatives up to order $2r+3$ outside 0, there exist points α^* and $\tilde\alpha$ between α_0 and α such that the following Taylor expansions hold:

$$\ell(\alpha,x)-\ell(\alpha_0;x)=\sum_{k=1}^{2r+2}\frac{\partial_\lambda^k\ell(\alpha_0;x)}{k!}(\alpha-\alpha_0)^k+\frac{\partial_\lambda^{2r+3}\ell(\alpha^*;x)}{(2r+3)!}(\alpha-\alpha_0)^{2r+3}, \tag{6.13}$$

and

$$\partial_\lambda\ell(\alpha,x)=\sum_{k=0}^{2r+1}\frac{\partial_\lambda^{k+1}\ell(\alpha_0;x)}{k!}(\alpha-\alpha_0)^k+\frac{\partial_\lambda^{2r+3}\ell(\tilde\alpha;x)}{(2r+2)!}(\alpha-\alpha_0)^{2r+2}, \tag{6.14}$$

where

$$\left|\partial_\lambda^{2r+3}\ell(\alpha^*;x)\right|,\left|\partial_\lambda^{2r+3}\ell(\tilde\alpha;x)\right|\le W(x), \tag{6.15}$$

by Assumption 5.2(ii). Next, the conclusions in the lemma will be obtained taking lateral limits as α_0 goes to zero. Recall that $\ell(\cdot;x)$ is continuous and consider the following exhaustive cases:

Case 1: $\alpha > 0$. taking the limit as α_0 decreases to zero, the above displayed relations and Lemma 6.2(i) together yield

$$\begin{aligned}
\ell(\alpha,x) - \ell(0;x) &= \sum_{k=r+1}^{2r+2} \frac{\partial_\lambda^k \ell(0+;x)}{k!}\alpha^k + \frac{W^*(\alpha;x)}{(2r+3)!}\alpha^{2r+3} \\
&= \sum_{k=r+1}^{2r+1} \frac{\partial_\lambda^k \ell(0+;x)}{k!}\alpha^k + \frac{\partial_\lambda^{2r+2}\ell(0+;x)}{(2r+2)!}\alpha^{2r+2} + \frac{W^*(\alpha;x)}{(2r+3)!}\alpha^{2r+3} \\
&= \alpha^{r+1}\left\{ \sum_{k=r+1}^{2r+1} \frac{\partial_\lambda^k \ell(0+;x)}{k!}\alpha^{k-r-1} \right. \\
&\qquad \left. + \left(\frac{\partial_\lambda^{2r+2}\ell(0+;x)}{(2r+2)!} + \frac{W^*(\alpha;x)}{(2r+3)!}\alpha \right)\alpha^{r+1} \right\},
\end{aligned}$$

and

$$\begin{aligned}
\partial_\lambda \ell(\alpha;x) &= \sum_{k=r}^{2r+1} \frac{\partial_\lambda^{k+1} \ell(0+;x)}{k!}\alpha^k + \frac{\tilde{W}(\alpha;x)}{(2r+2)!}\alpha^{2r+2} \\
&= \sum_{k=r}^{2r} \frac{\partial_\lambda^{k+1} \ell(0+;x)}{k!}\alpha^k + \frac{\partial_\lambda^{2r+2}\ell(0+;x)}{(2r+1)!}\alpha^{2r+1} + \frac{\tilde{W}(\alpha;x)}{(2r+2)!}\alpha^{2r+2} \\
&= \alpha^r\left\{ \sum_{k=r}^{2r} \frac{\partial_\lambda^{k+1} \ell(0+;x)}{k!}\alpha^{k-r} \right. \\
&\qquad \left. + \left(\frac{\partial_\lambda^{2r+2}\ell(0+;x)}{(2r+1)!} + \frac{\tilde{W}(\alpha;x)}{(2r+2)!}\alpha \right)\alpha^{r+1} \right\},
\end{aligned}$$

where $W^*(\alpha,x)$ is given by $W^*(\alpha,x) := \lim_{\alpha_0 \searrow 0} \partial_\lambda^{2r+3}\ell(\alpha^*;x)$ and, similarly, $\tilde{W}(\alpha,x) := \lim_{\alpha_0 \searrow 0} \partial_\lambda^{2r+3}\ell(\tilde{\alpha};x)$, so that

$$|W^*(\alpha,x)|, |\tilde{W}(\alpha,x)| \le W(x),$$

by (6.15); since α is positive, so that $I_{(-\infty,0)}(\alpha) = 0$, these last three displays are equivalent to (6.12)–(6.12).

Case 2: $\alpha < 0$. In this context, taking the limit as α_0 increases to zero in (6.13) and (6.14), Lemma 6.2(i) yields that

$$\ell(\alpha,x) - \ell(0;x) = \sum_{k=r+1}^{2r+2} \frac{\partial_\lambda^k \ell(0-;x)}{k!}\alpha^k + \frac{W^*(\alpha;x)}{(2r+3)!}\alpha^{2r+3}$$

and

$$\partial_\lambda \ell(\alpha;x) = \sum_{k=r}^{2r+1} \frac{\partial_\lambda^{k+1} \ell(0-;x)}{k!}\alpha^k + \frac{\tilde{W}(\alpha;x)}{(2r+2)!}\alpha^{2r+2}$$

where, analogously to the previous case, $W^*(\alpha,x) := \lim_{\alpha_0 \nearrow 0} \partial_\lambda^{2r+3}\ell(\alpha^*;x)$ and $\tilde{W}(\alpha,x) := \lim_{\alpha_0 \nearrow 0} \partial_\lambda^{2r+3}\ell(\tilde{\alpha};x)$ so that, again, (6.15) implies that (6.12) is valid. Observe now that Lemma 6.2(iii) allows to write

$$\sum_{k=r+1}^{2r+2} \frac{\partial_\lambda^k \ell(0-;x)}{k!}\alpha^k + \frac{W^*(\alpha;x)}{(2r+3)!}\alpha^{2r+3}$$

$$= \sum_{k=r+1}^{2r+1} \frac{\partial_\lambda^k \ell(0+;x)}{k!}(-1)^{k-1}\alpha^k + \frac{\partial_\lambda^{2r+2}\ell(0-;x)}{(2r+2)!}\alpha^{2r+2} + \frac{W^*(\alpha;x)}{(2r+3)!}\alpha^{2r+3}$$
$$= \sum_{k=r+1}^{2r+1} \frac{\partial_\lambda^k \ell(0+;x)}{k!}|\alpha|^{k-1}\alpha + \left(\frac{\partial_\lambda^{2r+2}\ell(0-;x)}{(2r+2)!} + \frac{W^*(\alpha;x)}{(2r+3)!}\alpha\right)(|\alpha|^r\alpha)^2$$
$$= |\alpha|^r\alpha\left\{\sum_{k=r+1}^{2r+1} \frac{\partial_\lambda^k \ell(0+;x)}{k!}|\alpha|^{k-r-1} + \left(\frac{\partial_\lambda^{2r+2}\ell(0-;x)}{(2r+2)!} + \frac{W^*(\alpha;x)}{(2r+3)!}\alpha\right)|\alpha|^r\alpha\right\}$$

and

$$\sum_{k=r}^{2r+1} \frac{\partial_\lambda^{k+1}\ell(0-;x)}{k!}\alpha^k + \frac{\tilde{W}(\alpha;x)}{(2r+2)!}\alpha^{2r+2}$$
$$= \sum_{k=r}^{2r} \frac{\partial_\lambda^{k+1}\ell(0+;x)}{k!}(-1)^k\alpha^k + \frac{\partial_\lambda^{2r+2}\ell(0-;x)}{(2r+1)!}\alpha^{2r+1} + \frac{\tilde{W}(\alpha;x)}{(2r+2)!}\alpha^{2r+2}$$
$$= \sum_{k=r}^{2r} \frac{\partial_\lambda^{k+1}\ell(0+;x)}{k!}|\alpha|^k + |\alpha|^r\left(\frac{\partial_\lambda^{2r+2}\ell(0-;x)}{(2r+1)!} + \frac{\tilde{W}(\alpha;x)}{(2r+2)!}\alpha\right)|\alpha|^r\alpha$$
$$= |\alpha|^r\left\{\sum_{k=r}^{2r} \frac{\partial_\lambda^{k+1}\ell(0+;x)}{k!}|\alpha|^{k-r} + \left(\frac{\partial_\lambda^{2r+2}\ell(0-;x)}{(2r+1)!} + \frac{\tilde{W}(\alpha;x)}{(2r+2)!}\alpha\right)|\alpha|^r\alpha\right\};$$

using that $\partial_\lambda^{2r+2}\ell(0-;x) = \partial_\lambda^{2r+2}\ell(0+;x) + I_{(-\infty,0)}(\alpha)\Delta(x)$, by (6.1), the last four displays together yield that (6.12) and (6.12) are also valid for $\alpha < 0$.

After the above preliminaries, the main result of this section can be established as follows.

PROOF OF THEOREM 6.1. (i) Since $L_n(\cdot; X_1^n)$ is the average of $\ell(\cdot; X_i)$, $i = 1, 2, \ldots, n$, by Lemma 6.3 the two indicated expansions hold with

$$W_n^*(\alpha) = \frac{1}{n}\sum_{i=1}^n W^*(\alpha, X_i) \quad \text{and} \quad \tilde{W}_n(\alpha) = \frac{1}{n}\sum_{i=1}^n \tilde{W}(\alpha, X_i),$$

so that (6.4) is satisfied, by (6.12).

(ii) Under $\mathcal{H}_0 : \lambda = 0$, X_i has symmetric distribution around zero with finite moment of order $4r+2$, by Assumption 5.2(i), so that a random variable of the form $|X_i|^{k-1}X_i$ has zero expectation and finite second moment when $1 \le k < 2r+2$. Thus, from Lemma 6.2, for all $k = r+1, \ldots 2r+1$,

$$E_0[\partial_\lambda^k \ell(0+; X_i)] = 0 \quad \text{and} \quad E_0[(\partial_\lambda^k \ell(0+; X_i))^2] = v_k < \infty,$$

as well as $E_0[\Delta(X_1)] = 0$. From this point, (5.9) and (6.9) together yield

$$\frac{2}{(2r+2)!}E[\partial_\lambda^{2r+2}\ell(0+; X_i)] = -\left(\frac{1}{(r+1)!}\right)^2 E[(\partial_\lambda^{r+1}\ell(0+; X_i))^2]$$
$$= -4\left(\frac{D^r g(0+)}{(r+1)!}\right)^2 E[X_i^{2r+2}] = -V_{r+1}.$$

where Lemma 6.2(ii) was used to set the second equality (notice that this shows that V_{r+1} is the variance of $\partial_\lambda^{r+1}\ell(0+;X_i)/(r+1)!$). Now, (6.5) and (6.6) follow from the central limit theorem and the strong law of large numbers, respectively.

7. Proof of Theorem 5.1

After the previous preliminaries, Theorem 5.1 is finally established in this section. The core of the argument has been decoupled into two lemmas showing that, under $\mathcal{H}_0 : \lambda = 0$, along a consistent sequence $\{\hat\lambda_n\}$ of maximum likelihood estimators, (i) the expansions in Theorem 6.1 can be simplified substantially, and (ii) that $\hat\lambda_n$ is no null that with probability converging to 1.

Lemma 7.1. *Suppose that Assumptions 5.1 and 5.2 hold and let $\{\hat\lambda_n\}$ be a consistent sequence of maximum likelihood estimators of λ. In this case, assertions (i) and (ii) below occur under the null hypothesis $\mathcal{H}_0 : \lambda = 0$.*
(i) On the event $[\,|\hat\lambda_n| < \delta]$, the following expressions are valid:

$$L_n(\hat\lambda_n; X_1^n) - L_n(0; X_1^n) = \hat\lambda_n|\hat\lambda_n|^r\left[\frac{\partial_\lambda^{r+1}L_n(0+;X_1^n)}{(r+1)!} + A_n - \frac{B_n}{2}\hat\lambda_n|\hat\lambda_n|^r\right], \tag{7.1}$$

and

$$\frac{\partial_\lambda L_n(\hat\lambda_n; X_1^n)}{r+1} = |\hat\lambda_n|^r\left[\frac{\partial_\lambda^{r+1}L_n(0+;X_1^n)}{(r+1)!} + \tilde A_n - \tilde B_n\hat\lambda_n|\hat\lambda_n|^r\right], \tag{7.2}$$

where

$$A_n = O_p\left(\frac{\hat\lambda_n}{\sqrt{n}}\right), \quad \tilde A_n = O_p\left(\frac{\hat\lambda_n}{\sqrt{n}}\right), \tag{7.3}$$

and

$$\lim_{n\to\infty} B_n = \lim_{n\to\infty}\tilde B_n = V_{r+1} \quad P_0\text{-a.s.} \tag{7.4}$$

(ii) Consequently,

$$2n[L_n(\hat\lambda_n; X_1^n) - L_n(0; X_1^n)] = 2n\hat\lambda_n|\hat\lambda_n|^r\left[\frac{\partial_\lambda^{r+1}L_n(0+;X_1^n)}{(r+1)!} + A_n - \frac{B_n}{2}\hat\lambda_n|\hat\lambda_n|^r\right] + o_p(1).$$

Proof. (i) Setting

$$A_n := \sum_{k=r+2}^{2r+1}\frac{\partial_\lambda^k L_n(0+;X_1^n)}{k!}|\hat\lambda_n|^{k-r-1}$$

and

$$B_n := -\left(\frac{2\partial_\lambda^{2r+2}L_n(0+;X_1^n) + 2\Delta_n I_{(-\infty,0)}(\hat\lambda_n)}{(2r+2)!} + \frac{2W_n^*(\hat\lambda_n)}{(2r+3)!}\hat\lambda_n\right),$$

(7.1) is equivalent to (6.2) with $\alpha = \hat{\lambda}_n$. Similarly, defining

$$\tilde{A}_n := \frac{1}{r+1} \sum_{k=r+1}^{2r} \frac{\partial_\lambda^{k+1} L_n(0+; X_1^n)}{k!} |\hat{\lambda}_n|^{k-r}$$

and

$$\begin{aligned} \tilde{B}_n &:= -\frac{1}{r+1} \left(\frac{\partial_\lambda^{2r+2} L_n(0+; X_1^n) + \Delta_n I_{(-\infty,0)}(\hat{\lambda}_n)}{(2r+1)!} + \frac{\tilde{W}_n(\hat{\lambda}_n)}{(2r+2)!} \hat{\lambda}_n \right) \\ &= -\left(\frac{2\partial_\lambda^{2r+2} L_n(0+; X_1^n) + 2\Delta_n I_{(-\infty,0)}(\hat{\lambda}_n)}{(2r+2)!} + \frac{\tilde{W}_n(\hat{\lambda}_n)}{(r+1)(2r+2)!} \hat{\lambda}_n \right) \end{aligned}$$

it follows that (7.2) is equivalent to (6.3) with $\alpha = \hat{\lambda}_n$. Therefore, since (6.2) and (6.3) are valid for $|\alpha| < \delta$, (7.1) and (7.2) hold on the event $[\,|\hat{\lambda}_n| < \delta]$. To conclude, it will be shown that (7.3) and (7.4) are satisfied. First, notice that A_n and $\tilde{A}_n$ defined above are null for $r = 0$, so that (7.3) certainly occurs in this case. On the other hand, if $r > 0$, then convergences (6.5) established in Theorem 6.1(ii) yield that $\partial_\lambda^k L_n(0+; X_1^n) = O_p(1/\sqrt{n})$ for $r+1 \le k < 2r+1$ and then (7.3) follows, since the above expressions for A_n and $\tilde{A}_n$ involve factors $|\hat{\lambda}_n|^s$ with $s \ge 1$ and

$$P_0[\hat{\lambda}_n \to 0] = 1, \tag{7.7}$$

by consistency. Next, observe that

$$|W_n^*(\hat{\lambda}_n)|, \ |\tilde{W}_n(\hat{\lambda}_n)| \le W_n = \frac{1}{n} \sum_{i=1}^{n} W(X_i),$$

by (6.4), and then the strong law of large numbers and Assumption 5.2(ii) yield that

$$\limsup_{n\to\infty} |W_n^*(\hat{\lambda}_n)|, \ \limsup_{n\to\infty} |\tilde{W}_n(\hat{\lambda}_n)| \le \int_{\mathbb{R}} W(x) f(x), \, dx < \infty \quad P_0\text{-a.s.},$$

so that

$$\lim_{n\to\infty} W_n^*(\hat{\lambda}_n)\hat{\lambda}_n = 0 = \lim_{n\to\infty} \tilde{W}_n(\hat{\lambda}_n)\hat{\lambda}_n \quad P_0\text{-a.s.},$$

by (7.7). From this point, (6.6) in Theorem 6.1(ii) and the specifications of B_n and $\tilde{B}_n$ lead to (7.4).

(ii) Since expansion (7.1) is valid on $[\,|\hat{\lambda}_n| < \delta]$, the conclusion follows from (7.7).

Lemma 7.2. *Suppose that Assumptions 5.1 and 5.2 are valid, let $\{\hat{\lambda}_n\}$ be a sequence of maximum likelihood estimators of λ, and define*

$$\Omega_n^{**} := \left[L_n(\hat{\lambda}_n; X_1^n) \ge L_n(\lambda; X_1^n), \ \lambda \in \mathbb{R} \right] \cap \left[\partial_\lambda^{r+1} L_n(0+; X_1^n) \ne 0 \right].$$

With this notation, assertions (i) and (ii) below occur.

(i) $\hat{\lambda}_n \ne 0$ on Ω_n^{**}.

Consequently,

(ii) $P_0[\hat{\lambda}_n \ne 0] \to 1$ as $n \to \infty$.

Proof. (i) The expansion for $L_n(\cdot; X_1^n) - L_n(0; X_1^n)$ in Theorem 6.1(i) yields that

$$\lim_{\alpha\to 0} \frac{L_n(\alpha; X_1^n) - L_n(0; X_1^n)}{|\alpha|^r \alpha} = \partial_\lambda^{r+1} L_n(0+; X_1^n).$$

It follows that if $\partial_\lambda^{r+1} L_n(0+; X_1^n) > 0$, then $L_n(\alpha; X_1^n) - L_n(0; X_1^n) > 0$ when α is positive and small enough, whereas if $\partial_\lambda^{r+1} L_n(0+; X_1^n) < 0$, then $L_n(\alpha; X_1^n) - L_n(0; X_1^n) > 0$ when $\alpha < 0$ and $|\alpha|$ is sufficiently small. Thus, $\partial_\lambda^{r+1} L_n(0+; X_1^n) \neq 0$ implies that 0 is not a maximizer of $L_n(\cdot; X_1^n)$, so that, if $\hat\lambda_n$ maximizes the average likelihood $L_n(\cdot; X_1^n)$ and $\partial_\lambda^{r+1} L_n(0+; X_1^n) \neq 0$ then $\hat\lambda_n$ is no null, *i.e.*, $\Omega_n^{**} \subset [\hat\lambda_n \neq 0]$.

(ii) Since $D^r g(0+) \neq 0$, it follows that $\partial_\lambda^{r+1} \ell(0+; X_i) = 2D^r g(0+)|X_i|^r X_i$ has a density, and then their average $\partial_\lambda^{r+1} L_n(0+; X_1^n)$ is absolutely continuous. It follows that $P_0[\partial_\lambda^{r+1} L_n(0+; X_1^n) \neq 0] = 1$, and then $P_0[\Omega_n^{**}] \to 1$, by Definition 3.1 (see Remark 3.1(ii)) and, via part (i), the conclusion follows.

Proof of Theorem 5.1. Suppose that Assumptions 5.1 and 5.2 hold, that the hypothesis $\mathcal{H}_0 : \lambda = 0$ occurs, and let $\{\hat\lambda_n\}$ be a consistent sequence of maximum likelihood estimators. In this context, define

$$\Omega_{n,*} := \left[L_n(\hat\lambda_n; X_1^n) \geq L_n(\lambda; X_1^n),\ \lambda \in \mathbb{R}\right] \cap \left[0 < |\hat\lambda_n| < \delta\right],$$

and notice that the conclusions in Lemma 7.1 occur on this event, since $\Omega_{n,*} \subset [\,|\hat\lambda| < \delta]$. Also, the consistency of $\{\hat\lambda_n\}$, Definition 3.1 and Lemma 7.2(ii) together imply that

$$\lim_{n\to\infty} P_0[\Omega_{n,*}] = 1. \tag{7.8}$$

Observe now that, on the event $\Omega_{n,*}$, the estimator $\hat\lambda_n$ is no null and maximizes $L_n(\cdot; X_1^n)$, so that the likelihood equation $\partial_\lambda L_n(\hat\lambda_n; X_1^n) = 0$ holds; see Remark 5.1(ii). Via (7.2) it follows that

$$\text{On } \Omega_{n,*}, \quad \tilde B_n \sqrt{n}\hat\lambda_n |\hat\lambda_n|^r = \sqrt{n}\frac{\partial_\lambda^{r+1} L_n(0+; X_1^n)}{(r+1)!} + \sqrt{n}\tilde A_n,$$

and then, from (7.8)

$$\begin{aligned}
\tilde B_n \sqrt{n}\hat\lambda_n |\hat\lambda_n|^r &= \sqrt{n}\frac{\partial_\lambda^{r+1} L_n(0+; X_1^n)}{(r+1)!} + \sqrt{n}\tilde A_n + o_p(1) \\
&= \sqrt{n}\frac{\partial_\lambda^{r+1} L_n(0+; X_1^n)}{(r+1)!} + O_p(\hat\lambda_n) + o_p(1) \\
&= \sqrt{n}\frac{\partial_\lambda^{r+1} L_n(0+; X_1^n)}{(r+1)!} + o_p(1)
\end{aligned}$$

where (7.3) was used to set the second equality, and the third one stems from $P_0[\hat\lambda_n \to 0] = 1$, by consistency; via (6.5) this yields that $\tilde B_n \sqrt{n}\hat\lambda_n |\hat\lambda_n|^r = O_p(1)$, and then (7.4) leads to

$$V_{r+1}\sqrt{n}\ \hat\lambda_n |\hat\lambda_n|^r = \sqrt{n}\frac{\partial_\lambda^{r+1} L_n(0+; X_1^n)}{(r+1)!} + o_p(1). \tag{7.9}$$

(i) Using that $\sqrt{n}\dfrac{\partial_\lambda^{r+1} L_n(0+;X_1^n)}{(r+1)!} \xrightarrow{\mathrm{d}} \mathcal{N}(0,V_{r+1})$ (see (5.9) and (6.5)), the above display yields

$$\sqrt{nV_{r+1}}\ \hat{\lambda}_n|\hat{\lambda}_n|^r \xrightarrow{\mathrm{d}} Z \quad \text{where } Z \text{ has standard normal distribution;} \tag{7.10}$$

since the inverse of the function $x \mapsto x|x|^r$ is the continuous mapping $x \mapsto |x|^{1/(r+1)}$ sign(x), it follows that

$$(nV_{r+1})^{1/(2(r+1))}\hat{\lambda}_n \xrightarrow{\mathrm{d}} |Z|^{1/(r+1)}\mathrm{sign}(Z).$$

(ii) Since $\sqrt{n}\hat{\lambda}_n|\hat{\lambda}_n|^r = O_p(1)$, by part (i), Lemma 7.1(ii), (7.3), (7.4) and (7.9) together yield

$$\begin{aligned}
&2n[L_n(\hat{\lambda}_n;X_1^n) - L_n(0;X_1^n)] \\
&\quad = 2n\hat{\lambda}_n|\hat{\lambda}_n|^r\left[\frac{\partial_\lambda^{r+1}L_n(0+;X_1^n)}{(r+1)!} + A_n - \frac{B_n}{2}\hat{\lambda}_n|\hat{\lambda}_n|^r\right] + o_p(1) \\
&\quad = 2\sqrt{n}\hat{\lambda}_n|\hat{\lambda}_n|^r\left[\sqrt{n}\frac{\partial_\lambda^{r+1}L_n(0+;X_1^n)}{(r+1)!} + \sqrt{n}A_n - \frac{B_n}{2}\sqrt{n}\hat{\lambda}_n|\hat{\lambda}_n|^r\right] + o_p(1) \\
&\quad = 2\sqrt{n}\hat{\lambda}_n|\hat{\lambda}_n|^r\left[\sqrt{n}\frac{\partial_\lambda^{r+1}L_n(0+;X_1^n)}{(r+1)!} + O_p(\hat{\lambda}_n)\right. \\
&\qquad\qquad \left. - \frac{V_{r+1}+o_p(1)}{2}\sqrt{n}\hat{\lambda}_n|\hat{\lambda}_n|^r\right] + o_p(1) \\
&\quad = 2\sqrt{n}\hat{\lambda}_n|\hat{\lambda}_n|^r\left[\sqrt{n}\frac{\partial_\lambda^{r+1}L_n(0+;X_1^n)}{(r+1)!} - \frac{V_{r+1}}{2}\sqrt{n}\hat{\lambda}_n|\hat{\lambda}_n|^r\right] + o_p(1) \\
&\quad = 2\sqrt{n}\hat{\lambda}_n|\hat{\lambda}_n|^r\left[V_{r+1}\sqrt{n}\hat{\lambda}_n|\hat{\lambda}_n - \frac{V_{r+1}}{2}\sqrt{n}\hat{\lambda}_n|\hat{\lambda}_n|^r\right] + o_p(1) \\
&\quad = \left(\sqrt{nV_{r+1}}\ \hat{\lambda}_n|\hat{\lambda}_n|^r\right)^2 + o_p(1);
\end{aligned}$$

together with (7.10), this yields that $2n[L_n(\hat{\lambda}_n;X_1^n) - L_n(0;X_1^n)] \xrightarrow{\mathrm{d}} Z^2$, completing the proof.

References

[1] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.*, **12**, 171–178.

[2] Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, **46**, 199–208.

[3] Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *J. R. Stat. Soc.*, Ser. B, **61**, 579–602.

[4] Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715–726.

[5] Azzalini, A. (2005). The skew-normal distribution and related multivariate families (with discussion). *Scand. J. Statist.*, **32** 159–188 (CR: 189–200).

[6] Billingsley, P. (1995). *Probability and Measure*. Third Edition, Wiley, New York.

[7] Cavazos-Cadena R. and González-Farías G. (2007). Necessary and sufficient conditions for the consistency of maximum likelihood estimators. Submitted.

[8] Chiogna, M. (2005). A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution. *Stat. Meth. & Appl.*, **14**, 331–341.

[9] DiCiccio, T. and Monti, A. C. (2004). Inferential aspects of the skew exponential power distribution. *J. Amer. Statist. Assoc.*, **99**, 439–450.

[10] Genton M. G.(Editor) (2004). *Skew Elliptical Distributions and Their Applications*. Chapman & Hall, London.

[11] Lehmann E. L. and Casella G. (1998). *Theory of Point Estimation*, Second Edition. Springer, New York.
[12] Newey, W. and McFadden, D. (1993). Estimation in large samples. In D. McFadden and R. Engler (eds), *Handbook of Econometrics*, Vol. 4. North-Holland, Amsterdam.
[13] Pewsey, A. (2000). Problems of inference for Azzalini's skew-normal distribution. *J. Appl. Statist.*, **27.7**, 859–870.
[14] Rotnitzky, A., Cox, D. R., Bottai, M., and Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli*, **6**, 243–284.
[15] Sartori, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *J. Statist. Plann. Inference*, **136**, 4259–4275.
[16] Shao, J. (1999). *Mathematical Statistics*. Springer, New York.

IMS Collections
Vol. 0 (2009) 304–325
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000016

Parametric Mixture Models for Estimating the Proportion of True Null Hypotheses and Adaptive Control of FDR

Ajit C. Tamhane[1] **and Jiaxiao Shi**[2]

Northwestern University

Abstract: Estimation of the proportion or the number of true null hypotheses is an important problem in multiple testing, especially when the number of hypotheses is large. Wu, Guan and Zhao (2006) found that nonparametric approaches are too conservative. We study two parametric mixture models (normal and beta) for the distributions of the test statistics or their p-values to address this problem. The components of the mixture are the null and alternative distributions with mixing proportions π_0 and $1-\pi_0$, respectively, where π_0 is the unknown proportion to be estimated. The normal model assumes that the test statistics from the true null hypotheses are i.i.d. $N(0,1)$ while those from the alternative hypotheses are i.i.d. $N(\delta,1)$ with $\delta \neq 0$. The beta model assumes that the p-values from the null hypotheses are i.i.d. $U[0,1]$ and those from the alternative hypotheses are i.i.d. Beta(a,b) with $a<1<b$. All parameters are assumed to be unknown. Three methods of estimation of π_0 are developed for each model. The methods are compared via simulation with each other and with Storey's (2002) nonparametric method in terms of the bias and mean square error of the estimators of π_0 and the achieved FDR. Robustness of the estimators to the model violations is also studied by generating data from other models. For the normal model, the parametric methods perform better compared to Storey's method with the EM method (Dempster, Laird and Rubin 1977) performing best overall when the assumed model holds; however, it is not very robust to significant model violations. For the beta model, the parametric methods do not perform as well because of the difficulties of estimation of parameters, and Storey's nonparametric method turns out to be the winner in many cases. Therefore the beta model is not recommended for use in practice. An example is given to illustrate the methods.

Contents

1 Introduction 305
2 Normal Model 307
2.1 Hsueh, Chen, and Kodell (HCK) Method 307
2.2 Test Statistics (TS) Method 308
2.3 EM Method 310
3 Beta Model 310
3.1 Hsueh, Chen, and Kodell (HCK) Method 311
3.2 Test Statistics (TS) Method 311

[1]Department of IEMS, Northwestern University, Evanston, IL 60208
[2]Department of Statistics, Northwestern University, Evanston, IL 60208
AMS 2000 subject classifications: Primary 62F10; secondary 62F12
Keywords and phrases: Beta model; Bias-correction; EM method; Mixture model; False discovery rate (FDR); Least squares method; Maximum likelihood method; Normal model; p-values.

3.3 EM Method . . . 312
4 Adaptive Control of FDR . . . 313
5 Simulation Results . . . 314
5.1 Simulation Results for Normal Model . . . 315
5.2 Simulation Results for Beta Model . . . 319
6 Example . . . 320
7 Concluding Remarks . . . 322
Acknowledgment . . . 323
Appendix A . . . 323
References . . . 325

1. Introduction

Suppose that m null hypotheses, $H_{01}, \ldots, H_{0m}$, are to be tested against alternatives, $H_{11}, \ldots, H_{1m}$. Let $X_1, \ldots, X_m$ be the test statistics and $p_1, \ldots, p_m$ their p-values. Throughout we assume that the X_i's and hence the p_i's are mutually independent. Suppose that some unknown number m_0 of the hypotheses are true and $m_1 = m - m_0$ are false. We wish to estimate m_0 or equivalently the proportion $\pi_0 = m_0/m$ of the true hypotheses based on the X_i's or equivalently the p_i's. The estimate $\widehat{m}_0$ is useful for devising more powerful adaptive multiple comparison procedures (MCPs) to control an appropriate type I error rate, e.g., the familywise error rate (FWE) (Hochberg and Tamhane 1987) in the Bonferroni procedure or the false discovery rate (FDR) in the Benjamini and Hochberg (1995) procedure. These procedures normally use the total number m as a conservative upper bound on the number of true hypotheses. Adaptive procedures based on $\widehat{m}_0$ are especially useful in large-scale multiplicity testing problems arising in microarray data involving m of the order of several thousands.

A number of methods have been proposed for estimating m_0 starting with Schweder and Spjøtvoll (1982); see, e.g., Hochberg and Benjamini (1990), Benjamini and Hochberg (2000), Turkheimer, Smith and Schmidt (2001), Storey (2002), Storey et al. (2004), Jiang and Doerge (2005) and Langaas et al. (2005). Many of these methods reject the p-values that differ significantly from the null $U[0,1]$ distribution as non-null and exclude them from the estimation process. Different formal or graphical tests are used for this purpose. For example, consider Storey's (2002) method with a fixed λ-level test for a sufficiently large λ (e.g., $\lambda = 0.5$) to reject any p-value $\leq \lambda$ as non-null. (It should be noted that λ is not fixed but is a tuning parameter whose value is determined from the data to minimize the mean square error of the estimate of π_0 using bootstrap.) Let $N_r(\lambda) = \sharp(p_i \leq \lambda)$ denote the number of rejected hypotheses and $N_a(\lambda) = \sharp(p_i > \lambda)$ the number of accepted hypotheses at level $\lambda \in (0,1)$. If type II errors are ignored for a sufficiently large λ then

$$E[N_a(\lambda)] \approx m_0(1-\lambda). \tag{1.1}$$

Storey's (ST) estimator is given by

$$\widehat{\pi}_0(\lambda) = \frac{N_a(\lambda)}{m(1-\lambda)} \text{ or } \widehat{m}_0(\lambda) = \frac{N_a(\lambda)}{1-\lambda}. \tag{1.2}$$

Schweder and Spjøtvoll's (1982) method visually fits a straight line through the origin to the plot of $N_a(p_{(i)}) = m - i$ vs. $1 - p_{(i)}$ $(1 \leq i \leq m)$ for large values

of the $p_{(i)}$. The slope of the fitted line is taken as an estimate of m_0 according to Equation (1.1). Because these estimators attribute all nonsignificant p-values to the true null hypotheses (type II errors are ignored) and do not explicitly model the non-null p-values, they tend to be positively biased which results in conservative adaptive control of any type I error rate.

To get a handle on type II errors, so that both the null and non-null p-values can be utilized to estimate π_0, the mixture model approach has been proposed by several authors. The mixture model differs from the setup given in the first paragraph in that the number of true hypotheses is a random variable (r.v.) and m_0 is its expected value. Specifically, let Z_i be a Bernoulli r.v. which equals 1 with probability π_0 if H_{0i} is true and 0 with probability $\pi_1 = 1 - \pi_0$ if H_{0i} is false. Assume that the Z_i $(1 \leq i \leq m)$ are i.i.d. Then the number of false hypotheses, $M_0 = \sum_{i=1}^{m} Z_i$, is a binomial r.v. with parameters m and π_0, and $E(M_0) = m_0 = m\pi_0$.

A parametric mixture model was considered by Hsueh, Chen, and Kodell (2003) (HCK). They assumed the following simple hypothesis testing setup. Suppose that all m hypotheses pertain to the means of the normal distributions with $H_{0i} : \mu_i = 0$ versus $H_{1i} : \mu_i > 0$. (HCK considered a two-sided alternative, but that is not germane to their method.) Conditional on Z_i, the test statistic $X_i \sim N(\delta_i, 1)$, where δ_i is the standardized μ_i with $\delta_i = 0$ if $Z_i = 1$ and $\delta_i = \delta > 0$ if $Z_i = 0$ where HCK assumed that δ is *known*. We refer to this model as the *normal model*, which was also used by Black (2004) to study the bias of Storey's (2002) estimator. An expression for the expected number of X_i's that are greater than any specified threshold can be derived using this setup. By plotting the corresponding observed number of X_i's against the threshold, m_0 could be estimated as the slope of the straight line through the origin using least squares (LS) regression.

The normal model is the topic of Section 2. We first extend the HCK estimation method to the unknown δ case, which is a nonlinear least squares (NLS) regression problem. Next we note that the HCK method makes use of only the *number* of X_i's that are greater than a specified threshold; it does not make use of the *magnitudes* of the X_i's. Therefore we propose two alternative methods of estimation which utilize the magnitudes of the X_i's in an attempt to obtain a better estimate of δ and thereby a better estimate of m_0. The first of these alternative methods is similar to the LS method of HCK, but uses the sample mean (instead of the number) of the X_i's that are greater than a specified threshold. We refer to it as the test statistics (TS) method. The second method is the EM method of Dempster, Laird and Rubin (1977) which finds the maximum likelihood estimators (MLEs) of the mixture distribution of the X_i's.

This normal mixture model approach in conjunction with the EM algorithm was proposed by Guan, Wu and Zhao (2004) and most recently by Iyer and Sarkar (2007). So, although the use of the EM algorithm for estimation in the context of the present problem is not new, we perform a comprehensive comparison of it with the other two new methods, and find that it performs best when the assumed model is correct, but is not robust to model violations.

In the second approach discussed in Section 3, the non-null p-values are modeled by a beta distribution with unknown parameters a and b (denoted by Beta(a, b)). We refer to this model as the *beta model*. Here we restrict to estimation methods based on p-values since the X_i's can have different null distributions. All three estimators (HCK, TS and EM) are also derived for the beta model.

We stress that both the normal and beta models are simply "working" models intended to get a handle on type II errors. We do not pretend that these models are strictly true. Therefore robustness of the estimators to the model assumptions is an

important issue. In the simulation comparisons for the normal model, robustness of the fixed δ assumption is tested by generating different δ_i's for the false hypotheses from a normal distribution. Robustness of the normal model assumption is tested by generating p_i's for the false hypotheses using the beta model and transforming them to the X_i's using the inverse normal transformation. Similarly, the robustness of the beta model is tested by generating X_i's using the normal model and transforming them to p_i's.

Adaptive control of FDR using different estimators of m_0 is the topic of Section 4. The ST, HCK, TS and EM estimators are compared in a large simulation study in Section 5. The performance measures used in the simulation study are the biases and mean square errors of the estimators of π_0 and FDR. An example illustrating application of the proposed methods is given in Section 6. Conclusions are summarized in Section 7. Proofs of some technical results are given in the Appendix.

2. Normal Model

The normal mixture model can be expressed as

$$f(x_i) = \pi_0\phi(x_i) + \pi_1\phi(x_i - \delta), \tag{2.1}$$

where $f(x_i)$ is the p.d.f. of X_i and $\phi(\cdot)$ is the p.d.f. of the standard normal distribution. Although δ will need to be estimated, we are not too concerned about its estimation accuracy since, after all, it is a parameter of a working model.

2.1. Hsueh, Chen, and Kodell (HCK) Method

Let

$$\beta(\delta, \lambda) = P_{H_{1i}}\{p_i > \lambda\} = P_{H_{1i}}\{X_i < z_\lambda\} = \Phi\left(z_\lambda - \delta\right) \tag{2.2}$$

denote the type II error probability of a test performed at level λ where $\Phi(\cdot)$ is the standard normal c.d.f. and $z_\lambda = \Phi^{-1}(1-\lambda)$. Then $E[N_r(\lambda)] = m_0\lambda + (m - m_0)[1 - \beta(\delta, \lambda)]$, and hence

$$E[N_r(\lambda)] - m\Phi\left(-z_\lambda + \delta\right) = m_0[\lambda - \Phi\left(-z_\lambda + \delta\right)]. \tag{2.3}$$

For $\lambda = p_{(i)},\ i = 1, 2, \ldots, m$, the term inside the square brackets in the R.H.S. of the above equation is

$$x_i = p_{(i)} - \Phi\left(-z_{p_{(i)}} + \delta\right) \tag{2.4}$$

and the L.H.S. can be estimated by

$$y_i = i - m\Phi\left(-z_{p_{(i)}} + \delta\right). \tag{2.5}$$

If δ is assumed to be known then we can calculate $(x_i, y_i),\ i = 1, 2, \ldots, m$, and using (2.3) fit an LS straight line through the origin by minimizing $\sum_{i=1}^{m}(y_i - m_0x_i)^2$ with respect to (w.r.t.) m_0. The LS estimator of m_0 is given by

$$\widehat{m}_0 = \frac{\sum_{i=1}^{m} x_iy_i}{\sum_{i=1}^{m} x_i^2}. \tag{2.6}$$

We first extend the HCK estimator to the unknown δ case by incorporating estimation of δ as part of the NLS problem of minimizing $\sum_{i=1}^{m}(y_i - m_0 x_i)^2$ w.r.t. m_0 and δ. The iterative algorithm for this purpose is given below. The initial values for this algorithm as well as the algorithms for the TS and EM estimators were determined by solving the following two moment equations for m_0 and δ:

$$\sum_{i=1}^{m} X_i = (m - m_0)\delta \text{ and } \sum_{i=1}^{m} X_i^2 = m_0 + (m - m_0)(\delta^2 + 1). \tag{2.7}$$

HCK Algorithm
Step 0: Compute initial estimates $\widehat{m}_0$ and $\widehat{\delta}$ by solving (2.7). Let $\widehat{\pi}_0 = \widehat{m}_0/m$.
Step 1: Set $\delta = \widehat{\delta}$ and compute (x_i, y_i), $i = 1, 2, \ldots, m$, using (2.4) and (2.5).
Step 2: Compute $\widehat{m}_0$ using (2.6) and $\widehat{\pi}_0 = \widehat{m}_0/m$.
Step 3: Find $\widehat{\delta}$ to minimize $\sum_{i=1}^{m}(y_i - m_0 x_i)^2$.
Step 4: Return to Step 1 until convergence.

Remark: One could use weighted least squares to take into account the heteroscedasticity of the y_i's. We tried this, but the resulting NLS problem was computationally much more intensive without a collateral gain in the efficiency of the estimators.

2.2. Test Statistics (TS) Method

As noted in Section 1, we hope to improve upon the HCK estimator by utilizing the information in the magnitudes of the X_i's. Toward this end we first propose an estimator analogous to the HCK estimator except that it uses the sample mean (rather than the number) of the X_i's that are significant at a specified level λ.

Define

$$S_a(\lambda) = \{i : p_i > \lambda\} = \{i : X_i < z_\lambda\} \text{ and } S_r(\lambda) = \{i : p_i \le \lambda\} = \{i : X_i \ge z_\lambda\}.$$

Then $N_a(\lambda) = |S_a(\lambda)|$ and $N_r(\lambda) = |S_r(\lambda)|$. Finally define

$$\overline{X}_a(\lambda) = \frac{1}{N_a(\lambda)} \sum_{i \in S_a(\lambda)} X_i \text{ and } \overline{X}_r(\lambda) = \frac{1}{N_r(\lambda)} \sum_{i \in S_r(\lambda)} X_i.$$

To derive the expected values of these sample means the following lemma is useful.

Lemma 1. *Define*

$$c_{0a}(\lambda) = E_{H_{0i}}\left(X_i | X_i < z_\lambda\right), c_{0r}(\lambda) = E_{H_{0i}}\left(X_i | X_i \ge z_\lambda\right),$$

and

$$c_{1a}(\delta, \lambda) = E_{H_{1i}}\left(X_i | X_i < z_\lambda\right), c_{1r}(\delta, \lambda) = E_{H_{1i}}\left(X_i | X_i \ge z_\lambda\right).$$

Then

$$c_{0a}(\lambda) = -\frac{\phi(z_\lambda)}{1 - \lambda}, c_{0r}(\lambda) = \frac{\phi(z_\lambda)}{\lambda}$$

and

$$c_{1a}(\delta, \lambda) = \delta - \frac{\phi(z_\lambda - \delta)}{\Phi(z_\lambda - \delta)}, c_{1r}(\delta, \lambda) = \delta + \frac{\phi(\delta - z_\lambda)}{\Phi(\delta - z_\lambda)}. \quad \blacksquare$$

Proof. The proof follows from the following expressions for the conditional expectations of $X \sim N(\mu, 1)$:

$$E(X|X \le x) = \mu - \frac{\phi(x-\mu)}{\Phi(x-\mu)} \text{ and } E(X|X > x) = \mu + \frac{\phi(\mu - x)}{\Phi(\mu - x)}.$$

The desired expected values of $\overline{X}_a(\lambda)$ and $\overline{X}_r(\lambda)$ are then given by the following lemma.

Lemma 2. *Let*

$$g(\pi_0, \delta, \lambda) = P\left\{Z_i = 1 | X_i < z_\lambda\right\} = \frac{\pi_0(1-\lambda)}{\pi_0(1-\lambda) + \pi_1 \Phi\left(z_\lambda - \delta\right)} \tag{2.8}$$

and

$$h(\pi_0, \delta, \lambda) = P\left\{Z_i = 1 | X_i \ge z_\lambda\right\} = \frac{\pi_0 \lambda}{\pi_0 \lambda + \pi_1 \Phi\left(-z_\lambda + \delta\right)}. \tag{2.9}$$

Then

$$E[\overline{X}_a(\lambda)] = g(\pi_0, \delta, \lambda) c_{0a}(\lambda) + [1 - g(\pi_0, \delta, \lambda)] c_{1a}(\delta, \lambda) \tag{2.10}$$

and

$$E[\overline{X}_r(\lambda)] = h(\pi_0, \delta, \lambda) c_{0r}(\lambda) + [1 - h(\pi_0, \delta, \lambda)] c_{1r}(\delta, \lambda), \tag{2.11}$$

where $c_{0a}(\lambda), c_{0r}(\lambda), c_{1a}(\delta, \lambda)$ and $c_{1r}(\delta, \lambda)$ are as given in Lemma 1.

Proof. Given in the Appendix.

To develop an estimation method analogous to the HCK method note that from (2.3) and (2.11) we get

$$E[N_r(\lambda)] E[\overline{X}_r(\lambda)] - m\Phi\left(-z_\lambda + \delta\right) c_{1r}(\delta, \lambda) = m_0 \left[\lambda c_{0r}(\lambda) - \Phi\left(-z_\lambda + \delta\right) c_{1r}(\delta, \lambda)\right]. \tag{2.12}$$

For $\lambda = p_{(i)},\ i = 1, 2, \ldots, m$, the term inside the square brackets in the R.H.S. of the above equation is

$$x_i = p_{(i)} c_{0r}(p_{(i)}) - \Phi\left(-z_{p_{(i)}} + \delta\right) c_{1r}(\delta, p_{(i)}) \tag{2.13}$$

and the L.H.S. can be estimated by

$$\begin{aligned} y_i &= i\overline{X}_r(p_{(i)}) - m\Phi\left(-z_{p_{(i)}} + \delta\right) c_{1r}(\delta, p_{(i)}) \\ &= \sum_{j=m-i+1}^{m} X_{(j)} - m\Phi\left(-z_{p_{(i)}} + \delta\right) c_{1r}(\delta, p_{(i)}). \end{aligned} \tag{2.14}$$

Then from (2.12) we see that a regression line of y_i versus x_i can be fitted through the origin with slope m_0 by minimizing $\sum_{i=1}^{m}(y_i - m_0 x_i)^2$ w.r.t. m_0 and δ. The algorithm to solve this NLS regression problem is exactly analogous to the HCK algorithm.

2.3. EM Method

Whereas the HCK and TS methods compute the LS estimators of π_0 and δ (for two different regression models), the EM method computes their MLEs. For these MLEs to exist, it is necessary that π_0 be bounded away from 0 and 1. The steps in the EM algorithm are as follows.

EM Algorithm
Step 0: Compute initial estimates $\widehat{m}_0$ and $\widehat{\delta}$ by solving (2.7). Let $\widehat{\pi}_0 = \widehat{m}_0/m$.
Step 1 (E-step): Calculate the posterior probabilities:

$$\widehat{\pi}_0(X_i) = \frac{\widehat{\pi}_0\phi(X_i)}{\widehat{\pi}_0\phi(X_i) + \widehat{\pi}_1\phi(X_i - \widehat{\delta})}$$

and $\widehat{\pi}_1(X_i) = 1 - \widehat{\pi}_0(X_i),\ i = 1, 2, \ldots, m$.
Step 2 (M-step): Calculate new estimates:

$$\widehat{\pi}_0 = \frac{\sum_{i=1}^m \widehat{\pi}_0(X_i)}{m} \text{ and } \widehat{\delta} = \frac{\sum_{i=1}^m \widehat{\pi}_1(X_i)X_i}{\sum_{i=1}^m \widehat{\pi}_1(X_i)}.$$

Step 3: Return to Step 1 until convergence.

3. Beta Model

In many applications the normal model may be inappropriate because the test statistics may not be normally distributed or different types of test statistics (e.g., normal, t, chi-square, Wlicoxon, log-rank) may be used to test different hypotheses or only the p-values of the test statistics may be available. In these cases we use the p-values to estimate π_0.

We propose to model the non-null p-values by a Beta(a, b) distribution given by

$$g(p|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1 - p)^{b-1}$$

with unknown parameters a and b with $a < 1$ and $b > 1$. This restriction is imposed in order to ensure that $g(p|a, b)$ is decreasing in p. It is well-known that the non-null distribution of the p-values must be right-skewed and generally decreasing in shape (see Hung, O'Neill, Bauer and Kohne 1997). Langaas et al. (2005) imposed the same restriction in their nonparametric estimate of the non-null distribution of p-values.

Of course, the null distribution of a p-value is Beta$(1, 1)$, i.e., the $U[0, 1]$ distribution. As in the case of the normal model, the beta model can be represented as a mixture model for the distribution of the p_i:

$$f(p_i) = \pi_0 \times 1 + \pi_1 g(p_i|a, b). \tag{3.1}$$

The parameters a and b must be estimated along with π_0. This problem is analogous to that encountered for the normal model with the difference that in addition to π_0, we have to estimate two parameters, a and b, instead of a single parameter δ. We first extend the HCK method for the normal model discussed in Section 2.1 to this beta model.

3.1. Hsueh, Chen, and Kodell (HCK) Method

Denote the type II error probability of a test performed at level λ by

$$\beta(a,b,\lambda) = P_{H_{1i}}\{p_i > \lambda\} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_\lambda^1 p^{a-1}(1-p)^{b-1}dp = 1 - I_\lambda(a,b), \tag{3.2}$$

where $I_\lambda(a,b)$ is the incomplete beta function. Put

$$x_i = p_{(i)} - I_{p_{(i)}}(a,b) \text{ and } y_i = i - mI_{p_{(i)}}(a,b). \tag{3.3}$$

Then the HCK method amounts to solving the NLS problem of minimizing $\sum_{i=1}^m (y_i - m_0 x_i)^2$ w.r.t. m_0 and (a,b) (subject to $a < 1 < b$). Gauss-Newton method (Gill et al. 1981) was used to perform minimization w.r.t. (a,b). The initial starting values for this algorithm as well as the algorithms for the TS and EM estimators described below were determined by solving the following three moment equations for m_0 and (a,b):

$$\begin{aligned}
\sum_{i=1}^m p_i &= \frac{1}{2}m_0 + \frac{a}{a+b}m_1, \\
\sum_{i=1}^m p_i^2 &= \frac{1}{3}m_0 + \frac{a(a+1)}{(a+b)(a+b+1)}m_1, \\
\sum_{i=1}^m p_i^3 &= \frac{1}{4}m_0 + \frac{a(a+1)(a+2)}{(a+b)(a+b+1)(a+b+2)}m_1.
\end{aligned} \tag{3.4}$$

3.2. Test Statistics (TS) Method

Here the TS estimator is based on the average of the "accepted" or "rejected" p-values defined as

$$\overline{p}_a(\lambda) = \frac{1}{N_a(\lambda)}\sum_{i\in S_a(\lambda)} p_i \text{ and } \overline{p}_r(\lambda) = \frac{1}{N_r(\lambda)}\sum_{i\in S_r(\lambda)} p_i.$$

Analogous to Lemma 1, we have the following lemma.

Lemma 3. *Define*

$$d_{0a}(\lambda) = E_{H_{0i}}\left(p_i | p_i > \lambda\right), d_{0r}(\lambda) = E_{H_{0i}}\left(p_i | p_i \le \lambda\right),$$

and

$$d_{1a}(a,b,\lambda) = E_{H_{1i}}\left(p_i | p_i > \lambda\right), d_{1r}(a,b,\lambda) = E_{H_{1i}}\left(p_i | p_i \le \lambda\right).$$

Then we have

$$d_{0a}(\lambda) = \frac{\lambda+1}{2}, d_{0r}(\lambda) = \frac{\lambda}{2}$$

and

$$d_{1a}(a,b,\lambda) = \frac{1-I_\lambda(a+1,b)}{1-I_\lambda(a,b)}\cdot\frac{a}{a+b}, d_{1r}(a,b,\lambda) = \frac{I_\lambda(a+1,b)}{I_\lambda(a,b)}\cdot\frac{a}{a+b}.$$

Proof. Straightforward.

The next lemma gives $E[\overline{p}_a(\lambda)]$ and $E[\overline{p}_r(\lambda)]$; its proof is exactly analogous to that of Lemma 2.

Lemma 4. *Let*

$$g(\pi_0, a, b, \lambda) = P\{Z_i = 1 | p_i > \lambda\} = \frac{\pi_0(1-\lambda)}{\pi_0(1-\lambda) + \pi_1[1 - I_\lambda(a,b)]}$$

and

$$h(\pi_0, a, b, \lambda) = P\{Z_i = 1 | p_i \leq \lambda\} = \frac{\pi_0 \lambda}{\pi_0 \lambda + \pi_1 I_\lambda(a,b)}.$$

Then

$$E[\overline{p}_a(\lambda)] = g(\pi_0, a, b, \lambda) d_{0a}(\lambda) + [1 - g(\pi_0, a, b, \lambda)] d_{1a}(a, b, \lambda) \tag{3.5}$$

and

$$E[\overline{p}_r(\lambda)] = h(\pi_0, a, b, \lambda) d_{0r}(\lambda) + [1 - h(\pi_0, a, b, \lambda)] d_{1r}(a, b, \lambda), \tag{3.6}$$

where $d_{0a}(\lambda), d_{0r}(\lambda), d_{1a}(a,b,\lambda)$ and $d_{1r}(a,b,\lambda)$ are as given in Lemma 3.

The equations for the TS estimator are derived as follows. Analogous to (2.12), we obtain

$$E[N_r(\lambda)] E[\overline{p}_r(\lambda)] - m I_\lambda(a,b) d_{1r}(a,b,\lambda) = m_0[\lambda d_{0r}(\lambda) - I_\lambda(a,b) d_{1r}(a,b,\lambda)].$$

For $\lambda = p_{(i)},\ i = 1, 2, \ldots, m$, the term in the square brackets in the R.H.S. of the above equation equals

$$x_i = \frac{p_{(i)}^2}{2} - \frac{a}{a+b} I_{p_{(i)}}(a+1, b)$$

and the L.H.S. can be estimated by

$$y_i = \sum_{j=1}^{i} p_{(j)} - \frac{a}{a+b} I_{p_{(i)}}(a+1, b).$$

The TS algorithm for the normal model can be modified to minimize $\sum_{i=1}^{m} (y_i - m_0 x_i)^2$ by replacing the minimization with respect to δ by minimization with respect to (a, b).

3.3. EM Method

The steps in the EM algorithm, which gives the MLEs of π_0 and (a, b), are as follows. As in the case of the normal model, for these MLEs to exist, it is necessary that π_0 be bounded away from 0 and 1.

Step 0: Initialize $\widehat{m}_0$ and $(\widehat{a}, \widehat{b})$ by solving (3.5). Let $\widehat{\pi}_0 = \widehat{m}_0/m$.
Step 1 (E-Step): Calculate the posterior probabilities:

$$\widehat{\pi}_0(p_i) = \frac{\widehat{\pi}_0}{\widehat{\pi}_0 + \widehat{\pi}_1 g(p_i | \widehat{a}, \widehat{b})}$$

and $\widehat{\pi}_1(p_i) = 1 - \widehat{\pi}_0(p_i),\ i = 1, 2, \ldots, m.$

Step 2 (M-Step): Calculate $\widehat{a}$ and $\widehat{b}$ as solutions of the equations (see Equations (21.1) and (21.2) in Johnson and Kotz 1970):

$$\psi(a) - \psi(a+b) = \frac{\sum_{i=1}^m \widehat{\pi}_1(p_i) \ln p_i}{\sum_{i=1}^m \widehat{\pi}_1(p_i)},$$

$$\psi(b) - \psi(a+b) = \frac{\sum_{i=1}^m \widehat{\pi}_1(p_i) \ln(1-p_i)}{\sum_{i=1}^m \widehat{\pi}_1(p_i)},$$

where $\psi(\cdot)$ is the digamma function (i.e., the derivative of the natural logarithm of the gamma function). Also calculate

$$\widehat{\pi}_0 = \frac{\sum_{i=1}^m \widehat{\pi}_0(p_i)}{m}.$$

Step 3: Return to Step 1 until convergence.

4. Adaptive Control of FDR

We now discuss the use of the estimate $\widehat{m}_0$ for adaptively controlling the FDR. The control is assumed to be strong control (Hochberg and Tamhane 1987), i.e., FDR $\leq \alpha$ for some specified $\alpha < 1$ for all possible combinations of true and false null hypotheses and the respective parameter values. Let R be the total number of rejected hypotheses and let V be the number of true hypotheses that are rejected. Benjamini and Hochberg (1995) introduced the definition

$$\text{FDR} = E\left[\frac{V}{R}\right] = E\left[\frac{V}{R}\middle| R > 0\right] P(R > 0),$$

where $0/0$ is defined as 0. Benjamini and Hochberg (1995) gave a step-up (SU) procedure that controls FDR $\leq \alpha$.

Storey (2002) considered a single-step (SS) procedure (which he referred to as the fixed rejection region method) that rejects H_{0i} if $p_i \leq \gamma$ for some common fixed threshold γ. His focus was on estimating the FDR. He proposed the following nonparametric estimator:

$$\widehat{\text{FDR}}_\lambda(\gamma) = \frac{\widehat{\pi}_0(\lambda)\gamma}{\{N_r(\lambda) \vee 1\}/m}, \tag{4.1}$$

where $\widehat{\pi}_0(\lambda)$ is given by (1.2). The solution $\widehat{\gamma}$ to the equation $\widehat{\text{FDR}}_\lambda(\gamma) = \alpha$ can be used in an MCP that tests each hypothesis at the $\widehat{\gamma}$-level. Storey, Taylor and Siegmund (2004, Theorem 3) have shown that this heuristic procedure (which uses a slightly modified estimator of π_0) controls the FDR. The heuristic procedures proposed below along the same lines (which use parametric estimators of the FDR) have not been rigorously shown to control the FDR.

We propose the following parametric estimator of the FDR:

$$\widehat{\text{FDR}}(\gamma) = \frac{\widehat{\pi}_0\gamma}{\widehat{\pi}_0\gamma + \widehat{\pi}_1[1 - \beta(\cdot, \gamma)]}, \tag{4.2}$$

where $\beta(\cdot, \gamma)$ is either $\beta(\widehat{\delta}, \gamma)$ given by (2.2) for the normal model or $\beta(\widehat{a}, \widehat{b}, \gamma)$ given by (3.2) for the beta model. To adaptively control the FDR at level α, we use

the same heuristic procedure as above except that $\widehat{\gamma}$ is obtained by setting this parametric estimator equal to α.

We may confine attention to $\alpha \leq \pi_0$ since if $\alpha > \pi_0$ then one can choose $\widehat{\gamma} = 1$, and reject all hypotheses while still controlling the FDR $= \pi_0 < \alpha$. Existence and uniqueness of $\widehat{\gamma}$ for $\alpha \in (0, \pi_0]$ is proved in the following two lemmas for the normal and beta models, respectively.

Lemma 5. *For the normal model, the solution $\widehat{\gamma}$ to the equation $\widehat{FDR}(\gamma) = \alpha$, where $\widehat{FDR}(\gamma)$ and $\beta(\widehat{\delta}, \gamma)$ are given by (4.2) and (2.2), respectively, exists and is unique for $\alpha \in (0, \pi_0]$.*

Proof. Given in the Appendix.

Lemma 6. *For the beta model, assuming $0 < \widehat{a} < 1 < \widehat{b}$, the solution $\widehat{\gamma}$ to the equation $\widehat{FDR}(\gamma) = \alpha$, where $\widehat{FDR}(\gamma)$ and $\beta(\widehat{\delta}, \gamma)$ are given by (4.2) and (3.2), respectively, exists and is unique for $\alpha \in (0, \pi_0]$.*

Proof. Given in the Appendix.

To develop an adaptive FDR-controlling procedure for the normal mixture model, Iyer and Sarkar (2007) took a somewhat different approach via the following asymptotic result of Genovese and Wasserman (2002): Assume that the p_i are independent $U[0, 1]$ when the H_{0i} are true and have a common distribution F when the H_{0i} are false. Then the nominal α-level Benjamini and Hochberg SU procedure is asymptotically (as $m \to \infty$) equivalent to Storey's SS procedure that rejects H_{0i} if $p_i \leq \widehat{\gamma}$ where $\widehat{\gamma}$ is the solution to the equation

$$F(\gamma) = \rho\gamma \text{ and } \rho = \frac{1 - \alpha\pi_0}{\alpha(1 - \pi_0)}.$$

Furthermore, since the SU procedure actually controls the FDR conservatively at $\pi_0\alpha$ level, exact control at level α is achieved by replacing α in the expression for ρ by α/π_0, which results in the following equation for γ:

$$F(\gamma) = \rho\gamma \text{ and } \rho = \frac{\pi_0(1 - \alpha)}{\alpha(1 - \pi_0)}. \tag{4.3}$$

By writing $F(\gamma) = 1 - \beta(\cdot, \gamma)$, we see that $\widehat{\text{FDR}}(\gamma) = \alpha$ and (4.3) are identical if π_0 is replaced by $\widehat{\pi}_0$ in (4.3). Iyer and Sarkar (2007) used the solution $\widehat{\gamma}$ from (4.3) in Storey's SS procedure with $F(\gamma) = \Phi(\delta - z_\gamma)$, and δ and π_0 replaced by their estimates $\widehat{\delta}$ and $\widehat{\pi}_0$ obtained from the EM method, which results in an adaptive FDR-controlling procedure, which is identical to the one proposed before.

5. Simulation Results

We compared different estimators by conducting an extensive simulation study. The ST estimator was used with $\lambda = 0.5$ throughout. The estimators were compared in terms of their accuracy of estimation of π_0 and control of FDR at a nominal $\alpha = 0.10$ using the SS procedure. The bias and mean square error (MSE) of the estimators were used as the performance measures. The results for the normal model are presented in Section 5.1 and for the beta model in Section 5.2. Throughout we used $m = 1000$ and the number of replications was also set equal to 1000. We varied

π_0 from 0.1 to 0.9 in steps of 0.1. The values $\pi_0 = 0$ and 1 were excluded because $\widehat{\pi}_0$ exhibits erratic results in these extreme cases; also FDR $= 0$ when $\pi_0 = 0$.

In each simulation run, first the random indexes of the true and false hypotheses were generated by generating Bernoulli r.v.'s Z_i. Then the respective X_i's or the p_i's were generated using the appropriate null or alternative distributions. The bias of each $\widehat{\pi}_0$ estimator was estimated as the difference between the average of the $\widehat{\pi}_0$ values over 1000 replicates and the true value of π_0. In the case of FDR, the bias was estimated as the difference between the average of the $\widehat{\text{FDR}}$ values over 1000 replicates and the nominal $\alpha = 0.10$. The MSE was computed as the sum of the square of the bias and the variance of the $\widehat{\pi}_0$ (or $\widehat{\text{FDR}}$) values averaged over 1000 replicates. The detailed numerical results are given in Shi (2006); here we only present graphical plots to save space.

5.1. Simulation Results for Normal Model

Simulations were conducted in three parts. In the first part, the true model for the non-null hypotheses was set to be the same as the assumed model by generating the X_i's for the false hypotheses from a $N(\delta, \sigma^2)$ distribution with a fixed $\delta = 2$ and $\sigma = 1$. In the other two parts of simulations, robustness of the assumed model was tested by generating the X_i's for the false hypotheses from different distributions than the assumed one. In the second part, the X_i's for the false hypotheses were generated from $N(\delta_i, \sigma^2)$ distributions where the δ_i's were themselves drawn from a $N(\delta_0, \sigma_0^2)$ distribution with $\delta_0 = 2$ and $\sigma_0 = 0.25$ corresponding to an approximate range of $[1, 3]$ for the δ_i. In the third part, the p_i's for the false hypotheses were generated from a Beta(a, b) distribution with $a = 0.5$ and $b = 2$, and the X_i's were computed using the inverse normal transformation $X_i = \Phi^{-1}(1 - p_i)$.

Results for Fixed δ: The bias and the square root of the mean square error ($\sqrt{\text{MSE}}$) of $\widehat{\pi}_0$ for ST, HCK, TS and EM estimators are plotted in Fig. 1. Note from Equation (2.3) that the bias of the ST estimator is given by

$$\text{Bias}[\widehat{\pi}_0(\lambda)] = \frac{1 - \pi_0}{1 - \lambda}\Phi(z_\lambda - \delta). \tag{5.1}$$

Also, using the fact that $N_a(\lambda)$ has a binomial distribution with number of trials m and success probability,

$$p = P\{p_i > \lambda\} = \pi_0(1 - \lambda) + (1 - \pi_0)\Phi(z_\lambda - \delta),$$

and using Equation (1.2) for $\widehat{\pi}_0(\lambda)$, we have

$$\text{Var}[\widehat{\pi}_0(\lambda)] = \frac{p(1 - p)}{m(1 - \lambda)^2}. \tag{5.2}$$

These formulae were used to compute the bias and MSE of the ST estimator instead of estimating them by simulation. Note that the bias of the ST estimator decreases linearly in π_0. The $\sqrt{\text{MSE}}$ plot for the ST estimator is also approximately linear because the bias is the dominant term in MSE. This is true whenever the alternative is fixed for all false null hypotheses.

The TS estimator does not offer an improvement over the HCK estimator, as we had hoped, and in fact performs slightly worse in terms of MSE for $\pi_0 \leq 0.5$. We suspect that this result is due to the bias introduced when the term $E[N_r(\lambda)]E[\overline{X}_r(\lambda)]$

in Equation (2.12) is estimated by $i\overline{X}_r(p_{(i)})$ for $\lambda = p_{(i)}$ because of the fact that the product of the expected values does not equal the expected value of the product of two dependent r.v.'s. The EM estimator has consistently the lowest bias and the lowest MSE.

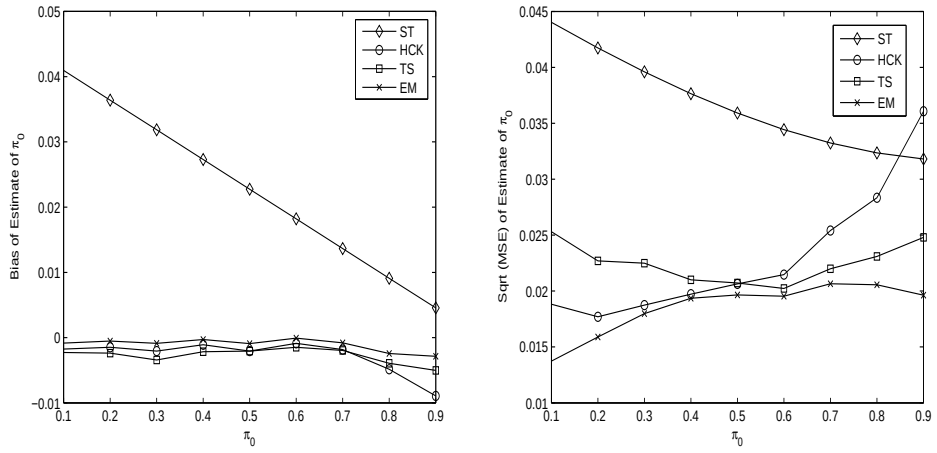


FIG 1. *Bias and $\sqrt{MSE}$ of $\widehat{\pi}_0$ for ST, HCK, TS and EM Estimators for Normal Model (Data Generated by Normal Model with Fixed δ)*

The bias and $\sqrt{\text{MSE}}$ of $\widehat{\text{FDR}}$ for ST, HCK, TS and EM estimators are plotted in Fig. 2. We see that the ST estimator leads to a large negative bias which means that, on the average, $\widehat{\text{FDR}}$ is less than the nominal $\alpha = 0.10$ resulting in conservative control of FDR. The other three estimators yield approximately the same level of control. Surprisingly, there is very little difference in the MSEs of the four estimators. The EM estimator still is the best choice with the lowest bias and the lowest MSE throughout the entire range of π_0 values.

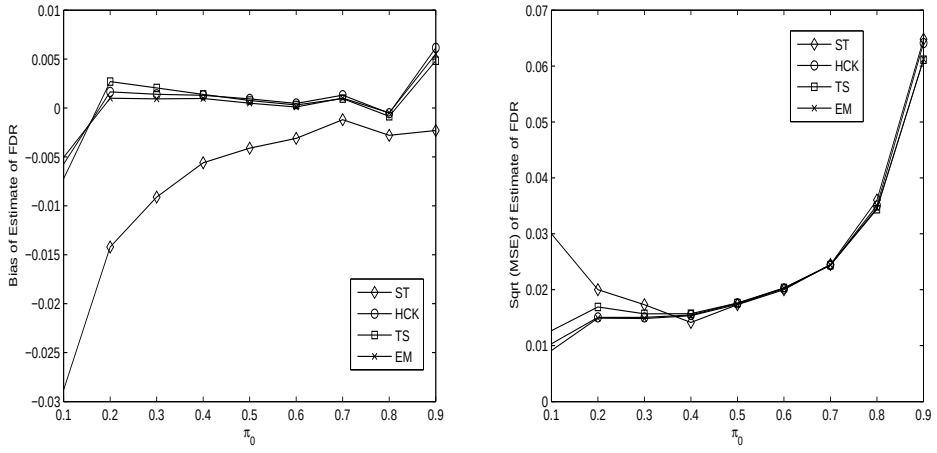


FIG 2. *Bias and $\sqrt{MSE}$ of $\widehat{FDR}$ for ST, HCK, TS and EM Estimators for Normal Model (Data Generated by Normal Model with Fixed δ)*

Results for Random δ: The bias and $\sqrt{\text{MSE}}$ of $\widehat{\pi}_0$ and of $\widehat{\text{FDR}}$ for ST, HCK, TS and EM estimators are plotted in Figs. 3 and 4, respectively. By comparing these results with those for fixed $\delta = 2$, we see that, as one would expect, there is a slight degradation in the performance of every estimator because the assumed model does not hold. The comparisons between the four estimators here are similar to those for fixed δ with the estimators ranked as EM > HCK > TS > ST.

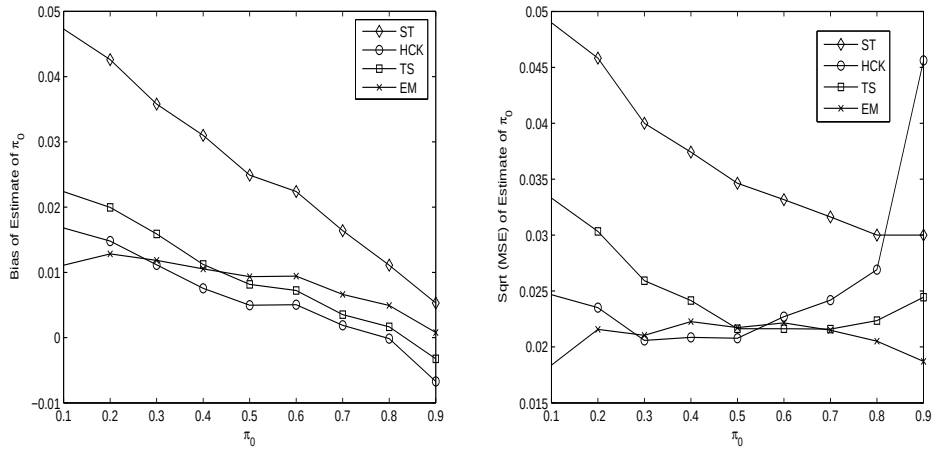


FIG 3. *Bias and $\sqrt{MSE}$ of $\widehat{\pi}_0$ for ST, HCK, TS and EM Estimators for Normal Model (Data Generated by Normal Model with Random δ)*

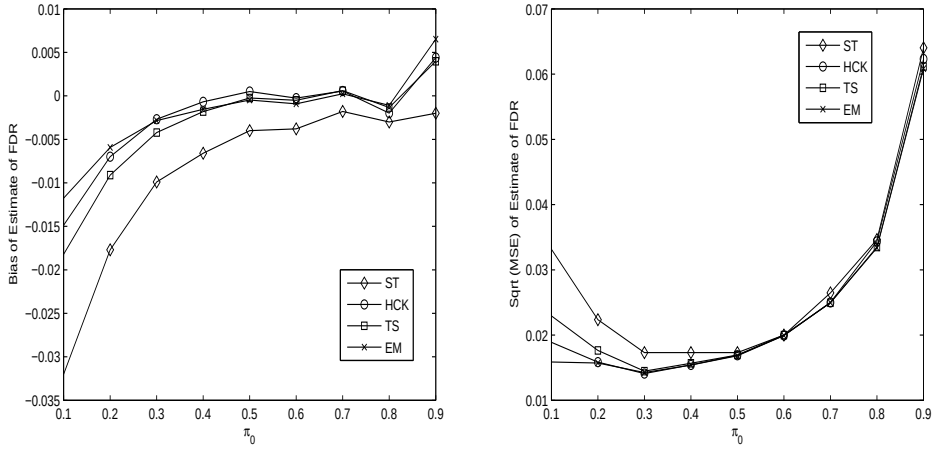


FIG 4. *Bias and $\sqrt{MSE}$ of $\widehat{FDR}$ for ST, HCK, TS and EM Estimators for Normal Model (Data Generated by Normal Model with Random δ)*

Robustness Results for Data Generated by Beta Model: The bias and $\sqrt{\text{MSE}}$ of $\widehat{\pi}_0$ and of $\widehat{\text{FDR}}$ for ST, HCK, TS and EM estimators are plotted in Figs. 5 and 6, respectively. Looking at Fig. 5 first, we see that the biases and MSEs of all four estimators are an order of magnitude higher compared to the normal model data which reflects lack of robustness.

It is interesting to note that the EM estimator is no longer uniformly best for estimating π_0. In fact, the HCK estimator has a lower bias and MSE for $0.2 \leq \pi_0 \leq 0.7$. The lack of robustness of the EM estimator is likely due to the strong dependence of the likelihood methods on distributional assumptions. On the other hand, for the least squares methods, the dependence on the assumed distribution is only through its first moment and hence is less strong. As far as control of FDR is concerned, there are not large differences between the proposed estimators. However, when $\pi_0 = 0.9$ the proposed estimators exceed the nominal FDR by as much as 0.05, while the ST estimator still controls FDR conservatively. In conclusion, the HCK estimator performs best for the middle range of π_0 values.

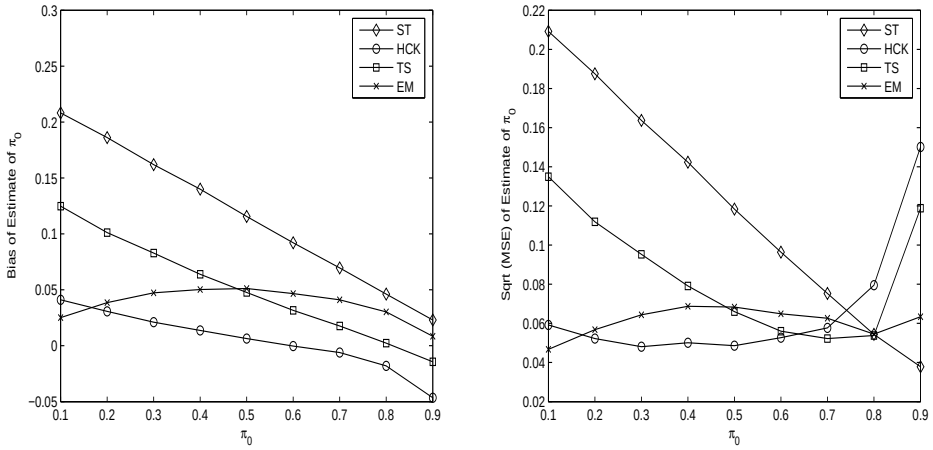


FIG 5. *Bias and $\sqrt{MSE}$ of $\widehat{\pi}_0$ for ST, HCK, TS and EM Estimators for Normal Model (Data Generated by Beta Model)*

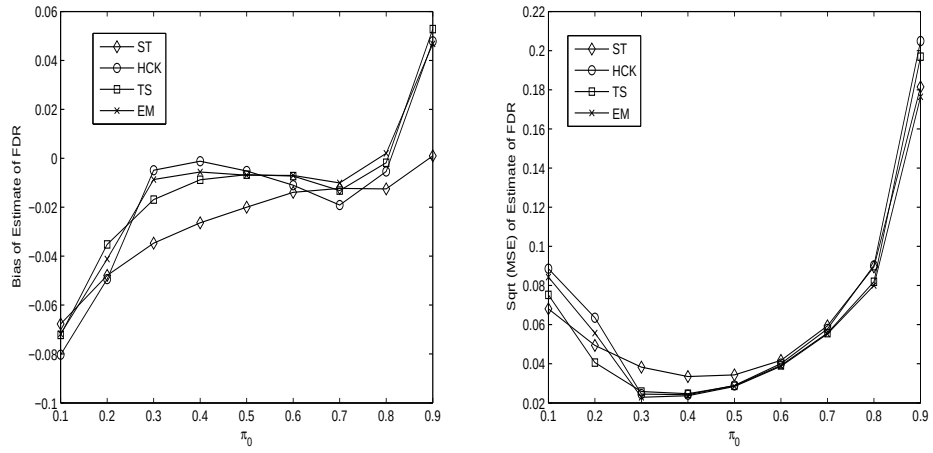


FIG 6. *Bias and $\sqrt{MSE}$ of $\widehat{FDR}$ for ST, HCK, TS and EM Estimators for Normal Model (Data Generated by Beta Model)*

5.2. Simulation Results for Beta Model

Results for Beta$(0.5, 2)$ **Data:** In this case the non-null p-values were generated from a Beta(a, b) distribution with $a = 0.5, b = 2.0$ and the null p-values were generated from the $U[0, 1]$ distribution. As before, the bias and variance of the ST estimator were not estimated from simulations, but were computed using Equations (5.1) and (5.2) with $\Phi(z_\lambda - \delta)$ replaced by $1 - I_\lambda(a, b)$. Note that the bias of the ST estimator decreases linearly in π_0 in this case as well and $\sqrt{\text{MSE}}$ decreases approximately linearly. From Fig. 7 we see that all estimators of π_0, except ST, have significant negative biases particularly over the interval $[0.2, 0.5]$ and for $\pi_0 \geq 0.7$, resulting in the achieved FDR significantly exceeding the nominal value of $\alpha = 0.10$ over the corresponding ranges of π_0 as can be seen from Fig. 8. Comparing the results here with those for the normal model with the fixed δ case, we see that the biases and MSEs of all estimators are an order of magnitude higher in the present case. The reason behind this poor performance of the beta model probably lies in the difficulty of estimating the parameters a, b of the beta distribution. Only the ST estimator controls FDR conservatively and has the smallest MSE for $0.2 \leq \pi_0 \leq 0.7$. Thus the ST estimator has the best performance since it is a nonparametric estimator (and the performance would be even better if λ is not fixed, but is used as a tuning parameter). In other words, the benefits of using a parametric model are far outweighed by the difficulty of estimating the parameters of the model resulting in less efficient estimators.

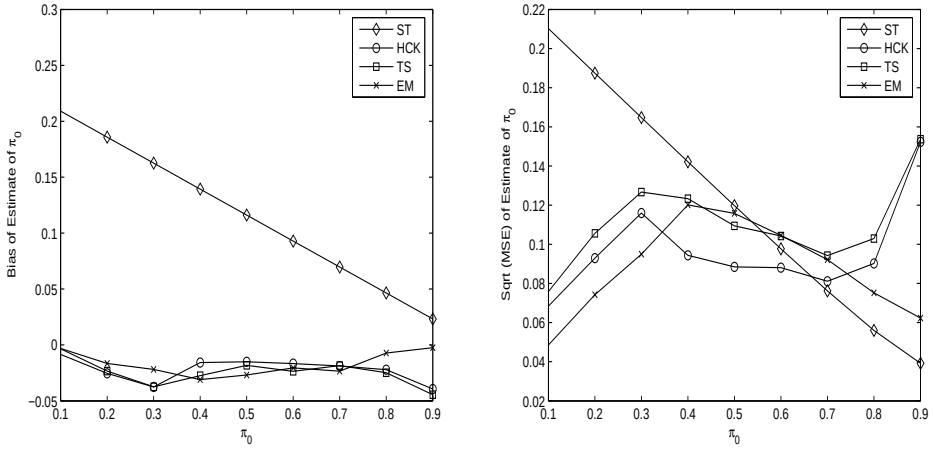


FIG 7. *Bias and* $\sqrt{MSE}$ *of* $\widehat{\pi}_0$ *for ST, HCK, TS and EM Estimators for Beta Model (Data Generated by Beta Model)*

Robustness Results for Data Generated by Normal Model: In this case we generated the data by the normal model with $N(2, 1^2)$ as the alternative distribution. The p-values were then computed and all four methods of estimation were applied. The results are plotted in Figs. 9 and 10. From these figures we see that none of the proposed estimators exhibit consistent negative bias as they did when the data were generated according to the beta model. This is somewhat surprising since one would expect these estimators to perform more poorly when the assumed model does not hold as in the present case. We also see that the EM estimator performs worse than other estimators. Thus lack of robustness of the EM estimator

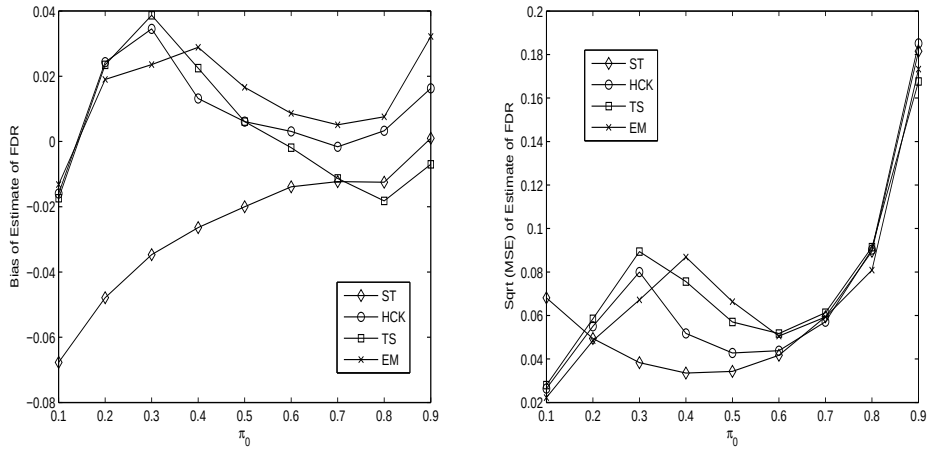


FIG 8. *Bias and* $\sqrt{MSE}$ *of* $\widehat{FDR}$ *for ST, HCK, TS and EM Estimators for Beta Model (Data Generated by Beta Model)*

to the model assumptions is demonstrated again, and for the same reason. The TS estimator generally has the lowest bias for estimating π_0 and its achieved FDR is closest to the nominal α; the ST estimator has the second best performance.

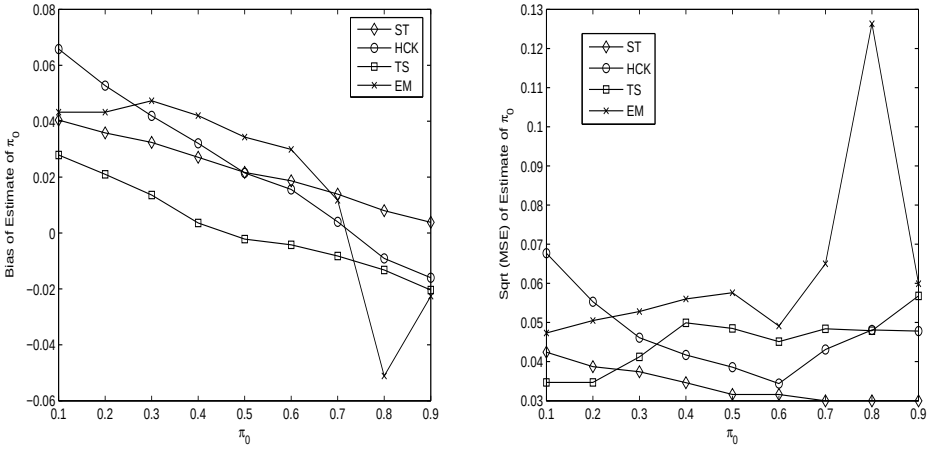


FIG 9. *Bias and* $\sqrt{MSE}$ *of* $\widehat{\pi}_0$ *for ST, HCK, TS and EM Estimators for Beta Model (Data Generated by Normal Model)*

6. Example

We consider the National Assessment of Educational Progress (NAEP) data analyzed by Benjamini and Hochberg (2000). The data pertain to the changes in the average eighth-grade mathematics achievement scores for the 34 states that participated in both the 1990 and 1992 NAEP Trial State Assessment. The raw p-values for the 34 states are listed in the increasing order in Table 2. The FWE controlling

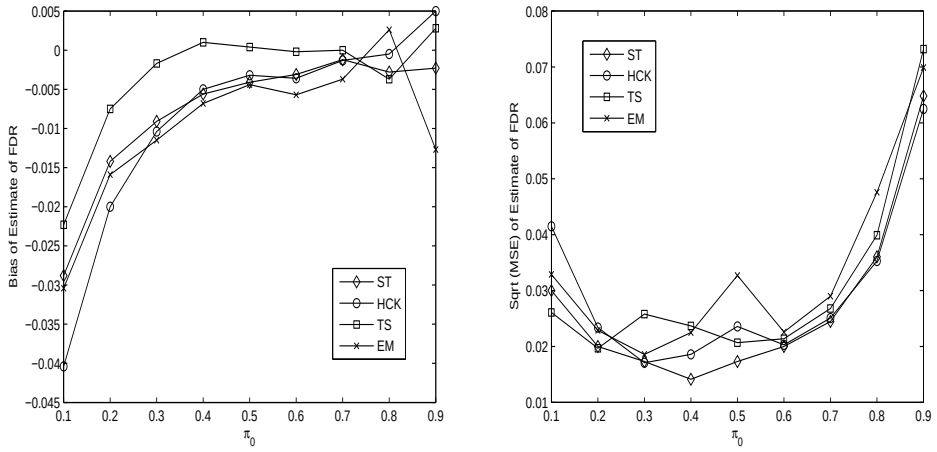


FIG 10. *Bias and $\sqrt{MSE}$ of $\widehat{FDR}$ for ST, HCK, TS and EM Estimators for Beta Model (Data Generated by Normal Model)*

Bonferroni procedure and the Hochberg (1988) procedure both identified only 4 significant results (those with p-values $\leq p_{(4)} = 0.0002$) Application of the FDR controlling non-adaptive Benjamini-Hochberg SU procedure resulted in 11 significant results. By applying their method they estimated $\widehat{m}_0 = 7$ ($\widehat{\pi}_0 = 0.2059$); using this value in the adaptive version of their procedure yielded 24 significant results.

We applied the three methods of estimation considered in this paper to these data under both the normal and beta models. The estimates $\widehat{\pi}_0$ and the associated $\widehat{\delta}$ or $(\widehat{a}, \widehat{b})$ values are given in Table 1. We see that for both models, the HCK and EM methods give smaller estimates of π_0 than does the TS method. The $\widehat{\gamma}$-values obtained by solving the equation $\widehat{\text{FDR}}(\gamma) = \alpha$ for $\alpha = 0.05$ are inversely ordered.

TABLE 1
Estimates of the Parameters for the Normal and Beta Models, Value of $\widehat{\gamma}$ and Number of Rejected Hypotheses for the HCK, TS and EM Methods

	Normal Model			Beta Model		
	HCK	TS	EM	HCK	TS	EM
$\widehat{\pi}_0$	0.1317	0.3233	0.1407	0.0096	0.1307	0.0160
$\widehat{\gamma}$	0.3163	0.0918	0.2946	1.0000	0.3092	1.0000
$\widehat{\delta}$	1.8285	2.2657	1.9221	–	–	–
$\widehat{a}$	–	–	–	0.3291	0.4474	0.3210
$\widehat{b}$	–	–	–	2.0764	3.2842	1.9313
N_r	28	21	27	34	27	34

N_r = Number of rejected hypotheses

The p-values $\leq \widehat{\gamma}$ are declared significant. From Table 2, we see that the number of significant p-values for HCK, TS and EM for the normal model are 28, 21 and 27, respectively. Thus, HCK and EM methods give more rejections than Benjamini and Hochberg's (2000) adaptive SU procedure.

Before fitting the beta mixture model, it is useful to plot a histogram of the p-values. This histogram is shown in Fig. 11. It has a decreasing shape, and assuming that the majority of the p-values are non-null, it corresponds to $a < 1$ and $b > 1$.

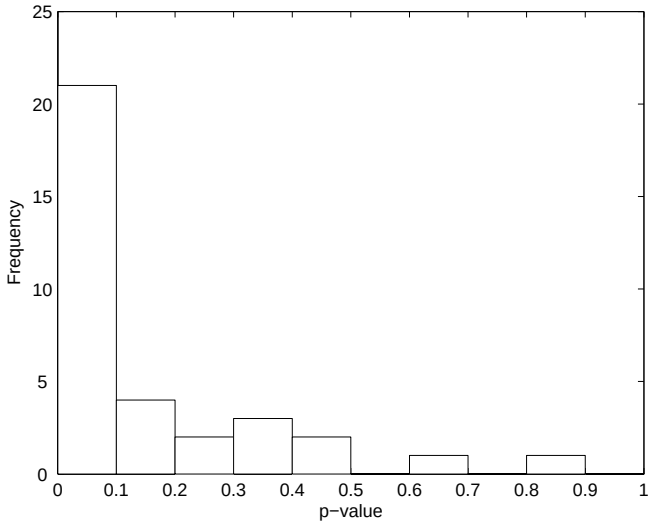


FIG 11. *Histogram of the p-Values for the NAEP Data*

HCK and EM methods yield $\widehat{\pi}_0 < \alpha = 0.05$, hence $\widehat{\gamma} = 1$ which means that all 34 hypotheses are rejected. This evidently liberal result is likely due to underestimation of π_0 using the beta model as noted in Section 5.2. The TS method yields $\widehat{\pi}_0 = 0.1307$ and $\widehat{\gamma} = 0.3092$, which are close to the estimates produced by the HCK and EM methods for the normal model and it rejects the same 27 hypotheses.

Rejections of hypotheses with large p-values will justifiably raise many eyebrows. This appears to be a problem with FDR-controlling procedures when there are many hypotheses that are clearly false (with p-values close to zero) which lowers the bar for rejection for other hypotheses. Shaffer (2005) has discussed this problem and has suggested imposing additional error controlling requirements in order to limit such dubious rejections. This is a topic for further research.

7. Concluding Remarks

In this paper we offered two different mixture models for estimating the number of true null hypotheses by modeling the non-null p-values. For each model (the normal and beta), three methods of estimation were developed: HCK, TS and EM. Generally speaking, these parametric estimators outperform (in terms of the accuracy of the estimate of π_0 and control of the FDR) the nonparametric ST estimator for the normal model but not for the beta model. The reason for this is that the normal model is easier to estimate and so the benefits of the parametric estimators are not significantly compromised by the errors of estimation. On the other hand, the beta model is difficult to estimate and so the benefits of the parametric estimators are lost. Therefore we do not recommend the use of the beta model in practice.

For normally distributed test statistics, the EM estimator generally performs best followed by the HCK and TS estimators. However, the EM estimator is not robust to the violation of the model assumptions. If the EM estimator for the normal model is applied to the data generated from the beta model or vice versa, its performance is often worse than that of the HCK estimator, and sometimes even that of the ST

TABLE 2
NAEP Trial State Assessment: Test Results for the HCK, TS and EM Methods (Normal Model)

State	p-value	HCK	TS	EM	State	p-value	HCK	TS	EM
RI	0.00000	*	*	*	NY	0.05802	*	*	*
MN	0.00002	*	*	*	OH	0.06590	*	*	*
HI	0.00002	*	*	*	CA	0.07912	*	*	*
NC	0.00002	*	*	*	MD	0.08226	*	*	*
NH	0.00180	*	*	*	WV	0.10026	*		*
IA	0.00200	*	*	*	VA	0.14374	*		*
CO	0.00282	*	*	*	WI	0.15872	*		*
TX	0.00404	*	*	*	IN	0.19388	*		*
ID	0.00748	*	*	*	LA	0.20964	*		*
AZ	0.00904	*	*	*	MI	0.23522	*		*
KY	0.00964	*	*	*	DE	0.31162	*		
OK	0.02036	*	*	*	ND	0.36890			
CT	0.04104	*	*	*	NE	0.38640			
NM	0.04650	*	*	*	NJ	0.41998			
WY	0.04678	*	*	*	AL	0.44008			
FL	0.05490	*	*	*	AR	0.60282			
PA	0.05572	*	*	*	GA	0.85628			

*Significant p-values are indicated by asterisks. For the beta model, the HCK and EM methods find all p-values significant, while the TS method finds the p-values less than $\widehat{\gamma} = 0.3093$ significant, i.e., the same as those under the EM column in this table.

estimator. The TS estimator did not improve on the HCK estimator in all cases as we had hoped. Thus our final recommendation is to use the normal model with the EM method if the test statistics follow approximately normal distributions and the HCK method otherwise. If only the p-values calculated from various types of test statistics are available then the ST method is recommended; alternatively the p-values may be transformed using the inverse normal transform and then the HCK method may be applied.

Acknowledgment

This research was partially supported by the National Heart, Lung and Blood Institute Grant 1 R01 HL082725-01A1 and the National Security Agency Grant H98230-07-1-0068. The authors are extremely grateful especially to one of the two referees who pointed out some crucial errors in the earlier version of the paper.

Appendix A

Proof of Lemma 2. We have

$$\begin{aligned} E[\overline{X}_a(\lambda)] &= E\left\{\frac{1}{N_a}\sum_{i\in S_a(\lambda)} X_i\right\} \\ &= E\left\{E\left[\frac{1}{n_a}\sum_{i\in s_a} X_i \,\middle|\, S_a(\lambda)=s_a, N_a(\lambda)=n_a\right]\right\} \\ &= E\left\{\frac{1}{n_a}\cdot n_a[g(\pi_0,\delta,\lambda)c_{0a}(\lambda)+[1-g(\pi_0,\delta,\lambda)]c_{1a}(\delta,\lambda)]\right\} \\ &= g(\pi_0,\delta,\lambda)c_{0a}(\lambda)+[1-g(\pi_0,\delta,\lambda)]c_{1a}(\delta,\lambda). \end{aligned}$$

In the penultimate step above, we have used the fact that conditionally on $X_i \le z_\lambda$, the probability that $Z_i = 1$ is $g(\pi_0, \delta, \lambda)$ and the probability that $Z_i = 0$ is $1 - g(\pi_0, \delta, \lambda)$. Furthermore, the conditional expectation of X_i in the first case is $c_{0a}(\lambda)$ and in the second case it is $c_{1a}(\delta, \lambda)$. The expression for $E[\overline{X}_r(\lambda)]$ follows similarly.

Proof of Lemma 5. By substituting for $\beta(\cdot, \gamma)$ from (2.2) and dropping carets on $\widehat{\mathrm{FDR}}(\gamma), \widehat{\pi}_0, \widehat{\pi}_1$ and $\widehat{\delta}$ for notational convenience, the equation to be solved is

$$\mathrm{FDR}(\gamma) = \frac{\pi_0}{\pi_0 + \pi_1 \Phi(\delta - z_\gamma)/\gamma} = \alpha.$$

It is easy to check that $\mathrm{FDR}(0) = 0$ and $\mathrm{FDR}(1) = \pi_0$. We shall show that $\mathrm{FDR}(\gamma)$ is an increasing function of γ which will prove the lemma. Thus we need to show that $u(\delta, \gamma) = \Phi(\delta - z_\gamma)/\gamma$ is decreasing in γ. By implicit differentiation of the equation $\Phi(z_\gamma) = 1 - \gamma$, we get

$$\frac{dz_\gamma}{d\gamma} = -\frac{1}{\phi(z_\gamma)}.$$

Hence,

$$\frac{du(\delta, \gamma)}{d\gamma} = \frac{\gamma\phi(\delta - z_\gamma) - \phi(z_\gamma)\Phi(\delta - z_\gamma)}{\gamma^2 \phi(z_\gamma)}.$$

Therefore we need to show that

$$v(\delta, \gamma) = \phi(z_\gamma)\Phi(\delta - z_\gamma) - \gamma\phi(\delta - z_\gamma) > 0 \ \ \forall \ \ \delta > 0.$$

Now $v(0, \gamma) = 0$. Therefore we must show that

$$\frac{dv(\delta, \gamma)}{d\delta} = \phi(\delta - z_\gamma)[\phi(z_\gamma) + \gamma(\delta - z_\gamma)] > 0,$$

which reduces to the condition: $w(\delta, \gamma) = \phi(z_\gamma) + \gamma(\delta - z_\gamma) > 0$. Since $w(\delta, \gamma)$ is increasing in δ, it suffices to show that

$$w(0, \gamma) = \phi(z_\gamma) - \gamma z_\gamma > 0.$$

By putting $x = z_\gamma$ and hence $\gamma = \Phi(-x)$ the above inequality becomes

$$\frac{\Phi(-x)}{\phi(x)} < \frac{1}{x},$$

which is the Mills' ratio inequality (Johnson and Kotz 1970, p. 279). This completes the proof of the lemma.

Proof of Lemma 6. By substituting for $\beta(\cdot, \gamma)$ from (3.2) and dropping carets on $\widehat{\mathrm{FDR}}(\gamma), \widehat{\pi}_0, \widehat{\pi}_1, \widehat{a}$ and $\widehat{b}$ for notational convenience, the equation to be solved is

$$\mathrm{FDR}(\gamma) = \frac{\pi_0}{\pi_0 + \pi_1 I_\gamma(a, b)/\gamma} = \alpha. \tag{A.1}$$

Note that $\mathrm{FDR}(0) = 0$ and $\mathrm{FDR}(1) = \pi_0$. To show that $\mathrm{FDR}(\gamma)$ is an increasing function of γ we need to show that $I_\gamma(a, b)/\gamma$ decreases in γ. To see this, note that the derivative of $I_\gamma(a, b)/\gamma$ w.r.t. γ is proportional to $\gamma g(\gamma|a, b) - I_\gamma(a, b)$, which is negative since the beta p.d.f. $g(\gamma|a, b)$ is strictly decreasing in γ for $a < 1$ and $b > 1$, and so $\gamma g(\gamma|a, b) < I_\gamma(a, b)$. It follows therefore that the equation $\mathrm{FDR}(\gamma) = \alpha$ has a unique solution in $\gamma \in (0, 1)$ for $\alpha \in (0, \pi_0]$.

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Ser. B*, **57**, 289–300.

[2] Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational Statistics*, **25**, 60–83.

[3] Black, M. A. (2004). A note on the adaptive control of false discovery rates. *Journal of Royal Statistical Society, Ser. B*, **66**, 297–304.

[4] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society, Ser. B*, **39**, 1–38.

[5] Finner, H. and Roters, M. (2001). On the false discovery rate and expected type I error. *Biometrical Journal*, **8**, 985-1005.

[6] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of Royal Statistical Society, Ser. B*, **64**, 499–517.

[7] Gill, P. E., Murray, W. and Wright, M. H. (1981). *Practical Optimization*. Academic Press, London and New York.

[8] Guan, Z., Wu, B. and Zhao, H. (2004). Model-based approach to FDR estimation. *Tech. Report 2004-016*. Division of Biostatistics, Univ. of Minnesota, Minneapolis, MN.

[9] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800–803.

[10] Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine*, **9**, 811–818.

[11] Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley and Sons, New York.

[12] Hsueh, H., Chen, J. J. and Kodell, R. L. (2003). Comparison of methods for estimating the number of true null hypotheses in multiple testing. *Journal of Biopharmaceutical Statistics*, **13**, 675–689.

[13] Hung, H. M., O'Neill, R. T., Bauer, P. and Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, **53**, 11–22.

[14] Iyer, V. and Sarkar, S. (2007). An adaptive single-step FDR procedure with applications to DNA microarray analysis. *Biometrical Journal*, **49**, 127–135.

[15] Jiang, H. and Doerge, R. W. (2005). Estimating the proportion of the true null hypotheses for multiple comparisons. Preprint.

[16] Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions*, **I**. John Wiley & Sons, Inc., New York.

[17] Langaas, M., Lindquist, B. H. and Ferkingstad, E. (2004). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of Royal Statistical Society, Ser. B*, **67**, 555–572.

[18] Schweder, T. and Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, **69**, 493–502.

[19] Shaffer, J. P. (2005). Multiple requirements for multiple test procedures. Paper presented at the IVth International Conference on Multiple Comparison Procedures, Shanghai, China.

[20] Shi, J. (2006). Improved Estimation of the Proportion of True Null Hypotheses with Applications to Adaptive Control of FDR and Drug Screening. *Doctoral Dissertation*. Department of Statistics, Northwestern University, Evanston, IL.

[21] Storey, J. (2002). A direct approach to false discovery rates. *Journal of Royal Statistical Society, Ser. B*, **64**, 297–304.

[22] Storey, J., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of Royal Statistical Society, Ser. B*, **66**, 187–205.

[23] Turkheimer, F. E., Smith, C. B. and Schmidt, K. (2001). Estimation of the number of 'true' null hypotheses in multivariate analysis of neuroimaging data. *NeuroImage*, **13**, 920–930.

[24] Wu, B., Guan, Z. and Zhao, H. (2006). Parametric and nonparametric FDR estimation revisited. *Biometrics*, **62**, 735–744.

IMS Collections
Vol. 0 (2009) 326–332
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000017

Bayesian Decision Theory for Multiple Comparisons

Charles Lewis[1] **and Dorothy T. Thayer**[2]

Fordham University and Educational Testing Service

Abstract: Applying a decision theoretic approach to multiple comparisons very similar to that described by Lehmann (1950, 1957a, 1957b), we introduce a loss function based on the concept of the false discovery rate (*FDR*). We derive a Bayes rule for this loss function and show that it is very closely related to a Bayesian version of the original multiple comparisons procedure proposed by Benjamini and Hochberg (1995) to control the sampling theory *FDR*. We provide the results of a Monte Carlo simulation that illustrates the very similar sampling behavior of our Bayes rule and Benjamini and Hochberg's procedure when applied to making all pair-wise comparisons in a one-way fixed effects analysis of variance setup with 10 and with 20 means.

Contents

1 Introduction 326
2 Setup 327
3 Bayes Decision Rule 328
4 A Bayesian Version of Benjamini & Hochberg's Procedure 329
5 Simulation Results 330
6 Conclusions 331
References 332

1. Introduction

A previous paper by the authors (Lewis & Thayer, 2004) considered the application of Bayesian decision theory to the multiple comparisons problem for random effects designs, following the earlier work of Shaffer (1999), Duncan (1965), and Waller and Duncan (1969). In our paper, we demonstrated that the Bayes rule for a per-comparison "0-1" loss function controls a random effects version of the false discovery rate (*FDR*), thus supporting and extending Shaffer's (1999) results.

A recent paper by Sarkar and Zhou (2008) adopts a random effects setup very similar to that of our earlier paper. Rather than considering Bayes rules, they introduce a procedure that controls the random effects *FDR* discussed by us while maximizing the random effects per-comparison power rate that we had considered. This approach produces substantial power gains over other procedures (including ours), but it "declares even small differences significant when τ [the between-groups standard deviation] is large, thereby achieving [even] greater power than the unadjusted (per-comparison) procedure for large values of τ" (Sarkar & Zhou, 2008, p. 692). We view this as a weakness, rather than a strength, of their method, as it seems to ignore the basic principle behind multiple comparisons procedures, namely

[1]Fordham University and Educational Testing Service
[2]Educational Testing Service

that making multiple inferences calls for increased conservatism relative to making a single inference.

The present study considers a more general setting for making multiple comparisons and introduces a new loss function that is more directly tied to the *FDR*. We derive a Bayes rule for this loss function and show that it is very closely related to a Bayesian version of the original multiple comparisons procedure proposed by Benjamini and Hochberg (1995) to control the sampling theory *FDR*. We provide the results of a Monte Carlo simulation that illustrates the very similar sampling behavior of our Bayes rule and Benjamini and Hochberg's procedure when applied to testing all pairwise comparisons for a one-way fixed effects analysis of variance setup with 10 and with 20 means.

2. Setup

We start with a general likelihood $p(\mathbf{y}|\theta)$, prior $p(\theta)$, and resulting posterior $p(\theta|\mathbf{y})$. Let $\psi = \mathbf{f}(\theta)$ be a vector of m "contrasts" among the elements of θ. Suppose our goal is to identify the sign of each of the elements of ψ, given $\mathbf{y}$. In the language of decision theory, for each $\psi_i,\ i = 1, \cdots, m$, we will take action a_i, with $a_i = +1$ used to indicate that we declare ψ_i to be positive, $a_i = -1$ indicating that we declare ψ_i to be negative, and $a_i = 0$ used to indicate that we are unable to determine sign of ψ_i. Although directly inspired by Williams, Jones and Tukey (1999), and Jones and Tukey (2000), this approach to (multiple) hypothesis testing has its origins in the much earlier work of Lehmann (1950, 1957a, 1957b).

To continue, we introduce two component loss functions: $L_1(\psi_i, a_i) = 1$ if the signs of ψ_i and a_i disagree, and $L_1 = 0$ otherwise (used to count the number of incorrect sign declarations); $L_2(\psi_i, a_i) = 1$ if $a_i = 0$, and $L_2 = 0$ otherwise (used to count the signs not declared). These actions and losses are very similar to those given by Lehmann (1957b, p. 549). They differ from conventional treatments of hypothesis testing in the sense that they focus on identifying the sign of each contrast and do not formally consider the possibility that the value of the contrast could be (exactly) 0. The reasonableness of this approach, compared with conventional point hypothesis testing is emphasized by Jones and Tukey (2000), among others.

We now propose a loss function that combines L_1 and L_2 as follows:

$$L_{DFDR}(\psi, \mathbf{a}) = \frac{\sum_{i=1}^{m} L_1(\psi_i, a_i)}{\max\left\{1, m - \sum_{i=1}^{m} L_2(\psi_i, a_i)\right\}} + \left(\frac{\alpha}{2}\right) \frac{\sum_{i=1}^{m} L_2(\psi_i, a_i)}{m}, \tag{1}$$

for a fixed choice of $0 < \alpha < 1$ (such as $\alpha = 0.05$). Here, *DFDR* (in notation introduced by Shaffer, 2002) stands for Directional False Discovery Rate. The first term in Equation 1 is the sample value of the *DFDR* for a given ψ and a vector of actions $\mathbf{a}$, namely the number of incorrect sign declarations, divided by the total number of signs declared (or divided by 1 if no signs are declared by $\mathbf{a}$).

The second term in Equation 1 is $\alpha/2$ times the sample proportion of signs not declared. This term may be interpreted as a sample per-comparison Type II error rate, weighted by a relative importance factor of $\alpha/2$. Using a per-comparison formulation, as well as assigning this loss component a small weight, serves to emphasize that failure to declare a sign is considered to be much less serious than declaring that sign incorrectly. This emphasis is in keeping with the concern about controlling Type I errors (at the expense of making Type II errors) in traditional

treatments of the multiple comparisons problem. The Bayes decision rule for the complete loss function in Equation 1 minimizes its posterior expected value, in this sense balancing the two types of losses against each other, with the major focus being on reducing the *DFDR*.

3. Bayes Decision Rule

To identify the actions that minimize the posterior expected value of the loss function given in Equation 1, we begin by introducing some notation.
If $\Pr(\psi_i > 0 \,|\mathbf{y}) > 0.5$, define $a_i^* = +1$ and $p_i = \Pr(\psi_i < 0 \,|\mathbf{y})$; if $\Pr(\psi_i > 0 \,|\mathbf{y}) \leqslant 0.5$, define $a_i^* = -1$ and $p_i = \Pr(\psi_i > 0 \,|\mathbf{y})$. Note that a_i^* and p_i are related by the following result: $E_{\theta|\mathbf{y}}\left[L_1(\psi_i, a_i^*)\,|\mathbf{y}\right] = p_i$. Now order the p_i so that $p_{(1)} \leqslant \cdots \leqslant p_{(m)}$. Define $\mathbf{a}^{(k)}$ for $k = 1, \cdots, m$ as $a_{(i)}^{(k)} = a_{(i)}^*$, for $i = 1, \cdots, k$, and $a_{(i)}^{(k)} = 0$, for $i = k+1, \cdots, m$. For $k = 0$, take $a_{(i)}^{(0)} = 0$, for $i = 1, \cdots, m$.

The posterior expected loss for $\mathbf{a}^{(k)}$ is given by

$$E_{\theta|\mathbf{y}}\left[L_{DFDR}\left(\psi, \mathbf{a}^{(k)}\right)|\mathbf{y}\right] = \frac{\sum_{i=1}^{k} p_{(i)}}{\max\{1,k\}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k}{m}\right). \tag{2}$$

Clearly, $\mathbf{a}^{(k)}$ minimizes the posterior expected loss among all action vectors $\mathbf{a}$ that declare exactly k signs. Let k_{DFDR} be the value of k for which the posterior expected loss given in Equation 2 is minimized. (The value of k_{DFDR} in a given setting would normally be determined by an exhaustive search over all values of $k = 0, \cdots, m$.) The Bayes decision rule for this problem is given by $\delta_{DFDR}(\mathbf{y}) = \mathbf{a}^{(k_{DFDR})}$, and the corresponding Bayes risk is

$$\begin{aligned} r(\delta_{DFDR}) &= E_{\theta,\mathbf{y}}\left[L_{DFDR}(\psi, \delta_{DFDR}(\mathbf{y}))\right] \\ &= E_{\mathbf{y}}\left[\frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{\max\{1,k_{DFDR}\}} + \left(\tfrac{\alpha}{2}\right)\left(1 - \tfrac{k_{DFDR}}{m}\right)\right]. \end{aligned} \tag{3}$$

The latter expectation in Equation 3 is taken with respect to the predictive distribution of $\mathbf{y}$, and it should be noted that the $p_{(i)}$ and, consequently, k_{DFDR} all depend on $\mathbf{y}$.

Since $E_{\theta|\mathbf{y}}\left[L_{DFDR}\left(\psi, \mathbf{a}^{(0)}\right)|\mathbf{y}\right] = \alpha/2$ for all $\mathbf{y}$, it follows that $r(\delta_{DFDR}) \leqslant \alpha/2$. Consequently,

$$E_{\theta,\mathbf{y}}\left[\frac{\sum_{i=1}^{m} L_1(\psi_i, \delta_{DFDR,i}(\mathbf{y}))}{\max\left\{1, m - \sum_{i=1}^{m} L_2(\psi_i, \delta_{DFDR,i}(\mathbf{y}))\right\}}\right] \leqslant \frac{\alpha}{2}. \tag{4}$$

Equation 4 says that a Bayesian version of the *DFDR* is bounded by $\alpha/2$ when the Bayes rule δ_{DFDR} is used. It may be worth observing that these results apply to a very general class of multiple comparison problems. Essentially the only restriction is that the set of contrasts be finite. Indeed, these do not even have to be contrasts in the usual sense of that term. They could also, for example, be a set of independent parameters that formed a family of interest.

4. A Bayesian Version of Benjamini & Hochberg's Procedure

Next, we consider the multiple comparisons procedure proposed by Benjamini and Hochberg (1995) and modified for directional testing by Williams, Jones and Tukey (1999).

However, we will translate the procedure into our Bayesian framework. Define $k_{DB\,\&\,H}$ to be the largest value of $k = 1, \cdots, m$, such that

$$p_{(k)} \leqslant \left(\frac{\alpha}{2}\right)\left(\frac{k}{m}\right),$$

with $k_{DB\,\&\,H} = 0$ if no such value of k exists. Define $\delta_{DB\,\&\,H}(\mathbf{y}) = \mathbf{a}^{(k_{DB\,\&\,H})}$.

If $k_{DB\,\&\,H} = 0$, then the posterior expected loss for $\mathbf{a}^{(k_{DB\,\&\,H})}$ is equal to $\alpha/2$. If $k_{DB\,\&\,H} > 0$, the posterior expected loss for $\mathbf{a}^{(k_{DB\,\&\,H})}$ is given by Equation 2 as

$$E_{\theta|\mathbf{y}}\left[L_{DFDR}\left(\psi, \mathbf{a}^{(k_{DB\,\&\,H})}\right)|\mathbf{y}\right] = \frac{\sum_{i=1}^{k_{DB\,\&\,H}} p_{(i)}}{k_{DB\,\&\,H}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DB\,\&\,H}}{m}\right). \tag{5}$$

From the definition of $k_{DB\,\&\,H}$, it follows that

$$p_{(i)} \leqslant \left(\frac{\alpha}{2}\right)\left(\frac{k_{DB\,\&\,H}}{m}\right) \text{ for } i = 1, \cdots, k_{DB\,\&\,H}. \tag{6}$$

Consequently, applying Inequality 6 to Equation 5, it follows that

$$E_{\theta|\mathbf{y}}\left[L_{DFDR}\left(\psi, \mathbf{a}^{(k_{DB\,\&\,H})}\right)|\mathbf{y}\right] \leqslant \left(\frac{\alpha}{2}\right)\left(\frac{k_{DB\,\&\,H}}{m}\right) + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DB\,\&\,H}}{m}\right) = \frac{\alpha}{2}.$$

Since this inequality holds for all $\mathbf{y}$, it implies that $r(\delta_{DB\,\&\,H}) \leqslant \alpha/2$, and so, just as with δ_{DFDR},

$$E_{\theta,\mathbf{y}}\left[\frac{\sum_{i=1}^{m} L_1(\psi_i, \delta_{DB\,\&\,H,i}(\mathbf{y}))}{\max\left\{1, m - \sum_{i=1}^{m} L_2(\psi_i, \delta_{DB\,\&\,H,i}(\mathbf{y}))\right\}}\right] \leqslant \frac{\alpha}{2}. \tag{7}$$

Equation 7 says that $\delta_{DB\,\&\,H}$ also controls our Bayesian *DFDR*.

This seems like an appropriate place to note that what has just been established (namely the fact that $\delta_{DB\,\&\,H}$ controls the *DFDR* for an arbitrary set of contrasts) is a Bayesian, rather than a sampling theory result. Indeed, Benjamini and Hochberg's (1995) sampling theory procedure has only been shown to control the sampling theory *FDR* in special circumstances, such as the case of independent tests. In particular, it has not been shown to provide sampling theory control of the *FDR* when making all pairwise comparisons among a set of means in a one-way, fixed effects analysis of variance setup.

Since δ_{DFDR} is a Bayes decision rule, it must be the case that $r(\delta_{DFDR}) \leqslant r(\delta_{DB\,\&\,H})$. Moreover, it is also possible to show that $k_{DFDR} \geqslant k_{DB\,\&\,H}$ for all $\mathbf{y}$. To see this, suppose the contrary: $k_{DB\,\&\,H} = k_{DFDR} + d$ with $d > 0$. By the definition of k_{DFDR}, we must have

$$\frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{\max\{1, k_{DFDR}\}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DFDR}}{m}\right) < \frac{\sum_{i=1}^{k_{DB\,\&\,H}} p_{(i)}}{k_{DB\,\&\,H}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DB\,\&\,H}}{m}\right)$$

or

$$\frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{\max\{1, k_{DFDR}\}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DFDR}}{m}\right) < \frac{\sum_{i=1}^{k_{DFDR}} p_{(i)} + \sum_{i=k_{DFDR}+1}^{k_{DFDR}+d} p_{(i)}}{k_{DFDR}+d} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DFDR}+d}{m}\right). \tag{8}$$

Note that a strict inequality has been used here, implying that, in case of ties, k_{DFDR} would be chosen to be the largest value of k that minimizes the posterior expected loss. To continue, using the definition of $k_{DB\ \&\ H}$,

$$\sum_{i=k_{DFDR}+1}^{k_{DFDR}+d} p_{(i)} \leqslant d\left(\frac{\alpha}{2}\right)\left(\frac{k_{DFDR}+d}{m}\right). \tag{9}$$

Combining Inequalities 8 and 9 gives

$$\frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{\max\{1, k_{DFDR}\}} + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DFDR}}{m}\right) < \frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{k_{DFDR}+d} + \left(\frac{\alpha}{2}\right)\left(\frac{d}{m}\right) + \left(\frac{\alpha}{2}\right)\left(1 - \frac{k_{DFDR}+d}{m}\right)$$

or

$$\frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{\max\{1, k_{DFDR}\}} < \frac{\sum_{i=1}^{k_{DFDR}} p_{(i)}}{k_{DFDR}+d}, \tag{10}$$

If $k_{DFDR} > 0$, Inequality 10 implies that $d = 0$, contrary to our initial assumption. If $k_{DFDR} = 0$, both sides of Inequality 10 would be 0, contradicting the strict inequality. Thus, we have demonstrated that $k_{DFDR} \geqslant k_{DB\ \&\ H}$ for all $\mathbf{y}$. In other words, the Bayes rule δ_{DFDR} will always declare at least as many signs as $\delta_{DB\ \&\ H}$.

5. Simulation Results

It is important to recall that the procedure actually proposed by Benjamini and Hochberg (1995) uses sampling theory p-values (one-tailed values in Williams, Jones and Tukey's 1999 version), rather than posterior tail probabilities and controls the sampling theory version of the *FDR* (or *DFDR* in Willams et al.'s version). Now consider a standard multiple comparisons problem: the one-way, fixed effects analysis of variance setup, with the ψ_i chosen to be all pair-wise differences among the group means. In this case, the relevant sampling theory and Bayesian (based on a vague prior for all parameters) p-values are identical tail probabilities from the appropriate Student's t-distribution (see, for instance, Box & Tiao, 1973, p. 140).

Tables 1 and 2 give the results of sampling theory simulations (based on 25,000 replications for each condition) of one-way, fixed effects ANOVA setups, considering all pair-wise differences for 10 evenly spaced means, and for 25 evenly spaced means. In these simulations, the within-group variance was set at 3.0 with $n = 3$ sample observations per group (so the sampling variances of the sample means are all equal to 1.0 and the within-group degrees of freedom equals 20 in the first case and 50 in

the second case). In addition, we chose $\alpha = 0.05$, with the intention of controlling the sampling theory *DFDR* at $\alpha/2 = 0.025$, although we emphasize again that there is no theory to support that control for pair-wise comparisons.

The *DFDR* values in Table 1 are sampling theory averages of the sample *DFDR* used as the first term of our loss function for the two rules with two numbers of means, and a range of spacings among the means. To make the results comparable across the two setups, we used the population standard deviation (denoted here by τ) to index the spread of the means. The Average Power values in Table 2 are sampling theory averages of the sample per-comparison correct sign declaration rate. All four quantities for a given number of means are computed from the same 25,000 replications at each spread of the means. Note that the spread is effectively given in units equal to the standard errors of the sample means. For the spread labeled "0.00+" all population mean values were set equal, and an arbitrary ordering was chosen to evaluate the "wrong sign" errors. Both procedures conservatively control the *DFDR* for all conditions considered. The Bayes rule procedure provides slightly greater per-comparison power than that of Benjamini and Hochberg in these conditions, but the actual differences are trivial.

Table 1. Sampling Theory $DFDR$ for All Pair-wise Comparisons Made Using Our Bayes Rule and Benjamini & Hochberg's Procedures.

τ : Spread	10 Means		25 Means	
of Means	δ_{DFDR}	$\delta_{DB\ \&\ H}$	δ_{DFDR}	$\delta_{DB\ \&\ H}$
0.00+	0.0204	0.0171	0.0206	0.0176
0.721	0.0062	0.0044	0.0067	0.0046
3.606	0.0005	0.0005	0.0013	0.0012
5.408	0.0001	0.0001	0.0006	0.0006
7.211	0.0000	0.0000	0.0003	0.0003
14.422	0.0000	0.0000	0.0000	0.0000

Table 2. Sampling Theory Average Power for All Pair-wise Comparisons Made Using Our Bayes Rule and Benjamini & Hochberg's Procedure.

τ : Spread	10 Means		25 Means	
of Means	δ_{DFDR}	$\delta_{DB\ \&\ H}$	δ_{DFDR}	$\delta_{DB\ \&\ H}$
0.00+	0.002	0.002	0.001	0.000
0.721	0.022	0.016	0.012	0.007
3.606	0.634	0.621	0.604	0.594
5.408	0.783	0.778	0.741	0.737
7.211	0.860	0.857	0.813	0.811
14.422	0.984	0.984	0.924	0.924

Note: 25,000 replications for each condition, $n = 3$ observations per group, within degrees of freedom $\nu = 20$ for 10 means and $\nu = 50$ for 25 means, within variance $\sigma^2 = 3.0$, $\alpha/2 = .025$.

6. Conclusions

The decision rule δ_{DFDR} has been shown to be optimal (from a Bayesian perspective) relative to the loss function L_{DFDR} for a wide class of multiple comparison problems involving sign declarations. It has also been shown to control a Bayesian version of the directional false discovery rate (*DFDR*), as has a Bayesian version of the procedure proposed by Benjamini and Hochberg ($\delta_{DB\ \&\ H}$). There is no guarantee that δ_{DFDR} or $\delta_{DB\ \&\ H}$ will control a sampling theory *DFDR* for the case

of pair-wise comparisons, although that appears to occur in the ANOVA examples given, where the two rules behave very similarly.

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

[2] Box, G. E. P. and Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Addison-Wesley, Reading, MA.

[3] Duncan, D. B. (1965). A Bayesian approach to multiple comparisons. *Technometrics*, **7**, 171–222.

[4] Jones, L. V. and Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, **5**, 411–414.

[5] Lehmann, E. L. (1950). Some principles of the theory of testing hypotheses. *The Annals of Mathematical Statistics*, **21**, 1–26.

[6] Lehmann, E. L. (1957a). A theory of some multiple decision problems. I. *The Annals of Mathematical Statistics*, **28**, 1–25.

[7] Lehmann, E. L. (1957b). A theory of some multiple decision problems. II. *The Annals of Mathematical Statistics*, **28**, 547–572.

[8] Lewis, C. and Thayer, D. T. (2004). A loss function related to the FDR for random effects multiple comparisons. *Journal of Statistical Planning and Inference*, **125**, 49–58.

[9] Sakar, S. K. and Zhou, T. (2008). Controlling Bayes directional false discovery rate in random effects model. *Journal of Statistical Planning and Inference*, **138**, 682–693.

[10] Shaffer, J. P. (1999). A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure. *Journal of Statistical Planning and Inference*, **82**, 197–213.

[11] Shaffer, J. P. (2002). Multiplicity, directional (Type III) errors and the null hypothesis. *Psychological Methods*, **7**, 356–369.

[12] Waller, R. A. and Duncan, D. B. (1969). A Bayes rule for symmetric multiple comparisons problems. *Journal of the American Statistical Association*, **64**, 1484–1503.

[13] Williams, V. S. L., Jones, L. V. and Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, **24**, 42–69.

IMS Collections
Vol. 0 (2009) 333–346
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/0000008

The Challenges of Model Objective Selection and Estimation for Ad-hoc Network Data Sets

Farinaz Koushanfar[1,2] **and Davood Shamsi**[2]

Rice University

Abstract: We introduce a new methodology for determining the difficulty of selecting the modeling objective function and estimating the parameters for an ad-hoc network data set. The method utilizes formulation of the underlying optimization problem instance that consists of an objective function and a set of constraints. The method is illustrated on real distance measurement data used for estimating the locations of wireless nodes that is the most studied and a representative problem for ad-hoc networks estimation. The properties of the data set that could affect the quality of optimization are categorized. In large optimization problems with multiple properties (characteristics) that contribute to the solution quality, it is practically impossible to analytically study the effect of each property. A number of metrics for evaluating the effectiveness of the optimization on each data set are proposed. Using the well known Plackett and Burmann fast simulation methodology, for each metric, the impact of the categorized properties of the data are determined for the specified optimization. A new approach for combining the impacts resulting from different properties on various metrics is described. We emphasize that the method is generic and has the potential to be more broadly applicable to other parameter estimation problems.

Contents

1 Introduction 334
2 Preliminaries 335
3 Metrics 337
3.1 Error Metrics 337
3.2 Objective Function (OF) Metrics 338
3.2.1 Drifting of Objective Function (OF) 338
3.2.2 Nearest Local Minimum 339
3.2.3 Measuring the Slope of OF Around the Solution 339
3.2.4 Depth of the Non-Global Local Minima 340
4 Simulation Methodology 340
5 Combining Different Ranks 341
6 Evaluation Results 342
7 Conclusion 345
Acknowledgement 345
References 346

[1]Electrical and Computer Engineering Department, Rice University, Houston, TX 77005
[2]Computer Science Department, Rice University, Houston, TX 77005

1. Introduction

Wireless adhoc networks consist of multiple wireless nodes distributed in an implementation area. To be power efficient, the wireless nodes only directly communicate with the nodes in their short local range (neighbor nodes). Communication between the non-neighbor nodes is enabled by successive usage of (one or more) local neighbors as forwarding relays. Several problems in this domain include modeling and estimation of data sets that only contain pairwise exchanged data between the neighboring nodes.

Years of continuous research in building statistical models and parameter estimation has produced a multitude of readily available methods and tools that can be employed for the problems in ad-hoc networks [10]. One limitation of the available methods is that majority of the ad-hoc modeling and estimation problems concern a large body of data and do not conform with typical assumptions needed to analytically declare the known theoretical optimality criteria. In such scenarios, the quality of the modeling and estimation methods are typically evaluated by how they perform on sets of real or simulated data. For example, some statistics of the resulting prediction error and/or a defined criterion (e.g., Bayesian information criterion (BIC)) is used for experimental evaluation of the method on the adhoc network measurements. A relevant question to answer is if indeed modeling and estimation of the pertinent data set requires introduction of a new model or an estimator, or the data could have been just as well addressed by the other known methods.

Our objective is to quantify the difficulty of model selection and estimation for a given adhoc network data set. This would provide impetus for inventing newer modeling and estimation objectives and tools that can address the difficult-to-characterize data. Simultaneously, formation of new tools would depend upon finding truly challenging network data sets that need to be addressed, as opposed to building new models that have a limited practical usage. Devising sets of challenging data would also build a foundation for comparing the various modeling objective functions and estimators for the ad-hoc network data sets. The problem of finding challenging data is complicated by variations in properties of the underlying data sets collected by different sources. This includes difference in size, format, wireless ranges, hidden covariates, and the form of noise present in the collected data. Thus, it is not easy to find unique metrics that could be used for comparison of different modeling objective functions and estimation methods.

In statistics literature, sensitivity of estimation error or other discrepancy metrics to the underlying noise in data has been widely studied for a number of modeling methods [24][3]. Also, the consistency of estimators based on a number of strong assumptions on the distribution of the data has been pursued [14]. However, no generic method or tool for determining the difficulty in modeling a data set free of imposing strong assumptions – such as normality or other closed-form distributions of noise – is available for use in adhoc networks. Note that the runtime complexity of a problem is an orthogonal concept. The complexity measures the worst-case computational time for the algorithm used for addressing the problem. Analyzing the worst-case runtime complexity does not help in understanding the complexity of characterizing a specific data set.

In adhoc network scenario, after the variable selection is done and the noise models are assumed, modeling is typically done by selecting a model form (e.g., nonlinear regression) and then estimating the model parameters on the data set. For analyzing the modeling objective function and estimation performance on the data, we study the pertinent optimization problem that consists of an objective

function (OF) and a number of constraints. The data set is considered as the input to the optimization problem. We introduce a number of metrics that measure the complexity of the optimization problem based on the problem OF properties and constraints. The challenge in most optimization problems is the existence of nonlinearities that make the solution space coarse, causing bumpiness and multiple local minimums. We propose a number of measures for the smoothness of the OF and constraints space that estimate the feasibility of reaching the global minimum.

To enable studying the effectiveness of the optimization on an adhoc network data set, one should characterize the properties of the pertinent data set. The properties are specific to each data set and the problem. In this article, we focus on the problem of finding the location of nodes (localization) in an adhoc wireless network by using erroneous mutual distance measurements between a number of node pairs. However, we emphasize that our method is generic and can be used for determining the challenge in addressing many adhoc data set model objective selection and estimation that includes forming an optimization problem. The localization problem is selected for four reasons. First, it is a very well addressed problem in the literature and there are several methods that are developed for this problem [6][9][19][2][20][26]. Second, there are a number of publicly available data sets for the measured distance data in the networks [5][21][8]. Third, the nonlinear relationship between noise in measurements data and the location of nodes makes the modeling problem extremely challenging. Fourth, localization problem is an NP-complete problem, i.e., in the worst case, there is no algorithm that can solve it in polynomial time [25][6]. Lastly, location discovery is a precursor for a number of other problems in ad hoc networks including sleeping coordination [12, 13], sensor coverage [15], and sensing exposure [16].

We characterize a number of properties of the measurement data set that could affect the quality of location estimation. Studying the interaction between the identified data properties and optimization metrics requires long simulations and analysis. We use the well-known Plackett and Burmann [23] simulation methodology to rapidly study the pairwise linear interactions of properties. A new approach for combining the impacts resulting from different properties of data on various optimization metrics is described. The sensitivity of optimization with respect to the various parameter ranks are presented.

To the best of our knowledge, this is the first work that systematically studies the impact of the adhoc network data set on the optimization employed for finding the modeling objectives and estimations. Most of the previous work are devoted to modeling and analysis of the worst case complexity. The results of our analysis could be directly used for constructing benchmarks for the problem. The proposed work aims at creating a unified framework based on real data that can help evaluation and comparison of desperate efforts that address the same problem.

The remainder of the paper is organized a follows. In the next section, location estimation problem and our notations are formally defined. In Section 3, we devise a number of metrics that are used for OF evaluation. The simulation methodology is described in Section 4. In Section 5, we illustrate how the results of different metrics can be combined. We have applied the derived method on the measurements from a real network in Section 6. We conclude in Section 7.

2. Preliminaries

In this section, we present the formal definition of the problem. We also describe the notations that are used throughout the paper.

Location estimation problem: Given a set of N nodes denoted by $V = \{v_1, v_2, \dots v_N\}$ in $\mathbb{R}^d$ $(d = 2, 3)$. For a given subset of node pairs denoted by $E \subset V \times V$, mutual distance of nodes are measured, *i.e.*, for all $(v_i, v_j) \in E$, $l(v_i, v_j) = d(v_i, v_j) + \epsilon_{i,j}$ is known; $d(v_i, v_j)$ is the Euclidean distance between the nodes v_i and v_j; $\epsilon_{i,j}$ is the distance measurement error. This error is only known if the real and measured location are both available. Moreover, there is a subset with $M(> 2)$ nodes denoted by $V_B = \{v_1, \dots v_M\}$, $V_B \subset V$ such that the nodes in V_B have their exact location information (coordinates). The nodes in the set V_B are called the *beacon* nodes.

Question: find the location of all possible nodes.

In this paper, we focus on two-dimensional networks. Extension to three-dimensional networks is straight forward. Coordinates of the node v_i are denoted by (x_i, y_i).

The location estimation problem can be formulated as an optimization problem. The goal is to find the coordinates of $K = N - M$ non-beacon nodes such that the discrepancy (error) between the measured distance data and the nodes' distances estimated from the final coordinates is minimized. In other words,

$$F_L(x_{M+1}, y_{M+1}, x_{M+2}, y_{M+2}, \dots x_N, y_N) = \sum_{(v_i,v_j)\in E} L(e_{v_i,v_j}) \tag{1}$$

$$e_{v_i,v_j} = l(v_i, v_j) - \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Where $L : \mathbb{R} \to \mathbb{R}^+$ is a function that is typically a metric (measure) of error. $F_L : \mathbb{R}^{2K} \to \mathbb{R}^+$ is known as objective function (OF) of the optimization problem.

Note that the OF of the location estimation problem is not necessarily a linear or convex function. There are a number of fast and efficient tools that are developed for linear and convex programming. However, there is no oracle algorithm that can solve all optimization problems. To find the minimum of a nonlinear problem like location estimation, there are a number of heuristic methods that may be employed. The nonlinear system solvers have a tendency to get trapped in a local minimum and do not necessarily lead to the global minimum. Although there are a variety of minimization algorithms, most of them are common in one subcomponent that starts from an initial point and follow the steepest decent to reach the minimum. The algorithms differ in how they choose the starting point, how they select the direction in the search space, and how they avoid local (non-global) minima. Thus, the shape of the OF around the global minimum is an important factor in finding the solution.

Data set: The measurement data used in this problem consists of measured distances between a number of static nodes in the plane. Measurements are noisy; there are multiple measurements for each distance. The true location of the nodes is known and will be known as the ground truth. As explained in Section 1, we sample the data set to obtain instances with specific properties.

Parameters: We will define a number of parameters that can be extracted from the data set. The sensitivity of the location estimation to the variations in each parameter will be studied. The analysis results will be used for identifying the hard instances of measurement data. Ten parameters are studied:

- P_1 – *Number of nodes* (N): the total number of nodes in the network.
- P_2 – *Number of beacons* (B): the number of beacon nodes with known locations.

- P_3 – *Mean squared error* ($\overline{\epsilon^2}$): mean squared error of distance measurements.
- P_4 – *Maximum allowed squared error* ($\text{MAX}_{\epsilon_m^2}$): the maximum squared error that can possibly exist in distance measurements.
- P_5 – *Percent of large errors* ($PER_{\epsilon_0^2}$): percentage of squared distance measurement noises that are higher than a specific value ϵ_0^2.
- P_6 – *Mean degree* ($\overline{D}$): mean degree of the nodes in the network. Degree of a node v_i is define as number of nodes that have their mutual distance to v_i.
- P_7 – *Minimum length* (MINL): possible minimum length of the measured distances between nodes in the network.
- P_8 – *Maximum length* (MAXL): possible maximum length of the measured distances between nodes in the network.
- P_9 – *Mean length* ($\bar{l}$): mean length of the measured distances between nodes in the network.
- P_{10} – *Minimum degree* (MIND): possible minimum degree of the nodes in the network.

To study the effect of the parameters, we construct a variety of network instances with different properties. The networks are constructed by selecting subsets of an implemented network. Having specific values for parameters, we use Integer Linear Programming (ILP) to extract each subset such that it meets specified conditions. To do so, we model parameter constraints as linear equalities and inequalities. Some parameters such as the mean squared error, $\overline{\epsilon^2}$, can be easily stated by linear equalities and inequalities. But some parameters such as the mean degree of the nodes, $\overline{D}$, need a mapping to be stated in linear terms. The description of the exact procedure of modeling by linear constraints is beyond the scop of this paper [8].

3. Metrics

In this section, we introduce metrics for error and OF that are used for evaluating the importance of different parameters for location estimation. Three error metrics and four OF metrics are presented. Thus, a total of twelve combined metrics are used to evaluate the importance of parameters.

3.1. Error Metrics

The three error metrics studied in this paper are: L_1, L_2, and the maximum likelihood (ML). L_1 and L_2 are the common error norms in the L_p family defined as:

$$L_p(e_{v_n,v_m} \in E) = (\sum_{(v_n,v_m)\in E} |e_{v_n,v_m}|^p)^{1/p} \quad if \quad 1 \leq p < \infty.$$

To find the error metric corresponding to ML, we need to model the noise in distance measurements. To model the noise, the probability density function (PDF) of errors, f_m, for the distance measurements should be approximated. Different methods are developed to approximate PDF of noise, f_m [8]. We have used kernel fitting that is a simple and known PDF approximation method [10]. To have the maximum likelihood estimation for the nodes' locations, we find the nodes' coordinates such that they maximize

$$\prod_{(v_n,v_m)\in E} f_m(e_{v_n,v_m}) = \exp\{\sum_{(v_n,v_m)\in E} \ln(f_m(e_{v_n,v_m}))\} \tag{2}$$

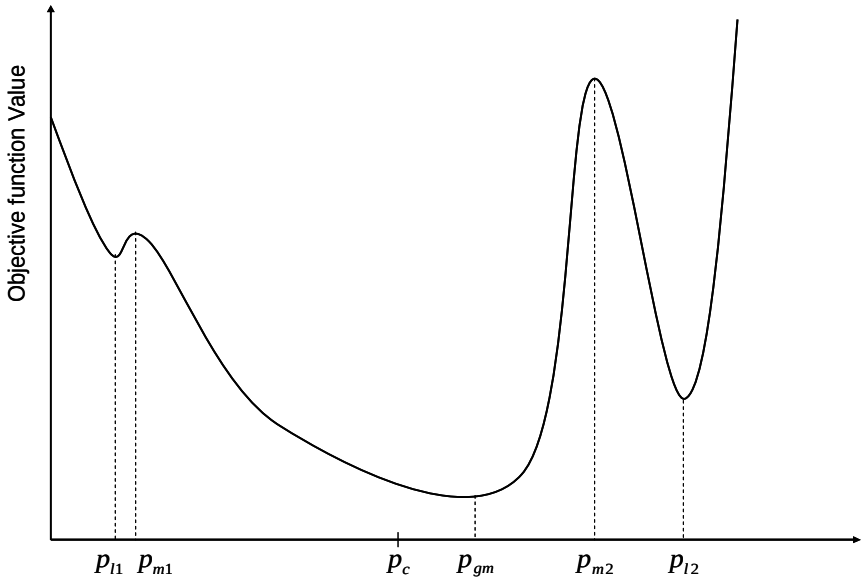


FIG 1. *Metrics and objective function (OF).*

or equivalently minimize

$$\sum_{(v_n,v_m)\in E} -\ln(f_m(e_{v_n,v_m})). \tag{3}$$

Note that we assume noise in distance measurements are independently identically distributed. Using the same notations as the Equation 1 and Equation 3, for the ML estimation we consider the following error metric:

$$L_{ML}(e_{v_n,v_m}) = -\ln(f_m(e_{v_n,v_m})). \tag{4}$$

3.2. Objective Function (OF) Metrics

We describe metrics that are used for evaluating OFs. The metrics are introduced based on the properties of OF that are effective in optimization. These metrics are such that they assign larger values to the more difficult-to-optimize OFs. For example, if one selects a convex OF, it may be possible to utilize convex programming depending on the form of the constraints. In defining the OF metrics, we assume that there is a fixed instance of location estimation data. Thus, for a fixed error metric, the OF would be fixed. Metrics of OF are denoted by $M : \mathcal{C} \rightarrow R^+$ where $\mathcal{C}$ is the functional space that contains all OFs.

3.2.1. Drifting of Objective Function (OF)

Since there is noise in distance measurements, true location of the nodes is often not the global minimum of the OF. Location of the OF's global minimum is a measure of the goodness of the OF. Figure 1 illustrates the effect of noise on the OF. For the sake of presentation simplicity, an one-dimensional OF is shown. In this figure, p_c is the correct nodes' location. However, the global minimum of the

OF is displaced at p_{gm} because of the noise. We consider the distance between p_c and its displaced location p_{gm} as an OF metric and denote it by *drifting*.

To find the drifting distance, we start from the true locations as the initial point. Next, the steepest descent direction of the OF is followed until a local minimum is reached. The Euclidean distance between the true locations and this local minimum quantifies the drifting metric (denoted by M_1) for the pertinent OF.

3.2.2. Nearest Local Minimum

Having a number of local minimums around the global minimum in an OF may cause the optimization algorithm to get trapped in one of the non-global local minimums. It is challenging to minimize such an OF since the global minimum is hard to reach. Figure 1 illustrates the phenomena. The OF has multiple local minima at points p_{m1}, p_{m1} and so on. The steepest decent method leads to the global minimum if and only if we start from a point between p_{m1} and p_{m2}. Hence, having a small distance between p_{m1} and p_{m2} would complicate the selection of the initial starting point.

We introduce a method to measure the distance of the true locations from the local minimums around the global minimum. Because of curse of dimensionality, it is not possible to find all the local minimums around the global minimum. We randomly sample the OF in multiple directions. The nearest local minimum is computed for each randomly selected direction. We statistically find the distance to the nearest local minimum by using multiple samples.

Assume $F : \mathbb{R}^{2K} \to \mathbb{R}^+$ is the OF. A random direction in $\mathbb{R}^{2K}$ is a vector in this space. Let us denote it by $v \in \mathbb{R}^{2K}$. First, we define a new function $h : \mathbb{R}^+ \to \mathbb{R}^+$ such that $h(t) = F(p_c + tv)$ where p_c is a vector containing the true locations of nodes. Second, we find the local minimum of h with the smallest positive t and denote it by t_1. We repeat this procedure for T times and find all t_i's. T is the number of samples. Finally, since it is expected that the defined metric has a larger value for more difficult-to-optimize OF, we define the nearest local minimum metric to be

$$M_2(F) = \left(\frac{1}{T} \sum_{i=1}^{T} t_i \right)^{-1}. \tag{5}$$

3.2.3. Measuring the Slope of OF Around the Solution

The Slope of OF (*i.e.*, the norm of OF's gradient) around the global minimum is a very important parameter in the convergence rate of the optimization algorithm. OFs with a small slope around the true location converge to the global minimum very slowly.

Thus, measuring the slope of the OF around the global minimum can be used to quantify the goodness of OF. Again, we measure slope of the OF in multiple random directions around the true locations, and statistically compute this metric. OFs with sharp slopes around the global minimum are easier to optimize. This can be seen in Figure 2 where the right side of the global minimum, p_{gm}, has a sharp slope. If the initial point of steepest descent algorithm is between p_{gm} and p_{m2}, it converges to the global minimum very fast. However, on the left side of global

minimum, p_{gm}, there is a gradual slope. Thus, the steepest descent algorithm would converge very slowly on the left side. We define the true locations' slope metric as

$$M_3(F) = \left(\frac{1}{T} \sum_{i=1}^{T} \text{slope in i-th random direction} \right)^{-1}. \quad (6)$$

Note that the slope of the i-th random direction, v_i, is measured at $p_{gm} + \sigma v_i$ where σ is a small number and is a user's defined criterion.

3.2.4. Depth of the Non-Global Local Minima

Optimization problems that have an OF with deep local minimums around the global minimum are difficult to solve. A number of heuristic optimization methods take advantage of the shallow local minimums to avoid non-global local minimums, e.g., simulated annealing [11]. In figure 2, avoiding the local minimum at p_{l1} is much easier than local minimum at p_{l2}.

We define the forth metric for quantifying the goodness of an OF on the data, as the depth of the non-global local minimums. We randomly select T local minimums around the true locations. Assuming that m_i is the OF value at the randomly selected local minimums, define

$$M_4(F) = \left(\frac{1}{T} \sum_{i=1}^{T} m_i \right)^{-1}. \quad (7)$$

4. Simulation Methodology

We find the linear effect of each parameter by studying all combinations of parameters. Assume each parameter has just two values. If we have k parameters then we have to study 2^k combinations that is computationally intractable. Instead, we use Plackett and Burman (PB) [23] fast simulation methodology that is a very well known method for reducing the number of simulations. Number of simulation in PB is proportional to the number of parameters. Although the PB method has not been used for the adhoc modeling and estimation problems, it was used for the simulations speedup in a number of other adhoc network problems [1][22][27][28].

In PB design, two values are assigned to each parameter: a normal value and an extreme value. The normal value is the typical value of the parameter while the extreme value is the value that is outside the typical range of the parameter. The extreme value often makes the problem either harder or easier to solve. A number of experiments with normal and extreme values of parameters are conducted.

Experiments are arranged based on a given matrix denoted by the *design matrix*. Design matrix has k columns (k is the number of parameters) and s rows where s is the number of experiments the should be set up as follows. The elements of the design matrix are either 0 or 1. We set up an experiment for each row. Values of the parameters depend on the elements on the row: 0 indicates that the normal value of the parameter is used and 1 indicates that the extreme value of the parameter is used in the experiment corresponding to the row.

Assume that we have selected an error metric, L_i, and an objective function metric, M_j. The OF itself denoted by F_{L_i} would be fixed. For each row of the

design matrix, h, we setup an experiment based on the elements of that row and measure the goodness of the objective function $M_j(F_{L_i})$ and save it in another array element denoted by $r_{i,j,h}$. The corresponding values are summed up for computing the importance factor (IF) of each parameter. For each parameter P_t, we define

$$\text{IF}_{t,i,j} = |\sum_{h=1}^{s} \alpha_{h,t} r_{i,j,h}| \tag{8}$$

where s is the number of experiments (number of rows in the design matrix), and $\alpha_{h,t}$ is 1 if the extreme value of the parameter P_t is used in the h-th experiment; otherwise, $\alpha_{h,t}$ is -1. The absolute value of IF is used to evaluate the effect of each parameter. The largest value indicates the most important parameter. For i-th error metric and j-th OF metric, $\text{IF}_{t,i,j} > \text{IF}_{u,i,j}$ means that the parameter P_t is more important than P_u. Thus, for each error metric, L_i, and for each objective function metric, M_j, we can rank parameters based on their effect on the estimated location. This ranking is denoted by $R_{i,j}$.

More precise results can be obtained by using the foldover design matrix [18]. In the foldover design matrix, all rows of the single design are repeated after its last row but 0s and 1s are exchanged in the repeated rows.

5. Combining Different Ranks

In this section, we explain how to combine the rankings of the parameters under study to obtain a global order for them. Using the ranking method in the previous section, we would have different rankings for various error metrics and OF metrics. Since there are three error metrics and four objective function metrics, there would be twelve different ranking lists for the importance of parameters; each parameter may have a different rank in each ranking list.

Each rank is obtained based on a specific property of the optimization problem. As it is explained in Section 3, for each error and objective function metric, the parameters are ranked based on the importance factor obtained from PB-design. IFs with large discrepancies lead to a stronger ranking compared to IFs with small discrepancies. Simply summing up the rankings would not necessarily determine which of the importance factors were better differentiating among the parameters.

For each ranking, $R_{i,j}$, and for each pair of parameters, P_s, P_t, we find the probability that P_s is more important than P_t. Based on the probabilities, we construct the global ranking.

Consider a specific error metric, L_i, and a specific objective function metric, M_j. Assume that the importance factor of the parameter P_t, $\text{IF}_{t,i,j}$, is normally distributed $\mathcal{N}(\lambda_{t,i,j}, \sigma^2)$. The observed value of $\text{IF}_{t,i,j}$ in a specific experiment is denoted by $if_{t,i,j}$. We normalize the importance factors to have a maximum value W. The mean of IFs are assumed to be uniformly distributed in $[0, W]$.

For each two parameters, P_s and P_t, given the BP-design experiment importance values $if_{s,i,j}$, and $if_{t,i,j}$, we find the probability: $Pr(\lambda_{s,i,j} \geq \lambda_{t,i,j} | IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j})$. The conditional probability can be written in the Bayesian format as

$$\begin{aligned} &\beta_{s,t,i,j} = Pr(\lambda_{s,i,j} \geq \lambda_{t,i,j} | IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j}) = \\ &\frac{Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j} | \lambda_{s,i,j} \geq \lambda_{t,i,j}) Pr(\lambda_{s,i,j} \geq \lambda_{t,i,j})}{Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j})}. \end{aligned} \tag{9}$$

Parameter	N_S	B_S	$\overline{\epsilon_S^2}$	$\text{MAX}_{\epsilon_m^2}$	$PER_{\epsilon_0^2}$	$\overline{D_S}$	MINL_S	MAXL_S	$\overline{l_S}$	MIND_S
Normal Value	55	12	10 (m^2)	200 (m^2)	50	10	5 (m)	40 (m)	20 (m)	4
Extreme Value	80	3	50 (m^2)	500 (m^2)	20	6	10 (m)	60 (m)	30 (m)	3

TABLE 1
Normal and extreme values for the parameters.

Since there is no prior information about the distributions of $\lambda_{s,i,j}$ and $\lambda_{t,i,j}$, we assume that $Pr(\lambda_{s,i,j} \geq \lambda_{t,i,j}) = \frac{1}{2}$. Furthermore,

$$Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = v_{t,i,j} | \lambda_{s,i,j} \geq \lambda_{t,i,j}) =$$
$$\int_{x=0}^{W} \int_{y=x}^{W} Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j} | \lambda_{s,i,j} = y, \lambda_{t,i,j} = x) \frac{dy}{W} \frac{dx}{W} =$$
$$\frac{1}{W^2} \int_{x=0}^{W} \int_{y=x}^{W} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y-if_{s,i,j})^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-if_{t,i,j})^2}{2\sigma^2}} dy\ dx. \quad (10)$$

Similarly, one can find

$$\begin{aligned} &Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j}) \\ &\quad = Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j} | \lambda_{s,i,j} \geq \lambda_{t,i,j}) Pr(\lambda_{s,i,j} \geq \lambda_{t,i,j}) \\ &\quad + Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j} | \lambda_{s,i,j} < \lambda_{t,i,j}) Pr(\lambda_{s,i,j} < \lambda_{t,i,j}). \end{aligned}$$

Now, for each parameter, P_t, we define the global importance factor, if_t,

$$if_t = \sum_{i=1}^{N_{em}} \sum_{j=1}^{N_{om}} \sum_{s=1, s \neq t}^{N_p} \beta_{s,t,i,j}. \quad (11)$$

Parameters with a larger if_t have a higher probability of being important compared to the other parameters. We sort the parameters based on their corresponding if_t values.

6. Evaluation Results

We have applied the developed method to real distance measurement data for location estimation problem. Parameters that were described in Section 2 are ranked using our methodology. We illustrate how the various ranking lists differ. Then, we combine the rankings to obtain a global ranking.

The distance measurements data from the CENS lab [4] is used to evaluate the effect of each parameter. This database is based on the real distance measurements for SH4 nodes [8]. 91 nodes are located in fixed locations. Distance measurement is done multiple times and in different days. The distance measurements are based on the time of flight (ToF)[17] of the signals. In this method, the time of flight of an acoustic signal is used to determine the distance between two nodes. It was previously shown that the noise in the measurements is strongly non-static [7]. Therefore, parametric methods based on optimizing the results according to a fixed noise distribution do not yield good location estimations.

Parameter	DOF			NLM			SMAS			DNGLM		
	L_1	L_2	ML	L_1	L_2	ML	L_1	L_2	ML	L_1	L_2	ML
N_S	4	4	2	6	5	6	2	2	2	3	2	1
B_S	2	1	1	4	2	3	4	9	4	1	4	3
$\overline{\epsilon_S^2}$	1	2	3	2	3	4	3	3	3	5	9	6
$\text{MAX}_{\epsilon_m^2}$	6	8	7	9	10	10	6	4	5	7	7	8
$PER_{\epsilon_0^2}$	7	9	10	10	8	5	7	8	7	9	10	9
$\overline{D_S}$	3	3	4	1	1	1	1	1	1	2	1	2
MINL_S	8	6	8	7	6	9	8	5	10	4	3	10
MAXL_S	10	10	9	5	9	8	10	10	8	10	5	4
$\overline{l_S}$	9	5	5	8	7	7	9	7	9	6	6	5
MIND_S	5	7	6	3	4	2	5	6	6	8	8	7

TABLE 2
Importance of different parameters for different objective functions and metrics.

We have used Integer Linear Programming (ILP) to sample the database for drawing instances with specific properties. In each experiment, the PB-design matrix implies a specific value for each parameter. Extreme and normal values for parameters are shown in Table 1. The values are determined based on the real measurements' error. In all experiments, ϵ_0^2 is equal to $20(m^2)$.

The following abbreviations are used in this section.

- ML : Maximum Likelihood
- DOF : Drifting of the Objective Function (M_1)
- NLM : Nearest Local Minimum (M_2)
- SMAS : Slope Measurement Around the Solution (M_3)
- DNGLM: Depth of Non-Global Local Minimum (M_4)

Table 2 shows the result of PB-based evaluations. Each parameter is ranked based on the specific error metric and the specific OF metric. It can be seen that a specific parameter has different rankings under various error metrics and OF metrics. For example, the total number of nodes, N_S, is ranked 1, 2, 3, 4, 5, and 6 in different cases. Thus, a specific parameter does not have the same importance under various metrics. It can be seen that the number of nodes, N_S, and the number of beacons, B_S, are the two important parameters in most evaluations; $PER_{\epsilon_0^2}$ and MAXL_S have overall low rankings.

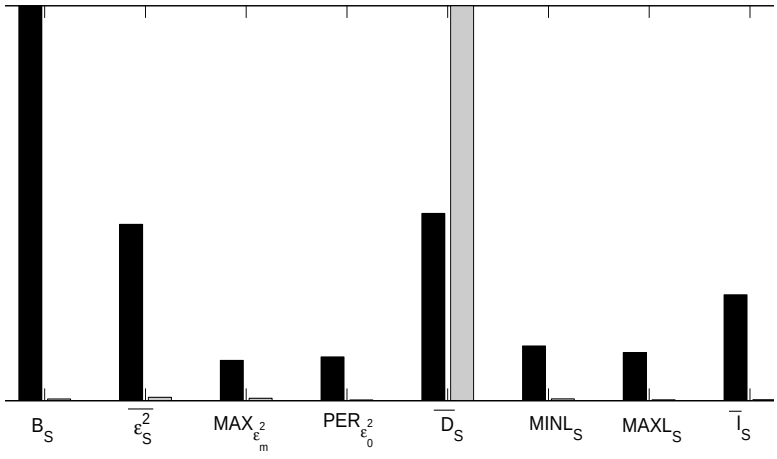


FIG 2. *Importance of different parameters for different objective functions and metrics.*

Parameter	N_S	B_S	$\overline{\epsilon_S^2}$	$\text{MAX}_{\epsilon_m^2}$	$PER_{\epsilon_0^2}$	$\overline{D_S}$	MINL_S	MAXL_S	$\overline{l_S}$	MIND_S
N_S	0	0.071	0.417	0.725	0.716	0.403	0.708	0.691	0.598	0.748
B_S	0.929	0	0.899	0.993	0.984	0.884	0.977	0.981	0.939	0.984
$\overline{\epsilon_S^2}$	0.583	0.101	0	0.787	0.786	0.476	0.756	0.754	0.660	0.798
$\text{MAX}_{\epsilon_m^2}$	0.275	0.007	0.213	0	0.490	0.193	0.464	0.477	0.354	0.515
$PER_{\epsilon_0^2}$	0.284	0.016	0.214	0.510	0	0.202	0.469	0.499	0.357	0.528
$\overline{D_S}$	0.597	0.116	0.524	0.807	0.798	0	0.785	0.795	0.678	0.821
MINL_S	0.292	0.023	0.244	0.536	0.531	0.215	0	0.519	0.392	0.545
MAXL_S	0.309	0.019	0.246	0.523	0.501	0.205	0.481	0	0.371	0.537
$\overline{l_S}$	0.402	0.061	0.340	0.646	0.643	0.322	0.608	0.629	0	0.671
MIND_S	0.252	0.016	0.202	0.485	0.472	0.179	0.455	0.463	0.329	0

TABLE 3

Drifting of objective function and L_2 metric: $Pr(\lambda_{i,j,s} \geq \lambda_{i,j,t} | V_{i,j,s} = v_{i,j,s}, V_{i,j,t} = v_{i,j,t})$ where the first column is P_s and the first row is P_t.

The comparative ranks of parameter pairs tend to vary as well. Figure 2 shows the normalized importance factor (IF) for two cases: DOF and SMAS with L_2 error metric. For DOF, the number of beacons B_S is strongly more important than the mean squared error $\overline{\epsilon_S^2}$. The mean degree of nodes, $\overline{D_S}$, is weakly more important than the mean squared error $\overline{\epsilon_S^2}$. The same behavior can be seen in SMAS. From our visual inspections, the number of nodes N_S and the mean degree of nodes $\overline{D_S}$ are the most important while others almost have the same importance factor (IF). The ranks of the mean squared error $\overline{\epsilon_S^2}$ and maximum edge length MAXL_S are 3 and 10 respectively. However, their importance factors are very close.

The discrepancy in the rank and comparative ranks confirms our postulation that averaging the parameter ranks is not the best way for combining them. Thus, we use the combining method that was introduced in Section 5. The probability comparisons for the values in Figure 2 are shown in Tables 3 and 4. The tables compare the importance of parameters. For example, for the DOF-L_2, Figure 2 states that B_S is strongly more important than $PER_{\epsilon_0^2}$. Table 3 shows that the probability that the mean of B_S is larger than the mean of $PER_{\epsilon_0^2}$ is 0.984. Similarly, $\text{MAX}_{\epsilon_m^2}$ and $PER_{\epsilon_0^2}$ have approximately the same importance. The probability that the mean of $\text{MAX}_{\epsilon_m^2}$ is larger than the mean of $PER_{\epsilon_0^2}$ is 0.49. This probability value is close to 0.5, meaning that there is not enough information to compare the values.

Table 4 compares the importance factors of SMAS for the L_2 error metric. Table 4 confirms the result. The rows corresponding to N_S, and $\overline{D_S}$ have values close to 1 confirming the high importance of the two parameters. When comparing other parameters, the probability that one parameter is greater than the other is about 0.5. It confirms our previous postulation that simple rankings are not sufficient for concluding the global parameter ordering and the importance factors are significant as well.

The global ranking based on the introduced combining method in Section 5 is shown in Table 5. The table indicates that the mean degree of nodes $\overline{D_S}$ is the most important parameter. This result is consistent with Table 2 where the mean degree of nodes $\overline{D_S}$ is the most important parameter in the seven scenarios.

The global ranking results could be used to improve the goodness of location estimations in ad-hoc networks. To deploy a network or on an already deployed network, one could exploit the results by considering the analyzed effect of each parameter on the estimated location's accuracy. Based on the constraints of the problem, the best parameters for improving the estimated locations could be determined. For example, when there are limitations for the mean degree of the graph,

Parameter	N_S	B_S	$\overline{\epsilon^2_S}$	$\text{MAX}_{\epsilon^2_m}$	$PER_{\epsilon^2_0}$	$\overline{D_S}$	MINL_S	MAXL_S	$\overline{l_S}$	MIND_S
N_S	0	0.947	0.947	0.936	0.944	0.285	0.939	0.931	0.937	0.958
B_S	0.053	0	0.493	0.506	0.504	0.017	0.501	0.496	0.500	0.504
$\overline{\epsilon^2_S}$	0.053	0.507	0	0.509	0.505	0.018	0.504	0.516	0.507	0.511
$\text{MAX}_{\epsilon^2_m}$	0.064	0.494	0.491	0	0.505	0.017	0.504	0.506	0.499	0.499
$PER_{\epsilon^2_0}$	0.056	0.496	0.495	0.495	0	0.017	0.492	0.496	0.502	0.500
$\overline{D_S}$	0.715	0.983	0.982	0.983	0.983	0	0.975	0.984	0.974	0.980
MINL_S	0.061	0.499	0.496	0.496	0.508	0.025	0	0.506	0.501	0.488
MAXL_S	0.069	0.504	0.484	0.494	0.504	0.016	0.494	0	0.502	0.494
$\overline{l_S}$	0.063	0.500	0.493	0.501	0.498	0.026	0.499	0.498	0	0.505
MIND_S	0.042	0.496	0.489	0.501	0.500	0.020	0.512	0.506	0.495	0

TABLE 4
SMAS and L_2 metric: $Pr(\lambda_{i,j,s} \geq \lambda_{i,j,t} | V_{i,j,s} = v_{i,j,s}, V_{i,j,t} = v_{i,j,t})$ where the first column is P_s and the first row is P_t.

Parameter	N_S	B_S	$\overline{\epsilon^2_S}$	$\text{MAX}_{\epsilon^2_m}$	$PER_{\epsilon^2_0}$	$\overline{D_S}$	MINL_S	MAXL_S	$\overline{l_S}$	MIND_S
Rank	2	3	4	8	10	1	6	9	7	5

TABLE 5
Global ranks.

one can increase the number of nodes in the network to increase the accuracy of the estimated location. Note that, changing one parameter typically only improves the accuracy up to a certain point; further changing the parameter would not yield an improvement in the estimation accuracy.

7. Conclusion

We introduce a systematic methodology for determining the challenge of modeling a pertinent adhoc network data set. The complex modeling problem is studied as an instance of a nonlinear optimization problem that consists of an objective function (OF) and a set of constraints. The data set is the optimization input and the estimated model is the output. We characterize the input by a set of its characteristic parameters. We define four new metrics that can be used to evaluate the goodness of an input for being optimized by a specific OF. The introduced metrics are: (1) drifting of the OF, (2) distance to the nearest local minimum, (3) the slope of the OF around the solution, and (4) the depth of the non-global local minima. We employ Plackett and Burmann simulation methodology to systematically evaluate the linear impact of various input parameters under each metric. Finally, we present a method for combining the effect of parameters under different metrics to determine the global impact of each parameter. We utilize the new methodology for estimating the locations of the nodes in an ad-hoc network where the distance measurement data is available. Three common forms of OF are considered: L_1, L_2 and L_∞. Our evaluations show that the mean degree on the nodes and the number of nodes in the network are the two most important parameters for estimating the locations.

Acknowledgement

This work is partly supported by the National Science Foundation (NSF) CAREER Award under grant number 0644289.

References

[1] Barrett, C., Marathe, A., Marathe, M. V. and Drozda, M. (2002). Characterizing the interaction between routing and MAC protocols in ad-hoc networks. *MobiHoc*, 92-103.

[2] Biswas, P. and Ye, Y. (2004). Semidefinite programming for ad hoc wireless sensor network localization. *Information Processing in Sensor Networks (IPSN)*, 2673-2684.

[3] Bullard, C. and Sebald, A. (1988). Monte carlo sensitivity analysis of input-output models. *The Review of Economics and Statistics*, **22.4**, 708-712.

[4] *CENS: Center for Embedded Networked Sensing at UCLA*. http://research.cens.ucla.edu/.

[5] *The Cricket indoor location system at MIT*. http://cricket.csail.mit.edu.

[6] Eren, T., Goldenberg, D., Whiteley, W., Yang, Y. R., Morse, A., Anderson, B. and Belhumeur, P. (2004). Rigidity, computation, and randomization in network localization. *INFOCOM*, 2673-2684.

[7] Feng, J., Girod, L. and Potkonjak, M. (2006). Consistency-based on-line localization in sensor networks. *Distributed Computing in Sensor Systems (DCOSS)*, 529-545.

[8] Feng, J., Girod, L. and Potkonjak, M. (2006). Location discovery using data-driven statistical error modeling. *INFOCOM*, 1-14.

[9] Goldenberg, D., Krishnamurthy, A., Maness, W., Yang, Y., Young, A., Morse, A. and Savvides, A. (2006). Network localization in partially localizable networks. *INFOCOM*, 313-326.

[10] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, Germany.

[11] Kirkpatrick, S., Gelatt, C. D. Jr. and Vecchi, M.P. (1983). Optimization by simulated annealing. *Science*, **220.4598**, 671-680.

[12] Koushanfar, F., Davare, A., Nguyen, D., Sangiovanni-Vincentelli, A. and Potkonjak, M. (2007). Techniques for maintaining connectivity in wireless ad-hoc networks under energy constraints. *ACM Trans. on Embedded Computing Systems*, **6.3**, 16–37.

[13] Koushanfar, F., Taft, N. and Potkonjak, M. (2006). Sleeping coordination for comprehensive sensing using isotonic regression and domatic partitions. *INFOCOM*, 1-13.

[14] Lehmann, E. (2006). *Nonparametrics: Statistical Methods Based on Ranks*. Springer-Verlag, Germany.

[15] Megerian, S., Koushanfar, F., Qu, G., Veltri, G. and Potkonjak, M. (2002). Exposure in wireless sensor networks: Theory and practical solutions. *ACM Journal of Wireless Networks*, **8.5**, 443454.

[16] Megerian, S., Koushanfar, F., Potkonjak, M. and Srivastava, M. (2005). Worst- and best-case coverage in sensor networks. *IEEE Transactions on Mobile Computing*, **4.1**, 84-92.

[17] Lanzisera, S., Lin, D. and Pister, K. (2006). Rf time of flight ranging for wireless sensor network localization. *WISES*, 1-12.

[18] Montgomery, D. (2001). *Design and Analysis of Experiments*, 5th ed. Wiley, New York.

[19] Patwari, N., Ash, J., Kyperountas, S., Hero, A. III, Moses, R. and Correal, R. (2005). Locating the nodes: cooperative localization in wireless sensor networks. *IEEE Signal Processing Magazine*, **22.4**, 54-69.

[20] Priyantha, N., Chakraborty, A. and Balakrishnan, H. (2000). The cricket location-support system. *MOBICOM*, 32–43.

[21] *OSL: Open Systems Laboratory at UIUC*. http://www-osl.cs.uiuc.edu/research?action=topic\&topic=Sensor+Networks.

[22] Perkins, D., Hughes, H. and Owen, C. (2002). Factors affecting the performance of ad hoc networks. *ICC*, 2048-2052.

[23] Plackett, R.L. and Burman, J.P. (1946). The design of optimum multifactorial experiments. *Biometrika*, **33.4**, 305-325.

[24] Saltelli, A., Tarantola, S., Campolongo, F. and Ratto, M. (2004). *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. Wiley, England.

[25] Saxe, J. (1979) Embeddability of weighted graphs in k-space is strongly np-hard. *Allerton Conf. in Communications, Control, and Computing*, 480-489.

[26] Slijepcevic, S., Megerian, S. and Potkonjak, M. (2002). Location errors in wireless embedded sensor networks: sources, models, and effects on applications. *Mobile Computing and Communications Review*, **6.3**, 67-78.

[27] Totaro, M. and Perkins, D. (2005). Using statistical design of experiments for analyzing mobile ad hoc networks. *MSWiM*, 159-168.

[28] Vadde, K. and Syrotiuk, V. (2004). Factor interaction on service delivery in mobile ad hoc networks. *IEEE Journal on Selected Areas in Communications (JSAC)*, **22.7**, 13351346.

IMS Collections
Vol. 0 (2009) 347–351
© Institute of Mathematical Statistics, 2009
arXiv: math.PR/00000019

A Note on the Investment Proportions of a Minimum-Variance Equity Portfolio

Wilhelmine von Türk[1,*]

American Century Investments

Abstract: It is shown that the capitalization-weighted portfolio is mathematically required to coincide with the minimum-variance portfolio, provided both portfolios are defined with respect to the same (arbitrary) collection of equities having linearly independent returns. This result is a logical consequence of the law of iterated expectations and has important implications for equity return covariance structure.

Contents

1 Introduction 347
2 Notation and Preliminary Results 348
3 The Population View 349
4 The Law of Iterated Expectations 350
5 Proof of the Main Result 350
6 Conclusion 351
Technical Appendix A 351
References 351

1. Introduction

Samuelson (1965) was the first to recognize the relevance of the Law of Iterated Expectations (LIE) in the area of finance. His celebrated paper shows how this law can be applied in the context of the price of a single risky asset. Essentially, the LIE reveals that forecasting error is not predictable. An illuminating discussion of this law can be found, e.g., in Campbell, Lo and MacKinlay (1997).

This paper extends Samuelson's insight by developing an application of the LIE in the context of an arbitrary collection of equity returns. Specifically, it will be shown that the LIE implies that the capitalization-weighted portfolio is mathematically required to coincide with the minimum-variance portfolio, provided both portfolios are defined with respect to the same (arbitrary) collection of equities having linearly independent returns. This result has important implications for equity return covariance structure, as summarized in technical appendix A.

[1]43 Del Mar Avenue, Berkeley, CA 94708, email: wvonturk@sbcglobal.net

*The author gratefully acknowledges her intellectual indebtedness to the tradition of the Department of Statistics, University of California at Berkeley, where she completed the dissertation with the guidance and advice of Professor Erich L. Lehmann. The responsibility for any shortcomings in the present paper rests squarely on the shoulders of the author.

AMS 2000 subject classifications: Primary 91B28; secondary 62H25

Keywords and phrases: minimum-variance portfolio, capitalization-weighted portfolio, equity return covariance structure, Law of Iterated Expectations, equity portfolio optimization.

The paper is organized as follows. Section 2 introduces the notation and presents a sufficient condition for proving the main result. The population view of the surprises (forecasting errors) is developed in Section 3. This population view yields an important building block for the proof of the main result. Section 4 applies the LIE in the context of a cross-section of equities, and exhibits the implication of this law from which the main result logically flows. The proof of the main result is summarized in Section 5. Section 6 concludes.

Technical appendix A summarizes the implications of this paper's main result for equity return covariance structure.

2. Notation and Preliminary Results

Consider an arbitrary collection of n equities with linearly independent returns. Denote their returns by $\mathbf{R} = (R_1 \dots R_n)'$ and denote the covariance matrix of $\mathbf{R}$ by Σ. Since the returns are assumed to be linearly independent, Σ is nonsingular, so that its inverse Σ^{-1} is well defined. Let $\mathbf{1}$ denote the n x 1 vector all of whose elements are equal to 1.

Let $\mathbf{M} = (M_1 \dots M_n)'$ denote the market capitalizations of the equities. Let $w_i = M_i/\mathbf{M}'\mathbf{1}$ and let $\mathbf{w} = (w_1 \dots w_n)'$. The vector $\mathbf{w}$ corresponds to the *capitalization weights*. Note that all of the elements of $\mathbf{w}$ are positive. Let $\mathbf{p} = (p_1 \dots p_n)'$ denote an n x 1 vector of constants such that $\mathbf{p}'\mathbf{1} = 1$. Then the elements of $\mathbf{p}$ correspond to the investment proportions of a portfolio fully invested in the n equities, and $\mathbf{p}'\mathbf{R}$ denotes its return. If $\mathbf{p} = \mathbf{w}$, the investment proportions are those of the capitalization-weighted portfolio. The expression $\mathbf{w}'\mathbf{R}$ denotes the return of the capitalization-weighted portfolio.

Among all possible fully invested portfolios that can be formed from this collection of equities, there is one portfolio whose return has minimum variance. This portfolio is the minimum-variance portfolio. It is well known that, if Σ is nonsingular, the investment proportions of the minimum-variance portfolio are given by $\Sigma^{-1}\mathbf{1}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$ (see, e.g., Roll (1977), Campbell, Lo and MacKinlay (1997) or Grinold and Kahn (2000)).

The desired result is that

$$\mathbf{w} = \Sigma^{-1}\mathbf{1}/\mathbf{1}'\Sigma^{-1}\mathbf{1}.$$

For the purpose of finding a proof of this result, it is convenient to define

$$\beta_i = \mathrm{Cov}(R_i, \mathbf{w}'\mathbf{R})/\mathrm{Var}(\mathbf{w}'\mathbf{R}).$$

Let $\boldsymbol{\beta} = (\beta_1 \dots \beta_n)'$. Then

$$\boldsymbol{\beta} = \Sigma\mathbf{w}/\mathbf{w}'\Sigma\mathbf{w}.$$

Note that $\mathbf{w}'\mathbf{1} = 1$, so that

$$\mathbf{w}'\Sigma\mathbf{w} = 1/\boldsymbol{\beta}'\Sigma^{-1}\mathbf{1}$$

and

$$\mathbf{w} = \Sigma^{-1}\boldsymbol{\beta}/\boldsymbol{\beta}'\Sigma^{-1}\mathbf{1}.$$

With this notation it is possible to establish the following preliminary result.

Lemma 2.1. *The following assertions are equivalent*:
(i) $\mathbf{w} = \Sigma^{-1}\mathbf{1}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$
(ii) $\boldsymbol{\beta} = \mathbf{1}$

Proof. $\mathbf{w} = \Sigma^{-1}\boldsymbol{\beta}/\boldsymbol{\beta}'\Sigma^{-1}\mathbf{1}$ implies that $\Sigma\mathbf{w} = \boldsymbol{\beta}/\boldsymbol{\beta}'\Sigma^{-1}\mathbf{1}$, and $\mathbf{w} = \Sigma^{-1}\mathbf{1}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$ implies that $\Sigma\mathbf{w} = \mathbf{1}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$. Therefore, $\mathbf{w} = \Sigma^{-1}\mathbf{1}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$ if and only if $\boldsymbol{\beta} = c\mathbf{1}$ for $c = \boldsymbol{\beta}'\Sigma^{-1}\mathbf{1}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$. Since $\mathbf{w}'\boldsymbol{\beta} = \mathbf{w}'\mathbf{1} = 1$, this constant c is equal to 1.

Next, consider an investment period which begins at time $t = 0$ and ends at time $t = T$. Note that at time $t = 0$, $\mathbf{w}$ is a vector of known constants while $\mathbf{R}$ is a vector of random variables. Let $\mathrm{E}[\mathbf{R}]$ denote the n x 1 vector whose elements correspond to $\mathrm{E}[R_i]$, where $\mathrm{E}[R_i]$ is the expected value of R_i at time $t = 0$ $(i = 1, \ldots, n)$. The expression $\mathrm{E}[R_i]$ corresponds to the best forecast at time $t = 0$ of the realized value of R_i observed at time $t = T$.

Now shift attention from the returns R_i to the surprises S_i, where $S_i = R_i - E[R_i]$. The surprises correspond to *forecasting errors*. Let $\mathbf{S} = (S_1, \ldots, S_n)'$, let $\mathcal{A}$ denote the σ-field induced by $\mathbf{S}$, and let $\mathcal{F}$ denote the σ-field induced by $\mathbf{w}'\mathbf{S}$. Since $\mathbf{w}'\mathbf{S}$ is a function of $\mathbf{S}$, $\mathcal{F} \subset \mathcal{A}$. The conditional expectation of S_i given $\mathbf{w}'\mathbf{S}$ can then be written as $\mathrm{E}^{\mathcal{F}}[S_i]$, which is defined almost surely (a.s.) in the sense that any two versions agree, except possibly on a null set in $\mathcal{A}$. With this notation, it is possible to simplify the task of proving the desired result as follows.

Lemma 2.2. *For the purpose of showing that* $\mathbf{w} = \Sigma^{-1}\mathbf{1}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$, *it is sufficient to show that* $\mathrm{E}^{\mathcal{F}}[S_i] = \mathbf{w}'\mathbf{S}$ $(a.s.)$ $(i = 1, \ldots, n)$.

Proof. In light of Lemma 2.1, it is sufficient to show that $\mathrm{E}^{\mathcal{F}}[S_i] = \mathbf{w}'\mathbf{S}$ $(a.s.)$ $(i = 1, \ldots, n)$ implies that $\boldsymbol{\beta} = \mathbf{1}$.

By definition,

$$\mathrm{Cov}(R_i, \mathbf{w}'\mathbf{R}) = \mathrm{E}[S_i\mathbf{w}'\mathbf{S}]$$

and

$$\mathrm{Var}(\mathbf{w}'\mathbf{R}) = \mathrm{E}\left[(\mathbf{w}'\mathbf{S})^2\right].$$

Therefore it suffices to show that $\mathrm{E}^{\mathcal{F}}[S_i] = \mathbf{w}'\mathbf{S}$ $(a.s.)$ $(i = 1, \ldots, n)$ implies that

$$\mathrm{E}[S_i\mathbf{w}'\mathbf{S}] = \mathrm{E}\left[(\mathbf{w}'\mathbf{S})^2\right] \quad (i = 1, \ldots, n).$$

This is easily accomplished by recalling the usual properties of conditional expectation operators, which yields

$$\mathrm{E}[S_i\mathbf{w}'\mathbf{S}] = \mathrm{E}\left[\mathbf{w}'\mathbf{S}\mathrm{E}^{\mathcal{F}}[S_i]\right],$$

so that $\mathrm{E}^{\mathcal{F}}[S_i] = \mathbf{w}'\mathbf{S}$ (a.s.) implies $\mathrm{E}[S_i\mathbf{w}'\mathbf{S}] = \mathrm{E}\left[(\mathbf{w}'\mathbf{S})^2\right]$ $(i = 1, \ldots, n)$.

Before it can be shown that the assertion $\mathrm{E}^{\mathcal{F}}[S_i] = \mathbf{w}'\mathbf{S}$ (a.s.) $(i = 1, \ldots, n)$ flows logically from the LIE, it is first necessary to develop the population view of the surprises (forecasting errors). This is done in the next section.

3. The Population View

Consider an arbitrary collection of n publicly traded companies and an investment period which begins at time $t = 0$ and ends at time $t = T$. The development of the population view of the surprises depends on the operation of repricing the shares of each company at the beginning of the investment period, while adjusting the shares outstanding so as to leave the market capitalization of each company unchanged. This operation, very familiar to equity investors, guarantees that there is no loss of generality in repricing the shares of each company at the beginning of the investment period to have a value of one dollar.

As in Section 2, let M_i denote the market capitalizations at the beginning of the investment period ($i = 1, \ldots, n$). If the price of one share is one dollar (and if the market capitalizations are rounded to the nearest dollar), then M_i corresponds to the number N_i of shares outstanding for the i-th company. Each share of the i-th company has return R_i and associated surprise (forecasting error) S_i. Therefore, at the end of the investment period, the observed values s_i of the random variables S_i form a population in which s_i occurs with frequency N_i. The capitalization weights w_i then correspond to the relative frequencies with which the s_i are observed. This population has mean $\mathbf{w}'\mathbf{s}$, where $\mathbf{s} = (s_1, \ldots, s_n)'$ is the observed value of the random vector $\mathbf{S} = (S_1, \ldots, S_n)'$.

It is well known that the population mean is the expected value of a randomly selected element from that population. From the standpoint of time $t = T$, the population mean is observed with certainty, while from the standpoint of time $t = 0$, the population mean corresponds to the random variable $\mathbf{w}'\mathbf{S}$. This means that

$$\mathrm{E}_T[\text{a randomly selected surprise}] = \mathbf{w}'\mathbf{s},$$

where "E_T" is to be read as "the expected value at time $t = T$, conditional on the observed value $\mathbf{w}'\mathbf{s}$ of $\mathbf{w}'\mathbf{S}$." Translating this into σ-field notation yields the following lemma.

Lemma 3.1. $\mathrm{E}^{\mathcal{F}}[\text{a randomly selected surprise}] = \mathbf{w}'\mathbf{S}$ *(a.s.), where $\mathcal{F}$ denotes the σ-field induced by* $\mathbf{w}'\mathbf{S}$.

As will be seen in Section 5, Lemma 3.1 is an important building block for the proof of the main result.

4. The Law of Iterated Expectations

A key insight provided by the LIE is that forecasting error is not predictable. The present paper contemplates an arbitrary collection of n equities with returns R_i and associated forecasting errors $S_i = R_i - \mathrm{E}[R_i]$ ($i = 1, \ldots, n$). In this context, the LIE implies that the expected value of any surprise S_i corresponds to the expected value of a *randomly selected surprise*. From the standpoint of time $t = 0$, this yields the set of equations

$$\mathrm{E}[S_i] = \mathrm{E}[\text{a randomly selected surprise}] \ (i = 1, \ldots, n).$$

Conditional on $\mathbf{w}'\mathbf{S}$, the LIE similarly implies the following lemma.

Lemma 4.1. $\mathrm{E}^{\mathcal{F}}[S_i] = \mathrm{E}^{\mathcal{F}}[\text{a randomly selected surprise}]$ *(a.s.)* ($i = 1, \ldots, n$)*, where $\mathcal{F}$ denotes the σ-field induced by* $\mathbf{w}'\mathbf{S}$.

Lemma 4.1 exhibits the implication of the LIE from which the main result of the present paper logically flows, as summarized in the next section.

5. Proof of the Main Result

The main result is stated as the proposition below.

PROPOSITION: *For an arbitrary collection of n equities having linearly independent returns,* $\mathbf{w} = \Sigma^{-1}\mathbf{1}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$.

Proof. From Lemma 4.1,

$$\mathrm{E}^{\mathcal{F}}[S_i] = \mathrm{E}^{\mathcal{F}}[\text{a randomly selected surprise}] \text{ (a.s.) } (i = 1, \ldots, n).$$

From Lemma 3.1,

$$\mathrm{E}^{\mathcal{F}}[\text{a randomly selected surprise}] = \mathbf{w}'\mathbf{S} \text{ (a.s.)}.$$

Combining Lemmas 4.1 and 3.1 yields

$$\mathrm{E}^{\mathcal{F}}[S_i] = \mathbf{w}'\mathbf{S} \text{ (a.s.) } (i = 1, \ldots, n).$$

In light of Lemma 2.2, this completes the proof.

6. Conclusion

It has been shown that the capitalization-weighted portfolio is mathematically required to coincide with the minimum-variance portfolio, provided both portfolios are defined with respect to the same (arbitrary) collection of equities having linearly independent returns. This result is a logical consequence of the LIE, and has important implications for equity return covariance structure, as summarized in technical appendix A.

Technical Appendix A

The main result of this paper has important implications for equity return covariance structure, as summarized in the following proposition.

PROPOSITION: *For any collection of n equities with linearly independent returns, the covariance matrix* Σ *of* $\boldsymbol{R} = (R_1, \ldots, R_n)'$ *is of the form*

$$\Sigma = \mathbf{1}\mathbf{1}'k + \mathbf{U}^2$$

where $\mathbf{U^2}$ *is a diagonal matrix such that the i-th diagonal element is positive and inversely proportional to* M_i *(*$i = 1, \ldots, n$*), and where k corresponds to a constant which can be positive, negative or zero.*

Note that for positive values of k, general equity return covariance structure exhibited in the proposition above corresponds to the covariance matrix of a 1-factor model in which

(i) the factor loadings of the unique common factor are all equal to one; and
(ii) the variances of the specific factors are inversely proportional to the market capitalizations of the equities.

References

[1] Campbell, J. Y., Lo, A. W. and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets.* Princeton University Press, Princeton, New Jersey.
[2] Grinold, R. C. and Kahn, R. N. (2000). *Active Portfolio Management.* McGraw-Hill, New York.
[3] Roll, R. (1977). A critique of the asset pricing theory's tests. *Journal of Financial Economics* **4**, 129–176.
[4] Samuelson, P. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* **6**, 41–49.