Interface '97

29th Symposium on the Interface:
Computing Science and Statistics

Mining and Modeling Massive Data Sets in
Science, Engineering, and Business

Subtheme in Environmental Statistics

May 14-17, 1997
Houston, Texas
Holiday Inn – Houston Medical Center

Pictures/lgLogo.PS

PROGRAM

Sponsored by:
THE INTERFACE FOUNDATION OF NORTH AMERICA

Hosted by:

WILLIAM MARSH RICE UNIVERSITY
David W. Scott, Program Chair

Co-hosted by:
UNIVERSITY OF TEXAS M. D. ANDERSON CANCER CENTER
UNIV. OF TEXAS HEALTH SCIENCES CENTER AT HOUSTON, SCHOOL OF PUBLIC HEALTH
W. M. KECK CENTER FOR COMPUTATIONAL BIOLOGY

Cooperating Institutions:
ASA, IMS, IASC, SIAM, INFORMS, ENAR/WNAR

Program Schedule

Pictures/lgLogo.PS

Mining and Modeling Massive Data Sets in Science, Engineering, and Business

Subtheme in Environmental Statistics

Abstracts

Pictures/lgLogo.PS

Mining and Modeling Massive Data Sets in Science, Engineering, and Business

Subtheme in Environmental Statistics

/home/interf97/SESSIONS/Pictures/meeting.ps

Interface 1997

Program Chair

David W. Scott
Department of Statistics, MS-138
Rice University
6100 Main Street
Houston, TX 77005-1892
(713) 527-6037
scottdw@rice.edu

Business Office

Diane J. Brown
Department of Statistics, MS-138
Rice University
6100 Main Street
Houston, TX 77005-1892
(713) 527-6032
Fax: (713) 285-5476

Conference Arrangements

Diane Brown, David Scott, John Salch, David Andrews, Keith Baggerly, Peter Thall, Luciano Batista, Michael Pearlman, J. Lynn Palmer, the Rice University Statistics Department

Grant Support

Grant support provided by:
The National Science Foundation
The U.S. Environmental Protection Agency
The U.S. National Security Agency

Other Support

Other support provided by:
UTMDACC (Barry Brown)
UTSPH (Ralph Frankowski)
Keck/NSF (George Phillips)
SAS, Inc. (Thursday Lunch)
Visual Numerics Inc. (Friday Lunch)
SGI, Inc. (Computers)
Southwestern Bell (ISDN Line)

Conference Organizers

The Program committee consists of:
David W. Scott (Rice University)
Keith A. Baggerly (Rice University)
Clark R. Abrahams (NationsBank)
Dennis D. Cox (Rice University)
Jan DeLeeuw (Univ. of California, Los Angeles)
Tony Elam (CITI, Rice University)
Katherine B. Ensor (Rice University)
Arnie Goodman (University of California, Irvine)
Peter Guttorp (University of Washington)
William J. Kennedy (Iowa State University)
Albert M. Liebetrau (Battelle Labs Northwest)
James R. Maar (Department of Defense)
Sally C. Morton (Rand Corporation)
Michael Schimek (KFU Graz, Austria)
Willy Zwaenepoel (Rice University)

Proceedings Editor

David W. Scott, at the business office address

Correspondence to IFNA

IFNA, Pat Joyce
Center for Computational Statistics
157 Sci-Tech II George Mason University
Fairfax, VA 22030

Contacts

e-mail: interface97@stat.rice.edu
phone: (713) 527-6037
fax: (713) 285-5476
mail: Interface '97
Department of Statistics, MS-138
Rice University
6100 Main Street
Houston, TX 77005-1892

General Information

Conference Facilities

The Holiday Inn – Houston Medical Center Hotel conference facilities have been recently remodeled and are located on the first and second floors. The registration desk is located off the lobby on the second floor.

Exhibition Hall

The exhibition hall is located in the San Antonio room on the second floor and will be open Wednesday afternoon until Saturday morning. Exhibitors this year include:

- SAS Software for Statistical Analysis
- SAS JMP Stat Discovery Software
- Springer-Verlag
- Visual Numerics, Inc.
- Salford Systems
- Captial One Financial Corporation
- Interface Foundation of North America
- SIAM/ASA

Please visit and support this year's exhibitors.

Sessions

All technical sessions will be held on the second floor plus the Austin and Galveston rooms on the first floor. A map of the meeting space may be found on the inside cover of this program guide. The Wednesday evening mixer, the Keynote Address, and the Thursday Plenary Luncheon Address will be in the combined Houston I, II, and III rooms. These rooms will be subdivided Thursday afternoon for the remainder of the conference. The Thursday evening Banquet will also be held in the combined Houston I, II, and III rooms. A string quartet from the Rice Shepherd School of Music will perform during pre-dinner and dinner periods.

Snacks

Breaks will be held in the area adjacent to the registration desk outside Houston II. There will be a mid-morning and mid-afternoon break for thirty minutes each. Please check the program guide for exact times each day.

Lunch

This year, the Organizing Committee is pleased to announce that box lunches will be furnished Thursday and Friday, along with special talks in the Houston meeting rooms. After the last morning session, you will have 15 minutes to get your box lunch and drink and return to the Houston meeting room.

The Thursday lunch is provided courtesy of **SAS, Inc.**, and our speaker is Bruce Nelson from Cisco Systems.

The Friday lunch is provided courtesy of **Visual Numerics, Inc.**, and you may select from among three speakers: Richard J. Hanson (VNI), Jon Kettenring (Bellcore), or George Phillips (Rice).

Banquet

Thursday evening's banquet will be held in the Houston meeting rooms. An outstanding menu of Tenderloin and Jumbo Shrimp and wine has been selected for your enjoyment. If you wish to have a vegetarian meal, please indicate your preference at the registration desk as early as possible. After dinner, there will be a thirty minute banquet address by James R. Thompson, Professor of Statistics at Rice University. Afterwards, you are invited to remain and continue your conversation. A cash bar will be available.

Dining

Other meals will be on your own. Houston has an extensive selection of high quality and ethnic dining experiences within two miles of the hotel. The hotel shuttle bus can take you to the Rice Village or the Galleria. An annotated list (courtesy of Peter Thall) is available at the registration counter.

Internet Access

The Dallas III meeting room will be equipped with a number of computer workstations (courtesy of **SGI, Inc**) connected to an ISDN line (courtesy of **Southwestern Bell Company**). This arrangement should allow adequate access to electronic mail via the internet. The 128 Kb capacity means that users should not expect to run anything more than telnet without affecting other users. The room will be locked after hours. Check with registration for details.

Sponsor

The Interface Foundation of North America is a nonprofit educational corporation founded in 1987 to sponsor the symposium and publish the proceedings. The IFNA is also a co-publisher of the *Journal of Computational and Graphical Statistics.*

Proceedings

Invited and contributed papers should be submitted to David Scott by June 16, 1997. See **Manuscript Preparation** guidelines below.

Emergency Access

If it is necessary for someone to reach you while you are attending sessions, ask at the Registration Area for our cellular phone number. We request that you limit the use of this number to emergencies, as the phone charges can be expensive. Non-emergency messages can be left at the Hotel (713-797-1110). If it is necessary for you to receive a fax, the hotel maintains a guest fax number (713-797-1034).

Keynote: Jerry Friedman

The keynote address will be given by Jerry Friedman, Professor of Statistics at Stanford University (since 1982) and Leader of the Computational Research Group at the Stanford Linear Accelerator Center (since 1972). Professor Friedman received his Ph.D. in High Energy Physics in 1967 from U.C. Berkeley. His research interests include data mining, machine learning, prediction algorithms, classification, modern regression, chemometrics, and computationally intensive methods for nonparametric data exploration. He is a Fellow of the American Statistical Association.

Some features of this program

To make distinguishing between the two types of sessions convenient, we have marked invited sessions with filled circles (●●) and contributed sessions with open circles (○○). Also note that in the Program Schedule, only the first author's affiliation is given. The complete affiliation information is given in the Abstracts section.

Manuscript Preparation Instructions for Interface '97 Authors

The proceedings volume is produced from camera-ready copy that the authors prepare according to the guidelines given below.

1. Sample LaTeX and TeX files are available at the web site:
`http://www.stat.rice.edu/interface97.html`.

2. Contributed papers should not exceed five pages and invited papers should not exceed ten pages. These pages include everything - bibliography, tables, graphics, title, etc.

3. Originals must be made by laser printer (not copies) on standard paper (8-1/2 × 11 inches). Refer to item 6 below for a description of the typing area.

4. Number the pages in LIGHT BLUE on the back of each page. Do not number the pages on the front.

5. Center your title on the middle six inches of the first page, and list the authors and addresses under the title. If you cite support, do so in a footnote. Keywords and phrases are not required.

6. There will be two columns on each page. The side margins are 3/4 inches to the left of the first column, $1/4$ inches between columns, and 3/4 inches to the right of the second column. The typing area for each column is 3 and 3/8 inches wide. Text should begin 1 and 1/4 inches from the top of the page, and leave one inch at the bottom of the page. If possible, use right justification for each column.

7. To ensure uniformity of appearance, use 10 point type, preferably computer modern or Times Roman.

8. The manuscript should be single spaced. If you are familiar with TeX, this means a baseline skip of 12 points.

9. The abstract should begin the first column.

10. The publisher will number the pages and insert a running head outside the typing area. On a separate sheet of paper, please indicate your choice of running head.

11. Other formatting is left to the discretion of the author(s).

12. Manuscripts prepared according to these guidelines are due by June 16, 1997. Send the manuscript to:

    Interface '97
    Department of Statistics - MS-138
    Duncan Hall, Room 1083
    Rice University
    6100 Main Street
    Houston, TX 77005-1892
    phone: (713) 527-6037
    fax: (713) 285-5476
    email: `interface97@stat.rice.edu`

The following LaTeX header has been suggested recently:

```
\documentstyle[11pt,twocolumn]{article}
\setlength{\textwidth}{17.5cm}
\setlength{\textheight}{25cm}
\setlength{\topmargin}{-1cm}
\setlength{\headheight}{0cm}
\setlength{\headsep}{0cm}
%\setlength{\topskip}{???}
%\setlength{\parskip}{???}
\setlength{\oddsidemargin}{-.75cm}
```

Wednesday

Wednesday, 8:00 a.m. - 9:00 a.m.

Registration for Short Courses Gallery

Wednesday, 9:00 a.m. - 4:30 p.m.

Short Course I: Tools for Discovering Patterns in Data Galveston
John Elder, Chief Scientist, Quantitative Solutions.

Short Course II: Simulation: A Modeler's Approach El Paso
James R. Thompson, Department of Statistics, Rice University.

Wednesday, 3:00 p.m. - 8:00 p.m.

Registration for Interface '97 Gallery

Exhibition Hall Open San Antonio

Wednesday, 7:45 p.m. - 10:00 p.m.

Evening Mixer Houston I-II-III
Light refreshments and cash bar.

Thursday

Thursday, 7:00 a.m. - 8:15 a.m.

ASA Statistical Graphics Executive Committee Breakfast (closed) Boardroom
Sally C. Morton, Chair.

Thursday, 7:15 a.m. -

Registration for Interface '97 Gallery

Thursday, 7:45 a.m. -

Exhibition Hall Open San Antonio

Thursday, 8:15 a.m. - 9:45 a.m.

Opening and Welcome Houston I-II-III
President Malcolm Gillis, Rice University
David W. Scott, Rice University

•• **Keynote Address** Houston I-II-III
Data Mining and Statistics: What's the Connection?, Jerome H. Friedman, Stanford University

Thursday, 10:15 a.m. - 12:00 p.m.

•• **Uncertainty Analysis for Computer Modeling Systems** Galveston
Organized by Al Liebetrau, Battelle Labs Northwest.
Chaired by Art Owen, Stanford University

10:15 *Uncertainty Analysis for Computer Modeling Systems*, Albert M. Liebetrau, Guang Chen, and Feng Gao, Pacific Northwest National Laboratory, Richland, Washington

10:45 *Structural Model Uncertainty in Stochastic Simulation Models*, Michael D. Mckay, Los Alamos National Laboratory, Los Alamos, NM, USA

11:15 *Estimating the Response, Derivatives, and Transmitted Variance using Computer Experiments*, James R. Koehler, University of Colorado at Denver

11:45 *Floor Discussion*

•• **New Directions for Statistical Packages and Environments** Houston I-II-III
Chaired by Doug Martin, U. Washington and MathSoft

10:15 *The Theme of 'Context' in Statistics Software*, John Sall, SAS Institute, Cary NC

10:45 *Tables and Graphs*, Leland Wilkinson, SPSS Inc. and Northwestern University

11:15 *Matrix – The Mother of all Data Structures*, Cleve Moler, The MathWorks, Inc.

11:45 *Discussion*, R. Douglas Martin, University of Washington and MathSoft, Inc.

11:55 *Floor Discussion*

Computational Environmental Geostatistics Austin ••
Organized by Dennis Cox and Katherine Ensor, Rice University. Chaired by Dennis Cox.

10:15 *Spatial vs. Ensemble Statistics of Spatial Simulation Algorithms*, Donald E. Myers, Department of Mathematics, University of Arizona

10:45 *Geostatistical Methods for Estimating Total Mass of Mercury in the Waters of the Florida Everglades*, Stephen L. Rathbun, Department of Statistics, University of Georgia

11:15 *Statistical Modeling of Vector-Valued Spatial Data Using Gradient Processes*, David A. Andrews and Dennis D. Cox, Rice University, Houston, Texas

11:45 *Floor Discussion*

Numerical Methods and Algorithms Dallas II ◦◦
Chaired by Shane Pankratz, Rice University

10:15 *Computing Cook's Distance for Nonlinear Regression in S-Plus*, C. Shane Reese, Texas A&M University, College Station, TX

10:35 *An Extension of the Mixed Primal-Dual Bases Algorithm to the Case of More Constraints than Dimensions, with Applications to Nonparametric Shape Constrained Problems in Regression, Maximum Likelihood Density Estimation, and the Bioassay Problem*, Mary C. Meyer, University of Michigan

10:55 *Derivatives of Fortran Programs, Automatically*, Alan Carle and Mike Fagan(*), Rice University

11:15 *An Efficient Method for Representing, Analyzing and Visualizing Massive, High Dimensional Data*, Tamraparni Dasu and Theodore Johnson, AT&T Labs - Research

11:35 *Efficient Estimation and Model Selection in Graphical Models*, Dag Wedelin, Chalmers University of Tech, Gothenburg, Sweden

11:55 *Floor Discussion*

Thursday, 12:00 p.m. - 1:30 p.m.

Executive Meeting of the Statistical Computing Section Lunch (closed) Boardroom
Daryl Pregibon, Chair

Thursday, 12:15 p.m. - 1:00 p.m.

Box lunch furnished courtesy of SAS, Inc.

Plenary Address Houston I-II-III
Chaired by Farrell Gerbode, Rice University

An Insider's View of Networking Trends, Bruce Nelson, Chief Science Officer, Cisco Systems, San Jose, CA

Thursday, 1:30 p.m. - 3:15 p.m.

•• **Numerical Methods for Large Problems** Houston I
Chaired by Karen Kafadar, UC Denver

- **1:30** *Interior-Point Methods: New Algorithmic Methodology for Large-Scale Inequality Constrained Optimization*, Richard A. Tapia, Rice University
- **2:00** *Software and Algorithms for Estimation Problems in Reflection Seismology*, William W. Symes, Rice University
- **2:30** *Latin Supercube Sampling for Very High Dimensional Simulations*, Art B. Owen, Statistics Department, Stanford University
- **3:00** *Floor Discussion*

•• **Educational Issues and Examples** Houston II
Chaired by Tony Rossini, University of South Carolina

- **1:30** *Graphs in Statistical Data Analysis: Is the Medium the Message?*, R. Dennis Cook and Sanford Weisberg, University of Minnesota
- **2:00** *Hubs and Signposts: Incremental Implementation of Statistical Strategy.*, R.W. Oldford, University of Waterloo
- **2:30** *Technology in the Teaching and Learning of Statistics*, Adrian Bowman, Statistics Dept., University of Glasgow, UK
- **3:00** *Floor Discussion*

•• **Business Data Mining: Methodological Viewpoints** Houston III
Organized and chaired by Clark R. Abrahams, NationsBank Consumer Credit Policy

- **1:30** *Introduction*, Clark R. Abrahams, NationsBank Consumer Credit Policy
- **1:35** *A Systems Approach to Knowledge Discovery*, Kenneth L. Reed, HyperParallel, Inc., Mission Viejo, California
- **2:05** *The Importance of Preprocessing and Exploratory Data Analysis in Data Mining*, Troy Haines, American Management Systems, Inc., Redwood City, California
- **2:35** *Operational Data Considerations in Data Mining*, Jim Sattler, Saber Consulting Services, San Francisco, CA
- **3:05** *Floor Discussion*

◦◦ **Spatial Statistical Problems** Austin
Chaired by David Andrews, Rice University

- **1:30** *Estimating Magnitude and Frequency of Flood Peaks at Ungaged Sites Using Regression*, Gary D. Tasker, U.S. Geological Survey, Reston, VA
- **1:50** *Applications of Spatial Statistics to the Quantitative Analysis of Genetic Instability in Carcinogenesis*, J.Jack Lee, Ren Yu, Waun Ki Hong, and Walter N. Hittelman, Department of Biomathematics, University of Texas M.D. Anderson Cancer Center
- **2:10** *Classification of Mixtures of Spatial Point Processes*, Nandini Raghavan, Prem Goel, and Somnath Ghosh, Statistics, The Ohio State University
- **2:30** *The Use of Threshold Statistical Methods for Detecting Long-Term Trends in Urban Ozone Levels*, Humberto Vaquera, Jose A. Villasenor, and Janet Hughes, Colegio de Posgraduados, Mexico
- **2:50** *Persistence and Recurrence in Atmospheric Circulation*, Andrew M. Fraser, Richard Smith, Kvein Vixie, and Padhraic Smyth, Systems Science, Portland State University
- **3:10** *Floor Discussion*

Wavelets Galveston ◦◦
Chaired by Sam Conley, Rice University

- **1:30** *Some Findings on the Covariance Structure of Wavelet Coefficients: Theory and Models in a Bayesian Perspective*, Marina Vannucci and Fabio Corradi, University of Kent at Canterbury, UK
- **1:50** *Wavelet Estimation Based on Quasi-Linear Approximation*, Sam Efromovich, The University of New Mexico, Albuquerque
- **2:10** *Detection of Additive Outliers in Time Series Data Using Wavelets*, S. Huzurbazar and Canan Bilen, University of Wyoming
- **2:30** *Testing for Abrupt Jumps with Wavelets*, R. Todd Ogden and Cheng Cheng, University of South Carolina
- **2:50** *Floor Discussion*

Thursday, 3:45 p.m. - 5:30 p.m.

Visualization: Applications and Environments Houston I ••
Organized and chaired by Sally C. Morton, RAND Corp.

- **3:45** *Analyzing Complex Data Sets: Visualization and Modeling*, Linda Clark, William S. Cleveland, Lorraine Denby(*), and Chuanhai Liu, Bell Labs, Lucent Technologies, Murray Hill, NJ
- **4:15** *Visualizing Multi-Dimensional Spaces of Simulation Models*, Steve Bankes, RAND
- **4:45** *A Java-Based Graphics Production Class Library*, Daniel J. Rope, Bureau of Labor Statistics
- **5:15** *Floor Discussion*

Software for Data Mining and Data Warehousing Houston II ••
Organized by Willy Zwaenepoel and Tony Elam, Rice University. Chaired by Tony Elam.

3:45 *TreadMarks: Software Support for Distributed Data Mining on Workstation Clusters*, Alan Cox, Department of Computer Science, Rice University

4:15 *Data Mining for Internet Applications*, David C. Martin, IBM Almaden Research Center

4:45 *Knowledge-Based Pre- and Post-Processors*, Arun Jain, Assoc. VP Information Technology, Rice Univ.

5:15 *Floor Discussion*

•• **Massive Data Sets in an Industrial Setting and Finance** Houston III

Chaired by James R. Thompson, Rice University

3:45 *Robust Statistics and Data Mining in Finance and Financial Engineering*, R. Douglas Martin, Professor of Statistics, University of Washington and Chief Scientist, Data Analysis Products Division of MathSoft, Inc.

4:15 *Applications of Positive-Breakdown Methods*, Peter J. Rousseeuw, University of Antwerp, Belgium

4:45 *War Stories from the Data Trenches*, Richard A. Becker and Allan R. Wilks, AT&T Labs - Research, Florham Park, New Jersey

5:15 *Floor Discussion*

◦◦ **Multivariate Methods** Austin

Chaired by John Salch, Rice University

3:45 *Partitioning and Subsampling to Uncover Subtle Structure in Massive Data Sets*, David M. Rocke, University of California, Davis

4:05 *Applications of Multidimensional Scaling to Molecular Conformation*, Michael W. Trosset, Rice University

4:25 *A Software Package for Interactive Correspondence Analysis*, Jason Bond and George Michailidis, UCLA

4:45 *Testing Partial Homogeneity for Multinomial Proportions*, W. D. Kaigh and E. F. Schuster(*), Department of Mathematical Sciences, The University of Texas at El Paso

5:05 *Data Mining, Data Perturbation and Degrees of Freedom*, Mohsen Pourahmadi, Statistics, Northern Illinois University, DeKalb

5:25 *Floor Discussion*

◦◦ **Monte Carlo, Simulation, and Bayesian Methods** Galveston

Chaired by Peter Olofsson, Rice University

3:45 *On Rules For Conducting Statistically Analyzable Experiments with Seeds for Stochastic Simulation Models*, Danielle Mihram, University of Southern California

4:05 *Further Monte Carlo Investigations of a Model for Non Gaussian Radar Clutter Generation*, Jorge L. Romeu, SUNY-Cortland, NY

4:25 *Asymptotic Analysis in Bayesian Inference*, Joseph N. Ladalla, Mathematical Sciences, University of Illinois at Springfield

4:45 *The Effects of Non-Gaussian Noise and Different Boundary Rules on Two-Dimensional Linear and Nonlinear Smoothers*, Karen Kafadar and Markus Emsermann, University of Colorado-Denver

5:05 *Floor Discussion*

Thursday, 6:30 p.m. - 10:30 p.m.

Banquet Houston I-II-III

6:45 Cocktails

7:30 Doors open

7:45 Dinner

8:30 *Has High Speed Computing Caused A Paradigm Shift in Statistical Science?*, James R. Thompson, Department of Statistics, Rice University

Friday

Friday, 8:15 a.m. - 10:00 p.m.

•• **Wavelets and Image Processing** Houston I
Chaired by Guy Nason, University of Bristol
8:15 *Alternative to the Basis Pursuit Algorithm for Wavelet Denoising*, Andrew Bruce, Sylvain Sardy, and Paul Tseng, MathSoft, Inc.
8:45 *Generalized Coiflets, A New Wavelet System for DSP*, C. Sidney Burrus and Jan E. Odegard, Rice University
9:15 *Statistical Signal and Image Processing using Wavelet-Domain Hidden Markov Models*, Richard Baraniuk, Rice University
9:45 *Floor Discussion*

•• **Virtual Reality and Data Analysis** Houston II
Chaired by Webster West, University of South Carolina
8:15 *Immersive Methods for Exploratory Analysis*, Edward J. Wegman, Center for Computational Statistics, George Mason University
8:45 *Exploring Associations Among Mid-Atlantic Stream Indicators Using Dynamic Multivariate Graphics and Geographic Mapping in a Highly Immersive Virtual Reality Environment*, Dianne Cook, Laura Nelson, Carolina Cruz-Neira, Jürgen Symanzik, Anthony R. Olsen, Sue Pierson, and Noel Cressie, Department of Statistics, Iowa State University
9:15 *Virtual Reality as a Tool for Statistical Research*, A. J. Rossini and R. W. West, Department of Statistics, University of South Carolina
9:45 *Floor Discussion*

•• **Semi-Parametric Modeling — State of the Art and Computational Issues** Houston III
Organized and chaired by Michael Schimek, KFU Graz, Austria, with discussants
8:15 *Computer-Assisted Semiparametric Generalized Linear Models*, Marlene Mueller, Humboldt University, Berlin, Germany
8:45 *Discussion*, Joan G. Staniswalis, University of Texas, El Paso
8:50 *Semiparametric Regression Analysis for Recurrent Event Interval Counts*, Joan G. Staniswalis, Peter F. Thall, and John D. Salch, MD Anderson Cancer Center, Houston, Texas
9:20 *Discussion*, Doug Nychka, North Carolina State University
9:25 *A Bayesian Approach to Bivariate Nonparametric Regression*, Robert Kohn and Mike Smith, Australian Graduate School of Management, University of New South Wales, Sydney, Australia
9:55 Discussants: Michael Schimek, KFU Graz and Leonhard Knorr-Held, LMU Munich.

Data Augmentation Techniques El Paso ○○
Chaired by Scott Baggett, Rice University
8:15 *Fitting Latent Changepoint Models for Serial Biomarkers Using the EM Algorithm*, Stephen W. Gulyas, Elizabeth H. Slate, and Bruce W. Turnbull, Cornell University Statistics Center
8:35 *Missing Value Imputation with Iterated Binary Recursive Partitioning*, Dan Steinberg, Richard Carson, and Leo Breiman, San Diego State University
8:55 *Griddy-Gibbs Sampler in a Bayesian Hierarchical Mixed Linear Model*, Mithat Gönen, Applied Logic Associates, Houston, Texas
9:15 *Gibbs Sampling for Estimating Linear Models for Ranks*, George R. Terrell, Statistics, Virginia Polytechnic Institute
9:35 *Multivariate Bayesian Variable Selection and Prediction*, Philip J. Brown, Marina Vannucci(*), and Tom Fearn, University of Kent at Canterbury, UK
9:55 *Floor Discussion*

Friday, 10:30 a.m. - 12:15 p.m.

National Research Center on Statistics and the Environment Houston I ••
Organized by Peter Guttorp, University of Washington. Chaired by Doug Nychka, North Carolina State University
10:30 *NRCES: The University of Washington vision*, Peter Guttorp, University of Washington
11:00 *Establishing the National Research Center on Statistics and the Environment*, Lawrence H. Cox, US Environmental Protection Agency, Research Triangle Park, NC
11:30 *Discussion*, Abdel H. El-Shaarawi, National Water Research Institute, Burlington, Ontario, Canada
12:00 *Floor Discussion*

Math/Stat Massive Information Retrieval Houston II ••
Organized and chaired by Jim Maar, U.S. Department of Defense
10:30 *Intelligent Information Management Using Latent Semantic Indexing*, Michael W. Berry, University of Tennessee, Knoxville
11:00 *Statistical Aspects of the WEBSOM System in Organizing Document Collections*, Samuel Kaski, Timo Honkela, Krista Lagus, and Teuvo Kohonen, Helsinki University of Technology, Espoo, Finland

11:30 *Document Routing and Filtering: Experiments in the Application of Classification to Textual Documents*, Jan Pedersen, Verity, Inc.
12:00 *Floor Discussion*

•• **Imaging** Houston III
Chaired by Laszlo Pallos, CDC
10:30 *Ultrasonic Imaging of Cast Ductile Iron Projectiles*, Wendy L. Poston, O. Thomas Holland, and Edward J. Wegman, Naval Surface Warfarce Center, Dahlgren, Virginia
11:00 *Application of Statistical Visualization Techniques to Image Classification*, J. L. Solka, G. W. Rogers, and W. L. Poston, Naval Surface Warfare Center, Dahlgren, Virginia
11:30 *Synthetic Resampling Methods for Variability Estimation in Functional Images*, Ranjan Maitra, Statistics and Data Mining Research Group, Bellcore
12:00 *Floor Discussion*

◦◦ **Education and the Web** El Paso
Chaired by Keith Baggerly, Rice University
10:30 *WebStat: An Environment for Statistical Analysis on the World Wide Web*, R. Webster West and R. Todd Ogden, University of South Carolina
10:50 *GUIs for Power and Sample Size*, Russell V. Lenth, The University of Iowa
11:10 *A Web-enhanced Introductory Statistics Course*, E. James Harner and William C. Wojciechowski, Statistics and Computer Science, West Virginia University
11:30 *Learning Statistical Concepts using Web-based Dynamic Graphics*, William C. Wojciechowski and E. James Harner, Statistics and Computer Science, West Virginia University
11:50 *Floor Discussion*

Friday, 12:15 p.m. - 1:30 p.m.

Box lunch furnished courtesy of Visual Numerics, Inc.

Plenary Address Houston I
Chaired by John Salch, Rice University
Fortran Mathematical Software: A Distributed Network Library, Richard J. Hanson, Visual Numerics, Inc. Houston, TX

Plenary Address Houston II
Chaired by Roy Welsch, MIT
Massive Data Sets, Data Mining, and Decision Support, Siddartha Dalal, Susan Dumais, Jon Kettenring(*), Verghese Kurien, Allen Mcintosh, and Ranjan Maitra, Bellcore

Plenary Address Houston III
Chaired by Arun Jain, Rice University
Computer Models of Proteins in Action, George N. Phillips, Biochemistry and Cell Biology, Rice University

Friday, 1:30 p.m. - 3:15 p.m.

Dimension Reduction and Projection Pursuit Houston I ••
Chaired by Brian Ripley, Oxford University
1:30 *Functional Projection Pursuit*, Guy P. Nason, University of Bristol, Bristol, U.K.
2:00 *Approximate Distance Clustering*, Carey E. Priebe, The Johns Hopkins University
2:30 *Dimension Reduction via Kernel Bias Decomposition*, Mark Marson and David W. Scott, Department of Defense
3:00 *Floor Discussion*

Business Data Mining: Application Views Houston II ••
Organized and chaired by Clark R. Abrahams, NationsBank Consumer Credit Policy
1:30 *Introduction*, Clark R. Abrahams, NationsBank Consumer Credit Policy
1:35 *Business Data Mining in Practice*, Michael J. A. Berry and Gordon Linoff(*), MRJ Technology Solutions, Cambridge, MA
2:05 *Data Mining for Simulation in a Process-Control Domain*, Marc Goodman, Continuum Software, Inc., Woburn, MA
2:35 *Strategic Integration: Utilizing Data Mining to Create Intelligent Customer Communication Programs*, Robert Singer, CMS
3:05 *Floor Discussion*

Statistics and Statisticians in the 21st Century: In Appreciation of Herman Chernoff's Distinguished Contributions and Unique Style as He Retires in 1997 Houston III ••
Organized by Arnie Goodman. Chaired by Barry Brown, M.D. Anderson Cancer Center
1:30 *Data Mining, Statistical Methods Mining, and History of Statistics*, Emanuel Parzen, Department of Statistics, Texas A&M University
2:00 *Using Statistics: The Perspective From Funders and Consumers*, A.C. Petersen, W.K. Kellogg Foundation
2:30 *Career Insurance for the 21st Century*, Arnold Goodman, University of California at Irvine
3:00 *Floor Discussion*

Sampling and Time Series El Paso ••
Chaired by Katherine Ensor, Rice University
1:30 *Novel General Features of the Census Bureau's X-12-ARIMA Seasonal Adjustment Program*, David Findley and Bor-Chung Chen, U.S. Census Bureau

2:00 *The Use of Weighing Designs and Inverse Theory on Composite Sampling Problems*, Vicki Lancaster and Sallie Keller-Mcnulty, Kansas State University, Manhattan, Kansas

2:30 *Forecasting with Stable Seasonal Pattern Models with an Application of Hawaiian Tourist Data*, Rong Chen and Thomas B. Fomby, Texas A&M University

3:00 *Floor Discussion*

∘∘ **Evolutionary Computation and Statistics** San Marcos

Chaired by Chad Shaw, Rice University

1:30 *A Parallel Program for 3-Arm Bandits*, Robert Oehmke, Janis Hardwick(*), and Quentin F. Stout, University of Michigan

1:50 *From Evolutionary Operation to Parallel Direct Search: Pattern Search Algorithms for Numerical Optimization*, Virginia Torczon and Michael W. Trosset, The College of William & Mary

2:10 *Cooperative Coevolutionary Subspace Pursuit*, Bradley C. Wallet, David J. Marchette, and Edward J. Wegman, Naval Surface Warfare Center, Dahlgren, VA

2:30 *New Analysis of the Nelder-Mead Algorithm*, C. T. Kelley, NC State University, Raleigh, NC, USA

2:50 *Floor Discussion*

Friday, 3:15 p.m. - 4:15 p.m.

∘∘ **Poster session** San Antonio

Statistics for Environmental Risk Assessment, Tony R. Fagan, Parsons Engineering Science, Inc.

Friday, 3:45 p.m. - 5:30 p.m.

•• **Computing Issues in Environmental Statistics** Houston I

Organized and chaired by Peter Guttorp, University of Washington

3:45 *Computational Issues in Fitting Spatial Deformation Models for Heterogeneous Spatial Correlation*, Wendy Meiring, Peter Guttorp, and Paul D. Sampson, Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, CO

4:15 *Design and Evaluation of Networks for Monitoring Air Quality*, Douglas Nychka and Nancy Saltzman, North Carolina State University and the National Institute of Statistical Sciences

4:45 *The Kriged Kalman Filter*, Colin Goodall, Kanti Mardia, Ed Redfern, and Paco Javier Alonso, Health Process Management, Doylestown PA and Penn State University

5:15 *Floor Discussion*

•• **Virtual Departments** Houston II

Organized by Jan De Leeuw. Chaired by Luke Tierney, University of Minnesota

3:45 *Virtual Departments: The Technology*, Balasubramanian Narasimhan, Department of Statistics, Stanford University

4:15 *Teaching in the Virtual Department*, Jan De Leeuw, Interdivisional Program in Statistics, UCLA

4:45 *What Is There In Research?*, A.J. Rossini, University of South Carolina

5:15 *Floor Discussion*

Hardware Architectures and Networks to Support Data Mining Applications Houston III ••

Organized by Tony Elam and David Scott, Rice Univ. Chaired by Tony Elam, Rice University

3:45 *MDAS- A Massive Data Analysis System*, Arcot K. Rajasekar, Chaitan Baru, Richard Frost, Richard Marciano, Reagan Moore, and Mike Wan, San Diego Supercomputer Center, La Jolla, California

4:15 *Data Engines for Large Scale Mining*, Robert Hollebeek, University of Pennsylvania

4:45 *Remarks on Data Mining at Baylor College of Medicine*, Denise Moreau, Baylor College of Medicine

5:15 *Floor Discussion*

Graphical Methods El Paso ∘∘

Chaired by Roxy Walsh, Rice University

3:45 *An Interactive Graphical Tool for Exploring Sequential Dependencies in Categorical Data*, Mark Fitzgerald, Los Alamos National Laboratory

4:05 *Selection Sequences - Interactive Analysis of Massive Data Sets*, Martin Theus and Heike Hofmann, AT&T Labs, Murray Hill, NJ

4:25 *Computational Aspects of Data Analysis Using Flowgraph Methods*, Aparna V. Huzurbazar, Dept. of Statistics, University of New Mexico

4:45 *Interactive Visual Mining of Massive Multidimensional Datasets*, Ted Mihalisin and John Timlin, Department of Physics, Temple University

5:05 *Generalised Linking as a Means of Analysing Complex Data Sets*, Adalbert F.X. Wilhelm, Center for Computational Statistics, George Mason University

5:25 *Floor Discussion*

Density Estimation and Regression San Marcos ∘∘

Chaired by David Marchette, NSWC

3:45 *A New Data Driven Mixture Estimator for Spatially Dependent Observations*, Jeffrey L. Solka and David J. Marchette, Naval Surface Warfare Center, Dahlgren, Virginia

4:05 *Local Polynomial Regression for Binary Response*, Aaron Aragaki and Naomi Altman(*), Cornell University

4:25 *The Bumpy Road to the Mode Forest*, Michael C. Minnotte, David J. Marchette, and Edward J. Wegman, Utah State University

4:45 *A Smoothed Bootstrap Adaptive Density Estimator via The Fast Gauss Transform*, Mark Farmen, The Ohio State University

5:05 *Floor Discussion*

Saturday

Saturday, 8:15 a.m. - 10:00 a.m.

•• **Topics in Biomathematics and Biostatistics** Houston I
Chaired by Ralph Frankowski, U.T. School of Public Health

8:15 *Cox Models for Subsampled Data*, Terry M. Therneau, Mayo Clinic
8:45 *Smoothing Spline Models for Stratified Longitudinal Data*, Babette A. Brumback and John A. Rice, Harvard University, Boston, Massachusetts
9:15 *Selection of Covariates for Population Pharmacokinetic Models*, William M. Sallas and Jerry R. Nedelman, Biostatistics Dept., Novartis Pharmaceuticals Corp., East Hanover, NJ
9:45 *Floor Discussion*

•• **Highlights of JCGS** Houston II
Organized by Bill Kennedy. Chaired by Sallie Keller-McNulty, Kansas State University

8:15 *Finding Modes with the Filtered Kernel Estimator*, David J. Marchette and Edward J. Wegman, Naval Surface Warfare Center, Dahlgren, Va
8:45 *Flexible Response Surface Methods via Spatial Regression and EBLUPS*, Michael A. O'connell and Russell D. Wolfinger, Becton Dickinson Research Center, RTP, NC.
9:15 *Visual Explanation Tools for Graphical Models*, Russell G. Almond and David Madigan, Educational Testing Service, Princeton, NJ
9:45 *Floor Discussion*

•• **Smoothing** Houston III
Chaired by Peter Rousseeuw, University of Antwerp

8:15 *Three Sides of Smoothing: Categorical Data Smoothing, Nonparametric Regression, and Density Estimation*, Jeffrey S. Simonoff, Department of Statistics and Operations Research, New York University, New York, NY
8:45 *The Error and the Computational Complexity of a Multivariate Binned Kernel Density Estimator*, Lasse Holmstrom, Rolf Nevanlinna Institute, University of Helsinki, Finland
9:15 *Using Weighted, Smooth Quantile Estimates to Establish Medical Normal Ranges*, Michael D. Lock, Douglas Haney, and Michael E. Tarter, Becton Dickinson Immunocytometry Systems, San Jose, CA
9:45 *Floor Discussion*

○○ **Classification, Trees, and Optimization** El Paso
Chaired by Nancy Glenn, Rice University

8:15 *Variable Selection in Regression Via Repeated Data Splitting*, Peter F. Thall, Kathy E. Russell, and Richard M. Simon, M. D. Anderson Cancer Center
8:35 *Bagging in the Presence of Outliers*, Peter Hall and Berwin A. Turlach(*), CMA, Australian National University, Australia
8:55 *An MLE Strategy for Combining CART Models*, William D. Shannon and David Banks, Washington University School of Medicine, St. Louis
9:15 *Evaluating Decision Trees Using Trade-offs*, Padraic G. Neville, SAS Institute, Inc.
9:35 *Floor Discussion*

Saturday, 10:30 a.m. - 12:15 p.m.

Bayesian Methods Houston I ••
Chaired by J. Lynn Palmer, M.D. Anderson Cancer Center

10:30 *Analysis of High Frequency Data by Markov Chain Monte Carlo Methods*, Siddhartha Chib, Washington University
11:00 *Some Adaptive Monte Carlo Methods for Bayesian Inference*, Luke Tierney, University of Minnesota, Minneapolis, MN
11:30 *A Process-Convolution Approach for Spatial Modeling*, David M. Higdon, Institute of Statistics and Decision Sciences, Duke University
12:00 *Floor Discussion*

Neural Nets and Pattern Recognition Houston II ••
Chaired by Lasse Holmstrom, University of Helsinki

10:30 *Polychotomous Classifiers and PACTS*, Trevor. J Hastie, Statistics Department, Stanford University, California
11:00 *Fusing Diverse Algorithms*, John F. Elder, Quantitative Solutions, Charlottesville, Virginia
11:30 *Can Statistical Theory Help Us Use Neural Nets Better?*, Brian D. Ripley, University of Oxford, Oxford, UK
12:00 *Floor Discussion*

Linking Software for Environmental and Spatial Data Sets Houston III ••
Chaired by Debby Swayne, Bellcore

10:30 *Linking Statistical Summaries and Maps: An Approach To Understand Massive Spatial Data Sets*, Daniel B. Carr, Anthony R. Olsen, Jean-Yves (Pip) Courbois, and Suzanne Pierson, George Mason University, Fairfax, VA
11:00 *Spatial Data Analysis in the Dynamically Linked ArcView/XGobi/XploRe Environment*, Juergen Symanzik, Thomas Koetter, Swetlana Schmelzer,

Sigbert Klinke, Dianne Cook, and Debby Swayne, Iowa State University

11:30 *Applications of the Average Shifted Histogram in Environmental Economics*, Gerald Whittaker and David W. Scott, Economic Research Service/USDA, Washington, DC

12:00 *Floor Discussion*

◦◦ **Business Applications and Experimentation** El Paso
Chaired by Rachel Boeckenhauer, Rice University

10:30 *Detecting Outliers in Statistical Data using Cook's D Statistic*, Donald R. Jensen and Donald E. Ramirez(*), Virginia Polytechnic Institute, Blacksburg, Virginia

10:50 *Upper Confidence Bounds for Soft Error Rates in Integrated Circuits*, José G. Ramirez and Isabel S. Llatas, Digital Semiconductors, Hudson, MA

11:10 *An Empirical Study on the Investment Risk of Taiwanese Stock Market*, Ken Hung and David Cheng, Western Washington University

11:30 *Efficient Testing of Electronic Devices*, Hans Engler, Michael Souders, and Gerard N. Stenbakken, Georgetown University, Washington, D.C.

11:50 *Robust Design of Experiments and Information Set of Experimenting*, Anatoly A. Naumov, Novosibirsk State Technical University,Russia

12:10 *Floor Discussion*

◦◦ **Biometric and Survival Methods** San Marcos
Chaired by Peter Thall, M.D. Anderson Cancer Center

10:30 *O'Brien's Extended t-test for Paired Observations*, Peter C. Wollan, Biostatistics Section, Mayo Clinic, Rochester MN

10:50 *Parameter Estimation in Compartmental Models Using the Laplace Transform with Applications to Pharmacokinetics*, Martha Contreras and Changmei Liu, Biometrics Unit, Cornell University, Ithaca, NY

11:10 *Air Force Aging Aircraft Research Programs*, Peter W. Hovey, Wright Laboratories, US Air Force

11:30 *Estimation of Multistate Survival Model*, Rafiq I. Chowdhury, M. Ataharul Islam, M. A. Shah, and Syed A. Mohsin, Kuwait University, Kuwait

11:50 *Martingale Inference of Multi-state Progressive Stochastic Processes*, John J. Hsieh, University of Toronto, Toronto, Canada

12:10 *Floor Discussion*

Thursday

Invited Session: Keynote Address

Data Mining and Statistics: What's the Connection?

Jerome H. Friedman
Stanford University

Data Mining is used to discover patterns and relationships in data, with an emphasis on large observational data bases. It sits at the common frontiers of several fields including Data Base Management, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization. From a statistical perspective it can be viewed as

Thursday, 10:15 a.m. - 12:00 p.m.

Invited Session: Uncertainty Analysis for Computer Modeling Systems

Uncertainty Analysis for Computer Modeling Systems

Albert M. Liebetrau, Guang Chen, and Feng Gao
Pacific Northwest National Laboratory, Richland, Washington

Complex natural and human phenomena are being studied with increasingly complex computer modeling systems in which a number of component models are linked to replicate the underlying phenomenon. The computer models in such systems are often sophisticated models in their own right, with heavy computing requirement. When joined with other similar models, they may produce a system model with very long computing times, often on the order of several days. New methods of uncertainty analysis are required for system models constructed in this manner. Conventional methods of sensitivity and uncertainty analysis are intractable for modeling systems that require so much computer time.

In this paper, we propose a strategy for uncertainty analysis that employs a simplified model of the underlying model system. This simplified system model employs analogues of major component models that are linked so as to retain the essential structure of the underlying system. A key requirement of the simplified system model is that it require much less computing time than its underlying parent. The simplified system model is then employed for nearly all exploratory sensitivity and uncertainty analysis, thereby greatly reducing the number of runs of the parent model. The strategy is particularly effective when a structured logic diagram is used to represent the underlying system. With such a representation, the value of obtaining additional information (refined models, additional data, etc.) at key modeling points to reduce uncertainty in model outputs is readily assessed. In this way, various alternatives for gathering information and model development can be examined until an effective variance reduction strategy is found. We have found the SIMULINK code to be particularly well suited for the computation required to implement this strategy. We illustrate our methods with an example drawn from a program to resolve safety issues concerning the waste storage tanks at the Hanford site.

Structural Model Uncertainty in Stochastic Simulation Models

Michael D. Mckay
Los Alamos National Laboratory, Los Alamos, NM, USA

Prediction uncertainty in stochastic simulation models can be described by a hierarchy of components, beginning at the top with structural model uncertainty and moving down through model parameter uncertainty, model input uncertainty and, finally, stochastic variability. This paper investigates structural uncertainty for a prototypical discrete event simulation model. It looks at the variability of model prediction associated with level of detail or aggregation. The research objective is to develop a general methodology for assessing contributions from various submodels to overall prediction uncertainty.

Estimating the Response, Derivatives, and Transmitted Variance using Computer Experiments

James R. Koehler
University of Colorado at Denver

The Kriging approach to modelling deterministic simulators is easily extended to use gradient information and to estimate directional derivatives. Additionally, from these directional derivative estimates an estimate of transmitted variance can be obtained. During this talk I will develop the formula for the mean-squared-error of the transmitted variance estimator. To assess the trade-offs in modelling the response and the transmitted variance I will present two performance criteria: the integrated-mean-squared-error of the response and the square-root of the averaged-mean-squared-error of the transmitted variance. Designs based on Latin Hypercube Sampling that estimate either total gradients, directional derivatives, or no derivative information are presented along with a cost function. Measures of efficiency are discussed for comparing these designs in terms of their predicting the response surface or its gradient. Finally, the results of these comparisons for some selected sample sizes, number of input variables, and type of response surface are discussed.

Invited Session: New Directions for Statistical Packages and Environments

The Theme of ‘Context’ in Statistics Software

John Sall
SAS Institute, Cary NC

The theme ’context’ unifies a number of ideas for improving statistical software tools for researchers.

1. Table variables for keeping better track of data.
2. An outliner document surface with special scrolling rules.
3. An Interactive surface that unfolds from general to specific.
4. Analysis objects that are self-aware, in order to work better with the user interface, the scripting system, and state storage.

These ideas of context are needed to help the next generation of statistical software maintain a competitive edge over business spreadsheets as the software of choice for engineers and researchers.

Tables and Graphs

Leland Wilkinson
SPSS Inc. and Northwestern University

There are tables of graphs (e.g. trellises) and graphs of tables (e.g. mosaics). Tabulation and quantitative graphics can be formalized by a single language with three operators. Examples of nesting, crossing, and concatenating are presented for both tables and graphs.

Matrix – The Mother of all Data Structures

Cleve Moler
The MathWorks, Inc.

The nonmathematical uses of the word ’matrix’ reflect its Latin origins in ’mater’, or mother. For MathWorks, the word has two meanings – a representation of a linear mapping and the basis of our entire existence. This talk shows how the use of matrices has evolved in MATLAB to become the general purpose data structures introduced in MATLAB 5. The examples illustrate our view of the role of statistical computing in a general purpose technical computing environment.

Invited Session: Computational Environmental Geostatistics

Spatial vs. Ensemble Statistics of Spatial Simulation Algorithms

Donald E. Myers
Department of Mathematics, University of Arizona

Spatial simulation algorithms are designed to generate realizations of random functions and to reproduce certain statistical characteristics of the random function. In practice only a small number of realizations are generated and hence the characteristics of the realizations are more important than those of the random function. Most algorithms reproduce the statistical characteristics, such as the covariance function, mean and univariate distribution, only in an average sense. At least second order stationarity is an implicit assumption in using the algorithms. Empirical and theoretical results for convergence are given for several of the commonly used algorithms such as LU, Sequential Gaussian and Simulated Annealing. The latter is an example where each realization reproduces the statistical properties with large probability.

Geostatistical Methods for Estimating Total Mass of Mercury in the Waters of the Florida Everglades

Stephen L. Rathbun
Department of Statistics, University of Georgia

Determination of the efficacy of efforts to reduce mercury contamination in the Florida Everglades requires estimates of the total mass of mercury in the waters of that region from observations of mercury concentration and water depth on a network of sample sites. In general, total mass can be estimated by numerical integration of a predictor for the product of water depth and mercury concentration over the region, and multiplication of the resulting estimate by a constant to obtain the appropriate units of measurement. Optimal predictors of water depth and mercury concentration can be obtained via cokriging, which turn requires a model for the joint distribution of these two variables. During the dry season, a portion of the Everglades is dry, so the distribution of water depth is truncated at zero. Simply treating water depth at dry sites as being equal to zero and ignoring the effects of this truncation results in overestimates of water depth, especially in dry and shallow regions. This in turn will result in overestimates of total mass. A method is introduced to reduce this bias. The proposed method is then compared to various ad hoc procedures using data from the Florida Everglades.

Statistical Modeling of Vector-Valued Spatial Data Using Gradient Processes

David A. Andrews and Dennis D. Cox
Rice University, Houston, Texas

We explore the statistical modeling of wind in a city-sized area using the gradient of a spatial “potential” process. Gradients of stochastic processes are defined. We also discuss some innovative methods for the determination of an appropriate covariance model in this setting. The final model is to be used in a ground-level ozone analysis for Houston, Texas.

Contributed Session: Numerical Methods and Algorithms

Computing Cook’s Distance for Nonlinear Regression in S-Plus

C. Shane Reese
Texas A&M University, College Station, TX

Cook’s Distance for linear regression is available as standard output in most regression software packages. We suggest a simple S-Plus function to compute Cook’s Distance for nonlinear regression. Two examples are used to illustrate the use of the function.

An Extension of the Mixed Primal-Dual Bases Algorithm to the Case of More Constraints than Dimensions, with Applications to Nonparametric Shape Constrained Problems in Regression, Maximum Likelihood Density Estimation, and the Bioassay Problem

Mary C. Meyer
University of Michigan

The problem of minimizing a nonlinear function with linear inequality constraints is considered, where the set defined by the constraints is a closed convex cone. The mixed primal-dual bases algorithm (Fraser and Massam, 1989) for regression under inequality constraints minimizes a quadratic function over such a cone in a finite number of steps, with the restriction that the number of constraints does not exceed the number of dimensions in the space. Some applications, however, require more constraints than dimensions, and in this paper the algorithm is extended to this more general case. Properties of the constraint cone and its polar cone are presented in the generality necessary in this situation. One suprising result is that the number of generators of the constraint cone can be much larger than the number of generators for the polar cone. Applications are presented for least-squares regression, and regression in which a roughness penalty term is included in the function to minimize. It is shown that shape constrained maximum likelihood estimators for densities and the bioassay problem can be efficiently found by iteratively minimizing a sum of squares using the extended algorithm.

Derivatives of Fortran Programs, Automatically

Alan Carle
Rice University
Mike Fagan
Rice University, Houston, Texas

We will describe the ADIFOR automatic differentiator for Fortran and some of our experiences with using ADIFOR. ADIFOR, developed through a collaboration of Argonne National Laboratory and Rice University, transforms a Fortran program that computes a function f into a new Fortran program that computes JS, for J the Jacobian of f and S a suitably-dimensioned matrix. In particular, we can compute the entire Jacobian, or we can compute columns of the Jacobian, or a matrix-vector product JV (without first forming J). The talk will conclude by recasting ADIFOR as a tool for flexibly overloading the floating point operations in FORTRAN programs and mention some potential applications of this tool.

An Efficient Method for Representing, Analyzing and Visualizing Massive, High Dimensional Data

Tamraparni Dasu and Theodore Johnson
AT&T Labs - Research

The purpose of this paper is to develop an inexpensive nonparametric method to the analysis of massive high dimensional data, with a strong emphasis on application in the industrial setting of automated data collection and storage. The data are first "ordered" using a distance distance metric such as the Euclidean distance. A partition of the data set into ordered sectors is created based on distance quantiles. The distance is measured from a robust measure of location such as the vector of trimmed means of each of the variables. The trimming is only used to establish a robust "center" of the data cloud. The computation of the quantiles is done efficiently using a recent algorithm that makes only one pass over the data. This property is important because a very large data set cannot be stored in memory during processing.

Each of the distance-based sectors is assigned a summary consisting of:

- easily computable multivariate statistics such as the mean vector and variance matrix,
- the marginal distributions and their characteristics such as quantiles, histograms and other summaries.

An important aspect of the problem is that the data representations or summaries used to represent each sector should be such that they can be meaningfully combined across sectors or data sets. This property ensures that we can compute the summaries in only two passes over the data set.

The summary vector is used as a “footprint” for the multivariate data cloud. Footprints can be used to compare two multivariate data clouds as well to discriminate between groups within the same data cloud. A battery of well known multivariate tests is applied within each sector as well as to the entire footprint vector. In addition, more expensive multidimensional scaling and clustering methods might now be applied to the footprints, if appropriate. A simple form of visualization of the data is facilitated by the use of spider web charts, with each radial axis representing a dimension and a desired statistic from the summary vector of each sector plotted along these axes.

Efficient Estimation and Model Selection in Graphical Models

Dag Wedelin
Chalmers University of Tech, Gothenburg, Sweden

Graphical models have under different names been widely studied in statistics (graphical models, markov graphs, log-linear models) as well as in AI (belief networks, causal networks). While such models have many desirable properties, they also give rise to complex computational (NP-hard) problems, making them very difficult to apply to realistic problems with many variables.

We present a set of efficient algorithms for determining the interaction structure in a multidimensional binary sample. This is made possible by an algorithm for approximate ML estimation in a given undirected structure, working together with a heuristic algorithm for determining the structure, using the MDL criterion.

We show that the algorithms are able to reliably reconstruct the structure for problems with more than 100 variables, including many of the directions.

Thursday, 12:15 p.m. - 1:00 p.m.

Plenary Address

An Insider’s View of Networking Trends

Bruce Nelson
Chief Science Officer, Cisco Systems, San Jose, CA

This talk will discuss realistic three-year trends, including topics like ”Is network video real?”, ”When does NT really make sense?”, ”What’s better than ATM?”, ”What are Microsoft and Intel really doing?”, ”Should I care about Internet 2?”, ”Can my network ever be secure?”, ”Will I ever get cheap megabits to my house?”, ”What’s Cisco really doing?”.

Thursday, 1:30 p.m. - 3:15 p.m.

Invited Session: Numerical Methods for Large Problems

Interior-Point Methods: New Algorithmic Methodology for Large-Scale Inequality Constrained Optimization

Richard A. Tapia
Rice University

As applications become more complicated and more demanding we are witnessing an increasing need for algorithms that can handle large optimization problems with a large number of inequality constraints. The standard (classical) approaches are not serving us well. A new methodology for inequality constraints, referred to as interior-point methodology, has recently appeared on the computational methods horizon. In this talk the speaker will give a brief survey of the standard approaches and point out their short-comings. The new interior-point methodology will then be discussed.

Software and Algorithms for Estimation Problems in Reflection Seismology

William W. Symes
Rice University

Contemporary reflection seismology generates massive amounts of data encoding very detailed information about the Earth's subsurface. Decoding this information presents redoubtable problems of algorithm design and software engineering. This presentation will discuss some novel estimation procedures and software designs supporting them.

Latin Supercube Sampling for Very High Dimensional Simulations

Art B. Owen
Statistics Department, Stanford University

Some very high dimensional integrands arise in problems such as graphical rendering, particle transport, computational finance and queuing systems. The methods of choice for such problems are usually Monte Carlo (MC) or quasi-Monte Carlo (QMC). In theory QMC methods are far more accurate than MC, at least asymptotically, but they can lose their effectiveness in high dimensions. Recent work has centered on using QMC on a small set of key dimensions and MC on the others. This talk considers how to use QMC on more than one set of input dimensions, by randomizing the run order within each set, generalizing the method of Latin hypercube sampling. In combination with methods of engineering the integrand to have more low dimensional structure, the result is to extend the reach of QMC methods to harder problems, including some with infinite dimension. When the QMC methods themselves are randomized it is possible to get a data based estimate of the accuracy of the integration.

Invited Session: Educational Issues and Examples

Graphs in Statistical Data Analysis: Is the Medium the Message?

R. Dennis Cook and Sanford Weisberg
University of Minnesota

In an influential article in the February 1973 *American Statistician*, Frank Anscombe remarked that "[m]ost textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs." This general observation no longer holds. Graphs of data are everywhere. Statisticians often bemoan the poor construction of the graphs produced, but even when the construction is adequate, graphs can be of little value. To be useful, we suggest that useful graphs must have a *context* induced by associated theory, and that a graph without the well-understood statistical context is hardly worth drawing.

Hubs and Signposts: Incremental Implementation of Statistical Strategy.

R.W. Oldford
University of Waterloo

This paper addresses the problem of incorporating more information on analysis strategies into quantitative programming environments. Among the many difficulties associated with the problem is the balance of decision control between the user and the software.

Given a programming environment that permits generic functions (i.e. functions which dispatch to different methods based on the value of their arguments), the paper proposes that designers develop methods which implement a generic function called *signposts* to operate on every individual data structure in the environment. The function is to return a displayable data structure such as a collection of "signposts" that would lead the user to further investigations of the data structure. Selecting a signpost invokes the corresponding investigation. Every object can be displayed with its signposts alongside.

It is also proposed that higher level guidance be organized into interactive visual displays called *hubs*. Hubs display pertinent information on some stage of the analysis and offer the user the opportunity to proceed to other stages. These might in turn be investigated using other hubs. Hubs might also lead to signposted displays of relevant data structures. Reciprocally, signposts could lead to hubs.

It is envisioned that through signposts and hubs, elements of strategy could be incrementally implemented in any such programming environment. Examples using the *Quail* environment will be shown.

Technology in the Teaching and Learning of Statistics

Adrian Bowman
Statistics Dept., University of Glasgow, UK

A good statistical computing environment is an essential provision for a training in applied statistics. In addition to tools for data analysis, the graphical facilities provided by modern software environments allow them to be used in increasingly flexible ways in teaching, including animations and dynamic illustrations of concepts and techniques. There is now also a generation of software tools which make it relatively easy to put attractive text, images and dynamic graphics (of a non-statistical type) on a computer screen. However, there is much still to be learnt about the design of effective educational materials using these tools.

Experiences of developing material under the STEPS (STatistical Education through Problem Solving) project will be described. This collaborative work involved several groups in the UK, in Glasgow, Lancaster, Sheffield, Leeds, UMIST, Nottingham and Reading. The project was motivated by the need to provide strong motivation to students who take statistics courses as part of another subject. This can, of course, be done through good lecture material. The aim of the STEPS project is to support this, by allowing students to explore problems within their own disciplines in computer-based form. It is intended that through this exploration statistical issues arise naturally, within a computing environment which seeks to offer a certain amount of educational support and which attempts to provide a structure within which students can work there way through a problem. Examples will include the design of a clinical trial, and how to tell the sex of a seagull.

Invited Session: Business Data Mining: Methodological Viewpoints

A Systems Approach to Knowledge Discovery

Kenneth L. Reed
HyperParallel, Inc., Mission Viejo, California

Data mining or knowledge discovery are new terms for what natural scientists have been doing for years, only we call it ecological modeling and analysis. During the late 1960's and early 1970's, hundreds of scientists world wide participated in the International Biological Program to develop mathematical models of ecosystems and ecological processes.

Ecological models were attempts to understand and predict ecosystem behavior based on theoretical models, heuristics, and available data. The systems were complex, understood at the wrong level of resolution, poorly represented in data, and difficult to measure and validate. The tools we used consisted of mathematics and statistics. Most of the models were deemed to be failures. From the mid 1970s on it has been difficult to obtain funds for ecological modeling. But in fact many valuable lessons were learned.

The data mining community is falling into some of the same traps the ecologists of the IBP days fell into. This could result in a general discrediting of data mining, similar to the collapse of AI and CASE. Some of the traps include overselling, proliferation of tools whose use is poorly understood, failure to consider the complete system, and expecting all the required information to be data resident. I will illustrate these points with examples and show how a systems approach can yield better results.

The Importance of Preprocessing and Exploratory Data Analysis in Data Mining

Troy Haines
American Management Systems, Inc., Redwood City, California

Preprocessing is an often over looked step in the knowledge discovery process. Generally, preprocessing consists of any task that is undertaken prior to the mining run. In practice, preprocessing occurs throughout the knowledge discovery process and includes techniques that assist in outlier detection, variable transformation, discretization, sampling, and dimension reduction. Missing and noisy data are especially prevalent in business databases which are not designed with predictive modeling or discovery in mind. Data mining products today are inconsistent in the way they address preprocessing. Many products supply only cursory preprocessing capabilities, either minimizing the importance, or insisting that preprocessing occur outside of the mining tool. Some data mining products adopt a philosophy of extreme automation, allowing very little, if any, user intervention. This paper reviews our experience with different data mining products preprocessing capabilities as applied to real world data. Our conclusions are two fold: (1) preprocessing capabilities should be better integrated with data mining tools and (2) analysts should not take preprocessing for granted. By employing many of the robust and proven exploratory data analysis techniques commonly employed by statisticians, analysts gain a familiarity with the data and an improved understanding of its peculiarities and unique properties. Ultimately this understanding yields better models and better business results.

Operational Data Considerations in Data Mining

Jim Sattler
Saber Consulting Services, San Francisco, CA

Computer technology in financial institutions is deployed principally to manage daily transactions efficiently. These on-line transaction processing systems (OLTP) serve certain business accounting needs but have consequences that make analysis difficult:

- massive amounts of data are entered into computerized storage;
- data are fragmented and poorly structured;
- data consist of elementary facts.

This paper pays close attention to specific technical problems involved in processing these data and discusses how to convert operational data for the purpose of data mining. It recommends a specific multi-step strategy to facilitate the creation of analysis data. This strategy involves variable transformation, sampling, summarization, and data warehousing.

Contributed Session: Spatial Statistical Problems

Estimating Magnitude and Frequency of Flood Peaks at Ungaged Sites Using Regression

Gary D. Tasker
U.S. Geological Survey, Reston, VA

A common hydrologic problem is the estimation of the magnitude and frequency of floods at sites where no flood data are available. An approach to the solution of this problem is to develop a group of regression equations relating specific quantile estimates, such as the 2-year, 10-year, 50-year, and 100-year recurrence interval peak flows determined at gaged sites to a set of physiographic and climatic basin characteristics, such as drainage area, basin slope, and mean annual precipitation. Generalized least squares (GLS) regression is used to account for differences in record length and the spatial correlation structure of sample estimates at the gaged sites for each individual equation. Next, the variables and error terms for these equations are transformed to a OLS regression model by multiplying by the square root of the inverse of the GLS covariance matrix. Seemingly unrelated regression (SUR) is used then to estimate the final regression parameters for the group of equations when there are significant correlations in the error terms of independently estimated regression equations and the explanatory variables are not the same for each equation.

Applications of Spatial Statistics to the Quantitative Analysis of Genetic Instability in Carcinogenesis

J.Jack Lee
Department of Biomathematics, University of Texas M.D. Anderson Cancer Center
Ren Yu and Waun Ki Hong
Department of Head and Neck, Thoracic Medical Oncology, University of Texas M.D. Anderson Cancer Center
Walter N. Hittelman
Department of Clinical Investigation, University of Texas M.D. Anderson Cancer Center

Genetic instability and the accumulation of genetic changes are driving forces in tumor development. To visualize this process and to quantify the degree of genetic instability, we utilized chromosome in-situ hybridization with centromere markers for chromosomes 7, 9, and 17 to detect the spatial array of chromosome copy numbers in tissues at risk for cancer development. Our goal was to detect both infrequent events (i.e., three or more chromosome copies per cell) as a measure of generalized genomic instability and to detect the outgrowth of clones with either over- or under-representation of chromosome copies. Specifically, we will present methods and findings in the following areas. (1) Averaged chromosome copies were computed by nonparametric smoothing. Cutoffs of abnormal chromosome copies to identify hotspot regions were selected based on the percentiles from lymphocyte data. Both generalized instability and localized clonal outgrowth were found in lung and cervix premalignant lesions. (2) Patterns of the hotspot regions were analyzed to determine whether they were distributed in complete spatial randomness or as a result of clonal outgrowth. Tissues from primary untreated lung cancer presented patterns of clonal outgrowth. On the other hand, tissues from metastatic lung tumor with radiation and/or chemotherapy (Normal lung tissue from non-lung cancer individuals treated with chemo- and/or radiotherapy) were more likely to exhibit patterns of complete spatial randomness. (3) Principal component analysis was applied to standardize the coordinates and to align tissues. The spatial correlations between chromosomes 7, 9, and 17 were studied. (4) Random coefficient regression models were applied to examine the amount of genetic instability as a function of the histology grade. Our results showed that the generalized instability increased during multistep tumorigenesis. (5) This quantitative approach may be useful in individual tumor risk assessment.

Classification of Mixtures of Spatial Point Processes

Nandini Raghavan and Prem Goel
Statistics, The Ohio State University
Somnath Ghosh
Aerospace Engg., Applied Mechanics and Aviation, The Ohio State University

We propose a methodology for comparing several competing models for spatial point processes; to assess goodness-of-fit to a given spatial point pattern. Our motivation for this problem comes from the need to classify spatial distributions of locations of second-phase inclusions and voids in heterogeneous materials. The existing literature in spatial analysis, devoted to significance tests for randomness (i.e. homogeneous Poisson process) versus specific alternative models (e.g. self-inhibiting or cluster processes), is not appropriate for this task.

We propose a class of spatial point processes to model the spatial distribution of locations of inclusions in these composite materials. We construct a supervised pattern recognition scheme using functionals based on nearest neighbor distances, second order statistics and spatial tessellations. Some of these functionals are familiar in the literature and we propose some others that appear to extract important differences among the various spatial point processes considered. We study the distributions of functionals of these statistics using data visualization tools like XGobi and Monte Carlo simulations. We demonstrate the feasibility of using a supervised pattern recognition scheme to classify a spatial point pattern and evaluate the performance of the procedure via computer experiments.

The Use of Threshold Statistical Methods for Detecting Long-Term Trends in Urban Ozone Levels

Humberto Vaquera and Jose A. Villasenor
Colegio de Posgraduados, Mexico
Janet Hughes
Tulane University, New Orleans, LA

An important issue in the analysis of ozone data is to determine if there is any trend in the ozone levels, particularly in high levels (harmful for people in large cities). One way to investigate for trends is by using threshold methods. The threshold methods are based on the assumption that the number of observations over a relatively high level threshold follows a Poisson distribution, and the magnitude of the exceedances over such threshold follows a generalized Pareto distribution. In this paper, the use of three threshold based statistical methods is considered. Some examples and a comparative study is given.

Persistence and Recurrence in Atmospheric Circulation

Andrew M. Fraser, Richard Smith, and Kvein Vixie
Systems Science, Portland State University
Padhraic Smyth
UC Irvine, Irvine, California

We examine persistence and recurrence of atmospheric circulation patterns in models fit to a record of Northern Hemisphere 700-mb heights since 1947. Our findings are consistent with those published by Kimoto and Ghil in 1993, in that a small number of "regimes" are identified. We consider a variety of model classes, the most complex being autoregressive hidden Markov models (ARHMMs) which we found in the speech research literature.

Contributed Session: Wavelets

Some Findings on the Covariance Structure of Wavelet Coefficients: Theory and Models in a Bayesian Perspective

Marina Vannucci
University of Kent at Canterbury, UK
Fabio Corradi
University of Florence, Italy

This contribution refers to a wavelet shrinkage technique: the wavelet transform, linear and orthogonal, permits a decomposition of noisy data into a set of wavelet coefficients; noise is then removed by shrinking the coefficients. For this purpose, a Bayesian approach is possible based on imposing a statistical model on the wavelet coefficients of the unknown signal and replacing them with the results of a Bayesian analysis. In contrast with other Bayesian models recently proposed, the adopted model does not assume that the wavelet coefficients of the unknown signal are independent. Theoretical results regarding the covariance structure of random wavelet coefficients are presented and used to specify the Bayesian model. Moreover, a hierarchical structure is implemented to learn about some of the parameters of the prior distributions. Applications to simulated signals are presented. The technique is generally applicable to the problem of modeling wavelet coefficients and has also been used for density estimation involving wavelets.

Wavelet Estimation Based on Quasi-Linear Approximation

Sam Efromovich
The University of New Mexico, Albuquerque

We discuss both theoretically and via intensive Monte-Carlo study a new procedure of spatially adaptive minimax estimation of functions from Besov spaces. The new procedure is based on a quasi-linear wavelet estimation that is a classical linear estimate enriched by a minimal number of nonlinear terms on finest scales. This estimate offers advantages of linear estimates, such as simplicity of adaptation and best estimation of smooth functions, together with good compression of data and spatial adaptivity of strongly nonlinear procedures.

Detection of Additive Outliers in Time Series Data Using Wavelets

S. Huzurbazar
University of Wyoming
Canan Bilen
University of Wyoming, Laramie, Wyoming

We propose a method based on wavelets for detecting additive outliers in time series data. The proposed method is easy to implement and requires no model fitting. Use of the method is illustrated through simulated examples as well as applications to ground water elevation data. The method is also compared with existing outlier detection procedures in time series.

Testing for Abrupt Jumps with Wavelets

R. Todd Ogden
University of South Carolina
Cheng Cheng
Johns Hopkins University

Various wavelet-based procedures for testing for a change-point are considered, with power computations under local alternatives. Such tests have broader application than the standard change-point problem, in that they are easily adapted to testing for jumps in an otherwise smooth function.

Thursday, 3:45 p.m. - 5:30 p.m.

Invited Session: Visualization: Applications and Environments

Analyzing Complex Data Sets: Visualization and Modeling

Linda Clark, William S. Cleveland, Lorraine Denby, and Chuanhai Liu
Bell Labs, Lucent Technologies, Murray Hill, NJ

Complex data sets with many observations often reveal complicated relationships among variables and complex statistical variation. Modeling such data is challenging. Searching the data to determine plausible models requires visualization methods that can reveal complexity and handle many observations. Modeling requires approaches that can account for the complexity, often requiring many parameters, but constraining them in some way to prevent using too many degrees of freedom.

Trellis display has proven useful as an approach to visualization of complex data sets because the multipanel conditioning methods can reveal the structure of complex relationships among variables and the multipage display creates sufficient screen real estate to allow the display of many observations.

Bayesian modeling has proven useful because modern computational methods allow great flexibility in modeling which in turn enables a feasible approach to accounting for complex relationships and complex statistical variation.

These issues are discussed in the context of data from a study of customer value analysis, or CVA. The data consist of over 5000 ratings by survey respondents of performance of 22 companies on 48 attributes. Visualization and modeling must face random location effects from nonnormal distributions, random variance effects, many covariates, error variables with a nonnormal distribution, scales of error variables that change with the values of covariates, time effects, repeated-measure effects, and missing values.

Visualizing Multi-Dimensional Spaces of Simulation Models

Steve Bankes
RAND

The analysis of public policy for complex systems such as the earth's climate, military combat, or the future of higher education, can be supported by performing large numbers of computational experiments with simulation models. Each such computational experiment can be thought of as a sample drawn from a multi-dimensional space of alternative plausible models for the system of interest. This approach to understanding complex and incompletely understood systems is known as "exploratory modeling."

The utility of exploratory modeling depends upon the use of adaptive sampling methods to efficiently allocate the computational experiments and in the use of graphical visualizations to reveal patterns in the results. A variety of search or adaptive sampling techniques can be useful in generating constellations of computational experiments that best support decisionmaking within computational constraints. Of these, algorithms that oversample near boundaries or thresholds in the outputs of experiments are particularly important for graphical display. Among visualization tools we utilize is an "automated scorecard" that allows browsing through multidimensional policy landscapes.

This approach has resulted in significant improvements in the study of several policy problems. As part of this talk I will describe several of these, including studies of global climate policy, military procurement, science and technology investment, corporate strategic planning, and higher education.

A Java-Based Graphics Production Class Library

Daniel J. Rope
Bureau of Labor Statistics

The GPL (Graphics Production Library) stems from a need for convenient tools for producing quality statistical graphics. Feedback from seminars shows that there is interest in applying new ideas for statistical graphics, but the reality is that people will use what is most convenient to create graphs. The problem worsens when the graphics are included in World Wide Web pages. These rendered bitmaps are often barely legible on low resolution computer screens. To solve both of these problems at once we are developing a platform-independent, extensible Java class library for producing graphics.

The library itself is composed of a set of reusable components (objects) that encapsulate drawing graphics. These objects can be "glued" together to create applets or applications, and the objects' underlying methods are used to change high level behavior and appearance. Applets are miniature software applications that run within Web browsers, and applications are standalone software products.

Beyond the ability to draw directly on the user's computer screen, one of the most significant advantages of this approach is that we can introduce a whole new level of user interaction with statistical graphics and we can promote understanding by providing meta-data via hypertext links, icons and pop-up windows. The user will be able to directly manipulate the graphic to more quickly answer his or her questions about the data. In a sense, the graphic will become the interface, since the interface is the graphic.

The demonstration will contain graphics using data from the academic (W.S. Cleveland's "barley" data) as well as the more practical, such as time series plots from the Current Population Survey and the Consumer Price Index.

Invited Session: Software for Data Mining and Data Warehousing

TreadMarks: Software Support for Distributed Data Mining on Workstation Clusters

Alan Cox
Department of Computer Science, Rice University

Treadmarks supports parallel computing on networks of workstations. Its main novel feature is that it provides a global shared address space across the different machines on a cluster. The shared address space distinguishes it from other well-known packages such as MPI that provide a message passing interface between machines. There is growing consensus in the parallel computing community that a shared memory interface is more desirable from the application programmer's viewpoint, allowing him or her to focus on algorithmic development rather than on managing communication. The challenge in providing a shared memory interface is to do so efficiently. To this end, TreadMarks incorporates several innovative features, including release consistency and multiple-writer protocols.

Data Mining for Internet Applications

David C. Martin
IBM Almaden Research Center

The IBM net.Mining Solutions group is developing applications of data-mining and machine learning for the Internet and emerging information dissemination areas. This paper describes the overall architecture of the net.Mining software and discusses the uses of that infrastructure to route people and content to an appropriate rendezvous.

Knowledge-Based Pre- and Post-Processors

Arun Jain
Assoc. VP Information Technology, Rice Univ.

Invited Session: Massive Data Sets in an Industrial Setting and Finance

Robust Statistics and Data Mining in Finance and Financial Engineering

R. Douglas Martin
Professor of Statistics, University of Washington and Chief Scientist, Data Analysis Products Division of MathSoft, Inc.

Cross-sectional and time series data arising in finance and financial engineering are often non-Gaussian by virtue of containing outliers. We begin with a number of simple but striking examples concerning cross-sectional data in finance, and financial time series *returns* (e.g., relative differences or log ratios) of equities and foreign exchange rates, where the classical LS slope estimates and correlation coefficients are rendered essentially misleading and useless by outliers, and where robust estimates give a very good description of the behavior of the bulk of the data. For *data mining* applications where many slopes and correlations need to be computed (e.g., when computing the beta of the CAPM theory for thousands of instruments), it will be important to compute both standard and robust estimates, and raise the appropriate red flag when there are significant differences. Then we focus on three applications of robust regression and robust covariance matrix estimation: (i) Linear modeling of monthly equity returns in terms of variables such as size (market capitalization in millions of dollars) and book-to-market; (ii) The capital asset pricing model (CAPM) and estimation of *beta*, and (iii) Portfolio optimization. Then we show how Trellis Graphics can be combined with classical and robust estimates to visualize multi-dimensional financial data in a particularly insightful way, that might be regarded as *true* data mining. We also illustrate how Dan Carr's hexagonal binning technique can be essential to support such visualization on large financial data sets.

Applications of Positive-Breakdown Methods

Peter J. Rousseeuw
University of Antwerp, Belgium

Positive-breakdown techniques can resist (and thus detect) a substantial fraction of contamination in the data. This property is often useful in practice. For instance, positive-breakdown regression methods are applied to electric power systems, where one needs to estimate the system's state variables robustly. They are also applied to the automatic coupling of fiber cables. Several other applications are in computer vision, e.g. for image recovery, surface reconstruction, and to detect moving objects in video from a mobile camera. Robust estimators of multivariate location and scatter have been used in geochemistry, to detect hidden mineralizations. Applications of such estimators to computer vision include image segmentation and determining shapes from color images.

War Stories from the Data Trenches

Richard A. Becker and Allan R. Wilks
AT&T Labs - Research, Florham Park, New Jersey

Data analysis for very large datasets is particularly challenging when the arrival of the data is relentless; it's like drinking from a fire hose. For two years we have been consuming about 300 megabytes per day, summarizing and storing it in a fully indexed database. In this talk we will describe our experiences, illustrate them with anecdotes, and formulate general rules for dealing with huge amounts of data.

Contributed Session: Multivariate Methods

Partitioning and Subsampling to Uncover Subtle Structure in Massive Data Sets

David M. Rocke
University of California, Davis

Some aspects of the statistical structure of a data set are automatically uncovered as the sample size rises. Examples incluse the mean of a population or the value of a regression coeficient. Other problems, particularly those that have a combinatorial aspect, become so difficult with rising problem size that exact computation becomes essentially impossible. This paper presents a probabalistic strategy based on partitioning and subsampling of the data set that has a feasible computation time even for massive data sets and that can uncover non-obvious structures such as clusters, mixtures, and multivariate outliers. The current prototype for this strategy is a method for robust estimation of multivariate location and shape and multivariate outlier identification (Rocke 1996, Annals of Statistics; Rocke and Woodruff 1996, JASA; Woodruff and Rocke 1994, JASA).

Applications of Multidimensional Scaling to Molecular Conformation

Michael W. Trosset
Rice University

Multidimensional scaling (MDS) is a collection of data analytic techniques for constructing configurations of points from information about interpoint distances. Such constructions arise in computational chemistry when one endeavors to infer the conformation of a molecule from information about its interatomic distances. For a number of reasons, this application of MDS poses computational challenges not encountered in more traditional applications. This presentation will sketch the mathematical formulation of MDS for molecular conformation problems, discuss some algorithms that have been proposed for their solution, and report some numerical results.

A Software Package for Interactive Correspondence Analysis

Jason Bond and George Michailidis
UCLA

A highly interactive, user-friendly object oriented software package written in Lisp-Stat is introduced that performs simple correspondence analysis, multiple correspondence analysis (also known as homogeneity analysis), and profile analysis (see Gifi (1990)). These three techniques are integrated into a single enviornment driven by a graphical interface which focuses on ease of use and takes advantage of Lisp-Stats advanced plotting capabilities. New techniques that assess the stability of the solution are also introduced. Some of the features of the package include colored graphics, incremental graph zooming capabilities, manual point seperation to determine identities of overlapping points, and stability and fit measures. The features of the package are used to show some interesting trends in a large educational dataset (a subset of about 2 dozen variables and over ten thousand observations from the National Educational Longitudinal Study of 1988).

Testing Partial Homogeneity for Multinomial Proportions

W. D. Kaigh and E. F. Schuster
Department of Mathematical Sciences, The University of Texas at El Paso

Likelihood ratio tests are derived and developed to assess partial homogeneity of estimated multinomial proportions. The general hypothesis of partial homogeneity postulates the equality of some specified proper subset of the multinomial proportions. Although the proper subset is predetermined, the values for all proportions are unspecified. Asymptotic chi-squared null distribution and power efficiency comparisons for these likelihood ratio procedures are explored through Monte Carlo power studies to validate chi-squared null distribution approximations for the likelihood ratio based test statistics, to explore their power efficiencies under alternatives, and to investigate stepwise procedures proposed to automate the process of finding partial homogeneity when it exists in a multinomial model. Testing and partial homogeneity model selection and goodness-of-fit methods are illustrated with both educational and real data applications including the popular unbalanced dice-like game *Pass the Pigs* and in analyzing survey data.

Data Mining, Data Perturbation and Degrees of Freedom

Mohsen Pourahmadi
Statistics, Northern Illinois University, DeKalb

Nowadays the same data set is used routinely in the formulation, selection, estimation, diagnostics and prediction of nonlinear and nonparametric models by relying on the power of modern computers. Is it possible to assess and compare performances of such models? What is the "size" or degrees of freedom (df) of a model in the face of excessive data mining? We address these questions in the context of the Canadian lynx data which has attained benchmark status in the time series literature. While no universal answer can be expected, a data perturbation procedure (Wong and Ye, 1996) is used to assess and adjust the impact of data mining on the (in-sample) prediction error variance by assigning an appropriate generalized degrees of freedom to each fitted model. A PPR model with the smallest (in-sample) prediction error variance among all models fitted to this data, turned out to be comparable and nearly equivalent to other models after adjustment for data mining. However, based on the MSE of out-of-sample prediction error, the semiparametric model $X_t = 1.08 + 1.37X_{t-1} + f(X_{t-2}) + \epsilon_t$, and Tong's self-exciting threshold autoregressive (SETAR) model perform much better than the PPR and other models known for the lynx data.

Contributed Session: Monte Carlo, Simulation, and Bayesian Methods

On Rules For Conducting Statistically Analyzable Experiments with Seeds for Stochastic Simulation Models

Danielle Mihram
University of Southern California

This paper examines and defines the conditions under which we statisticians can invoke designed experimentation with the seeds themselves of a stochastic and computerized simulation model. Since any computer-based statistical sampling experiment can be viewed as a simulation model of a statistician conducting such an experiment and since any computer-based Monte Carlo evaluation of an integral (i.e., solution of a differential equation) can be viewed as simulation model of an applied mathematician conducting such a numerical evaluation, the results (applicable to algorithmic simulation models) become quite general. We show how experimentation among the (non-stochastic) input variables of a simulation model divides them into environmental conditions and treatments, and even defines blocks (of experimental units). How then can we justify conducting experiments with the remaining (the stochastic) input variables, the seeds? We show how variance components resolves the issue.

Further Monte Carlo Investigations of a Model for Non Gaussian Radar Clutter Generation

Jorge L. Romeu
SUNY-Cortland, NY

In a 1993 Interface paper, the first stage of this study was presented. There, the generation problem of a Spherically Invariant Random Process (SIRP) X, obtained as x=s*Z, with Z a multivariate gaussian, was studied for the unidimensional and some multidimensional special cases. Here, we complete such investigation by first studying the case of a Bi Dimensional, K-Distributed SIRP that has a closed form solution to compare with. Then, we study the general, multivariate case obtaining via Monte Carlo the empirical (unidimensional) distribution of the quadratic form p=X'*(Sigma)-1*X of the SIRP, X. Since p characterizes uniquely any process X, we test the goodness of fit of p, via the Ozturk-Dudewicz statistic. We assess the distribution identification of SIRP X, via this method, for specific sample sizes, number of variates and intercorrelations. Finally, we use this characterization of SIRP X, to study the effects of the factors aforementioned on the identification procedure. Numerical results are presented.

Asymptotic Analysis in Bayesian Inference

Joseph N. Ladalla
Mathematical Sciences, University of Illinois at Springfield

Ladalla (l991) had developed an asymptotic expansion for the characteristic function of the postererior distribution of the parameter vector when sampling from multiparameter exponential family models. This tool (expansion for characteristic function) has a vital role to play in asymptotic analysis. We demonstrate this in providing a solution in closed form for the Behrens-Fisher problem. Even though asymptotic, the solution holds good for samples of sizes as small as 10.

The Effects of Non-Gaussian Noise and Different Boundary Rules on Two-Dimensional Linear and Nonlinear Smoothers

Karen Kafadar
University of Colorado-Denver
Markus Emsermann
University of Colorado-Denver, Denver, Colorado

Methods of two-dimensional smoothers usually involve kernels or locally weighted regression, sometimes constructed adaptively depending upon the apparent behavior of the function and the density of data points in local neighborhoods. One reason for their popularity is the ability to derive analytical results concerning their bias and variance, since they are linear smoothers. Nonlinear smoothers are also contenders but have been applied much less often since their performance is analytically intractable and hence cannot be compared theoretically with linear smoothers. In this study we compare loess with headbanging, a nonlinear smoother, on simulated data of a known function subject to various types of noise distributions (Gaussian, mixture of Gaussians, gamma, and slash). and using three different types of boundary rules. Finally, we illustrate these smoothers using elevation data in U.S. counties.

Thursday, 6:30 p.m. - 10:30 p.m.

Has High Speed Computing Caused A Paradigm Shift in Statistical Science?

James R. Thompson
Department of Statistics, Rice University

The paradigm shift in statistical science facilitated by high speed computing started over thirty years ago. The revolution began with John W. Tukey and his associates at Bell Labs and Princeton University, who were also active in the early development of computer science. To the original school of Exploratory Data Analysis have been added other related movements: those of nonparametric density estimation, nonparametric regression, neural nets, data mining, etc. This enlarged constituency of Exploratory Data Analysis has been generally disdainful of models and relies heavily on visualization. It remains to be seen whether the radical empiricism of EDA will prevail. The revolution continues.

Friday

Friday, 8:15 a.m. - 10:00 p.m.

Invited Session: Wavelets and Image Processing

Alternative to the Basis Pursuit Algorithm for Wavelet Denoising

Andrew Bruce
MathSoft, Inc.
Sylvain Sardy and Paul Tseng
Dept. of Mathematics, University of Washington

Penalized least squares regression methodology, such as ridge regression, plays a prominant role in statistical prediction. Recently, considerable attention has focused on the use of an L_1 penalty term instead of L_2 measures. Of particular note is the "basis pursuit" procedure of Chen, Donoho, and Saunders for nonparametric signal de-noising using over-complete adaptive waveform bases (e.g., wavelet packets). While an L_1 penalty term offers several potential advantages, it poses a difficult computational challenge to obtain the optimal solution. A key contribution of basis pursuit is a novel application of primal-dual log-barrier interior-point optimization techniques. Their optimization algorithm provides a way to solve the L_1 penalized regression problem in nearly-linear time, and can be applied on problems of high dimension where the number of basis functions is much larger than the number of data values. In this paper, we investigate an alternative optimization approach for a special class of problems based on "block coordinate relaxation" (BCR) techniques. BCR is a general method for nonlinear function optimization, and works by successively optimizing over coordinate blocks. In the signal de-noising problem, BCR can be applied whenever the overcomplete set of basis functions is the union of sets of orthogonal basis functions. Examples of when BCR can be applied include translation invariant wavelets, wavelet packets, and cosine packets. We study two approaches towards block selection: cyclical rules and optimal descent rules (Gauss-Southwell). For the optimal descent rules, the best basis algorithm is used to find the best block. We show that BCR is globally convergent, and empirically, BCR is faster than interior point methods for a variety of signal de-noising problems.

Generalized Coiflets, A New Wavelet System for DSP

C. Sidney Burrus and Jan E. Odegard
Rice University

By constraining both scaling function and wavelet moments to be zero, one can obtain reasonably smooth and approximatly symmetric wavelets that have interesting properties. The systems are not unique and that characteristic is particularly interesting.

Statistical Signal and Image Processing using Wavelet-Domain Hidden Markov Models

Richard Baraniuk
Rice University

Wavelet-based statistical signal processing techniques such as denoising and detection typically model the wavelet coefficients as independent or jointly Gaussian. These models are unrealistic for many real-world signals. In this paper, we develop a new framework for statistical signal processing based on wavelet-domain hidden Markov models (HMMs). The framework enables us to concisely model the statistical dependencies and nonGaussian statistics encountered with real-world signals. Wavelet-domain HMMs are designed with the intrinsic properties of the wavelet transform in mind and provide powerful yet tractable probabilistic signal models. Efficient Expectation Maximization algorithms are developed for fitting the HMMs to observational signal data. The new framework is suitable for a wide range of applications, including signal estimation, detection, classification, prediction, and even synthesis. To demonstrate the utility of wavelet-domain HMMs, we develop novel algorithms for signal denoising, classification, and detection.

Invited Session: Virtual Reality and Data Analysis

Immersive Methods for Exploratory Analysis

Edward J. Wegman
Center for Computational Statistics, George Mason University

This paper focuses on our experience with immersive techniques, i.e. virtual reality. We use both a group VR setting, with 3-D stereoscopic projection techniques and an individual VR setting, with head mounted displays. We discuss the hardware involved in these settings, some of the problems with each type of display, and some of our successes, including some work in progress. We plan to illustrate some of our immersive methods with video tape.

Exploring Associations Among Mid-Atlantic Stream Indicators Using Dynamic Multivariate Graphics and Geographic Mapping in a Highly Immersive Virtual Reality Environment

Dianne Cook, Jürgen Symanzik, and Noel Cressie
Department of Statistics, Iowa State University
Laura Nelson and Carolina Cruz-Neira
Iowa Center for Emerging Manufacturing Technology, Iowa State University
Anthony R. Olsen
US EPA National Health and Environmental Effects Research Laboratory, Western Ecology Division, Corvallis, Oregon
Sue Pierson
OAO, c/o USEPA NHEERL Western Ecology Division, Corvallis, Oregon

Geography inherently fills a 3-D space and yet we struggle with displaying geography using, primarily, 2-D display devices. Virtual environments offer a more realisticly-dimensioned display space and this is being realized in the expanding area of research on 3-D geographic information systems (GISs). Traditionally a GIS has only limited tools for statistical analysis, and 3-D GIS research has concentrated on the visualization of the geography. Here we discuss linking multivariate statistical graphics to geography in the highly immersive C2 environment at Iowa State University using mid-atlantic streams data.

Virtual Reality as a Tool for Statistical Research

A. J. Rossini and R. W. West
Department of Statistics, University of South Carolina

Virtual Reality (VR), as a model of the real world, is a natural environment for statisticians. With the advent of the World Wide Web, (WWW), the Virtual Reality Modelling Language (VRML), and the programming language Java, one can construct virtual statistical worlds which can be accessed by users from machines, across the Internet. In this talk, we discuss the use of virtual reality for data analysis as well as for research in methods and theory. While the obvious use of VR is as quick means of visualizing 3-dimensional objects, the addition of behaviors and interactions of the objects and the user extend the realm of possibilities for data analysis and description. Some examples of these methods will be interactively demonstrated. Another line of research is to consider the construction of a virtual statistical worlds which is tuned for the growth, development, and evaluation of statistical ideas and concepts. One preliminary attempt of this is demonstrated, and pitfalls to overcome and paradigms shifts which might be necessary are discussed.

Invited Session: Semi-Parametric Modeling — State of the Art and Computational Issues

Computer-Assisted Semiparametric Generalized Linear Models

Marlene Mueller
Humboldt University, Berlin, Germany

We introduce semiparametric extensions of the generalized linear model. One particular model of interest is the generalized partially linear model (GPLM) which allows a nonparametric modelling of the influence of the continuous covariables. The estimation procedure is introduced and a test on the correct specification of this model (vs. a parametric generalized linear model) is presented. The application to a data set on East–West German migration illustrates the use of this technique.

Semiparametric methods are hightly demanding on software. Flexibility for extensions, tools for efficient computation on the user level as well as interactive graphics to display the resulting curves/surfaces are such requirements. We thus complete our presentation by indicating the practical implementation in new version of the statistical computing environment XploRe.

Semiparametric Regression Analysis for Recurrent Event Interval Counts

Joan G. Staniswalis and Peter F. Thall
MD Anderson Cancer Center, Houston, Texas
John D. Salch
Rice University, Houston, Texas

This paper deals with analysis of data from longitudinal studies where the rate of a recurrent event characterizing morbidity is the primary criterion for treatment evaluation. We consider clinical trials which require patients to visit their clinical center at successive scheduled times as part of followup. At each visit, the patient reports the number of events which occurred since the previous visit, or an examination reveals the number of accumulated events, such as skin cancers. The exact occurrence times of the events are unavailable and the actual patient visit times typically vary randomly about the scheduled followup times. Each patient's record thus consists of a sequence of clinic visit dates, event counts corresponding to the successive time intervals between clinic visits, and baseline covariates. We propose a semiparametric regression model, extending the fully parametric model of Thall (1988), to estimate and test for covariate effects on the rate of events over time while also accounting for the possibly time-varying nature of the underlying event rate. Covariate effects enter the model parametrically, while the underlying time-varying event rate is modelled nonparametrically. The method of Severini and Wong (1992) is used to construct asymptotically efficient estimators of the parametric component and to specify their asymptotic distribution. A simulation study and application to a data set are provided.

A Bayesian Approach to Bivariate Nonparametric Regression

Robert Kohn
Australian Graduate School of Management, University of New South Wales, Sydney, Australia
Mike Smith
Department of Econometrics, Monash University, Clayton, Australia

This paper outlines a general approach to estimating a bivariate regression function in a nonparametric fashion. It utilises a Bayesian variable selection algorithm previously introduced by the authors to choose significant terms from a bivariate regression spline basis. Extensive simulated comparisons are provided which demonstrate the competitive performance of this approach against other data driven bivariate surface estimators prominent in the literature. It is then shown how the procedure can be extended to provide a general approach to nonparametric bivariate surface estimation in two problematic regression settings– the case where there are outlying observations and when the errors are thought to potentially auto-correlated. Simulated and real data examples illustrate the effectiveness of tackling such difficult problems using this methodology.

Contributed Session: Data Augmentation Techniques

Fitting Latent Changepoint Models for Serial Biomarkers Using the EM Algorithm

Stephen W. Gulyas, Elizabeth H. Slate, and Bruce W. Turnbull
Cornell University Statistics Center

We consider the analysis of serial biomarker readings used to screen and monitor individuals in a given population for a specific disease of interest. These readings are subject to measurement/misclassification error. For disease onset and progression, a latent stochastic changepoint process is proposed. This can be viewed as a generalization of a hidden Markov model. Techniques for computing standard errors and incorporating covariates in the disease and marker processes are considered. The EM algorithm is used to estimate the parameters of the changepoint and marker distributions. This study was motivated by the use of prostate specific antigen as a biomarker for prostate cancer.

Missing Value Imputation with Iterated Binary Recursive Partitioning

Dan Steinberg
San Diego State University
Richard Carson
Economics, University of California, San Diego
Leo Breiman
Statistics, University of California

Experiments with various methods of imputing missing values for all variables in a database in one fell swoop with CART are reported. Using a series of repeated CART runs, missing values are imputed for variables, then the imputed values are themselves used in further iterations to re-impute missing values until a convergence criterion is met. The methods differ in (a)whether or not predictors are provided with provisional imputed start values such as the sample median or most common categorical level; (b)whether imputed values are made available immediately for predicting other variables only after all variables subject to imputation have been analyzed; and (c) whether the variable being filled in retains its last round imputation or has its missing values temporarily reset to missing during the imputation run. Various degrees of missingness (minimal, moderate, heavy) and various patterns of missing values (random, informative) are used in the test runs.

Griddy-Gibbs Sampler in a Bayesian Hierarchical Mixed Linear Model

Mithat Gönen
Applied Logic Associates, Houston, Texas

We present an example of how one can significantly improve computational efficiency using a Bayesian hierarchical mixed linear model as opposed to modeling without conditioning on random effects. Our example uses non-standard prior distributions and leads to Gibbs conditionals with unknown normalizing constants, so we employ the griddy-Gibbs sampler to facilitate the computations.

Gibbs Sampling for Estimating Linear Models for Ranks

George R. Terrell
Statistics, Virginia Polytechnic Institute

Linear Models for ranks simultaneously generalize linear rank statistics and hierarchical multinomial logistic models, while vastly enriching both. However, until recently they were computationally intractable to estimate accurately; since the maximum likelihood equations require us to extract properties of nonidentically distributed order statistics. It turns out, though, that since the full conditional distributions of order statistics are trivial, Gibbs sampling is a natural way to simulate these distributions. Therefore, we can now apply linear rank models to real problems.

Multivariate Bayesian Variable Selection and Prediction

Philip J. Brown and Marina Vannucci
University of Kent at Canterbury, UK
Tom Fearn
University College, London, UK

The multivariate linear regression model is considered with p regressors. A latent vector with p binary entries serves to identify one of two types of regression coefficients, those close to zero and those not. Given this vector we assume a natural conjugate Normal framework, and mix over the 2^p latent vectors. The practical context is situations where it is important to choose a subset (or subsets) of p regressor variables which are good for prediction of all q responses simultaneously. In the application we wish to estimate the proportions of $q = 3$ components of a mixture of sugars (Sucrose, Glucose and Fructose) in varying amounts in aqueous solution from their near-infrared spectrum of $p = 160$ absorbances.

We give a simple method of eliciting hyperparameters of the prior distribution.

For prediction purposes we average over a set of likely *a posteriori* subsets. We give algorithms for fast computing of the posterior distribution, although our envisaged application involves such a large number of regressors, $(p = 160)$, that Markov Chain Monte Carlo approximations to the posterior distribution are required.

Friday, 10:30 a.m. - 12:15 p.m.

Invited Session: National Research Center on Statistics and the Environment

NRCES: The University of Washington vision

Peter Guttorp
University of Washington

We outline some of the research projects currently under way at the National Research Center for Statistics and the Environment. One of the main eff orts is to develop methodology for assessing deterministic models, taking into a ccount the uncertainty in inputs and outputs as well as in the model itself. We also work on ideas in ecological assessment, in space-time modeling of air pollu tion, in environmental sampling methods, and in risk assessment.

Establishing the National Research Center on Statistics and the Environment

Lawrence H. Cox
US Environmental Protection Agency, Research Triangle Park, NC

The U.S. Environmental Protection Agency recently established the National Research Center on Statistics and the Environment at the University of Washington-Seattle. USEPA's goal in establishing the Center is to provide a national presence for research and graduate education in environmental statistics. The University of Washington was selected based on a national call for proposals and competition. USEPA established the Center under a five-year, cooperative research agreement, and anticipates providing a total of $5 million to the Center during the five-year period. This presentation will highlight potential areas of research collaboration between USEPA and the Center and provide an amalgam of EPA views, visions, etc. on the Center (skewed toward those of the presenter), and other information about USEPA environmental statistics programs and research.

Invited Session: Math/Stat Massive Information Retrieval

Intelligent Information Management Using Latent Semantic Indexing

Michael W. Berry
University of Tennessee, Knoxville

Common approaches to retrieving textual materials from digital libraries depend on a lexical match between words in users' requests and keywords assigned to documents in large database. Because of the tremendous diversity in the words people use to describe the same document, such lexical-matching methods are necessarily incomplete and imprecise. Using mathematical models based on orthogonal decompositions (SVD, ULV, or URV), the implicit higher-order semantic structure or association of terms (or keywords) with documents can be derived from large, sparse "term-by-document" matrices. Terms and documents can be encoded for matching user queries in a high-dimensional vector space using the largest approximate singular vectors. This retrieval method is commonly referred to as Latent Semantic Indexing or LSI since the sub- spaces spanned by the approximate singular vectors represent important associative relationships between terms and documents that are not evident in individual documents. This talk will focus on up- and down-dating procedures which are needed to maintain a consistent LSI model in the case of time-varying text collections.

Statistical Aspects of the WEBSOM System in Organizing Document Collections

Samuel Kaski, Timo Honkela, Krista Lagus, and Teuvo Kohonen
Helsinki University of Technology, Espoo, Finland

WEBSOM is a novel method for organizing document collections onto map displays to enhance the interactive browsing and retrieval of the documents. The map is organized totally automatically according to the contents of the full-text documents by the Self-Organizing Map algorithm. The map display provides a visual overview of the whole document collection, and similar documents are located close to each other on the map display, which aids in the exploration. In this paper we concentrate on comparing the WEBSOM with other related methods. Particular emphasis is put on how effective the methods are in treating large document collections. The two-phase architecture of the WEBSOM system makes it possible to build contextual information about the relations of words off-line into a word category representation, which can then be utilized quickly on-line, when the documents are being encoded. In principle it is a computationally intensive operation to construct large map displays from the encoded document representations, but there exist several effective computational shortcuts.

Document Routing and Filtering: Experiments in the Application of Classification to Textual Documents

Jan Pedersen
Verity, Inc.

The application of classification methods to the problem of document routing and filtering faces several challenges: first the native dimensionality (the size of the collection vocabulary) is very large (tens of thousands), second the number of positive examples (typically a few hundred) is small relative to the size collection (typically hundreds of thousands). In addition, the evaluation criterion, average precision, is not differentiable, and hence is difficult to directly optimize. I will discuss dimensionality reduction and training set sampling techniques required to directly apply error minimizing methods to this problem. It will emerge that careful application will produce a classifier that out performs simpler, prototype and nearest-neighbor methods.

Invited Session: Imaging

Ultrasonic Imaging of Cast Ductile Iron Projectiles

Wendy L. Poston and O. Thomas Holland
Naval Surface Warfarce Center, Dahlgren, Virginia
Edward J. Wegman
George Mason University, Fairfax, Virginia

Current techniques for visualizing ultrasound data include B-scans (a 2-D representation of ultrasonic data from a slice of the object) and C-scans (a 2-D representation of ultrasonic data from the entire object). Each of these use aspects of the A-scan to envision flaws in the object. We propose using the A-scans to reconstruct a 3-D representation of the object, which makes flaw detection easier. In this paper, we apply our visualization method to the ultrasonic non-destructive testing of cast ductile iron projectiles.

Application of Statistical Visualization Techniques to Image Classification

J. L. Solka, G. W. Rogers, and W. L. Poston
Naval Surface Warfare Center, Dahlgren, Virginia

This paper detail our recent work in the application of existing and the development of new statistical visualization techniques to the problem of the identification of optimal feature sets for use in low-level vision processing. In particular we are interested in displaying the relationship of extracted feature sets that are often multi-class and high dimensional in nature. The identified feature sets are ultimately employed within a statistical classification framework to identify regions of interest within images. The techniques discussed will include parallel coordinates, the grand tour and a new visualization technique for the display of six dimensional feature sets.

Synthetic Resampling Methods for Variability Estimation in Functional Images

Ranjan Maitra
Statistics and Data Mining Research Group, Bellcore

The ability to obtain pixelwise quantitation of tissue biologic parameters such as glucose metabolic rate is one of the most powerful features of Positron Emission Tomography (PET). Many techniques have been suggested in the literature with regard to extracting these images of functional parameters from dynamic time-course sequences of reconstructed PET scans. Estimating variability in these images is important for the development of inference procedures in diagnostic settings. The complexity of the radio-tracer models as well as the non-linearity of the estimation algorithms used in the generation of a functional image make analytic formulae for the dispersions intractable. A natural extension of the usual resampling approach is impractical to implement because each image reconstruction in PET is a computationally demanding solution to a high-dimensional linear inverse problem. This paper suggests a computationally practical approximate simulation approach which approximates the distribution of the reconstructed PET scans by a Gaussian random field and performs a synthetic parametric bootstrap in the imaging domain. Results of experiments done to evaluate the approach in a simplified model one-dimensional problem are very encouraging. Smoothing the pointwise variance estimators is seen to improve their root-mean-squared (RMS) error characteristics. The suggested methodology is developed here in the context of functional images extracted by mixture analysis; however, the approach is general enough to extend to other functional imaging methods.

Contributed Session: Education and the Web

WebStat: An Environment for Statistical Analysis on the World Wide Web

R. Webster West
University of South Carolina
R. Todd Ogden
University of South Carolina, Columbia, SC

The field of statistics has already seen the benefits of the World Wide Web (WWW) through the transfer of various forms of static information such as technical reports, course materials, etc. With the advent of the Java computing language, dynamic applications (applets) can now be transfered over the WWW as well. Applets place virtually no requirements on the user's hardware and only minimal requirements on the user's WWW browser. In order to make use of this very general technology, WebStat, a basic statistical package written in the form of a Java applet, has been constructed and is now freely available over the WWW. A demonstration of WebStat's abilities is given, and the benefits and limitations of this package for computing in a basic statistics course are discussed. The ability to expand WebStat for new and advanced statistical procedures is also addressed.

GUIs for Power and Sample Size

Russell V. Lenth
The University of Iowa

Power and sample-size problems cannot be solved by a statistician in isolation; it is crucial that the investigator (usually a non-statistician) play the leading role, so that judgments about appropriate effect sizes and practical sample sizes are available. This talk concerns graphical user interfaces that can facilitate the process. With the right add-in functions, spreadsheets can be quite effective. In situations where several interrelated tests need to be examined at once, more sophisticated software is helpful. In the latter case, the Java object model can be used to great advantage. Both spreadsheet methods and Java applets will be demonstrated.

A Web-enhanced Introductory Statistics Course

E. James Harner and William C. Wojciechowski
Statistics and Computer Science, West Virginia University

The Department of Statistics and Computer Science at West Virginia University is developing Web enhancements to its algebra-based undergraduate introductory statistics course (Statistics 101). The long-term objective is to develop enough material to support distance learning using the Web. The Web content centers around an Interactive Workbook consisting of a series of 15 units each organized around a theme. Each unit consists of explanatory material containing static graphics, animations, demonstrations, examples and exercises. The animations and demonstrations are used to teach important statistical concepts whereas the examples focus on statistical problem solving. The demonstrations and examples step the students through a series of statistical operations using embedded Java applets or Lisp-Stat linked software code. The student then can work exercises which are implemented as Java-based forms. An exercise/answer sheet, keyed to the student ID and the instructor, is automatically returned as an HTML document, e-mailed to the student, and stored in a grader-accessible Java database on the server. The persistence of the database entries allows the student to modify answers until the due date and rejects entries for students not registered in the course. After the due date, the grader-annotated answers, along with the correct answers, are e-mailed to the student. The Java database also contains student grades and entries have appropriate access permissions.

Learning Statistical Concepts using Web-based Dynamic Graphics

William C. Wojciechowski and E. James Harner
Statistics and Computer Science, West Virginia University

Introductory students rarely master important statistical concepts such as the central limit theorem or power. The Web, combined with interactive statistical visualizations, provides an improved means of learning these principles. A series of interactive demonstrations, some written in Lisp-Stat and some in Java, have been developed as teaching aids. The underlying code is activated from an explanatory hypertextual environment—directly as Java applets or by using Lisp-Stat as a helper application. Students can visualize the effects of changing the parameters (or the underlying situation) defining probability distributions, confidence intervals, power, the central limit theorem, etc. The controls are reusable components and thus become familiar to the students within different statistical contexts. Controls are always embedded within a single graphics object to reduce clutter and eliminate the complexity of manipulating multiple windows. The Lisp-Stat demonstrations can also be used independently of a Web browser.

Friday, 12:15 p.m. - 1:30 p.m.

Plenary Address

Fortran Mathematical Software: A Distributed Network Library

Richard J. Hanson
Visual Numerics, Inc. Houston, TX

The use of Fortran 90 operators and function modules has resulted in a novel Visual Numerics, Inc. library product. Use of the VNI Fortran 90 library results in a readable and compact code. An example for solving linear algebraic systems and evaluating residuals is expressed as $y = A.ix.b$ for the solving step, and $r = b - (A.x.y)$ for residuals. This makes it possible to write code with a shorter development cycle than using ordinary subprograms.

Most of the operators and functions are generic in terms of precision and data type. In particular they support a box data type. This data type adds an additional dimension to an array. For example if the arrays A, b, and y are declared with sizes $A(n, n, k), b(n), y(n, k)$, then the code fragment $y = A.ix.b$ results in task-parallel solution of the problems.

Our support for parallel computing also includes an error processor. The output of error text, should it be required, is synchronized by the processing group reaching an empty state in the subprogram name call stack. The error processor also supports non-parallel use of the IMSL Fortran Libraries, which are a part of the DNFL product. This design feature allows for maintenance and migration of legacy applications.

Plenary Address

Massive Data Sets, Data Mining, and Decision Support

Siddartha Dalal, Susan Dumais, Jon Kettenring, Verghese Kurien, Allen Mcintosh, and Ranjan Maitra
Bellcore

Statisticians are zeroing in on the mysteries of massive data sets while computer scientists are offering data mining as an interdisciplinary approach to tackling such problems. Meanwhile practitioners need to get on with their work. In this talk, we will describe our approaches to several large problems which we have encountered at Bellcore and also outline future research challenges based on these experiences.

Plenary Address

Computer Models of Proteins in Action

George N. Phillips
Biochemistry and Cell Biology, Rice University

More and more of biology is being shown to be 'computable', including the molecular basis of many fundamental processes. The interface between the behavior of individual molecules and statistical phenomena is currently a very active area. Some examples in oxygen binding to heme proteins and the switching on and off of muscle contractions will be described in this context.

Friday, 1:30 p.m. - 3:15 p.m.

Invited Session: Dimension Reduction and Projection Pursuit

Functional Projection Pursuit

Guy P. Nason
University of Bristol, Bristol, U.K.

Functional projection pursuit is the extension of exploratory projection pursuit to curved or functional data. Projection pursuit attempts to obtain the optimal view of high-dimensional data. In the curved data setting projection becomes integration, the projection vectors are functions and the purpose of FPP is to find the projection function that gives the most insightful view of the data. Fortunately all the existing wisdom concerning projection indices is still valid.

Interesting questions are raised by this extension: how does one differentiate the projection index with respect to the projection functions? What is the role of sphering?

The method is illustrated using simulated and real data. The real set comes from a study investigating the differences in biomedical time-series for babies that sleep with their mother (co-sleep) and for those that don't (solitary sleep).

Approximate Distance Clustering

Carey E. Priebe
The Johns Hopkins University

We address the problem of statistical pattern recognition (clustering, classification, discriminant analysis) for n observations in d-dimensional Euclidean space when $n << d$. As such, we are concerned with circumventing the curse of dimensionality. Our approach, termed *approximate distance clustering,* is motivated by dimensionality reduction methods which embed points from a high-dimensional Euclidean space into a lower dimensional space while approximately preserving pairwise distances between points. The methodology involves non-linear projections based on inter-point distances to random subsets. Combinatorial algorithms are developed and theoretical, simulation, and experimental results are presented indicating that approximate distance clustering has application in many challenging pattern recognition scenarios. This work is joint with Lenore J. Cowen.

Dimension Reduction via Kernel Bias Decomposition

Mark Marson
Department of Defense
David W. Scott
Rice University

In this talk, we discuss two new formulas for computing the asymptotic integrated squared bias (AISB) of a general multivariate kernel density estimator. These formulas are then used to compute orthogonal projections of data onto lower dimensional subspaces in which as much as possible of the AISB is retained. The methods we use for dimension reduction favor those subspaces on which the clustering information of the data is prominent. We will contrast our method with informative components analysis and principle components analysis.

Invited Session: Business Data Mining: Application Views

Business Data Mining in Practice

Michael J. A. Berry and Gordon Linoff
MRJ Technology Solutions, Cambridge, MA

The talk will give examples of how data mining is being used today in several industries including retail, banking, telecommunications and manufacturing. For each case studied, I will describe the problem that is to be resolved by data mining, the approach used, and the reasons for selecting that approach.

The retail case describes work that is just getting started with the catalog division of a major apparel retailer. Data mining is being applied to gain a better understanding of existing customers so as to be able to mail them smaller, more focused catalogs.

The banking case examines the way that historical transaction data can be used to develop profiles of different classes of demand deposit account holders and to recognize when an individual customer changes classes which may represent a cross selling opportunity or provide a warning flag for impending attrition.

The telecommunications case shows how call detail data can be coaxed into revealing valuable information such as who sends faxes from a home telephone number and is therefore likely to be working at home.

The manuafacturing case shows how decision tress can be used to build simulations of environments that change over time. Data mining based simulations of this kind have been used for process control in coffee roasting and barley malting applications.

The talk closes with an examination of the obstacles to successful data mining in today's businesses with today's commercial data mining products and suggests ways that these products could be improved.

Data Mining for Simulation in a Process-Control Domain

Marc Goodman
Continuum Software, Inc., Woburn, MA

An important problem in the Industrial Process Control is the detection of situations that will result in a forced shutdown of the process. Often, in complex real-world processes, there is no model for describing the process in sufficient detail to identify such situations. Projective Visualization, a Data Mining technique for learning to simulate a process based on observations of that process, has been successfully applied to this problem in the domain of continuous-feed coffee roasting in a factory setting.

Data in the coffee roasting domain consists of low-level sensor reading captured at 30 second intervals. These sensor readings are noisy and incomplete. Further, understanding temporal relationships in the data is crucial for projecting future states of the roaster and predicting shutdowns. All of these factors play a significant role in the choice of appropriate Data Mining techniques for predicting roaster shutdown.

This talk will describe Projective Visualization and how it was applied to the roaster control problem. It will also describe key features and limitations in Data Mining techniques as they are currently applied to process control problems.

Strategic Integration: Utilizing Data Mining to Create Intelligent Customer Communication Programs

Robert Singer
CMS

This session will focus on leveraging data mining and predictive modeling to create data driven customer communications programs. The presentation will focus on the integration of advanced data mining techniques in the areas of marketing and collections.

The talk will focus on how businesses leverage their data to understand their customer base and apply the learning gained through mining and data discovery to develop relationships with their individual customers. A high level overview of modeling in industry, including various techniques and applications, will be presented. The remainder of the talk will focus on integrating the output of data analysis into program development, execution, and evaluation.

The objectives of the session are: 1. To present how data mining is used in marketing and collections departments. 2. To present how predictive modeling is used in marketing and collections departments. 3. To present the integration process of linking data mining and modeling techniques to the creation and evaluation of creative customer communications programs.

Invited Session: Statistics and Statisticians in the 21st Century: In Appreciation of Herman Chernoff's Distinguished Contributions and Unique Style as He Retires in 1997

Data Mining, Statistical Methods Mining, and History of Statistics

Emanuel Parzen
Department of Statistics, Texas A&M University

The emerging field of "data mining" is (according to Pregibon (1996)) a blend of statistics, artificial intelligence, and database research. It is projected to be a multi-billion dollar industry by the year 2000. I propose the name "statistical methods mining" for the process of developing various maps of the world of statistical knowledge in order to apply to data mining the enormous virtual encyclopaedia of statistical knowledge that exists in the literature. I propose that planning modern core, applied, and computational statistical activities should be based on statistical history whose mottos are "the purpose of statistical computing is insight not numbers" and "the purpose of statistical history is insight (about future influences) not (of past influences)". The goal of statistical methods mining is to promote the optimum practice of analysis of diverse (possibly massive) data sets by enabling researchers in all fields to be made aware that statisticians have important roles to play to provide expertise about classical and modern statistical methods, theory, practice, applications, and computational techniques for data mining and statistical inference (extraction of information and identification of models from data). Contents of this paper are: 1. Future of statistics is bright, future of statisticians needs planning; 2. What went wrong in the recent history of statistical science; 3. Balancing core research and interdisciplinary research; 4. Statistical education and statistical modeling are analogous iterative cycles; 5. Statistical history was never neglected in past thinking about the future of statistics.

Using Statistics: The Perspective From Funders and Consumers

A.C. Petersen
W.K. Kellogg Foundation

Like many fields, academic statistics has, over the past two decades, become increasingly focused on the basic discipline rather than continuing to build from the needs of other fields of research and practice. With the recognition that real problems are complex, and the advent of massive databases incorporating that complexity, the field of statistics is more important than ever before. This presentation will describe the needs of funders of research and programs for valid statistical approaches and inferences in order to satisfy the ultimate consumers of funded work. It will also present suggestions for effective roles for statisticians.

Career Insurance for the 21st Century

Arnold Goodman
University of California at Irvine

If the 21st century becomes the century for statistics, will it also become the century for statisticians? Likely not. The world is driven by chance and by data, but not by us. Our last four ASA presidents have voiced their concerns.

Where are the technologies of databases and networks plus software engineering driving statistics in 21st century? Where are the economics of markets and the presures of employment forcing statisticians in the 21st century?

However, there is a way to protect yourself from risk by providing your own insurance with attitude and action. What insurance can be provided by personal attitude? What insurance can be provided by personal action?

Invited Session: Sampling and Time Series

Novel General Features of the Census Bureau's X-12-ARIMA Seasonal Adjustment Program

David Findley and Bor-Chung Chen
U.S. Census Bureau

X-12-ARIMA is the Census Bureau's new time series modeling and seasonal adjustment program. It was released in April, 1997 and can be obtained via ftp. The program has some innovations that can be recommended for incorporation, after appropriate modification, into a variety of statistical programs. For example, there is a "history" option that causes the program to run on a sequence of truncated versions of a given time series for the purpose of obtaining statistics from "historical" records of out-of-sample forecast performance or of seasonal adjustment revisions. These statistics permit the comparison of competing models and competing seasonal adjustment options without the assumption that any model is correct or that any adjustment is optimal. Also there is a simple facility to cause the program to run with the same options on a set of hundreds of series. This is useful for simulation or bootstrap analyses. We shall present a variety of applications of these features.

The Use of Weighing Designs and Inverse Theory on Composite Sampling Problems

Vicki Lancaster and Sallie Keller-Mcnulty
Kansas State University, Manhattan, Kansas

The definition of a composite sample is a sample that is formed from two or more samples. The subsequent measurement is taken on the entire composite or an aliquot removed from the composite. The constituent information on the individual pre-composite samples is observed indirectly through the composite measurement. Combinatorial, retesting schemes are commonly employed after the composites are measured to identify anomalous pre-composite samples. In many environmental appications these retesting schemes are not a viable option because the entire contents of the pre-composite sample is used in forming the composite.

This paper discusses the feasibility of combining composite sampling with the inverse theory techniques of linear regularization in an environmental site characterization investigation. This application is important because Federal legislation has mandated the cleanup of hazardous waste sites, thus creating the need to characterize these sites for various chemical constituents. An abundance of samples, high measurement costs, and limited budgets create the appeal of compositing samples.

Forecasting with Stable Seasonal Pattern Models with an Application of Hawaiian Tourist Data

Rong Chen
Texas A&M University
Thomas B. Fomby
Southern Methodist University, Dallas, Texas

We propose several variations of the Stable Seasonal Pattern (SSP) model first introduced by Marshall and Oliver (1970) and study their prediction procedures. Depending on the type of data (count data or continuous variable) we propose different treatments. Previously SSP models have been applied to trendless data or adapted to trending data in as hoc ways. In the models considered here conditional independence allows the seasonal pattern and trend to be modeled separately while prediction utilizes both efficiently. In an out-of-sample forecasting experiment conducted on Hawaiian tourism data, one of the proposed variations demonstrates its long-term forecasting potential relative to a multiplicative, seasonal Box-Jenkins model.

Contributed Session: Evolutionary Computation and Statistics

A Parallel Program for 3-Arm Bandits

Robert Oehmke, Janis Hardwick, and Quentin F. Stout
University of Michigan

We will discuss a new parallel program for optimizing and analyzing 3-arm Bernoulli bandit problems with sample sizes of 100 or more. With simple modifications, the program can also be used for 2-arm bandits with trichotomous responses. Our results are the first to permit one to optimize these complex problems with practical sample sizes. We will explain the computing issues involved in developing this program. A parallel computer was used because the dynamic programming and forward induction employed require $\Theta(n^6)$ time and $\Theta(n^5)$ space for a sample of size n.

From Evolutionary Operation to Parallel Direct Search: Pattern Search Algorithms for Numerical Optimization

Virginia Torczon
The College of William & Mary
Michael W. Trosset
Rice University, Houston, Texas

G.E.P. Box's seminal suggestions for Evolutionary Operation (1957) led other statisticians to propose algorithms for numerical optimization that rely exclusively on the direct comparison of function values. These contributions culminated in the development of the widely-used simplex algorithm of Nelder and Mead (1965).

Recent examination of these popular methods by the numerical optimization community has produced new insights. Numerical experiments and carefully constructed examples have revealed that the Nelder-Mead algorithm may be unreliable even in fairly simple situations. In contrast, many of the original methods, which we collectively describe as *pattern search methods*, are guaranteed to converge to a stationary point of the objective function under conventional nonlinear programming assumptions. In addition, the structure of these algorithms is such that they are easily parallelized.

We will briefly survey the history of pattern search methods and explicate their common structure, pointing out the key features that the Nelder-Mead simplex algorithm lacks. We will close with some practical suggestions for using pattern search methods in serial and in distributed computing environments.

Cooperative Coevolutionary Subspace Pursuit

Bradley C. Wallet
Naval Surface Warfare Center, Dahlgren, VA
David J. Marchette
Naval Surface Warfare Center, Dahlgren, VA USA
Edward J. Wegman
Center for Computational Statistics, George Mason University, Fairfax, VA USA

We develop a method for using cooperative coevolution for performing local dimensionality reduction. Lower dimensional linear subspaces are found which most closely fit data in a higher dimensional space. Data points are then projected orthogonally to the appropriate subspace. This allows for dimensionality reduction while still preserving the local structure. Our method is applied to a number of data sets, and we discuss how this method could be used for discriminant analysis, clustering, and data compression.

New Analysis of the Nelder-Mead Algorithm

C. T. Kelley
NC State University, Raleigh, NC, USA

The Nelder-Mead algorithm can stagnate and converge to a non-optimal point, even for very simple problems. In this note we propose a test for sufficient decrease which, if passed for the entire iteration, will guarantee convergence of the Nelder-Mead iteration to a stationary point if the objective function is smooth. Failure of this condition is an indicator of potential stagnation. As a remedy we propose a new step, which we call an oriented restart, which reinitializes the simplex to a smaller one with orthogonal edges which contains an approximate steepest descent step from the current best point.

Friday, 3:15 p.m. - 4:15 p.m.

Contributed Session: Poster session

Statistics for Environmental Risk Assessment

Tony R. Fagan
Parsons Engineering Science, Inc.

U.S. environmental statues (e.g., CERCLA) require that remedial actions at hazardous waste sites be protective of human health and the environment. Quantitative risk evaluation is a framework for estimating the potential risks to human health and the environment. Statisitics are an integral step in the risk evaluation process. Historically, environmental data were assumed to be normally or lognormally distributed. The corresponding t-test was used to compare data from potentially site-impacted areas and background data for developing a list of chemicals of potential concern for further risk analysis. The maximum site concentration was then used to represent the average exposure concentration of those chemicals above background in the risk analysis. A graphical distributional analysis and a two-tiered approach to comparing site and background data are presented using (1) a suite of central tendency tests, and (2) an application of upper tolerance limits. Methods based on censored data and the boostrap are presented. This case study illustrates the statistical analysis performed for several waste disposal and ordnance detonation sites at the US Army National Training Center, Fort Irwin, California.

Friday, 3:45 p.m. - 5:30 p.m.

Invited Session: Computing Issues in Environmental Statistics

Computational Issues in Fitting Spatial Deformation Models for Heterogeneous Spatial Correlation

Wendy Meiring
Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, CO
Peter Guttorp and Paul D. Sampson
University of Washington, Seattle, Washington

Environmental monitoring networks are recording pollutant levels, weather, and a myriad of other factors. It is often of interest to estimate these values at locations where records are not available. Many spatial estimation procedures rely on spatial covariance models. Assumptions of spatial isotropy or stationarity may be violated due to factors such as topography and local emissions structures. In this talk we discuss computational issues for a heterogeneous (spatially non-stationary) model for spatial correlations between point monitoring sites. The modeling procedure involves deforming the geographic space into a new space (D-space) where inter-site correlations depend only on distances. Correlations between unmonitored sites are then estimated as a function of distance in the D-space. The estimation of the D-space locations of the monitoring sites, and of the parameters of the isotropic D-space variogram model is a difficult multidimensional problem. The dimensionality increases with the number of monitoring sites. We use examples to review the modeling approach and illustrate computational complexities. We indicate directions for future work necessary for modeling massive data sets.

Design and Evaluation of Networks for Monitoring Air Quality

Douglas Nychka
North Carolina State University and the National Institute of Statistical Sciences
Nancy Saltzman
National Institute of Statistical Sciences

An important problem in monitoring environmental pollutants is how to place a network of measuring instruments to make most efficient use of resources. Based on the usual models for Gaussian random fields, choosing design points to minimize the prediction variance is difficult numerically and is not feasible for large numbers of observation points. This work presents several strategies, based on subset selection or a space-filling geometric criteria, for constructing designs that are readily computable. Given a set of candidate points it is shown that there is an equivalence between choosing an optimal subset for predicting the average and a regression subset selection problem. Besides using the exhaustive branch and bound algorithm for subset selection (in SPLUS the "leaps" function), a new method, the lasso, is also considered. The lasso is a recent development by Robert Tibshirani and is based on a regularization/Bayesian principle. Space-filling designs are found by a naive, but fast, algorithm based on swapping nearest neighbors. This startegy is very useful for adapting to the complicated design constraints imposed by geographic boundaries. The methods are illustrated by shrinking an ozone monitoring network for the Chicago urban area and designing a large network for the Midwest region of the US. Here the design performance is evaluated by nonstationary models that are estimated from observed data and numerical output (Regional Oxidant Model).

The Kriged Kalman Filter

Colin Goodall
Health Process Management, Doylestown PA and Penn State University
Kanti Mardia and Ed Redfern
University of Leeds, UK
Paco Javier Alonso
Univercidad de Granada, Spain

For time-series of scalar or multivariate observations at a set of spatial-distributed sites, cast into the framework of the general state space model, estimation may be performed using the multivariate Kalman filter. For spatial-distributed data, modeled as a (sum of) random fields, kriging provides optimal conditional estimates of the field at any interpolated spatial location. We describe the Kriged Kalman filter, by which predictions are obtained at points throughout both space and time.

The Kriged Kalman filter extends the parameter matrix of the observation equation of the multivariate Kalman filter to a vector of spatial fields, known as principal fields. A scalar observation at any site is modeled as a linear combination of the principal fields at the site — the linear combination is specified by the state vector — plus spatial-correlated error.

There is considerable art to the choice of principal fields. One approach is to utilize the first several eigenvectors of the particular partial information matrix that provides a basis for the space of all possible kriging solutions for the given sites. A modification is to use the complete basis for an optimal subset of sites.

The computational requirement for the Kriged Kalman filter is linked software for (i) the multivariate Kalman filter and (ii) for geostatistical analysis, including site selection and covariance modeling.

Invited Session: Virtual Departments

Virtual Departments: The Technology

Balasubramanian Narasimhan
Department of Statistics, Stanford University

Advances in modern technology and new media such as the WWW have made remote instruction, research and collaboration possible. Several virtual universities have sprung up with the aim of devlivering courses over the Web. Although these are first steps, developments in some key areas of video, audio aand multicasting are poised to make an impact on education.

This talk will focus on these developments, standards, tools, and illustrate how to set up some of these services. A collection of resources for further reference will also be provided.

Teaching in the Virtual Department

Jan De Leeuw
Interdivisional Program in Statistics, UCLA

We discuss principles and tools to effectively implement and organize teaching materials on the World Wide Web. Advantages and disadvantages of client-side and server-side tools are discussed. Illustrations are taken from the UCLA Statistics site at `http:\\www.stat.ucla.edu`.

What Is There In Research?

A.J. Rossini
University of South Carolina

The advent of the Internet and the World Wide Web (WWW) has shortened the distances between departments. Department homepages provides a virtual presence for the department in a manner which allows for the increased interaction of researchers as well as better access to information and data. We discuss some of the tools, data, and useful capabilities of the WWW in assisting statistical research.

Invited Session: Hardware Architectures and Networks to Support Data Mining Applications

MDAS- A Massive Data Analysis System

Arcot K. Rajasekar, Chaitan Baru, Richard Frost, Richard Marciano, Reagan Moore, and Mike Wan
San Diego Supercomputer Center, La Jolla, California

MDAS - Massive Data Analysis System - is a project for facilitating a software infrastructure to enable resource discovery and scheduling of computations in a heterogeneous, distributed system. The creation of a Massive Data Analysis System (MDAS) will enable new modes of science through improved data management of scientific data sets. This requires a semantically-enabled scalable software infrastructure that can manage petabytes of data, support rapid access of selected data sets, and provide support for subsequent computationally intensive analyses. To accomplish this, MDAS provides a metadata-enabled computational framework that can inter-operate with diverse resources, applications, methods and datasets. The system is designed with a comprehensive ontology for storing a core system-level metadata that allows it to handle a variety of entities. Moreover, it provides the means to seamlessly integrate application- and resource-specific metadata that can be effectively utilized to deal with peculiarities of individual systems. In this talk we discuss the design and implementation of MDAS.

Data Engines for Large Scale Mining

Robert Hollebeek
University of Pennsylvania

Data in the many Terabyte range and above can be mined by parallel and distributed hardware architectures. This Data Intensive Computing presents many interesting challenges in both hardware and software. The National Scalable Cluster Project (NSCP) at the University of Pennsylvania is studying Terabyte and Petabyte scale mining and designing regional infrastructure for the greater Philadelphia region as well as wide area distributed computing using the vBNS to investigate such problems.

Remarks on Data Mining at Baylor College of Medicine

Denise Moreau
Baylor College of Medicine

Contributed Session: Graphical Methods

An Interactive Graphical Tool for Exploring Sequential Dependencies in Categorical Data

Mark Fitzgerald
Los Alamos National Laboratory

In attempting to model a complicated system, one may need to understand the nature of the dependence between hundreds of different components in the system. Even searching for simple dependencies can be an overwhelming task. This report addresses the problem of searching for sequential dependencies between components when there is data available on the sequence of events occurring at the various components — i.e., the data is a time series of categorical data in which the number of categories is large. A simple contingency table approach can be used to identify these sequential dependencies (even in the presence of covariates), so the task then becomes one of displaying the huge table which results. An interactive graphical tool has been developed in XLISP-STAT. The display highlights the "strongest" dependencies and allows the user the freedom to change the level of "strength" and to focus on either the whole dependency structure or only those related to particular events of interest.

Selection Sequences - Interactive Analysis of Massive Data Sets

Martin Theus
AT&T Labs, Murray Hill, NJ
Heike Hofmann
University of Augsburg, Augsburg, Germany

One of the main tasks in analysing massive data sets is to condition views, i.e. graphs and/or statistics, of the data on subsets. One recent approach to do so are Trellis Displays, introduced by Cleveland et al. But whereas Trellis Displays are static, and show all views of all subgroups at a time, the exploratory analysis process is iterative and often focussed on particular subgroups.
Selection Sequences are an intuitive extension to the established linked-highlighting paradigm, which e.g. is implemented in DataDesk or SAS-Insight. Selection Sequences store the whole hierarchical path of a selection and allow an easy editing, redefinition and interrogation of each selection in that path. This enables the user to analyse even masive data sets without losing the aim he was heading for. Since all selection parameters are stored in a seqeuence, a generalizition to graphical data base queries is easy to achieve.
A first implementation of Selection Sequences can be found in the research software MANET.

Computational Aspects of Data Analysis Using Flowgraph Methods

Aparna V. Huzurbazar
Dept. of Statistics, University of New Mexico

An innovative approach to data analysis involves modelling based on flowgraph methods. These methods, although fairly new to statistics, are commonly used to design and analyze complex stochastic systems in the engineering sciences. We introduce them in the context of system reliability. A flowgraph is a graphical representation of a set of relations describing a stochastic system. It consists of a set of nodes connected by directed line segments (branches) that model the dependence of an output variable on an input variable. A practical application consists of using flowgraph methods to access the system moment generating function. This moment generating can be inverted using methods such as saddlepoint approximations to give Bayesian predictive densities, cumulative distribution functions as well as system reliability and hazard functions. These methods are all computationally intensive in terms of numerical methods, simulation, and symbolic algebra. We demonstrate the viability of these methods by considering systems in series, parallel, and with feedback loops. Practical analysis of a cellular telephone network is also presented.

Interactive Visual Mining of Massive Multidimensional Datasets

Ted Mihalisin
Department of Physics, Temple University
John Timlin
Mihalisin Associates, Inc.

Recent extensions to the TempleMVV multidimensional visualization system (U.S. Patent No. 5,228,119) greatly enhance its analysis and data mining capabilities. The system can now be used to visually examine datasets containing hundreds of millions of records and thousands of variables in a highly interactive fashion. One can now perform fits and view the original data, fit predictions and residuals in a space of up to 10 dimensions. Two, three or thousands of dependent variables (response variables or measures) can be plotted in a 10 dimensional manner against independent variables (explanatory variables or dimensions). Boolean operator criteria can be set that control which dependent variables or sets of dependent variables are displayed in the multidimensional space. One can visually detect highly conditional domain specific correlations and associations, multidimensional filamentary clusters and isolated cells or regions of abnormally high frequency. If suitable graphic data representations are used, human pattern recognition skills allow one to quickly discover effects that often elude algorithms. Moreover the multidimensional plotting capabilities allow one to view isolated "nuggets" of information in context so that one can ascertain not only their statistical significance but also their significance relative to other "nuggets."

Generalised Linking as a Means of Analysing Complex Data Sets

Adalbert F.X. Wilhelm
Center for Computational Statistics, George Mason University

Linking has been recognized as a key concept in interactive statistical graphics. The standard form, also known as empirical linking, uses the natural relation of various measurements that all have been made on the same individual. But linking can also be defined to build up relation between observations made on different points, between data coming from different sources, between raw data and some models. The first extension of linking that we present is cross-layer-linking, a useful tool especially for spatial data analysis. Here we build up a relation between point and area data using some neighborhood definitions as link function. For massive data sets linking observations between different aggregation levels promises good results. We give a formal definition of generalised linking and point towards some applications in KDD (knowledge discovery in databases).

Contributed Session: Density Estimation and Regression

A New Data Driven Mixture Estimator for Spatially Dependent Observations

Jeffrey L. Solka and David J. Marchette
Naval Surface Warfare Center, Dahlgren, Virginia

This paper describes a new adapative mixtures type estimator that is especially designed to process random samples obtained from an inherently spatial process. The new estimator allows for the incorporation of the spatial information into both the parameter estimation and term creation process. A description of the new algorithm along with pedagogical results on a few examples are discussed within the paper.

Local Polynomial Regression for Binary Response

Aaron Aragaki and Naomi Altman
Cornell University

Nonparametric regression methods can provide nonparametric modeling, guidance in selection of parametric models and diagnostic tools. This is particularly important for binary regression due to the lack of simple graphical tools for data exploration.

In this article, we discuss the application of local polynomial regression to the binary regression problem. We show that local polynomial regression is consistent, and is a simpler alternative to generalized smooth models. Bandwidth selection for good small sample performance remains problematic. We show by simulation that methods such as cross-validation and "plug-in" estimators, which perform well for continuous response, do poorly for binary data. However, bootstrap bandwidth selection, although very computer-intensive, appears to work well for binary response.

The Bumpy Road to the Mode Forest

Michael C. Minnotte
Utah State University
David J. Marchette
NSWC, Dahlgren, Virginia
Edward J. Wegman
George Mason University, Fairfax, Virginia

The mode tree of Minnotte and Scott (1993) provides a valuable method of investigating features such as modes and bumps in a unknown density. By examining kernel density estimates for a range of bandwidths, we can learn a lot about the structure of a data set. Unfortunately, the basic mode tree can be strongly affected by small changes in the data, and gives no way to differentiate between important modes and those caused by, for example, outliers. The mode forest overcomes these difficulties by looking simultaneously at a large collection of mode trees, all based on some variation of the original data, by means such as resampling or jittering. The result is both visually appealing and informative.

A Smoothed Bootstrap Adaptive Density Estimator via The Fast Gauss Transform

Mark Farmen
The Ohio State University

In kernel density estimation, a Smoothed Bootstrap (SB) estimate of the pointwise Mean Squared Error (MSE) can be minimized to produce a useful location dependent bandwidth. The SB estimate of MSE can be written explicitly in terms of the bandwidth, h, and a second bandwidth g associated with the degree of bootstrap smoothing. The SB MSE can be computed quickly when the normal kernel is used. However, a fast method is needed to compute the density estimator on a set of m grid points using n data points. Fast methods require $O(nlog(n) + mlog(m))$ numeric operations as opposed to $O(mn)$ operations. Traditional fast methods based on updating cannot be used with the normal kernel but an approach that is used to calculate the Fast Gauss transform can be modified to produce a fast method for computing the adaptive density estimator. The method relies on power series expansions of the kernel in a set of regions and careful summarization of the data. The method can actually compute the adaptive kernel estimator faster than the usual kernel density estimator. The method is used in one dimension but it is easily adapted to higher dimensions.. The method can supply a specified number of density derivatives along with the density.

Saturday

Saturday, 8:15 a.m. - 10:00 a.m.

Invited Session: Topics in Biomathematics and Biostatistics

Cox Models for Subsampled Data

Terry M. Therneau
Mayo Clinic

Traditionally, the Cox proportional hazards model has been most used in smaller clinical trial data sets. There is increasing interest, however, in its application to large data sets; examples include the analysis of mortgage defaults (1.5 million records, 1.6% default), survey sampling, and population based research efforts such as the Women's Health Study.

Several estimators have been proposed for these situations, particularly for the variance. We will show that most of them can be viewed in a simple way as a combination of a choice based and a finite sampling process. The estimators themselves and their variances can then easily be computed using standard statistical packages such as Splus or SAS. By using only a sample of the full data set, considerable computational efficiencies ensue.

Smoothing Spline Models for Stratified Longitudinal Data

Babette A. Brumback
Harvard University, Boston, Massachusetts
John A. Rice
University of California, Berkeley, California

Longitudinal studies generate sequences of observations measured over time. Typically, the sequences originate from different subjects and are modeled as independent realizations from related stochastic processes. In this talk, however, the focus is on sequences sampled in a stratified manner, thereby violating the usual independence assumption. The sequences are assumed to arise from an additive combination of smooth curves representing levels of the stratifying factors. This leads us to generalize smoothing spline technology for such samples, which we accomplish using the duality of smoothing splines and mixed effects models. The resulting methodology accomodates both missing data and unbalanced stratification in making estimates of smooth underlying means. It is demonstrated to be useful for data imputation and exploratory analyses of variance. Furthermore, the methodology can be implemented very efficiently with the help of eigenvector techniques, which also make practical a bootstrap procedure for assessing variability of estimated quantities. These methods were developed for analyses of biomarker data from studies of reproductive function, where sequences are stratified by menstrual cycles, subjects, and groups.

Selection of Covariates for Population Pharmacokinetic Models

William M. Sallas and Jerry R. Nedelman
Biostatistics Dept., Novartis Pharmaceuticals Corp., East Hanover, NJ

The development of population pharmacokinetic models is often reported as an art. The process of moving from a preliminary to final model is often unclear to the reader and a giant hurdle for the practitioner. Here we specify the procedure adopted for selecting covariates for the final population pharmacokinetic model in the following application. Our goal is that a clearer specification can lead to a better understanding and thus improvements in the model-building process.

In two randomized, double blind, placebo-controlled, parallel group studies for a new drug for Alzheimers disease, patients in the active drug treatment groups received SDZ ENA 713 capsules for up to 26 weeks twice daily in doses of 0.5 to 6 mg, depending on their randomized treatment group and their tolerability to the drug. Patients had at most three blood samples collected for analysis of ENA 713 and a metabolite ZNS 114-666. One sample was scheduled at weeks 2, 12, and 18 for each patient in a subset of all those under study. The pharmacokinetic data set analyzed consisted of a total of 1341 observations for ENA 713 and 1341 observations on ZNS 114-666 from 625 patients.

For ENA and ZNS separately, population pharmacokinetic models were developed 1) to estimate systemic exposure to ENA and ZNS for each patient for subsequent analyses of safety and efficacy variables and 2) to identify subpopulations whose pharmacokinetics differ. The structural model was a one-compartment model with first-order absorption and first-order elimination. The fraction bioavailable was modeled as an increasing function of dose. The statistical model incorporated fixed and random effects for the subjects. Intersubject fixed effects were incorporated by modeling the dependence of pharmacokinetic parameters on covariates. In all 37 covariate parameters (associated with demographic variables, concomitant medications, prior and concurrent diseases) for each of the four pharmacokinetic parameters needed to be considered.

Invited Session: Highlights of JCGS

Finding Modes with the Filtered Kernel Estimator

David J. Marchette
Naval Surface Warfare Center, Dahlgren, Va
Edward J. Wegman
George Mason University, Fairfax, Va

The mode tree is a useful tool for visualizing the modal structure of a density. Locations of modes of the density are plotted as a function of the bandwidth used in a kernel estimate of the density. Since the mode tree uses a single bandwidth in the kernel estimator, it exhibits all the drawbacks that a single bandwidth kernel estimator has, particularly for densities with large tails or differences in the scales of the modes. A modification is presented that uses the filtered kernel estimator, a version of the kernel estimator using a small number of bandwidths. The different bandwidths allow different amounts of smoothing in different regions, which allows the reduction of spurious modes in the tails while at the same time allowing better resolution in selected areas.

Flexible Response Surface Methods via Spatial Regression and EBLUPS

Michael A. O'connell
Becton Dickinson Research Center, RTP, NC.
Russell D. Wolfinger
SAS Institute, Cary, NC.

Spatial regression models are developed as a complementary alternative to second-order polynomial response surfaces in the context of process optimization. These models provide estimates of design variable effects and, via EBLUPS, smooth, data-faithful approximations to the unknown response function over the design space. The predicted response surfaces are driven by the covariance structures of the models. Several structures, isotropic and anisotropic, are considered and connections with thin plate splines are reviewed. Estimation of covariance parameters is achieved via maximum likelihood and residual maximum likelihood.

A feature of the spatial regression approach is the visually appealing graphical summaries that are produced. These allow rapid and intuitive identification of process windows on the design space for which the response achieves target performance. Relevant design issues are briefly discussed and spatial designs such as the packing designs available in Gosset (Hardin and Sloane, 1994), are suggested as a suitable design complement. The spatial regression models also perform well with no global design, for example with data obtained from series of designs on the same space of design variables.

The approach is illustrated with an example involving the optimization of components in a DNA amplification assay. A Monte Carlo comparison of the spatial models with both thin plate splines and second-order polynomial response surfaces for a scenario motivated by the example is also given. This shows superior performance of the spatial models to the second-order polynomials with respect to both prediction over the complete design space and for cross-validation prediction error in the region of the optimum. An anisotropic spatial regression model performs best for a high noise case and both this model and the thin plate spline for a low noise case. Spatial regression is recommended for construction of response surfaces in all process optimization applications.

Visual Explanation Tools for Graphical Models

Russell G. Almond
Educational Testing Service, Princeton, NJ
David Madigan
University of Washington, Seattle, WA

The process of risk analysis in a complex domain involves more than just building a mathematical model of a system; it involves communicating the ideas embodied in that model. Communication must exist both among engineers building the model and the modellers and the decision makers who must understand the results when articulating policy. As we discussing in our JCGS paper, graphical models (also Bayes nets and influence diagrams) provide a natural graphical explanation facility.

This talk describes how those explanation facilities can be embodied in software. In particular, it provides a guided tour of some of the graphical explanation features built into the GRAPHICAL-BELIEF prototype. An on-line tour of GRAPHICAL-BELIEF is available at:

`http:\\bayes.stat.washington.edu\almond\gb\graphical-belief.html`.

Invited Session: Smoothing

Three Sides of Smoothing: Categorical Data Smoothing, Nonparametric Regression, and Density Estimation

Jeffrey S. Simonoff
Department of Statistics and Operations Research, New York University, New York, NY

The past 40 years have seen a great deal of research into the construction and properties of nonparametric estimates of smooth functions. This research has focused primarily on two sides of the smoothing problem: nonparametric regression and density estimation. Theoretical results for these two situations are similar, and multivariate density estimation was an early justification for the Nadaraya-Watson kernel regression estimator.

A third, less well-explored, strand of applications of smoothing is to the estimation of probabilities in categorical data. In this talk the position of categorical data smoothing as a bridge between nonparametric regression and density estimation is explored. Nonparametric regression provides a paradigm for the construction of effective categorical smoothing estimates, and use of an appropriate likelihood function yields cell probability estimates with many desirable properties. Such estimates can be used to construct regression estimates when one (or more) of the categorical variables are viewed as target variables. They also lead naturally to the construction of well-behaved density estimates using local likelihood estimation. Several real data sets will be used to illustrate these points.

The Error and the Computational Complexity of a Multivariate Binned Kernel Density Estimator

Lasse Holmstrom
Rolf Nevanlinna Institute, University of Helsinki, Finland

Prebinning data is a way to reduce the computational cost of multivariate kernel density estimation. The data is discretized on a mesh and a kernel type estimator is formed on the grid points. We report a mathematically precise result on the integrated squared error of the binned estimator and discuss the computational complexity of the estimator as measured by the average number of nonzero terms that need to be summed.

Using Weighted, Smooth Quantile Estimates to Establish Medical Normal Ranges

Michael D. Lock and Douglas Haney
Becton Dickinson Immunocytometry Systems, San Jose, CA
Michael E. Tarter
University of California, Berkeley

A laboratory test result is often compared to a normal range as an aid in making a medical diagnosis. For example, an individual with a fasting glucose level that greatly exceeds the normal range of 65 to 110 mg/dL may be suspected as having diabetes. Normal ranges are usually estimated by calculating the 2.5% and 97.5% sample quantiles from a sample of healthy individuals; that is, the limits of the range are estimated using the sample cumulative distribution function (cdf). Given that the sample cdf is a step function, an obvious alternative to this method is to estimate the quantiles using a smooth estimate of the cdf. Here a Fourier series-based estimator is used in place of the sample cdf. The estimator is smoothed to optimize a weighted error criterion where the weights are chosen to increase estimator accuracy in the tails, around the quantiles of interest. Confidence intervals for the estimated quantiles are derived via resampling, and the results are compared to traditional methodology through simulation.

Contributed Session: Classification, Trees, and Optimization

Variable Selection in Regression Via Repeated Data Splitting

Peter F. Thall, Kathy E. Russell, and Richard M. Simon
M. D. Anderson Cancer Center

A new algorithm, backward elimination via repeated data splitting (BERDS) is proposed for variable selection in regression. The data are partitioned into two sets $\{E, V\}$, and an exhaus tive backward elimination (BE) is performed in E. For each p–value cutoff α used in the BE, the corresponding fitted model from E is validated in V by computing a validation sum of sq uares. This is repeated m times, and the α minimizing the sum of the m sums of squares i s used as the cutoff in a final BE on the entire data set. BERDS is a modification of the algorith m BECV proposed by Thall, Simon and Grier (1992). An extensive simulation study shows that, compar ed to BECV, BERDS has a smaller model error and higher probabilities of excluding noise variables, of selecting each of several uncorrelated true predictors, and of selecting exactly one of two or t hree highly correlated true predictors. BERDS is also superior to standard BE with cutoffs .05 or .10, and this superiority increases with the number of noise variables in the data and the degree o f correlation among true predictors. An application is provided for illustration.

Bagging in the Presence of Outliers

Peter Hall
CMA, Australian National University, Australia
Berwin A. Turlach
ACSys CRC and CMA, Australian National University, Australia

It is shown that bagging methods can improve the resistance of classification rules to outlier contamination. In settings where discrimination is difficult this effect is of second order, since in such cases the bagged and non-bagged rules are identical to first order. The outlier-reduction property does not improve performance to such an extent that outliers are no longer noticeable, and in fact their second-order presence is felt even in the limit, as sample sizes diverge. It is also shown that in the absence of contamination by outliers, and again for difficult discrimination problems, the effects of bagging are negligible to second order. Therefore, when bagging is not really needed, deleterious effects that result from employing it are quite small.

An MLE Strategy for Combining CART Models

William D. Shannon
Washington University School of Medicine, St. Louis
David Banks
Carnegie Mellon University, Pittsburgh

Because of the computationally intensive strategy for fitting CART to data, these models lose their appeal in mining very large databases. Selecting a random sample small enough to be modeled by CART has the potential risk of missing important features and/or producing a model with poor generalizability due to the inherent variability in classification trees. Fitting a tree independently to each of multiple random samples and combining the results through bagging produces significant improvement in classifying at the expense of losing the interpretability of the single tree. We propose a strategy where trees independently fit to multiple random samples are 'averaged' to give a single tree. This tree is not necessarily one of the trees fit to the samples. In this talk we will define a distribution function and distance metric on the set of classification trees and present an ascent search strategy for finding a maximum liklelihood tree. We will show some performance measures resulting from simulated data experiments, and apply our approach to a biomedical problem.

Evaluating Decision Trees Using Trade-offs

Padraic G. Neville
SAS Institute, Inc.

Classification Trees often fail to isolate rare events because the non-rare events dominate the classification rate. Common practice is to change the training data so that the proportions of the rare and non-rare events are equal, or to ignore the classification measure.

The researcher is really seeking to find a large subset containing a high proportion of rare events. A trade-off is made between the two objectives of "large" and "high".

Another common goal is to predict the rare events with some accuracy (sensitivity), and simultaneously predict the non-rare events with some other accuracy (specificity). A trade-off is made between these two objectives. Common practice using trees is to use asymmetric classification costs, or ignore the classification measure.

Optimization based on trade-off measures means finding, for each value of one objective, the best fit for the other objective. Trade-off measures thus produce a curve instead of a single number such as a classification rate. The trade-off is sometimes resolved using profit and loss formulae. This paper describes a tree pruning method for finding optimal trade-offs.

Saturday, 10:30 a.m. - 12:15 p.m.

Invited Session: Bayesian Methods

Analysis of High Frequency Data by Markov Chain Monte Carlo Methods

Siddhartha Chib
Washington University

From a Bayesian perspective, models are usually compared using Bayes factors (ratios of marginal likelihoods). In recent years, interest in computing these model evidence measures by simulation has grown and a number of methods are now available. This talk provides some unifying perspectives and introduces new ideas for computing these measures based on the approach introduced by Chib (1995). The efficacy of the different approaches is dicussed and illustrated in the context of two complex models.

Some Adaptive Monte Carlo Methods for Bayesian Inference

Luke Tierney
University of Minnesota, Minneapolis, MN

Monte Carlo methods, in particular Markov chain Monte Carlo methods, have become increasingly important as a tool for practical Bayesian inference in recent years. A wide range of algorithms is available, and choosing a particular algorithm that will work well on a particular problem is challenging. It is therefore important to explore the possibility of developing adaptive strategies that choose and adjust the algorithm to a particular problem based on information obtained during sampling as well as information provided with the problem. This talk outlines some of the issues in developing adaptive methods and presents some preliminary results.

A Process-Convolution Approach for Spatial Modeling

David M. Higdon
Institute of Statistics and Decision Sciences, Duke University

In the environmental arena, many large data bases have a spatial as well as a temporal component to them. A novel "process-convolution" approach is developed here which gives a model that can account for spatial and temproal dependence. Unlike standard tools for statistical modeling of spatially dependent data such as kriging, which rely on matrix inversions, the process-convolution model requires no inversions and hence can handle larger datasets. In addition, this approach is sufficiently flexible to allow for the possibility of the dependence structure "evolving" over space and time. This talk develops the process-convolution model and proposes an estimation scheme which uses MCMC for inference. Applications in modeling toxic waste dispersal as well and trends in the mean temperature field of the North Atlantic Ocean will be considered.

Invited Session: Neural Nets and Pattern Recognition

Polychotomous Classifiers and PACTS

Trevor. J Hastie
Statistics Department, Stanford University, California

Classification problems get harder as the number of classes increase. For example, the null error rate is $1 - 1/K$ where K is the number of classes. Classes can mask each other in structured parametric settings.

Some recent proposals can be viewed as *Plug in Classifier* techniques or PACTS. One solves a multi-class problem by solving many smaller problems, involving a much smaller number of classes. These are then combined in an appropriate way to perform the final classification.

In this talk we discuss two different PACTS:

- Dietrich and Bakhiri(1995) create many different two-class problems by grouping the classes into super-classes in many different ways. They motivate their procedure using coding theory concepts, and each binary classifier identifies a particular bit in the binary representation of the class labels. We discuss some properties of this approach and generalizations.
- Friedman(1996) suggests solving all $\binom{K}{2}$ pairwise classification problems, and then combining to them to form a K-class classifier. We describe a technique based on combining all the pairwise probability estimates. Each pairwise problem can be customized separately to take care of the particular aspects of their decision boundary.

Fusing Diverse Algorithms

John F. Elder
Quantitative Solutions, Charlottesville, Virginia

Despite the varied pedigrees of emerging tools for Data Mining, the underlying algorithms seem to fall into a handful of distinct families, whose properties provide hints as to their performance on a given dataset. This can help identify the technique likely to be best for an application. Yet, which modeling method is employed seems to have less impact on performance than the expertise level of the analyst. (This suggests that one might standardize on a sufficiently powerful inductive modeling method and concentrate resources on the data analysis tasks external to it.)

It appears however, that one can employ a simple procedure to nearly match the performance benefits of "tender loving care" in modeling: combine competing candidate models within a family. By bootstrapping the training set (e.g. *bagging*; Breiman, 1996), or varying the guidance parameters, a variety of simple models can be induced which, when bundled, seem to generalize as well as, or better than, the primary model alone. (And, as the component models have less riding on them, their construction can be more automated.)

Further, there remains reason to master a range of techniques: models from different families are more distinct, and the *fusion boost* seems to increase with model variety. The effects of combining models within and between families will be demonstrated in a couple of case studies.

Can Statistical Theory Help Us Use Neural Nets Better?

Brian D. Ripley
University of Oxford, Oxford, UK

If we view neural nets as class of statistical models with high-dimensional parameters, we can consider how to apply the ideas of statistical theory, in particular ideas for model choice and the concepts of predictive Bayesian inference. It turns out that these ideas do give considerable insight, and enable us to find more powerful solutions with reduced computational load. This will be illustrated by a couple of case studies.

Invited Session: Linking Software for Environmental and Spatial Data Sets

Linking Statistical Summaries and Maps: An Approach To Understand Massive Spatial Data Sets

Daniel B. Carr
George Mason University, Fairfax, VA
Anthony R. Olsen
US EPA, Corvallis, OR
Jean-Yves (Pip) Courbois
Oregon State University, Corvallis, OR
Suzanne Pierson
OAO Corporation, Corvallis, OR

We describe selected methods for, and challenges in, linking multivariate attribute information where geographic location is important. Map applications typically involve at least two spatial coordinates, plus one or more attributes. Linked graphics provide a class of tools helpful in understanding such higher-dimensional geographic relationships. In recent times, linking has become almost synonymous with dynamic color linking in a interactive low-resolution workstation environment. The first part of the talk, introduces color micromaps as a positional link between row-labeled plot statistical summaries of attributes and their spatial locations. Our methods give equal consideration to presenting data in attribute space and in geographic space. The example connects continental U.S. ecoregions, seasonal landuse derived from AVHRR imagery, and digital terrain data (8 million cases by over 150 attributes) to summarize their characteristics in linked-map graphics. The basic display is high-resolution and static: a large poster. The second part of the talk introduces challenges in binning and displaying an example massive data application of spectral intensities from a 54 million pixel six-band Landsat image of the Willamette Valley.

Spatial Data Analysis in the Dynamically Linked ArcView/XGobi/XploRe Environment

Juergen Symanzik
Iowa State University
Thomas Koetter, Swetlana Schmelzer, Sigbert Klinke, Dianne Cook, and Debby Swayne
Bellcore, Iowa State, Humboldt State

It was always possible to display spatial locations and concomitant geographic variables in a Geographic Information System (GIS), explore multivariate data in a Dynamic Statistical Graphics (DSG) program, and conduct spatial data analysis in a statistical software package. However, to our best knowledge, this is the first time that three such programs, the GIS ArcView 2.1, the DSG program XGobi (Swayne, Cook, and Buja, 1991), and the statistical package XploRe (Härdle, Klinke, and Turlach, 1995) have been dynamically linked into one environment to provide all facilities required for an efficient analysis of spatial data. The ArcView/XGobi/XploRe environment allows us to directly issue and use XploRe commands from within any of the other two packages and it dynamically links different types of views uniquely available in each program. This environment has been successfully used to analyze spatial data sets such as satellite imagery and precision agriculture data.

Applications of the Average Shifted Histogram in Environmental Economics

Gerald Whittaker
Economic Research Service/USDA, Washington, DC
David W. Scott
Rice University, Houston, Texas

The Average Shifted Histogram (ASH) provides a very fast, robust method of estimating multivariate nonparametric regressions. The ASH algorithm is fast enough that large (500,000 observations), georeferenced economic and physical data sets can be interactively analyzed. Applications include linking economic models to physical environmental models, effects of alternative policies on water quality and effects of government policy on agricultural land values. The results of ASH estimation are conveniently visualized with three dimensional slices from $\Re^n$ using a geographic information system (GIS).

Contributed Session: Business Applications and Experimentation

Detecting Outliers in Statistical Data using Cook's D Statistic

Donald R. Jensen
Virginia Polytechnic Institute, Blacksburg, Virginia
Donald E. Ramirez
University of Virginia, Charlottesville, Virginia

We introduce a generalized F distribution, and we show how to numerically compute p-values for this family of distributions. Natural versions of Cook's D statistic are distributed as such distributions. We apply these techniques to determine the statistical outliers (for single observations and joint observations) for the Intercountry Life-Cycle Savings Data from Belsley, Kuh, and Welsh (1980).

Upper Confidence Bounds for Soft Error Rates in Integrated Circuits

José G. Ramirez
Digital Semiconductors, Hudson, MA
Isabel S. Llatas
Universidad Simón Bolívar
Rice University, Houston. TX.

Soft errors in integrated circuits are spontaneous errors, or changes in stored information that are not caused by permanent damage to the circuits. It is known that energetiiiica nuclear particles, such as alpha particles, can produce soft errors. In order to qualify a IC for manufacturing it is necessary that it meets certain requirements among them soft error rate (SER) specifications. In this paper we discuss several methods to compute an upper bound of a given confidence for SER. These methods are based on the fact that the distribution of SER can be approximated by the product of two Poisson distributions

An Empirical Study on the Investment Risk of Taiwanese Stock Market

Ken Hung
Western Washington University
David Cheng
Nat'l Dong Hwa U

Data are drawn from AREMOS Stock Market Statistical Data Bank maintained by Ministry of Education, Taiwan. Daily returns of 151 companies in 20 industries from January 1990 to September 1995 are converted into 69 monthly returns for each company. The investment risk of the Taiwanese stock market is decomposed into the systematic risk and the unsystematic risk. It is found that the unsystematic risk of this market is only 41.89% of the entire investment risk which is low compared to 73% of USA market. One can conclude that the Taiwanese stock market is subject to a higher level of systematic market risk as most stock markets of developed countries are compared.

Efficient Testing of Electronic Devices

Hans Engler
Georgetown University, Washington, D.C.
Michael Souders and Gerard N. Stenbakken
NIST, Gaithersburg, Maryland

Testing electronic devices in order to assure the quality of individual products can be time-consuming and expensive. To speed up this process without compromising on reliability, testing strategies have been developed at NIST. They are based on the extraction of low-dimensional linear error models, using a priori knowledge of the structure of the device or data from an exhaustively tested training and validation set. Production testing then is performed only on a small subset of all possible test points which must be chosen carefully. We shall discuss statistical and computational aspects of these procedures and report on the theoretical background, heuristics, algorithms, and practical experiences.

Robust Design of Experiments and Information Set of Experimenting

Anatoly A. Naumov
Novosibirsk State Technical University,Russia

The problems of software elaboration, concerning experimenting strategies synthesis and data processing in a real-time systems are considered. The elaborated method is a variation of self organization method, combining with e-invariant imbedding method. It is also considered one of possible approaches of regularization of incorrectly - raised active identification problem. This one is based on a self-organized algorithm. It means cyclical mathematical model synthesis experimenting plan sequential generation and evolutionary selection of optimal plans points. The investigations , carried by us , had testified that system active identification problem in general and experimenting strategy synthesis problem in particular is incorrectly formulated ones , and the using of Kiefer-Wolfowitz-Vedorov statements analogous to pointed above in practice leads to falling down of efficiency of experimental plans. It may be said ,that the active identification problem in a classic form is a twice incorrectly raised one. If the first level of incorrectness is associated with unsteadiness of model parameters evaluation problem, based on least-square method , then the second one is a consequence of basis spatial vector of mathematical model being a argument of experimenting strategy optimality criteria functional.

Contributed Session: Biometric and Survival Methods

O'Brien's Extended t-test for Paired Observations

Peter C. Wollan
Biostatistics Section, Mayo Clinic, Rochester MN

O'Brien (1988) proposed a two-sample comparison based on quadratic regression of group membership (as Y) vs. the response variable (as X). The advantage of the method is that it is sensitive to differences in either location or scale. The procedure can be modified to allow for correlation between paired observations, using GEE (binomial family). Using simulation, we show that the paired version is dramatically more powerful than the corresponding test which ignores the pairing, but is less sensitive to pure loction effects than the paired t-test. We present an application to a study of diurnal cycles in certain hormones.

Parameter Estimation in Compartmental Models Using the Laplace Transform with Applications to Pharmacokinetics

Martha Contreras and Changmei Liu
Biometrics Unit, Cornell University, Ithaca, NY

We present methodology and discuss the implementation of integral transform methods when estimating parameters from an implicitly defined objective function. In particular, this talk will focus on the use of the Laplace transform when estimating the flow rates, in a Compartmental model system, from observations gathered at discrete times. To this, we present a feasible algorithm and analyze its properties when compared to the standard least squares estimator. We conclude with a discussion involving the use of Compartmental models in the analysis of Pharmacokinetic data.

Air Force Aging Aircraft Research Programs

Peter W. Hovey
Wright Laboratories, US Air Force

The current emphasis on reducing the Federal budget has led all Federal agencies to look for ways to reduce spending. In the United States Air Force, one method of economizing is to delay buying new aircraft while extending the useful life of existing aircraft. Developing a plan for using an aircraft longer than the original design life poses a different set of problems that cannot be readily solved using standard deterministic engineering tools. (New aircraft are relatively damage free while aging aircraft have growing fatigue cracks and corrosion.) The Air Force, and other Federal agencies, are looking for probabilistic tools to help evaluate life extension programs for aircraft. This paper describes some of the unique engineering problems associated with aging aircraft, some of the data collection efforts that have been implemented to track damage accumulation in aging aircraft, and some of the probabilistic tools that have been developed to help solve these problems using the available data.

Estimation of Multistate Survival Model

Rafiq I. Chowdhury, M. A. Shah, and Syed A. Mohsin
Kuwait University, Kuwait
M. Ataharul Islam
Dhaka University, Dhaka, Bangladesh

The longitudinal data analysis generally involves the problems of censoring and repeated observations. The proportional hazards model (Cox, 1972) deals with the modeling of longitudinal data for partially censored data for a single time observation for each individual. Kay (1982) showed an extension of the proportional hazards model for a number of transient states. Islam and Singh (1992) showed the multistate generalization of proportional hazards model. Islam and Singh (1992) did not provide any application in their paper. In this paper, an outline for solving equations for estimating parameters of the multistate survival model (Islam and Singh, 1992) with an application to a diabetes data set from a hospital in Bangladesh is discussed. The multistate survival model for partially censored data involves a number of inter communicating and absorbing states and a large number of parameters to be estimated on the basis of interactive solution of simultaneous equations. For estimating parameters, the Interactive Scientific Processor (ISP), a statistical software package, has been used. The multistate survival model estimated in this paper can be extensively used for solving real life problems.

Martingale Inference of Multi-state Progressive Stochastic Processes

John J. Hsieh
University of Toronto, Toronto, Canada

This article presents estimation and hypothesis testing of transition rates for multi-state stochastic processes using martingale methods. The methods are developed for any discrete-state and continuous-time process consisting of a finite number of progressive transient states and a finite number of absorbing states. The procedure can incorporate censoring, competing risks and time-varying covariates. First, various counting processes allowing for competing risks and covariates are formulated in terms of the state variables for transitions into each state. Second, intensity processes are derived as compensators of the corresponding counting processes by conditioning on appropriate filtrations. Third, useful martingales are formed with variances derived from the martingale transform theorem and asymptotic distributions from martingale central limit theorem. Fourth, martingale estimators of transition rates and their standard errors as well as various test statistics for the transition rates are derived as direct consequences of the above. The results so obtained are used for analyzing competing- risk life tables, multi-state life tables, staging of diseases, epidemic models as well as other types of survival studies.

Index