Fast and Accurate Inference for the Smoothing Parameter in Semiparametric Models

Alex Trindade

Dept. of Mathematics & Statistics, *Texas Tech University*

Joint work with **Rob Paige**, *Missouri University of Science and Technology*

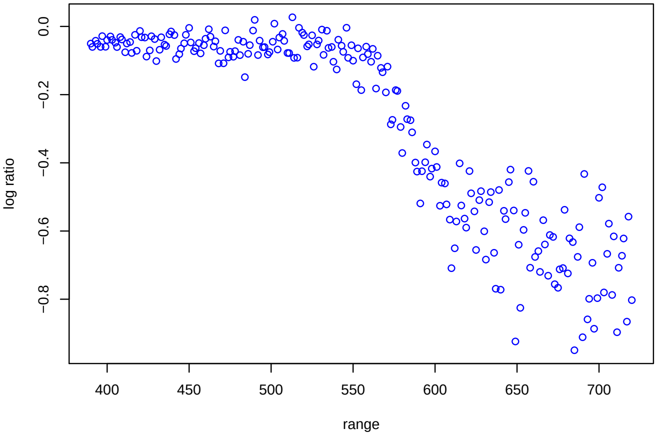
Funded in part by the *National Security Agency*
Grants: H98230-09-1-0071 (Paige) & H98230-08-1-0071 (Trindade)

May 2011

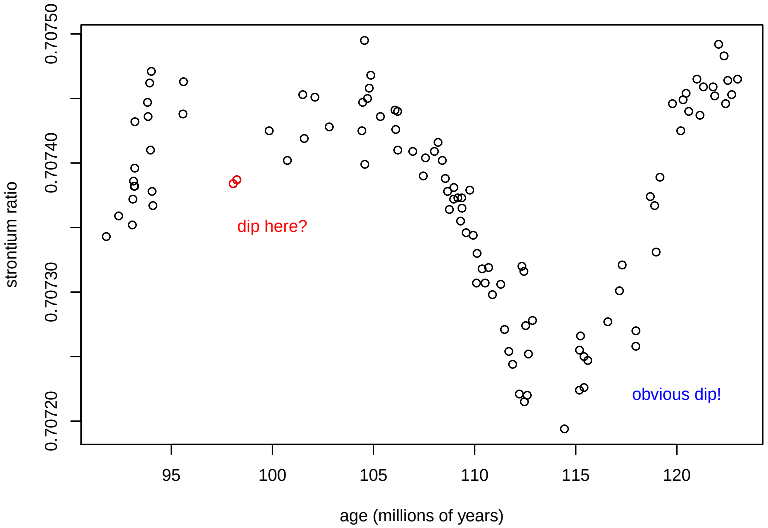
Outline

1. Motivation
   - The need for a semiparametric smoothing approach...
   - Penalized spline models
   - Penalized splines as linear mixed models (LMMs)
2. Main Result: Inference on Smoothing Parameter
   - Estimators as roots of quadratic estimating equations (QEEs)
   - Saddlepoint-based bootstrap (SPBB) inference for QEEs
   - Exact ML & REML inference in LMMs
3. Simulations: Confidence Intervals
   - Coverages, lengths, compute times
4. Application: The Fossil Data

- **Model:** $y = \mu(x) + \text{error}$.
- **Goal:** estimate mean function $\mu(x)$, i.e. smooth data.



The Fossil Data (Ruppert, Wand, & Carroll, 2003)



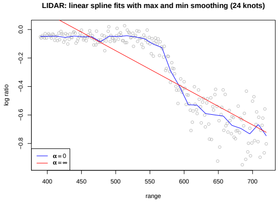
Penalized Spline Model (degree p, with K knots)

$$\mu(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} u_k (x - \kappa_k)_+^p$$

- For n obs (x_i, y_i), write in matrix form: $\boldsymbol{\mu} = X\boldsymbol{\beta} + Z\mathbf{u} \equiv B\boldsymbol{\theta}$.
- Model can allow for autocorrelation, R, in residuals (e.g. time series).
- Estimate $\boldsymbol{\theta}$ by minimizing

$$\hat{\boldsymbol{\theta}}_{\text{PS}} = \arg\min_{\boldsymbol{\theta}} \left\{ (\mathbf{y} - B\boldsymbol{\theta})' R^{-1} (\mathbf{y} - B\boldsymbol{\theta}) + \alpha \mathbf{u}'\mathbf{u} \right\}$$

- α is a **smoothing parameter** controlling balance between:
  - fidelity to data $(\alpha = 0)$
  - smoothness of fit $(\alpha = \infty)$



Linear Mixed Model (LMM) Formulation & BLUP's

Penalized spline can be recast as LMM with one variance component (Brumback, Ruppert, & Wand, 1999)

$$\mathbf{y} = \underbrace{X\boldsymbol{\beta}}_{\text{fixed effects}} + \underbrace{Z\mathbf{u} + \boldsymbol{\varepsilon}}_{\text{random effects}}$$

- BLUP of $\mathbf{y}$ in this context is $\tilde{\mathbf{y}} = B\tilde{\boldsymbol{\theta}}$, where

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ (\mathbf{y} - B\boldsymbol{\theta})' R^{-1} (\mathbf{y} - B\boldsymbol{\theta}) + \frac{\sigma_\varepsilon^2}{\sigma_u^2} \mathbf{u}'\mathbf{u} \right\}.$$

- Implies BLUP-optimal value for α is:

$$\alpha = \sigma_\varepsilon^2 / \sigma_u^2$$

Estimation of Smoothing Parameter

- Since α is ratio of variance components in LMM, many parametric methods available.
- Also have several nonparametric methods.

Examples (Parametric)

- Maximum Likelihood (ML)
- **REstricted Maximum Likelihood (REML)**

Examples (Nonparametric)

- Akaike's Information Criterion (AIC)
- **Generalized Cross-Validation (GCV)**

A Unified View of Smoothing Parameter Estimators (New)

- Above estimators can be viewed as roots of a quadratic estimating equation (QEE) in normal random variables

$$\mathcal{Q}(\alpha) = \mathbf{y}' A_\alpha \mathbf{y}$$

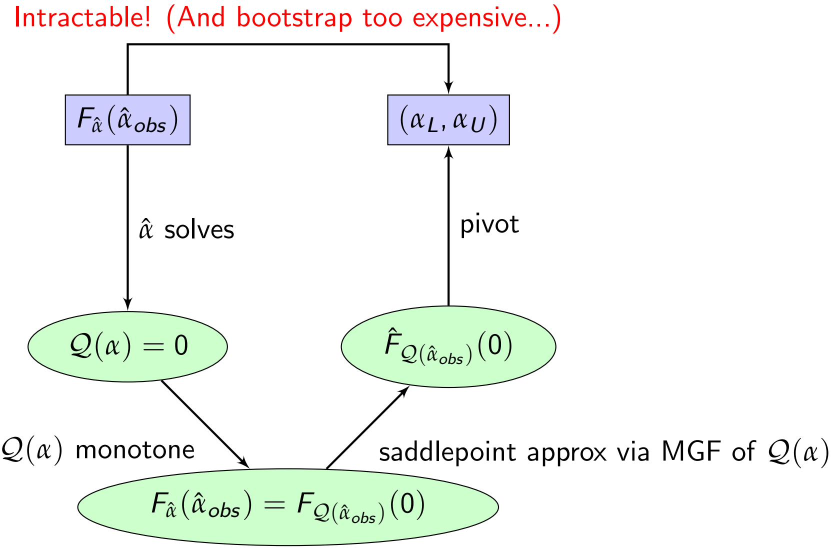
- The $n \times n$ matrix A_α has a (complicated, but) closed form expression in each case...
- Theorem (Paige & Trindade, 2010): REML QEE is unbiased.
- Krivobokova & Kauermann (2007): REML less sensitive to misspecification of residual correlation than AIC or GCV.

Saddlepoint-Based Bootstrap (SPBB) Inference for QEEs

Pioneered by Paige, Trindade, & Fernando (2009):

- Relate distribution of root of QEE to that of estimator.
- Under normality have closed form for MGF of QEE.
- Use to saddlepoint approximate distribution of estimator.
- Now invert distribution to get CI... numerically!
- Leads to 2nd order accurate CIs: coverage is $O(n^{-1})$.
- Works for: ML, REML, AIC, GCV, etc.!

SPBB: An Approximate Parametric Bootstrap



Exact ML & REML Inference for α

Exact finite sample inference for $\alpha = \sigma_\varepsilon^2/\sigma_u^2$ in LMMs with one variance component (Crainiceanu, Ruppert, Claeskens, & Wand, 2005):

- **Note:** asymptotic χ^2 dist is poor approx in finite samples due to substantial point mass at 0 (Crainiceanu & Ruppert, 2004).
- Invert (restricted) likelihood ratio test.
- Grid search needed to locate endpoints of CI (α_L, α_U).
- Only works for ML & REML...

Simulations: Mimic Extensive Study of Lee (2003)

- Simulate datasets of sample size $n = 200$ from curves

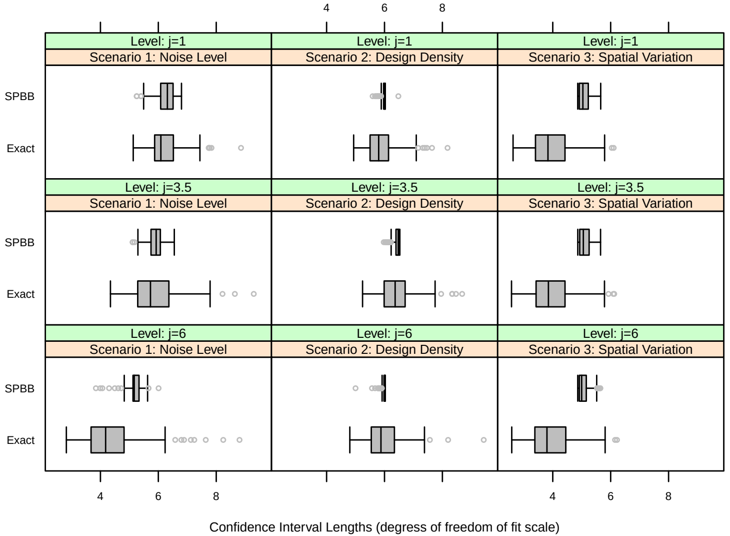
$$y = f(x) + \varepsilon, \qquad \varepsilon \sim \text{IID N}(0, \sigma_\varepsilon^2)$$

- Vary 3 factors:
  - **noise level** (σ_ε^2);
  - **design density** (number of x's);
  - **spatial variation** (type of curve).
- Each factor at 3 levels ($j = 1, 3.5, 6$).
- Each scenario (factor-level combo) replicated 200 times.
- REML-Fit linear penalized spline: O-spline basis with 35 knots placed at empirical quantiles of $x \in (0, 1)$ (Wand & Ormerod, 2008).

Results: Empirical Coverage of Nominal 95% CIs

		Empirical Probabilities (**Exact**, SPBB)		
Scenario	Level	Underage	Coverage	Overage
Noise Level	$j = 1$	**0.065** 0.055	**0.915** 0.925	**0.020** 0.020
	$j = 3.5$	**0.035** 0.025	**0.950** 0.945	**0.015** 0.030
	$j = 6$	**0.000** 0.000	**0.987** 0.970	**0.013** 0.030
Design Density	$j = 1$	**0.040** 0.040	**0.945** 0.935	**0.015** 0.025
	$j = 3.5$	**0.045** 0.035	**0.925** 0.920	**0.030** 0.045
	$j = 6$	**0.040** 0.040	**0.945** 0.945	**0.015** 0.015
Spatial Variation	$j = 1$	**0.000** 0.000	**0.934** 0.970	**0.066** 0.030
	$j = 3.5$	**0.000** 0.000	**0.928** 0.965	**0.072** 0.035
	$j = 6$	**0.000** 0.000	**0.883** 0.960	**0.117** 0.040

CI Lengths: Trellis Boxplots of SPBB vs. Exact



Comparison: Exact, SPBB, and Bootstrap CIs

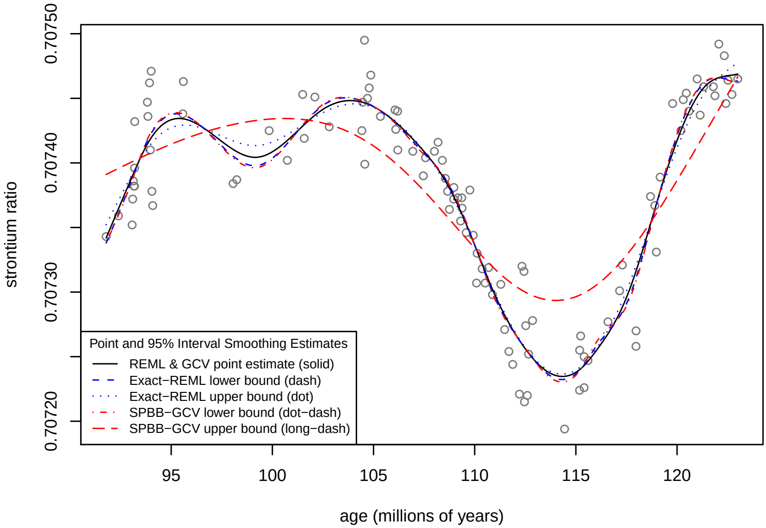
For the 200 simulated datasets with Noise Level factor at level $j = 1$

Method and (minutes/CI)	Coverage Probability	Interval Length Statistics Min	Q_1	Median	Q_3	Max
SPBB-REML (15)	0.925	5.25	6.09	6.31	6.51	6.80
Exact-REML (105)	0.915	5.13	5.87	6.09	6.52	8.86
Bootstrap (2,100)	1.000	8.84	13.18	15.48	18.13	28.57

The Smoothed Fossil Data

- Chaudhuri & Marron (1999): SiZer method to assess significance of small dip around 100 MY ago (NOT sig. at 95% level).
- Ruppert *et al.* (2003): fit penalized spline models with truncated polynomial bases with a variety of knots, degrees, and amounts of smoothing.
- Wand & Ormerod (2008): showcase "natural boundary" properties of O-splines; use judiciously chosen set of 20 interior knots.
- Our analysis: fit O-spline of Wand & Ormerod (2008); get 95% Exact-REML, SPBB-REML, and SPBB-GCV CIs.

Application: The Fossil Data



Summary of SPBB Inference

- Can be used under a variety of different criteria: ML, REML, GCV, and AIC.
- Performance: nearly exact.
- Computing:
  - 1 order of magnitude faster than exact;
  - 2 orders of magnitude faster than bootstrap.
- Only computationally feasible alternative when no known exact or asymptotic methods exist, e.g. GCV and AIC.
- Smoothing parameter is tuning parameter; but can be used to uncover features in data...

Key References

- Chaudhuri, P. & Marron J.S. (1999). SiZer for exploration of structures in curves. *J. Amer. Statist. Assoc.* 94, 807-823.
- Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, M. (2005), "Exact likelihood ratio tests for penalized splines", *Biometrika*, 92, 91-103.
- Krivobokova, T., & Kauermann, G. (2007), "A Note on Penalized Spline Smoothing with Correlated Errors", *J. Amer. Statist. Assoc.*, 102, 1328-1337.
- Lee, T.C.M. (2003). Smoothing parameter selection for smoothing splines: a simulation study. *Comp. Statist. Data Anal.* 42, 139-148.
- Paige, R.L., Trindade, A.A. and Fernando, P.H. (2009), "Saddlepoint-based bootstrap inference for quadratic estimating equations", *Scand. J. Stat.*, 36, 98-111.
- Paige, R.L., & Trindade, A.A., "Fast and Accurate Inference for the Smoothing Parameter in Semiparametric Models", *Aust. & New Zeal. J. Stat.*, (to appear).
- Ruppert, D., Wand, M.P., & Carroll, R.J. (2003), *Semiparametric Regression*. London: Cambridge.