Slide 1

Semiparametric Modeling, Penalized Splines, and Mixed Models

David Ruppert
Cornell University

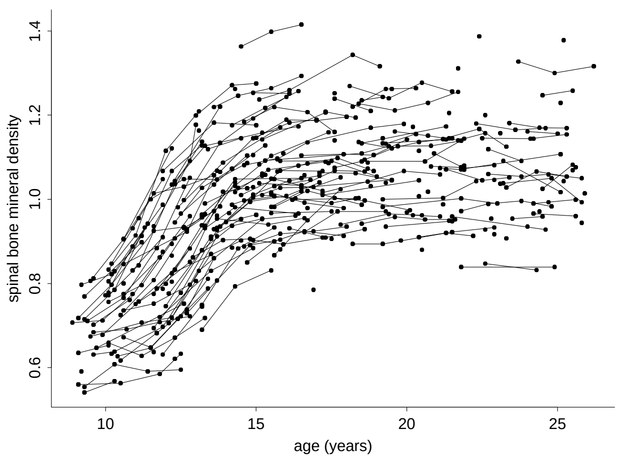
http://www.orie.cornell.edu/~davidr

January 2004

Joint work with Babette Brumback, Ray Carroll, Brent Coull, Ciprian Crainiceanu, Matt Wand, Yan Yu, and others

Slide 2

Example (data from Hastie and James, this analysis in RWC)



Slide 3

Possible Model

$\texttt{SBMD}_{i,j}$ is spinal bone mineral density on ith subject at age equal to $\texttt{age}_{i,j}$.

$$\texttt{SBMD}_{i,j} = U_i + m(\texttt{age}_{i,j}) + \epsilon_{i,j},$$

$$i = 1, \ldots, m = 230, \quad j = i, \ldots, n_i.$$

U_i is the random intercept for subject i.

$\{U_i\}$ are assumed i.i.d. $N(0, \sigma_U^2)$.

Slide 4

Underlying philosophy

1. minimalist statistics
   - keep it as simple as possible
2. build on classical parametric statistics
3. modular methodology

Slide 5

Reference

Semiparametric Regression by Ruppert, Wand, and Carroll (2003)

- Lots of examples from biostatistics.

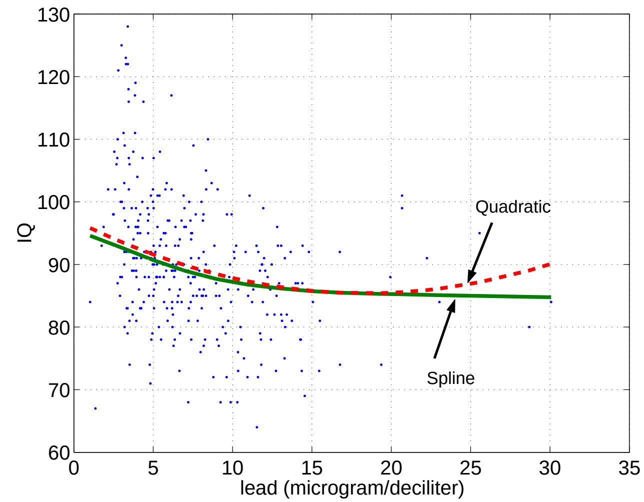
Slide 6

Recent Example — April 17, 2003

Canfield et al. (2003) — Intellectual impairment and blood lead.

- longitudinal (mixed model)
- nine covariates (modelled linearly)
- effect of lead modelled as a spline (semiparametric model)
  - disturbing conclusion

Slide 7



Thanks to Rich Canfield for data and estimates.

Slide 8

Semiparametric regression

Partial linear or partial spline model:

$$Y_i = \mathbf{W}_i^{\mathsf{T}} \boldsymbol{\beta}_W + m(X_i) + \epsilon_i.$$

$$m(x) = \mathbf{X}_i^{\mathsf{T}} \boldsymbol{\beta}_X + \mathbf{B}^{\mathsf{T}}(x)\mathbf{b}.$$

$$\mathbf{B}^{\mathsf{T}}(x) = \left(B_1(x) \quad \cdots \quad B_K(x) \right).$$

E.g.,

$$\mathbf{X}_i^{\mathsf{T}} = \left(X_i \quad \cdots \quad X_i^p \right)$$

$$\mathbf{B}^{\mathsf{T}}(x) = \left\{ (x-\kappa_1)_+^p \quad \cdots \quad (x-\kappa_K)_+^p \right\}$$

Slide 9

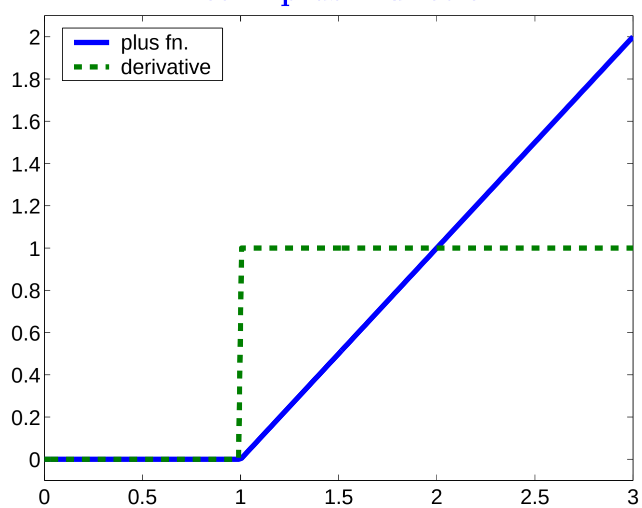
Example

$$m(x) = \beta_0 + \beta_1 x + b_1(x-\kappa_1)_+ + \cdots + b_K(x-\kappa_K)_+$$

- slope jumps by b_k at κ_k

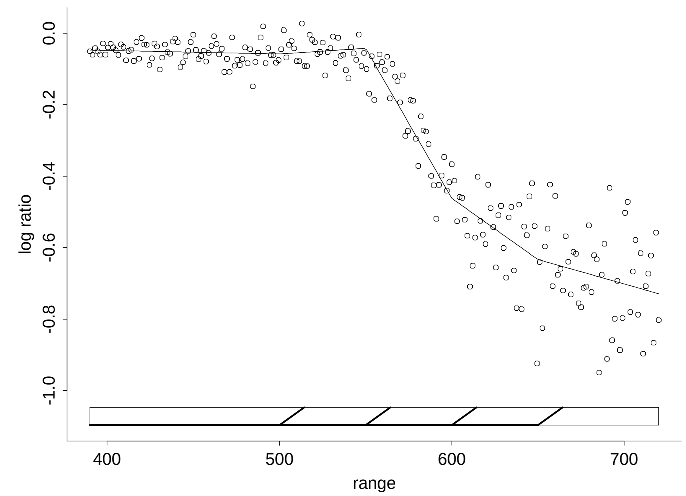
ide 10

Linear “plus” function



Slide 11

Fitting LIDAR data with plus functions



Slide 12

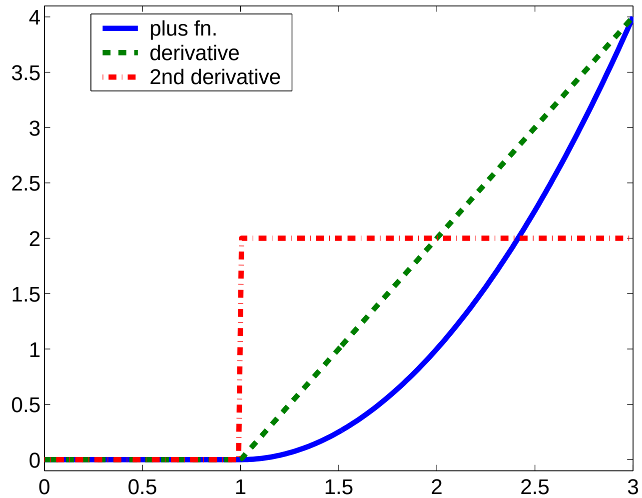
Generalization

$$m(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + b_1(x-\kappa_1)_+^p + \cdots + b_K(x-\kappa_K)_+^p$$

- pth derivative jumps by $p!\, b_k$ at κ_k
- first $p-1$ derivatives are continuous

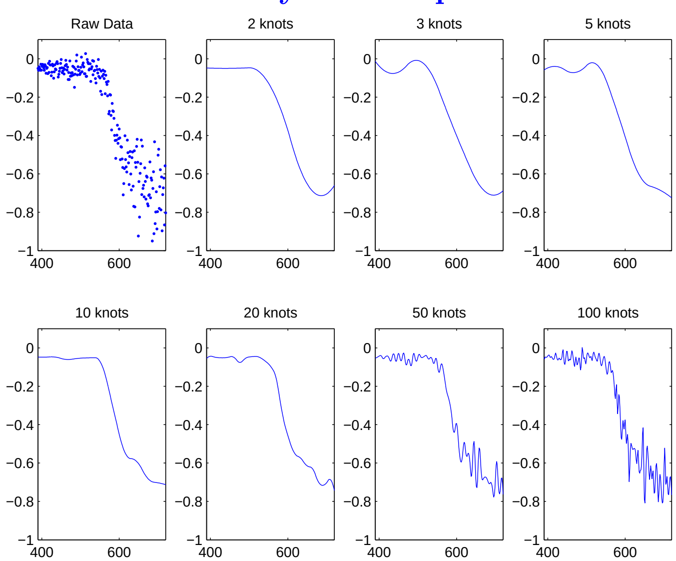
ide 13

Quadratic “plus” function



ide 14

Ordinary Least Squares



Slide 15

Penalized least-squares

Minimize

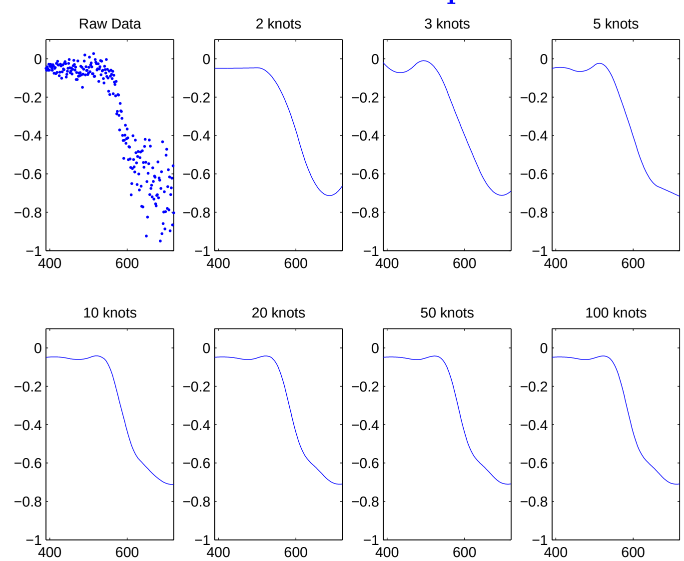
$$\sum_{i=1}^{n}\left\{Y-(\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\beta}_W+\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\beta}_X+\mathbf{B}^{\mathsf{T}}(X_i)\mathbf{b})\right\}^2+\lambda\,\mathbf{b}^{\mathsf{T}}\mathbf{D}\mathbf{b}.$$

E.g.,

$$\mathbf{D}=\mathbf{I}.$$

Slide 16

Penalized Least Squares



ide 17

Ridge Regression

From previous slide:

$$\sum_{i=1}^{n}\left\{Y-(\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\beta}_W+\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\beta}_X+\mathbf{B}^{\mathsf{T}}(X_i)\mathbf{b})\right\}^2+\lambda\,\mathbf{b}^{\mathsf{T}}\mathbf{D}\mathbf{b}.$$

Let $\mathcal{X}$ have row $\begin{pmatrix}\mathbf{W}_i^{\mathsf{T}} & \mathbf{X}_i^{\mathsf{T}} & \mathbf{B}^{\mathsf{T}}(X_i)\end{pmatrix}$. Then

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}_W\\ \widehat{\boldsymbol{\beta}}_X\\ \widehat{\mathbf{b}}\end{pmatrix}=\left\{\mathcal{X}^{\mathsf{T}}\mathcal{X}+\lambda\ \mathrm{blockdiag}(\mathbf{0},\mathbf{0},\mathbf{D})\right\}^{-1}\mathcal{X}^{\mathsf{T}}\mathbf{Y}.$$

- Also, a BLUP in a mixed model and an empirical Bayes estimator.

Slide 19

From previous slides:

Let $\mathcal{X}$ have row $\begin{pmatrix}\mathbf{W}_i^{\mathsf{T}} & \mathbf{X}_i^{\mathsf{T}} & \mathbf{B}^{\mathsf{T}}(X_i)\end{pmatrix}$. Then

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}_W\\ \widehat{\boldsymbol{\beta}}_X\\ \widehat{\mathbf{b}}\end{pmatrix}=\left\{\mathcal{X}^{\mathsf{T}}\mathcal{X}+\lambda\ \mathrm{blockdiag}(\mathbf{0},\mathbf{0},\mathbf{D})\right\}^{-1}\mathcal{X}^{\mathsf{T}}\mathbf{Y}.$$

Linear mixed model:

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\mathbf{b}}\end{pmatrix}=\begin{pmatrix}\mathbf{X}^{\mathsf{T}}\mathbf{X} & \mathbf{X}^{\mathsf{T}}\mathbf{Z}\\ \mathbf{Z}^{\mathsf{T}}\mathbf{X} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z}+\lambda\boldsymbol{\Sigma}_b^{-1}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{X}^{\mathsf{T}}\mathbf{Y}\\ \mathbf{Z}^{\mathsf{T}}\mathbf{Y}\end{pmatrix}$$

$$=\left\{\begin{pmatrix}\mathbf{X} & \mathbf{Z}\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}\mathbf{X} & \mathbf{Z}\end{pmatrix}+\lambda\,\mathrm{blockdiag}(\mathbf{0},\boldsymbol{\Sigma}_b^{-1})\right\}^{-1}\begin{pmatrix}\mathbf{X} & \mathbf{Z}\end{pmatrix}^{\mathsf{T}}\mathbf{Y}$$

ide 18

Linear Mixed Models

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{b}+\boldsymbol{\varepsilon}$$

where $\mathbf{b}$ is $N(0,\sigma_b^2\boldsymbol{\Sigma}_b)$.

$\mathbf{X}\boldsymbol{\beta}$ are the "fixed effects" and $\mathbf{Z}\mathbf{b}$ are the "random effects."

Henderson's equations.

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\mathbf{b}}\end{pmatrix}=\begin{pmatrix}\mathbf{X}^{\mathsf{T}}\mathbf{X} & \mathbf{X}^{\mathsf{T}}\mathbf{Z}\\ \mathbf{Z}^{\mathsf{T}}\mathbf{X} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z}+\lambda\boldsymbol{\Sigma}_b^{-1}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{X}^{\mathsf{T}}\mathbf{Y}\\ \mathbf{Z}^{\mathsf{T}}\mathbf{Y}\end{pmatrix}.$$

$$\lambda=\frac{\sigma_\epsilon^2}{\sigma_b^2}.$$

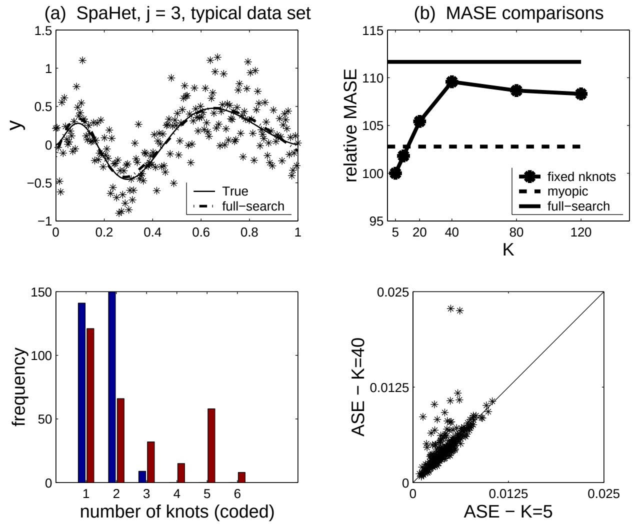
Slide 20

Selecting λ

1. cross-validation (CV)
2. generalized cross-validation (GCV)
3. ML or REML in mixed model framework

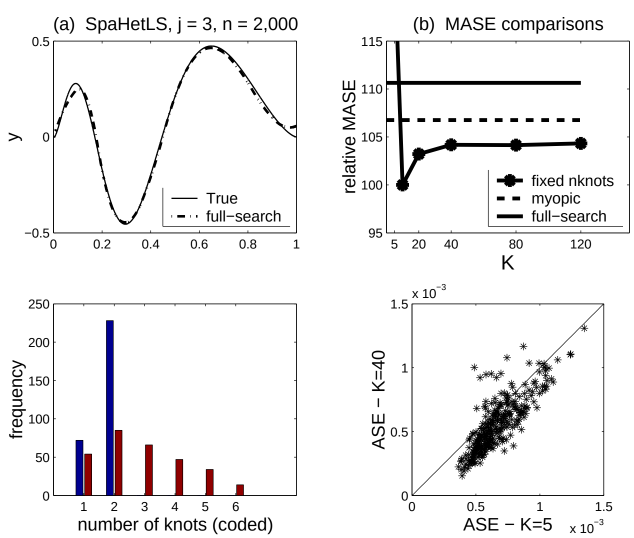
ide 21

Selecting the Number of Knots



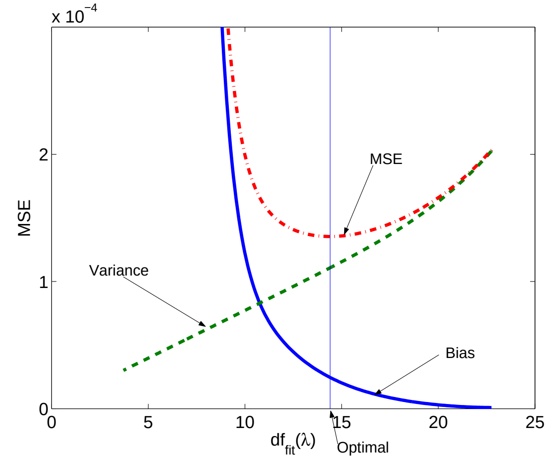
$n = 200$

ide 22



$n = 2,000$

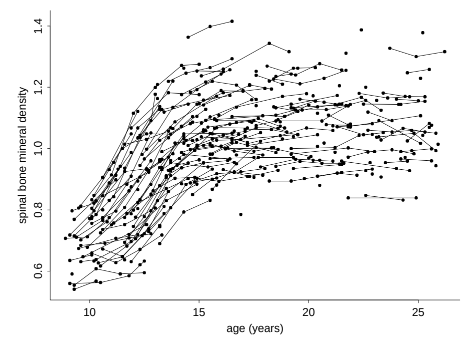
Slide 23



$n = 10,000$, 20 knots, quadratic spline

Slide 24

Return to spinal bone mineral density study



$$\mathtt{SBMD}_{i,j} = U_i + m(\mathtt{age}_{i,j}) + \epsilon_{i,j},$$

$$i = 1, \ldots, m = 230, \quad j = i, \ldots, n_i.$$

ide 25

$$\mathbf{X} = \begin{bmatrix} 1 & \mathtt{age}_{11} \\ \vdots & \vdots \\ 1 & \mathtt{age}_{1n_1} \\ \vdots & \vdots \\ 1 & \mathtt{age}_{m1} \\ \vdots & \vdots \\ 1 & \mathtt{age}_{mn_m} \end{bmatrix}$$

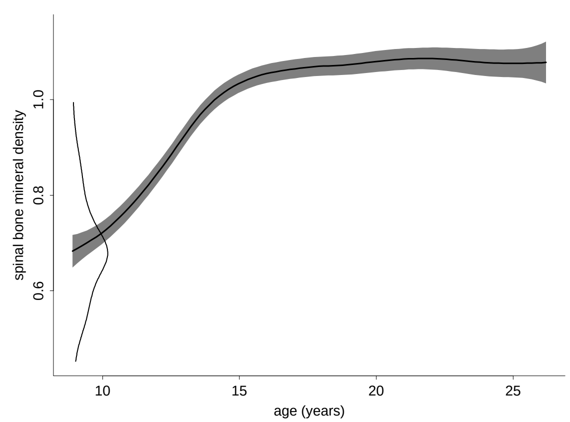
Slide 27

$$\mathbf{u} = \begin{bmatrix} U_1 \\ \vdots \\ U_m \\ b_1 \\ \vdots \\ b_K \end{bmatrix}$$

ide 26

$$\mathbf{Z} = \begin{bmatrix} 1 & \cdots & 0 & (\mathtt{age}_{11} - \kappa_1)_+ & \cdots & (\mathtt{age}_{11} - \kappa_K)_+ \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 0 & (\mathtt{age}_{1n_1} - \kappa_1)_+ & \cdots & (\mathtt{age}_{1n_1} - \kappa_K)_+ \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & (\mathtt{age}_{m1} - \kappa_1)_+ & \cdots & (\mathtt{age}_{m1} - \kappa_K)_+ \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & (\mathtt{age}_{mn_m} - \kappa_1)_+ & \cdots & (\mathtt{age}_{mn_m} - \kappa_K)_+ \end{bmatrix}$$

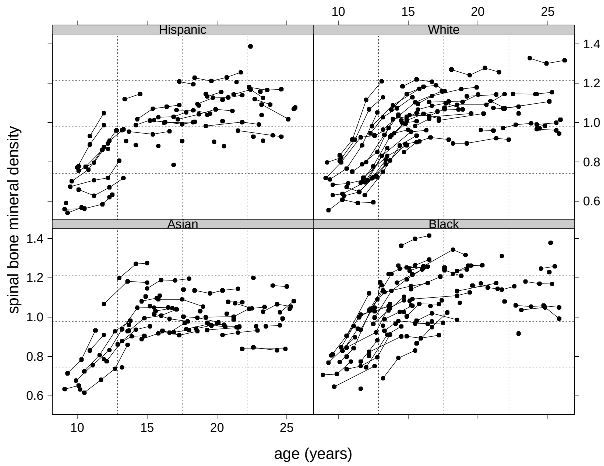
Slide 28



Variability bars on $\widehat{m}$ and estimated density of U_i

ide 29

Broken down by ethnicity



ide 30

Model with ethnicity effects

$$
\begin{aligned}
\texttt{SBMD}_{ij} \;&=\; U_i + m(\texttt{age}_{ij}) + \beta_1 \texttt{black}_i + \beta_2 \texttt{hispanic}_i \\
&\quad + \beta_3 \texttt{white}_i + \varepsilon_{ij}, \quad 1 \le j \le n_i, \quad 1 \le i \le m.
\end{aligned}
$$

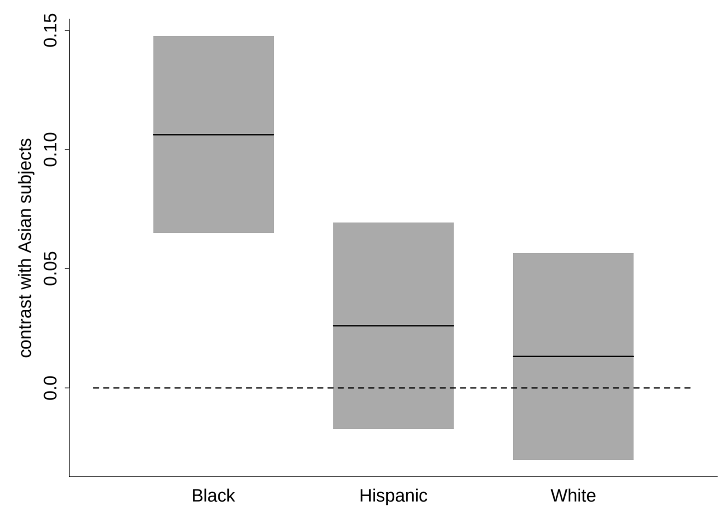
Asian is the reference group.

Slide 31

Only requires an expansion of the fixed effects by adding the columns

$$
\begin{bmatrix}
\texttt{black}_1 & \texttt{hispanic}_1 & \texttt{white}_1 \\
\vdots & \vdots & \vdots \\
\texttt{black}_1 & \texttt{hispanic}_1 & \texttt{white}_1 \\
\vdots & \vdots & \vdots \\
\texttt{black}_m & \texttt{hispanic}_m & \texttt{white}_m \\
\vdots & \vdots & \vdots \\
\texttt{black}_m & \texttt{hispanic}_m & \texttt{white}_m
\end{bmatrix}
$$

Slide 32



ide 33

- In this model, the age effects curve for the four ethnic groups are **parallel**.
- Could we model them as non-parallel?
- Might be problematic in this example because of the small values of the n_i.
- But the methodology should be useful in other contexts.

Slide 35

Penalized Splines and Additive Models

Additive model:

$$Y_i = m_1(X_{1,i}) + \ldots + m_P(X_{P,i}) + \epsilon_i$$

ide 34

- Add interactions between `age` and `black`, `hispanic`, and `white`.
  - These are fixed effects.
- Then add interactions between `black`, `hispanic`, `white`, and `asian` and the linear plus functions in `age`.
  - These are mean-zero random effects with their own variance component
  - This variance component control the amount of shrinkage of the enthicity-specific curves to the overall effect.

Slide 36

Bivariate additive spline model

$$Y_i = \beta_0 + \beta_{x,1} X_i + b_{x,1}(X_i - \kappa_{x,1})_+ + \cdots + b_{x,K}(X_i - \kappa_{x,K_x})_+$$

$$+ \beta_{z,1} Z_i + b_{z,1}(Z_i - \kappa_{z,1})_+ + \cdots + b_{z,K}(Z_i - \kappa_{z,K_z})_+ + \epsilon_i$$

- no need for backfitting
- computation very rapid
- no identifiability issues
- inference is simple

ide 37

Bayesian methods

The linear mixed model is half-Bayesian.

- The random effects have a prior.
- The parameters without a prior are:
  - fixed effects
    - give them diffuse normal priors
  - variance components
    - give them diffuse inverse gamma priors

ide 38

Bayesian methods

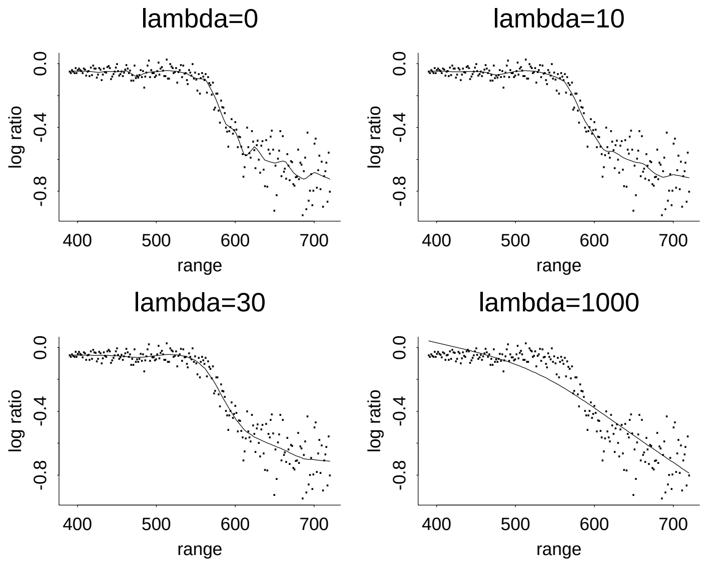
Can be easily implemented in WinBUGS or programmed in, say, MATLAB.

Allows Bayes rather than empirical Bayes inference.

- Uncertainty due to smoothing parameter selection is taken into account.

Slide 39

The Bias-Variance Trade-off and Confidence Bands

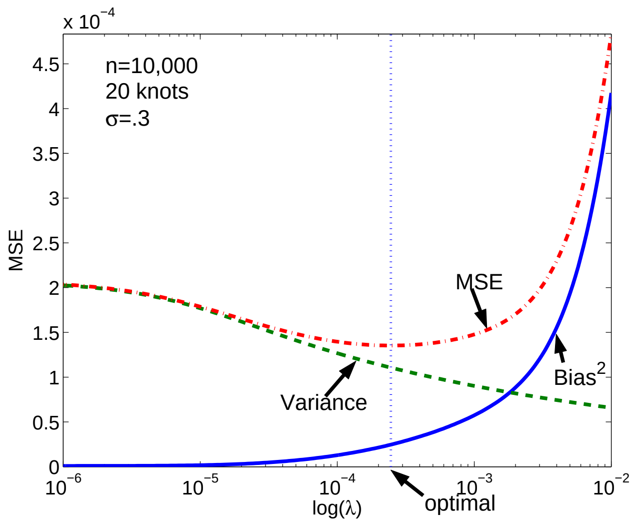


Slide 40

How does one adjust confidence intervals for bias?

- undersmooth — so variance dominates and bias can be safetly ignored.

ide 41



Slide 43

Wahba/Nychka Bayesian Intervals

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \mathrm{Cov}\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \sigma_u^2 \mathbf{I} & 0 \\ 0 & \sigma_\varepsilon^2 \mathbf{I} \end{bmatrix},$$

$$\mathbf{C} = (\, \mathbf{X} \quad \mathbf{Z}\,)$$

$\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\mathbf{u}}$ are BLUPs.

ide 42

Adjustment for bias continued

- estimate bias by a higher order method and subtract off bias (essentially the same as above)
- Wahba/Nychka Bayesian intervals
  - bias is random so adds to posterior variance
  - interval is widened but there is no "offset".

Slide 44

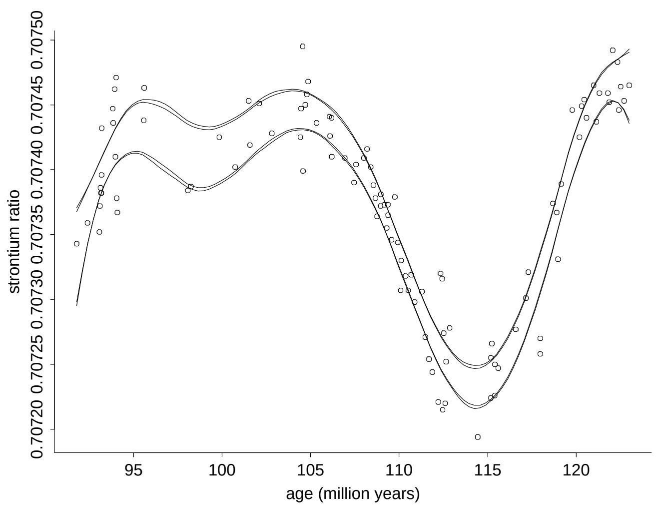
$$\mathrm{Cov}\left(\begin{bmatrix} \widetilde{\boldsymbol{\beta}} \\ \widetilde{\mathbf{u}} \end{bmatrix} \Big| \mathbf{u}\right) = \sigma_\varepsilon^2(\mathbf{C}^\mathsf{T}\mathbf{C} + \tfrac{\sigma_\varepsilon^2}{\sigma_u^2}\mathbf{D})^{-1}\mathbf{C}^\mathsf{T}\mathbf{C}(\mathbf{C}^\mathsf{T}\mathbf{C} + \tfrac{\sigma_\varepsilon^2}{\sigma_u^2}\mathbf{D})^{-1}$$

(Frequentist variance. Ignores bias)

$$\mathrm{Cov}\left(\begin{bmatrix} \widetilde{\boldsymbol{\beta}} \\ \widetilde{\mathbf{u}} - \mathbf{u} \end{bmatrix}\right) = \sigma_\varepsilon^2(\mathbf{C}^\mathsf{T}\mathbf{C} + \tfrac{\sigma_\varepsilon^2}{\sigma_u^2}\mathbf{D})^{-1}.$$

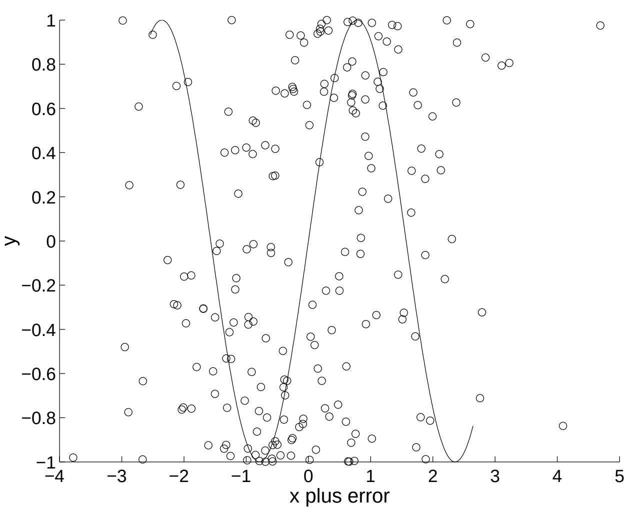
(Bayesian posterior variance. Takes bias into account.)

ide 45



ide 46

Effect of measurement error



$W = X + \text{error and Var}(X) = \text{Var(error)}.$

Slide 47

Correction for measurement error

Relatively little research in this area.

- Fan and Truong (1993): deconvolution kernels
  - first work
  - inefficient in finite-sample studies
  - no inference
  - strictly for 1-dimensional smoothing
- Carroll, Maca, Ruppert
  - functional SIMEX methods and structural spline methods
  - more efficient than Fan and Truong

Slide 48

- Berry, Carroll, and Ruppert (JASA, 2002)
  - fully Bayesian
  - smoothing or penalized splines
  - rather efficient in finite-sample studies
  - inference available
  - scales up — semiparametric inference is easy
  - structural

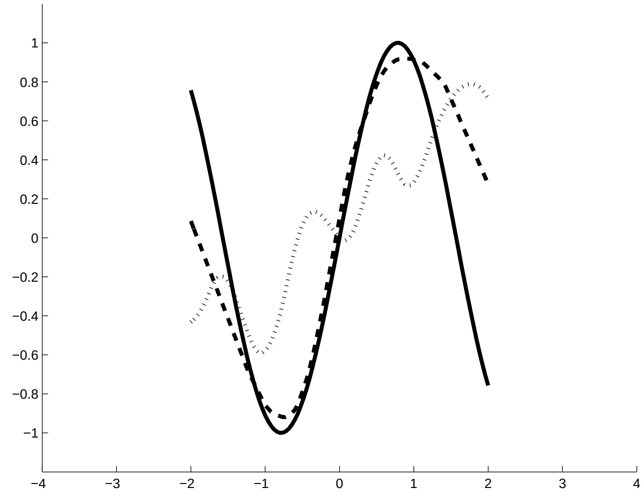
ide 49

Berry, Carroll, and Ruppert

- starts with mixed-model spline formulation
  - but fully Bayesian
- conjugate priors
- true covariates are i.i.d. normal
  - but surprisingly robust
- normal measurement error
- in Gibbs, only sampling of true (unknown) covariates requires a Hastings-Metropolis step

Slide 51

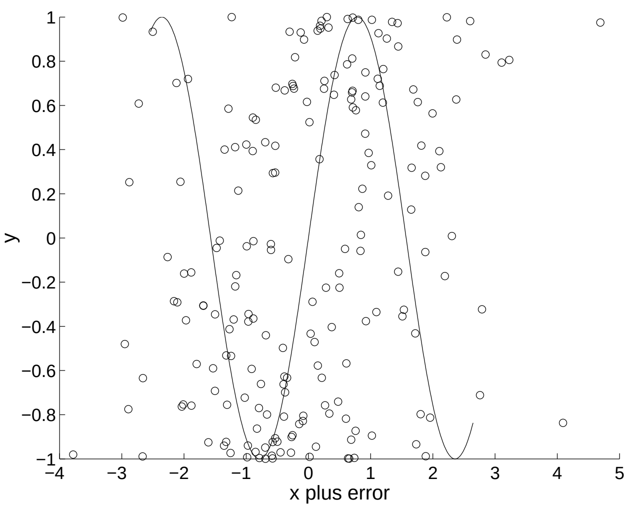
Correction for measurement error



Solid: true. **Dotted:** uncorrected. **Dashed:** corrected.

ide 50

Effect of measurement error



$W = X + \text{error}$ and $\text{Var}(X) = \text{Var}(\text{error})$.

Slide 52

Measurement Error, continued

Ganguli, Staudenmayer, Wand:

- EM maximum likelihood estimation in BCR model.
- Works about as well as the fully Bayesian approach.
- Extension to additive models.

Generalized Regression

ide 53

- Extension to non-Gaussian responses is conceptually easy.
- Get a GLLM.
  - However, GLIM's are not trivial. Can use:
    * Monte Carlo EM
    * Or MCMC

Single-Index Models

$$Y_i = g(\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\theta}) + \mathbf{Z}_i^{\mathsf{T}}\boldsymbol{\beta} + \epsilon_i.$$

Yu and Ruppert (2002, JASA).

ide 54

Let

$$g(x) = \gamma_0 + \gamma_1 x + \cdots + \gamma_p x^p$$
$$+c_1(x-\kappa_1)_+^p + \cdots + c_K(x-\kappa_K)_+^p.$$

Becomes a nonlinear regression model

$$Y_i = m(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{c}) + \epsilon_i.$$