Semiparametric Modeling with Splines: a Personal Perspective

David Ruppert
Cornell University

Sep 5, 2014

Outline

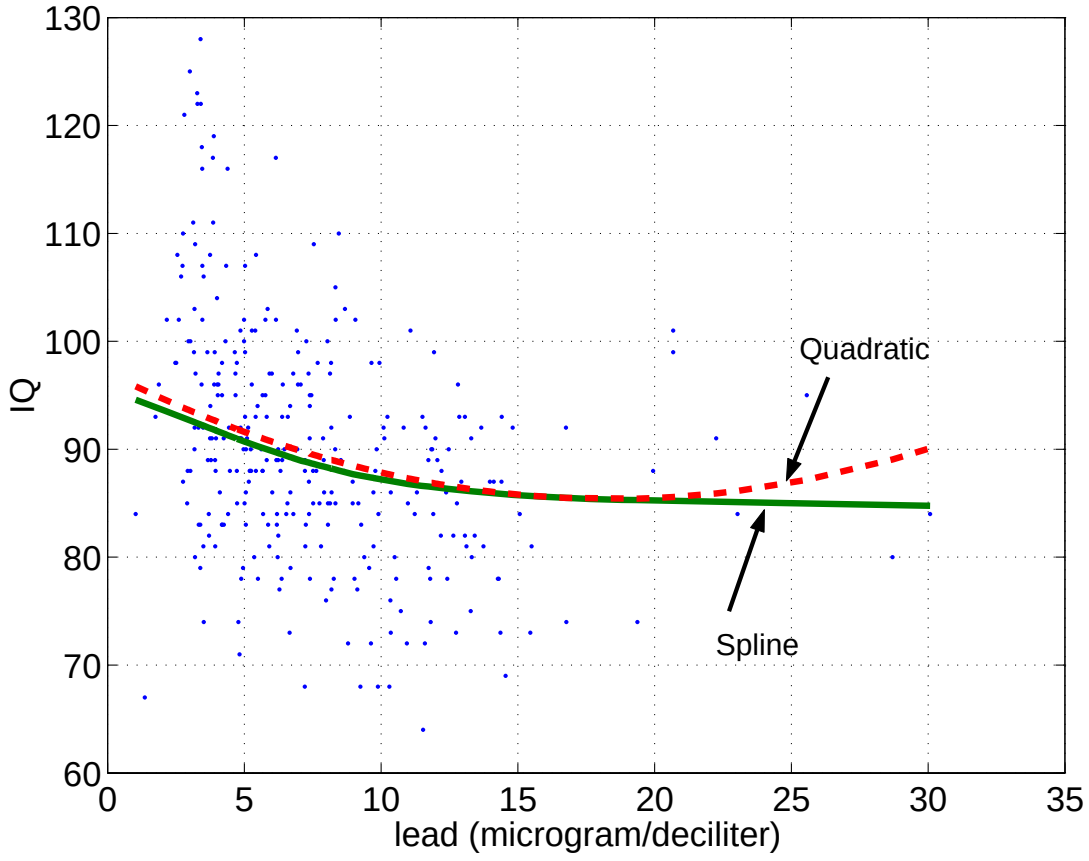
1. Two examples.
2. Splines.
3. Splines as Mixed Models.
4. Local CLT for P-splines.
5. Likelihood ratio tests of polynomial regression versus nonparametric regression.
6. Software.
7. Bivariate regression and estimation of covariance functions.
8. Functional generalized additive model.

Intellectual impairment and blood lead

Example I (courtesy of Rich Canfield, Nutrition, Cornell)

- Blood lead concentration and intelligence were measured on children.
- Question: How do low doses of lead affect IQ?
- Several IQ measurements per child
  - so longitudinal
- Nine "confounders"
  - e. g., maternal IQ
- Effect of lead appears nonlinear
- Rich contacted me in 1998
  - New methodology was needed.
  - Rich wanted something like a SAS Proc GAMMixed, which did not exist

Dose-response curve



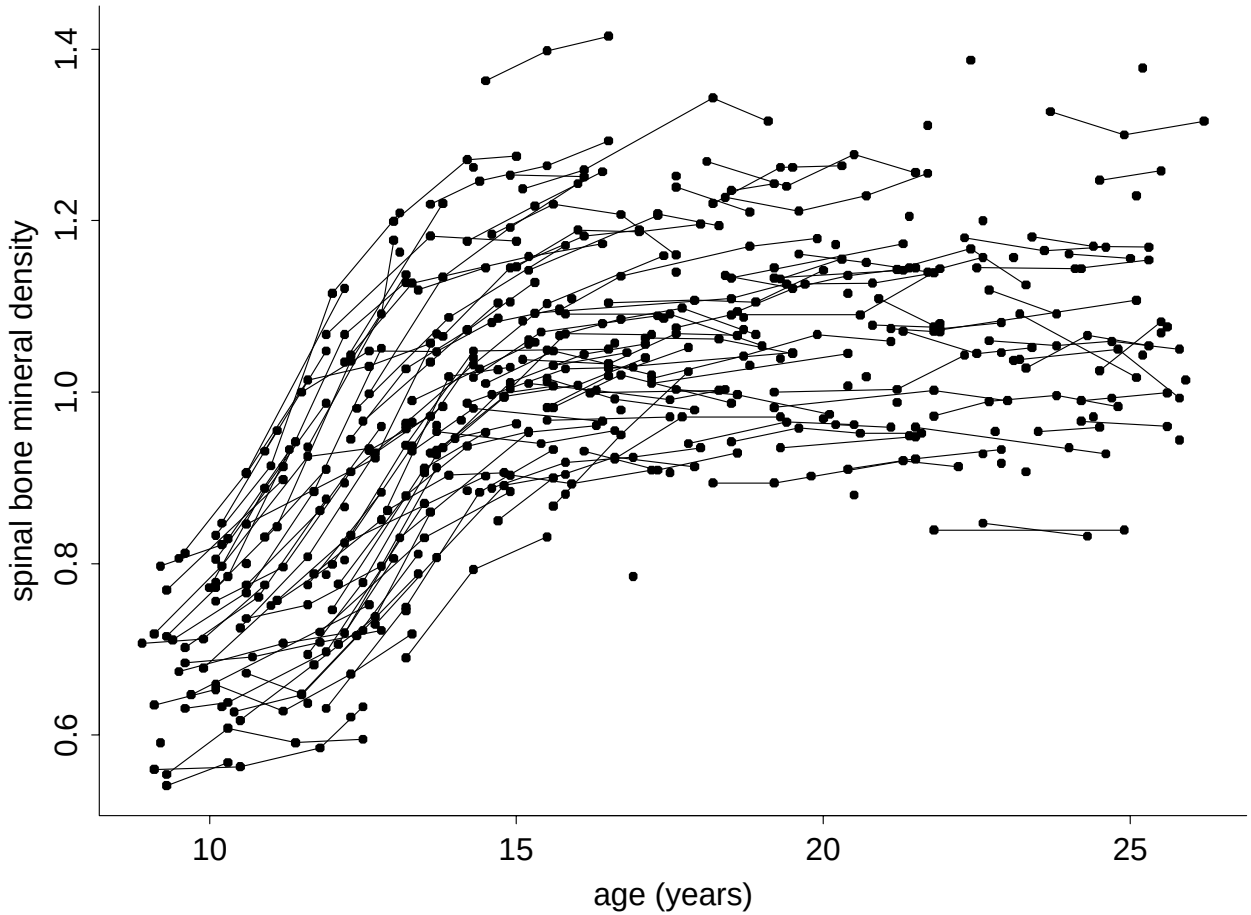
Thanks to Rich Canfield for data and estimates. In this plot, the IQ measurements have been adjusted for confounders.

Spinal bone mineral density example

Example II

- age and spinal bone mineral density measured on girls and young women
- several measurements on each subject
- increasing but nonlinear curves

Spinal bone mineral density data



What is needed to accommodate these examples?

We need a model with

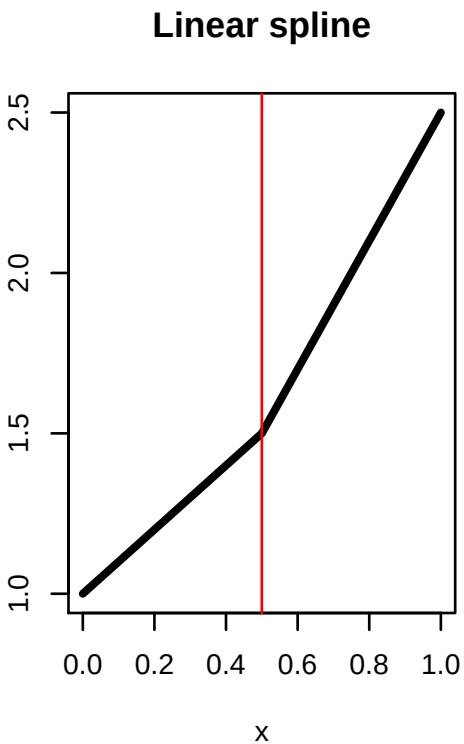
- potentially many variables
- possibility of nonlinear effects
- random subject-specific effects

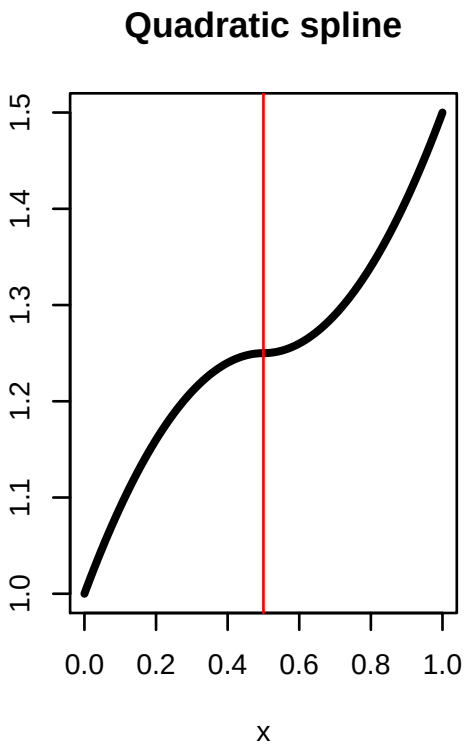
The model should be one that can be fit with readily available software such as R.

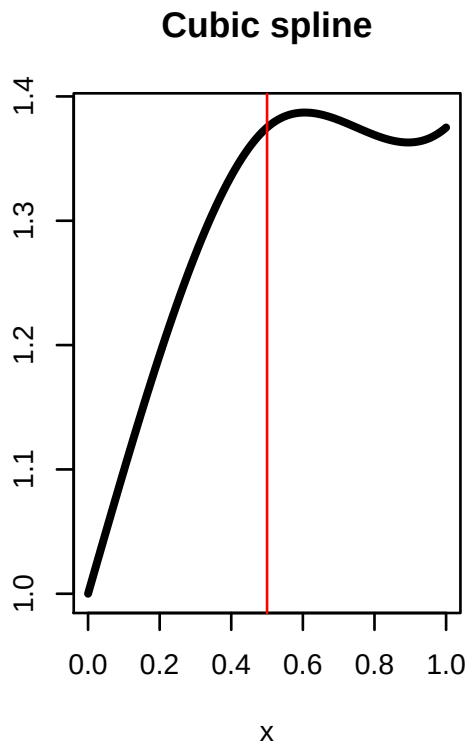
A spline is a piecewise polynomial with maximum smoothness

- The polynomial form changes at fixed locations called "knots."
- If the polynomials are of degree p, then the spline has $p-1$ derivatives at the knots.

Examples of Splines







Each spline has a knot at 0.5.

- Smoothing splines was developed by numerical analysts.
- Grace Wahba introduced them to statistics.
- Model: $Y_i = f(X_i) + \epsilon_i$.
- $\widehat{f}$ minimizes $\sum\{ Y_i - f(X_i)\}^2 + \lambda \int \{f''(x)\}^2 dx$.
- λ controls the smoothness of the spline.
    - Selecting an appropriate value λ is crucial.

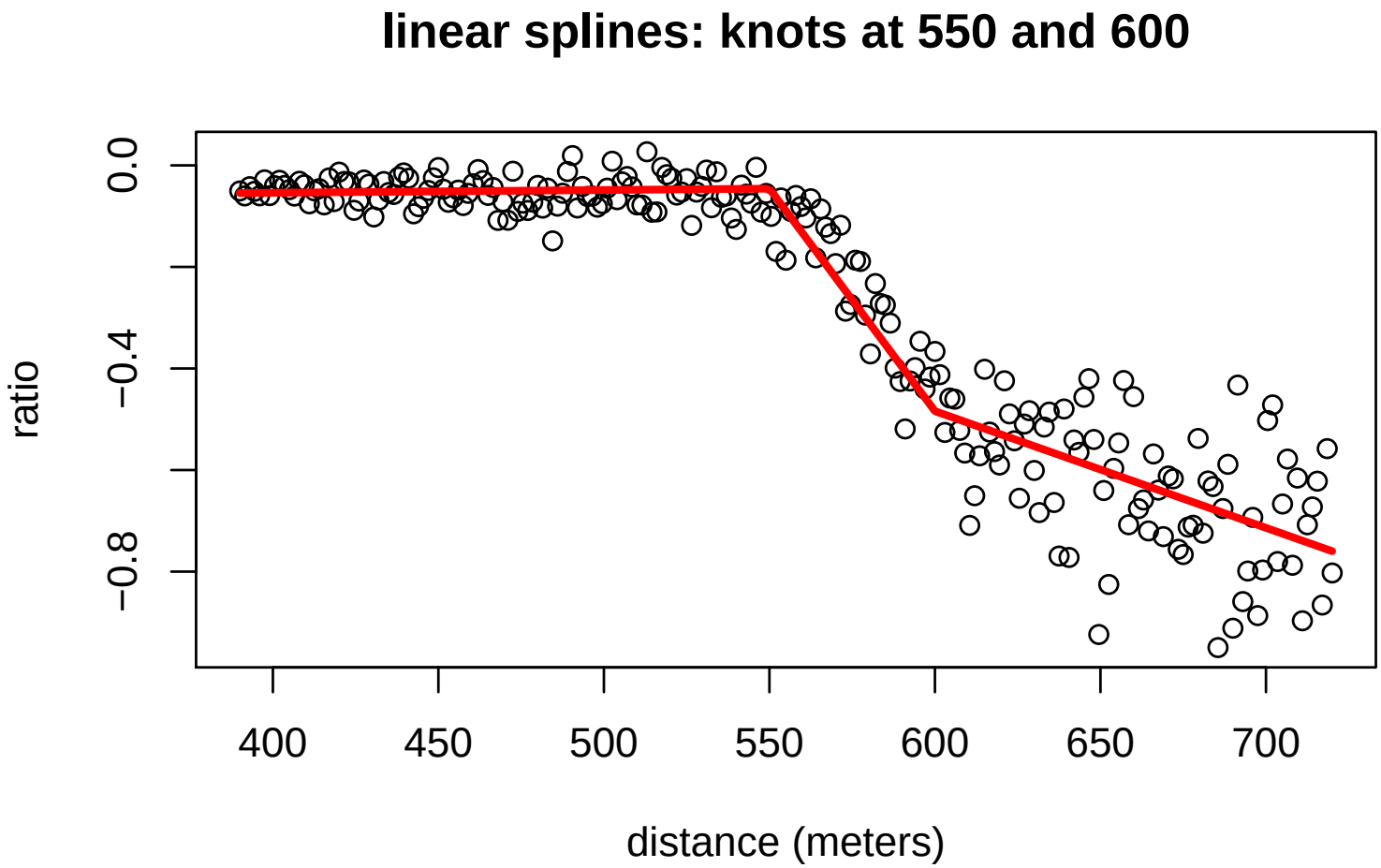
Limitations of Splines

- Smoothing splines are fantastic for univariate regression.
- There are fast $O(n)$ algorithms.
- They are less satisfactory for complex problems such as those introduced earlier.
    - The large number of knots can be a serious problem.

Regression Splines

- Regression splines are fit using standard software.
- For a linear spline, one uses a basis function $(x - \kappa)_+$ for each knot κ.
  - $x_+ := x\, I(x > 0)$ is the positive part function.
  - higher degree splines are similar.
- It is easy to embed regression splines into multiple regression models.
- A regression spline does not use a roughness penalty, so selecting the number and locations of the knots is crucial.

Regression Spline Example: Lidar



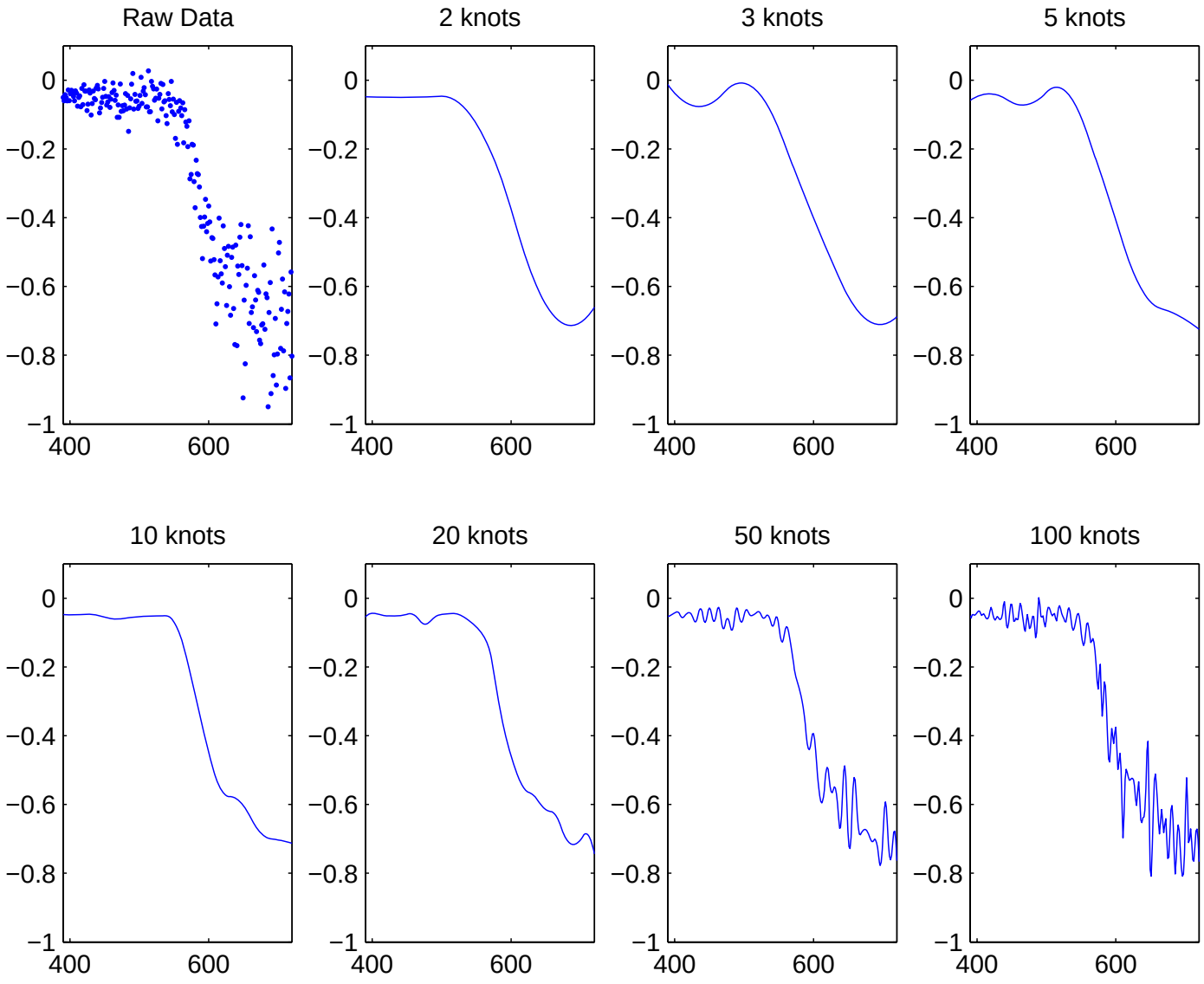
Model:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^j + \sum_{j=1}^{K} u_j (X_i - \kappa_j)_+^p + \epsilon_i := s(X_i) + \epsilon_i.$$

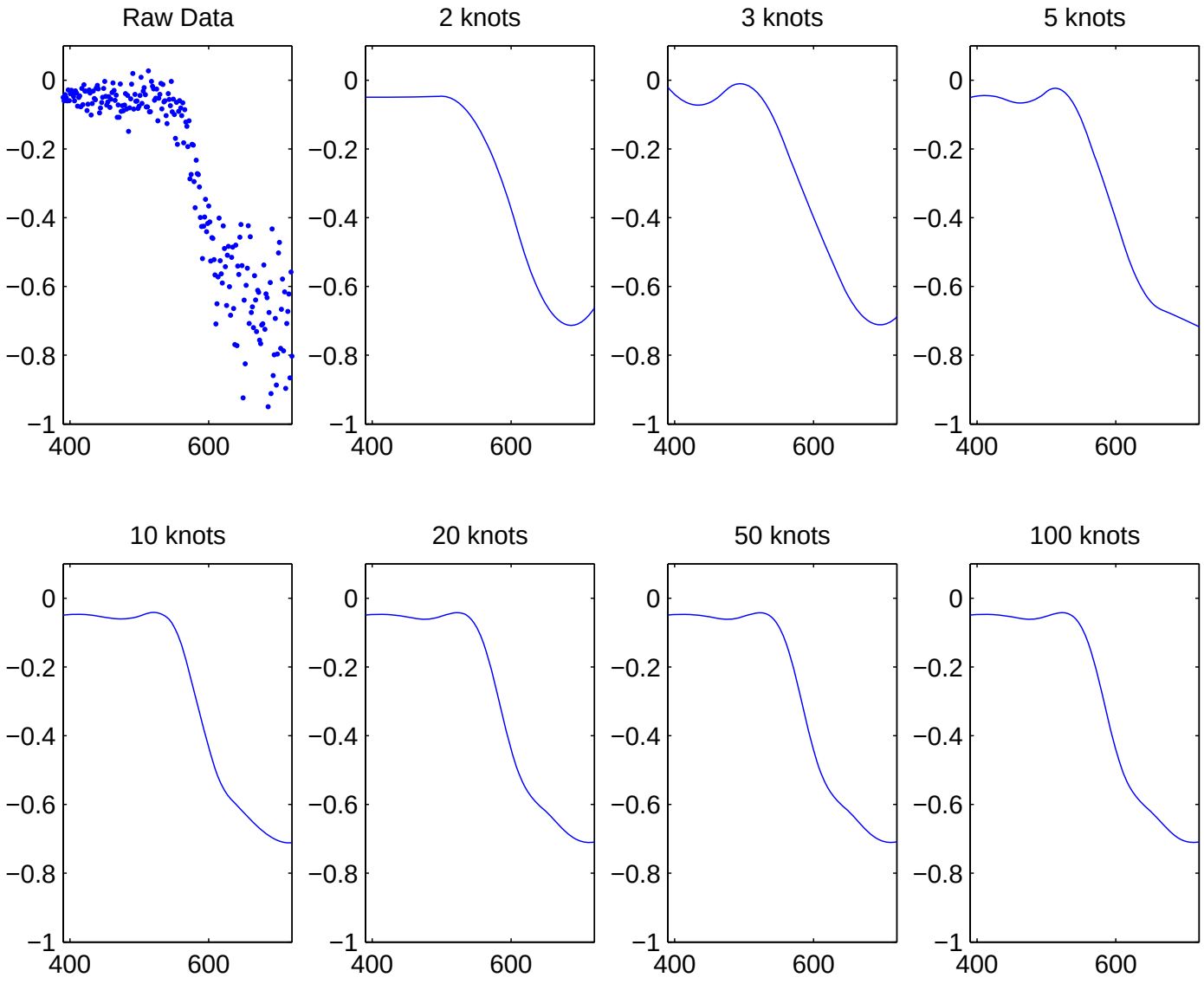
Objective function:

$$\sum_{i=1}^{n} \{ Y_i - s(X_i) \}^2 + \lambda \sum_{j=1}^{K} u_i^2.$$

Lidar data: unpenalized quadratic regression splines



Lidar data: penalized splines with λ selected by GCV



Model:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^j + \sum_{j=1}^{K} u_j (X_i - \kappa_j)_+^p + \epsilon_i := s(X_i) + \epsilon_i.$$

Objective function:

$$\sum_{i=1}^{n} \{ Y_i - s(X_i) \}^2 + \lambda \sum_{j=1}^{K} u_i^2.$$

Equivalent to assuming that $u_1, \ldots, u_K$ are iid $N(0, \sigma_u^2)$ and $\lambda = \sigma_\epsilon^2 / \sigma_u^2$.

- Moreover, σ_ϵ^2 and σ_u^2 can be estimated by REML.

Return to Blood Lead and IQ

Here's how Rich Canfield analyzed his data:

- Model the effect of blood lead concentration as a spline with random coefficients of the "positive part functions."
- The polynomial coefficients are fixed effects.
- Model the confounders as linear fixed effects.
- Add random subject-specific effects to model correlation.

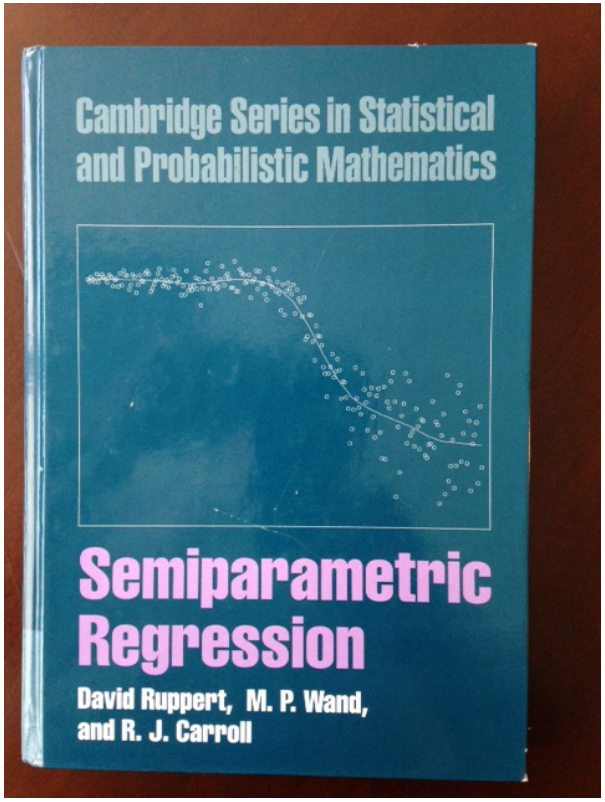
Mixed Model for Blood Lead and IQ

Model:

$$Y_{i,\ell} = \underbrace{b_i}_{\text{subject effect}} + \underbrace{\beta_0 + \sum_{j=1}^{p} \beta_j X_i^j}_{\text{polynomial: fixed effects}} + \underbrace{\sum_{k=1}^{K} u_k (X_i - \kappa_j)_+^p}_{\text{spline: random effects}} + \underbrace{\sum_{j=1}^{M} \alpha_j Z_{i,j}}_{\text{confounder effects}} + \epsilon_i.$$

- $Y_{i,\ell}$ is the ℓth IQ measurement on the ith subject.
- X_i is the lead concentration of the ith subject.
- $Z_{i,j}$ is the jth confounder variable measured on the ith subject.

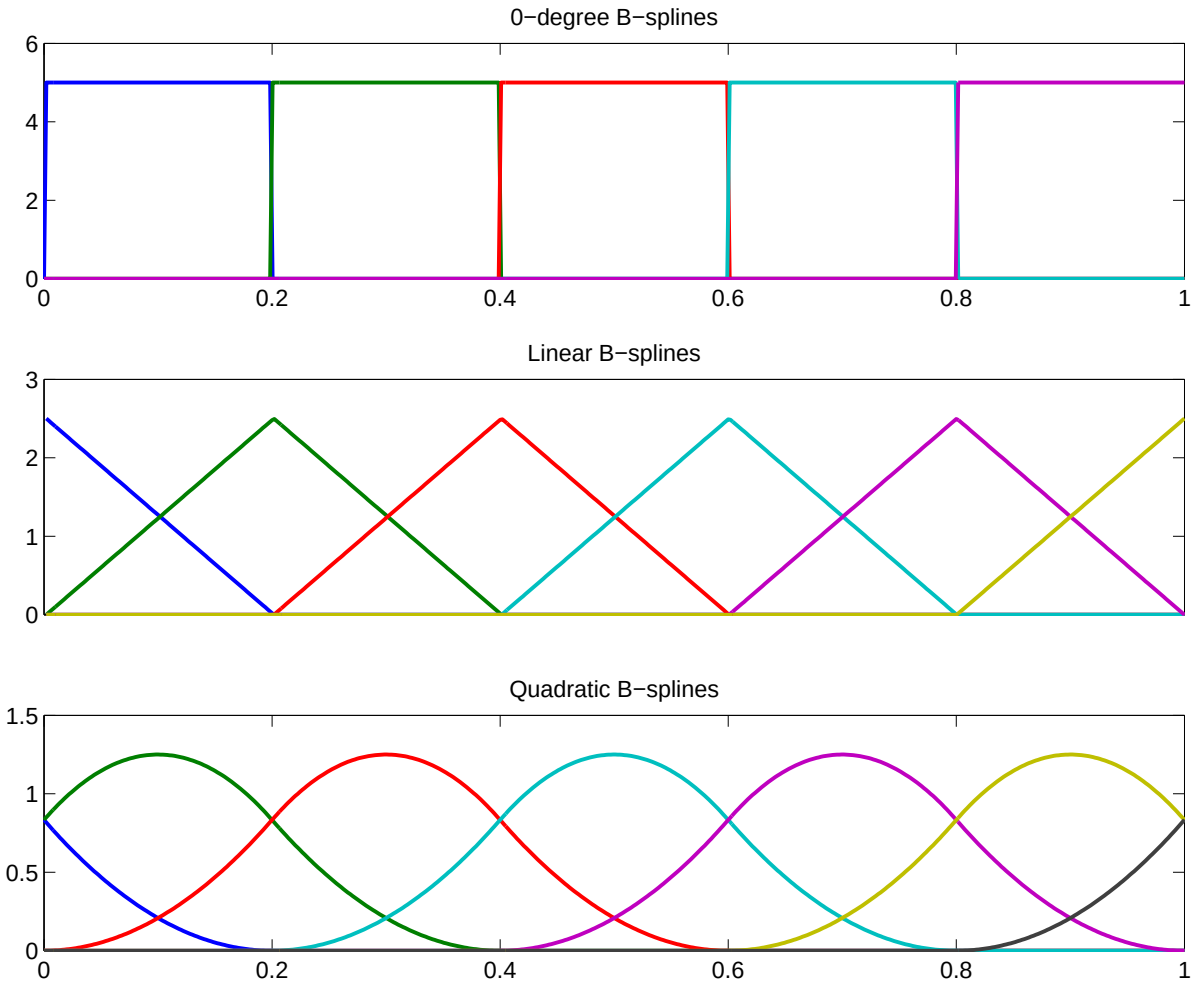
- In 1998 I visited Matt Wand at HSPH.
- We planned to start work on a book on smoothing.
- From discussions with Babette Brumback, we learned of the mixed model approach to splines.
- That approach become a major focus of the book.
- Right after I returned to Cornell, I learned of Rich Canfield's research.



Published in 2003.

- With the knots and the degree fixed, the set of splines is a vector space.
  - The dimension is $1 + p + K$ where p is the degree and K is the number of knots.
- A very convenient basis is the set of B-splines.
- B-splines have minimal support and are numerically stable.

Examples of B-splines



A P-spline is of the form $\sum_{j=1}^{1+p+K} b_j B_j(x)$.

- $B_1(x), \ldots, B_{1+p+K}(x)$ is the B-spline basis.
- $\kappa_1, \ldots, \kappa_K$ are equally-spaced knots.

The penalty used with a P-spline is $\sum_{k=1+m}^{K} \{\Delta^m b_k\}^2$.

- Δ is the differencing operator.
- m is the order of the differencing.
- Asymptotically, m is more important than p and K.
  - More later.

Li and Ruppert (2008, *Biometrika*) developed an asymptotic theory for P-splines, but only for special cases.

- There were technical difficulties when $q = \max(p, m) > 2$.
  - The difficulties were in studying the roots of certain polynomial of degree $2q$
  - Unpublished work with Tanya Apanosovich has solved this problem.
- Wang, Shen, and Ruppert (2011, *EJS*) obtained general results using Green's functions.

Under assumptions (e.g., enough derivatives) for any $x \in (0,1)$ [interior points], we have

$$n^{\frac{2m}{4m+1}}\{\hat{f}(x) - f(x)\} \to N\{\tilde{\mu}(x), V(x)\}, \text{ as } n \to \infty,$$

where

$$\tilde{\mu}(x) = \frac{1}{(2m)!}\mu^{(2m)}(x)h^{2m}\int t^{2m}H_m(t)\,dt$$

and

$$V(x) = h^{-1}\sigma^2(x)\int H_m^2(t)\,dt$$

- m is the order of the difference penalty.
- H_m is the "equivalent kernel."
- h is the "equivalent bandwidth" and depends on λ.

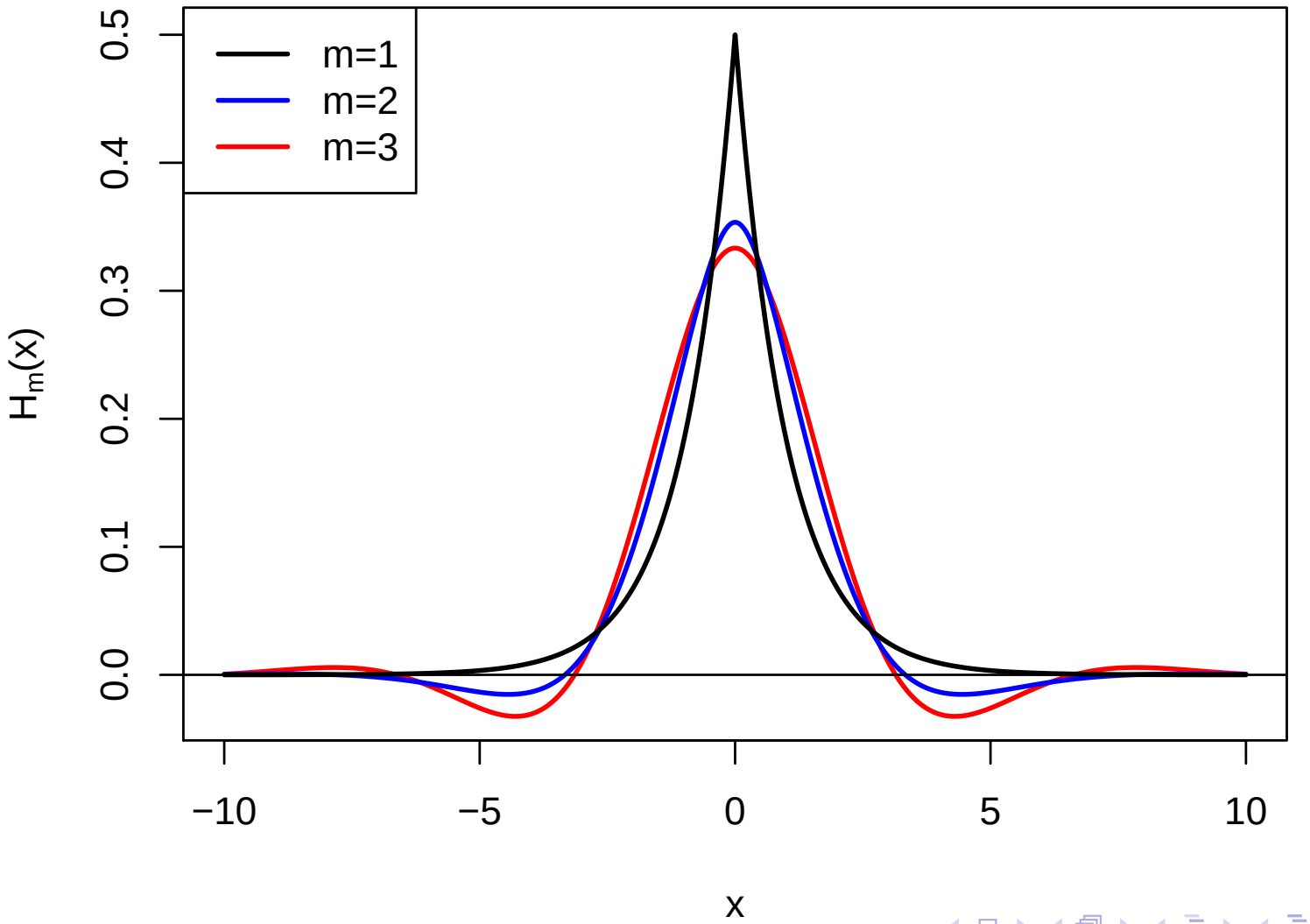
For any m: H_m is symmetric about 0,

$$\int H_m(x)\,dx = 1,$$

and

$$\int x^k H_m(x)\,dx = 0 \quad \text{for} \quad k = 1, \ldots, 2m-1$$

Equivalent kernels for $m = 1$, 2, and 3 (interior)



Assumptions

The assumptions of the theorem confirm some folklore.

- Folklore:

  Number of knots (K) not important, provided large enough.

- Confirmation:

  $K \sim K_0 n^{\gamma}$, where

  - $K_0 > 0$
  - $\gamma > 2m/\{\ell(4m+1)\}$
  - $\ell := \min(2m, p+1)$

Some folklore: penalty parameter

- Folklore:

  Value of the penalty parameter (λ) crucial.

- Confirmation:

  $\lambda \sim (Kh)^{2m}$ where $h \sim h_0 n^{-\frac{1}{4m+1}}$

- Folklore:

  Modeling bias (= bias due to approximating the regression function by a spline) small.

- Confirmation: Modeling bias does not appear in asymptotic bias

Testing Null Hypothesis of Polynomial Regression

Model: $Y_i = f(X_i) + \underbrace{\sum_{j=1}^{M} \alpha_j Z_{i,j}}_{\text{confounder effects}} + \epsilon_i.$

H_0: $f(x)$ is a pth degree polynomial $\Rightarrow \sigma_u^2 = 0$ in mixed model

$$Y_{i,\ell} = \underbrace{b_i}_{\text{subject effect}} + \underbrace{\beta_0 + \sum_{j=1}^{p} \beta_j X_i^j}_{\text{polynomial: fixed effects}} + \underbrace{\sum_{k=1}^{K} u_k (X_i - \kappa_j)_+^p}_{\text{spline: random effects}}$$

$$+ \underbrace{\sum_{j=1}^{M} \alpha_j Z_{i,j}}_{\text{confounder effects}} + \epsilon_i.$$

- H_0: $\sigma_u^2 = 0$ is mixed model is nonstandard.
- The null hypothesis is on the boundary of the parameter space.
  - This problem has been studied as far back as Herman Chernoff.
  - Self and Liang (1987, *JASA*) and Stram and Lee (1994, *Biometrics*) are more recent references.
  - In simulations, the Chernoff/Self & Liang limit theory did not appear to hold.

- Non-standard for a second reason.
  - The data are not independent under the alternative.
  - They do not even satisfy any of the usual mixing conditions.
- In his thesis, Ciprian Crainiceanu found the limiting null distribution of the LRT and RLRT test statistics.
- References:
  - Crainiceanu, Ruppert, Claeskens, and Wand (2003, *Biometrika*).
  - Crainiceanu and Ruppert (2004, *JRSS-B*).

We Now Have Great Software

The use of splines in mixed models is now relatively easy due to great software.

- Simon Wood's mgcv package and his book *Generalized Additive Models* are excellent.
- Models that cannot be fit by "canned software" can often be fit using a Bayesian analysis and MCMC implement using BUGS or Stan.

Slide From a Talk by Phil Reiss

Great Books

Forthcoming Book

Semiparametric Regression with R will be published next year.

- Authors:
  - Jarek Harezlak
  - David Ruppert
  - Matt Wand
- Will cover software in R, OpenBUGS, and Stan.

Bivariate Regression: $y_i = m(s_i, t_i) + \epsilon_i$

Tensor product spline: $m(s,t) = \sum_{k=1}^{c} \sum_{\ell=1}^{c} \beta_{k,l} B_k(s) B_\ell(t)$

- $B_1, \ldots, B_c$ is a univariate basis

Eilers and Marx's bivariate P-spline uses row penalties and column penalties.

Suppose the y_{ij} are observed on a $J_1 \times J_2$ rectangular grid, e.g., a covariance matrix where $J_1 = J_2 = J$.

- Put the $y_{i,j}$ in a matrix $\mathbf{Y}$.
- The sandwich smoother is $\widehat{\mathbf{Y}} = \mathbf{S}_1 \mathbf{Y} \mathbf{S}_2$. (Xiao et al., 2013, *JRSS-B*).
  - Here $\mathbf{S}_1$ and $\mathbf{S}_2$ are univariate hat matrices.
- The sandwich smoother can be derived from the Eilers and Marx bivariate P-spline by modifying the penalty.

Sandwich smoother in matrix notation:

$$\widehat{\mathbf{Y}} = \mathbf{S}_1 \mathbf{Y} \mathbf{S}_2$$

where $\widehat{\mathbf{Y}}$ and $\mathbf{Y}$ are rectangular matrices.

Sandwich smoother in vector notation:

$$\widehat{\mathbf{y}} = (\mathbf{S}_2 \otimes \mathbf{S}_1)\mathbf{y}$$

where $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$ and $\widehat{\mathbf{y}} = \mathrm{vec}(\widehat{\mathbf{Y}})$.

The Sandwich Smoother is Fast

- For fixed smoothing parameters, a bivariate spline can be computed as a generalized linear array model (GLAM) (Currie, Durban, and Eilers, 2006, JRSS-B).
- The bottleneck for GLAM is in computing the effective degrees of freedom (DF) needed for GCV (Generalized Cross Validtion) to select the smoothing parameters.

Recall: Sandwich smoother in vector notation:

$$\widehat{\mathbf{y}} = (\mathbf{S}_2 \otimes \mathbf{S}_1)\mathbf{y},$$

where $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$ and $\widehat{\mathbf{y}} = \mathrm{vec}(\widehat{\mathbf{Y}})$.

From the vector notation, we see that DF $= \mathrm{tr}(\mathbf{S}_2 \otimes \mathbf{S}_1) = \mathrm{tr}(\mathbf{S}_2)\mathrm{tr}(\mathbf{S}_1)$: fast to compute

Recall the sandwich smoother of the sample covariance matrix:

$$\widetilde{\mathbf{K}} = \mathbf{S}\widehat{\mathbf{K}}\mathbf{S}.$$

All four matrices are $J \times J$.

The rank of $\widetilde{\mathbf{K}}$ is at most $\min(J, c)$ where

- c is the dimension of the spline basis so $c = \text{rank}(\mathbf{S})$.

FACE (FAst Covariance Estimation) uses this low rank to improve the speed and reduce the storage requirements of the sandwich formula significantly.

- Due to Luo, Ruppert, Zipunnikov, and Crainiceanu

The Functional Generalized Additive Model (FGAM) is

$$g\{E(Y_i|X_i)\} = \theta_0 + \int_{\mathcal{T}} F\{X_i(t), t\}dt.$$

- $F(\cdot, \cdot)$ is an unknown function from $\mathcal{X} \times \mathcal{T}$ to $\Re$
- if $F(x, t) = x\beta(t)$, then we have a Functional linear model

Model introduced and studied by McLean, Hooker, Staicu, and Ruppert.

We use the bivariate tensor product B-spline model:

$$F(x,t) = \sum_{j=1}^{K_x} \sum_{k=1}^{K_t} \theta_{j,k} B_j^X(x) B_k^T(t).$$

- Here $\{B_j^X(x) : j = 1, \ldots, K_x\}$ and

  $\{B_k^T(x) : k = 1, \ldots, K_t\}$ are univariate B-spline bases.

- A roughness penalty is imposed on $\{\theta_{j,k}\}_{j=1\,k=1}^{K_x\ K_t}$.

From previous slide:

$$F(x,t) = \sum_{j=1}^{K_x} \sum_{k=1}^{K_t} \theta_{j,k} B_j^X(x) B_k^T(t).$$

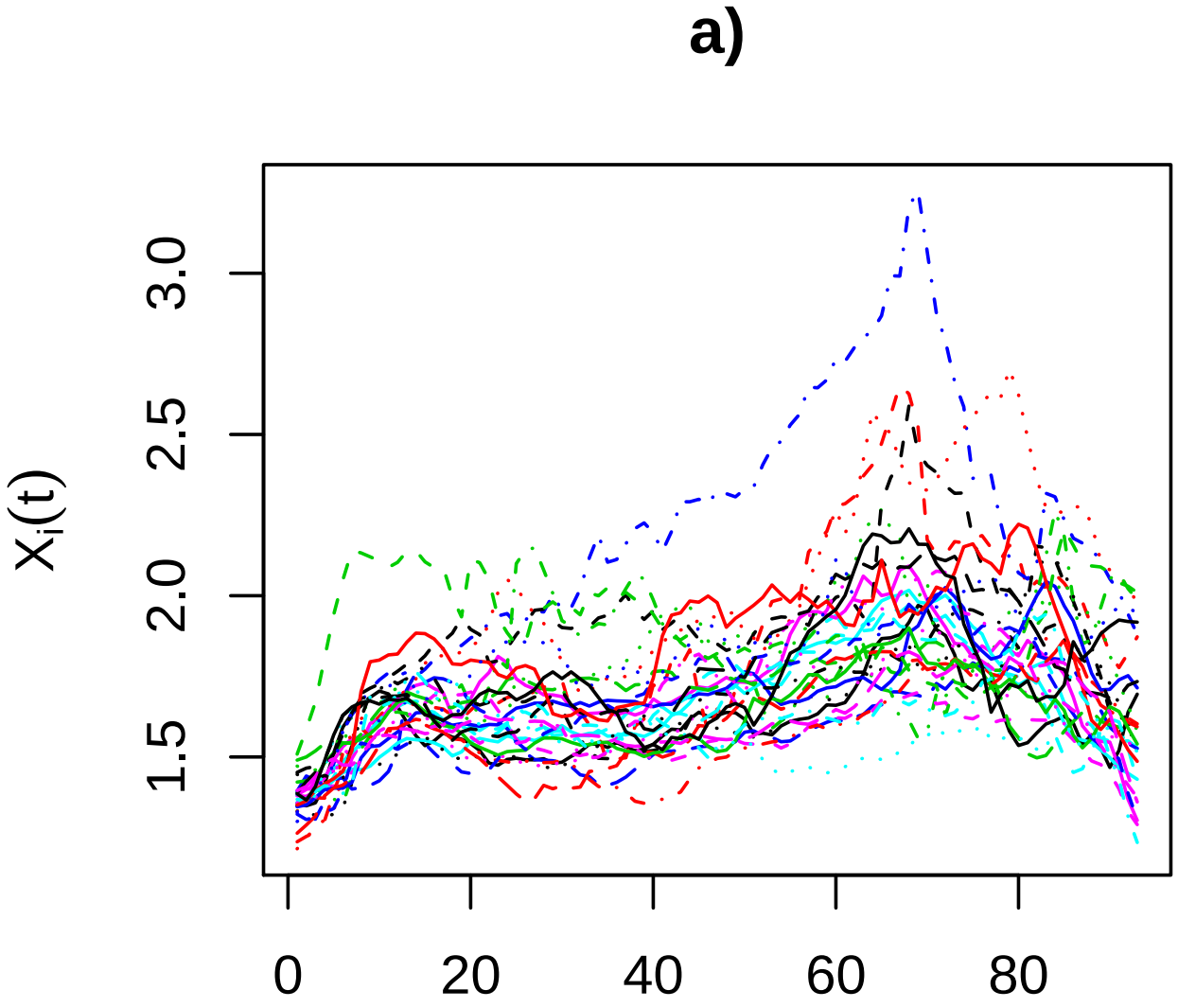
Therefore,

$$g\{E(Y_i|X_i)\} = \int_{\mathcal{T}} F\{X_i(t), t\}\, dt = \sum_{j=1}^{K_x} \sum_{k=1}^{K_t} \theta_{j,k}.Z_{j,k}(i)$$

- Here $Z_{j,k}(i) = \int_{\mathcal{T}} B_j^X\{X_i(t)\} B_k^T(t)\, dt$.
- The integral can be approximated numerically.

DTI data – parallel diffusivity along the corpus callosum

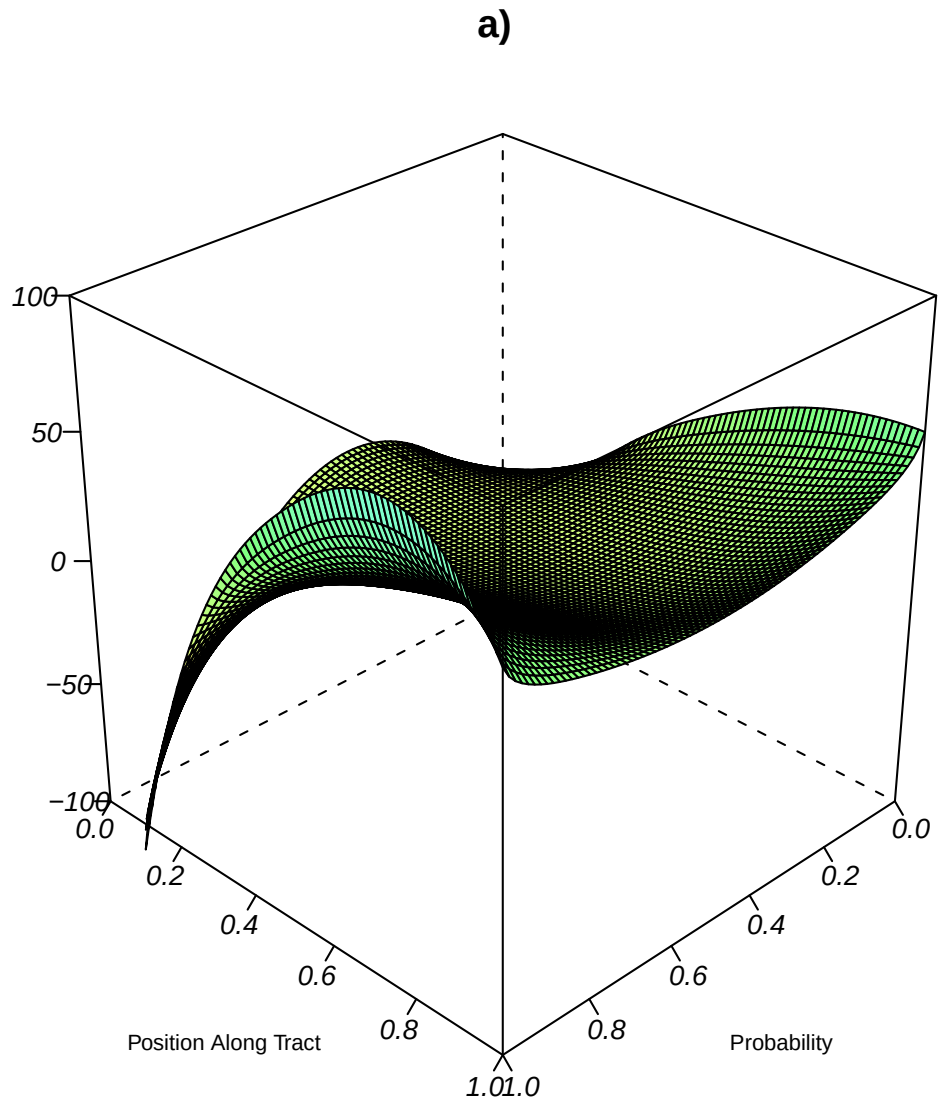
MS patients only – Untransformed



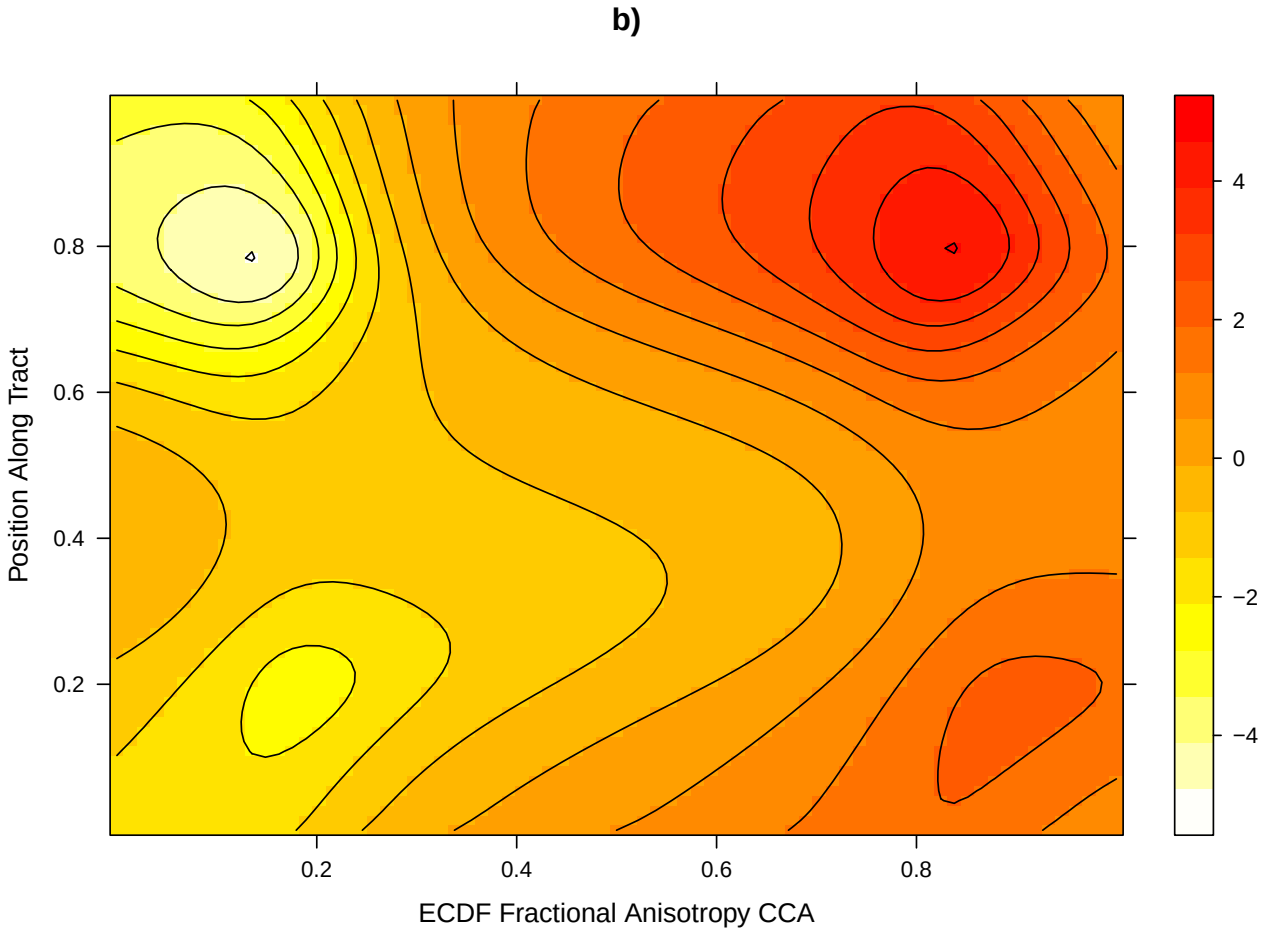
PASAT

- PASAT = Paced Auditory Serial Addition Test
- Subject given numbers at three second intervals
  - asked to add the current number to the previous one
- MS patients often perform significantly worse than controls

Estimated surface $\widehat{F}(p, t)$ for fractional anisotropy



t-statistics for fractional anisotropy



CCA = corpus callosum

Thanks for coming!