Penalized Splines, Mixed Models, and Recent Large-Sample Results

David Ruppert

Operations Research & Information Engineering, Cornell University

Feb 4, 2011

Collaborators

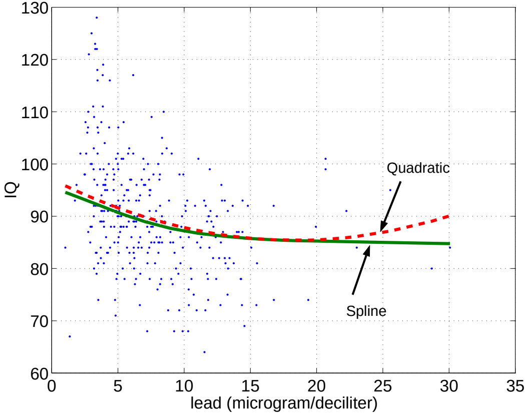
- Matt Wand, University of Wollongong
- Raymond Carroll, Texas A&M University University
- Yingxing (Amy) Li, Cornell University
- Tanya Apanosovich, Thomas Jefferson Medical College
- Xiao Wang, Purdue University
- Jinglai Shen, University of Maryland, Baltimore County
- Luo Xiao, Cornell University

Two parts

- Old: overview of the book *Semiparametric Regression* by Ruppert, Wand, and Carroll (2003)
  - Still an active area: 314 papers referenced in *Semiparametric Regression During 2003–2007* (*EJS*, 2009)
- New: asymptotics of penalized splines

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

- blood lead and intelligence measured on children
- Question: how do low doses of lead affect IQ?
  - important – doses decreasing since lead no longer added to gasoline
- several IQ measurements per child
  - so longitudinal
- nine "confounders"
  - e. g., maternal IQ
  - need to adjust for them
- effect of lead appears nonlinear
- important conclusion

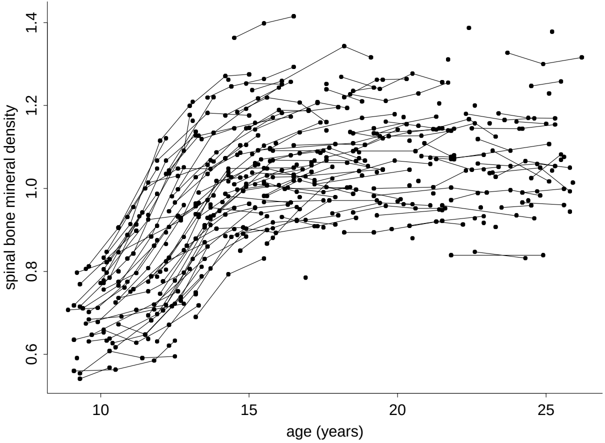


Thanks to Rich Canfield for data and estimates

Example II (in Ruppert, Wand, Carroll (2003), *Semiparametric Regression*

- 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 SAS, Splus, or R.

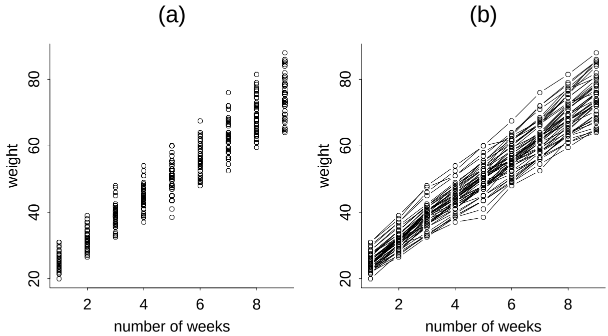
1. minimalist statistics
   - keep it as simple as possible
   - but need to accommodate features such as correlated data and confounders
2. build on classical parametric statistics
3. modular methodology
   - so we can add components to accommodate special features in data sets

Outline of the approach

- Start with linear mixed model
  - allows random subject-specific effects
  - fine for variables that enter linearly
- Expand the basis for those variables that have nonlinear effects
  - we will use a spline basis
  - treat the spline coefficients as random effects to induce empirical Bayes shrinkage = smoothing
- End result
  - linear mixed model from a software perspective, but
  - nonlinear from a modeling perspective

Example: pig weights (random effects)

Example III [from Ruppert, Wand, and Carroll (2003)]

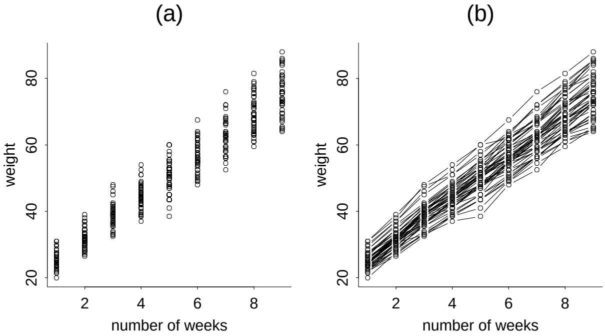


Random intercept model

$$Y_{ij} = (\beta_0 + b_{0i}) + \beta_1 \texttt{week}_j$$

- Y_{ij} = weight of ith pig at the jth week
- β_0 is the average intercept for pigs
- b_{0i} is an offset for ith pig
- So $(\beta_0 + b_{0i})$ is the intercept for the ith pig

Example III



$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\,\texttt{week}_j$$

- β_1 is the average slope
- b_{ii} is an adjustment to slope of the ith pig
- So $(\beta_1 + b_{1i})$ is the slope for the ith pig
- b_{0i} and b_{1i} seem positively correlated
  - makes sense: faster growing pigs should be larger at the start of data collection

- Model is:

$$Y_i = \mathbf{X}_i^\mathsf{T} \beta + \mathbf{Z}_i^\mathsf{T} \mathbf{b} + \epsilon_i$$

- $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ and $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iq})$ are vectors of predictor variables
- $\beta = (\beta_1, \ldots, \beta_p)$ is a vector of fixed effects
- $\mathbf{b} = (b_1, \ldots, b_q)$ is a vector of random effects
  - $\mathbf{b} \sim MVN\{0, \Sigma(\theta)\}$
  - θ is a vector of variance components

Estimation in linear mixed models

- $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are the parameter vectors
    - estimated by
        - ML (maximum likelihood), or
        - REML (maximum likelihood with degrees of freedom correction)
- $\mathbf{b}$ is a vector of random variables
    - predicted by a BLUP (Best linear unbiased predictor)
    - BLUP is shrunk towards zero (mean of $\mathbf{b}$)
    - amount of shrinkage depends on $\widehat{\boldsymbol{\theta}}$

- Random intercepts example:

$$Y_{ij} = (\beta_0 + b_{0i}) + \beta_1 \texttt{week}_j$$

- high variability among the intercepts $\Rightarrow$ less shrinkage of b_{0i} towards 0
  - extreme case: intercepts are fixed effects
- low variability among the intercepts $\Rightarrow$ more shrinkage
  - extreme case: common intercept (a simpler fixed effects model)

Comparison between fixed and random effects modeling

- fixed effects models allow only the two extremes:
  - no shrinkage
  - common intercept
- mixed effects modeling allows all possibilities between these extremes

- polynomials are excellent for local approximation of functions
- in practice, polynomials are relatively poor at global approximation
- a spline is made by joining polynomials together
  - takes advantage of polynomials' strengths without inheriting their weaknesses
- splines have "maximal smoothness"

"Positive part" notation:

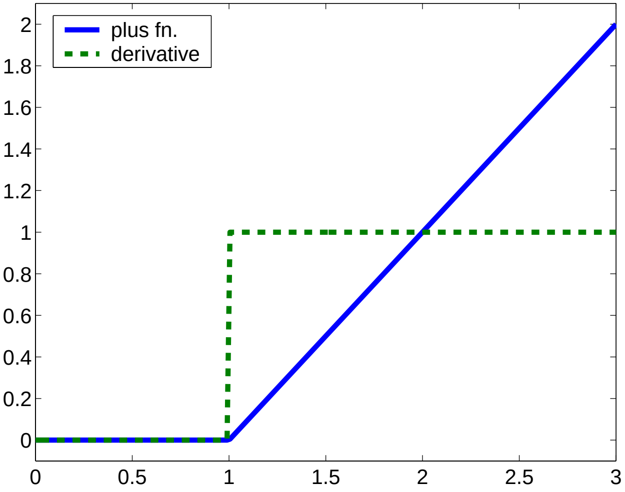
$$x_+ = x, \text{ if } x > 0 \tag{1}$$
$$= 0, \text{ if } x \le 0 \tag{2}$$

Linear spline:

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

- $\kappa_1, \ldots, \kappa_K$ are "knots"
- $b_1, \ldots, b_K$ are the spline coefficients

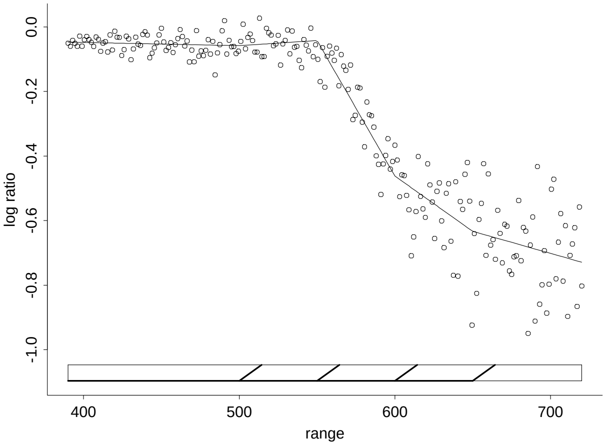
Linear “plus” function with $\kappa = 1$



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

- slope jumps by b_k at κ_k, $k = 1, \ldots, K$

Fitting LIDAR data with plus functions

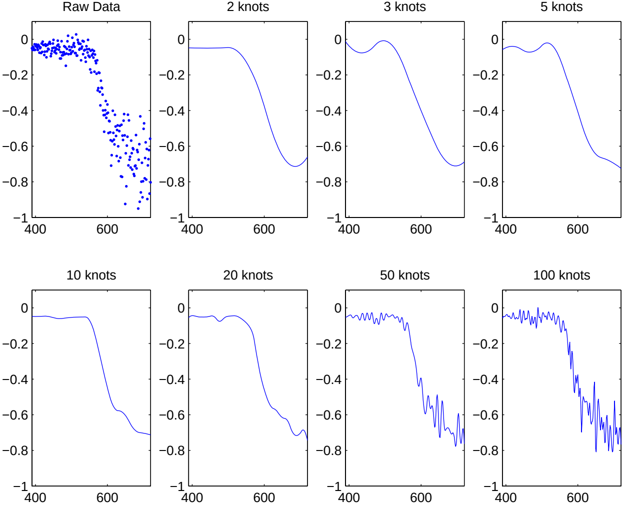


Generalization: higher degree splines

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

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

LIDAR data: ordinary Least Squares



- Use matrix notation:

$$\begin{aligned} m(X_i) &= \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i^p \\ &\quad + b_1 (X_i - \kappa_1)_+^p + \cdots + b_K (X_i - \kappa_K)_+^p \\ &= \mathbf{X}_i^\mathsf{T} \boldsymbol{\beta}_X + \mathbf{B}^\mathsf{T}(X_i)\mathbf{b} \end{aligned}$$

- Minimize

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

- From previous slide: minimize

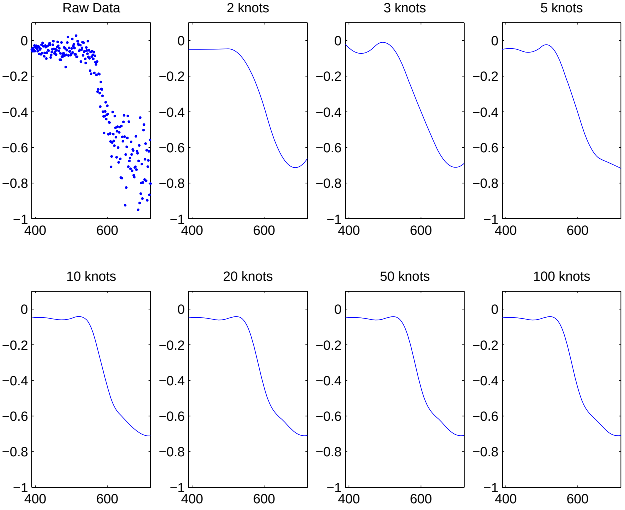
$$\sum_{i=1}^{n} \left\{ Y_i - (\mathbf{X}_i^\mathsf{T} \boldsymbol{\beta}_X + \mathbf{B}^\mathsf{T}(X_i)\mathbf{b}) \right\}^2 + \lambda\, \mathbf{b}^\mathsf{T}\mathbf{D}\mathbf{b}.$$

- $\lambda\, \mathbf{b}^\mathsf{T}\mathbf{D}\mathbf{b}$ is a penalty that prevents overfitting
- $\mathbf{D}$ is a positive semidefinite matrix
    - so the penalty is non-negative
    - Example:

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

- λ controls that amount of penalization
- the choice of λ is crucial

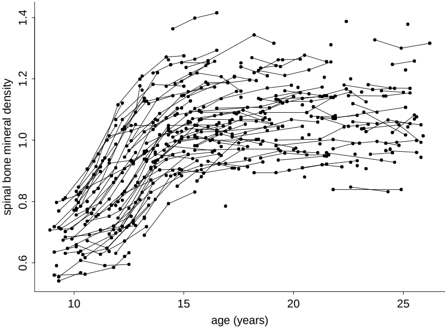
Penalized Least Squares



To choose λ use:

1. one of several model selection criteria:
   - cross-validation (CV)
   - generalized cross-validation (GCV)
   - AIC
   - C_P
2. ML or REML in mixed model framework
   - convenient because one can add other random effects
   - also can use standard mixed model software
3. Bayesian MCMC

Return to spinal bone mineral density study



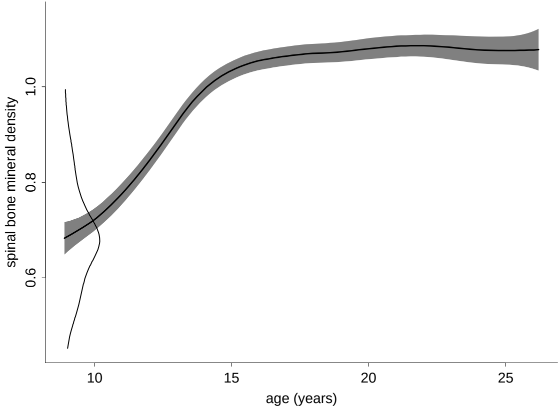
$$\texttt{SBMD}_{i,j} = U_i + m(\texttt{age}_{i,j}) + \epsilon_{i,j},$$

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

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

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

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



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

Modeling the blood lead and IQ data

For the jth measurements on the ith subject:

$$\mathrm{IQ}_{ij} = m(\mathrm{lead}_{ij}) + b_i + \beta_1 X_{ij}^1 + \cdots + \beta_L X_{ij}^L + \epsilon_{ij}$$

- $m(\cdot)$ is a spline
  - include the population average intercept
- b_i is a random subject-specific intercept
  - $E(b_i) = 0$
  - model assumes parallel curves
  - b_i models within-subject correlation
- X_{ij}^ℓ is the value of the ℓth confounder, $\ell = 1, \ldots, L$

Summary (overview of semiparametric regression)

- Semiparametric philosophy
  - use nonparametric models where needed
  - but only where needed
- LMMs and GLMMs are fantastic tools, but (apparently) totally parametric
- By basis expansion, LMMs and GLMMs become semiparametric
- Low-rank splines eliminate computational bottlenecks
- Smoothing parameters can be estimated as ratios of variance components

A smoother is linear if:

$$\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

- $\mathbf{Y}$ is the data vector
- $\widehat{\mathbf{Y}}$ contains the fitted values
- $\mathbf{H}$ is the smoother (or hat) matrix and does not depend on $\mathbf{Y}$

Note that

$$\widehat{Y}_i = \sum_{j=1}^{n} H_{i,j} Y_j$$

- $(H_{i,1}, \ldots, H_{i,n})$ [the ith row of $\mathbf{H}$] can be viewed as the finite sample kernel for estimation of $E(Y_i|X_i) = f(X_i)$

Nadaraya-Watson kernel estimator

$$\widehat{f}(X_i) = \frac{(nh_n)^{-1}\sum_{j=1}^{n} Y_j K\Big\{(X_j - X_i)/h_n\Big\}}{(nh_n)^{-1}\sum_{j=1}^{n} K\Big\{(X_j - X_i)/h_n\Big\}}$$

- $K(\cdot)$ is the kernel—it is symmetric about 0
- h_n is the bandwidth and $h_n \to 0$ as $n \to \infty$
- The denominator is a kernel density estimator
- Many smoother are asymptotically equivalent to a N-W estimator.
    - Then we want to find the “equivalent kernel” and “equivalent bandwidth” of penalized splines
- The “equivalent kernel” and “equivalent bandwidth” can be used to compare different estimators, for example, splines, kernel regression, and local regression

K is an mth order kernel if

$$\begin{aligned} \int y^k K(y) dy &= 1 \quad \text{if } k = 0 \\ &= 0 \quad \text{if } 0 < k < m \\ &\neq 0 \quad \text{if } k = m \end{aligned}$$

- m must be even because K is symmetric so that all odd moments are zero.
- $m = 2$ if K is nonnegative. Example: local linear regression

Assume that $X_1, \ldots, X_n$ are iid uniform(0,1). Then for the numerator we have

$$E\left[(nh_n)^{-1}\sum_{j=1}^{n} Y_j K\left\{h_n^{-1}(X_j - X_i)\right\}\right] =$$

$$(nh_n)^{-1}\sum_{j=1}^{n} f(X_j) K\left\{h^{-1}(X_j - X_i)\right\} \approx$$

$$h_n^{-1}\int f(x) K\left\{h^{-1}(x - X_i)\right\} dx =$$

$$\int f(x - h_n z) K(z) dz \approx$$

$$f(x) + h_n^m f^{(m)}(x) \int z^m K(z) dz$$

- The bias is of order $O(h_n^m)$ as $n \to \infty$

Framework for large-sample theory of penalized splines

- p-degree spline model:

$$f(x) = \sum_{k=1}^{K+p} b_k B_k(x), \;\; x \in (0, 1)$$

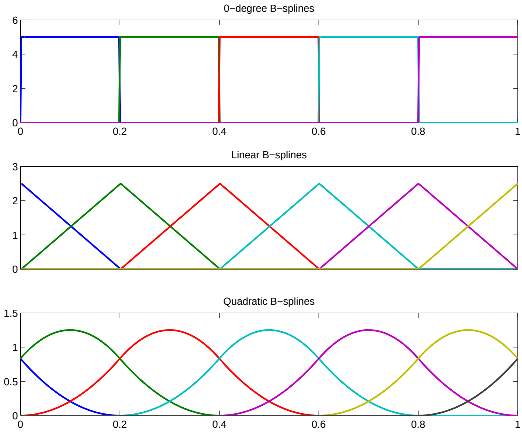
- pth degree B-spline basis:

$$\{B_k(x) : k = 1, \ldots, K + p\}$$

- knots:

$$\kappa_0 = 0 < \kappa_1 < \ldots < \kappa_K = 1$$

B-splines



0−degree B−splines
Linear B−splines
Quadratic B−splines
0
0.2
0.4
0.6
0.8
1
6
4
2
3
1
1.5
0.5

- Penalized spline estimators are approximately binned Nadaraya-Watson kernel estimators
- The order of the N-W kernel depends solely on m (order of penalty)
  - this was surprising to us
- order of kernel is $2m$ in the interior
- order is m at boundaries

Summary of main results, continued

- The spline degree p does not affect the asymptotic distribution, but
- p determines the type of binning and the minimum rate at which $K \to \infty$
- $p = 0 \Rightarrow$ usual binning
- $p = 1 \Rightarrow$ linear binning

Summary of main results, continued

- a higher value of p means that less knots are needed
  - because there is less modeling bias
  - modeling bias $=$ binning bias
- The rate at which $K \to \infty$ has no effect
  - except that it must be above a minimum rate

- Penalized least-squares minimizes

$$\sum_{i=1}^{n}\left\{y_i - \sum_{k=1}^{K+p}\widehat{b}_k B_k(x_i)\right\}^2 + \lambda \sum_{k=m+1}^{K+p}\{\Delta^m(\widehat{b}_k)\}^2,$$

- $\Delta b_k = b_k - b_{k-1}$ and $\Delta^m = \Delta(\Delta^{m-1})$
  - $m = 1 \Rightarrow$ constant functions are unpenalized
  - $m = 2 \Rightarrow$ linear functions are unpenalized

Assume:

- $x_1 = 1/n, x_2 = 2/n, \ldots, x_n = 1$
- $\kappa_0 = 0, \kappa_1 = 1/K, \kappa_2 = 2/K, \ldots, \kappa_K = 1$
- assume that $n/K := M$ is an integer

$$(\underbrace{B_p^T B_p}_{\text{B}-\text{splines}} + \underbrace{\lambda}_{\to\infty} \underbrace{D_m^T D_m}_{\text{penalty}})\,\hat{\mathbf{b}} = \underbrace{\left(B_p^T \mathbf{Y}\right)}_{\text{binned } Y\text{s}}$$

After dividing by M:

$$(\underbrace{\Sigma_p}_{O(1)} + \underbrace{\lambda}_{\to\infty} \underbrace{D_m^T D_m}_{O(1)})\,\hat{\mathbf{b}} = \underbrace{\left(B_p^T \mathbf{Y}/M\right)}_{\text{bin averages: } O_P(1)}$$

From previous slide:

$$(\underbrace{\Sigma_p}_{O(1)} + \underbrace{\lambda}_{\to\infty} \underbrace{D_m^T D_m}_{O(1)})\,\hat{\mathbf{b}} = \underbrace{\left(B_p^T \mathbf{Y}/M\right)}_{\text{bin averages: } O_P(1)}$$

We will (approximately) invert $\Sigma_p + \lambda D_m^T D_m$ (symmetric, banded)

Inverting: $\Sigma_p + \lambda D_m^T D_m$

$$q := \max(m, p) = (\#\text{ of bands above diagonal})$$

Typical column of $\Sigma_p + \lambda D_m^T D_m$ is

$$(0, \cdots, 0, \omega_q, \cdots, \omega_1, \omega_0, \omega_1, \cdots, \omega_q, 0, \cdots, 0)^{\mathsf{T}}$$

The polynomial that determines the asymptotic distribution

Define $P(x)$ as

$$P(x) = \omega_q + \omega_{q-1}x + \cdots + \omega_0 x^m + \cdots + \omega_{q-1}x^{2q-1} + \omega_q x^{2q}.$$

ρ is a root of $P(x)$ and

$$\mathbf{T}_i(\rho) = \left(\rho^{|1-i|}, \cdots, \rho, 1, \rho, \cdots, \rho^{|K-i|}\right)$$

$\mathbf{T}_i(\rho)$ orthogonal to columns of $(\Sigma_p + \lambda D_m^T D_m)$ except

- first and last q
- jth such that $|i-j| \leq q$

We can find $a_1, \ldots, a_q$ and $\rho_1, \ldots, \rho_q$ such that

$\mathbf{S}_i = \sum_{k=1}^{q} a_k \mathbf{T}_i(\rho_k)$ is orthogonal to all columns of

$(\Sigma_p + \lambda D_m^T D_m)$ except

1. ith
2. first and last q

For each k, $\rho_k \to 0$ or $|\rho_k| \uparrow 1$ sufficiently slowly, so $\mathbf{S}_i$ is asymptotically orthogonal to all columns except the ith

From earlier slides:

$$(\Sigma_p + \lambda D_m^T D_m)\,\hat{\mathbf{b}} = \left(B_p^T \mathbf{Y}/M\right)$$

$\mathbf{S}_i$ is (asymptotically) orthogonal to all columns of $(\Sigma_p + \lambda D_m^T D_m)$ except the ith

Therefore

$$b_i \approx \mathbf{S}_i \left(B_p^T \mathbf{Y}/M\right)$$

$\therefore$ $\mathbf{S}_i$ is (almost) the finite-sample kernel

Need explicit expression for $\mathbf{S}_i$, which depends on the roots of:

$$P(x) = \omega_q + \omega_{q-1}x + \cdots + \omega_0 x^m + \cdots + \omega_{q-1}x^{2q-1} + \omega_q x^{2q}.$$

- No roots have $|\rho| = 1$ or $\rho = 0$.
- If ρ is a root, then so is ρ^{-1}.
- So roots come in pairs (ρ, ρ^{-1}).
- q roots have $|\rho| < 1$

- For previous page: q roots have $|\rho| < 1$
  - of these, m of them converges upwards to 1
  - if $q > m$, then $q - m$ of them converge to 0

Asymptotic kernel (interior): m even

If m is even, then

$$H_m(x) = \sum_{i=1}^{m/2} \left\{ \frac{\alpha_{2i}}{m} \exp(-\alpha_{2i}|x|) \cos(\beta_{2i}|x|) \right.$$

$$\left. + \frac{\beta_{2i}}{m} \exp(-\alpha_{2i}|x|) \sin(\beta_{2i}|x|) \right\}$$

$\alpha_k + \beta_k\sqrt{-1}$, $i = 1, \ldots, m$, are

- roots of $x^{2m} + (-1)^m = 0$
- with $\alpha_i > 0$ (so magnitude > 1)

If m is odd, then

$$H_m(x) = \frac{\exp(-|x|)}{2m} + \sum_{i=1}^{\frac{m-1}{2}} \left\{ \frac{\alpha_{2i}}{m} \exp(-\alpha_{2i}|x|) \cos(\beta_{2i}|x|) \right.$$

$$\left. + \frac{\beta_{2i}}{m} \exp(-\alpha_{2i}|x|) \sin(\beta_{2i}|x|) \right\}$$

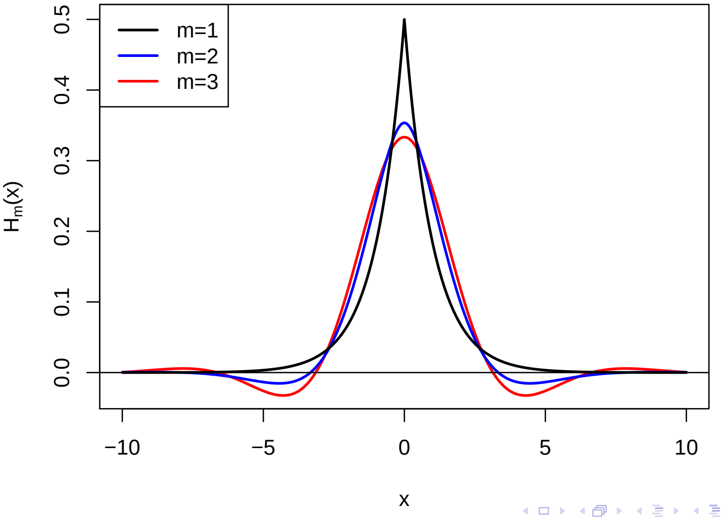
For any m:

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

and

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

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



Under assumptions (later) for any $x \in (0,1)$ [interior points], we have

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

where

$$\tilde{\mu}(x) = \frac{1}{(2m)!}\mu^{(2m)}(x)h_0^{2m}\int t^{2m}H_m(t)dt$$

and

$$V(x) = h_0^{-1}\sigma^2(x)\int H_m^2(t)dt$$

Assumptions

- The assumptions of the theorem confirm some folklore

- Folklore:

  Number of knots not important, provided large enough.

- Confirmation:

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

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

- Folklore:

  Value of the penalty parameter crucial.

- Confirmation:

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

- Folklore:

  Modeling bias small.

- Confirmation:

  Modeling bias does not appear in asymptotic bias

Comparison with Local Polynomial Regression

- Local polynomial regression (odd degree):
  - same order kernel at boundary and interior
  - order = degree + 1
- Penalized spline estimation
  - Kernel order lower at boundary
  - $2m$ in interior and m in boundary region
- Why haven't we noticed serious problems when using splines?

Comparison with Local Linear Regression

Let's look at the choices most used in practice

Local linear:

- 2nd order kernel everywhere

Penalized spline with $m = 2$:

- 2nd order kernel at boundary
- 4th order kernel in interior

- Wang, Shen, and Ruppert (2011, *EJS*) obtain the asymptotic kernel using Green's function
- Luo, Li, and Ruppert (2010, arXiv) show that a bivariate P-spline is asymptotically equivalent to a N-W estimator with a product kernel
  - they introduce a modified penalty to obtain this result
  - the new penalty also allows a much faster algorithm

Thanks for your attention