Large Sample Theory of Penalized Splines

David Ruppert

Operations Research and Information Engineering
Cornell University

(Joint work with Yingxing Li)

July 30, 2007

Main Results

- P-spline estimators are approximately binned Nadaraya-Watson kernel estimators
- The number of knots is not important, provided it is grows fast enough (confirms folklore)
- The degree of the spline does not affect the rate of convergence (surprising to me)
- Order of equivalent kernel and rate of convergence of estimator depend on the order of the penalty

Univariate nonparametric regression

- Assume a univariate nonparametric model

$$y_i = f(x_i) + \epsilon_i, \ \ i = 1, \ldots, n$$

Splines

- A spline is a piecewise polynomial
  - its polynomial form changes at "knots"
  - $p-1$ derivatives are continuous at knots
  - pth derivative jumps at knots
  - nonparametric if the knots become dense

Three methods of spline fitting

- Regression splines
  - fit by ordinary (unweighted) least squares
- Smoothing splines
  - knot at each unique value of x
  - excessive number of knots can be a problem with more complex models
- Penalized splines
  - knots, degree, and penalty chosen independently

Why Penalized Splines Have Become Popular

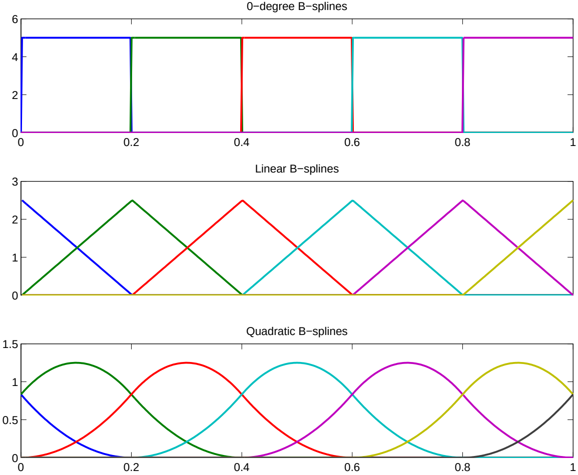
- Reasonably easy to understand
- Work well in practice
- Combines nicely with parametric models to form semiparametric models, e.g., the partially linear model

$$y_i = m(x_i) + \boldsymbol{\beta}^\mathsf{T}\mathbf{z}_i + \epsilon_i$$

Why Penalized Splines Have Become Popular

- Can be fit using parametric statistical software
  - by MCMC using, say, WinBUGS
  - by mixed model software
  - using Matt Wand's "SemiPar" package in R (a front-end to R's mixed model software)
  - for generalized regression by GLMM software

B-splines


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

p-degree spline model

- The model:

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

- pth degree B-spline basis:

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

- knots:

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

Penalized least-squares

- Penalized least-squares minimizes

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

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

$p = 0, \ m = 1$

Assume:

- $x_1 = 1/n, x_2 = 2/n, \ldots, x_n = 1$
- $\kappa_0 = 0, \kappa_1 = 1/K, \kappa_2 = 2/K, \ldots, \kappa_K = 1$

$p = 0, \; m = 1$

Assume:

- $x_1 = 1/n, x_2 = 2/n, \ldots, x_n = 1$
- $\kappa_0 = 0, \kappa_1 = 1/K, \kappa_2 = 2/K, \ldots, \kappa_K = 1$
- $B_k(x) = I\{\kappa_{k-1} < x \leq \kappa_k\}, \; 1 \leq k \leq K$

$p = 0, \; m = 1$

Assume:

- $x_1 = 1/n, x_2 = 2/n, \dots, x_n = 1$
- $\kappa_0 = 0, \kappa_1 = 1/K, \kappa_2 = 2/K, \dots, \kappa_K = 1$
- $B_k(x) = I\{\kappa_{k-1} < x \le \kappa_k\}, \; 1 \le k \le K$
- assume that n/K is an integer

$p = 0, \ m = 1$

Assume:

- $x_1 = 1/n, x_2 = 2/n, \ldots, x_n = 1$
- $\kappa_0 = 0, \kappa_1 = 1/K, \kappa_2 = 2/K, \ldots, \kappa_K = 1$
- $B_k(x) = I\{\kappa_{k-1} < x \leq \kappa_k\}$, $1 \leq k \leq K$
- assume that n/K is an integer
- then $X^{\mathsf{T}}X = MI_K$ where I_K

$p = 0, \; m = 1$, continued

Assume further:

- $m = 1$

Then

$$D^{\mathsf{T}} D = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix},$$

$p = 0, \ m = 1$, PLS estimator

The PLS estimator solves:

$$\Lambda \widehat{\mathbf{b}} = \mathbf{z} = \overline{\mathbf{y}}/(1+2\lambda)$$

where

$$\Lambda = \begin{pmatrix} \theta & \eta & 0 & 0 & \cdots & 0 & 0 \\ \eta & 1 & \eta & 0 & \cdots & 0 & 0 \\ 0 & \eta & 1 & \eta & \cdots & 0 & 0 \\ 0 & 0 & \eta & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & \eta \\ 0 & 0 & 0 & 0 & \cdots & \eta & \theta \end{pmatrix}, \quad \eta = -\frac{\lambda}{1+2\lambda}$$

$p = 0, \ m = 1$, PLS estimator, page 2

Let $\rho \in (0, 1)$ be a root of

$$\eta + \rho + \eta\rho^2 = 0.$$

Then

$$\rho = \frac{1 - \sqrt{1 - 4\eta^2}}{-2\eta} = \frac{1 + 2\lambda - \sqrt{1 + 4\lambda}}{2\lambda}.$$

Define

$$T_i = (\rho^{i-1}, \rho^{i-2}, \ldots, \rho, 1, \rho, \rho^2, \ldots, \rho^{K-i})^{\mathsf{T}}$$

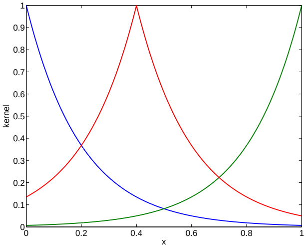
T_i is orthogonal to all columns of Λ except the first, last, and ith

Finite-sample kernel

Finite-sample kernel:

$$\widehat{f}(x) = \sum_{j=1}^{K} H(x, \overline{x}_j)\overline{y}_i$$

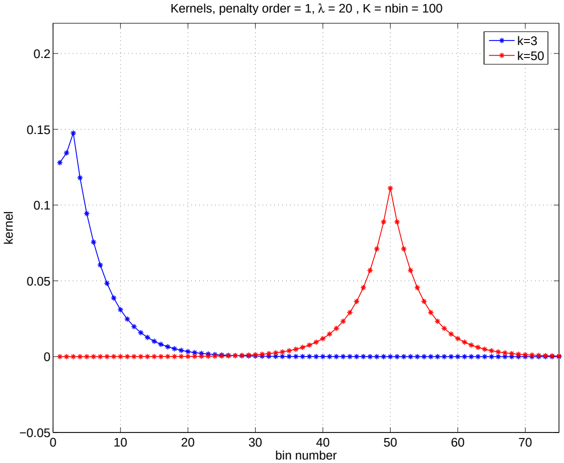
Three kernels corresponding to first-order penalty



- x is an "estimation point" (here fixed at 0.4)

- finite-sample kernel is linear combination of three kernels
- double exponential kernel centered at x
- boundary kernels are $\exp(-x)$ and $\exp(x)$
- weights for the boundary kernels are asymptotically negligible
- all kernels can be re-scaled by a bandwidth

Finite-sample kernels, first-order penalty



Connection with smoothing splines

We get the same equivalent kernel as for smoothing splines with a penalty on the first derivative

Finding $\widehat{b}_i$ – interior case

- Suppose $i/K \to x \in (0,1)$ (non-boundary case)
- After some algebra:

$$\widehat{b}_i \sim \frac{\sum_{j=1}^{K} \rho^{|i-j|}\overline{y}_j}{\sum_{j=1}^{K} \rho^{|i-j|}}.$$

- Note that

$$\widehat{f}(x) = \widehat{b}_i$$

for x in the ith bin

Equivalence to N-W kernel estimator

- After some more algebra

$$\rho^{|i-j|} \sim \exp\left\{-\frac{|\overline{x}_i - \overline{x}_j|}{hn^{-1/5}}\right\}$$

- Thus, $\widehat{f}_n$ is asymptotically equivalent to the Nadaraya-Watson estimator with
  - double exponential kernel $H(x) = (1/2)\exp(-|x|)$
  - bandwidth $hn^{-1/5}$

Nadaraya-Watson kernel estimators

Model:

$$y_i = f(x_i) + \epsilon_i$$

Nadaraya-Watson estimator:

$$\widehat{f}(x) = \frac{\sum_{i=1}^{n} H\{(x_i - x)/h_n\} y_i}{\sum_{i=1}^{n} H\{(x_i - x)/h_n\}}$$

- $H(\cdot)$ is called the kernel function
- h_n is the bandwidth

Binned Nadaraya-Watson kernel estimators

Binned Nadaraya-Watson estimator:

- range of the x_i divided into K subintervals (bins)
- $\overline{x}_j$ is average of x_i in ith bin
- $\overline{y}_j$ is average of y_i such that x_i is in the ith bin

$$\widehat{f}(x) = \frac{\sum_{j=1}^{K} H\{(\overline{x}_j - x)/h_n\}\overline{y}_j}{\sum_{j=1}^{K} H\{(\overline{x}_i - x)/h_n\}}$$

P-spline equivalent to a Nadaraya-Watson kernel estimator

- Thus, $\widehat{f}_n$ is asymptotically equivalent to a binned Nadaraya-Watson estimator with
  - double exponential kernel $H(x) = (1/2)\exp(-|x|)$
  - bandwidth $hn^{-1/5}$
- binned bias is negligible if $K = Cn^{\gamma}$ for $\gamma > 2/5$ and $C > 0$
- "negligible" means $o(n^{-2/5})$

Selecting λ to achieve desired bandwidth

- To get bandwidth $hn^{-1/5}$ we need λ chosen as

$$\lambda \sim \{(Cn^{\gamma})(hn^{-1/5})\}^2 = (\#\text{ knots } \times \text{ bandwidth})^2$$

Asymptotic Distribution

For $x \in (0,1)$, as $n \to \infty$ we have

$$n^{2/5}\{\hat{f}_n(x) - f(x)\} \Rightarrow N\{\mathcal{B}(x), \mathcal{V}(x)\}$$

where

- $\mathcal{B}(x) = h^2 f^{(2)}(x)$
- $\mathcal{V}(x) = 2^{-1} h^{-1} \sigma^2(x)$

Some folklore

- Folklore: The number of knots is not important, provided that it is large enough.
  - Confirmation:

$$K \sim Cn^{\gamma} \text{ with } C > 0 \text{ and } \gamma > 2/5. \tag{1}$$

- Folklore: The value of the penalty parameter is crucial.
  - Confirmation:

$$\lambda \sim C^2 h^2 n^{2\gamma - 2/5} = (\# \text{ knots } \times \text{ bandwidth})^2 \tag{2}$$

    for some $h > 0$.

- Folklore: Modeling bias is small.
  - Confirmation: Modeling bias does not appear in asymptotic bias provided (1) and (2) hold.

Order of a kernel and bias

Moments: kth moment is $\int x^k H(x)\,dx$

Order of kernel: A kernel is of kth order if the first non-zero moment is the kth

- Non-negative kernel: order is at most 2

Bias: bias $= O\{(\text{bandwidth})^k\}$

Variance:

variance $= O\left(\frac{1}{n\times\text{bandwidth}}\right)$

and

optimal RMSE $= O(n^{-k/(2k+1)})$

2nd order-penalty gives 4th order kernel (in interior)

Now let $m = 2$ (2nd order difference penalty)

- Assume:
  - $K \sim Cn^{\gamma}$ with $C > 0$ and $\gamma > 4/9$
  - $\lambda \sim 4C^4 h^4 n^{4\gamma - 4/9} \sim 4(Khn^{-1/9})^4$.

Then for any $x \in (0,1)$, when $n \to \infty$, we have

$$n^{4/9}\{\widehat{f}_n(x) - f(x)\} \Rightarrow N\{\mathcal{B}_1(x), \mathcal{V}_1(x)\},$$

where

- $\mathcal{B}_1(x) = (1/24)h^4 f^{(4)}(x) \int x^4 T(x) dx$
- $\mathcal{V}_1(x) = h^{-1} \left\{ \int T^2(x) dx \right\} \sigma^2(x)$
  - $T(x)$ is a fourth order kernel

Mathematical approach

Main technical device uses roots of the polynomial

$$w(\xi) = \lambda(1-4\xi+6\xi^2-4\xi^3+\xi^4)+\xi^2 = \lambda(1-\xi)^4+\xi^2, \ \lambda > 0$$

- No real roots and no roots of modulus one
- Roots are: r, conj(r), r^{-1}, conj$(r)^{-1}$ (all distinct)
- Use the conjugate pair with modulus less than one

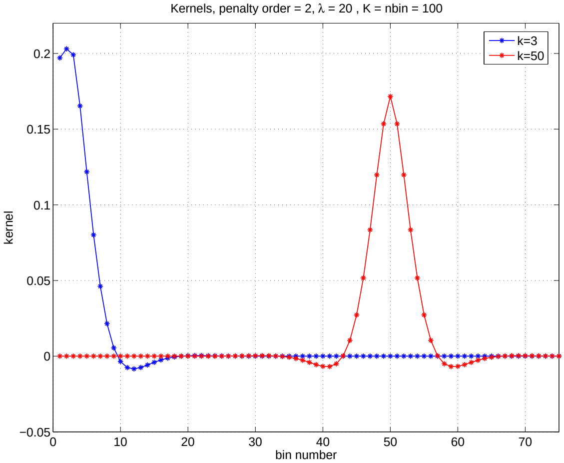
Asymptotic Kernel

- The asymptotic kernel is a linear combination of

$$\exp(-|x|)\cos(x) \quad \text{and} \quad \exp(-|x|)\sin(|x|)$$

- Same equivalent kernel as for smoothing splines with a penalty on the second derivative

Finite-sample kernels, second-order penalty



Linear splines need less knots

Assume $m = 1$ (1st-order difference penalty).

- If $p = 1$ (linear), then require
$$K \sim Cn^{\gamma} \text{ with } C > 0 \text{ and } \gamma > 1/5$$
- When p was 0 (piecewise constant), we required $\gamma > 2/5$
- Otherwise, results are the same as for 0-degree and linear splines

A similar result holds for $m = 2$.

Conjectures

- Conjecture: For x in the interior:

  $$\text{P-spline} \sim \text{N-W estimator with an } 2m\text{-order kernel}$$

  - Recall: m is order of difference penalty
  - Kernel order independent of $p =$ degree of spline
  - Shown to hold for $m = 1, 2$ and $p = 0, 1$
- $p = 1$ requires less knots than $p = 0$
  - What happens for $p > 1$?
  - Conjecture: Still less knots are needed

Unequally spaced X

- Assume $G(x_t) = t/n$ for a smooth G with $g = G'$
- Fit a spline to (Y_t, u_t) with regression function $f \circ G^{-1}$
  - evaluate this estimate at $G(x)$ to estimate $f(x)$
- Equally spaced knots for (Y_t, u_t) implies knots at sample quantiles for (Y_t, x_t)
- asymptotic bias is

$$h^2(f \circ G^{-1})^{(2)}\{G(x)\} = \frac{h^2}{g^2(x)}\left\{f^{(2)}(x) - \frac{f'(x)g'(x)}{g(x)}\right\}$$

- Nadaraya-Watson bias is

$$h^2\left\{f^{(2)}(x) + \frac{2f'(x)g'(x)}{g(x)}\right\}$$

Additive Models

Talk by Yingxing Li on bivariate additive model:

- Session 543
- Thursday morning

We use only one of two potential smoothing parameters

Both K and λ are potential smoothing parameters

- In our asymptotic theory, only λ plays the role of a smoothing parameter
- Could develop a theory where only K plays this role
  - would be similar to regression spline $(\lambda = 0)$ theory
- One could also choose K and λ so that both have a non-negligible effect
- Our theory mimics actual practice

Summary

- P-spline estimators ≈ binned N-W kernel estimators
- The number of knots unimportant if above a minimum
- Degree of spline
  - determines minimum convergence rate for number of knots
  - does not affect rate of convergence
- Order of penalty determines
  - order of equivalent kernel
  - convergence rate of estimator
- mth order penalty ⇔ smoothing spline with penalty on mth difference

Thanks for coming