Some Asymptotic Results on Penalized Spline Smoothing

Tatyana Krivobokova

Colloquium – Department of Statistics and Probability
Michigan State University, August 11, 2009

Outline

Based on (Y_i, x_i), $x_i \in [a, b]$, $i = 1, ..., n$ with true relationship

$$Y_i = f(x_i) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

we aim to estimate $f(\cdot) \in W^{p+1}[a, b]$.

Spline-based methods

- Regression splines
- Smoothing splines
- Penalized splines

One chooses

- some spline basis functions $N_i(\cdot)$ of degree p
- based on a set of l knots $\kappa_1, \ldots, \kappa_l$

and finds $\hat{f}_{\mathrm{reg}}(\cdot) = N_l(\cdot)\hat{\beta}$ solving

$$\min_{\beta} \sum_{i=1}^{n} \{Y_i - N_l(x_i)\beta\}^2.$$

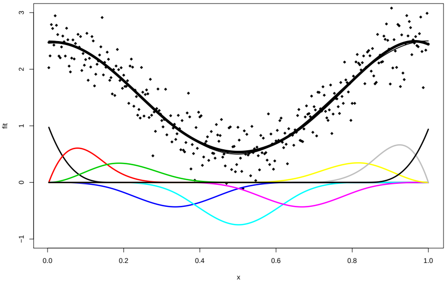
The resulting estimate is the LSE

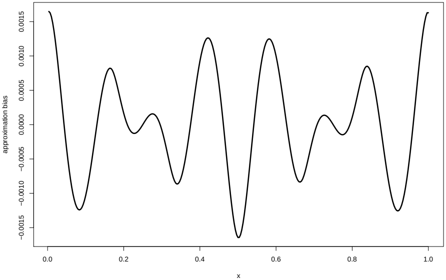
$$\hat{f}_{\mathrm{reg}}(\cdot) = N_l(\cdot)(N_l^T N_l)^{-1} N_l^T Y,$$

with N_l as a $n \times l$ dimensional spline basis matrix (e.g. B-splines), and $N_l(x_i)$ as the row vector of N_l evaluated at x_i.

Regression spline estimator

+ optimal rate of convergence
+ low parameter dimension
+ no boundary effects
– number and placements of knots problem





A $2q - 1$ degree smoothing spline $\hat{f}_{\mathrm{spl}}$ is the minimizer of

$$\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b \{f(x)^{(q)}\}^2 dx,$$

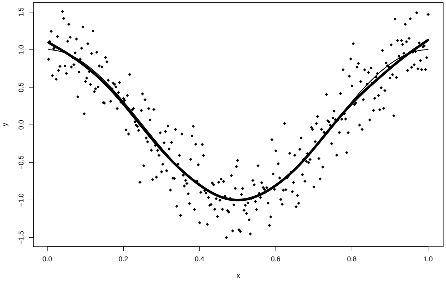
for $f(\cdot) \in W^q[a, b]$ and can be written as

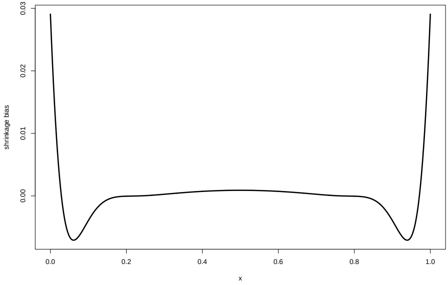
$$\hat{f}_{\mathrm{spl}}(\cdot) = N_n(\cdot)(N_n^T N_n + \lambda_n D_n)^{-1} N_n^T Y,$$

with N_n as a $n \times n$ natural $(2q - 1)$-degree spline model matrix, corresponding penalty matrix D_n and λ_n chosen with e.g. GCV.

Smoothing spline estimator

- $+$ no knots placement problem (knots equal observations)
- $+/-$ rate of convergence depends on the natural boundary conditions met
- $-$ high parameter dimension
- $-$ boundary effects





Between regression and smoothing splines

Choosing $l < k \ll n$ knots $\kappa_1, \dots, \kappa_k$ and solving

$$\min_{\beta} \left(\sum_{i=1}^{n} \{Y_i - N_k(x_i)\beta\}^2 + \lambda \int_a^b \left[\{N_k(t)\beta\}^{(q)} \right]^2 dt \right),$$

result in

$$\hat{f}_{\text{pen}}(\cdot) = N_k(\cdot)(N_k^T N_k + \lambda_k D_k)^{-1} N_k^T Y$$

with N_k as some $n \times k$ dimensional p-degree spline basis matrix, D_k as the corresponding penalty and λ_k chosen with e.g. GCV.

Penalized spline estimator

- $+$ no knots placement problem
- $+$ low parameter dimension
- $+$ flexible choice of bases and penalties
- $+$ links to mixed and Bayesian models
- ? asymptotic properties are not explored

Some first results

Hall and Opsomer (Biometrika, 2005)
Li and Ruppert (Biometrika, 2008)
Kauermann, Krivobokova and Fahrmeir (JRSSB, 2009)
Claeskens, Krivobokova and Opsomer (Biometrika, 2009)

Optimal rate of convergence

Stone (Ann. Statist., 1982):
For any nonparametric estimator $\hat{f}$ of $f \in C^{p+1}[a,b]$ the optimal rate of convergence for $\|\hat{f} - f\|_{L_q}$, $0 < q < \infty$ is

$$n^{-\frac{2p+2}{2p+3}}$$

Smoothing technique	Control parameter	Optimal order
Regression splines	number of knots	$k \sim C_1 n^{\frac{1}{2p+3}}$
Smoothing splines	smoothing parameter	$\lambda \sim C_2 n^{\frac{1}{2p+3}}$
Penalized splines	number of knots &	$k \sim ?$
	smoothing parameter	$\lambda \sim ?$

Two asymptotic scenarios

For a penalized spline estimator $\hat{f}_{\mathrm{pen}} = N(N^T N + \lambda D_q)^{-1} N^T Y$

$$AMSE(\hat{f}_{\mathrm{pen}}) = \begin{array}{l}\text{average}\\ \text{variance}\end{array} + \begin{array}{l}\text{average squared}\\ \text{shrinkage bias}\end{array} + \begin{array}{l}\text{average squared}\\ \text{approximation bias}\end{array}$$

and

$$K_q^{2q} = \text{maximum eigenvalue of } \lambda(N^T N)^{-1} D_q$$

defines the breakpoint between two asymptotic scenarios

- $K_q < 1$ leads to the regression splines type asymptotics
- $K_q \geq 1$ leads to the smoothing splines type asymptotics

Asymptotic scenarios with $K_q < 1$

For $K_q < 1$ and

$$k \sim C_1 n^{\frac{1}{2p+3}} \text{ and } \lambda = O\left(n^{\gamma}\right), \; \gamma \leq \frac{p+2-q}{2p+3}$$

we find

- $\hat{f}_{\mathrm{pen}}(\cdot)$ converges to $f(\cdot)$ with $n^{-\frac{2p+2}{2p+3}}$
- Average approximation and shrinkage bias are of the same order
- Asymptotic order of k is the same as for regression splines
- Shrinkage bias becomes negligible for small λ

Asymptotic scenarios with $K_q \geq 1$

For $K_q \geq 1$, $\lambda n^{2q-1} \to \infty$ and

$$\lambda = O\left(n^{\frac{1}{2q+1}}\right) \text{ and } k \sim C_2 n^{\nu},\ \nu \geq \frac{1}{2q+1}$$

we find

- $\hat{f}_{\mathrm{pen}}(\cdot)$ converges to $f(\cdot)$ with $n^{-\frac{2q}{2q+1}} > n^{-\frac{2p+2}{2p+3}}$ for $q \leq p$
- Shrinkage bias dominates the AMSE
- Asymptotic order of k and λ depend only on q
- Average approximation bias is negligible

Pointwise bias and variance

Representing

$$\hat{f}_{\rm pen}(x) = \hat{f}_{\rm reg}(x) - \lambda N(x)(N^T N + \lambda D_q)^{-1} D_q (N^T N)^{-1} N^T Y$$

under certain assumptions one finds

$$\begin{aligned} E\{\hat{f}_{\rm pen}(x)\} - f(x) &\approx b_a(x) + b_\lambda(x) \\ Var\{\hat{f}_{\rm pen}(x)\} &\approx \frac{\sigma^2}{n} N(x)(G + \lambda D_q/n)^{-1} G (G + \lambda D_q/n)^{-1} N^t(x) \end{aligned}$$

with $G = \int_a^b N(x)^T N(x) \rho(x) dx$

Two bias components

Approximation bias

$$b_a(x) = -\frac{f^{(p+1)}(x)}{(p+1)!} \sum_{j=0}^{K} I_{[\kappa_j,\kappa_{j+1})}(x)(\kappa_{j+1}-\kappa_j)^{p+1} B_{p+1}\left(\frac{x-\kappa_j}{\kappa_{j+1}-\kappa_j}\right),$$

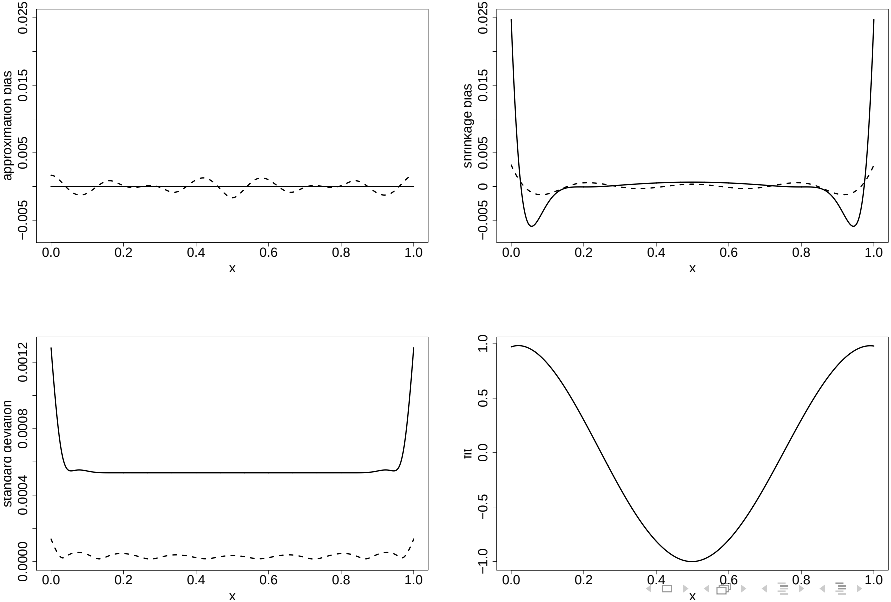
with $B_{p+1}(\cdot)$ denoting the $(p+1)$th Bernoulli polynomial.

Shrinkage bias

$$b_\lambda(x) = -\frac{\lambda}{n} N(x)(G+\lambda D_q)^{-1} D_q \beta,$$

where β is s.t. $N(\cdot)\beta$ is the best L_∞ approximation to $f(\cdot)$.

Pointwise bias and variance



Summary so far

- Penalized splines enjoy similarities to regression and smoothing splines
- K_q defines a clear breakpoint between two asymptotic scenarios
- Asymptotic scenarios with $K_q < 1$ can result in a smaller AMSE
- A guideline for choosing k is needed
- Pointwise expressions for bias and variance are available
- Equivalent kernel functions can provide more insights (ongoing work)

Mixed model representation

Representing

$$N\beta = N(N_b b + N_u u) = Xb + Zu,$$

with (N_b, N_u) is of full rank, $N_b^T N_u = N_u^T N_b = N_b^T D_q N_b = 0$, $N_u^T D_q N_u = I$ and assuming

$$Y|u \sim N(Xb + Zu, \sigma^2 I_n), \quad u \sim N(0, \sigma_u^2 I)$$

result in the linear mixed model with BLUP

$$\tilde{f}_{\text{pen}}(x) = N_m \left(N_m^T N_m + \frac{\sigma^2}{\sigma_u^2} D_m \right)^{-1} N_m^T Y$$

with $N_m = (X, Z)$ and $D_m = \text{diag}(0_{p+1}, 1_{k+p+1-q})$.

Two models

$$\hat{f}_{\text{pen}} = N(N^T N + \lambda D)^{-1} N^T Y \qquad \tilde{f}_{\text{pen}} = N\left(N^T N + \sigma^2/\sigma_u^2 D\right)^{-1} N^T Y$$

- $N\beta$ is fixed
- λ is estimated with e.g. GCV

- $N\beta \sim N(Xb, \sigma_u^2 Z Z^T)$
- σ^2/σ_u^2 is a (RE)ML estimate

It is known

- In general $\tilde{f}_{\text{pen}}$ tends to overfit f (current work)
- σ^2/σ_u^2 is very robust to the correlation misspecification (Krivobokova and Kauermann, JASA 2008)

REML and GCV based λ

We compare REML and GCV based λ for two cases

- $f \in W^q[a, b]$
- $f(x) \sim N\{X(x)b, \sigma_u^2 Z(x)Z(x)^t\}$

Define λ_{REML}, $\overline{\lambda}_{REML}$ and λ_{MSE}, $\overline{\lambda}_{MSE}$ as solutions to

$$
\begin{aligned}
E_{Y|u}\left(\frac{\partial l_p^R(\lambda)}{\partial \lambda}\right) &= 0 \quad \text{and} \quad E_{Y,u}\left(\frac{\partial l_p^R(\lambda)}{\partial \lambda}\right) = 0 \\
E_{Y|u}\left(\frac{\partial GCV(\lambda)}{\partial \lambda}\right) &= 0 \quad \text{and} \quad E_{Y,u}\left(\frac{\partial GCV(\lambda)}{\partial \lambda}\right) = 0
\end{aligned}
$$

Two smoothing parameter estimates

If $f \in W^q[a,b]$ then λ_{REML} and λ_{MSE} solve

$$
\begin{aligned}
0 &= E_{Y|u}\left(\frac{\partial l_p^R(\lambda)}{\partial\lambda}\right) = \frac{\partial AMSE(\lambda)}{\partial\lambda} + \frac{\partial b(x,\lambda)}{\partial\lambda} + o(n^{-1}) \\
0 &= E_{Y|u}\left(\frac{\partial GCV(\lambda)}{\partial\lambda}\right) = \frac{\partial AMSE(\lambda)}{\partial\lambda} + o(n^{-1}),
\end{aligned}
$$

with $b(x,\lambda) = f^t(S_\lambda - S_\lambda^2)f/n - \sigma_\epsilon^2 \mathrm{tr}(S_\lambda + S_\lambda^2)/n + \sigma_\epsilon^2 \log|VX^tV^{-1}X|/n$,
$V = I + ZZ^t/\lambda$

Two smoothing parameter estimates

Using the Taylor expansion, one obtains

$$\frac{\lambda_{REML}}{\lambda_{MSE}} = 1 + \frac{\sigma_\epsilon^2\{\mathrm{tr}(S_\lambda^2) - p - 1 + q\} - f^t(S_\lambda - S_\lambda^2)f}{\sigma_\epsilon^2 \mathrm{tr}(S_\lambda^2) - p - 1 + q} + o(1)$$

with $S_\lambda = S(\lambda_{MSE})$

With the Demmler-Reinsch decomposition $S_\lambda = A\,\mathrm{diag}(1+\lambda s)^{-1}A^t$ the numerator can be written as

$$\sigma_\epsilon^2\{\mathrm{tr}(S_\lambda^2) - p - 1 + q\} - f^t(S_\lambda - S_\lambda^2)f = \sigma_\epsilon^2 \sum_{i=1}^{k} \frac{1 - \lambda s_i c_i^2/\sigma_\epsilon^2}{(1+\lambda s_i)^2},$$

with $c = A^t f$.

Two smoothing parameter estimates

The term

$$\sigma_\epsilon^2 \sum_{i=1}^{k} \frac{1 - \lambda s_i c_i^2/\sigma_\epsilon^2}{(1+\lambda s_i)^2}$$

can be either positive, negative or zero, depending on f, σ_ϵ^2 and k

Note that $\max_i c_i/\sigma_\epsilon$ depends on the signal-to-noise ratio

Then

- for $\lambda s_1 = K_q^{2q} < 1$ and $\max_i c_i/\sigma_\epsilon < 1$ it holds $\lambda_{REML} > \lambda_{MSE}$
- if $\max_i c_i/\sigma_\epsilon < \text{tr}(S_\lambda^2)/\text{tr}(S_\lambda - S_\lambda^2)$ then $\lambda_{REML} > \lambda_{MSE}$
- for $\lambda s_1 = K_q^{2q} \geq 1$ and $k \to n$ it holds $\lambda_{REML} < \lambda_{MSE}$
- there can exist such k that $\lambda_{REML} \approx \lambda_{MSE}$

Two smoothing parameter estimates

If $f \in W^q[a,b]$ then

- REML is biased w.r.t. AMSE
- REML performance depends on k, f and σ_ϵ^2

If $f(x) \sim N\{X(x)b, \sigma_u^2 Z(x) Z(x)^t\}$ then

- $\overline{\lambda}_{REML} = \overline{\lambda}_{MSE}$ (Krivobokova and Kauermann, JASA, 2007)

Mixed models for generalized responses

For $Y_i|x_i \sim \exp\{y^T h^{-1}(x_i) - \rho\{h^{-1}(x_i)\} + c(Y_i)\}$ one models

$$E(Y|u) = h(Xb + Zu), \quad u \sim N(0, \sigma_u^2 I),$$

leading to the likelihood

$$L(b, \sigma_u^2) = \sigma_u^{-(k+p+1-q)} \int_{R^{k+p+1-q}} \exp[-g(u)]du,$$

with $g(u) = -y^T(Xb + Zu) + 1_n^T \rho(Xb + Zu) + u^T u/(2\sigma_u^2)$,

which is not available analytically and is usually solved with the Laplace approximation (Breslow & Clayton, JASA 1993)

Laplace approximation

The Laplace approximation is reliable for $n \to \infty$ and k "small" with the error term

$$\varepsilon_0 = -g_{jlrs} g^{jl} g^{rs}[3]/24 + g_{jlr} g_{stv} \left(g^{jl} g^{rs} g^{tv}[9] + g^{js} g^{lt} g^{rv}[6] \right) /72$$

It has been shown that if $k \sim C_1 n^{1/(2p+3)}$, then ε_0 is negligible (Kauermann, Krivobokova, Fahrmeir, JRSSB 2008).

Still to do: how big is ε_0 for $K_q \geq 1$?

Summary

- First asymptotic results in a unified framework
- Less knots implies less boundary effects
- Less knots implies $\lambda_{REML} \approx \lambda_{MSE}$
- More asymptotic results are needed for generalized framework
- Generalization to smoothing in R^d and its asymptotics is open