Smoothing parameter selection in two frameworks for spline estimators

Tatyana Krivobokova

Georg-August-Universität Göttingen

29th European Meeting of Statisticians

Budapest, 20 – 25 Juli 2013

Model

Nonparametric model for n data pairs (y_i, x_i)

$$Y_i = f(x_i) + \epsilon_i, \; i = 1, \ldots, n, \; \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

for fixed $x_i \in [0, 1]$ and an unknown smooth f

Smoothing parameter λ for any $\widehat{f}(x) = \widehat{f}(x; \lambda)$ can be chosen to minimize some unbiased estimator of the mean square risk

$$\mathrm{R}(\widehat{f}, f) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{f}(x_i; \lambda) - f(x_i)\right\}^2\right]$$

Choice of λ

In practice GCV (or asymptotic equivalent AIC, C_p) as $\widehat{\mathrm{R}}$ is used

$$\mathrm{E}\{\mathrm{GCV}(\lambda)\} = \mathrm{R}(\widehat{f}, f)\{1 + o(1)\} + \sigma^2 \left\{1 + o(n^{-1})\right\}$$

Known practical problems of GCV and similar criteria

- large variability of $\widehat{\lambda}$ obtained with GCV
- extremely sensitive to serial dependences in ϵ_i
- unstable in low signal-to-noise ratio situations

Two frameworks for splines

Frequentist model

$$Y_i = f(x_i) + \epsilon_i = \sum_{j=0}^{q-1} \beta_j x_i^j + \int_0^1 f^{(q)}(t) \frac{(x-t)_+^{q-1}}{(q-1)!} dt + \epsilon_i,$$

for $x_i \in [0,1], \quad f \in \mathcal{W}^q[0,1], \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Stochastic model

$$Y_i = F(x_i) + \epsilon_i = \sum_{j=0}^{q-1} \beta_j x_i^j + \sigma_u \int_0^1 \frac{(x_i-t)_+^{q-1}}{(q-1)!} dW(t) + \epsilon_i,$$

for $x_+ = \max\{0, x\}$, $W(t)$ standard Wiener process and $i = 1, \ldots, n$

Two frameworks

In the frequentist framework f is estimated from

$$\min_{f \in \mathcal{W}^q[0,1]} \left[\frac{1}{n} \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_0^1 \{f(x)^{(q)}\}^2 dx \right],$$

which is minimized by the smoothing spline estimator $\widehat{f}$

In the stochastic framework F is found as the best linear unbiased predictor, which equals to $\widehat{f}$ with $\lambda = \sigma^2/(n\sigma_u^2)$

Note that the sample paths of $F \notin \mathcal{W}^q[0,1]$ with probability 1 (are less smooth)

Low-rank splines

Let $\mathcal{S}(2q-1,k)$ be a spline space of degree $2q-1$ based on k knots τ_j

Making further assumptions on regularity of x_i and τ_j and on

$$k = \text{const } n^{\nu},\ \nu \in (1/(2q),1)\ ,\ \lambda \to 0,\ \lambda n \to \infty$$

one solves a lower dimensional problem $(k < n)$ to estimate $f \in \mathcal{W}^q$

$$\min_{s \in \mathcal{S}(2q-1,k)} \left[\frac{1}{n} \sum_{i=1}^{n} \{y_i - s(x_i)\}^2 + \lambda \int_0^1 \{s^{(q)}(x)\}^2 dx \right]$$

Similarly, the estimator in the stochastic framework is generalized

Two smoothing parameters

Estimators in both models are equal up to the smoothing parameter

Smoothing parameter λ

- relies on the frequentist model with $f \in \mathcal{W}^q[0,1]$
- estimated by minimizing criteria that estimate $\mathrm{R}(\widehat{f}, f)$

Smoothing parameter σ^2/σ_u^2

- relies on the stochastic model
- estimated by maximizing the corresponding likelihood

Frequentist model

Aim to answer

How both smoothing parameter estimators behave
if the data follow a frequentist model?

Available results

- Sun & Speckman (2000, unpublished) asymptotic distribution of the estimators (smoothing splines) for functions satisfying natural boundary conditions

Oracle smoothing parameters

Let denote

$$\lambda_f = \lambda_f(n) = \arg\min_{\lambda>0} \mathsf{E}\{\mathsf{GCV}(\lambda)\} = \arg\min_{\lambda>0} \mathsf{R}(\widehat{f}, f)\{1 + o(1)\}$$

and

$$\lambda_r = \lambda_r(n) = \arg\min_{\sigma^2/\sigma_u^2>0} \mathsf{E}_f\{-l_p(\sigma^2/\sigma_u^2; y)\}$$

with l_p as the profile (restricted) likelihood for σ^2/σ_u^2

λ_r is the smoothing parameter that one gets in the mean from the likelihood in case the data follow $Y_i = f(x_i) + \epsilon_i$, $f \in \mathcal{W}^q$

Oracle smoothing parameters

Let $f \in \mathcal{W}^{qm}[0,1]$, $m \in [1,2]$, where $\mathcal{W}^{qm}$ is a fractional order Sobolev (Besov) space with certain boundary conditions

$$\lambda_f \geq \mathrm{C}(f,q,m,\sigma^2)\, n^{-\frac{2q}{2qm+1}},\ m \in [1,2],$$

$$\lambda_r = \mathrm{C}(f,q,\sigma^2)\, n^{-\frac{2q}{2q+1}},\ \forall m$$

as shown in Wahba (1995, AoS)

- λ_f adapts to the unknown smoothness and boundary conditions (up to $2q$)
- performance of λ_r depends on n, q and $C(f,q,\sigma^2)$

Oracle smoothing parameters

It follows, that λ_r is suboptimal

- $\lambda_f/\lambda_r \to \infty$ with $n \to \infty$ for $f \in \mathcal{W}^{qm}[0,1]$, $m \in (1,2]$
- $\widehat{f}(\lambda_r)$ (asymptotically) undersmooths f compared to $\widehat{f}(\lambda_f)$

In many small-samples simulation studies (e.g. Kohn, JASA, 1991) $\widehat{\sigma}^2/\widehat{\sigma}_u^2$ appeared to perform better than $\widehat{\lambda}$

$\hookrightarrow$ look at the properties of estimators $\widehat{\sigma}^2/\widehat{\sigma}_u^2$ and $\widehat{\lambda}$

Smoothing parameter estimators

Under the frequentist model and mentioned assumptions on x_i, τ_i, k, λ

$$\frac{\widehat{\sigma}^2/\widehat{\sigma}_u^2}{\lambda_r} \xrightarrow{\mathcal{P}} 1 \quad \text{and} \quad \frac{\widehat{\lambda}}{\lambda_f} \xrightarrow{\mathcal{P}} 1.$$

Moreover,

$$\lambda_r^{-1/(4q)} \left(\frac{\widehat{\sigma}^2/\widehat{\sigma}_u^2}{\lambda_r} - 1 \right) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, 2C_1(q)\Big)$$

and

$$\lambda_f^{-1/(4q)} \left(\frac{\widehat{\lambda}}{\lambda_f} - 1 \right) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, 2C_2(q)\Big),$$

Smoothing parameter estimators

$$
\begin{aligned}
C_1(q) &= c_\rho \,\mathrm{sinc}\{\pi/(2q)\} \frac{q}{12q^2 - 3} \\
C_2(q) &= c_\rho \,\mathrm{sinc}\{\pi/(2q)\} \frac{q(12q^2 + 8q + 1)}{15(8q^2 - 2q - 1)}
\end{aligned}
$$

- c_ρ depends on the design density ρ
- $C_1(q)$ decreases with q
- $C_2(q)$ increases with q
- $C_2(q)/C_1(q)$ grows fast with q

Smoothing parameter estimators

$$\begin{aligned} \mathrm{var}\left(\frac{\widehat{\lambda}}{\lambda_f}\right) &= O\left(n^{-\frac{1}{2qm+1}}\right), \quad m \in [1,2] \\ \mathrm{var}\left(\frac{\widehat{\sigma}^2/\widehat{\sigma}_u^2}{\lambda_r}\right) &= O\left(n^{-\frac{1}{2q+1}}\right) \end{aligned}$$

- the convergence rate of $\widehat{\lambda}/\lambda_f$ and $\widehat{\sigma}^2/(\lambda_r\widehat{\sigma}_u^2)$ to 1 is very slow
- $\widehat{\lambda}/\lambda_f$ converge more slowly for smoother functions

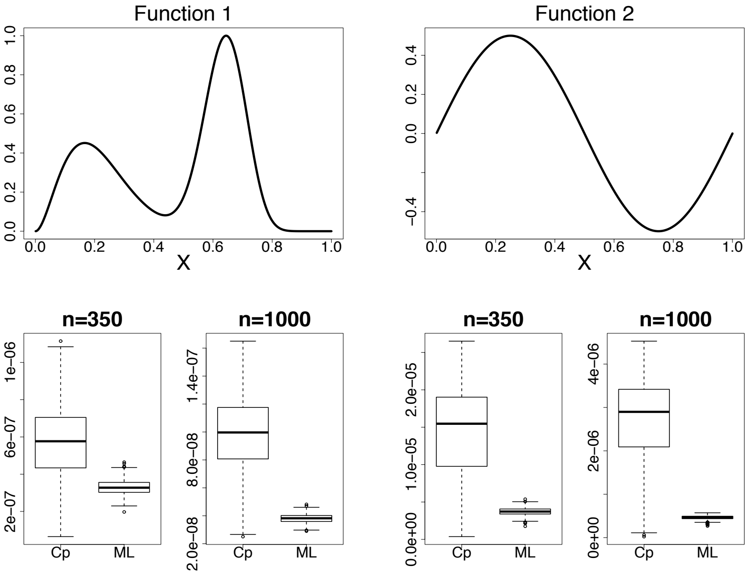
Smoothing parameter estimators

$$\frac{\mathrm{var}(\widehat{\lambda})}{\mathrm{var}(\widehat{\sigma}^2/\widehat{\sigma}_u^2)} \approx q(q+2)\left(\frac{\lambda_f}{\lambda_r}\right)^{2+1/(2q)}$$

That is due to $\lambda_f/\lambda_r \to \infty$, $n \to \infty$

- $\mathrm{var}(\widehat{\lambda}_f)/\mathrm{var}(\widehat{\sigma}^2/\widehat{\sigma}_u^2)$ is large and grows with q and n
- $\widehat{f}(\widehat{\sigma}^2/\widehat{\sigma}_u^2)$ is much more stable than $\widehat{f}(\widehat{\lambda}_f)$

Simulations $q = 2$



Simulation results show

- $\widehat{\lambda}$ is much more variable than $\widehat{\sigma}^2/\widehat{\sigma}_u^2$
- $\lambda_f/\lambda_r > 1$ and grows with n
- for a periodic f_2, λ_f/λ_r is larger than for f_1
- in small samples $\widehat{f}(\widehat{\lambda})$, $\widehat{f}(\widehat{\sigma}^2/\widehat{\sigma}_u^2)$ perform comparable for f_1

	$n = 350$		$n = 1000$	
	f_1	f_2	f_1	f_2
$\frac{R\left(\widehat{f}(\widehat{\lambda}),f\right)}{R\left(\widehat{f}(\widehat{\sigma}^2/\widehat{\sigma}_u^2),f\right)}$	1.02	0.99	0.90	0.80

Data-driven q

Large ratio λ_f/λ_r suggests that f is smoother than assumed $\mathcal{W}^q[0,1]$

If q can be chosen data-driven, such that $\lambda_f/\lambda_r \approx 1$ for given n and f, then $\widehat{\sigma}^2/\widehat{\sigma}_u^2$ should outperform $\widehat{\lambda}$ due to much smaller variance

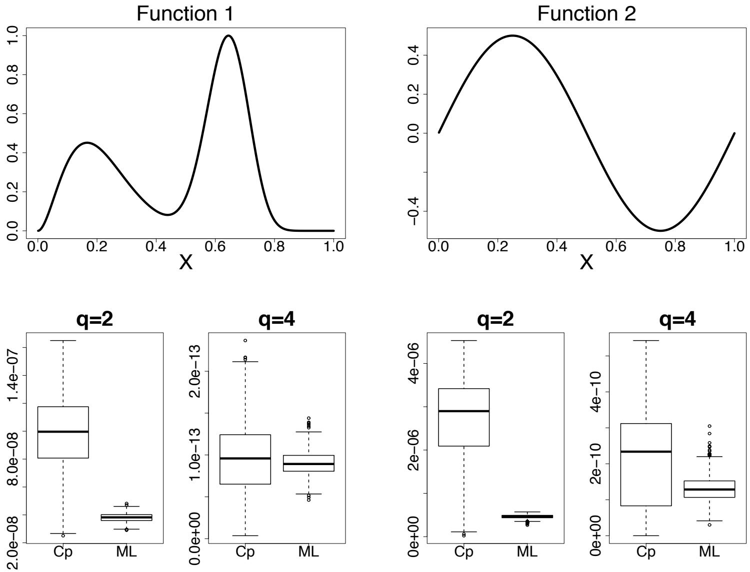
One possible way is to choose q such that

$$R(q) = \left| Y^t(I_n - S)S^2 Y - \widehat{\sigma}^2\{\mathrm{tr}(S^2) - q\}\right|$$

is smallest; here $S = S(\widehat{\sigma}^2/\widehat{\sigma}_u^2)$ is the smoother matrix

This criterion is obtained comparing estimating equations of both smoothing parameters

Simulations $n = 1000$



Conclusion

- λ_f/λ_r grows with n
- λ_f is able to adapt to the unknown smoothness (up to $2q$)
- performance of λ_r depends on n and q

- $\widehat{\lambda}$ and $\widehat{\sigma}^2/\widehat{\sigma}_u^2$ are both consistent and asymptotically normal
- convergence rate of $\widehat{\lambda}$ and $\widehat{\sigma}^2/\widehat{\sigma}_u^2$ is very slow
- $\widehat{\lambda}$ converges to λ slower for smoother functions
- constants in asymptotic variances of $\widehat{\lambda}$ and $\widehat{\sigma}^2/\widehat{\sigma}_u^2$ are obtained

- taking a larger q can improve the performance of $\widehat{\sigma}^2/\widehat{\sigma}_u^2$
- data-driven choice of q is interesting direction for further research