Smoothing parameter selection in two frameworks for penalized splines

Tatyana Krivobokova *
Georg-August-Universität Göttingen

12th October 2012

Abstract

There are two popular smoothing parameter selection methods for spline smoothing. First, smoothing parameters can be estimated minimizing criteria that approximate the average mean squared error of the regression function estimator. Second, the maximum likelihood paradigm can be employed, under the assumption that the regression function is a realization of some stochastic process. In this article the asymptotic properties of both smoothing parameter estimators for penalized splines are studied and compared. A simulation study and a real data example illustrate the theoretical findings.
Keywords: Average mean squared error minimizer; Maximum likelihood; Oracle parameters.

1 Introduction

Since works of Grace Wahba and co-authors, smoothing splines with the cross-validated smoothing parameter have become an established nonparametric regression tool. One of the attractive features of spline smoothing is its direct link to Bayesian estimation of stochastic processes, as first noticed in Kimeldorf and Wahba (1970). Employing this link, an unknown smoothing parameter can be estimated from the corresponding likelihood function. Large parameter dimension of smoothing spline estimators makes this

*Courant Research Center "Poverty, equity and growth" and Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Wilhelm-Weber-Str. 2, 37073 Göttingen, Germany

technique often unappealing in practice, so that instead, low-rank splines are employed, see e.g. Ruppert et al. (2003). The idea is to replace an unknown smooth function by a spline, with the number of knots being far less than the number of observations and use penalized least squares for estimation. Such low-rank spline estimators can also be interpreted as best linear unbiased predictors in a certain mixed model, making it possible to utilize the maximum likelihood paradigm for (smoothing parameter) estimation.

Hence, the smoothing parameter choice for spline estimators can be carried out not only minimizing criteria that approximate the average mean squared error of the regression function estimator (such as cross-validation), but also employing the maximum likelihood principle. Apparently, both approaches implicate different assumptions on the underlying data generating process. If the true regression function is a realization of a certain stochastic process, then Wahba (1985), Stein (1990) and Kou (2003) have shown for smoothing splines that the cross-validated and the maximum likelihood smoothing parameter estimators are asymptotically equal in the mean, but the maximum likelihood estimator, which is optimal in this case, is less variable. If the true regression function belongs to a Sobolev space, then Wahba (1985) have found that the maximum likelihood based smoothing spline estimator is asymptotically sub-optimal. In spite of this disappointing result, the small sample performance of the maximum likelihood estimators reported in the extensive simulation studies (see e.g. Kohn et al., 1991) appeared to be rather attractive. Therefore, the (asymptotic) properties of the smoothing parameter estimators are of interest. In their unpublished technical report, Speckman and Sun (2001) aimed to prove consistency and asymptotic normality of both smoothing parameter estimators, if the regression function belongs to a Sobolev class. However, some of these results are restricted to regression functions that satisfy natural boundary conditions.

The small-sample and asymptotic properties of both smoothing parameter estimators in the special case of low-rank spline smoothers have got less attention. Kauermann

(2005) obtained the probability for the maximum likelihood smoothing parameter estimator to undersmooth, in a setting with a fixed number of knots, while Reiss and Ogden (2009) concentrated on the geometry of both criteria. In this article general low-rank smoothers are considered. The consistency and asymptotic normality of both smoothing parameter estimators under fairly general assumptions on the underlying data generating processes are proved. In particular, obtained constants in the asymptotic variances of the smoothing parameter estimators shed light on the small-sample performance of both criteria.

The paper is organized as follows. Basic definitions and assumptions are presented in Section 2. Both smoothing parameter selectors are introduced and studied in Section 3. Some practical issues concerning the performance of the maximum likelihood estimator are discussed in Section 5. A small simulation study and a real data example are presented in Sections 4 and 6, respectively. Technical details are given in the Appendix and in the Supplementary materials.

2 Penalized splines

Let n observed values (y_i, x_i) originate from the model

$$Y_i = f(x_i) + \epsilon_i, \tag{1}$$

for some sufficiently smooth unknown function f, deterministic $x_i \in [0,1]$ and i.i.d. ϵ_i, such that $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{E}(\epsilon_i^2) = \sigma^2 > 0$ and $\mathrm{E}(\epsilon_i^4) < \infty$, $i = 1, \ldots, n$. The spline-based estimator of f is defined as the solution to

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

where $\mathcal{S}(p, \underline{\tau})$ is the $(k+p+1)$-dimensional set of spline functions of degree p based on a set of k inner knots $\underline{\tau} = \{0 = \tau_0 < \tau_1 < \ldots, \tau_k < \tau_{k+1} = 1\}$. For technical reasons, in the subsequent proofs the degree of the spline p is set to $p = 2q-1$, but in principle any $p \geq q$ can be used. A special case with $\tau_i = x_i$, $i = 1, \ldots n$ and spline functions satisfying the so-called natural boundary conditions $s^{(j)}(0) = s^{(j)}(1) = 0$, $j = q, \ldots 2q-1$ defines the smoothing spline estimator. In this article only low-rank spline smoothers with $k = o(n)$ and $\lambda > 0$ are considered and referred to as penalized spline estimators, see Wand and Ormerod (2008).

Any spline function of degree p can be represented as a sum of a polynomial of degree p and a linear combination of truncated polynomials of the same degree based on $\underline{\tau}$, that is

$$s(x) = \sum_{i=0}^{q-1} \beta_{i+1} x^i + \sum_{i=q}^{p} u_{i-q+1} x^i + \sum_{i=1}^{k} u_{p-q+1+i} (x - \tau_i)_+^p = \boldsymbol{X}(x)\boldsymbol{\beta} + \boldsymbol{Z}(x)\boldsymbol{u} = \boldsymbol{C}(x)\boldsymbol{\theta},$$

where $(x - \tau_i)_+ = \max(x - \tau_i, 0)$, $\boldsymbol{X}(x) = (1, x, \ldots, x^{q-1})$, $\boldsymbol{Z}(x) = \{x^q, \ldots, x^p, (x - \tau_1)_+^p, \ldots, (x - \tau_k)_+^p\}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^t$, $\boldsymbol{u} = (u_1, \ldots, u_{k+p+1-q})^t$, $\boldsymbol{C}(x) = \{\boldsymbol{X}(x), \boldsymbol{Z}(x)\}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \boldsymbol{u}^t)^t$. Plugging in this matrix representation of $s(x)$ into (2) and solving it with respect to $\boldsymbol{\theta}$ results in

$$\widehat{f}(x) = \boldsymbol{C}(x)\widehat{\boldsymbol{\theta}} = \boldsymbol{C}(x)(\boldsymbol{C}^t\boldsymbol{C} + \lambda n \widetilde{\boldsymbol{D}})^{-1}\boldsymbol{C}^t\boldsymbol{Y}, \tag{3}$$

where $\boldsymbol{C} = \{\boldsymbol{C}(x_1)^t, \ldots, \boldsymbol{C}(x_n)^t\}^t$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$, the penalty block-diagonal matrix $\widetilde{\boldsymbol{D}} = \text{diag}(\boldsymbol{0}_q, \boldsymbol{D})$, with $\boldsymbol{D} = \int_0^1 \boldsymbol{Z}^{(q)}(x)^t \boldsymbol{Z}^{(q)}(x) dx$. Note that in practice this particular decomposition of $\boldsymbol{C}(x) = \{\boldsymbol{X}(x), \boldsymbol{Z}(x)\}$ may be numerically unstable, since the truncated polynomial basis is often bad conditioned (see e.g. de Boor, 2001, p.84). Therefore, in practice, other equivalent representations of $\boldsymbol{C}(x)$ based on B-splines are used, see Durban and Currie (2003) and Fahrmeir et al. (2004).

Further, the following assumptions are made.

(A1) For deterministic design points $x_i \in [0,1]$, assume that $x_i = Q\{(2i-1)/(2n)\}$, $i = 1, \ldots, n$ for a strictly increasing $Q : [0,1] \to [0,1]$, $Q \in L^2[0,1]$, such that $\sup_{x\in[0,1]} Q'(x) \geq \text{const} > 0$. Let also $\rho(x) = 1/Q'\{Q^{-1}(x)\}$.

(A2) The number of equidistant knots k satisfies $k = \text{const}\, n^{\nu}$, $\nu \in (1/(2q), 1)$.

(A3) The smoothing parameter $\lambda = \lambda(n) > 0$ with $\lambda \to 0$ is such that $\lambda n \to \infty$, $n \to \infty$.

(A4) $f \in W^q[0,1] = \{f : f \in C^{q-1}[0,1],\ \int_0^1 \{f(x)^{(q)}\}^2 dx < \infty\}$.

Here and subsequently, "const" denotes a generic positive constant. Assumption (A1) ensures certain regularity of the data points and is adopted from Speckman (1985). Assumption (A2) follows from Theorem 1 of Claeskens et al. (2009) and is crucial in this study. Apparently, the penalized spline estimator (3) has two unknown control parameters – the number of knots k and the smoothing parameter λ. In practice, the typical approach is to first fix k to some value and then choose λ in a data-driven way. Claeskens et al. (2009) defined the variable $K_q = \text{const}\, \lambda k^{2q}$ as the maximum eigenvalue of $\lambda n(\boldsymbol{C}^t\boldsymbol{C})^{-1}\widetilde{\boldsymbol{D}}$ and showed that λ is identifiable (can be estimated consistently) if and only if $K_q \to \infty$, which is equivalent to (A2) and (A3). This choice of k ensures also that the approximation bias is negligible and two estimators with different number of knots (both satisfying (A2)) are indistinguishable in terms of the average mean squared error.

3 Smoothing parameter estimation

3.1 Two smoothing parameter selectors

First, let model (1) and assumptions (A1) – (A4) hold. Let also $A_n(\lambda) = n^{-1}\sum_{i=1}^n \{\widehat{f}(x_i) - f(x_i)\}^2$. Then, it is reasonable to choose a λ that minimizes the average mean squared error $\mathrm{E}_f\{A_n(\lambda)\}$, where E_f denotes the expectation under the model (1). Since in practice f is unknown, some asymptotically unbiased estimators of $\mathrm{E}_f\{A_n(\lambda)\}$ are used instead,

e.g. Mallows' C_p, defined by $\mathrm{C}_p(\lambda) = n^{-1}\boldsymbol{Y}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{Y}\{1 + 2\mathrm{tr}(\boldsymbol{S}_\lambda)/n\}$, where $\boldsymbol{S}_\lambda = \boldsymbol{C}(\boldsymbol{C}^t\boldsymbol{C} + \lambda n\widetilde{\boldsymbol{D}})^{-1}\boldsymbol{C}^t$ denotes the smoothing matrix. Note that

$$\mathrm{E}_f\{\mathrm{C}_p(\lambda)\} - \sigma^2\left\{1 - \frac{4\mathrm{tr}(\boldsymbol{S}_\lambda)^2}{n^2}\right\} = \mathrm{E}_f\{A_n(\lambda)\}\left\{1 + \frac{2\mathrm{tr}(\boldsymbol{S}_\lambda)}{n}\right\} = \mathrm{E}_f\{A_n(\lambda)\}\left\{1 + o(1)\right\},$$

due to $\mathrm{tr}(\boldsymbol{S}_\lambda)/n = \mathrm{const}\,\lambda^{-1/(2q)}n^{-1} = o(1)$, as follows from Lemma 1 in the Appendix and (A3). In fact, other well-known criteria like generalized cross validation (GCV, Craven and Wahba, 1978) and Akaike information criterion (AIC, Akaike, 1969) are also asymptotically unbiased estimators of $\mathrm{E}_f\{A_n(\lambda)\}$, see Supplementary materials. The smoothing parameter estimator under (1) is defined as $\widehat{\lambda}_f = \arg\min_{\lambda>0}\mathrm{C}_p(\lambda)$, and the corresponding estimating equation with $T_{Cp}(\widehat{\lambda}_f) = 0$ (see e.g. Section 5.1. in van der Vaart, 1998, for the formal definition) is given by

$$T_{Cp}(\lambda) = \frac{1}{n}\left[\boldsymbol{Y}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda\boldsymbol{Y}\left\{1 + \frac{2\mathrm{tr}(\boldsymbol{S}_\lambda)}{n}\right\} - \boldsymbol{Y}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{Y}\frac{\mathrm{tr}(\boldsymbol{S}_\lambda - \boldsymbol{S}_\lambda^2)}{n}\right].$$

Let now (A1) – (A2) still hold, but instead of (1) the data follow

$$Y_i = F(x_i) + \epsilon_i = \boldsymbol{X}(x_i)\boldsymbol{\beta} + \sigma_u\int_0^1\frac{(x_i - t)_+^{q-1}}{(q-1)!}dW(t) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{4}$$

$i = 1, \ldots, n$, where $\int_0^1(x_i - t)_+^{q-1}dW(t)$ is a $(q-1)$-fold integrated Wiener process (Wahba, 1990). The best linear unbiased predictor of F based on n data pairs (y_i, x_i) coincides with the smoothing spline estimator with the smoothing parameter $\sigma^2/(n\sigma_u^2)$, as shown in Kimeldorf and Wahba (1970). Under model (1), a smooth $f \in W^q[0,1]$ was estimated solving a minimization problem (2) over a finite dimensional spline space $\mathcal{S}(p, \underline{\tau})$. Under model (4), one can proceed in a similar way and estimate (predict) $F(x)$ from

$$Y_i = \boldsymbol{X}(x_i)\boldsymbol{\beta} + \boldsymbol{Z}(x_i)\boldsymbol{u} + \epsilon_i,\ i = 1, \ldots, n,\ \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \sigma_u^2\boldsymbol{D}^{-1}),\ \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n), \tag{5}$$

with $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^t$. Since $\boldsymbol{D}$ is a symmetric positive definite matrix by definition, its inverse exists and is unique. Model (5) is a linear mixed model. The best linear unbiased

predictor for $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \boldsymbol{u}^t)^t$ is known to be $\widetilde{\boldsymbol{\theta}} = (\boldsymbol{C}^t\boldsymbol{C} + \sigma^2/\sigma_u^2\widetilde{\boldsymbol{D}})^{-1}\boldsymbol{C}^t\boldsymbol{Y}$ (see e.g. Robinson, 1991). Thus, $\widetilde{f}(x) = \boldsymbol{C}(x)\widetilde{\boldsymbol{\theta}}$ equals $\widehat{f}(x)$ defined in (3) with the smoothing parameter $\sigma^2/(n\sigma_u^2)$. Parameters $\lambda = \sigma^2/(n\sigma_u^2)$ and σ^2 are estimated from the corresponding restricted log-likelihood (see Patterson and Thompson, 1971) $l_r(\sigma^2, \lambda; \boldsymbol{Y}) = l(\widetilde{\boldsymbol{\beta}}, \sigma^2, \sigma_u^2; \boldsymbol{Y}) - \log|\boldsymbol{X}^t\boldsymbol{V}_\lambda^{-1}\boldsymbol{X}|/2$, where $\boldsymbol{V}_\lambda = \boldsymbol{I}_n + \sigma_u^2\boldsymbol{Z}\boldsymbol{D}^{-1}\boldsymbol{Z}^t/\sigma^2$, $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{V}_\lambda^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}_\lambda^{-1}\boldsymbol{Y}$ and $l(\boldsymbol{\beta}, \sigma^2, \sigma_u^2; \boldsymbol{Y})$ is the log-likelihood for the mixed model (5). Plugging in $\widehat{\sigma}^2 = \boldsymbol{Y}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{Y}/(n-q)$ into $l_r(\sigma^2, \lambda; \boldsymbol{Y})$ results in the profile restricted log-likelihood for λ

$$-2l_p(\lambda; \boldsymbol{Y}) = (n-q)\log \boldsymbol{Y}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{Y} + \log|\boldsymbol{V}_\lambda||\boldsymbol{X}^t\boldsymbol{V}_\lambda^{-1}\boldsymbol{X}|.$$

More details on the (restricted profile) log-likelihood are provided in the Supplementary materials. The corresponding smoothing parameter estimator is denoted by $\widehat{\lambda}_r = \arg\min_{\lambda>0}\{-l_p(\lambda; \boldsymbol{Y})\}$ and the estimating equation for $\widehat{\lambda}_r$ is taken to be

$$T_{ML}(\lambda) = \frac{1}{n}\left[\boldsymbol{Y}^t(\boldsymbol{S}_\lambda - \boldsymbol{S}_\lambda^2)\boldsymbol{Y} - \boldsymbol{Y}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{Y}\,\frac{\mathrm{tr}(\boldsymbol{S}_\lambda) - q}{n-q}\right].$$

3.2 Oracle smoothing parameters

Before the asymptotic properties of both smoothing parameter estimators are considered, let us compare oracle smoothing parameters under both frameworks. Oracle smoothing parameters are explicitly defined in Table 1 and are not available in practice, since they depend on the unknown model parameters. According to this definition, λ_f and λ_r are the oracle smoothing parameters in case the data are modeled according to the true data generating processes. If the data are modeled according to (4), but originate from (1), the corresponding oracle smoothing parameter is denoted by $\lambda_{r|f}$. That is, $\lambda_{r|f}$ is a smoothing parameter one would get in the mean from the likelihood, in case the data are sampled from (1). The reverse situation defines $\lambda_{f|r}$. Since the data can either follow model (1) (frequentist framework) or be a realization of the stochastic process (4) (stochastic

	Model (1) is estimated	Model (4) is estimated
Data follow model (1)	$\lambda_f = \arg\min_{\lambda>0} \mathrm{E}_f\{A_n(\lambda)\}$	$\lambda_{r\|f} = \arg\min_{\lambda>0} \mathrm{E}_f\{-l_p(\lambda;\boldsymbol{Y})\}$
Data follow model (4)	$\lambda_{f\|r} = \arg\min_{\lambda>0} \mathrm{E}_\beta\{\mathrm{C}_p(\lambda)\}$	$\lambda_r = \sigma^2/(n\sigma_u^2)$

Table 1: Definition of oracle smoothing parameters. E_f and E_β are expectations under models (1) and (4), respectively.

framework), one is interested to compare the performance of $\lambda_{r|f}$ and λ_f, as well as of $\lambda_{f|r}$ and λ_r. All oracle smoothing parameters depend on the sample size n, which is omitted in the notation.

First, let the true data follow model (1) and compare $\lambda_{r|f}$ and λ_f. Note that

$$\mathrm{E}_f\{T_{ML}(\lambda)\} = \mathrm{E}_f\{T_{Cp}(\lambda)\} + R(\lambda), \tag{6}$$

with $R(\lambda) = n^{-1}[\boldsymbol{f}^t\boldsymbol{S}_\lambda^2(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{f} - \sigma^2\{\mathrm{tr}(\boldsymbol{S}_\lambda^3) - q\} + o(1)]$. Derivation of this equation is given in the Supplementary materials. Representation (6) makes clear that $\lambda_{r|f}$ is biased with respect to λ_f, unless $R(\lambda_{r|f})$ is not negligible. Following Wahba (1985), let $\mathcal{W}^{mq}[0,1] = \{f \in W^q[0,1] : \|\bar{s}_p^{(mq)}\|^2 < M < \infty\}$, $m \in [1,2]$, M independent of k, $\bar{s}_p$ as the best $L_2[0,1]$ approximation of $f \in W^q[0,1]$ by $\mathcal{S}(p,\underline{\tau})$ and $\|\bar{s}_p^{(mq)}\|^2$ is defined in the Appendix in (7). The set $\mathcal{W}^{mq}[0,1]$ is related to a Besov space with certain boundary conditions (see Section 3 in Cox, 1988) and a larger m corresponds to a smoother space. Then, up to constants, the result of Wahba (1985) holds also for penalized splines.

Theorem 1 *Let model (1) with $f \in \mathcal{W}^{mq}[0,1]$, $m \in [1,2]$ and (A1) – (A3) hold. Then,*

$$\begin{aligned}
\lambda_{r|f} &= \left[n\,\frac{\|\bar{s}_p^{(q)}\|^2\; c(\rho)\{1+o(1)\}}{\sigma^2 c(q,2,K_q)}\right]^{-\frac{2q}{2q+1}}, \\
\lambda_f &\geq \left[n\,\frac{\|\bar{s}_p^{(mq)}\|^2 4q\; c(\rho)\{1+o(1)\}}{\sigma^2 c(q,2,K_q)}\right]^{-\frac{2q}{2qm+1}},
\end{aligned}$$

with equality in the last expression for $m = 2$. Here $c(\rho)$ and $c(q, 2, K_q)$ are constants defined in Lemma 1 in the Appendix.

This theorem implies that $\lambda_f/\lambda_{r|f} \to \infty$ and, hence, $\mathrm{E}_f\{A_n(\lambda_{r|f})\}/\mathrm{E}_f\{A_n(\lambda_f)\} \to \infty$ as $n \to \infty$, for $m \in (1, 2]$. In a much more general framework this is also shown in Lukas (1993) and Lukas (1995). Summarizing, if f satisfies certain additional smoothness assumptions, then the average mean squared error minimizer λ_f automatically adapts, resulting in a faster rate of convergence for $\widehat{f}(\lambda_f)$. In contrast, $\lambda_{r|f}$ is not able to utilize these additional properties of f and the convergence of $\widehat{f}(\lambda_{r|f})$ depends on the used q only. Consequently, if f is any smoother than $W^q[0, 1]$, then $\widehat{f}(\lambda_{r|f})$ is asymptotically sub-optimal and undersmooths f compared to $\widehat{f}(\lambda_f)$ for $n \to \infty$.

Now, let the data follow (4) and consider $\lambda_{f|r}$ and λ_r.

Theorem 2 *Under model (4) and assumptions (A1) – (A2) it holds $\lambda_{f|r} = \lambda_r\{1 + o(1)\}$.*

Hence, in the stochastic framework both oracle smoothing parameters are asymptotically equal and optimal, which agrees with the result of Wahba (1985) for smoothing splines.

3.3 Properties of smoothing parameter estimators

The following theorem states the properties of $\widehat{\lambda}_r$ and $\widehat{\lambda}_f$ in the frequentist framework.

Theorem 3 *Let model (1) and assumptions (A1) – (A4) hold. Then,*

$$\frac{\widehat{\lambda}_r}{\lambda_{r|f}} \xrightarrow{\mathcal{P}} 1 \quad \textit{and} \quad \frac{\widehat{\lambda}_f}{\lambda_f} \xrightarrow{\mathcal{P}} 1, \quad \textit{as } n \to \infty.$$

Moreover, for $n \to \infty$

$$(\lambda_{r|f})^{-1/(4q)} \left(\frac{\widehat{\lambda}_r}{\lambda_{r|f}} - 1 \right) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, 2\, c(\rho)\, C_1(q)\Big)$$

and

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

where $c(\rho)$ is defined in Lemma 1 in the Appendix and

$$C_1(q) = \text{sinc}\{\pi/(2q)\}\ \frac{q}{12q^2-3},$$

$$C_2(q) = \text{sinc}\{\pi/(2q)\}\ \frac{q(12q^2+8q+1)}{15(8q^2-2q-1)},$$

with $\text{sinc}(x) = \sin(x)/x$.

According to this result and Theorem 1, $\text{var}_f(\widehat{\lambda}_r/\lambda_{r|f}) = O\left(n^{-1/(2q+1)}\right)$ and $\text{var}_f(\widehat{\lambda}_f/\lambda_f) = O\left(n^{-1/(2qm+1)}\right)$, $m \in [1,2]$, being larger for smoother functions. This exceedingly slow convergence rate of both smoothing parameter estimators agrees with known results for kernel regression, see e.g. Rice (1984) or Härdle et al. (1988). Apparently, the variances of $\widehat{\lambda}_f$ and $\widehat{\lambda}_r$ depend on the corresponding oracle smoothing parameters, implying that the ratio $\text{var}_f(\widehat{\lambda}_f)/\text{var}_f(\widehat{\lambda}_r)$ can grow with n and this rate is fastest for $f \in \mathcal{W}^{2q}[0,1]$. Also, the $C_2(q)/C_1(q)$ is relatively large and increases with q (e.g. $C_2(2)/C_1(2) = 65/9$ and $C_2(4)/C_1(4) = 405/17$). Consequently, in finite samples the variance of $\widehat{\lambda}_f$ can be hundreds times larger than the variance of $\widehat{\lambda}_r$, especially for larger sample sizes and smoother functions, see also Section 4. Another interesting finding is that the constant $C_1(q)$ decreases for growing q, while $C_2(q)$ is several times larger and increases with q. That is, using a larger q has a smaller effect on the variability of $\widehat{\lambda}_r$, but significantly increases the variance of $\widehat{\lambda}_f$, at least in small samples.

The properties of $\widehat{\lambda}_f$ and $\widehat{\lambda}_r$ in the stochastic framework (with $\lambda_{f|r} = \lambda_r\{1+o(1)\}$, according to Theorem 2) are given in the following theorem.

Theorem 4 *Let model (4) and assumptions (A1) – (A2) hold. Then,*

$$\frac{\widehat{\lambda}_r}{\lambda_r} \xrightarrow{\mathcal{P}} 1 \quad \textit{and} \quad \frac{\widehat{\lambda}_f}{\lambda_{f|r}} \xrightarrow{\mathcal{P}} 1, \quad \textit{as } n \to \infty.$$

Moreover, for $n \to \infty$

$$(\lambda_r)^{-1/(4q)} \left(\frac{\widehat{\lambda}_r}{\lambda_r} - 1\right) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, 2\, c(\rho)\, C_3(q)\Big)$$

and

$$(\lambda_{f|r})^{-1/(4q)} \left(\frac{\widehat{\lambda}_f}{\lambda_{f|r}} - 1 \right) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, 2\, c(\rho)\, C_4(q)\Big),$$

where $c(\rho)$ is defined in Lemma 1 in the Appendix and

$$C_3(q) = \text{sinc}\{\pi/(2q)\}\, \frac{2q}{2q-1},$$

$$C_4(q) = \text{sinc}\{\pi/(2q)\}\, \frac{4q(2q+1)}{3(2q-1)},$$

with $\text{sinc}(x) = \sin(x)/x$.

Smoothing parameter estimators in the stochastic framework for smoothing splines of order 1 and 2 have been already studied in the literature. Known results for smoothing splines give the ratio of two variances $\text{var}_\beta(\widehat{\lambda}_f)/\text{var}_\beta(\widehat{\lambda}_r)$ to be $2 = C_4(1)/C_3(1)$ for $q = 1$ (Stein, 1990) and $10/3 = C_4(2)/C_3(2)$ for $q = 2$ (Kou, 2003). This agrees with the results of Theorem 4, which also holds for low-rank smoothers and general q.

4 Simulations

In this section the theoretical findings are illustrated by a simulation study. Three functions are considered: the first $f_1(x) = 6\beta_{30,17}(x)/10 + 4\beta_{3,11}(x)/10$, with $\beta_{l,m}(x) = \Gamma(l+m)\{\Gamma(l)\Gamma(m)\}^{-1}x^{l-1}(1-x)^{m-1}$, also used in Wahba (1985) (left top plot in Figure 1), the second $f_2(x) = \sin(2\pi x)$ and the third function is obtained as a realization of a stochastic process (5) with $p = 2q-1 = 3$, $k = 40$ equidistant knots, $\boldsymbol{\beta} = (1.5, -0.02)^t$ and $\sigma = \sigma_u = 0.1$ (left bottom plot in Figure 1). In the frequentist framework the errors ϵ_i are assumed to be i.i.d. zero mean normal with $\sigma = 0.1$. The covariate values are taken to be $x_i = i/n$ with $i = 1, \ldots, n$ (similar results hold for non-equidistant, but sufficiently regular xs). All three functions are estimated with penalized splines of degree $p = 2q-1 = 3$ based on equidistant knots with k ranging from 10 to 40 with step size 1.

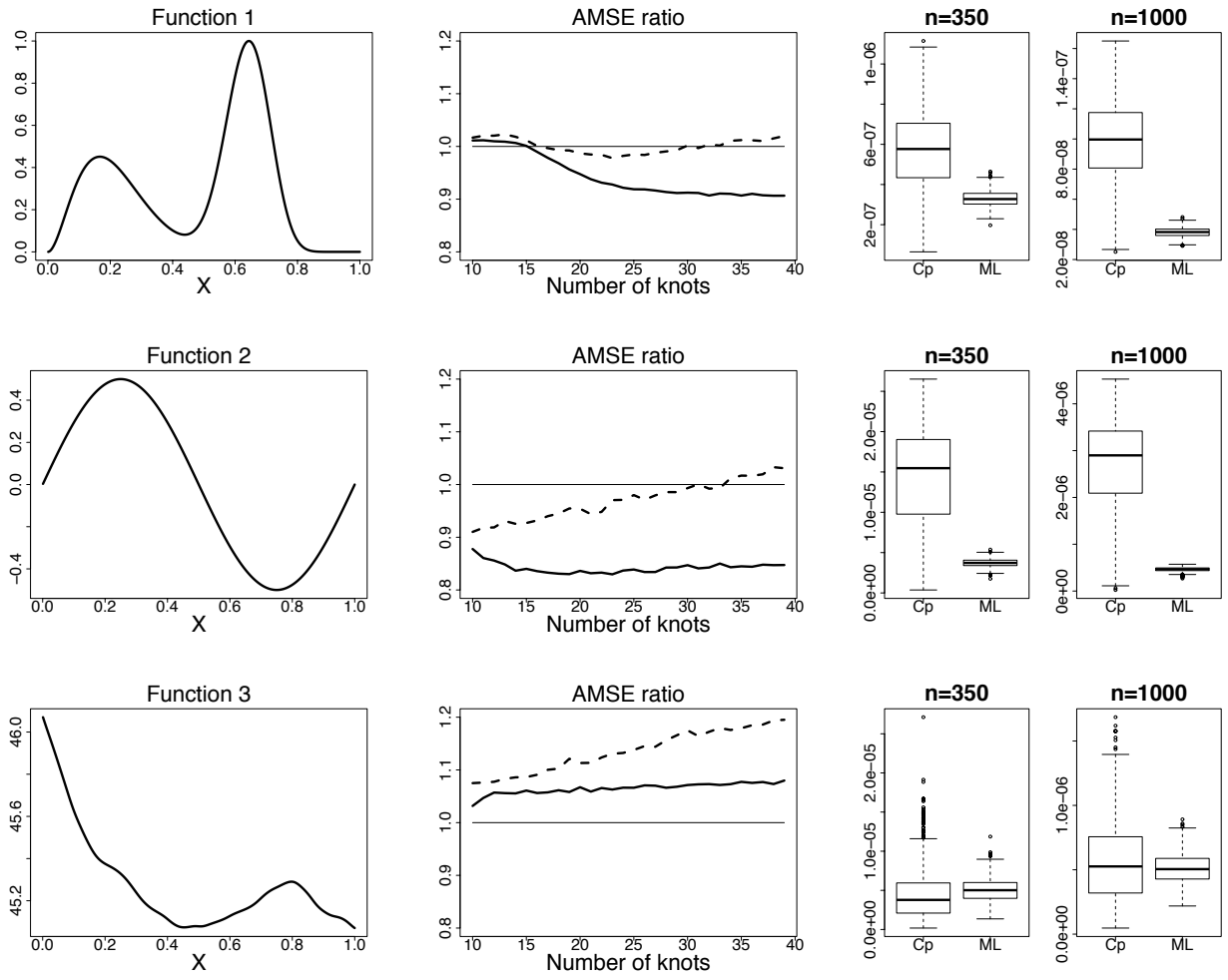

Figure 1: Effect of the sample size for $q = 2$: For $n = 1000$ (bold) and $n = 350$ (dashed) plots of $\bar{A}_n(\widehat{\lambda}_f)/\bar{A}_n(\widehat{\lambda}_r)$ depending on number of knots (middle plots) and boxplots for $\widehat{\lambda}_f$ and $\widehat{\lambda}_r$ averaged over number of knots (right plots) for f_1 (top left), f_2 (middle left) and f_3 (bottom left).

The number of Monte Carlo replications is 1000.

Figure 1 summarizes the simulation results for two sample sizes $n = 1000$ (bold) and $n = 350$ (dashed). The second column of Figure 1 shows the ratio $\bar{A}_n(\widehat{\lambda}_f)/\bar{A}_n(\widehat{\lambda}_r)$, where $\bar{A}_n(\cdot)$ denotes the sample mean of A_n. For $n = 1000$ Mallows' C_p performs much better than the maximum likelihood for the first two fixed functions and somewhat worse in the third stochastic setting. Note that as soon as enough knots are taken (about $15 - 20$ in this setting, which makes (A2) to fullfil), $\bar{A}_n(\cdot)$ remains nearly constant as a function of

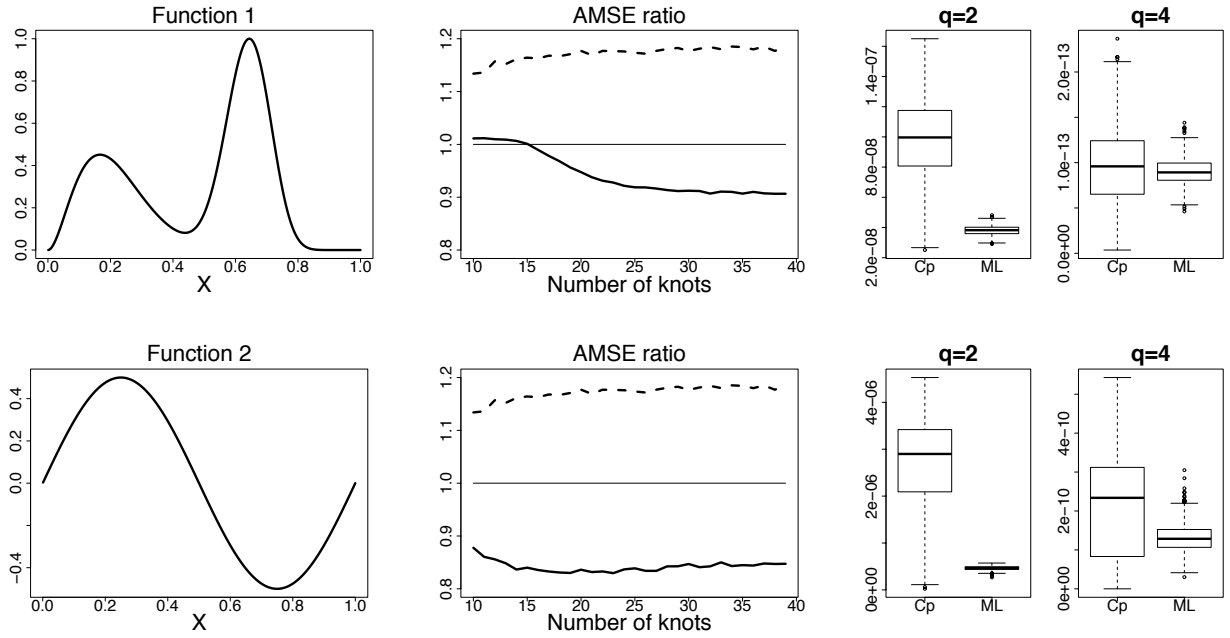

Figure 2: Effect of the penalty order for $n = 1000$: For $q = 2$ (bold) and $q = 4$ (dashed) plots of $\bar{A}_n(\widehat{\lambda}_f)/\bar{A}_n(\widehat{\lambda}_r)$ depending on number of knots (middle plots) and boxplots for $\widehat{\lambda}_f$ and $\widehat{\lambda}_r$ averaged over number of knots (right plots) for f_1 (top left) and f_2 (bottom left).

k. In a smaller sample of $n = 350$ the ratio $\bar{A}_n(\widehat{\lambda}_f)/\bar{A}_n(\widehat{\lambda}_r)$ for the first two functions is much closer to one, than for $n = 1000$. This can be attributed not only to a smaller variance of $\widehat{\lambda}_r$, but also to a smaller ratio $\lambda_f/\lambda_{r|f}$, as visible from the box plots of estimated smoothing parameters in the last column of Figure 1. As expected, the ratio of variances $\widehat{\text{var}}_f(\widehat{\lambda}_f)/\widehat{\text{var}}_f(\widehat{\lambda}_r)$ is found to increase with the sample size. In particular, for the smoothest periodic function f_2 this ratio is extremely large being around 480 for $n = 1000$ and 180 for $n = 350$. In the stochastic framework the influence of the sample size is less pronounced.

Further, additional simulations for $n = 1000$ and $q = 4$ were run. In Figure 2 ratios of the average mean squared errors $\bar{A}_n(\widehat{\lambda}_f)/\bar{A}_n(\widehat{\lambda}_r)$ for $q = 2$ (bold) and $q = 4$ (dashed) are shown. Apparently, the maximum likelihood estimator outperforms the Mallows' C_p based estimator, once q is increased up to 4, at least for this sample size $n = 1000$. In fact, $\bar{A}_n(\widehat{\lambda}_r)$ with $q = 4$ is also smaller than $\bar{A}_n(\widehat{\lambda}_f)$ with $q = 2$. Hence, the maximum

likelihood estimator with a larger q may be preferable in practice due its larger efficiency and stability in finite samples.

All together, the outcome of simulations, consistently with the theoretical results, confirms that in the frequentist framework $\lambda_f/\lambda_{r|f}$, as well as $\widehat{\mathrm{var}}_f(\widehat{\lambda}_f)/\widehat{\mathrm{var}}_f(\widehat{\lambda}_r)$, grow with the sample size. The ratio $\lambda_f/\lambda_{r|f}$ depends heavily on f, penalty order q and the sample size n and found to be closer to one for smaller ns and larger qs. In the stochastic framework both estimators perform similar, but the Mallows's C_p based smoothing parameter estimator is more variable.

5 Practical issues in the frequentist framework

Theorem 3 and results of Section 4 suggest that, if for a particular data set $\lambda_f/\lambda_{r|f}$ is close to one, then $\widehat{\lambda}_r$ is competitive to $\widehat{\lambda}_f$, since its variance is much smaller. On the other hand, a large ratio $\lambda_f/\lambda_{r|f}$ would indicate that f belongs to a smoother class of functions than assumed $W^q[0,1]$ (see Theorem 1) and a larger q can be used for $\widehat{\lambda}_r$. Hence, if one could choose q in a data-drive way, so that $\lambda_f/\lambda_{r|f}$ is closest to one, then $\widehat{f}(\widehat{\lambda}_r)$ would perform better than $\widehat{f}(\widehat{\lambda}_f)$, in terms of the average mean squared error. A natural way is to look at $R(\lambda)$ from (6). If for a given q the value $R(\lambda_{r|f}) = 0$, then $\lambda_{r|f} = \lambda_f$, while $R(\lambda_{r|f}) < 0$ implies $\lambda_{r|f} > \lambda_f$. Since $\mathrm{var}_f(\widehat{\lambda}_r)$ is small even for large qs, one can expect that an unbiased estimator of $R(\lambda_{r|f})$ has good small sample properties. Therefore, define

$$R^*(q) = \left[\boldsymbol{Y}^t \boldsymbol{S}_\lambda^2 (\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{Y} - \sigma^2\{\mathrm{tr}(\boldsymbol{S}_\lambda^2) - q\}\right]\Big|_{\lambda=\widehat{\lambda}_r}$$

to choose the penalty order q for the estimation of $\widehat{\lambda}_r$ as $\arg\min_{q\in\mathbb{N}} |R^*(q)|$. The detailed study of the criterion $|R^*(q)|$ is out of the scope of this paper, but in the Supplementary materials a small simulation study is reported, complementary to the one in Section 4. Such data-driven choice of q is an interesting direction for further research.

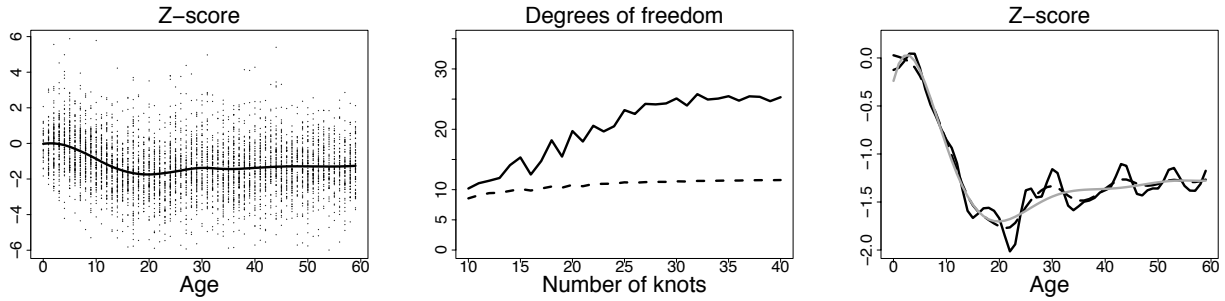

Figure 3: Left: The data, Mallows' C_p and ML based estimates with 10 knots (indistinguishable); Middle: Estimated degrees of freedom for Mallows' C_p (bold) and ML (dashed) based estimates; Right: Age effect estimates with Mallows' C_p (bold), ML with $q = 2$ (dashed) and ML with $q = 6$ (grey) smoothing parameters, all based on 40 knots.

6 Example

To illustrate high instability of the Mallows' C_p criterion and general difficulties of smoothing parameter selection in case of low signal-to-noise ratio in the data, the example on undernutrition of children in Kenya is presented. Information on weight and height of $n = 4651$ children in Kenya is obtained from the Kenyan Demographic and Health Survey (KDHS2003, see Central Bureau of Statistics (CBS) Kenya et al., 2004). The so-called Z-score for stunting is a common indicator for chronic undernutrition (see WHO, 1998) and is defined as $Z_i = \{H_i - \text{med}(H)\}\text{var}(H)^{-1/2}$, where H_i is the height of ith child at a certain age and med(H) and var(H) are the median and variance of the children heights at the same age in some reference population of healthy children, respectively. The data at hand are cross-sectional, that is, there are no repeated observations of the same individuals. A very low signal-to-noise ratio in the data shown in the left plot of Figure 3, makes estimation challenging. In spite of nearly five thousand observations, the Mallows' C_p criterion depends strongly on the number of knots chosen, selecting more complex models for larger ks. In contrast, the maximum likelihood estimator is much more robust, as clearly seen in the middle plot of Figure 3. For the estimation $p = 2q - 1 = 3$ and equidistant knots ranging from 10 to 40 with step size 1 were used. Estimates of the age

effect based on 40 knots with Mallows' C_p (bold line) and with the maximum likelihood (dashed) are shown in the right plot of Figure 3. Even though the maximum likelihood estimator is more robust, there are some structures in the estimated curve that seem to be implausible. The criterion $|R^*(q)|$, discussed in the previous section, has the smallest value at $q = 6$, suggesting that the regression function might be much smoother. Indeed, estimating the data using $q = 6$ (grey line in the right plot of Figure 3) gives a reasonable result. In fact, estimating the data using sophisticated techniques with an adaptive smoothing parameter yields the same fit.

7 Conclusion

Properties of oracle smoothing parameters for smoothing splines in the frequentist and stochastic frameworks are well-known and in this work also shown to hold for low-rank smoothers, under certain assumptions on the number of knots. In particular, in the stochastic framework both – the average mean squared error minimizer and the maximum likelihood oracle smoothing parameter – are asymptotically equal and optimal, while in the frequentist framework the maximum likelihood oracle smoothing parameter is asymptotically sub-optimal. In this article, both Mallows' C_p and maximum likelihood smoothing parameter estimators are shown to be consistent and asymptotically normal in the frequentist and stochastic frameworks for penalized splines. Obtained constants in the asymptotic variances deliver interesting insights into the small-sample behavior of smoothing parameter estimators. In both frameworks, the variance of the maximum likelihood estimator is found to be smaller than that of the Mallows' C_p estimator, with the constant decreasing for growing penalty order q. Therefore, in spite of the asymptotic sub-optimality in the frequentist framework, the maximum likelihood estimator in finite samples and for a certain choice of q appears to be superior to the Mallows' C_p estimator.

Acknowledgments

I am very grateful to the editor, associate editor and the referees of the paper for the constructive remarks that helped to improve the paper substantially. Also the support of the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of the Institutional Strategy of the University of Göttingen is acknowledged.

Appendix. Technical details

A.1 Preliminaries

To estimate f by solving (2), usually the so-called Demmler-Reinsch basis for $\mathcal{S}(p, \underline{\tau})$ is employed. This basis is uniquely defined by the conditions

$$\begin{aligned} \sum_{l=1}^{n} \phi_{k,i}(x_l)\phi_{k,j}(x_l) &= \delta_{ij}, \\ \int_0^1 \phi_{k,i}(x)^{(q)}\phi_{k,j}(x)^{(q)}dx &= \eta_i\delta_{ij}, \quad i,j = 1,\ldots,k+p+1, \end{aligned}$$

for the Kronecker delta δ_{ij}.

With this basis, $s_p(x) = \sum_{i=1}^{k+p+1} b_i\phi_{k,i}(x)$ is the best least squares approximation of f by $\mathcal{S}(p, \underline{\tau})$, with $b_i = \sum_{j=1}^{n} \phi_{k,i}(x_j)f(x_j)$. The best $L_2[0,1]$ projection of f onto $\mathcal{S}(p, \underline{\tau})$ can be written as $\bar{s}_p(x) = \sum_{i=1}^{k+p+1} b_{p,i}\phi_{k,i}(x)$, $b_{p,i} = \{\int_0^1 \phi_{k,i}(x)^2\rho(x)dx\}^{-1} \int_0^1 \phi_{k,i}(x)f(x)\rho(x)dx$. Under assumption (A1) it holds for any $g_1, g_2 \in W^q[0,1]$ (see e.g. Speckman, 1985),

$$\left|\frac{1}{n}\sum_{i=1}^{n} g_1(x_i)g_2(x_i) - \int_0^1 g_1(x)g_2(x)\rho(x)dx\right| \leq n^{-2}\text{const } \|g_1\|_{L_2[0,1]}\|g_2\|_{L_2[0,1]},$$

so that $b_{p,i}^2 = b_i^2\{1 + O(n^{-1})\}$ for any fixed i. With this, one can define

$$\|\bar{s}_p^{(mq)}\|^2 := \frac{1}{n}\sum_{i=q+1}^{k+p+1} b_{p,i}^2(n\eta_i)^m = \frac{1}{n}\sum_{i=q+1}^{k+p+1} b_i^2(n\eta_i)^m\{1 + o(1)\} =: \|s_p^{(mq)}\|^2\{1 + o(1)\}, \quad (7)$$

which is a usual $L_2[0,1]$ norm for $m = 1$, i.e. $\|\bar{s}_p^{(q)}\|^2 = \int_0^1 \{\bar{s}_p(x)^{(q)}\}^2 dx$. Following Claeskens et al. (2009), the approximation for η_i, derived in Speckman (1985) under (A1), will be used

$$\eta_1 = \cdots = \eta_q = 0,\ \eta_i \ = \ n^{-1}\widetilde{c}(\rho)^{2q}(i-q)^{2q},\ i = q+1, \ldots, k+p+1, \tag{8}$$

with $\widetilde{c}(\rho) = \pi \int_0^1 \rho(x)^{1/(2q)} dx\{1+o(1)\} = c(\rho)\{1+o(1)\}$, where $o(1)$ converges to zero as $n \to \infty$ and is independent of i for $i = o\left\{n^{2/(2q+1)}\right\}$.

In practice, the basis matrix $\boldsymbol{\Phi}_k = \{\boldsymbol{\phi}_k(x_1)^t, \ldots, \boldsymbol{\phi}_k(x_n)^t\}^t$, for a row vector $\boldsymbol{\phi}_k(x) = \{\phi_{k,1}(x), \ldots, \phi_{k,k+p+1}(x)\}$, can be obtained from the singular value decomposition of the matrix $(\boldsymbol{C}^t\boldsymbol{C})^{-1/2}\widetilde{\boldsymbol{D}}(\boldsymbol{C}^t\boldsymbol{C})^{-1/2} = \boldsymbol{U}\mathrm{diag}(\boldsymbol{\eta}_k)\boldsymbol{U}^t$, $\boldsymbol{\eta}_k = (\boldsymbol{0}_q, \eta_{q+1}, \ldots, \eta_{k+p+1})^t$. Then, $\boldsymbol{\Phi}_k = \boldsymbol{C}(\boldsymbol{C}^t\boldsymbol{C})^{-1/2}\boldsymbol{U}$ is a $n \times (k+p+1)$ semi-orthonormal matrix with $\boldsymbol{\Phi}_k^t\boldsymbol{\Phi}_k = \boldsymbol{I}_n$ and $\boldsymbol{S}_\lambda = \boldsymbol{C}(\boldsymbol{C}^t\boldsymbol{C} + \lambda n\widetilde{\boldsymbol{D}})^{-1}\boldsymbol{C}^t = \boldsymbol{\Phi}_k\{\boldsymbol{I}_n + \lambda n\mathrm{diag}(\boldsymbol{\eta}_k)\}^{-1}\boldsymbol{\Phi}_k^t$, so that $\mathrm{tr}(\boldsymbol{S}_\lambda^l) = \sum_{i=1}^{k+p+1}(1+\lambda n\eta_i)^{-l}$. Note that K_q defined as the maximum eigenvalue of $\lambda n(\boldsymbol{C}^t\boldsymbol{C})^{-1}\widetilde{\boldsymbol{D}}$ can be approximated as $K_q = \lambda\{\widetilde{c}(\rho)(k+p+1-q)\}^{2q}$. Employing (8), $\mathrm{tr}(\boldsymbol{S}^l)$ can be explicitly calculated.

Lemma 1 *If (A1) – (A3) hold, then for any $q, l \in \mathbb{N}$*

$$\mathrm{tr}(\boldsymbol{S}_\lambda^l) = \lambda^{-1/(2q)}\frac{c(q,l,K_q)}{c(\rho)}\{1+o(1)\},$$

where the constant $c(\rho)$ is defined after (8) and

$$c(q,l,K_q) \ = \ \frac{\Gamma\{l-1/(2q)\}\Gamma\{1/(2q)\}}{2q\Gamma(l)} - \frac{K_q^{1-2ql}}{2ql-1}\,{}_2F_1\left[\left\{l, l-\frac{1}{2q}\right\}; \left\{l+1-\frac{1}{2q}, -K_q^{-2q}\right\}\right],$$

with ${}_2F_1[\{\cdot\};\{\cdot\}]$ denoting a hypergeometric series.

For the proof see Theorem 1 of Claeskens et al. (2009). The constant $c(q,l,K_q)$ can be explicitly calculated for given values q, l and K_q and the hypergeometric series is converging for all $q, l \in \mathbb{N}$ with ${}_2F_1[\{l, l-1/(2q)\}; \{l+1-1/(2q), -K_q^{-2q}\}] \in (0,1]$, see e.g. Abramowitz and Stegun (1972, Ch. 15).

A.2 Proofs

The following lemma will be used in the proof of Theorem 1 and Lemma 3.

Lemma 2 *Under (A1) – (A3) for $f \in \mathcal{W}^{mq}[0,1]$, $m \in [1,2]$ it holds for any $l \in \mathbb{N}$*

$$\frac{1}{n}\sum_{i=q+1}^{k+p+1} \frac{\lambda b_i^2 n\eta_i}{(1+\lambda n\eta_i)^l} = \lambda \|\bar{s}_p^{(q)}\|^2\{1+o(1)\}, \tag{9}$$

$$\frac{1}{n}\sum_{i=q+1}^{k+p+1} \frac{(\lambda b_i n\eta_i)^2}{(1+\lambda n\eta_i)^l} \leq \lambda^m \|\bar{s}_p^{(mq)}\|^2\{1+o(1)\}, \tag{10}$$

with equality in (10) for $m = 2$.

Proof of Lemma 2

By definition $\|\bar{s}_p^{(mq)}\|^2 = n^{-1}\sum_{i=q+1}^{k+p+1} b_{p,i}^2 (n\eta_i)^m < M < \infty$ for M independent of $k \to \infty$, thereby $(n\eta_i)^m = c(\rho)^{2mq}(i-q)^{2mq}$ according to (8). Hence, $n^{-1}b_{p,i}^2(n\eta_i)^m$ decays exponentially with i and for any index $j = j(n)$, such that $j \to \infty$ as $n \to \infty$ and $j < k$, it holds for some $\varepsilon > 0$,

$$\|\bar{s}_p^{(mq)}\|^2 = \frac{1}{n}\sum_{i=q+1}^{k+p+1} b_{p,i}^2 (n\eta_i)^m = \frac{1}{n}\sum_{i=q+1}^{j} b_{p,i}^2 (n\eta_i)^m + O(j^{-\varepsilon}).$$

Let j be such that $\lambda n\eta_j = o(1)$, e.g. $j \propto \lambda^{-\alpha/(2q)}$ for any $\alpha \in (0,1)$ and $j < k$ following from (A2) and (A3). Then,

$$\begin{aligned}\frac{1}{n}\sum_{i=q+1}^{k+p+1} \frac{b_i^2 \lambda n\eta_i}{(1+\lambda n\eta_i)^l} &= \frac{\lambda}{n}\sum_{i=q+1}^{j} b_i^2 n\eta_i\{1+O(\lambda n\eta_j)\} + \frac{\lambda}{n}\sum_{i=j+1}^{k+p+1} \frac{b_i^2 n\eta_i}{(1+\lambda n\eta_i)^l} \\ &= \lambda\left[\sum_{i=q+1}^{j} b_i^2\eta_i\left\{1+O\left(\lambda^{1-\alpha}\right)\right\} + O\left(\lambda^{\varepsilon\alpha/(2q)}\right)\right] = \lambda\|\bar{s}_p^{(q)}\|^2\{1+o(1)\},\end{aligned}$$

where in the second sum $(1+\lambda n\eta_i)^{-l} \leq 1$ has been used. Also, for $f \in \mathcal{W}^{2q}[0,1]$

$$\frac{1}{n}\sum_{i=q+1}^{k+p+1} \frac{(b_i\lambda n\eta_i)^2}{(1+\lambda n\eta_i)^l} = \lambda^2\left[\sum_{i=q+1}^{j} b_i^2\eta_i^2 n\left\{1+O\left(\lambda^{1-\alpha}\right)\right\} + O\left(\lambda^{\varepsilon\alpha/(2q)}\right)\right] = \lambda^2\|\bar{s}_p^{(2q)}\|^2\{1+o(1)\}.$$

For $f \in \mathcal{W}^{mq}[0,1]$, $m \in [1,2)$ one can bound

$$\frac{1}{n}\sum_{i=q+1}^{k+p+1}\frac{(\lambda b_i n\eta_i)^2}{(1+\lambda n\eta_i)^l} \;\leq\; \frac{\lambda^m}{n}\sum_{i=q+1}^{k+p+1} b_i^2(n\eta_i)^m = \lambda^m\|\bar{s}_p^{(mq)}\|^2\{1+o(1)\},$$

proving the lemma. □

Proof of Theorem 1

For $f \in \mathcal{W}^{mq}[0,1]$ the lower bound for the smoothing parameter λ_f is found from

$$\begin{aligned}
0 &= \frac{\lambda^{1-m}}{2}\,\mathrm{E}_f\left\{\frac{\partial A_n(\lambda)}{\partial\lambda}\right\} = \frac{\boldsymbol{f}^t(\boldsymbol{I}_n-\boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda\boldsymbol{f}}{\lambda^m n} - \frac{\sigma^2\mathrm{tr}(\boldsymbol{S}_\lambda^2-\boldsymbol{S}_\lambda^3)}{\lambda^m n}\\
&= \frac{1}{n}\sum_{i=q+1}^{k+p+1} b_i^2(n\eta_i)^m\frac{(\lambda n\eta_i)^{2-m}}{(1+\lambda n\eta_i)^3} - \frac{\sigma^2}{\lambda^m n}\sum_{i=1}^{k+p+1}\frac{\lambda n\eta_i}{(1+\lambda n\eta_i)^3}\\
&\leq \|\bar{s}_p^{(mq)}\|^2\{1+o(1)\} - \lambda^{-(2qm+1)/(2q)}\frac{\sigma^2 c(q,2,K_q)}{4qn\;c(\rho)}\{1+o(1)\},
\end{aligned}$$

where Lemma 1 and Lemma 2 have been applied. Similarly, the smoothing parameter $\lambda_{r|f}$ is found as the solution to

$$\begin{aligned}
0 &= \frac{1}{\lambda}\mathrm{E}_f\left\{T_{ML}(\lambda)\right\} = \frac{1}{\lambda n}\boldsymbol{f}^t(\boldsymbol{S}_\lambda-\boldsymbol{S}_\lambda^2)\boldsymbol{f} - \frac{\sigma^2}{\lambda n}\left\{\mathrm{tr}(\boldsymbol{S}_\lambda^2)-q\right\}\{1+o(1)\}\\
&= \frac{1}{n}\sum_{i=q+1}^{k+p+1}\frac{b_i^2 n\eta_i}{(1+\lambda n\eta_i)^2} - \frac{\sigma^2}{\lambda n}\sum_{i=q+1}^{k+p+1}\frac{1}{(1+\lambda n\eta_i)^2}\{1+o(1)\}\\
&= \|\bar{s}_p^{(q)}\|^2\{1+o(1)\} - \lambda^{-(2q+1)/(2q)}\frac{\sigma^2 c(q,2,K_q)}{c(\rho)n}\{1+o(1)\}.
\end{aligned}$$

□

Proof of Theorem 2

For a $q\times q$ null matrix $\mathbf{0}_q$ and $\boldsymbol{\mathcal{R}} = \{\mathcal{R}(x_i,x_j)\}_{i,j=1}^n$, with

$$\mathcal{R}(x,s) = \mathrm{cov}\left\{\int_0^1\frac{(x-t)_+^{q-1}}{(q-1)!}dW(t), \int_0^1\frac{(s-t)_+^{q-1}}{(q-1)!}dW(t)\right\} = \int_0^1\frac{(x-t)_+^{q-1}}{(q-1)!}\frac{(s-t)_+^{q-1}}{(q-1)!}dt,$$

define a blockdiagonal matrix $\widetilde{\boldsymbol{\mathcal{R}}} = \mathrm{diag}(\mathbf{0}_q, \boldsymbol{\mathcal{R}})$ and for $l \in \{0, 1\}$

$$\widetilde{r}(l) = \frac{\mathrm{tr}\{(\widetilde{\boldsymbol{\mathcal{R}}} - \boldsymbol{C}\widetilde{\boldsymbol{D}}^{-}\boldsymbol{C}^t)(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda^l\}}{\mathrm{tr}\{\boldsymbol{C}\widetilde{\boldsymbol{D}}^{-}\boldsymbol{C}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda^l\}}.$$

Using the relationship $\mathrm{tr}\{\boldsymbol{C}\widetilde{\boldsymbol{D}}^{-}\boldsymbol{C}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda^l\} = \lambda n\ \mathrm{tr}\{(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda^{l+1}\}$ implies

$$\begin{aligned} n\ \mathrm{E}_\beta\{T_{Cp}(\lambda)\} &= \left[\sigma^2\mathrm{tr}\{(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda\} + \sigma_u^2\lambda n\ \mathrm{tr}(\boldsymbol{S}_\lambda^2 - \boldsymbol{S}_\lambda^3)\{1 + \widetilde{r}(1)\}\right]\{1 + o(1)\} \\ &- \left[\sigma^2\mathrm{tr}\{(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\} + \sigma_u^2\lambda n\ \mathrm{tr}(\boldsymbol{S}_\lambda - \boldsymbol{S}_\lambda^2)\{1 + \widetilde{r}(0)\}\right]\frac{\mathrm{tr}(\boldsymbol{S}_\lambda^2 - \boldsymbol{S}_\lambda^3)}{n} \\ &= \mathrm{tr}(\boldsymbol{S}_\lambda^2 - \boldsymbol{S}_\lambda^3)\left[\sigma_u^2\lambda n\{1 + o(1)\} - \sigma^2\{1 + o(1)\}\right]. \end{aligned}$$

Since $\lambda_{f|r}$ is defined as the solution to $\mathrm{E}_\beta\{T_{Cp}(\lambda)\} = 0$, the assertion of the theorem $\lambda_{f|r} = \sigma^2/(n\sigma_u^2)\{1 + o(1)\}$ follows, with $\widetilde{r}(1) = o(1)$ and $\widetilde{r}(0)\mathrm{tr}(\boldsymbol{S}_\lambda^2 - \boldsymbol{S}_\lambda^3)/n = o(1)$ shown in the Supplementary materials. □

The following lemma will be used in the proof of Theorem 3.

Lemma 3 *Let model (1) and assumptions (A1) – (A4) hold. Then, for any $l \in \mathbb{N}$*

$$\boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda^{1+l}\boldsymbol{f}\big|_{\lambda=\lambda_{r|f}} = \sigma^2\left\{\mathrm{tr}(\boldsymbol{S}_\lambda^2) - q\right\}\big|_{\lambda=\lambda_{r|f}}\{1 + o(1)\} \tag{11}$$

$$\boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^{1+l}\boldsymbol{S}_\lambda\boldsymbol{f}\big|_{\lambda=\lambda_{r|f}} = o\left(\lambda_{r|f}^{-1/(2q)}\right). \tag{12}$$

Moreover,

$$\boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda^{1+l}\boldsymbol{f}\big|_{\lambda=\lambda_f} = \sigma^2\mathrm{tr}\left(\boldsymbol{S}_\lambda^2 - \boldsymbol{S}_\lambda^3\right)\big|_{\lambda=\lambda_f}\{1 + o(1)\} \tag{13}$$

$$\boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^{2+l}\boldsymbol{S}_\lambda\boldsymbol{f}\big|_{\lambda=\lambda_f} = o\left(\lambda_f^{-1/(2q)}\right). \tag{14}$$

Proof of Lemma 3

From $\boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda^{1+l}\boldsymbol{f} = \sum_{i=q+1}^{k+p+1}\lambda b_i n\eta_i(1 + \lambda n\eta_i)^{-(l+2)}$ and (9) one concludes that $\boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda^{1+l}\boldsymbol{f} = \boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda\boldsymbol{f}\{1 + o(1)\}$. Equation (11) follows from the definition of $\lambda_{r|f}$ via $\left[\boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda\boldsymbol{f} - \sigma^2\left\{\mathrm{tr}(\boldsymbol{S}_\lambda^2) - q\right\} + o(1)\right]\big|_{\lambda=\lambda_{r|f}} = 0$, and (12) holds due to

$$\boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda\boldsymbol{f}\big|_{\lambda=\lambda_{r|f}} = \boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda\boldsymbol{f}\big|_{\lambda=\lambda_{r|f}} - \boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda^2\boldsymbol{f}\big|_{\lambda=\lambda_{r|f}} = o\left(\lambda_{r|f}^{-1/(2q)}\right).$$

Equations (13) and (14) are proved analogously, making use of the definition of λ_f via $\left\{\boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda\boldsymbol{f} - \sigma^2\mathrm{tr}(\boldsymbol{S}_\lambda^2 - \boldsymbol{S}_\lambda^3)\right\}\big|_{\lambda=\lambda_f} = 0$ and

$$\begin{aligned}\boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda^{1+l}\boldsymbol{f} &= \boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda^{1+l}\boldsymbol{f} - \boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda^{2+l}\boldsymbol{f} \\ &= \boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda\boldsymbol{f}\{1+o(1)\} - \boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda^2\boldsymbol{f}\{1+o(1)\} \\ &= \boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda\boldsymbol{f}\{1+o(1)\}.\end{aligned}$$

□

Proof of Theorem 3

The theorem is proved for $\widehat{\lambda}_f$ only, the proof for $\widehat{\lambda}_r$ is completely analogous and is given in the Supplementary materials to this article.

From the Taylor expansion $0 = T_{Cp}(\widehat{\lambda}_f) = T_{Cp}(\lambda_f) + T'_{Cp}(\widetilde{\lambda})(\widehat{\lambda}_f - \lambda_f)$, for some $\widetilde{\lambda}$ between $\widehat{\lambda}_f$ and λ_f, it holds $\widehat{\lambda}_f - \lambda_f = -T_{Cp}(\lambda_f)/T'_{Cp}(\widetilde{\lambda})$. Hence, one needs to show

$$\frac{T_{Cp}(\lambda_f) - \mathrm{E}_f\{T_{Cp}(\lambda_f)\}}{\sqrt{\mathrm{var}_f\{T_{Cp}(\lambda_f)\}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \quad \text{and} \quad \frac{T'_{Cp}(\widetilde{\lambda})}{\mathrm{E}_f\left\{T'_{Cp}(\lambda_f)\right\}} \xrightarrow{\mathcal{P}} 1.$$

Expressions for $\mathrm{E}_f\left\{T'_{Cp}(\lambda_f)\right\}$ and $\mathrm{var}_f\{T_{Cp}(\lambda_f)\}$, given by

$$\begin{aligned}\mathrm{E}_f\left\{T'_{Cp}(\lambda_f)\right\} &= \frac{\sigma^2}{\lambda_f n}\,\mathrm{tr}\{(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)\boldsymbol{S}_\lambda^2(4\boldsymbol{I}_n - 3\boldsymbol{S}_\lambda)\}\big|_{\lambda=\lambda_f}\{1+o(1)\}, \\ \mathrm{var}_f\{T_{Cp}(\lambda_f)\} &= \frac{2\sigma^4}{n^2}\,\mathrm{tr}\{(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^4\boldsymbol{S}_\lambda^2\}\big|_{\lambda=\lambda_f}\{1+o(1)\},\end{aligned}$$

are derived employing Lemma 3 and $\partial\boldsymbol{S}_\lambda/\partial\lambda = -\lambda^{-1}(\boldsymbol{S}_\lambda - \boldsymbol{S}_\lambda^2)$. More details on these equations are given in the Supplementary materials. Next, consider

$$\begin{aligned}n\,T_{Cp}(\lambda_f) &= \left[\boldsymbol{Y}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda\boldsymbol{Y}\{1+o(1)\} - \widehat{\sigma}^2\mathrm{tr}(\boldsymbol{S}_\lambda - \boldsymbol{S}_\lambda^2)\right]\big|_{\lambda=\lambda_f}, \\ \mathrm{E}_f\{n\,T_{Cp}(\lambda_f)\} &= \left[\boldsymbol{f}^t(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda\boldsymbol{f} + \sigma^2\mathrm{tr}\{(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2\boldsymbol{S}_\lambda\} - \sigma^2\mathrm{tr}(\boldsymbol{S}_\lambda - \boldsymbol{S}_\lambda^2)\right]\big|_{\lambda=\lambda_f} + o(1).\end{aligned}$$

Denoting $d_i = \sum_{j=1}^{n} \phi_{k,i}(x_j) y_j$ with $\mathrm{E}_f(d_i^2) = b_i^2 + \sigma^2$, and noting that $\widehat{\sigma}^2 = \sigma^2\{1 + O_p(n^{-1/2})\}$, define random variables ξ_i

$$n\left[T_{Cp}(\lambda_f) - \mathrm{E}_f\{T_{Cp}(\lambda_f)\}\right] = \sum_{i=q+1}^{k+p+1} \left(d_i^2 - b_i^2 - \sigma^2\right) \frac{(\lambda_f n \eta_i)^2}{(1+\lambda_f n \eta_i)^3} + o_p(1) =: \sum_{i=q+1}^{k+p+1} \xi_i,$$

such that $\mathrm{E}_f(\xi_i) = o(1)$ and $s_n^2 = \sum_{i=q+1}^{k+p+1} \mathrm{var}_f(\xi_i) = 2\sigma^4 \mathrm{tr}\{(\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^4 \boldsymbol{S}_\lambda^2\}\{1+o(1)\}$. Since $s_n^2 = \mathrm{const}\ \lambda_f^{-1/(2q)}$ and $(\lambda_f^{1/(2q)} k)^{-1} \to 0$, according to (A2) and (A3), each $\mathrm{var}_f(\xi_i) = o(1)$ and there exist a constant B, such that $\mathrm{E}_f|\xi_i|^2 = \mathrm{var}_f(\xi_i) + o(1) < B$, $i = q+1, \ldots, k+p+1$. With this, the Lyapunov's condition

$$s_n^{-4} \sum_{i=q+1}^{k+p+1} \mathrm{E}_f|\xi_i|^4 < B s_n^{-4} \sum_{i=q+1}^{k+p+1} \mathrm{E}_f|\xi_i|^2 = B s_n^{-2} = O\left(\lambda_f^{1/(2q)}\right)$$

converges to zero as $n \to \infty$. This proves $s_n^{-1} \sum_{i=q+1}^{k+p+1} \xi_i \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$, or equivalently, $[\mathrm{var}_f\{T_{Cp}(\lambda_f)\}]^{-1/2} \left[T_{Cp}(\lambda_f) - \mathrm{E}_f\{T_{Cp}(\lambda_f)\}\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.

Next is shown that $\widehat{\lambda}_f \xrightarrow{\mathcal{P}} \lambda_f$. From $\mathrm{var}_f\{T_{Cp}(\lambda)\} = O\left(\lambda^{-1/(2q)} n^{-2}\right) \to 0$, $n \to \infty$ it follows that $T_{Cp}(\lambda) \xrightarrow{\mathcal{P}} \mathrm{E}_f\{T_{Cp}(\lambda)\}$, for any λ satisfying (A3). It remains to verify that $\mathrm{E}_f\left[T_{Cp}\{\lambda_f(1-\varepsilon)\}\right] < 0 < \mathrm{E}_f\left[T_{Cp}\{\lambda_f(1+\varepsilon)\}\right]$, for any $\varepsilon \in (0,1)$ (see Lemma 5.10 in van der Vaart, 1998). As shown in the Supplementary materials, for $B_1(\lambda) = n^{-1} \boldsymbol{f}^t (\boldsymbol{I}_n - \boldsymbol{S}_\lambda)^2 \boldsymbol{S}_\lambda \boldsymbol{f} > 0$ and $n \to \infty$ it holds

$$\begin{aligned}
\mathrm{E}_f[T_{Cp}\{\lambda_f(1-\varepsilon)\}] &= (1-\varepsilon)^2 B_1(\lambda_f) \left\{1 - (1-\varepsilon)^{-2-1/(2q)} + o(1)\right\} < 0, \\
\mathrm{E}_f\left[T_{Cp}\{\lambda_f(1+\varepsilon)\}\right] &= (1+\varepsilon)^2 B_1(\lambda_f) \left\{1 - (1+\varepsilon)^{-2-1/(2q)} + o(1)\right\} > 0,
\end{aligned}$$

so that $\widehat{\lambda}_f \xrightarrow{\mathcal{P}} \lambda_f$ follows.

Let now consider $T'_{Cp}(\tau\lambda_f) / \mathrm{E}_f\{T'_{Cp}(\lambda_f)\}$, where $\tau \in [1-\varepsilon, 1+\varepsilon]$, for any bounded $\varepsilon > 0$. Since $\mathrm{var}_f\left\{T'_{Cp}(\tau\lambda_f)\right\} = (\tau\lambda_f)^{-2-1/(2q)} n^{-2} \mathrm{const}\{1 + o(1)\}$, it is easy to see that

$$\mathrm{var}_f\left[\frac{T'_{Cp}(\tau\lambda_f)}{\mathrm{E}_f\left\{T'_{Cp}(\lambda_f)\right\}}\right] = O\left(\lambda_f^{1/(2q)}\right) \to 0, \quad n \to \infty.$$

With Lemma 3, $\mathrm{E}_f\left\{T'_{Cp}(\tau\lambda_f)\right\} = \mathrm{E}_f\{T'_{Cp}(\lambda_f)\}(4q\tau + \tau^{-1-1/(2q)})/(4q+1)\{1+o(1)\}$, where

$$\frac{4q\,\tau + \tau^{-1-1/(2q)}}{4q+1} = \begin{cases} 1-\varepsilon\{1-1/(2q)\} + O(\varepsilon^2), & \text{for } \tau = 1-\varepsilon \\ 1+\varepsilon\{1-1/(2q)\} + O(\varepsilon^2), & \text{for } \tau = 1+\varepsilon \end{cases},$$

so that for any fixed $\tau \in [1-\varepsilon, 1+\varepsilon]$ it holds $T'_{Cp}(\tau\lambda_f)/\mathrm{E}_f\{T'_{Cp}(\lambda_f)\} \xrightarrow{\mathcal{P}} 1$, as $n \to \infty$. Since $P(|\widetilde{\lambda}/\lambda_f - 1| \leq \varepsilon) \to 1$ for $n \to \infty$ and any $\varepsilon > 0$ due to $\widehat{\lambda}_f \xrightarrow{\mathcal{P}} \lambda_f$, it follows that $T'_{Cp}(\widetilde{\lambda})/\mathrm{E}_f\left\{T'_{Cp}(\lambda_f)\right\} \xrightarrow{\mathcal{P}} 1$, $n \to \infty$. Putting all together and applying Slutsky's lemma proves the theorem. □

Proof of Theorem 4

The proof of Theorem 4 is analogous to that of Theorem 3 and is given in the Supplementary materials to this article.

References

Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institue of Statistical Mathematics*, 21:243 – 47.

Central Bureau of Statistics (CBS) Kenya, Ministry of Health (MOH) Kenya, and ORC Macro (2004). *Kenya Demographic and Health Survey 2003*. CBS, MOH, and ORC Macro, Calverton, Maryland.

Claeskens, G., Krivobokova, T., and Opsomer, J. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(6):529–544.

Cox, D. D. (1988). Approximation of method of regularization estimators. *Ann. Statist.*, 16:694–712.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. Estimating

the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31(4):377–403.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer, New York. Revised Edition.

Durban, M. and Currie, I. (2003). A note on p-spline smoothing with b-splines and penalties. *Comput. Statist.*, 11(2):89–121.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, 14:731–761.

Härdle, W., Hall, P., and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.*, 83(404):86–95.

Kauermann, G. (2005). A note on smoothing parameter selection for penalised spline smoothing. *J. Statist. Plann. Infer.*, 127:53–69.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Statist.*, 41(2):495–502.

Kohn, R., Ansley, C., and Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *J. Amer. Statist. Assoc.*, 86(416):1042–1050.

Kou, S. C. (2003). On the efficiency of selection criteria in spline regression. *Prob. Theory Relat. Fields*, 127:152–176.

Lukas, M. (1993). Asymptotic optimality of generalized cross-validation for choosing the regularization parameter. *Numer. Math.*, 66:41–66.

Lukas, M. (1995). On the discrepancy principle and generalized maximum likelihood for regularization. *Bull. Austral. Math. Soc.*, 52:399–424.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

Reiss, P. T. and Ogden, R. T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(2):505–523.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.*, 12:1215–1230.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statist. Science*, 6(1):15–51. With comments and a rejoinder by the author.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge, UK.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.*, 13(3):970–983.

Speckman, P. and Sun, D. (2001). Asymptotic properties of smoothing parameter selection in spline smoothing. Technical report, University of Missouri.

Stein, M. L. (1990). A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *Ann. Statist.*, 18(3):1139–1157.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press, New York.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.*, 13(4):1378–1402.

Wahba, G. (1990). *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Wand, M. and Ormerod, J. (2008). On semiparametric regression with O'Sullivan penalised splines. *ANZJS*, 50:179–198.

WHO (1998). WHO child growth standards based on length/height, weight, and age. *Acta Paediatrica*, 450.