Biometrika (2009), **xx**, x, *pp*. 1–26

Printed in Great Britain

Asymptotic properties of penalized spline estimators

BY GERDA CLAESKENS

Katholieke Universiteit Leuven, ORSTAT and Leuven Statistics Research Center, Naamsestraat 69, B-3000 Leuven, Belgium.

gerda.claeskens@econ.kuleuven.be

TATYANA KRIVOBOKOVA

Georg-August-Universität Göttingen, CRC Poverty, Equity and Growth, Platz der Göttinger Sieben 3, D-37073 Göttingen, Germany.

tatyana.krivobokova@wiwi.uni-goettingen.de

JEAN D. OPSOMER

Colorado State University Department of Statistics, Fort Collins, CO 80523, USA

jopsomer@stat.colostate.edu

SUMMARY

We study the class of penalized spline estimators, which enjoy similarities to both regression splines, without penalty and with less knots than data points, and smoothing splines, with knots equal to the data points and a penalty controlling the roughness of the fit. Depending on the number of knots, sample size and penalty, we show that the theoretical properties of penalized regression spline estimators are either similar to those of regression splines or to those of smoothing splines, with a clear breakpoint distinguishing the cases. We prove that using less knots results in better asymptotic rates than when using a large number of knots. We obtain expressions for bias and variance and asymptotic rates for the number of knots and penalty parameter.

Some key words: Mean squared error; Nonparametric regression; Penalty; Regression splines; Smoothing splines.

1. INTRODUCTION

Penalized spline smoothing has gained much popularity over the last decade. This smoothing technique with flexible choice of bases and penalties can be viewed as a compromise between regression and smoothing splines. In this paper we obtain asymptotic properties of such estimators and relate them to known asymptotic results for regression splines and smoothing splines, which can be seen as the two extreme cases, with penalized splines situated in between.

The combination of regression splines, with number of knots less than the sample size, and a penalty has been studied by several authors. O'Sullivan (1986) used penalized fitting with cubic B-splines for inverse problems. He used a set of knots different from the data and a penalty equal to the integrated squared second derivative of the spline function. O'Sullivan splines are discussed by Ormerod & Wand (2008). Kelly & Rice (1990) and Besse et al. (1997) used B-spline approximations to the smoothing splines, which they called hybrid splines. Schwetlick & Kunert (1993) decoupled the order of the B-spline and the derivative in the penalty function. This same idea has been promoted by Eilers & Marx (1996) who used a difference penalty on the spline coefficients. Many applications and examples of penalized splines are presented in Ruppert et al. (2003).

There is a rich literature on smoothing splines, which we shall only briefly touch here. Reference books are Wahba (1990), Green & Silverman (1994) and Eubank (1999). For smoothing splines, the penalty is the integrated squared qth derivative of the function, leading to a smoothing spline of degree $2q-1$, with $q=2$ a common choice. Rice & Rosenblatt (1981, 1983) study the estimator's integrated mean squared error and effects of boundary bias, see also Oehlert (1992) and Utreras (1988). Wahba (1975) and Craven & Wahba (1978) investigated the averaged mean squared error, in connection with the choice of the smoothing parameter. Cox (1983) studied convergence rates for robust smoothing splines. Speckman (1985) obtained the optimal rates of convergence for smoothing spline estimators, and Nychka (1995) obtained local properties of smoothing splines.

For regression splines, the integrated mean squared error was studied by Agarwal & Studden (1980), and Huang (2003a,b) who obtained local asymptotic results by considering the least squares estimator as an orthogonal projection. Important theoretical results on unpenalized regression splines are obtained by Zhou et al. (1998).

Theoretical properties of penalized spline estimators are less explored. Some first results can be found in Hall & Opsomer (2005), who used a white noise representation of the model to obtain the mean squared error and consistency of the estimator. Kauermann et al. (2008) work with generalized linear models. Li & Ruppert (2008) used an equivalent kernel representation for piecewise constant and linear B-splines and first or second order difference penalties. Their assumption on the relative large number of knots, thus close to the smoothing splines case, allowed them to ignore the approximation bias.

In this paper we provide a general treatment, any order of spline and general penalty, and we study with one theory the two asymptotic situations, either close to regression splines or close to smoothing splines. One of our main results is that we find a clear "breakpoint" in the asymptotic properties of the penalized splines, with the boundary between the two types of behavior depending on an explicitly defined function of the number of knots, the sample size and the penalty parameter. Depending on the value of this function, the asymptotic results are related to those of regression splines or to those of smoothing splines. An interesting finding is that it is better to use a smaller number of knots, thus close to the regression splines case, since that results in a smaller mean squared error.

2. ESTIMATION WITH SPLINES

2·1. *Notation and model assumptions*

Based on data (Y_i, x_i), with fixed $x_i \in [a, b]$, $i = 1, \ldots, n$ and $a, b < \infty$ with true relationship

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

we aim to estimate the unknown smooth function $f(\cdot) \in C^{p+1}([a, b])$, a $p + 1$ times continuously differentiable function, with penalized splines. The residuals ε_i are assumed to be uncorrelated with zero mean and variance $\sigma^2 > 0$.

2·2. *Penalized splines with B-spline basis functions*

The idea of penalized spline smoothing with B-spline basis functions traces back to O'Sullivan (1986), see also Schwetlick & Kunert (1993). Classically, B-splines are defined recursively, see de Boor (2001, ch. IX). Let the value p denote the degree of the B-spline, implying that the order equals $p + 1$. On an interval $[a, b]$, define a sequence of knots $a = \kappa_0 < \kappa_1 < \cdots < \kappa_K < \kappa_{K+1} = b$. In addition, define p knots $\kappa_{-p} = \kappa_{-p+1} = \cdots = \kappa_{-1} = \kappa_0$ and another set of p knots $\kappa_{K+1} = \kappa_{K+2} = \cdots = \kappa_{K+p+1}$. The B-spline basis functions are defined as

$$N_{j,1}(x) = \begin{cases} 1, \; \kappa_j \leq x < \kappa_{j+1} \\ 0, \text{ otherwise} \end{cases},$$

$$N_{j,p+1}(x) = \frac{x - \kappa_j}{\kappa_{j+p} - \kappa_j} N_{j,p}(x) + \frac{\kappa_{j+p+1} - x}{\kappa_{j+p+1} - \kappa_{j+1}} N_{j+1,p}(x),$$

for $j = -p, \ldots, K$. Thereby the convention $0/0 = 0$ is used. With the use of the additional knots, this gives precisely $K + p + 1$ basis functions.

We define the penalized spline estimator as the minimizer of

$$\sum_{i=1}^{n} \{Y_i - \sum_{j=-p}^{K} \beta_j N_{j,p+1}(x_i)\}^2 + \lambda \int_a^b [\{\sum_{j=-p}^{K} \beta_j N_{j,p+1}(x)\}^{(q)}]^2 dx, \tag{2}$$

where the penalty is the integrated squared qth order derivative of the spline function, which is assumed to be finite. Since the $(p+1)$st derivative of a spline function of degree $p+1$ contains Dirac delta functions, it is a natural condition to have $q \leq p$. However, in Section 5 we treat the case of truncated polynomial basis functions where $q = p+1$. The penalty constant λ plays the role of a smoothing parameter. For a fixed n, letting $\lambda \to 0$ implies an unpenalized estimate, while $\lambda \to \infty$ forces convergence of the qth derivative of the spline function to zero, with the consequence that the limiting estimator is a $(q-1)$th degree polynomial. From the derivative formula for B-spline functions (de Boor (2001), ch. X),

$$\{\sum_{j=-p}^{K} \beta_j N_{j,p+1}(x)\}^{(q)} = \sum_{j=-p+q}^{K} N_{j,p+1-q}(x)\beta_j^{(q)},$$

where the coefficients $\beta_j^{(q)}$ are defined recursively via

$$\begin{aligned}\beta_j^{(1)} &= p(\beta_j - \beta_{j-1})/(\kappa_{j+p} - \kappa_j),\\ \beta_j^{(q)} &= (p+1-q)(\beta_j^{(q-1)} - \beta_{j-1}^{(q-1)})/(\kappa_{j+p+1-q} - \kappa_j),\ q = 2, 3, \ldots.\end{aligned} \tag{3}$$

We rewrite the penalty term in (2) as $\lambda \beta^t \Delta_q^t R \Delta_q \beta$, where the matrix R has elements $R_{ij} = \int_a^b N_{j,p+1-q}(x) N_{i,p+1-q}(x) dx$, for $i, j = -p+q, \ldots, K$ and Δ_q denotes the matrix corresponding to the weighted difference operator defined in (3), i.e. $\beta^{(q)} = \Delta_q \beta$. For the special case of equidistant knots, i.e. $\kappa_j - \kappa_{j-1} = \delta$ for any $j = -p+1, \ldots, K$, there is an explicit expression of the matrix Δ_q in terms of the matrix ∇_q, corresponding to the qth difference operator, defined recursively via $\beta_j^{(1)} = \beta_j - \beta_{j-1}$, $\beta_j^{(q)} = \beta_j^{(q-1)} - \beta_{j-1}^{(q-1)}$, $q = 2, 3, \ldots$. In this case, $\Delta_q = \delta^{-q} \nabla_q$.

Further, define the spline basis vector of dimension $1 \times (K+p+1)$ as $N(x) = \{N_{-p,p+1}(x), \ldots, N_{K,p+1}(x)\}$, the $n \times (K+p+1)$ spline design matrix $N = \{N(x_1)^t, \ldots, N(x_n)^t\}^t$, and let $D_q = \Delta_q^t R \Delta_q$. With this notation, the penalized spline estimator takes the form of a ridge regression estimator

$$\widehat{f} = N(N^t N + \lambda D_q)^{-1} N^t Y, \tag{4}$$

where $\widehat{f} = \{\hat{f}(x_1), \ldots, \hat{f}(x_n)\}^t$ and $Y = (Y_1, \ldots, Y_n)^t$. This estimator has been considered in Ormerod & Wand (2008), who gave it the name O'Sullivan spline, or just O-spline, estimator and presented an efficient algorithm for computation of the matrix D_q. A slightly modified version of (4), known as the P-spline estimator, has been introduced by Eilers & Marx (1996). They used equidistant knots and a combination of cubic splines ($p = 3$) and second order penalty ($q = 2$). Moreover, only the diagonal elements of the tridiagonal matrix R were taken into account, resulting in the simpler penalty matrix $c\delta^{-4}\nabla_2^t\nabla_2$, with $c = \int_a^b \{N_{j,2}(x)\}^2 dx$. Since c and δ are constants, they can be absorbed in the penalty constant. Eilers & Marx (1996) motivated the difference penalty as a good approximation to the penalty D_q. Since these simplifications do not influence the asymptotic properties of the estimator, we use the general estimator (4) for our theoretical investigation.

2·3. *Regression splines*

An unpenalized estimator with $\lambda = 0$ in (4) is referred to as a regression spline estimator. More precisely, the regression spline estimator of order $(p+1)$ for $f(x)$ is the minimizer of

$$\sum_{i=1}^{n} \{Y_i - \hat{f}_{\mathrm{reg}}(x_i)\}^2 = \min_{s(x) \in S(p+1;\kappa)} \sum_{i=1}^{n} \{Y_i - s(x_i)\}^2,$$

where

$$S(p+1;\kappa) = \Big\{ s(\cdot) \in C^{p-1}[a,b] : \ s \text{ is a degree } p \text{ polynomial on each } [\kappa_j, \kappa_{j+1}] \Big\}, \ p > 0,$$

is the set spline functions of degree p with knots $\kappa = \{a = \kappa_0 < \kappa_1 < \cdots < \kappa_K < \kappa_{K+1} = b\}$ and $S(1;\kappa)$ is the set of step functions with jumps at the knots. Since $N_{j,p+1}(\cdot)$, $j = -p, \ldots, K$

form a basis for $S(p+1;\kappa)$, see Schumaker (1981, ch. 4),

$$\hat{f}_{\text{reg}}(x) = N(x)(N^tN)^{-1}N^tY \in S(p+1,\kappa). \tag{5}$$

Further, we denote with $s_f(\cdot) = N(\cdot)\beta \in S(p+1,\kappa)$ the best L_∞ approximation to function f. The asymptotic properties of the regression spline estimator $\hat{f}_{\text{reg}}(x)$ have been studied in Zhou et al. (1998), where the following assumptions are stated.

(A1) Let $\delta = \max_{0\le j\le K}(\kappa_{j+1}-\kappa_j)$. There exists a constant $M>0$, such that $\delta/\min_{0\le j\le K}(\kappa_{j+1}-\kappa_j) \le M$ and $\delta = o(K^{-1})$.

(A2) For deterministic design points $x_i \in [a,b]$, $i=1,\ldots,n$, assume that there exists a distribution function Q with corresponding positive continuous design density ρ such that, with Q_n the empirical distribution of $x_1,\ldots,x_n$, $\sup_{x\in[a,b]}|Q_n(x)-Q(x)| = o(K^{-1})$.

(A3) The number of knots $K = o(n)$.

Zhou et al. (1998) obtained the approximation bias and variance as

$$\mathrm{E}\{\hat{f}_{\text{reg}}(x)\} - f(x) = b_a(x) + o(\delta^{p+1}), \tag{6}$$

$$\mathrm{var}\{\hat{f}_{\text{reg}}(x)\} = \frac{\sigma^2}{n}N(x)G^{-1}N^t(x) + o\{(n\delta)^{-1}\}, \tag{7}$$

where $G = \int_a^b N(x)^tN(x)\rho(x)dx$ and the approximation bias

$$b_a(x;p+1) = -\frac{f^{(p+1)}(x)}{(p+1)!}\sum_{j=0}^{K} I_{[\kappa_j,\kappa_{j+1})}(x)(\kappa_{j+1}-\kappa_j)^{p+1}B_{p+1}\left(\frac{x-\kappa_j}{\kappa_{j+1}-\kappa_j}\right), \tag{8}$$

with $B_{p+1}(\cdot)$ the $(p+1)$th Bernoulli polynomial, see p. 804 of Abramowitz & Stegun (1972).

2·4. *Smoothing splines*

The smoothing spline estimator for $f(\cdot)$ in (1) arises as a solution of the minimization problem

$$\min_{f\in W^q[a,b]}\left[\sum_{i=1}^{n}\{Y_i - f(x_i)\}^2 + \lambda\int_a^b\{f^{(q)}(x)\}^2dx\right], \tag{9}$$

where $\lambda > 0$ and $W^q[a,b]$ denotes the Sobolev space of order q, i.e. $W^q[a,b] = \{f: \ f$ has $q-1$ absolute continuous derivatives, $\int_a^b \{f^{(q)}(x)\}^2 dx < \infty\}$. It turns out that $\hat{f}_{\rm ss}(x)$, the solution of (9), is the natural polynomial spline function of degree $2q-1$ with knots at x_i. Namely, $\hat{f}_{\rm ss}(x)$ is a polynomial of degree $q-1$ on $[x_1, x_2]$ and $[x_{n-1}, x_n]$ and of degree $2q-1$ on (x_i, x_{i+1}), $i = 2, \ldots, n-2$ with jumps in the $(2q-1)$st derivative only at the knots. It has been proven, see e.g. Utreras (1985), that $\mathrm{E}\{(f - \hat{f}_{\rm ss})^2\} = O(\lambda/n) + \sigma^2 O(n^{1/(2q)-1}\lambda^{-1/(2q)})$, so that $\lambda = O(n^{1/(1+2q)})$ provides the optimal rate of convergence, as long as $\lambda n^{2q-1} \to \infty$. The differentiability assumption for smoothing splines ($f \in W^q$) is weaker compared to regression splines case ($f \in C^{p+1}$) if $p \geq q$. We refer to Eubank (1999) for further discussion of the theoretical properties of smoothing splines.

3. AVERAGE MEAN SQUARED ERROR OF THE PENALIZED SPLINE ESTIMATOR

We investigate the average mean squared error (AMSE) of the penalized spline estimator and discuss the optimum choice of smoothing parameter λ and number of knots K. Similar asymptotic results could be obtained using the mean integrated squared error (MISE) instead of the average mean squared error. Compare, for example, Wahba (1975) for the average mean squared error and Rice & Rosenblatt (1981) for the mean integrated squared error for smoothing splines or Zhou et al. (1998) for the average mean squared error and Agarwal & Studden (1980) for the mean integrated squared error for regression splines. With the Demmler & Reinsch (1975) decomposition, the average bias and variance can be expressed in terms of the eigenvalues obtained from the singular value decomposition

$$(N^t N)^{-1/2} D_q (N^t N)^{-1/2} = U \mathrm{diag}(s) U^t, \tag{10}$$

where U is the matrix of eigenvectors and s is the vector of eigenvalues s_j. Denote $A = N(N^t N)^{-1/2} U$. This matrix is semi-orthogonal with $A^t A = I_{K+p+1}$ and $AA^t =$

$N(N^tN)^{-1}N^t$. We can rewrite the penalized spline estimator (4) as

$$\widehat{f} = A\{I_n + \lambda\,\mathrm{diag}(s)\}^{-1}A^tY \tag{11}$$

$$= \{I_n + \lambda\, A\mathrm{diag}(s)A^t\}^{-1}AA^tY = \{I_n + \lambda\, A\mathrm{diag}(s)A^t\}^{-1}\widehat{f}_{\mathrm{reg}}. \tag{12}$$

Equation (12) clearly shows the shrinkage effect of including the penalty term. Equality (11) provides an expression that is straightforward to use to obtain the average mean squared error

$$\begin{aligned}\mathrm{AMSE}(\widehat{f}) &= \frac{1}{n}E\{(\widehat{f}-f)^t(\widehat{f}-f)\} \\ &= \frac{\sigma^2}{n}\sum_{j=1}^{K+p+1}\frac{1}{(1+\lambda s_j)^2} + \frac{\lambda^2}{n}\sum_{j=1}^{K+p+1}\frac{s_j^2 b_j^2}{(1+\lambda s_j)^2} + \frac{1}{n}f^t(I_n - AA^t)f,\end{aligned}$$

where $f = \{f(x_1), \cdots, f(x_n)\}^t$ and $b = A^tf$ with components b_j. Since AA^t is idempotent and $AA^tf = \mathrm{E}(\hat{f}_{\mathrm{reg}})$ we obtain that

$$\begin{aligned}\mathrm{AMSE}(\widehat{f}) = &\sum_{j=1}^{K+p+1}\frac{\sigma^2}{n(1+\lambda s_j)^2} + \sum_{j=1}^{K+p+1}\frac{\lambda^2 s_j^2 b_j^2}{n(1+\lambda s_j)^2} \\ &+ \frac{1}{n}\sum_{j=1}^{n}\left[E\{\hat{f}_{\mathrm{reg}}(x_j)\} - f(x_j)\right]^2.\end{aligned} \tag{13}$$

The first term in (13) is the average variance, the second term is the average squared shrinkage bias which is due to the penalization, and the last term is the average squared approximation bias, which can be obtained from (6) and is due to representing an arbitrary function by a linear combination of spline functions.

We now study the optimal orders of the smoothing parameter λ and of the number of knots K. With the constant $\tilde{c}_1$ introduced in Lemma 3 in the Appendix, define

$$K_q = (K + p + 1 - q)(\lambda\tilde{c}_1)^{1/(2q)}n^{-1/(2q)}. \tag{14}$$

THEOREM 1. *Under assumptions (A1)–(A3) the following statements hold:*

(a) If $K_q < 1$ and $f(\cdot) \in C^{p+1}[a,b]$,

$$AMSE(\widehat{f}) = O\left(\frac{K}{n}\right) + O\left(\frac{\lambda^2}{n^2}K^{2q}\right) + O(K^{-2(p+1)}),$$

and for $K \sim C_1 n^{1/(2p+3)}$, *with* C_1 *a constant, and* $\lambda = O(n^{\gamma})$ *with* $\gamma \leq (p+2-q)/(2p+3)$, *the penalized spline estimator attains the optimal rate of convergence for* $f \in C^{p+1}[a,b]$ *with AMSE*$(\widehat{f}) = O(n^{-(2p+2)/(2p+3)})$.

(b) If $K_q \geq 1$ *and* $f(\cdot) \in W^q[a,b]$,

$$AMSE(\widehat{f}) = O\left(\frac{n^{1/(2q)-1}}{\lambda^{1/(2q)}}\right) + O\left(\frac{\lambda}{n}\right) + O(K^{-2q}),$$

and for $\lambda = O(n^{1/(2q+1)})$, *such that* $\lambda n^{2q-1} \to \infty$ *and* $K \sim C_2 n^{\nu}$ *with* $\nu \geq 1/(2q+1)$ *and* C_2 *a constant, the penalized spline estimator attains the optimal rate of convergence for* $f \in W^q[a,b]$ *with AMSE*$(\widehat{f}) = O(n^{-2q/(1+2q)})$.

Case (a) with $K_q < 1$ results in the asymptotic scenario similar to that of regression splines. The average mean squared error is determined by the average asymptotic variance and squared approximation bias. The shrinkage bias becomes negligible for small λ, that is for $\gamma < (p+2-q)/(2p+3)$. The asymptotically optimal number of knots has the same order as that for regression splines, that is $K \sim C_1 n^{1/(2p+3)}$. Case (b) with $K_q \geq 1$ results in the asymptotic scenario close to that of smoothing splines. The average mean squared error is dominated by the average asymptotic variance and squared shrinkage bias. The average squared approximation bias is of the same asymptotic order as the average shrinkage bias for $K_q = 1$ and of negligible order for $K_q > 1$. The asymptotic order of the average mean squared error depends only on the order of the penalty q and the bound of the average mean squared error is precisely the same as known from the smoothing spline theory, up to the average squared approximation bias, which is negligible for $K_q > 1$.

The assumption on the smoothness of the function f can be somewhat weakened in case (a). The assumption $f \in C^{p+1}$ can be replaced by a slightly weaker assumption $f \in W^{p+1}$, since according to Barrow & Smith (1978) the expression for the approximation bias (8) holds for $f(\cdot) \in W^{p+1}$ as well. See also the discussion in Agarwal & Studden (1980), Remark 3.3.

The result of Theorem 1 suggests that the convergence rate of penalized spline estimators is faster if $K_q < 1$, since $q \leq p$ is assumed. Thus, it is advisable to prefer a small number of knots in practice. However, there is still a need for a practical guideline for choosing K and λ, so that $K_q < 1$ is satisfied. This is planned to be addressed in a separate work.

4. ASYMPTOTIC BIAS AND VARIANCE

We derive the pointwise asymptotic bias and variance in both asymptotic scenarios.

THEOREM 2. *Under assumptions $(A1)-(A3)$, the following statements hold:*

(a) If $K_q < 1$ and $f(\cdot) \in C^{p+1}[a,b]$,

$$E\{\hat{f}(x)\} - f(x) = b_a(x; p+1) + b_\lambda(x) + o(\delta^{p+1}) + o(\lambda n^{-1}\delta^{-q}),$$

$$var\{\hat{f}(x)\} = \frac{\sigma^2}{n} N(x)(G + \lambda D_q/n)^{-1} G (G + \lambda D_q/n)^{-1} N^t(x) + o\{(n\delta)^{-1}\},$$

(b) If $K_q \geq 1$ and $f(\cdot) \in W^q[a,b]$,

$$E\{\hat{f}(x)\} - f(x) = b_a(x; q) + b_\lambda(x) + o(\delta^q) + o\{(\lambda/n)^{1/2}\},$$

$$var\{\hat{f}(x)\} = \frac{\sigma^2}{n} N(x)(G + \lambda D_q/n)^{-1} G (G + \lambda D_q/n)^{-1} N^t(x) + o(n^{-1}(\lambda/n)^{-1/(2q)}).$$

The shrinkage bias b_λ is defined as $b_\lambda(x) = -\lambda n^{-1}\, N(x)(G + \lambda D_q/n)^{-1} D_q \beta$, where G and β are given in Section 2·3.

To better understand the shrinkage bias $b_\lambda(x)$, we show in the Appendix that $b_\lambda(x) = -\lambda\, N(x) H^{-1} \Delta_q^t W s_f^{(q)}(\tau)/n$ with $H = G + \lambda D_q/n$, $W = \text{diag}\left(\sum_{l=j}^{j+p-q} \int_{\kappa_l}^{\kappa_{l+1}} N_{j,q}(t)dt\right)$ and $s_f^{(q)}(\tau) = \{s_f^{(q)}(\tau_{-p+q}), \ldots, s_f^{(q)}(\tau_K)\}^t$ for some $\tau_j \in [\kappa_j, \kappa_{j+p+1-q}]$, $j = -p+q, \ldots, K$.

For equidistant knots and $p = q = 1$, this simplifies to

$$b_\lambda(x) = \frac{\lambda}{n} s_f^{(1)} \sum_{j=0}^{K} I_{[\kappa_j,\kappa_{j+1})}(x) \Big[(\kappa_{j+1} - x) \left\{ (H^{-1})_{j+1,1} + (H^{-1})_{j+1,K+2} \right\} + (x - \kappa_j) \left\{ (H^{-1})_{j+2,1} + (H^{-1})_{j+2,K+2} \right\} \Big],$$

where $s_f^{(1)}(x) = s_f^{(1)}$ is a constant for $s_f(\cdot) \in S(2;\kappa)$. Since $|(H^{-1})_{i,j}| = r^{|i-j|} O(\delta^{-1})$ for some $r \in (0,1)$, see Lemma 1, the $(H^{-1})_{j,1}$ decrease exponentially with growing j, while the $(H^{-1})_{j,K+2}$ increase with growing j. Thus, for j close to $[K/2]$, both $(H^{-1})_{j,K+2}$ and $(H^{-1})_{j,1}$ are small, implying that $b_\lambda(x)$ has much bigger values for x near the boundaries. Similar, but somewhat more complicated expressions can be obtained for more general settings. In contrast to the approximation bias, the shrinkage bias $b_\lambda(x)$ depends on the design density $\rho(x)$.

As already discussed in the previous section, the approximation and shrinkage bias play different roles in the two asymptotic scenarios. To show this, we plotted both bias terms together with the standard deviation of the penalized spline estimator for scenarios with $K_q < 1$ and $K_q \geq 1$ in Figure 1. The true function $f(x) = \cos(2\pi x)$ is evaluated at $n = 15000$ equally spaced points on $(0,1)$ and the errors are taken to be independent with distribution $N(0, 0.3^2)$. We used B-splines of degree three and a second order penalty, based on $K = 5$ equidistant knots for $K_q < 1$, and based on $K = 1000$ for $K_q \geq 1$. The penalty λ was determined by Generalized Cross-Validation (GCV) in both cases. For $K_q < 1$, one observes that the order of both bias components is the same. If $K_q \geq 1$, the approximation bias is extremely small, while the shrinkage bias is about 10 times larger than that for $K_q < 1$. In both cases, the shrinkage bias has bigger values near the boundaries. The variance of the estimator is bigger in case $K_q \geq 1$. In general, the variance of the penalized spline estimator is bigger near the boundaries, due to the structure of the matrix H^{-1}, see Lemma 1 in the Appendix.

5. PENALIZED SPLINES USING TRUNCATED POLYNOMIAL BASIS FUNCTIONS

Ruppert & Carroll (2000) used truncated polynomials as basis functions. For truncated polynomials of degree p based on K inner knots $a < \kappa_1 < \cdots < \kappa_K < b$, the penalized spline estimator is defined as the solution to the penalized least squares criterion

$$\sum_{i=1}^{n}\{Y_i - F(x_i)\alpha\}^2 + \lambda_p \sum_{j=1}^{K} \alpha_{j+p}^2,$$

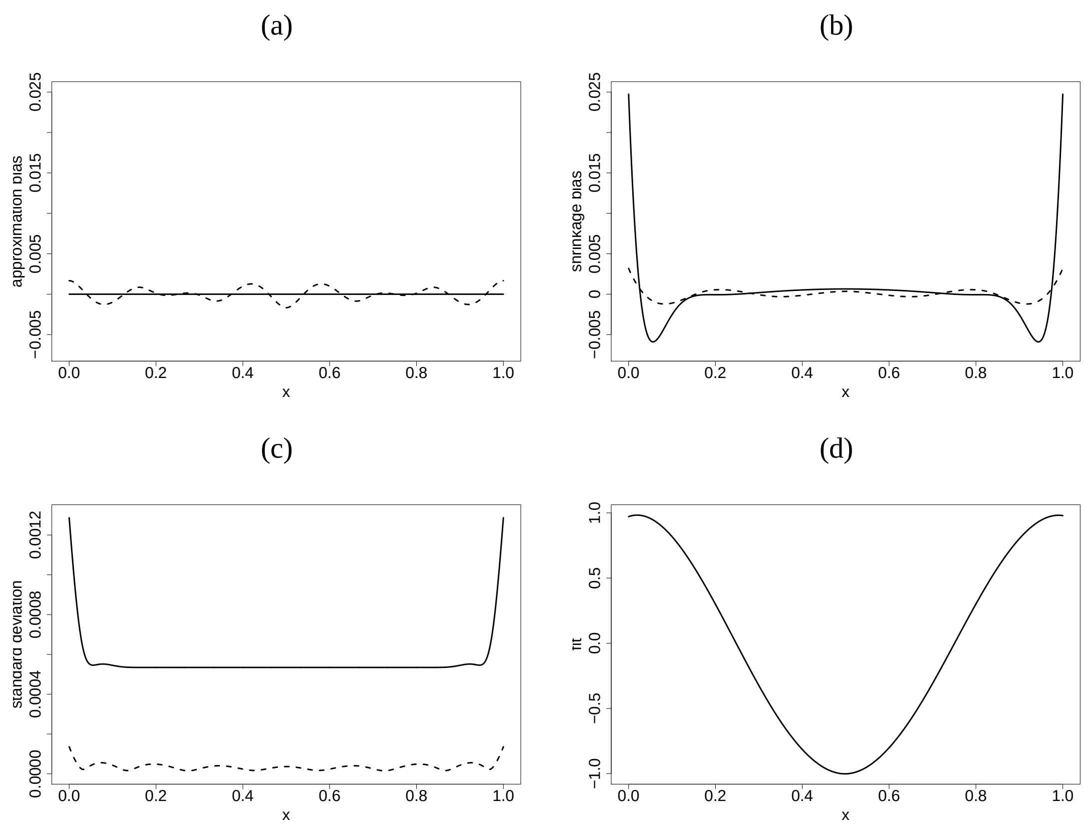

Fig. 1. Example of pointwise bias and variance of two penalized spline estimators with $K_q < 1$ (dashed line) and $K_q > 1$ (solid line). Panel (a) shows the approximation bias, (b) the shrinkage bias, (c) the standard deviation and (d) the true mean function $\cos(2\pi x)$.

with $F(x) = \{1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_K)_+^p\}$ and $\alpha = (\alpha_0, \ldots, \alpha_{K+p})$. The resulting estimator is a ridge regression estimator given by

$$\widehat{f}_p = F(F^{\mathrm{t}}F + \lambda_p \tilde{D}_p)^{-1}F^{\mathrm{t}}Y, \tag{15}$$

where $F = \{F(x_1)^t, \ldots, F(x_n)^t\}^t$ and $\tilde{D}_p$ is the diagonal matrix diag$(0_{p+1}, 1_K)$, indicating that only the spline coefficients are penalized.

The ridge penalty imposed on the spline coefficients can also be viewed as a penalty containing the integrated squared $(p+1)$th derivative of the spline function. Indeed,

$$\{F(x)\alpha\}^{(p)} = p!\,\alpha_p + p!\sum_{j=1}^{K}\alpha_{k+p}I_{[\kappa_j,\infty)}(x).$$

Since the derivative of an indicator function is a Dirac delta function (see e.g. Bracewell, 1999, p. 94), which integrates to one, it follows that

$$\int_a^b \left[\{F(x)\alpha\}^{(p+1)}\right]^2 dx = (p!)^2\sum_{j=1}^{K}\alpha_{j+p}^2.$$

In general, the results of Theorem 1 are not applicable to penalized splines with truncated polynomials since Lemma 3 does not hold for $q = p+1$. We use the equivalence of truncated polynomial and B-spline basis functions to arrive at asymptotic bias and variance expressions, see the appendix for more details. We obtain that for $K_q < 1$,

$$\begin{aligned} \mathrm{E}\{\hat{f}_p(x)\} - f(x) &= b_a(x; p+1) - \frac{\lambda_p\delta^{-p+1}}{(p!)^2 n}N(x)H^{-1}\nabla_{p+1}^t s_f^{(p+1)}(\kappa) + o(\delta^{p+1}) + o(\lambda n^{-1}\delta^{-p}) \\ &= O(\delta^{p+1}) + O(\lambda n^{-1}\delta^{-p}), \end{aligned} \tag{16}$$

$$\mathrm{var}\{\hat{f}_p(x)\} = \frac{\sigma^2}{n}N(x)H^{-1}GH^{-1}N^t(x) + o\{(n\delta)^{-1}\} = O\{(n\delta)^{-1}\}, \tag{17}$$

where $s_f^{(p+1)}(\kappa) = \delta^{-1}\{s_f^{(p)}(\kappa_1), s_f^{(p)}(\kappa_2) - s_f^{(p)}(\kappa_1), \ldots, s_f^{(p)}(\kappa_K) - s_f^{(p)}(\kappa_{K-1})\}^t$. It follows that taking $K \sim C_1 n^{1/(2p+3)}$ and $\lambda_p = O(n^\gamma)$ with $\gamma \le 2/(2p+3)$ leads to the optimal rate of convergence. For $K_q \geq 1$, we obtain that

$$\begin{aligned}
\mathrm{E}\{\hat{f}_p(x)\} - f(x) &= b_a(x; p+1) - \frac{\lambda_p \delta^{-p+1}}{(p!)^2 n} N(x) H^{-1} \nabla^t_{p+1} s_f^{(p+1)}(\kappa) + o(\delta^{p+1}) \\
&\quad + o\{(\lambda n^{-1})^{(p+1)/(2p+1)}\} = O(\delta^{p+1}) + O\{(\lambda n^{-1})^{(p+1)/(2p+1)}\}, \quad (18) \\
\mathrm{var}\{\hat{f}_p(x)\} &= \frac{\sigma^2}{n} N(x) H^{-1} G H^{-1} N^t(x) + o\{n^{-1}(\lambda n^{-1})^{(2p)/(2p+1)}\} \quad (19) \\
&= O\{n^{-1}(\lambda n^{-1})^{(2p)/(2p+1)}\},
\end{aligned}$$

Taking $\lambda \sim C_3 n^{2/(2p+3)}$ and $K = O(n^{\tilde{\nu}})$ with $\tilde{\nu} \geq 1/(2p+3)$ leads to the optimal rate of convergence, which is the same as in case $K_q < 1$, that is $n^{-(2p+2)/(2p+3)}$. Thus, if the truncated polynomials basis is used, there is no difference between two asymptotic scenarios and the optimal rate of convergence is reached in either case.

6. Discussion

The results in this paper and in particular Theorem 1 provide a theoretical justification that a smaller number of knots leads to a smaller averaged mean squared error. Moreover, we are able to characterize through K_q in (14) the relation between K, λ and n which determines the breakpoint between a "small" and "large" number of knots, or in other words, between the asymptotic scenario close to that of regression splines on the one hand and that of smoothing splines on the other hand. Results of this paper also show that using truncated polynomial basis functions leads to the optimal rate of convergence independent of the assumption made on the number of knots.

Penalized splines gained a lot of their popularity because of the link to mixed models where the spline coefficients are modeled as random effects, see Brumback et al. (1999), and earlier Speed (1991) for the case of smoothing splines. An interesting topic of further research would be a detailed study of the asymptotic properties of the estimators in this setting, building further on Kauermann et al. (2008) who verified the use of the Laplace approximation for a generalized mixed model with a growing number of spline basis functions for $K_q < 1$, but not for $K_q \geq 1$. Since mixed models are related to Bayesian models using a prior distribution on the spline coefficients, this could also bring additional insight in Bayesian spline estimation, see e.g. Carter & Kohn (1996); Speckman & Sun (2003).

The results of this paper are expected to hold for the more general class of likelihood based models, in particular for the generalized linear models as in Kauermann et al. (2008); a detailed study is interesting, though beyond the scope of the current paper. Other worthwhile routes of further investigation include models for spatial data, incorporating correlated errors and heteroscedasticity.

APPENDIX. TECHNICAL DETAILS

For use in the subsequent proofs, we define $G_{K,n} = (N^t N)/n$, $H_{K,n} = G_{K,n} + \lambda D_q/n$ and $H = G + \lambda D_q/n$ and state the following results:

(R1) Lemmas 6.3 and 6.4 in Zhou et al. (1998). $\|G_{K,n}^{-1}\|_\infty = \max_{1\leq i\leq K+p+1} \sum_{j=1}^{K+p+1} |\{G_{K,n}^{-1}\}_{i,j}| = O(\delta^{-1})$, $\max_{1\leq i,j\leq K+p+1} |\{G_{K,n}^{-1} - G^{-1}\}_{i,j}| = o(\delta^{-1})$, $\max_{1\leq i,j\leq K+p+1} |\{G_{K,n} - G\}_{i,j}| = o(\delta)$.

(R2) Under (A1)–(A3), $\max_{-p+q\leq j\leq K} \int_a^b N_{j,p+1}(u)\{f(u) - s_f(u)\}dQ_n(u) = o(\delta^{p+2})$, see Lemma 6.10 in Agarwal & Studden (1980) and thus $\mathrm{E}\{\hat{f}_{\mathrm{reg}}(x) - s_f(x)\} = N(x)G_{K,n}^{-1}\frac{1}{n}N(f - s_f) = o(\delta^{p+1})$, with $f = \{f(x_1), \ldots, f(x_n)\}^t$ and $s_f = \{s_f(x_1), \ldots, s_f(x_n)\}^t$. If $f \in W^q[a,b]$, then $\mathrm{E}\{\hat{f}_{\mathrm{reg}}(x) - s_f(x)\} = o(\delta^q)$.

(R3) $|\{G_{K,n}^{-1}\}_{ij}| \leq c\delta^{-1} r^{|i-j|}$ for some constants $c > 0$ and $r \in (0,1)$, see Lemma 6.3 in Zhou et al. (1998).

Before proving the two Theorems, we need the following three Lemmas.

LEMMA 1. *There exist some constants $r \in (0,1)$ and $c_0 > 0$ independent of K and n such that $|\{H_{K,n}^{-1}\}_{i,j}| \leq c_0\delta^{-1} r^{|i-j|}$ for $K_q < 1$ and $|\{H_{K,n}^{-1}\}_{i,j}| \leq c_0\delta^{-1}(1 + K_q^{2q})^{-1} r^{|i-j|}$ for $K_q \geq 1$.*

Proof. We apply Theorem 2.2 of Demko (1977) to $h_{\max}^{-1}H_{K,n}$, with $h_{\max}$ the maximum eigenvalue of $H_{K,n}$. First we verify the necessary conditions. The band diagonal matrix $H_{K,n}$ has $\{H_{K,n}^{-1}\}_{i,j} = 0$ for $|i-j| > p$, with $p \le q$. Since $H_{K,n}$ is a symmetric positive definite matrix, its spectral norm equals its maximum eigenvalue $h_{\max}$, so that $\|h_{\max}^{-1}H_{K,n}\|_2 = h_{\max}^{-1}\|H_{K,n}\|_2 = h_{\max}^{-1}(\max_{z:z^tz=1} z^tH_{K,n}^tH_{K,n}z)^{1/2} = 1$. Similarly, $\|h_{\max}H_{K,n}^{-1}\|_2 = h_{\max}/h_{\min}\|h_{min}H_{K,n}^{-1}\|_2 = h_{\max}/h_{\min}$. Thus, Theorem 2.2 of Demko (1977) applies and $h_{\max}|\{H_{K,n}^{-1}\}_{i,j}| \le c^*r^{|i-j|}$ for some $c^* > 0$ which depends only on p and $h_{\max}/h_{\min}$. It remains to find the lower bound for $h_{\max}$. The matrix $H_{K,n}$ is similar to $\tilde{H}_{K,n} = G_{K,n}(I_{K+p+1} + \lambda/nG_{K,n}^{-1/2}D_qG_{K,n}^{-1/2})$ and thus has the same eigenvalues. According to Corollary 2.4 of Lu & Pearce (2000) we can bound h_{max} from below with the product of the minimum eigenvalue of $G_{K,n}$ and the maximum eigenvalue of $(I_{K+p+1} + \lambda/nG_{K,n}^{-1/2}D_qG_{K,n}^{-1/2})$. The minimum eigenvalue of $G_{K,n}$ has the lower bound $\tilde{c}_o\delta$ for some $\tilde{c}_0$ independent of K and n, according to Lemma 6.2 of Zhou et al. (1998). The maximum eigenvalue of $(I_{K+p+1} + \lambda/nG_{K,n}^{-1/2}D_qG_{K,n}^{-1/2})$ is $(1 + K_q^{2q})$. With this we find $h_{\max} \ge \tilde{c}_0\delta$ for $K_q < 1$ and $h_{\max} \ge \tilde{c}_0\delta(1 + K_q^{2q})$ for $K_q \ge 1$. Setting $c_0 = c^*/\tilde{c}_0$ proves the lemma. □

From Lemma 1, it immediately follows that $\|H_{K,n}^{-1}\|_\infty = O(\delta^{-1})$ for $K_q < 1$ and $\|H_{K,n}^{-1}\|_\infty = O\{\delta^{-1}(1 + K_q^{2q})^{-1}\}$ for $K_q \ge 1$.

LEMMA 2. *The following statements hold:* $\max_{1\le i,j\le K+p+1} |\{H_{K,n}^{-1} - H^{-1}\}_{i,j}| = o(\delta^{-1})$ *for* $K_q < 1$ *and* $\max_{1\le i,j\le K+p+1} |\{H_{K,n}^{-1} - H^{-1}\}_{i,j}| = o\{\delta^{-1}(1 + K_q^{2q})^{-1}\}$ *for* $K_q \ge 1$.

Proof. First, we represent

$$\begin{aligned}(G + \lambda D_q/n)^{-1} &= (G - G_{K,n} + G_{K,n} + \lambda D_q/n)^{-1} \\ &= (G_{K,n} + \lambda D_q/n)^{-1} + (G_{K,n} + \lambda D_q/n)^{-1}(G_{K,n} - G) \\ &\quad \times\{I - (G_{K,n} + \lambda D_q/n)^{-1}(G_{K,n} - G)\}^{-1}(G_{K,n} + \lambda D_q/n)^{-1}.\end{aligned}$$

Applying Lemma 1 and (R1), one finds $\max_{1\le i,j\le K+p+1}|\{H_{K,n}^{-1}-H^{-1}\}_{i,j}| = \max_{1\le i,j\le K+p+1}|[H_{K,n}^{-1}(G_{k,n}-G)\{I_{K+p+1}-H_{K,n}^{-1}(G_{K,n}-G)\}^{-1}H_{K,n}^{-1}]_{i,j}|$, from which the result is immediate. □

A study of asymptotic properties of spline estimators via eigenvalues goes back to at least Utreras (1980), see also Utreras (1981, 1983). Speckman (1981, 1985) extended these results and a version of that we use below. Lemma 3 is adopted from Speckman (1985, eqn. 2.5d), see also Eubank (1999, p. 237).

LEMMA 3. *Under design condition (A2) and for the eigenvalues obtained in (10),*

$$s_1 = \cdots = s_q = 0, \quad s_j = n^{-1}(j-q)^{2q}\tilde{c}_1,\ j = q+1,\ldots,K+p+1,$$

where $\tilde{c}_1 = c_1\{1+o(1)\}$ *with* c_1 *is a constant that depends only on* q *and the design density and* $o(1)$ *converges to* 0 *as* $n\to\infty$ *uniformly for* $j_{1n}\le j\le j_{2n}$ *for any sequences* $j_{1n}\to\infty$ *and* $j_{2n}=o(n^{2/(2q+1)})$.

With a slightly different assumption on the design density, namely that the design density is regular in the sense that for $i=1,\ldots,n$, $\int_a^{x_i}\rho(x)dx = (2i-1)/(2n)$, Speckman (1985) obtained the exact expression of the constant as $c_1 = \pi^{2q}(\int_a^b \rho(x)^{1/(2q)}dx)^{-2q}$.

Proof of Theorem 1. Let us begin with case (a), that is $K_q<1$. First, we rewrite

$$\sum_{j=1}^{K+p+1}\frac{1}{(1+\lambda s_j)^2} = q + \sum_{j=q+1}^{K+p+1}\frac{1}{\{1+\lambda n^{-1}\tilde{c}_1(j-q)^{2q}\}^2} \tag{A1}$$

$$= \left(\frac{\tilde{c}_1\lambda}{n}\right)^{-1/(2q)}\int_0^{K_q}\frac{du}{(1+u^{2q})^2} + q - 1 + r_q, \tag{A2}$$

with K_q defined in (14) and $r_q = O(1)$ as the remainder term of the Euler-Maclaurin formula. Now using a series expansion around zero of $(1+x)^{-2} = \sum_{j=0}^{\infty}(-1)^j(j+1)x^j$ for $0<x<1$ we easily find

$$\int_0^{K_q}\frac{du}{(1+u^{2q})^2} = K_q\sum_{j=0}^{\infty}(-1)^j(j+1)\frac{K_q^{2qj}}{2qj+1} = K_q c_2,$$

where $c_2 = {}_2F_1(2,1/(2q);1+1/(2q),-K_q^{2q})$ denotes the hypergeometric series, see Abramowitz & Stegun (1972, Ch. 15), converging for any $K_q<1$ and $q>0$. With this, we obtain that the average variance in case (a) equals

$$\frac{\sigma^2}{n}\sum_{j=1}^{K+p+1}\frac{1}{(1+\lambda s_j)^2}=\frac{\sigma^2}{n}\{c_2(K+p+1-q)+q-1+r_q\}=O\left(\frac{K}{n}\right).$$

Consider now the second term in (13). Bearing in mind that $K_q^{2q}=\lambda n^{-1}\tilde{c}_1(K+p+1-q)^{2q}<1$ and that the function $x(1+x)^{-2}\leq x$ for $0<x<1$, we can bound the average squared shrinkage bias with

$$\frac{\lambda}{n}\sum_{j=1}^{K+p+1}s_jb_j^2\frac{\lambda s_j}{(1+\lambda s_j)^2}\leq\frac{\lambda}{n}K_q^{2q}\sum_{j=1}^{K+p+1}s_jb_j^2=\frac{\lambda^2}{n^2}(K+p+1-q)^{2q}\beta_f^tD_q\beta_f,$$

with $\beta_f=(N^tN)^{-1}N^tf$. Further, adding and subtracting s_f from f in β_f we find

$$\begin{aligned}\beta_f^tD_q\beta_f&=\beta^tD_q\beta+2(f-s_f)^tN(N^tN)^{-1}D_q(N^tN)^{-1}N^ts_f\\&\quad+(f-s_f)^tN(N^tN)^{-1}D_q(N^tN)^{-1}N^t(f-s_f)\\&=\beta^tD_q\beta+o(\delta^{p+1})+o(\delta^{2p+2}),\end{aligned}$$

where (R2) was applied to obtain the orders of two last terms. Since the penalty $\beta^tD_q\beta$ was assumed to be finite, see below (2), the average shrinkage bias in (13) has the order $O(\lambda^2n^{-2}K^{2q})$. Finally, the average squared approximation bias in (13), has the asymptotic order $O(K^{-2(p+1)})$ for a function $f\in C^{p+1}[a,b]$, as follows from (8). We now choose orders of K and λ, so that they ensure the best possible rate of convergence. As shown in Stone (1982), a $p+1$ times continuously differentiable function has the optimal rate of convergence $n^{-(2p+2)/(2p+3)}$. It is straightforward to see that choosing $K\sim C_1n^{1/(2p+3)}$, with C_1 a constant, implies the average variance and the average squared approximation bias to have the same order $O(n^{-(2p+2)/(2p+3)})$. The shrinkage bias is controlled by the smoothing parameter λ. Choosing $\lambda=O(n^{(p+2-q)/(2p+3)})$ balances both bias components, while λ values of a smaller asymptotic order make the shrinkage bias negligible.

Let us now consider case (b) with $K_q\geq1$ and first find the order of the average variance. Since the expansion $(1+x)^{-2}$ diverges for $x=1$, we first exclude this value from the sum in (A1) as follows

$$\sum_{j=1}^{K+p+1}\frac{1}{(1+\lambda s_j)^2}=\sum_{j=1}^{j^*-1}\frac{1}{(1+\lambda s_j)^2}+\frac{1}{4}+\sum_{j=j^*+1}^{K+p+1}\frac{1}{(1+\lambda s_j)^2},$$

where j^* is such that $\lambda n^{-1}\tilde{c}_1(j^*-q)^{2q}=1$. The integral representation of the average variance is

$$\frac{\sigma^2}{n}\sum_{j=1}^{K+p+1}\frac{1}{(1+\lambda s_j)^2}=\frac{\sigma^2}{n}\left(\frac{\tilde{c}_1\lambda}{n}\right)^{-1/(2q)}\int_0^{1-(\lambda n^{-1}\tilde{c}_1)^{1/(2q)}}\frac{du}{(1+u^{2q})^2} \tag{A3}$$

$$+\frac{\sigma^2}{n}\left(\frac{\tilde{c}_1\lambda}{n}\right)^{-1/(2q)}\int_{1+(\lambda n^{-1}\tilde{c}_1)^{1/(2q)}}^{K_q}\frac{du}{(1+u^{2q})^2}+\frac{\sigma^2}{n}\tilde{r}_q, \tag{A4}$$

with $\tilde{r}_q=O(1)$ as a constant, including $1/4$ and two remainder terms of the Euler-Maclaurin formula. For $K_q=1$ only the first integral and a constant are present. If there is no such j^* that $\lambda n^{-1}\tilde{c}_1(j^*-q)^{2q}=1$, then we obtain one integral with the upper bound less than one and another integral with the lower bound larger than one directly, with $\tilde{r}_q$ updated correspondingly. Since the upper limit of the integral is less than one, we use the series expansion of $(1+x)^{-2}$ as in case (a) and obtain for the integral in (A3),

$$\frac{\sigma^2}{n}\left(\frac{\tilde{c}_1\lambda}{n}\right)^{-1/(2q)}\left\{1-\left(\frac{\tilde{c}_1\lambda}{n}\right)^{1/(2q)}\right\}\tilde{c}_2=O\left(n^{1/(2q)-1}\lambda^{-1/(2q)}\right),$$

with $\tilde{c}_2={}_2F_1(2,1/(2q);1+1/(2q),-\{1-(\lambda n^{-1}\tilde{c}_1)^{1/(2q)}\}^{2q})$ as a converging hypergeometric series. Changing the integration variable to its reciprocal, one gets for the integral in (A4),

$$\frac{\sigma^2}{n}\left(\frac{\tilde{c}_1\lambda}{n}\right)^{-1/(2q)}\left[K_q^{1-4q}c_3-\tilde{c}_3\{1-(\lambda n^{-1}\tilde{c}_1)^{1/(2q)}\}^{4q-1}\right](4q-1)^{-1}=O\left(n^{1/(2q)-1}\lambda^{-1/(2q)}\right),$$

where $c_3={}_2F_1(2,(4q-1)(2q)^{-1};(6q-1)(2q)^{-1},-K_q^{-2q})$ and $\tilde{c}_3={}_2F_1(2,(4q-1)(2q)^{-1};(6q-1)(2q)^{-1},-\{1+(\lambda n^{-1}\tilde{c}_1)^{1/(2q)}\}^{-2q})$ both are hypergeometric series converging for any $K_q>1$ and $q>0$. Thus, for case (b) with $K_q\geq 1$ the average variance has the asymptotic order $O(n^{1/(2q)-1}\lambda^{-1/(2q)})$. Since $x(1+x)^{-2}\leq 1/4$ for any $x\geq 1$, the average squared shrinkage bias for $K_q\geq 1$ is bounded by

$$\frac{\lambda}{n}\sum_{j=q+1}^{K+p+1}b_j^2s_j\frac{\lambda s_j}{(1+\lambda s_j)^2}\leq\frac{\lambda}{4n}\sum_{j=q+1}^{K+p+1}b_j^2s_j=\frac{\lambda}{4n}\beta_f D_q\beta_f=\frac{\lambda}{4n}\{\beta D_q\beta+o(\delta^q)\}.$$

With this, the average squared shrinkage bias is of order $O(\lambda/n)$ for $K_q\geq 1$. It is straightforward to see that $\lambda=O(n^{1/(2q+1)})$ balances the average squared shrinkage bias and the average variance. Finally the average squared approximation bias will not dominate the average mean squared error if the number of knots satisfies $K\sim C_2n^\nu$, with $\nu\geq 1/(2q+1)$ and C_2 as a constant. This implies that the average approximation bias is of the same order as the average squared shrinkage bias if $K_q = 1$ and is negligible with the order $O(n^{-\nu'})$, with $\nu' > 2q/(2q+1)$ for $K_q > 1$. Thus, $\mathrm{AMSE}(\widehat{f}) = O(n^{-2q/(1+2q)})$. □

Proof of Theorem 2. Let us first consider the bias. We represent

$$\hat{f}(x) = \hat{f}_{\rm reg}(x) - \frac{\lambda}{n} N(x) H_{K,n}^{-1} D_q G_{K,n} \frac{1}{n} N^t Y$$

with $\hat{f}_{\rm reg}(x)$ defined in (5) and find

$$\begin{aligned} \mathrm{E}\{\hat{f}(x)\} - f(x) &= \{s_f(x) - f(x)\} + E\{\hat{f}_{\rm reg}(x) - s_f(x)\} \\ &\quad + \frac{\lambda}{n} N(x) H_{K,n}^{-1} D_q G_{K,n}^{-1} N^t \frac{1}{n}(f - s_f + s_f). \end{aligned}$$

According to Barrow & Smith (1978), it holds that $s_f(x) - f(x) = b_a(x; p+1) + o(\delta^{p+1})$ for $K_q < 1$ and $b_a(x;q) + o(\delta^q)$ for $K_q \geq 1$, due to different smoothness assumptions made on $f(\cdot)$. The order of the second component is given by (R2). Let us consider $\lambda N(x) H_{K,n}^{-1} D_q \beta / n$ with $\beta = G_{K,n}^{-1} N^t s_f / n = (N^t N)^{-1} N^t s_f$. Using the definition of penalty D_q and noting that $s_f^{(q)}(x) = (N(x)\beta)^{(q)} = N_q(x)\Delta_q \beta$ with $N_q(x) = \{N_{-p+q,p+1-q}(x), \ldots, N_{K,p+1-q}(x)\}$, we can apply the mean value theorem and rewrite

$$-\frac{\lambda}{n} N(x) H_{K,n}^{-1} D_q \beta = -\frac{\lambda}{n} N(x) H_{K,n}^{-1} \Delta_q^t \int_a^b N_q^t(x) s_f^{(q)}(x) dx = -\frac{\lambda}{n} N(x) H_{K,n}^{-1} \Delta_q^t W s_f^{(q)}(\tau),$$

where $W = \mathrm{diag}\left(\sum_{l=j}^{j+p-q} \int_{\kappa_l}^{\kappa_{l+1}} N_{j,q}(x) dx\right)$ and $\tau = (\tau_{-p+q}, \ldots, \tau_K)^t$ with some $\tau_j \in [\kappa_j, \kappa_{j+p+1-q}]$, $j = -p+q, \ldots, K$. Further, we represent

$$\begin{aligned} & -\frac{\lambda}{n} N(x) H^{-1} \Delta_q^t W s_f^{(q)}(\tau) - \frac{\lambda}{n} N(x)(H_{K,n}^{-1} - H^{-1}) \Delta_q^t W s_f^{(q)}(\tau) \\ &= -\frac{\lambda}{n} N(x)(G + \lambda D_q/n)^{-1} D_q \beta - \frac{\lambda}{n} N(x)(H_{K,n}^{-1} - H^{-1}) \Delta_q^t W s_f^{(q)}(\tau) \\ &= b_\lambda - \frac{\lambda}{n} N(x)(H_{K,n}^{-1} - H^{-1}) \Delta_q^t W s_f^{(q)}(\tau). \end{aligned}$$

It remains to show that $\lambda N(x)(H_{K,n}^{-1} - H^{-1})\Delta_q^t W s_f^{(q)}(\tau)/n$ and $\lambda H_{K,n}^{-1} D_q G_{K,n}^{-1} N^t (f - s_f)/n$ are of negligible asymptotic order for both $K_q < 1$ and $K_q \geq 1$. Since $N_{j,q}(\cdot) \leq 1$, one finds $\|W\|_\infty = O(\delta)$. Moreover, by definition $\|\Delta_q\|_\infty = O(\delta^{-q})$, see also Lemma 6.1 in Cardot (2000). Thus, with Lemmas 1, 2 and $\|s_f^{(q)}(\tau)\|_\infty = O(1)$ it is straightforward to see that for $K_q < 1$, $\lambda N(x)(H_{K,n}^{-1} - H^{-1})\Delta_q^t W s_f^{(q)}(\tau)/n = o(\lambda n^{-1} \delta^{-q})$ and for $K_q \geq 1$,

$\lambda N(x)(H_{K,n}^{-1} - H^{-1})\Delta_q^t W s_f^{(q)}(\tau)/n = o\{\lambda n^{-1}\delta^{-q}(1+K_q^{2q})^{-1}\} = o\{(\lambda/n)^{1/2}K_q^q(1+K_q^{2q})^{-1}\} = o\{(\lambda/n)^{1/2}\}$, since $K_q^q(1+K_q^{2q})^{-1} \leq 1/2$ for $K_q \geq 1$. From (R2) follows that $G_{K,n}^{-1}N^t(f-s_f)/n$ is a vector with elements of order $o(\delta^{p+1})$ for $f \in C^{p+1}[a,b]$ and $o(\delta^q)$ for $f \in W^q[a,b]$. Using the same arguments as above we obtain $\lambda N(x)H_{K,n}^{-1}D_q G_{K,n}^{-1}N^t(f-s_f)/n = o(\lambda n^{-1}\delta^{p+1-q})$ for $K_q < 1$ and $\lambda N(x)H_{K,n}^{-1}D_q G_{K,n}^{-1}N^t(f-s_f)/n = o\{(\lambda/n)^{1/2}\}$ for $K_q \geq 1$. Thus, if $K_q < 1$,

$$\mathrm{E}\{\hat{f}(x)\} - f(x) = b_a(x;p+1) + b_\lambda(x) + o(\delta^{p+1}) + o(\lambda n^{-1}\delta^{-q}) = O(\delta^{p+1}) + O(\lambda n^{-1}\delta^{-q})$$

and if $K_q \geq 1$,

$$\mathrm{E}\{\hat{f}(x)\} - f(x) = b_a(x;q) + b_\lambda(x) + o(\delta^q) + o\{(\lambda/n)^{1/2}\} = O(\delta^q) + O\{(\lambda/n)^{1/2}\}.$$

The differentiability assumption of f is not crucial here and is made only for consistency with Theorem 1. Finally, let us consider the variance $\mathrm{var}\{\hat{f}(x)\} = \sigma^2 N(x)H_{K,n}^{-1}G_{K,n}H_{K,n}^{-1}N^t(x)/n$. Adding and subtracting in the same fashion as above H^{-1} and G, one finds for $K_q < 1$,

$$\mathrm{var}\{\hat{f}(x)\} = \frac{\sigma^2}{n}N(x)(G+\lambda D_q/n)^{-1}G(G+\lambda D_q/n)^{-1}N^t(x) + o(\{n\delta\}^{-1}) = O(\{n\delta\}^{-1})$$

and for $K_q \geq 1$,

$$\begin{aligned}\mathrm{var}\{\hat{f}(x)\} &= \frac{\sigma^2}{n}N(x)H^{-1}GH^{-1}N^t(x) + o(\{n^{-1}(\lambda/n)^{-1/(2q)}K_q(1+K_q^{2q})^{-2}\}) \\ &= \frac{\sigma^2}{n}N(x)(G+\lambda D_q/n)^{-1}G(G+\lambda D_q/n)^{-1}N^t(x) + o(\{n^{-1}(\lambda/n)^{-1/(2q)}\}) \\ &= O(\{n^{-1}(\lambda/n)^{-1/(2q)}\}). \qquad \square\end{aligned}$$

Proof of (16)–(19). From the alternative definition of B-splines as scaled $(p+1)$th order divided differences of truncated polynomials, see de Boor (2001, Ch. IX),

$$N_{j,p+1}(x) = (-1)^{(p+1)}(\kappa_{j+p+1} - \kappa_j)[\kappa_j, \dots, \kappa_{j+p+1}](x-\cdot)_+^p, \;\; j = -p, \dots, K, \tag{A5}$$

where $[\kappa_j, \dots, \kappa_{j+p+1}](x-\cdot)_+^p$ denotes the $(p+1)$th order divided difference of $(x-\cdot)_+^p$ as a function of knots κ_j for fixed x. In case of equidistant knots, (A5) simplifies to $N_{j,p+1}(x) = (-1)^{(p+1)}\delta^{-p}\nabla_{p+1}(x-\cdot)_+^p/p!$. B-spline and truncated polynomial basis functions span the same set of spline functions (de Boor, 2001, Ch. IX), thus there exists a square and invertible transition matrix L, such that $N = FL$.

The equivalence of the *penalized* spline estimators $\widehat{f}$ and $\widehat{f}_p$ is not automatic, but will follow when there is equality of the penalties. We work out the case of fitting with B-splines and obtaining the same penalized estimator as $\widehat{f}_p$ in (15) with $\tilde{D}_p$ as penalty matrix. Using the equality $N = FL$ for the penalized estimator $\widehat{f}_p$ implies that we can write it as $\widehat{f}_p = N(N^{\mathrm{t}}N + \lambda_p L^t \tilde{D}_p L)^{-1} N^{\mathrm{t}} Y$. Thus, fitting with B-splines yields an equivalent estimator to $\widehat{f}_p$ if we use the penalty term $\lambda_p L^t \tilde{D}_p L$ instead of λD_q. This penalty matrix can be obtained as follows. By writing $(N(x)\beta)^{(p)} = \sum_{j=0}^{K} N_{j,1}(x)\beta_j^{(p)} = \sum_{j=1}^{K} I_{[\kappa_j,\infty)}(x)(\beta_j^{(p)} - \beta_{j-1}^{(p)})$ we find that

$$\int_a^b \left[\{N(x)\beta\}^{(p+1)}\right]^2 dx = \sum_{j=1}^{K} (\beta_j^{(p)} - \beta_{j-1}^{(p)})^2.$$

Thus, L can be found from the equation $(p!)^2 \beta^t L^t \tilde{D}_p L \beta = \sum_{j=1}^{K} (\beta_j^{(p)} - \beta_{j-1}^{(p)})^2$. For equidistant knots $\beta_j^{(p+1)} = (\beta_j^{(p)} - \beta_{j-1}^{(p)})/\delta$, according to (3), and one obtains that

$$(p!)^2 \beta^t L^t \tilde{D}_p L \beta = \sum_{i=1}^{K} (\delta \beta_j^{(p+1)})^2 = \delta^{-2p} \beta^t \nabla_{p+1}^t \nabla_{p+1} \beta.$$

Thus, for equivalence of the estimators the penalty matrix using B-splines with equidistant knots should be $L^t \tilde{D}_p L = \delta^{-2p} \nabla_{p+1}^t \nabla_{p+1}/(p!)^2$. We can find the optimal asymptotic orders for K and λ as well as the pointwise bias and variance, following the arguments in the proof of Theorem 2, though by replacing λD_q by $\lambda_p \delta^{-2p} \nabla_{p+1}^t \nabla_{p+1}/(p!)^2$. For $K_q > 1$, then due to the penalty matrix $\|H_{K,n}^{-1}\|_\infty = O\{\delta^{-1}(1 + \lambda n^{-1} \delta^{-2p-1})^{-1}\}$. Proceeding in the same manner as in the proof of Theorem 2, we obtain (18) and (19). □

ACKNOWLEDGEMENTS

The authors wish to thank Maarten Jansen for helpful hints concerning some of the calculations. They are also grateful to all reviewers of this paper for their constructive remarks.

REFERENCES

ABRAMOWITZ, M. & STEGUN, I. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover.

AGARWAL, G. & STUDDEN, W. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.* **8**, 1307–1325.

BARROW, D. L. & SMITH, P. W. (1978). Asymptotic properties of best $L_2[0, 1]$ approximation by splines with variable knots. *Quart. Appl. Math.* **36**, 293–304.

BESSE, P., CARDOT, H. & FERRATY, F. (1997). Simultaneous nonparametric regression of unbalanced longitudial data. *Comp. Statist. Data Anal.* **24**, 255–270.

BRACEWELL, R. (1999). *The Fourier Transform and Its Applications*. New York: McGraw-Hill.

BRUMBACK, B. A., RUPPERT, D. & WAND, M. P. (1999). Comment on Shively, Kohn and Wood. *J. Amer. Statist. Assoc.* **94**, 794–797.

CARDOT, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *J. Nonpar. Statist.* **12**, 503–538.

CARTER, C. K. & KOHN, R. (1996). Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika* **83**, 589–601.

COX, D. D. (1983). Asymptotics for M-type smoothing splines. *Ann. Statist.* **11**, 530–551.

CRAVEN, P. & 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**, 377–403.

DE BOOR, C. (2001). *A Practical Guide to Splines*. New York: Springer. Revised Edition.

DEMKO, S. (1977). Inverses of band matrices and local convergence of spline projections. *SIAM. J. Numer. Anal.* **14**, 616–619.

DEMMLER, A. & REINSCH, C. (1975). Oscillation matrices with spline smoothing. *Numer. Math.* **24**, 375–382.

EILERS, P. H. C. & MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Stat. Science* **11**, 89–121. With comments and a rejoinder by the authors.

EUBANK, R. L. (1999). *Nonparametric regression and spline smoothing*, vol. 157 of *Statistics: Textbooks and Monographs*. New York: Marcel Dekker Inc., 2nd ed.

GREEN, P. J. & SILVERMAN, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*, vol. 58 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

HALL, P. & OPSOMER, J. (2005). Theory for penalized spline regression. *Biometrika* **92**, 105–118.

Huang, J. Z. (2003a). Asymptotics for polynomial spline regression under weak conditions. *Statist. Probab. Lett.* **65**, 207–216.

Huang, J. Z. (2003b). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31**, 1600–1635.

Kauermann, G., Krivobokova, T. & Fahrmeir, L. (2008). Some asymptotic results on generalized penalized spline smoothing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* To appear.

Kelly, C. & Rice, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics* **46**, 1071–1085.

Li, Y. & Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika* **95**, 415–436.

Lu, L.-Z. & Pearce, C. (2000). Some new bounds for singular values and eigenvalues of matrix products. *Ann. Oper. Res.* **98**, 141–148.

Nychka, D. (1995). Splines as local smoothers. *Ann. Statist.* **23**, 1175–1197.

Oehlert, G. W. (1992). Relaxed boundary smoothing splines. *Ann. Statist.* **20**, 146–160.

Ormerod, J. & Wand, M. (2008). On semiparametric regression with O'Sullivan penalised splines. *ANZJS* To appear.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Stat. Science* **1**, 505–527. With discussion.

Rice, J. & Rosenblatt, M. (1981). Integrated mean squared error of a smoothing spline. *J. Approx. Theory* **33**, 353–369.

Rice, J. & Rosenblatt, M. (1983). Smoothing splines: regression, derivatives and deconvolution. *Ann. Statist.* **11**, 141–156.

Ruppert, D. & Carroll, R. (2000). Spatially-adaptive penalties for spline fitting. *Aust. N. Z. J. Statist.* **42**, 205–224.

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

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. New York: Wiley.

Schwetlick, H. & Kunert, V. (1993). Spline smoothing under constraints on derivatives. *BIT* **33**, 512–528.

Speckman, P. (1981). The asymptotic integrated mean square error for smoothing noisy data by splines. Technical report, University of Oregon.

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

SPECKMAN, P. L. & SUN, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika* **90**, 289–302.

SPEED, T. (1991). Comment on “that BLUP is a good thing: The estimation of random effects,” by G. K. Robinson. *Statist. Science* **6**, 42–44.

STONE, C. J. (1982). Optimal rate of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–1053.

UTRERAS, F. (1980). Sur le choix du paramètre d’ajustement dans le lissage par fonctions spline. *Numer. Math.* **34**, 15–28.

UTRERAS, F. (1981). Optimal smoothing of noisy data using spline functions. *SIAM J. Sci. Stat. Comput.* **2**, 349–362.

UTRERAS, F. (1983). Natural spline functions, their associated eigenvalue problem. *Numer. Math.* **42**, 107–117.

UTRERAS, F. (1985). Smoothing noisy data under monotonicity constraints existence, characterization and convergence rates. *Numer. Math.* **47**, 611–625.

UTRERAS, F. (1988). Boundary effects on convergence rates for Tikhonov regularization. *J. Approx. Theory* **54**, 235–249.

WAHBA, G. (1975). Smoothing noisy data with spline functions. *Numer. Math.* **24**, 383–393.

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

ZHOU, S., SHEN, X. & WOLFE, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.* **26**, 1760–1782.