

Asymptotic properties of penalized spline estimators

GERDA CLAESKENS

Katholieke Universiteit Leuven, ORSTAT, Naamsestraat 69, B-3000 Leuven, Belgium

gerda.claeskens@econ.kuleuven.be

TATYANA KRIVOBOKOVA

Katholieke Universiteit Leuven, ORSTAT, Naamsestraat 69, B-3000 Leuven, Belgium

tatyana.krivobokova@econ.kuleuven.be

JEAN D. OPSOMER

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

jopsomer@stat.colostate.edu.

May 28, 2008

Summary

In this paper 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 an assumption 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.

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 in a unified way 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. It is known that both ‘extremes’ enjoy different asymptotic properties. Important theoretical results on (unpenalized) regression

splines are obtained by Zhou et al. (1998). We employ these results in an elegant way by constructing an additive expansion of the penalized spline estimator about the regression spline estimator, for the situation of a relatively small number of knots. Speckman (1985) obtained the optimal rates of convergence for smoothing spline estimators. Based on this result and following Utreras (1980, 1981) who showed the convergence of eigenvalues associated to smoothing splines to the eigenvalues of a differential operator, we obtain the asymptotic properties for the penalized splines for the case of a relatively large number of knots, similar to those of smoothing spline estimators.

One of the main results of this paper 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 K , the sample size n and the penalty parameter λ (see further). Depending on the value of this function, the asymptotic results are related to those of regression splines or to those of smoothing splines.

Another interesting finding is that it is better to use a smaller number of knots (close to regression splines case), since that results in a smaller mean squared error if the function is smoother, that is, has more continuous derivatives than the order of the penalty.

The combination of regression splines (with K less than n) and a penalty has been studied by several authors. O’Sullivan (1986) used penalized fitting with cubic B-splines in the context of inverse problems. For cubic B-splines he used a set of knots that is different from the data points and a penalty that is equal to the integrated squared second derivative of the spline function. Kelly and Rice (1990) and Besse et al. (1997) used similar B-spline approximations to the smoothing splines, which they called hybrid splines. Schwetlick and Kunert (1993) decoupled the order of the B-spline and the derivative used in the penalty function. This same idea has further been promoted by Eilers and Marx (1996) who propose to use 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 on this topic are Wahba (1990), Green and Silverman (1994) and Eubank (1999). For smoothing splines, the penalty is the integrated squared q th derivative of the

function, leading to a smoothing spline of degree $2q - 1$ (with $q = 2$ a common choice in practice). Rice and Rosenblatt (1981, 1983) study the estimator's integrated mean squared error and effects of boundary bias, see also Oehlert (1992) and Utreras (1988). In a series of papers, of which Wahba (1975) and Craven and Wahba (1978) are early references, the averaged mean squared error is used, also in connection to the choice of the smoothing parameter. Cox (1983) studies rates of convergence for robust smoothing splines, and local properties of smoothing splines are studied by Nychka (1995).

For regression splines, the integrated mean squared error was studied by Agarwal and Studden (1980), and Huang (2003a,b) obtained local asymptotic results by considering the least squares estimator as an orthogonal projection.

Theoretical properties of penalized spline estimators are less explored. Some first results can be found in Hall and Opsomer (2005) who used a white noise representation of the model to obtain mean squared error and consistency of the estimator. Kauermann et al. (2007) work with generalized linear models. Li and 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 (close to 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. In the same time, we obtain the breakpoint between the two cases as a function of the number of knots, sample size and penalty parameter. We include in our study both the approximation and shrinkage bias. We study both B-splines and truncated polynomial spline basis functions and relate the corresponding results.

The paper is organized as follows. Section 2 defines the penalized spline estimators, gives relations between the use of different types of spline basis functions and states some results on regression splines that will be used later. Section 3 contains the results on the average mean squared error of the estimator and obtains the breakpoint that separates both asymptotic cases. Pointwise bias and variance are obtained in Section 4. We conclude in Section 5. The proofs are given in the Appendix.

2 Penalized splines

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

$$Y_i = f(x_i) + \varepsilon_i, \quad (1)$$

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 $\sigma^2 > 0$.

2.1 Penalized splines with B-spline basis functions

The idea of penalized spline smoothing traces back to O'Sullivan (1986) (see also Schwetlick and Kunert, 1993), but it was Eilers and Marx (1996) who introduced the combination of B-splines and difference penalties which they called P-splines. 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 $a = \kappa_0 < \kappa_1 < \dots < \kappa_K < \kappa_{K+1} = b$. In addition, define p knots $\kappa_{-p} = \kappa_{-p+1} = \dots = \kappa_{-1} = \kappa_0$ and another set of p knots $\kappa_{K+1} = \kappa_{K+2} = \dots = \kappa_{K+p+1}$. The B-spline basis functions are defined as

$$\begin{aligned} N_{j,1}(x) &= \begin{cases} 1, & \kappa_j \leq x < \kappa_{j+1} \\ 0, & \text{otherwise} \end{cases}, \\ N_{j,p+1}(x) &= \frac{x - \kappa_j}{\kappa_{j+p} - \kappa_j} N_{j,p}(x) + \frac{\kappa_{j+p+1} - x}{\kappa_{j+p+1} - \kappa_{j+1}} N_{j+1,p}(x), \end{aligned}$$

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

We define the P-spline estimator as the minimizer of

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

where the sum of squared differences is penalized with the integrated squared q th 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 (see also Section 2.2), it is a natural condition to have $q \leq p$. However, we give a separate treatment to the case of truncated polynomial basis functions where $q = p + 1$, see Section 2.2 and the end of Section 4. The penalty constant λ plays the role of a smoothing parameter. Note that letting $\lambda \rightarrow 0$ implies an unpenalized estimate, while $\lambda \rightarrow \infty$ forces convergence of the q th derivative of the spline function to zero, with the consequence that the limiting estimator is a $(q - 1)$ th degree polynomial. The derivative formula for B-spline functions, as given in de Boor (2001, ch. X), states that

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

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

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

Thus we can rewrite the penalty term in (2) as $\lambda \boldsymbol{\beta}^t \boldsymbol{\Delta}_q^t \mathbf{R} \boldsymbol{\Delta}_q \boldsymbol{\beta}$, where matrix \mathbf{R} has elements $R_{ij} = \int_a^b N_{j,p+1-q}(x) N_{i,p+1-q}(x) dx$, for $i, j = -p + q, \dots, K$ and $\boldsymbol{\Delta}_q$ denotes the matrix corresponding to the weighted difference operator defined in (3), i.e. $\boldsymbol{\beta}^{(q)} = \boldsymbol{\Delta}_q \boldsymbol{\beta}$. Note that for the special case of equidistant knots, i.e. $\kappa_j - \kappa_{j-1} = \delta$ for any $j = -p + 1, \dots, K$, there is an explicit expression of the matrix $\boldsymbol{\Delta}_q$ in terms of the q th backward difference operator ∇_q . This latter matrix is defined recursively via $\nabla_q = \nabla_1(\nabla_{q-1})$, $\nabla_1 \beta_j = \beta_{j-1} - \beta_j$. For equidistant knots it holds $\boldsymbol{\Delta}_q = \delta^{-q} \nabla_q$. For the special case $p = q$, matrix \mathbf{R} reduces to a diagonal matrix with diagonal elements $(\kappa_j - \kappa_{j-1})$, which are further simplified to δ for equidistant knots. Another special case of $p = q + 1$ results in a tridiagonal matrix \mathbf{R} , with the integrals of the squared linear (order equal to 2) B-splines on the main diagonal.

Further, define the spline basis vector of dimension $1 \times (K + p + 1)$ as $\mathbf{N}(x) = \{N_{-p,p+1}(x), \dots, N_{K,p+1}(x)\}$, the $n \times (K + p + 1)$ spline design matrix $\mathbf{N} = \{\mathbf{N}(x_1)^t, \dots, \mathbf{N}(x_n)^t\}^t$, and let $\mathbf{D}_q = \boldsymbol{\Delta}_q^t \mathbf{R} \boldsymbol{\Delta}_q$. With this notation the P-spline estimator takes the

form of a ridge regression estimator

$$\widehat{\mathbf{f}} = \mathbf{N}(\mathbf{N}^t \mathbf{N} + \lambda \mathbf{D}_q)^{-1} \mathbf{N}^t \mathbf{Y}, \quad (4)$$

where $\widehat{\mathbf{f}} = \{\widehat{f}(x_1), \dots, \widehat{f}(x_n)\}^t$ and $\mathbf{Y} = (Y_1, \dots, Y_n)^t$.

Originally, Eilers and Marx (1996) simplified (4) by suggesting to use equidistant knots and a combination of cubic splines ($p = 3$) and second order penalty ($q = 2$). Moreover, they only took into account the diagonal elements of \mathbf{R} , resulting in the simpler penalty matrix $c\delta^{-4}\nabla_2^t \nabla_2$, with $c = \int_a^b \{\mathbf{N}_{j,2}(x)\}^2 dx$. Since c is a constant, it can be absorbed in the penalty constant. For our theoretical investigation we use the general definition of penalty matrix \mathbf{D}_q .

2.2 Penalized splines using truncated polynomial basis functions

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

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

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

$$\widehat{\mathbf{f}}_p = \mathbf{F}(\mathbf{F}^t \mathbf{F} + \lambda_p \tilde{\mathbf{D}}_p)^{-1} \mathbf{F}^t \mathbf{Y}, \quad (5)$$

where $\mathbf{F} = \{\mathbf{F}(x_1)^t, \dots, \mathbf{F}(x_n)^t\}^t$ and $\tilde{\mathbf{D}}_p$ is the diagonal matrix $\text{diag}(0_{p+1}, 1_K)$, indicating that only the spline coefficients are penalized. Note that $\lambda_p \rightarrow \infty$ results in the p th degree polynomial limiting fit.

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, where the $(p + 1)$ th derivative is a generalized function. Indeed,

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

Since the derivative of an indicator function is a Dirac delta function, one finds that

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

Truncated polynomials and B-splines are directly connected, which can be seen from the alternative definition of B-splines as appropriately scaled $(p+1)$ th order divided differences of the truncated polynomials (see de Boor, 2001),

$$N_{j,p+1}(x) = (-1)^{(p+1)}(\kappa_{j+p+1} - \kappa_j)[\kappa_j, \dots, \kappa_{j+p+1}](x - \cdot)_+^p, \quad j = -p, \dots, K, \quad (6)$$

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 (6) 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 exist coefficient vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that $\mathbf{N}\boldsymbol{\beta} = \mathbf{F}\boldsymbol{\alpha}$, implying equivalence of unpenalized estimators. In other words, there exists a square and invertible transition matrix \mathbf{L} (see further), such that $\mathbf{N} = \mathbf{F}\mathbf{L}$.

The equivalence of the *penalized* spline estimators $\widehat{\mathbf{f}}$ and $\widehat{\mathbf{f}}_p$ 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 $\widehat{\mathbf{f}}_p$ in (5) with $\tilde{\mathbf{D}}_p$ as penalty matrix. Using the equality $\mathbf{N} = \mathbf{F}\mathbf{L}$ for the penalized estimator $\widehat{\mathbf{f}}_p$, implies that we can write it as $\widehat{\mathbf{f}}_p = \mathbf{N}(\mathbf{N}^t\mathbf{N} + \lambda_p\mathbf{L}^t\tilde{\mathbf{D}}_p\mathbf{L})^{-1}\mathbf{N}^t\mathbf{Y}$. Thus, fitting with B-splines yields an equivalent estimator to $\widehat{\mathbf{f}}_p$ if we use the penalty term $\lambda_p\mathbf{L}^t\tilde{\mathbf{D}}_p\mathbf{L}$ instead of $\lambda\mathbf{D}_q$. This penalty matrix can be explicitly obtained for equidistant knots. By writing $(\mathbf{N}(x)\boldsymbol{\beta})^p = \sum_{j=0}^K N_{j,1}(x)\beta_j^{(p)} = \sum_{j=1}^K I_{[\kappa_j, \infty)}(x)(\beta_j^{(p)} - \beta_{j-1}^{(p)})$ we find that

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

We formally generalize (3) to $\beta_j^{(p+1)} = (\beta_j^{(p)} - \beta_{j-1}^{(p)})/\delta$ and obtain that

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

Thus, in this case, for equivalence of the estimators, the penalty matrix using B-splines should be $\mathbf{L}^t \tilde{\mathbf{D}}_p \mathbf{L} = \delta^{-2p} \nabla_{p+1}^t \nabla_{p+1} / (p!)^2$. In addition, these calculations provide a method to obtain the transformation matrix \mathbf{L} . In general, also for unequid spaced knots, \mathbf{L} can be found from the equation $\beta^t \mathbf{L}^t \tilde{\mathbf{D}}_p \mathbf{L} \beta = \sum_{j=1}^K (\beta_k^{(p)} - \beta_{j-1}^{(p)})^2 / (p!)^2$.

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

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

where

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

is the set of spline functions of degree p with knots $\underline{\kappa} = \{a = \kappa_0 < \kappa_1 < \dots < \kappa_K < \kappa_{K+1} = b\}$ and $S(1, \underline{\kappa})$ is the set of functions with jumps at the knots. Since $N_{j,p+1}(\cdot)$, $j = -p, \dots, K$ form a basis for $S(p + 1, \underline{\kappa})$ (see Schumaker, 1981), $\hat{f}_{\text{reg}}(x) = \mathbf{N}(x)(\mathbf{N}^t \mathbf{N})^{-1} \mathbf{N}^t \mathbf{Y} \in S(p + 1, \underline{\kappa})$. Further we denote with $s_f(\cdot) = \mathbf{N}(\cdot) \beta \in S(p + 1, \underline{\kappa})$ the best L_∞ approximation to function $f(\cdot)$.

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 $\delta = \max_{0 \leq j \leq K} (\kappa_{j+1} - \kappa_j)$. There exists a constant $M > 0$, such that

$$\delta / \min_{0 \leq j \leq K} (\kappa_{j+1} - \kappa_j) \leq M \text{ and } \delta = o(K^{-1}).$$

(A2) For deterministic design points $x_i \in [a, b]$, $i = 1, \dots, n$ 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, \dots, 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 of the regression spline estimator as

$$E\{\hat{f}_{\text{reg}}(x)\} - f(x) = b_a(x) + o(\delta^{p+1}), \quad (7)$$

$$\text{Var}\{\hat{f}_{\text{reg}}(x)\} = \frac{\sigma^2}{n} \mathbf{N}(x) \mathbf{G}^{-1} \mathbf{N}^t(x) + o\{(n\delta)^{-1}\}, \quad (8)$$

where $\mathbf{G} = \int_a^b \mathbf{N}(x)^t \mathbf{N}(x) \rho(x) dx$ and the approximation bias

$$b_a(x) = -\frac{f^{(p+1)}(x)}{(p+1)!} \sum_{j=0}^K I_{[\kappa_j, \kappa_{j+1})}(x) (\kappa_{j+1} - \kappa_j)^{p+1} B_{p+1}\left(\frac{x - \kappa_j}{\kappa_{j+1} - \kappa_j}\right), \quad (9)$$

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

We will come back to these results in Section 4 where they will be used for the case of penalized splines.

3 Average mean squared error of a penalized spline estimator

In this section we investigate the average mean squared error (AMSE) of a penalized spline estimator and discuss the optimum choice of smoothing parameter λ and number of knots K .

With the Demmler-Reinsch decomposition (Demmler and Reinsch, 1975) the average bias and variance can be expressed in terms of the eigenvalues obtained from the singular value decomposition

$$(\mathbf{N}^t \mathbf{N})^{-t/2} \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1/2} = \mathbf{U} \text{diag}(\mathbf{s}) \mathbf{U}^t, \quad (10)$$

where \mathbf{U} is the matrix of eigenvectors and \mathbf{s} is the vector of eigenvalues s_j . Denote $\mathbf{A} = \mathbf{N}(\mathbf{N}^t \mathbf{N})^{-1/2} \mathbf{U}$. This matrix is semi-orthogonal with $\mathbf{A}^t \mathbf{A} = \mathbf{I}_{K+p+1}$ and $\mathbf{A} \mathbf{A}^t = \mathbf{N}(\mathbf{N}^t \mathbf{N})^{-1} \mathbf{N}^t$. We can rewrite the penalized spline estimator (4) as

$$\hat{\mathbf{f}} = \mathbf{A} \{\mathbf{I}_n + \lambda \text{diag}(\mathbf{s})\}^{-1} \mathbf{A}^t \mathbf{Y} \quad (11)$$

$$= \{\mathbf{I}_n + \lambda \mathbf{A} \text{diag}(\mathbf{s}) \mathbf{A}^t\}^{-1} \mathbf{A} \mathbf{A}^t \mathbf{Y} = \{\mathbf{I}_n + \lambda \mathbf{A} \text{diag}(\mathbf{s}) \mathbf{A}^t\}^{-1} \hat{\mathbf{f}}_{\text{reg}}. \quad (12)$$

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

$$\begin{aligned} \text{AMSE}(\hat{\mathbf{f}}) &= \frac{1}{n} E\{(\hat{\mathbf{f}} - \mathbf{f})^t(\hat{\mathbf{f}} - \mathbf{f})\} \\ &= \frac{\sigma^2}{n} \sum_{j=1}^{K+p+1} \frac{1}{(1 + \lambda s_j)^2} + \frac{\lambda^2}{n} \sum_{j=1}^{K+p+1} \frac{s_j^2 b_j^2}{(1 + \lambda s_j)^2} + \frac{1}{n} \mathbf{f}^t (\mathbf{I}_n - \mathbf{A}\mathbf{A}^t) \mathbf{f}, \end{aligned}$$

where $\mathbf{f} = \{f(x_1), \dots, f(x_n)\}^t$ and $\mathbf{b} = \mathbf{A}^t \mathbf{f}$ with components b_j . Since $\mathbf{A}\mathbf{A}^t$ is an idempotent matrix and $\mathbf{A}\mathbf{A}^t \mathbf{f} = E(\hat{\mathbf{f}}_{\text{reg}})$ we obtain that

$$\text{AMSE}(\hat{\mathbf{f}}) = \frac{\sigma^2}{n} \sum_{j=1}^{K+p+1} \frac{1}{(1 + \lambda s_j)^2} + \frac{\lambda^2}{n} \sum_{j=1}^{K+p+1} \frac{s_j^2 b_j^2}{(1 + \lambda s_j)^2} + \frac{1}{n} \sum_{j=1}^n \left[E\{\hat{f}_{\text{reg}}(x_j)\} - f(x_j) \right]^2.$$

The first term in this equation is the average variance, the second term is the average squared shrinkage bias and the last component is the average squared approximation bias, which can be obtained from (7).

We now study optimal orders of the smoothing parameter λ and of the number of knots K . 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 1 is adopted from Speckman (1985, eqn. 2.5d), see also Eubank (1999, p. 237).

Lemma 1 *Under design condition (A2) and for the eigenvalues obtained in (10),*

$$\begin{aligned} s_1 &= \dots = s_q = 0 \\ s_j &= n^{-1}(j - q)^{2q} c_1 \{1 + o(1)\}, \quad j = q + 1, \dots, K + p + 1, \end{aligned}$$

where c_1 is a constant that depends only on q and the design density.

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

originally obtained for smoothing splines, this result remains valid for this situation with $K < n$.

Using

$$\begin{aligned} \sum_{j=1}^{K+p+1} \frac{1}{(1 + \lambda s_j)^2} &= q + \sum_{j=q+1}^{K+p+1} \frac{1}{(1 + \lambda n^{-1} \tilde{c}_1 (j - q)^{2q})^2} \\ &= \left(\frac{\tilde{c}_1 \lambda}{n} \right)^{-1/2q} \int_0^{K_q} \frac{du}{(1 + u^{2q})^2} + q - 1 + r_q, \end{aligned}$$

with

$$K_q = (K + p + 1 - q)(\lambda \tilde{c}_1)^{1/2q} n^{-1/2q}, \quad (13)$$

$\tilde{c}_1 = c_1 \{1 + o(1)\}$ and $r_q = O(1)$ as the remainder term of the Euler-Maclaurin formula, we can rewrite

$$\begin{aligned} \text{AMSE}(\hat{\mathbf{f}}) &= \frac{\sigma^2}{n} \left(\frac{\tilde{c}_1 \lambda}{n} \right)^{-1/2q} \int_0^{K_q} \frac{du}{(1 + u^{2q})^2} + \frac{\lambda^2}{n} \sum_{j=q+1}^{K+p+1} \frac{b_j^2 s_j^2}{(1 + \lambda s_j)^2} \\ &+ \sum_{i=1}^n \{b_a(x_i)\}^2 / n + o(\delta^{2(p+1)}) + \sigma^2(q - 1 + r_q) / n. \end{aligned} \quad (14)$$

This expression is exact, not a bound. Depending on whether K_q in (13) is smaller or larger than one (and thus depending on K , λ and q), we obtain a different value for the integral in (14) and thus the following results. First, define W^q as the q th order Sobolev space, i.e. $W^q = \{f : f \text{ has } q - 1 \text{ absolute continuous derivatives, } \int_a^b \{f^{(q)}(x)\}^2 dx < \infty\}$.

Theorem 1 *Under assumptions (A1)–(A3) and for $p \leq 2q - 1$ it holds*

(a) *If $K_q < 1$*

$$\begin{aligned} \text{AMSE}(\hat{\mathbf{f}}) &= \frac{\sigma^2}{n} c_2 (K + p + r_q) + \frac{\lambda^2}{n} c_3 \boldsymbol{\beta}^t \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q \boldsymbol{\beta} \\ &+ o(\lambda^2 n^{-2} \delta^{-2q}) + \frac{1}{n} \sum_{i=1}^n \{b_a(x_i)\}^2 + o(\delta^{2(p+1)}) \\ &= O\left(\frac{K}{n}\right) + O\left(\frac{\lambda^2}{n^2} K^{2q}\right) + O(K^{-2(p+1)}), \end{aligned}$$

for $c_2 = c_2(K_q, q) \in (0.6, 1]$ for any $K_q \in (0, 1)$ and $q > 0$ and some constant $c_3 \in [1/4, 1]$. For $K \sim Cn^{1/(2p+3)}$ (with C a constant) and $\lambda = O(n^\gamma)$ with $\gamma \leq (p + 2 - q)/(2p + 3)$ one gets $\text{AMSE}(\hat{\mathbf{f}}) = O(n^{-(2p+2)/(2p+3)})$.

(b) If $K_q > 1$ and $f \in W^q$,

$$\begin{aligned} AMSE(\widehat{\mathbf{f}}) &\leq \frac{\sigma^2}{n} \left\{ q - 1 + r_q + (c_4 - K_q^{1-4q}c_5) \left(\frac{\tilde{c}_1\lambda}{n} \right)^{-1/2q} \right\} \\ &\quad + \frac{\lambda}{4n} \boldsymbol{\beta}^t \mathbf{D}_q \boldsymbol{\beta} + o(\lambda n^{-1}) + \frac{1}{n} \sum_{i=1}^n \{b_a(x_i)\}^2 + o(\delta^{2q}) \\ &= O\left(\frac{n^{1/2q-1}}{\lambda^{1/2q}}\right) + O\left(\frac{\lambda}{n}\right) + O(K^{-2q}), \end{aligned}$$

for constants $c_4 = \int_0^\infty (1 + u^{2q})^{-2} du$ and $c_5 = c_5(K_q, q) \in (0, 1/3]$. For $\lambda = O(n^{1/(2q+1)})$ and $K \sim C_\nu n^\nu$ with $\nu > 1/(2q + 1)$ and C_ν as a constant, one gets $AMSE(\widehat{\mathbf{f}}) = O(n^{-2q/(1+2q)})$.

Case (a) with $K_q < 1$ results in the asymptotic scenario similar to that of regression splines. The AMSE is determined by the average asymptotic variance and the average 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 $K \sim C n^{1/(2p+3)}$. Case (b) with $K_q > 1$ results in the asymptotic scenario close to that of smoothing splines. The AMSE is dominated by the average asymptotic variance and average squared shrinkage bias, while the average squared approximation bias is negligible for the chosen asymptotic order of the number of knots K . The asymptotic order of the AMSE 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 $K_q > 1$.

Note that the assumption on the smoothness of 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 and Smith (1978) the expression for the approximation bias (9) holds for $f(\cdot) \in W^{p+1}$ as well. See also the discussion in Agarwal and Studden (1980), Remark 3.3.

The result of Theorem 1 suggests that the convergence rate of penalized spline estimators in both frameworks depends on the differentiability of the true function. For

example, if the true function f is at least 4 times continuously differentiable (or just $f \in W^4$) and is fitted with cubic B-splines ($p = 3$) and second order penalty ($q = 2$), then $AMSE(\hat{\mathbf{f}}) = O(n^{-7/8})$ for $K_q < 1$ and $AMSE(\hat{\mathbf{f}}) = O(n^{-4/5})$ for $K_q > 1$, implying the faster convergence rate in the first asymptotic scenario. However, if $f \in W^2$ only, then in both frameworks one obtains $AMSE(\hat{\mathbf{f}}) = O(n^{-4/5})$. In general, if the true function $f \in W^l$, with $l > q$, then the AMSE is smaller for the case $K_q < 1$ (K small), than for $K_q > 1$ (K large), since $q \leq p$ is assumed.

Since in practice the smoothness of the function is usually unknown, it is advisable to prefer a small number of knots ($K_q < 1$). This result supports the simulation study of Ruppert (2002) who “found examples where having too many knots degrades the performance of the spline estimator”. This seems to be the first rigorous proof of those empirical findings. However, there is still need for a practical guideline for choosing K and λ , so that $K_q < 1$ is satisfied. Apparently, one first would need to choose K somewhat bigger than needed for estimating the data with unpenalized regression splines. Once this is done, the corresponding smoothing parameter could be estimated with any data-driven criterion like (generalized) cross-validation or AIC. The practical choice of the number of knots and smoothing parameter is important and is planned to be addressed in a separate work.

Similar asymptotic results could be obtained using the mean integrated squared error (MISE) instead of the AMSE (compare for example Wahba (1975) for the AMSE and Rice and Rosenblatt (1981) for the MISE for smoothing splines).

4 Asymptotic bias and variance

4.1 Penalized splines with a “small” number of knots

In the case of a penalized spline estimator $\hat{f}(x)$ with a relatively small number of knots, determined by $K_q < 1$, we are able to derive expressions for the pointwise asymptotic bias and variance. First we relate the penalized spline estimator (4) with $q \leq p$ to a regression spline estimator. We define $\mathbf{G}_{K,n}(x) = \mathbf{N}^t \mathbf{N} / n$ and make the following assumption.

(A4) Eigenvalues of $\lambda n^{-1} \mathbf{G}_{K,n}^{-1} \mathbf{D}_q$ are less than 1.

This allows us to relate regression and penalized spline estimators using a series expansion around $\lambda = 0$, as was done in Wand (1999) in the context of obtaining a plug-in estimator for the optimal value of λ . We start with

$$\begin{aligned} \hat{f}(x) &= \mathbf{N}(x) \left(\frac{1}{n} \mathbf{N}^t \mathbf{N} + \frac{\lambda}{n} \mathbf{D}_q \right)^{-1} \frac{1}{n} \mathbf{N}^t \mathbf{Y} \\ &= \mathbf{N}(x) \left\{ \mathbf{I}_{K+p+q} - \frac{\lambda}{n} \mathbf{G}_{K,n}^{-1} \mathbf{D}_q + \left(\frac{\lambda}{n} \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \right)^2 - \dots \right\} \frac{1}{n} \mathbf{G}_{K,n}^{-1} \mathbf{N}^t \mathbf{Y} \end{aligned} \quad (15)$$

$$= \hat{f}_{\text{reg}}(x) - \frac{\lambda}{n} \mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \mathbf{G}_{K,n}^{-1} \frac{1}{n} \mathbf{N}^t \mathbf{Y} + r_n. \quad (16)$$

The first term in (16) is the regression spline estimator and the second term gives the difference between the penalized and unpenalized (regression spline) estimator, which also contributes to the bias and variance. Assumption (A4) ensures convergence of the series in (15). Note that (A4) is equivalent to the assumption $K_q < 1$ in case (a) of Theorem 1, since K_q^{2q} is the maximum eigenvalue of matrix $\lambda n^{-1} \mathbf{G}_{K,n}^{-t/2} \mathbf{D}_q \mathbf{G}_{K,n}^{-1/2}$, which as a matrix is similar to $\lambda n^{-1} \mathbf{G}_{K,n}^{-1} \mathbf{D}_q$, and thus has the same eigenvalues.

From this expansion we obtain the following result.

Theorem 2 *Suppose $f(\cdot) \in C^{p+1}[a, b]$, assumptions (A1) – (A4) hold and $p \leq 2q - 1$. Then,*

$$E\{\hat{f}(x)\} - f(x) = b_a(x) + b_\lambda(x) + o(\delta^{p+1}) + o(\lambda n^{-1} \delta^{-q}), \quad (17)$$

$$\begin{aligned} \text{Var}\{\hat{f}(x)\} &= \frac{\sigma^2}{n} \mathbf{N}(x) \mathbf{G}^{-1} \mathbf{N}^t(x) - \lambda c_7 \frac{\sigma^2}{n^2} \mathbf{N}(x) \mathbf{G}^{-1} \mathbf{D}_q \mathbf{G}^{-1} \mathbf{N}^t(x) \\ &+ o\{(n\delta)^{-1}\} + o(\lambda n^{-2} \delta^{-(2q+1)}), \end{aligned} \quad (18)$$

with penalization or shrinkage bias

$$b_\lambda(x) = -\frac{\lambda}{n} c_6 \mathbf{N}(x) \mathbf{G}^{-1} \mathbf{D}_q \boldsymbol{\beta},$$

where $c_6 \in [1/2, 3/2]$ and $c_7 \in [3/4, 2]$ are constants.

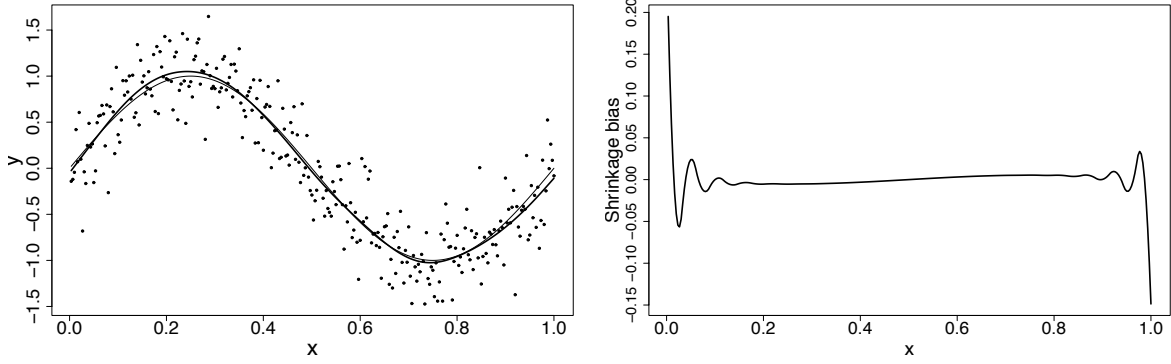


Figure 1: True function $\sin(2\pi x)$ (solid) and its estimate (bold) based on 35 equidistant knots and $p = q = 2$ (left) and the shrinkage bias (right).

Thus, apart from the approximation bias $b_a(x)$ the asymptotic bias of a penalized spline estimator has an additional component – the shrinkage bias $b_\lambda(x)$. As shown in Appendix A.2, the shrinkage bias can be represented as $b_\lambda(x) = -\frac{\lambda}{n}c_6 \mathbf{N}(x)\mathbf{G}^{-1}\Delta_q^t \mathbf{W} \mathbf{s}_f^{(q)}(\boldsymbol{\tau})$ with $\mathbf{W} = \text{diag}\left(\sum_{l=j}^{j+p-q} \int_{\kappa_l}^{\kappa_{l+1}} \mathbf{N}_{j,q}(t)dt\right)$ and $\mathbf{s}_f^{(q)}(\boldsymbol{\tau}) = \{s_f^{(q)}(\tau_{-p+q}), \dots, s_f^{(q)}(\tau_K)\}^t$ for some $\tau_j \in [\kappa_j, \kappa_{j+p+1-q}]$, $j = -p + q, \dots, K$. It is easy to see that for equidistant knots and $p = q = 1$

$$\begin{aligned}
 b_\lambda(x) &= \frac{\lambda}{n}c_6 s_f^{(1)} \sum_{j=0}^K I_{[\kappa_j, \kappa_{j+1})}(x) \left[(\kappa_{j+1} - x) \{(\mathbf{G}^{-1})_{j+1,1} + (\mathbf{G}^{-1})_{j+1,K+2}\} \right. \\
 &\quad \left. + (x - \kappa_j) \{(\mathbf{G}^{-1})_{j+2,1} + (\mathbf{G}^{-1})_{j+2,K+2}\} \right],
 \end{aligned}$$

where $s_f^{(1)}(x) = s_f^{(1)}$ is a constant for $s_f(\cdot) \in S(2, \underline{\kappa})$. Since $|(\mathbf{G}^{-1})_{k,l}| = \varrho^{|k-l|}O(\delta^{-1})$ for some $\varrho \in (0, 1)$ (see Lemma 6.3 in Zhou et al., 1998), the elements $(\mathbf{G}^{-1})_{j,1}$ decrease exponentially with growing j , while the $(\mathbf{G}^{-1})_{j,K+2}$ increase with growing j . Thus, for j close to $[K/2]$, both $(\mathbf{G}^{-1})_{j,K+2}$ and $(\mathbf{G}^{-1})_{j,1}$ 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 $b_a(x)$, the shrinkage bias $b_\lambda(x)$ depends on the design density $\rho(x)$.

Figure 1 shows the shrinkage bias and the penalized spline estimator for the true function $f(x) = \sin(2\pi x)$. The function is evaluated at $n = 300$ equally spaced points on

the $(0, 1)$ interval and errors are taken to be i.i.d $N(0, 0.3^2)$. We used B-splines of degree two and a second order penalty, based on $K = 35$ equidistant knots. The larger bias near the boundaries is clearly observed.

Unpenalized splines do not suffer from boundary effects (Gasser and Müller, 1984) and their approximation bias $b_a(x)$ has the largest values in the regions with large $|f^{(p+1)}(x)|$. Huang (2003b) notices that when the degree of the spline approximation is larger or equal to $(p + 1)$, the number of continuous derivatives of f , then the term $b_a(x)$ is not present. For example, if the twice continuously differentiable function is estimated by linear splines, the approximation bias is given by $b_a(x)$. If, for the same function, we use splines of second or third degree the approximation bias is of a smaller asymptotic order. This motivates to use in practice splines of higher degree. Thus, it is recommendable to choose the upper bound in the condition on p of Theorem 1, that is $p = 2q - 1$, which corresponds exactly to the degree of a smoothing spline estimator penalized with integrated squared q th derivative.

Finally, we investigate the asymptotic orders of variance and bias components. We use the optimal asymptotic orders of the smoothing parameter λ and number of knots K , obtained in Section 3 for case (a). Since under assumptions (A1)-(A4) and for $p \leq 2q - 1$ it holds

$$\begin{aligned} -[E\{\hat{f}(x)\} - f(x)] &= O(\delta^{p+1}) + O(\lambda n^{-1}\delta^{-q}) \\ \text{Var}\{\hat{f}(x)\} &= O(n^{-1}\delta^{-1}) - O(\lambda n^{-2}\delta^{-(2q+1)}), \end{aligned}$$

we find for

$$K \sim Cn^{\frac{1}{2p+3}}, \quad \lambda = O(n^\gamma) \text{ with } \gamma \leq \frac{p+2-q}{2p+3}$$

that for $\gamma = (p+2-q)/(2p+3)$ both bias components b_a and b_λ have the same order, which is $O(n^{-(p+1)/(2p+3)})$, while the second component in the variance is of negligible order. If $\gamma < (p+2-q)/(2p+3)$, then penalization loses its influence and asymptotic orders of the bias and variance are determined by those of an unpenalized estimator. Moreover, Zhou et al. (1998) note that the asymptotic variance of $\hat{f}_{\text{reg}}(x)$ has much bigger values near the boundary, than in the interior. Apparently, adding penalization has no effect on the asymptotic variance.

The case of a penalized estimator (5) using a truncated polynomial basis needs a separate treatment since a different penalty matrix is used. A closed form expression of the penalization matrix is available for equidistant knots only. Replacing $\lambda \mathbf{D}_q$ by $\lambda_p \delta^{-2p} \nabla_{p+1}^t \nabla_{p+1} / (p!)^2$ and modifying (A4) accordingly, in full analogy we obtain that

$$\begin{aligned} E\{\hat{f}_p(x)\} - f(x) &= b_a(x) - \frac{\lambda_p \delta^{-p+1}}{(p!)^2 n} c_6 \mathbf{N}(x) \mathbf{G}^{-1} \nabla_{p+1}^t \mathbf{s}_f^{(p+1)}(\boldsymbol{\kappa}) + o(\delta^{p+1}) + o(\lambda n^{-1} \delta^{-p}) \\ &= O(\delta^{p+1}) + O(\lambda n^{-1} \delta^{-p}), \end{aligned} \quad (19)$$

$$\begin{aligned} \text{Var}\{\hat{f}_p(x)\} &= \frac{\sigma^2}{n} \mathbf{N}(x) \mathbf{G}^{-1} \mathbf{N}^t(x) - \frac{\lambda_p \delta^{-2p} \sigma^2}{(p!n)^2} c_7 \mathbf{N}(x) \mathbf{G}^{-1} \nabla_{p+1}^t \nabla_{p+1} \mathbf{G}^{-1} \mathbf{N}^t(x) \\ &+ o\{(n\delta)^{-1}\} + o(\lambda n^{-2} \delta^{-(2p+2)}) = O\{(n\delta)^{-1}\} + O(\lambda n^{-2} \delta^{-(2p+2)}), \end{aligned} \quad (20)$$

where $\mathbf{s}_f^{(p+1)}(\boldsymbol{\kappa}) = \delta^{-1} \{s_f^{(p)}(\kappa_1), s_f^{(p)}(\kappa_2) - s_f^{(p)}(\kappa_1), \dots, s_f^{(p)}(\kappa_K) - s_f^{(p)}(\kappa_{K-1})\}^t$. It follows that taking $K \sim Cn^{1/(2p+3)}$ and $\lambda_p = O(n^\gamma)$ with $\gamma \leq \frac{2}{2p+3}$ leads to the optimal rate of convergence. Moreover, in this setting not only approximation and shrinkage biases are balanced, both components of the asymptotic variance are of same asymptotic order. More details on the asymptotics of a generalized penalized estimator with truncated polynomial basis functions are available in Kauermann et al. (2007).

4.2 Penalized splines with a “large” number of knots

A penalized spline estimator with a relatively large number of knots, determined by $K_q > 1$, behaves asymptotically similarly to a smoothing spline estimator. Thus, a study of the asymptotic bias and variance of a penalized spline estimator in this case can be carried out through derivation of an equivalent kernel, similar to the classical results on smoothing splines (see e.g. Eubank, 1999). We consider this task beyond the scope of our paper, and plan to address this issue in a separate work. However, we discuss here the available results of Li and Ruppert (2008), who studied the asymptotic properties of a penalized spline estimator by deriving an equivalent kernel.

Note that the general setting of Li and Ruppert (2008) differs from the one adopted in the current paper. Using B-spline basis functions of degree p and equidistant knots, Li and Ruppert (2008) employ a simplified version of penalty \mathbf{D}_q , namely just $\nabla_q^t \nabla_q$.

Also, the spline degree and the penalty order are decoupled, so that cases with $q > p$ are possible, in contrast to our setting with $q \leq p$. Moreover, their true function is assumed to be $2q$ times continuously differentiable, compared to $f \in W^q$ adopted in our paper. However, the case of linear splines penalized with the first order penalty ($p = q = 1$) fits into both settings - the one of Li and Ruppert (2008) as well as of the current paper (under a slightly stronger assumption $f \in C^{p+1}$), making comparison between both results possible.

For interior points, Li and Ruppert (2008) obtain the asymptotic orders $K \sim C_\nu n^\nu$, $\nu > 1/5$ and $\lambda_n = O(n^{2\nu-2/5})$, where λ_n is connected with smoothing parameter λ corresponding to our setting through $\lambda_n = \lambda(K^2/n)$, resulting in $\lambda = O(n^{3/5})$. Clearly, $\nu > 1/5$ implies $K_q = K_1 = (K + p)(\lambda\tilde{c}_1/n)^{1/2} = O(n^\nu n^{-1/5}) > 1$ for n sufficiently large. The optimal rates for K and λ of Li and Ruppert (2008) are given for the non-boundary region and differ from those obtained in case (b) of Theorem 1, since the average shrinkage bias (and thus $AMSE(\hat{\mathbf{f}})$) is dominated by the boundary behavior of the estimator. However, since the average variance is not affected by the boundary behavior, one can observe that the asymptotic rates for K and λ obtained in Li and Ruppert (2008) imply the optimal rate of convergence for the average variance $O\{(n\lambda)^{-1/2}\} = O(n^{-4/5})$. Note that a further study of the estimator in the boundary is possible, using results on kernel smoothing for boundary values (in Li and Ruppert (2008), the local boundary behavior of a penalized spline estimator is considered only for the cases $p = 0$). For the global penalized spline estimator, the results at boundary and interior points would need to be combined to be able to obtain global rates of convergence for the AMSE.

5 Discussion

The results in this paper (and in particular Theorem 1) provide a theoretical justification of the empirical findings that a smaller number of knots can lead to a smaller averaged mean squared error (Ruppert, 2002). Moreover we were able to precisely characterize through K_q in (13) 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 of smoothing splines on the other hand. This paper made a unifying theory, valid for both situations. Results of this paper also made it clear that for the best rates the choices of the degree of the spline p , and the order of the penalty q are connected through $p \leq 2q - 1$, and with better rates for p as large as possible, leading to the familiar $p = 2q - 1$ from the smoothing spline theory.

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. (2007) who verified the use of the Laplace approximation for a mixed model with a growing number of spline basis functions. Since mixed models are tightly connected to Bayesian models using a prior distribution on the spline coefficients, this could also bring additional insight in Bayesian spline estimation (e.g. Carter and Kohn, 1996; Speckman and 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. (2007); 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 the subsequent proofs we make use of the following results

(R1) $\|\mathbf{G}_{K,n}^{-1}\|_{\infty} = O(\delta^{-1})$, $\|\mathbf{G}_{K,n}^{-1} - \mathbf{G}^{-1}\|_{\infty} = o(\delta^{-1})$ and $\|\mathbf{G}_{K,n} - \mathbf{G}\|_{\infty} = o(\delta)$ (Lemmas 6.3 and 6.4 in Zhou et al., 1998).

(R2) $\|\mathbf{D}_q\|_{\infty} = O(\delta^{-2q+1})$ (Lemma 6.1 in Cardot, 2000).

(R3) Under (A1) - (A3) it holds $\max_{q \leq j \leq K} \int_a^b N_{j,p+1}(u) \{f(u) - s_f(u)\} dQ_n(u) = o(\delta^{p+2})$

(Lemma 6.10 in Agarwal and Studden, 1980) and thus

$$E\{\hat{f}_{\text{reg}}(x) - s_f(x)\} = \mathbf{N}(x)\mathbf{G}_{K,n}^{-1}\frac{1}{n}\mathbf{N}(\mathbf{f} - \mathbf{s}_f) = o(\delta^{p+1}),$$

with $\mathbf{f} = \{f(x_1), \dots, f(x_n)\}^t$ and $\mathbf{s}_f = \{s_f(x_1), \dots, s_f(x_n)\}^t$.

Lemma 2 $\max_{1 \leq i, j \leq K} |(\mathbf{G}_{K,n}^{-1}\mathbf{D}_q\mathbf{G}_{K,n}^{-1})_{i,j} - (\mathbf{G}^{-1}\mathbf{D}_q\mathbf{G}^{-1})_{i,j}| = o(\delta^{-(2q+1)})$.

Proof. Rewriting

$$\begin{aligned} \mathbf{G}_{K,n}^{-1}\mathbf{D}_q\mathbf{G}_{K,n}^{-1} - \mathbf{G}^{-1}\mathbf{D}_q\mathbf{G}^{-1} &= \mathbf{G}_{K,n}^{-1}\mathbf{D}_q(\mathbf{G}_{K,n}^{-1} - \mathbf{G}^{-1}) + \mathbf{G}^{-1}\mathbf{D}_q(\mathbf{G}_{K,n}^{-1} - \mathbf{G}^{-1}) \\ &\quad + \mathbf{G}_{K,n}^{-1}\mathbf{D}_q\mathbf{G}^{-1} - \mathbf{G}^{-1}\mathbf{D}_q\mathbf{G}_{K,n}^{-1} \end{aligned}$$

and using (R1) – (R2) we obtain

$$\begin{aligned} \|\mathbf{G}_{K,n}^{-1}\mathbf{D}_q\mathbf{G}^{-1} - \mathbf{G}^{-1}\mathbf{D}_q\mathbf{G}_{K,n}^{-1}\|_\infty &= \|\mathbf{G}_{K,n}^{-1}(\mathbf{D}_q\mathbf{G}^{-1} - \mathbf{G}_{K,n}\mathbf{G}^{-1}\mathbf{D}_q\mathbf{G}_{K,n}^{-1})\|_\infty \\ &= \|\mathbf{G}_{K,n}^{-1}(\mathbf{D}_q\mathbf{G}^{-1} - (\mathbf{G}_{K,n} - \mathbf{G})\mathbf{G}^{-1}\mathbf{D}_q\mathbf{G}_{K,n}^{-1} - \mathbf{D}_q\mathbf{G}_{K,n}^{-1})\|_\infty \\ &= \|\mathbf{G}_{K,n}^{-1}\mathbf{D}_q(\mathbf{G}^{-1} - \mathbf{G}_{K,n}^{-1} + o(\delta^{-1}))\|_\infty = o(\delta^{-(2q+1)}), \end{aligned}$$

from which the result follows. \square

Lemma 3 For any $K + p + 1$ dimensional vectors \mathbf{v} and \mathbf{w} such that $\mathbf{v}^t\mathbf{w} \geq 0$ it holds

$$0 \leq \mathbf{v}^t(\mathbf{N}^t\mathbf{N})^{-1}\mathbf{D}_q\mathbf{w} \leq \frac{\tilde{c}_1}{n}(K + p + 1 - q)^{2q}\mathbf{v}^t\mathbf{w}.$$

Proof. Since matrices $(\mathbf{N}^t\mathbf{N})^{-1}\mathbf{D}_q$ and $(\mathbf{N}^t\mathbf{N})^{-1/2}\mathbf{D}_q(\mathbf{N}^t\mathbf{N})^{-1/2}$ are similar, they have the same eigenvalues. Thus, $(\mathbf{N}^t\mathbf{N})^{-1}\mathbf{D}_q = \tilde{\mathbf{U}}\text{diag}(\mathbf{s})\tilde{\mathbf{U}}^t$, with $\tilde{\mathbf{U}}$ a matrix of eigenvalues and where \mathbf{s} is defined in Lemma 1, so that $\min_{1 \leq j \leq K+p+1}(s_j) = 0$ and $\max_{q+1 \leq j \leq K+p+1}(s_j) = (K + p + 1 - q)^{2q}\tilde{c}_1n^{-1}$. Since $(\mathbf{N}^t\mathbf{N})^{-1}\mathbf{D}_q$ and $\mathbf{w}\mathbf{v}^t$ are positive semidefinite and $\mathbf{v}^t(\mathbf{N}^t\mathbf{N})^{-1}\mathbf{D}_q\mathbf{w} = \text{tr}\{(\mathbf{N}^t\mathbf{N})^{-1}\mathbf{D}_q\mathbf{w}\mathbf{v}^t\} = \text{tr}\{\text{diag}(\mathbf{s})\tilde{\mathbf{U}}^t\mathbf{w}\mathbf{v}^t\tilde{\mathbf{U}}\}$, the results follows from

$$0 \leq \text{tr}\{\text{diag}(\mathbf{s})\tilde{\mathbf{U}}^t\mathbf{w}\mathbf{v}^t\tilde{\mathbf{U}}\} \leq \frac{\tilde{c}_1}{n}(K + p + 1 - q)^{2q}\mathbf{v}^t\mathbf{w},$$

see Fang et al. (1994) or Xing et al. (2000). \square

If $\mathbf{v}^t\mathbf{w} \leq 0$, the lemma leads to the following result

$$0 \leq |\mathbf{v}^t(\mathbf{N}^t\mathbf{N})^{-1}\mathbf{D}_q\mathbf{w}| \leq \frac{\tilde{c}_1}{n}(K + p + 1 - q)^{2q}|\mathbf{v}^t\mathbf{w}|.$$

A.1 Proof of Theorem 1

Let us first consider case (a), that is $K_q < 1$. Using a series expansion around zero of $(1+x)^{-2} = \sum_{j=0}^{\infty} (-1)^j (j+1)x^j$ for $|x| < 1$ we easily find

$$\int_0^{K_q} \frac{du}{(1+u^{2q})^2} = K_q \sum_{j=0}^{\infty} (-1)^j (j+1) \frac{K_q^{2qj}}{2qj+1} = K_q {}_2F_1 \left(2, \frac{1}{2q}; \frac{1+2q}{2q}, -K_q^{2q} \right) =: K_q c_2,$$

where ${}_2F_1(2, 1/(2q); 1 + 1/(2q), -K_q^{2q})$ denotes the hypergeometric series (see Ch. 15 of Abramowitz and Stegun, 1972), which for any $K_q < 1$ and $q > 0$ converges to some $c_2 = c_2(K_q, q) \in (0.6, 1]$. With this we obtain that the average variance in case (a) has the asymptotic order $O(K/n)$.

Using the definitions of \mathbf{b} and \mathbf{s} we can represent the average squared shrinkage bias as

$$\frac{\lambda^2}{n} \sum_{j=q+1}^{K+p+1} b_j^2 s_j^2 \left\{ 1 - \frac{\lambda s_j (2 + \lambda s_j)}{(1 + \lambda s_j)^2} \right\} = \frac{\lambda^2}{n} \{ \boldsymbol{\beta}_f^t \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q \boldsymbol{\beta}_f - r_b \},$$

where $r_b = \sum_{j=1}^{\infty} \lambda^j (-1)^{j+1} (j+1) \boldsymbol{\beta}_f^t \{ \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \}^j \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q \boldsymbol{\beta}_f$ and $\boldsymbol{\beta}_f = (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{N}^t \mathbf{f}$. With Lemma 3 we can bound

$$0 \leq r_b \leq \frac{K_q^{2q} (2 + K_q^{2q})}{(1 + K_q^{2q})^2} \boldsymbol{\beta}_f^t \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q \boldsymbol{\beta}_f,$$

so that for $K_q < 1$ there exists a constant $c_3 \in [1/4, 1]$ such that average squared shrinkage bias equals $\lambda^2 c_3 \boldsymbol{\beta}_f^t \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q \boldsymbol{\beta}_f / n$. Adding and subtracting \mathbf{s}_f from \mathbf{f} in $\boldsymbol{\beta}_f$ we find

$$\begin{aligned} \boldsymbol{\beta}_f^t \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q \boldsymbol{\beta}_f &= \boldsymbol{\beta}^t \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q \boldsymbol{\beta} \\ &+ 2(\mathbf{f} - \mathbf{s}_f)^t \mathbf{N} (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{N}^t \mathbf{s}_f \\ &+ (\mathbf{f} - \mathbf{s}_f)^t \mathbf{N} (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{N}^t (\mathbf{f} - \mathbf{s}_f) \\ &= \boldsymbol{\beta}^t \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q \boldsymbol{\beta} + o(\delta^{p+1-4q} n^{-1}) + o(\delta^{2p+2-4q} n^{-1}), \end{aligned}$$

where (R3) and Lemma 3 where applied to obtain the orders of two last terms. To find the asymptotic order of $\boldsymbol{\beta}^t \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q \boldsymbol{\beta}$ we again apply Lemma 3 and obtain

$$\begin{aligned} \boldsymbol{\beta}^t \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q \boldsymbol{\beta} &\leq \frac{\tilde{c}_1 (K + p + 1 - q)^{2q}}{n} \boldsymbol{\beta}^t \mathbf{D}_q \boldsymbol{\beta} \\ &= \frac{\tilde{c}_1 (K + p + 1 - q)^{2q}}{n} \int_a^b \left[\{ \mathbf{N}(x) \boldsymbol{\beta} \}^{(q)} \right]^2 dx. \end{aligned}$$

Since the penalty was assumed to be finite (see below eqn. (2)) and $p \leq 2q - 1$ we obtain that

$$\frac{\lambda^2}{n} c_3 \boldsymbol{\beta}_f^t \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q \boldsymbol{\beta}_f = \frac{\lambda^2}{n} c_3 \boldsymbol{\beta}^t \mathbf{D}_q (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{D}_q \boldsymbol{\beta} + o(\lambda^2 n^{-2} \delta^{-2q}) = O\left(\frac{\lambda^2}{n^2} K^{2q}\right).$$

Finally, the average squared approximation bias has the asymptotic order $O(K^{-2(p+2)})$, as follows from (9). 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 $K \sim Cn^{1/(2p+3)}$ (with C a constant) implies the average variance and 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.

For values $K_q > 1$ we can find the integral in (14) as the difference

$$\int_0^\infty \frac{du}{(1+u^{2q})^2} - \int_{K_q}^\infty \frac{du}{(1+u^{2q})^2} =: c_4 - \int_{K_q}^\infty \frac{du}{(1+u^{2q})^2}.$$

Changing the integration variable to its reciprocal in the second integral and using a series expansion of $(1+x)^{-2}$ again, one easily obtains that

$$\begin{aligned} \int_{K_q}^\infty \frac{du}{(1+u^{2q})^2} &= \sum_{j=0}^{\infty} (-1)^j (j+1) \frac{K_q^{-2q(j+2)+1}}{2q(j+2)-1} \\ &= K_q^{1-4q} {}_2F_1\left(2, \frac{4q-1}{2q}; \frac{6q-1}{2q}, -K_q^{-2q}\right) (4q-1)^{-1} =: K_q^{1-4q} c_5, \end{aligned}$$

where ${}_2F_1(2, (4q-1)(2q)^{-1}; (6q-1)(2q)^{-1}, -K_q^{-2q})$ is a converging hypergeometric series, with $c_5 = c_5(K_q, q) \in (0, 1/3]$ for any $K_q > 1$ and $q > 0$. Thus, for case (b) the average variance has the asymptotic order $O(n^{1/2q-1} \lambda^{-1/2q})$.

Since the function $x(1+x)^{-2} \leq 1/4$ for any $x > 0$, one can bound the average shrinkage bias

$$\frac{\lambda}{n} \sum_{j=q+1}^{K+p+1} b_j^2 s_j \frac{\lambda s_j}{(1+\lambda s_j)^2} \leq \frac{\lambda}{4n} \sum_{j=q+1}^{K+p+1} b_j^2 s_j = \frac{\lambda}{4n} \boldsymbol{\beta}_f^t \mathbf{D}_q \boldsymbol{\beta}_f.$$

Again, adding and subtracting \mathbf{s}_f from \mathbf{f} in $\boldsymbol{\beta}_f$ we find

$$\boldsymbol{\beta}_f \mathbf{D}_q \boldsymbol{\beta}_f = \boldsymbol{\beta} \mathbf{D}_q \boldsymbol{\beta} + o(\delta^{p+1-2q}).$$

Thus, the average squared approximation bias is of order $O(\lambda/n)$. It is straightforward to see that $\lambda = O(n^{1/(2q+1)})$ balances the average squared shrinkage bias and the average variance. For the condition $K_q > 1$ to be fulfilled, the number of knots should satisfy $K \sim C_\nu n^\nu$, with $\nu > 1/(2q+1)$ and C_ν as a constant. This implies that the average approximation bias is negligible with the order $O(n^{-\nu'})$, with $\nu' > 2q/(2q+1)$. Thus, $\text{AMSE}(\hat{\mathbf{f}}) = O(n^{-2q/(1+2q)})$. \square

A.2 Proof of Theorem 2

From (16) we obtain that

$$E\{\hat{f}(x)\} - f(x) = E\{\hat{f}_{\text{reg}}(x)\} - f(x) - \frac{\lambda}{n} \mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \mathbf{G}_{K,n}^{-1} \mathbf{N}^t \frac{1}{n} \mathbf{f} + E(r_n).$$

Rewriting

$$E(r_n) = \frac{\lambda}{n} \sum_{j=1}^{\infty} (-1)^{j+1} \left(\frac{\lambda}{n}\right)^j \mathbf{N}(x) (\mathbf{G}_{K,n}^{-1} \mathbf{D}_q)^j \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \mathbf{G}_{K,n}^{-1} \mathbf{N}^t \frac{1}{n} \mathbf{f}$$

and using Lemma 3 we find

$$|E(r_n)| \leq \frac{K_q^{2q}}{1 + K_q^{2q}} \frac{\lambda}{n} |\mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \mathbf{G}_{K,n}^{-1} \mathbf{N}^t \frac{1}{n} \mathbf{f}|.$$

Since $K_q < 1$ according to assumption (A4), $K_q/(1 + K_q) < 1/2$ and thus there exists some constant $c_6 \in [1/2, 3/2]$, such that

$$\begin{aligned} E\{\hat{f}(x)\} - f(x) &= E\{\hat{f}_{\text{reg}}(x) - s_f(x)\} + \{s_f(x) - f(x)\} + \frac{\lambda}{n} c_6 \mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \mathbf{G}_{K,n}^{-1} \mathbf{N}^t \frac{1}{n} \mathbf{s}_f \\ &\quad - \frac{\lambda}{n} c_6 \mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \mathbf{G}_{K,n}^{-1} \mathbf{N}^t \frac{1}{n} (\mathbf{f} - \mathbf{s}_f). \end{aligned}$$

The order of the first component is given by (R3). According to Barrow and Smith (1978) it holds that $s_f(x) - f(x) = b_a(x) + o(\delta^{p+1})$. With Lemma 3, assumption (A4) and result (R3) we obtain

$$\frac{\lambda}{n} c_6 |\mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \mathbf{G}_{K,n}^{-1} \mathbf{N}^t \frac{1}{n} (\mathbf{f} - \mathbf{s}_f)| \leq \frac{\lambda \tilde{c}_1}{n} c_6 (K + p + 1 - q)^{2q} o(\delta^{p+1}) = o(\delta^{p+1}).$$

Thus,

$$E\{\hat{f}(x)\} - f(x) = b_a(x) + \frac{\lambda}{n} c_6 \mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \boldsymbol{\beta} + o(\delta^{p+1}),$$

with $\boldsymbol{\beta} = \mathbf{G}_{K,n}^{-1} \mathbf{N}^t \mathbf{s}_f / n = (\mathbf{N}^t \mathbf{N})^{-1} \mathbf{N}^t \mathbf{s}_f$. Using the definition of penalty \mathbf{D}_q and noting that $s_f^{(q)}(x) = (\mathbf{N}(x) \boldsymbol{\beta})^{(q)} = \mathbf{N}_q(x) \boldsymbol{\Delta}_q \boldsymbol{\beta}$, with $\mathbf{N}_q(x) = \{N_{-p+q, p+1-q}(x), \dots, N_{K, p+1-q}(x)\}$ we can apply the mean value theorem and rewrite

$$\begin{aligned} -\frac{\lambda}{n} c_6 \mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \boldsymbol{\beta} &= -\frac{\lambda}{n} c_6 \mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \boldsymbol{\Delta}_q^t \int_a^b \mathbf{N}_q^t(x) s_f^{(q)}(x) dx \\ &= -\frac{\lambda}{n} c_6 \mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \boldsymbol{\Delta}_q^t \mathbf{W} \mathbf{s}_f^{(q)}(\boldsymbol{\tau}), \end{aligned}$$

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

Let us consider the shrinkage bias term in more detail. First, we decompose $\boldsymbol{\Delta}_q^t = [\boldsymbol{\Delta}_{q,1}, \boldsymbol{\Delta}_{q,2}, \boldsymbol{\Delta}_{q,3}]^t$ with $\boldsymbol{\Delta}_{q,1}^t$ and $\boldsymbol{\Delta}_{q,3}^t$ as $q \times (K+p+1-q)$ dimensional matrices and $\boldsymbol{\Delta}_{q,2}^t$ as a $(K+p+1-2q) \times (K+p+1-q)$ matrix of weighted differences as defined in (3), so that $\boldsymbol{\Delta}_q^t \mathbf{W} \mathbf{s}_f^{(q)}(\boldsymbol{\tau}) = [\{\boldsymbol{\Delta}_{q,1}^t \mathbf{W} \mathbf{s}_f^{(q)}(\boldsymbol{\tau})\}^t, \{\boldsymbol{\Delta}_{q,2}^t \mathbf{W} \mathbf{s}_f^{(q)}(\boldsymbol{\tau})\}^t, \{\boldsymbol{\Delta}_{q,3}^t \mathbf{W} \mathbf{s}_f^{(q)}(\boldsymbol{\tau})\}^t]^t$. Since for equidistant knots $\{\mathbf{W}\}_{jj} = \tilde{\omega}$, a constant, for all j and $\boldsymbol{\Delta}_{q,2}^t$ is the matrix corresponding to a divided difference operator of order q , we can rewrite

$$\boldsymbol{\Delta}_q^t \mathbf{W} \mathbf{s}_f^{(q)}(\boldsymbol{\tau}) = \tilde{\omega} [\{\boldsymbol{\Delta}_{q,1}^t \mathbf{s}_f^{(q)}(\boldsymbol{\tau})\}^t, \{\mathbf{s}_f^{(2q)}(\tilde{\boldsymbol{\tau}})/q!\}^t, \{\boldsymbol{\Delta}_{q,3}^t \mathbf{s}_f^{(q)}(\boldsymbol{\tau})\}^t]^t,$$

for some $\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_{-p+q}, \dots, \tilde{\tau}_{K-q})$ with $\tilde{\tau}_j \in [\tau_j, \tau_{j+q}]$, $j = -p+q, \dots, K-q$. Since $s_f(x)$ is a piecewise polynomial of degree p , $s_f^{(2q)}(x)$ disappears if $p \leq 2q-1$. For $p > 2q-1$ we find $s_f^{(2q)}(x) \neq 0$ and thus the order of the shrinkage bias would increase with δ^{-1} . Using assumption (A1) one can show that $\{\boldsymbol{\Delta}_{q,2}^t \mathbf{W} \mathbf{s}_f^{(q)}(\boldsymbol{\tau})\}^t$ is asymptotically small for $p \leq 2q-1$. These considerations justify the assumption on the relationship between p and q made in Theorem 1.

Further, since $N_{j,q}(\cdot) \leq 1$, one finds $\|\mathbf{W}\|_\infty = O(\delta)$. Moreover, by definition $\|\boldsymbol{\Delta}_q\|_\infty = O(\delta^{-q})$ (see also Lemma 6.1 in Cardot, 2000). Thus, using (R1) and noting that $\|\mathbf{s}_f^{(q)}(\boldsymbol{\tau})\|_\infty = O(1)$ and $s_f^{(2q)}(x) = 0$ we obtain the shrinkage bias $b_\lambda(x)$ as

$$-\frac{\lambda}{n} c_6 \mathbf{N}(x) \mathbf{G}^{-1} \boldsymbol{\Delta}_q^t \mathbf{W} \mathbf{s}_f^{(q)}(\boldsymbol{\tau}) - \frac{\lambda}{n} c_6 \mathbf{N}(x) (\mathbf{G}_{K,n}^{-1} - \mathbf{G}^{-1}) \boldsymbol{\Delta}_q^t \mathbf{W} \mathbf{s}_f^{(q)}(\boldsymbol{\tau})$$

$$\begin{aligned}
&= -\frac{\lambda}{n} c_6 \mathbf{N}(x) \mathbf{G}^{-1} \mathbf{\Delta}_q^t \mathbf{W} \mathbf{s}_f^{(q)}(\boldsymbol{\tau}) + o(\lambda n^{-1} \delta^{-q}) \\
&\stackrel{\text{def}}{=} b_\lambda(x) + o(\lambda n^{-1} \delta^{-q}).
\end{aligned}$$

From (15) we find

$$\begin{aligned}
\text{Var}\{\hat{f}(x)\} &= \frac{\sigma^2}{n} \mathbf{N}(x) \left(\mathbf{I}_n - \frac{\lambda}{n} \mathbf{G}_{K,n}^{-1} \mathbf{D}_q + \dots \right)^2 \mathbf{G}_{K,n}^{-1} \mathbf{N}^t(x) \\
&= \text{Var}\{\hat{f}_{\text{reg}}(x)\} - 2\lambda \frac{\sigma^2}{n^2} \mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \mathbf{G}_{K,n}^{-1} \mathbf{N}^t(x) \\
&\quad + 3\lambda^2 \frac{\sigma^2}{n^3} \mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \mathbf{G}_{K,n}^{-1} \mathbf{N}^t(x) + \dots
\end{aligned}$$

with $\text{Var}\{\hat{f}_{\text{reg}}(x)\}$ as given in (8). Applying Lemma 3 again, we can rewrite

$$\text{Var}\{\hat{f}(x)\} = \text{Var}\{\hat{f}_{\text{reg}}(x)\} - \lambda \frac{\sigma^2}{n^2} c_7 \mathbf{N}(x) \mathbf{G}_{K,n}^{-1} \mathbf{D}_q \mathbf{G}_{K,n}^{-1} \mathbf{N}^t(x),$$

with some $c_7 \in [3/4, 2]$. Finally, using Lemma 2 we obtain that

$$\begin{aligned}
&- \lambda \frac{\sigma^2}{n^2} c_7 \mathbf{N}(x) \mathbf{G}^{-1} \mathbf{D}_q \mathbf{G}^{-1} \mathbf{N}^t(x) - \lambda \frac{\sigma^2}{n^2} c_7 \mathbf{N}(x) (\mathbf{G}_{K,n}^{-1} \mathbf{D}_q \mathbf{G}_{K,n}^{-1} - \mathbf{G}^{-1} \mathbf{D}_q \mathbf{G}^{-1}) \mathbf{N}^t(x) \\
&= -\lambda \frac{\sigma^2}{n^2} c_7 \mathbf{N}(x) \mathbf{G}^{-1} \mathbf{D}_q \mathbf{G}^{-1} \mathbf{N}^t(x) + o(\lambda n^{-2} \delta^{-(2q+1)}),
\end{aligned}$$

proving Theorem 2. □

Acknowledgements

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

References

- Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York.
- Agarwal, G. and Studden, W. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.*, 8:1307–1325.

- Barrow, D. L. and Smith, P. W. (1978). Asymptotic properties of best $L_2[0, 1]$ approximation by splines with variable knots. *Quart. Appl. Math.*, 36(3):293–304.
- Besse, P., Cardot, H., and Ferraty, F. (1997). Simultaneous nonparametric regression of unbalanced longitudinal data. *Comp. Statist. Data Anal.*, 24:255–270.
- Brumback, B. A., Ruppert, D., and 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. and Kohn, R. (1996). Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika*, 83(3):589–601.
- Cox, D. D. (1983). Asymptotics for M -type smoothing splines. *Ann. Statist.*, 11(2):530–551.
- 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.
- Demmler, A. and Reinsch, C. (1975). Oscillation matrices with spline smoothing. *Numer. Math.*, 24(5):375–382.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B -splines and penalties. *Stat. Science*, 11(2):89–121. With comments and a rejoinder by the authors.
- Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*, volume 157 of *Statistics: Textbooks and Monographs*. Marcel Dekker Inc., New York, second edition.
- Fang, Y., Loparo, K., and Feng, X. (1994). Inequalities for the trace of matrix product. *IEEE Trans. Automat. Control*, 39(12):2489–2490.
- Gasser, T. and Müller, H. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.*, 11:171–185.
- Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*, volume 58 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.
- Hall, P. and 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(3):207–216.
- Huang, J. Z. (2003b). Local asymptotics for polynomial spline regression. *Ann. Statist.*, 31(5):1600–1635.
- Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2007). Some asymptotic results on generalized penalized spline smoothing. Research report 0733, ORSTAT, K.U.Leuven.
- Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, 46:1071–1085.
- Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, 95(2):415–436.
- Nychka, D. (1995). Splines as local smoothers. *Ann. Statist.*, 23(4):1175–1197.
- Oehlert, G. W. (1992). Relaxed boundary smoothing splines. *Ann. Statist.*, 20(1):146–160.
- O’Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Stat. Science*, 1:505–527. With discussion.
- Rice, J. and Rosenblatt, M. (1981). Integrated mean squared error of a smoothing spline. *J. Approx. Theory*, 33(4):353–369.
- Rice, J. and Rosenblatt, M. (1983). Smoothing splines: regression, derivatives and deconvolution. *Ann. Statist.*, 11(1):141–156.
- Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.*, 11(4):735–757.
- Ruppert, D. and Carroll, R. (2000). Spatially-adaptive penalties for spline fitting. *Aust. N. Z. J. Statist.*, 42:205–224.
- Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge, UK.
- Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.
- Schwetlick, H. and Kunert, V. (1993). Spline smoothing under constraints on derivatives. *BIT*, 33(3):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(3):970–983.
- Speckman, P. L. and Sun, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, 90(2):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(4):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(3):349–362.
- Utreras, F. (1983). Natural spline functions, their associated eigenvalue problem. *Numer. Math.*, 42(1):107–117.
- Utreras, F. (1988). Boundary effects on convergence rates for Tikhonov regularization. *J. Approx. Theory*, 54(3):235–249.
- Wahba, G. (1975). Smoothing noisy data with spline functions. *Numer. Math.*, 24(5):383–393.
- 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. (1999). On the optimal amount of smoothing in penalized spline regression. *Biometrika*, 86:936–940.
- Xing, W., Zhang, Q., and Wang, Q. (2000). A trace bound for a general square matrix product. *IEEE Trans. Automat. Control*, 45(8):1563–1565.
- Zhou, S., Shen, X., and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.*, 26(5):1760–1782.