Biometrika (2013), **100**, 4, *pp.* 955–970
Printed in Great Britain

doi: 10.1093/biomet/ast031
Advance Access publication 30 August 2013

Smoothing splines with varying smoothing parameter

BY XIAO WANG

Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, Indiana 47907, U.S.A.

wangxiao@purdue.edu

PANG DU

Department of Statistics, Virginia Tech, 406-A Hutcheson Hall, Blacksburg, Virginia 24061, U.S.A.

pangdu@vt.edu

AND JINGLAI SHEN

Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, Maryland 21250, U.S.A.

shenj@umbc.edu

SUMMARY

This paper considers the development of spatially adaptive smoothing splines for the estimation of a regression function with nonhomogeneous smoothness across the domain. Two challenging issues arising in this context are the evaluation of the equivalent kernel and the determination of a local penalty. The penalty is a function of the design points in order to accommodate local behaviour of the regression function. We show that the spatially adaptive smoothing spline estimator is approximately a kernel estimator, and that the equivalent kernel is spatially dependent. The equivalent kernels for traditional smoothing splines are a special case of this general solution. With the aid of the Green's function for a two-point boundary value problem, explicit forms of the asymptotic mean and variance are obtained for any interior point. Thus, the optimal roughness penalty function is obtained by approximately minimizing the asymptotic integrated mean squared error. Simulation results and an application illustrate the performance of the proposed estimator.

Some key words: Equivalent kernel; Green's function; Nonparametric regression; Smoothing spline; Spatially adaptive smoothing.

1. INTRODUCTION

Smoothing splines play a central role in nonparametric curve-fitting. Recent surveys include Wahba (1990), Eubank (1999), Gu (2002), and Eggermont & LaRiccia (2009). Specifically, consider the problem of estimating the mean function from a regression model

$$y_i = f_0(t_i) + \sigma(t_i)\epsilon_i \quad (i = 1, \ldots, n),$$

where the t_i are the design points on [0, 1], the ϵ_i are independent and identically distributed random variables with zero mean and unit variance, $\sigma^2(\cdot)$ is the variance function, and f_0 is the underlying true regression function. The traditional smoothing spline is formulated as the solution f to the minimization of

$$\frac{1}{n}\sum_{i=1}^{n}\sigma^{-2}(t_i)\{y_i - f(t_i)\}^2 + \lambda\int_0^1 \{f^{(m)}(t)\}^2\,\mathrm{d}x,$$

where $\lambda > 0$ is the penalty parameter controlling the trade-off between the goodness-of-fit and the smoothness of the fitted function. Smoothing splines have a solid theoretical foundation and are among the most widely used methods for nonparametric regression (Cox, 1983; and a 1981 unpublished technical report by P. L. Speckman of the University of Oregon).

The traditional smoothing spline model has a major deficiency: it uses a global smoothing parameter λ, so the degree of smoothness of f_0 remains about the same across the design points. This makes it difficult to efficiently estimate functions with nonhomogeneous smoothness. Wahba (1995) suggested using a more general penalty term, where the constant λ is replaced by a roughness penalty function $\lambda(\cdot)$. Since $\lambda(\cdot)$ is then a function of t, the model becomes adaptive in the sense that it accommodates the local behaviour of f_0 and imposes a heavier penalty in regions of lower curvature of f_0. Pintore et al. (2006) used a piecewise constant approximation for $\lambda(\cdot)$, but this requires specification of the number of knots, the knot locations, and the values of $\lambda(\cdot)$ between those locations. Storlie et al. (2010) discussed some computational issues with spatially adaptive smoothing splines. Liu & Guo (2010) refined the piecewise constant idea and designed a data-driven algorithm to determine the optimal jump locations and sizes for $\lambda(\cdot)$. Besides adaptive smoothing splines, many other adaptive methods have been developed, including variable-bandwidth kernel smoothing (Müller & Stadtmüller, 1987), adaptive wavelet shrinkage (Donoho & Johnstone, 1994, 1995, 1998), local polynomials with variable bandwidth (Fan & Gijbels, 1996), local penalized splines (Ruppert & Carroll, 2000), regression splines (Friedman & Silverman, 1989; Stone et al., 1997; Luo & Wahba, 1997; Hansen & Kooperberg, 2002), and free-knot splines (Mao & Zhao, 2003). Further, Bayesian adaptive regression has been reported by Smith & Kohn (1996), DiMatteo et al. (2001), and Wood et al. (2002). Nevertheless, adaptive smoothing splines have the advantages of computational efficiency and easy extension to multi-dimensional covariates via the smoothing spline analysis of variance technique (Wahba, 1990; Gu, 2002). Moreover, the results in the present paper can be extended to the more general L-spline smoothing (Kimeldorf & Wahba, 1971; Kohn & Ansley, 1983; Wahba, 1985). Also, the usual Reinsch scheme can be easily modified to this case.

Let $W_2^m = \{f : f^{(m-1)}$ is absolutely continuous and $f^{(m)} \in L_2[0, 1]\}$, where $L_2[0, 1]$ denotes the space of Lebesgue square-integrable functions, endowed with the usual norm $\|\cdot\|_2$ and inner product $(\cdot\,,\cdot)_2$. The method of adaptive smoothing splines involves finding $f \in W_2^m$ to minimize the functional

$$\psi(f) = \frac{1}{n}\sum_{i=1}^{n}\sigma^{-2}(t_i)\{y_i - f(t_i)\}^2 + \lambda\int_0^1 \rho(t)\{f^{(m)}(t)\}^2\,\mathrm{d}t, \tag{1}$$

where $\lambda > 0$ is the penalty parameter and $\rho : [0, 1] \to (0, \infty)$ denotes the adaptive penalty function; more properties of ρ will be stated later. Here, by incorporating a function $\rho(t)$ into the roughness penalty, we generalize the traditional smoothing splines, which correspond to $\rho(t) \equiv 1$. A two-point boundary value problem technique has been developed to find the asymptotic mean squared error of the adaptive smoothing spline estimator with the aid of the Green's function. Thus, the optimal roughness penalty function is obtained explicitly by approximately minimizing the asymptotic integrated mean squared error. Asymptotic analysis of traditional smoothing splines using Green's functions was performed by Rice & Rosenblatt (1983), Silverman (1984), Messer (1991), Nychka (1995), Eggermont & LaRiccia (2009) and Wang et al. (2010); an extension to certain adaptive splines was presented in Abramovich & Grinshtein (1999). In this paper we take a different approach, and develop a general framework for asymptotic analysis of adaptive smoothing splines. This yields a systematic, yet simpler, method for obtaining closed-form expressions of equivalent kernels for interior points, as well as for asymptotic analysis. Our estimator possesses the interpretation of spatial adaptivity (Donoho & Johnstone, 1998), and the equivalent kernel may vary in shape and bandwidth from point to point, depending on the data.

2. Characterizations of the estimator

In this section, we derive optimality conditions for the solution that minimizes the functional (1). Let $\omega_n(t) = n^{-1}\sum_{i=1}^{n} I(t_i \leqslant t)$, where I is the indicator function, and let ω be a distribution function with a continuous and strictly positive density function q on $[0, 1]$. For a function g, define its norm by $\|g\| = \sup_{t\in[0,1]} |g(t)|$. Let $D_n = \|\omega_n - \omega\|$. If the design points t_i are equally spaced, then $D_n = O(n^{-1})$ with $q(t) = 1$ for $t \in [0, 1]$. If the t_i are independent and identically distributed regressors from a distribution with bounded positive density q, then $D_n = O\{n^{-1/2}(\log\log n)^{1/2}\}$ by the law of the iterated logarithm for empirical distribution functions.

Let h be a piecewise constant function such that $h(t_i) = y_i$ $(i = 1, \ldots, n)$. For any $t \in [0, 1]$ and $f \in L_1[0, 1]$, define

$$l_1(f, t) = \int_0^t \sigma^{-2}(s) f(s)\, \mathrm{d}\omega(s), \quad l_k(f, t) = \int_0^t l_{k-1}(f, s)\, \mathrm{d}s \quad (2 \leqslant k \leqslant m)$$

and

$$\check{l}_1(f, t) = \int_0^t \sigma^{-2}(s) f(s)\, \mathrm{d}\omega_n(s), \quad \check{l}_k(f, t) = \int_0^t \check{l}_{k-1}(f, s)\, \mathrm{d}s \quad (2 \leqslant k \leqslant m).$$

Theorem 1. *Necessary and sufficient conditions for $\hat{f} \in W_2^m$ to minimize ψ in (1) are that*

$$(-1)^m \lambda\, \rho(t) \hat{f}^{(m)}(t) + \check{l}_m(\hat{f}, t) = \check{l}_m(h, t) \quad (t \in [0, 1]) \tag{2}$$

almost everywhere and that

$$\check{l}_k(\hat{f}, 1) = \check{l}_k(h, 1) \quad (k = 1, \ldots, m). \tag{3}$$

Both $\check{l}_1(\hat{f}, t)$ and $\check{l}_1(h, t)$ are piecewise constant in t. Therefore $\check{l}_m(h, t) - \check{l}_m(\hat{f}, t)$ is a piecewise $(m-1)$th-order polynomial. Thus, Theorem 1 shows that $\rho(t)\hat{f}^{(m)}(t)$ is a piecewise $(m-1)$th-order polynomial. The exact form of $\hat{f}$ will depend on additional assumptions about $\rho(t)$. For example, Pintore et al. (2006) assumed $\rho(t)$ to be piecewise constant with possible jumps at a subset of the design points; then, the optimal solution is a polynomial spline of order $2m$. It is well known that the traditional smoothing spline is a natural spline of order $2m$, which corresponds to the case where $\rho(t) \equiv 1$.

3. Asymptotic properties of the estimator

We establish an equivalent kernel and an asymptotic distribution of the spatially adaptive smoothing splines at interior points using a two-point boundary value problem technique. The key idea is to represent the solution to (2) by using a Green's function. It will be shown that the adaptive smoothing spline estimator can be approximated by a kernel estimator using this Green's function.

Let $R_k(t) = l_k(\hat{f}, t) - \check{l}_k(\hat{f}, t)$ $(k = 1, \ldots, m)$. Specifically, when $k = m$, it follows from Theorem 1 that

$$R_m(t) = (-1)^m \lambda \rho(t) \hat{f}^{(m)}(t) + l_m(\hat{f}, t) - l_m(h, t).$$

Write $r(t) = \sigma^2(t)/q(t)$. Then $l_m(\hat{f}, t)$ solves the two-point boundary value problem

$$(-1)^m \lambda \rho(t) \frac{\mathrm{d}^m}{\mathrm{d}t^m} \left\{ r(t) \frac{\mathrm{d}^m}{\mathrm{d}t^m} l_m(\hat{f}, t) \right\} + l_m(\hat{f}, t) = \check{l}_m(h, t) + R_m(t), \tag{4}$$

subject to the $2m$ boundary conditions from (3):

$$l_k(\hat{f}, 0) = 0, \quad l_k(\hat{f}, 1) = l_k(h, 1) + R_k(1) \quad (k = 1, \ldots, m). \tag{5}$$

The solution to (4) can be obtained explicitly with the aid of the Green's function. For readers unfamiliar with Green's functions, operationally speaking, if $P(t, s)$ is the Green's function for

$$(-1)^m \lambda \rho(t) \{r(t) u^{(m)}(t)\}^{(m)} + u(t) = 0,$$

then $\int_0^1 P(t, s) \{\check{l}_m(h, s) + R_m(s)\} \,\mathrm{d}s$ will solve (4). This, taken together with the boundary conditions (5), yields the solution to the two-point boundary value problem (4)–(5). The derivation of the Green's function and discussion of the boundary conditions are given in the Supplementary Material. Specifically, let $\{C_k(t) : k = 1, \ldots, 2m\}$ be $2m$ linearly independent solutions to the homogeneous differential equation

$$(-1)^m \lambda \rho(t) \frac{\mathrm{d}^m}{\mathrm{d}t^m} \left\{ r(t) \frac{\mathrm{d}^m}{\mathrm{d}t^m} l_m(\hat{f}, t) \right\} + l_m(\hat{f}, t) = 0.$$

Then, $l_m(\hat{f}, t)$ in (4) can be expressed as

$$l_m(\hat{f}, t) = \int_0^1 P(t, s) \check{l}_m(h, s) \,\mathrm{d}s + \int_0^1 P(t, s) R_m(s) \,\mathrm{d}s + \sum_{k=1}^{2m} a_k C_k(t), \tag{6}$$

where the last term is due to the boundary conditions and the coefficients a_k $(k = 1, \ldots, 2m)$ can be shown to be unique and stochastically bounded for all sufficiently small λ; see the Supplementary Material. The expression (6) can be decomposed into three parts: the asymptotic mean $\int_0^1 P(t, s) l_m(f_0, s) \,\mathrm{d}s$, the random component $\int_0^1 P(t, s) \check{l}_m(h - f_0, s) \,\mathrm{d}s$, and the remainder term $\Gamma(t) = \sum_{k=1}^{2m} a_k C_k(t) + \int_0^1 P(t, s) \tilde{R}_m(s) \,\mathrm{d}s$, where $\tilde{R}_m(t) = l_m(\hat{f} - f_0, t) - \check{l}_m(\hat{f} - f_0, t)$. It will be shown that $\|\tilde{R}_m\|$ has a smaller order and that the remainder term is negligible in the asymptotic analysis. The crucial representation of the adaptive smoothing spline estimator is obtained by taking the mth derivative pointwise on both sides of (6); this gives

$$r^{-1}(t) \hat{f}(t) = \frac{\mathrm{d}^m}{\mathrm{d}t^m} \int_0^1 P(t, s) l_m(f_0, s) \,\mathrm{d}s + \frac{\mathrm{d}^m}{\mathrm{d}t^m} \int_0^1 P(t, s) \check{l}_m(h - f_0, s) \,\mathrm{d}s + \Gamma^{(m)}(t). \tag{7}$$

We now introduce the main assumptions of this paper.

Assumption 1. The functions $\rho(\cdot)$, $q(\cdot)$ and $\sigma(\cdot)$ are $m+1$ times continuously differentiable and strictly positive.

Assumption 2. The function f_0 is $2m$ times continuously differentiable.

Assumption 3. The smoothing parameter λ satisfies $\lambda \to 0$ as $n \to \infty$. Write

$$\Delta_n = D_n n^{-1/2} \lambda^{-(1+m)/(2m)} \max\left[\{\log(1/\lambda)\}^{1/2},\ (\log\log n)^{1/2}\right]$$

and assume that $\Delta_n \to 0$ as $n \to \infty$.

Assumption 4. The random errors ϵ_i have a finite fourth moment.

Assumption 3 ensures that the smoothing parameter λ tends to zero but not too quickly. In particular, it encompasses the case of equally spaced design variables and the case of independent and identically distributed regressors from a distribution with bounded positive density. In the former case we have $D_n = O(n^{-1})$, and in the latter case $D_n = O\{n^{-1/2}(\log\log n)^{1/2}\}$. The optimal choice of λ discussed below is of order $n^{-2m/(4m+1)}$, and it is easy to check that it satisfies Assumption 3.

THEOREM 2. *Suppose that Assumptions 1–4 hold. Let $\beta = \lambda^{-1/(2m)}$. For any given $t \in (0, 1)$, the adaptive smoothing spline estimator $\hat{f}$ can be written as*

$$\hat{f}(t) = f_0(t) + \lambda(-1)^{m-1} r(t)\{\rho(t) f_0^{(m)}(t)\}^{(m)} + o(\lambda) + \frac{1}{n}\sum_{i=1}^{n} \frac{\sigma(t_i)}{q(t_i)} J(t, t_i)\epsilon_i$$
$$+ O(\beta^m)\Delta_n + O(\beta^m)\exp\{-\beta O(1)\}$$

uniformly in λ, where $J(t, s)$ is as given in (8).

Remark 1. Eggermont & LaRiccia (2006) were the first to show in full generality that standard spline smoothing corresponds approximately to smoothing by a kernel method. A simple explicit formula for the equivalent kernel for all m, denoted by $K(t, s)$, was given by Berlinet & Thomas-Agnan (2004). For interior points, the kernel K is of the form $K(t, s) = \beta L(\beta|t - s|)$ for some function L, such that $L(|\cdot|)$ is a $2m$th-order kernel on $(-\infty, \infty)$. In particular, the shape of $K(t, \cdot)$ is determined by $L(\cdot)$ and is the same for different t. For example, the closed-form expressions for the first equivalent kernels are

$$m = 1:\quad L(|t|) = \frac{1}{2}\exp(-|t|),$$
$$m = 2:\quad L(|t|) = \frac{1}{2^{3/2}}\exp\left(-\frac{|t|}{2^{1/2}}\right)\left\{\cos\left(\frac{|t|}{2^{1/2}}\right) + \sin\left(\frac{|t|}{2^{1/2}}\right)\right\},$$
$$m = 3:\quad L(|t|) = \frac{1}{6}\exp(-|t|) + \exp\left(-\frac{|t|}{2}\right)\left\{\frac{1}{6}\cos\left(\frac{3^{1/2}|t|}{2}\right) + \frac{3^{1/2}}{6}\sin\left(\frac{3^{1/2}|t|}{2}\right)\right\},$$
$$m = 4:\quad L(|t|) = \exp(-0{\cdot}9239|t|)\{0{\cdot}2310\cos(0{\cdot}3827|t|) + 0{\cdot}0957\sin(0{\cdot}3827|t|)\}$$
$$+ \exp(-0{\cdot}3827|t|)\{0{\cdot}0957\cos(0{\cdot}9239|t|) + 0{\cdot}2310\sin(0{\cdot}9239|t|)\}.$$

Theorem 2 indicates that the spatially adaptive smoothing spline estimator is also approximately a kernel regression estimator. The equivalent kernel $J(t, s)$ is the corresponding Green's function. It is shown in the Supplementary Material that

$$J(t, s) = \beta\varrho(s)Q'_{\beta}(s)L\{\beta|Q_{\beta}(t) - Q_{\beta}(s)|\}, \tag{8}$$

where

$$Q_{\beta}(t) = \int_0^t \{r(s)\rho(s)\}^{-1/(2m)}\{1 + O(\beta^{-1})\}\,\mathrm{d}s$$

is an increasing function of t and $\|\varrho\| = 1 + O(\beta^{-1})$. This shows that the shape of $J(t, \cdot)$ varies with t. Our estimator possesses the interpretation of spatial adaptivity (Donoho & Johnstone, 1998); it is asymptotically equivalent to a kernel estimator with a kernel that varies in shape and bandwidth from point to point.

Remark 2. The number β^{-1} in (8) plays a role similar to the bandwidth h in kernel smoothing. Theorem 2 shows that the asymptotic mean has bias $(-1)^{m-1}\lambda r(t)\{\rho(t)f_0^{(m)}(t)\}^{(m)}$, which can be made negligible if λ is taken to be reasonably small. On the other hand, λ cannot be arbitrarily small, since that would inflate the random component. The admissible range for λ results from a compromise between these two opposing influences.

COROLLARY 1. *Given $\rho(\cdot)$ and $r(\cdot)$, and assuming Assumptions 1–4, if $\lambda = n^{-2m/(4m+1)}$, then for any $t \in (0, 1)$, $n^{2m/(4m+1)}\{\hat{f}(t) - f_0(t)\}$ converges to*

$$N\left[(-1)^{m-1}r(t)\{\rho(t)f_0^{(m)}(t)\}^{(m)},\ L_0\, r(t)^{1-1/(2m)}\rho(t)^{-1/(2m)}\right]$$

in distribution, where $L_0 = \int_{-\infty}^{\infty} L^2(|t|)\,\mathrm{d}t$.

The proof of Corollary 1 is given in the Supplementary Material. The asymptotic mean squared error of the spatially adaptive smoothing spline estimator is of order $n^{-4m/(4m+1)}$, which is the optimal rate of convergence given in Stone (1982).

4. OPTIMAL SELECTION OF ρ

The optimal λ and ρ are chosen to minimize the integrated asymptotic mean squared error

$$\int_0^1 \left(\lambda^2 r^2(t)\left[\{\rho(t)f_0^{(m)}(t)\}^{(m)}\right]^2 + \frac{L_0}{n\lambda^{1/(2m)}}r(t)^{1-1/(2m)}\rho(t)^{-1/(2m)}\right)\mathrm{d}t,$$

which is in fact a function of $\lambda\rho(t)$. We choose the optimal λ to be $\lambda^{\text{opt}} = n^{-2m/(4m+1)}$. The optimal roughness penalty function $\rho(t)$ minimizes the functional

$$\Pi(\rho) = \int_0^1 \left(r^2(t)\left[\{\rho(t)f_0^{(m)}(t)\}^{(m)}\right]^2 + L_0 r(t)^{1-1/(2m)}\rho(t)^{-1/(2m)}\right)\mathrm{d}t. \tag{9}$$

Without any further assumptions, the above minimization problem does not have an optimal solution, since any arbitrarily large and positive function ρ with $\{\rho(t)f_0^{(m)}(t)\}^{(m)} = 0$ on any subinterval of $[0, 1]$ will make $\Pi(\cdot)$ arbitrarily small. To deal with this problem, we first impose a technical assumption on f_0.

Assumption 5. The set $\mathcal{N} = \{t \in [0, 1] : f_0^{(m)}(t) = 0\}$ has zero measure.

Let $u(t) = \{\rho(t) f_0^{(m)}(t)\}^{(m)}$ and $z(t) = \rho(t) f_0^{(m)}(t)$, and let D^{-m} be the m-fold integral operator. Then $z^{(m)}(t) = u(t)$ and

$$z(t) = (D^{-m}u)(t) + \theta^{\mathrm{T}}(t)x_0, \tag{10}$$

for $\theta(t) = (1, t, t^2/2!, \ldots, t^{m-1}/(m-1)!)^{\mathrm{T}}$ and some $x_0 \in \mathbb{R}^m$. Moreover, we can define $z(t)/f_0^{(m)}(t)$ to be any positive constant for all $t \in \mathcal{N}$ at which $f_0^{(m)}(t) = 0$. This definition is assumed in what follows. Hence, the functional $\Pi(\rho)$ in (9) becomes

$$J(u, x_0) = \int_0^1 r^2(t)u^2(t)\,\mathrm{d}t + \int_0^1 L_0 r(t)^{1-1/(2m)} \left\{\frac{z(t)}{f_0^{(m)}(t)}\right\}^{-1/(2m)} \mathrm{d}t,$$

where $z(t)$ is defined by (u, x_0). We then introduce another technical assumption on $z(t)$ or, essentially, on ρ.

Assumption 6. There exist positive constants μ and ε such that $\|x_0\| \leqslant \mu$ and $z(t)/f_0^{(m)}(t) \geqslant \varepsilon$ for all t. Also, $\{z(t)/f_0^{(m)}(t)\}^{-1/(2m)}$ is Lebesgue integrable on $[0, 1]$.

Consider the following set in $L_2[0, 1] \times \mathbb{R}^m$:

$$\mathcal{P} = \Big\{(u, x_0) \in L_2[0, 1] \times \mathbb{R}^m : \|x_0\| \leqslant \mu,\ z(t)/f_0^{(m)}(t) \geqslant \varepsilon \text{ for all } t \in [0, 1], \text{ and}$$
$$\Big\{z(t)/f_0^{(m)}(t)\Big\}^{-1/(2m)} \text{ is Lebesgue integrable on } [0, 1]\Big\},$$

where $z(t)$ is as given in (10), dependent on (u, x_0). In the Supplementary Material we establish the following theorem, which says that the objective functional J attains a unique minimum in $\mathcal{P}$. Under Assumptions 5 and 6, the existence of an optimal solution is established. Then, since the objective functional J is strictly convex and the constraint set $\mathcal{P}$ is convex, the uniqueness of the optimal solution follows.

Theorem 3. *Under Assumptions 1, 2, 5 and 6, the optimization problem* $\inf_{(u,x_0)\in\mathcal{P}} J(u, x_0)$ *has a unique solution in* $\mathcal{P}$.

Remark 3. Given the optimal solution (u^*, x^*), $z_{(u^*,x^*)}(t)$ is bounded on $[0, 1]$ by virtue of its absolute continuity. The lower bound ε in Assumption 6 ensures that the optimal ρ is bounded below, away from zero. However, there is no guarantee that the optimal ρ is bounded above, due to the possibility of small values of $|f_0^{(m)}|$. To avoid this problem, one could impose an additional upper bound constraint in Assumption 6. The proof of existence and uniqueness would remain the same.

5. Implementation

Obtaining an explicit solution of (9) is difficult. Motivated by Pintore et al. (2006), we consider approximating ρ by a piecewise constant function such that $\rho(t) = \rho_j$ for $t \in (\tau_{j-1}, \tau_j]$, $j = 0, \ldots, S+1$. Here $\tau_0 = 0$, $\tau_{S+1} = 1$, and $0 < \tau_1 < \cdots < \tau_S < 1$ are interior adaptive smoothing knots whose selection will be described below. When the integral in (9) is taken ignoring the nondifferentiability at the jump points τ_j $(j = 1, \ldots, S)$, we obtain

$$\sum_{j=1}^{S+1} \left[\rho_j^2 \int_{\tau_{j-1}}^{\tau_j} r^2(t)\{f_0^{(2m)}(t)\}^2 \, \mathrm{d}t + \rho_j^{-1/(2m)} L_0 \int_{\tau_{j-1}}^{\tau_j} r(t)^{1-1/(2m)} \, \mathrm{d}t \right].$$

Therefore, the optimal ρ_j is

$$\rho_j = \left[\frac{L_0 \int_{\tau_{j-1}}^{\tau_j} r(t)^{1-1/(2m)} \, \mathrm{d}t}{4m \int_{\tau_{j-1}}^{\tau_j} r^2(t)\{f_0^{(2m)}(t)\}^2 \, \mathrm{d}t} \right]^{2m/(4m+1)} \qquad (j = 1, \ldots, S+1).$$

Unfortunately, the optimal values for the ρ_j depend on $r(t)$ and the $2m$th derivative of the underlying regression function $f_0(t)$. We replace them by estimates in practice.

Remark 4. Rigorously speaking, such a step-function approximation to ρ is not a valid solution of (9) due to nondifferentiability. However, simulations seem to suggest that such a simple approximation can yield good results. Furthermore, one can modify such a ρ, for instance, to make it satisfy Assumption 2. In a sufficiently small neighbourhood of each jump point, one can replace the steps on either side of the jump by a smooth curve connecting the two constants such that the resulting function satisfies Assumption 2. Hence the piecewise constant ρ can be viewed as a simple approximation to this smooth version of ρ.

We now describe the implementation steps. The first step is to select the interior smoothing knots τ_j $(j = 1, \ldots, S)$. An abrupt change in the smoothness of the function is often associated with a similar change in the conditional probability density of y given t. For example, a steeper part of the function often comes with sparser data, or smaller conditional probability densities of y given t. Hence, we first use the sscden function in the R package gss (R Development Core Team, 2013) to estimate the conditional probability densities of y given t on a dense grid, say $s_k = k/100$ for $k = 1, \ldots, 100$. Then, for a given S, we select the top S grid points s_k where the conditional probability density changes the most from s_k to s_{k+1}. A more accurate but considerably more time-consuming way of selecting the smoothing knots is by means of a binary tree search algorithm as proposed in Liu & Guo (2010).

Estimation of $\sigma^2(t)$ was first studied by Müller & Stadtmüller (1987). Here we use the local polynomial approach of Fan & Yao (1998); see Hall & Carroll (1989), Ruppert et al. (1997), and Cai & Wang (2008) for other methods. This provides the weights for obtaining a weighted smoothing spline estimate of $f(t)$, whose derivative yields an estimate of $f^{(2m)}(t)$. The function $q(t)$ can be replaced by an estimate of the density function of t_i $(i = 1, \ldots, n)$. All these computations can be conveniently carried out using the R packages locpol and gss.

Ideally, the optimal ρ_j computed as above would work well. However, similar to Storlie et al. (2010), we have found that a powered-up version ρ_j^{γ} for some $\gamma > 1$ often helps in practice. Intuitively, this power-up compensates a bit for the underestimated differences in $f^{(2m)}(t)$ across the predictor domain.

For the tuning parameters S and γ, we consider $S \in \{0, 2, 4, 8\}$ and $\gamma \in \{1, 2, 4\}$. Theoretically, a larger S might be preferred due to the better approximation of such step functions to the real function. However, as shown in Pintore et al. (2006) and Liu & Guo (2010), an S greater than 8 tends to overfit the data. The choices for γ are as suggested in Storlie et al. (2010). In traditional smoothing splines, smoothing parameters are selected by generalized crossvalidation

(Craven & Wahba, 1979) or the generalized maximum likelihood estimate (Wahba, 1985). As pointed out by Pintore et al. (2006), a proper criterion for selecting the piecewise constant $\rho(\cdot)$ should penalize on the number of segments of ρ. The generalized Akaike information criterion proposed in Liu & Guo (2010) serves this purpose; it is a penalized version of the generalized maximum likelihood estimate where S is penalized similarly to the degrees of freedom in the conventional Akaike information criterion. In this paper, we will use the generalized Akaike information criterion to select S and γ.

Once the piecewise constant penalty function ρ is determined, we compute the corresponding adaptive smoothing spline estimate as follows. By the representer theorem (Wahba, 1990), the minimizer of (1) lies in a finite-dimensional space of functions of the form

$$f(t) = \sum_{i=1}^{n} c_i K_\rho(t_i, t) + \sum_{j=0}^{m-1} d_j \phi_j(t), \tag{11}$$

where the c_i and d_j are unknown coefficients, $\phi_j(t) = t^j/j!$ for $j = 0, \ldots, m-1$, and K_ρ is the reproducing kernel function whose closed-form expressions at $(t_i, \cdot)$ with a piecewise constant ρ are given in § 2.2 of Pintore et al. (2006). Upon substituting (11) into (1), we solve for $c = (c_1, \ldots, c_n)^{\mathrm{T}}$ and $d = (d_0, \ldots, d_{m-1})^{\mathrm{T}}$ by the Newton–Raphson procedure with a fixed λ. Here λ can be selected by generalized crossvalidation or the generalized maximum likelihood estimate with the adaptive reproducing kernel function.

6. Simulations

This section compares the estimation performance of different smoothing spline methods. For traditional smoothing splines, we used the cubic smoothing splines from the function ssanova in the R package gss, and the smoothing parameter was selected by the generalized crossvalidation score. For the spatially adaptive smoothing splines of Pintore et al. (2006), we used an equally spaced five-step penalty function following their implementation, and the optimal penalty function was selected to minimize the generalized crossvalidation function (19) in Pintore et al. (2006). For the Loco-Spline of Storlie et al. (2010), we downloaded the authors' original program from the *Journal of Computational and Graphical Statistics* website. For the proposed adaptive smoothing splines, we used $m = 1$ and cubic smooth splines to compute the optimal ρ_j.

Two well-known functions with varying smoothness over the domain were considered under the model $y_i = f(t_i) + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$. We used $n = 200$ and $t_i = i/n$ $(i = 1, \ldots, n)$ in all the simulations and repeated each simulation on 100 randomly generated data replicates. The integrated squared error $\int_0^1 \{\hat{f}(t) - f_0(t)\}^2\,\mathrm{d}t$ and pointwise absolute errors at $t = 0{\cdot}2, 0{\cdot}4, 0{\cdot}6, 0{\cdot}8$ were used to evaluate the performance of an estimate $\hat{f}$. To visualize the comparison, we also selected for each example and each method a data replicate to represent the median performance as follows. The function estimates from each method yielded 100 integrated squared errors. Upon ranking these from lowest to highest, we chose the 50th integrated squared error and its corresponding data replicate to represent the median performance. We then plotted the function estimates from these selected data replicates in Fig. 1–2 to compare the median estimation performances of different methods. To assess the variability in estimation, we also superimposed on these plots the pointwise empirical 0·025 and 0·975 quantiles of the 100 estimates.

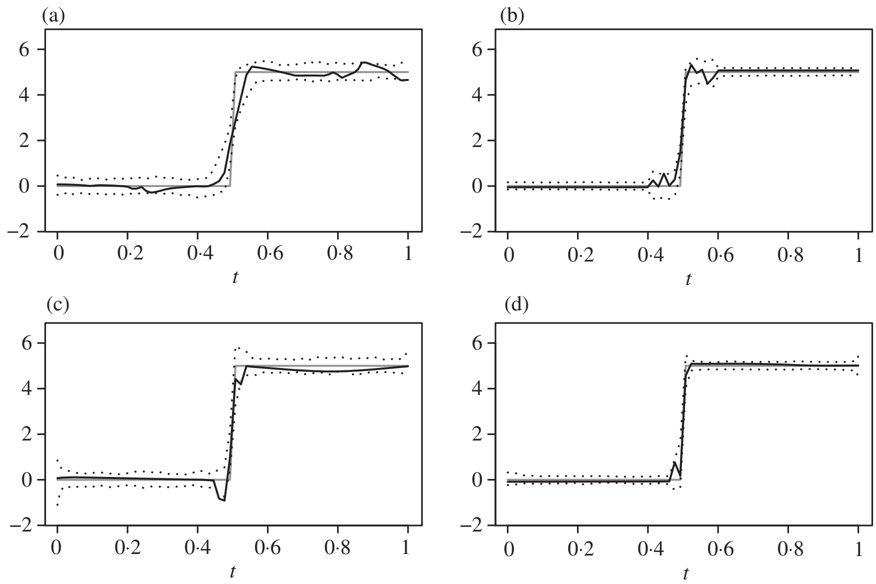

Fig. 1. Estimates of the Heaviside function produced by the data replicate with median integrated squared error, for four different methods: (a) traditional smoothing spline; (b) the method of Pintore et al. (2006); (c) the Loco-Spline of Storlie et al. (2010); (d) our proposed adaptive smoothing spline. The plotted curves are the true function (solid grey line), the spline estimate (solid black line), and the pointwise empirical 0·025 and 0·975 quantiles (dotted lines).

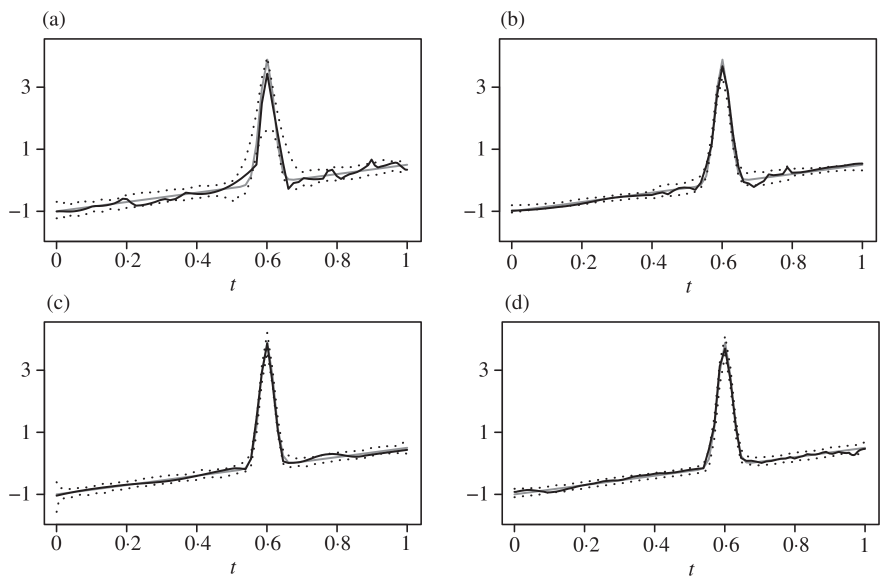

Fig. 2. Estimates of the Mexican hat function produced by the data replicate with median integrated squared error, for four different methods: (a) traditional smoothing spline; (b) the method of Pintore et al. (2006); (c) the Loco-Spline of Storlie et al. (2010); (d) our proposed adaptive smoothing spline. The plotted curves are the true function (solid grey line), the spline estimate (solid black line), and the pointwise empirical 0·025 and 0·975 quantiles (dotted lines).

Table 1. *Comparison of integrated squared errors and pointwise absolute errors for various estimates. Values are empirical means and standard deviations (in parentheses) multiplied by 100 based on 100 data replicates*

Method	ISE	PAE(0·2)	PAE(0·4)	PAE(0·6)	PAE(0·8)
			Heaviside function		
SS	18 (7)	15 (11)	17 (14)	16 (14)	16 (12)
PSH	5 (2)	6 (5)	6 (5)	7 (5)	7 (5)
Loco	7 (3)	10 (8)	13 (12)	11 (10)	12 (12)
ADSS	2 (2)	7 (5)	6 (5)	6 (5)	7 (6)
			Mexican hat function		
SS	6·6 (6·2)	8 (6)	8 (8)	96 (72)	8 (6)
PSH	1·1 (0·3)	4 (3)	8 (5)	35 (11)	8 (5)
Loco	0·6 (0·3)	4 (4)	5 (4)	13 (10)	5 (4)
ADSS	0·6 (0·2)	4 (3)	4 (3)	15 (10)	6 (4)

ISE, integrated squared error; PAE, pointwise absolute error; SS, smoothing splines; PSH, the splines of Pintore et al. (2006); Loco, the Loco-Splines of Storlie et al. (2010); ADSS, the adaptive smoothing splines proposed in this paper

We first consider data generated from the Heaviside function $f(t) = 5I(t \geqslant 0{\cdot}5)$ with $\sigma = 0{\cdot}7$. Based on the error summary statistics in Table 1, all the adaptive methods outperform the traditional smoothing splines, with our method and that of Pintore et al. (2006) displaying clear advantages in all the error measures. Furthermore, our method had the smallest mean integrated squared error, illustrated by the plots in Figs 1. While the median estimates from all three adaptive methods tracked the true function reasonably well, the Loco-Spline estimates showed greater variability in the flat parts of the Heaviside function than estimates obtained from the other two adaptive methods. Further, our method does the best job in tracking the jump. As shown in Fig. 1(b), the estimate of Pintore et al. (2006) oscillates around the jump of the Heaviside function, possibly because the equally spaced jump points for ρ suggested in their paper sometimes have difficulty characterizing the jump in the true function. This echoes the finding in Liu & Guo (2010) that the jump locations of ρ also need to be adaptive, an idea that is adopted in our method.

The second example is the Mexican hat function $f(t) = -1 + 1{\cdot}5t + 0{\cdot}2\phi_{0{\cdot}02}(t - 0{\cdot}6)$ with $\sigma = 0{\cdot}25$, where $\phi_{0{\cdot}02}(t - 0{\cdot}6)$ is the density function of $N(0{\cdot}6, 0{\cdot}02^2)$. From Table 1 and Fig. 2, the estimates from our method and the Loco-Spline have comparable performance, and both outperform the traditional smoothing spline and the method of Pintore et al. (2006). Again, the estimate of Pintore et al. (2006) shows oscillations around the places where the hat peak rises steeply from the brim.

For the estimates plotted in Figs. 1–2, we also plot in Fig. 3 the estimated log penalties. In general, the penalty functions from the three adaptive methods track the smoothness changes in the underlying functions reasonably well.

7. Application

In this section, we apply the proposed adaptive smoothing splines to data from electroencephalograms of epilepsy patients (Liu & Guo, 2010). Previous research (Qin et al., 2009) has shown that the 26–50 Hz frequency band is important in characterizing electroencephalograms and may help to determine the spatial-temporal initiation of seizures. Figure 4(a) shows the raw time-varying log spectral band power of 26–50 Hz calculated every half second for a 15-minute

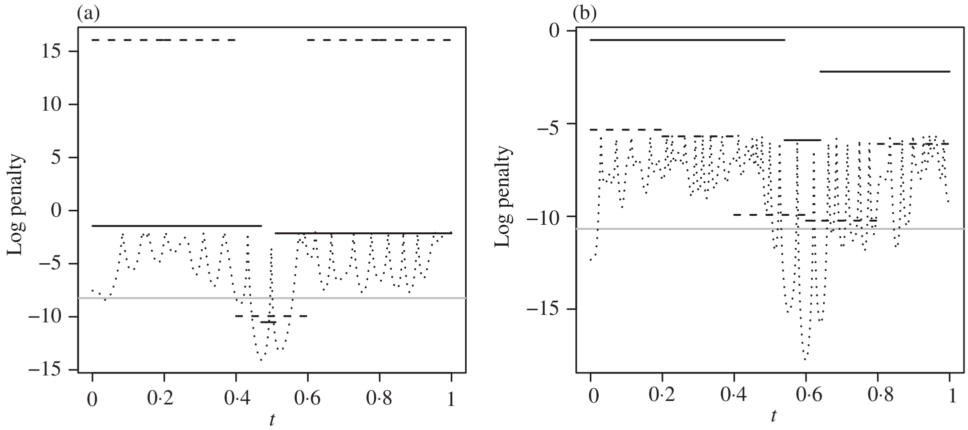

Fig. 3. Estimated log penalties for the estimates shown in Figs. 1–2: (a) Heaviside function; (b) Mexican hat function. The log penalties are for traditional smoothing splines (solid grey lines), the method of Pintore et al. (2006) (dashed steps), the Loco-Spline (dotted lines), and the proposed method (solid steps).

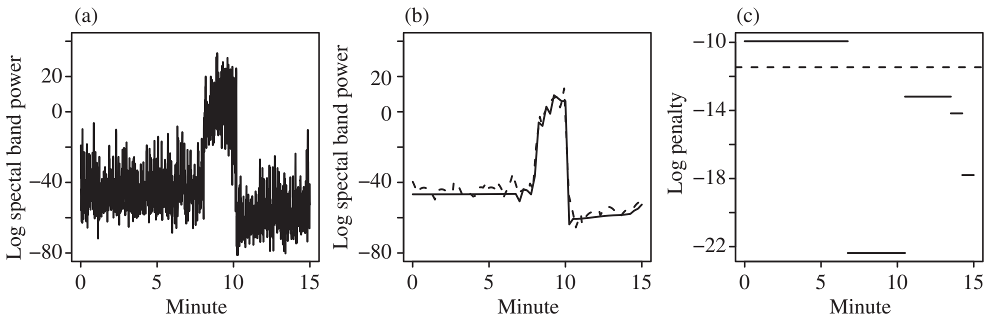

Fig. 4. EEG data example: (a) raw log spectral band power; (b) reconstructions obtained from the traditional smoothing splines (dashed line) and the proposed adaptive smoothing splines (solid line); (c) estimated log penalties for the traditional smoothing splines (dashed line) and the proposed adaptive smoothing splines (solid steps).

intracranial electroencephalogram series. The sampling rate was 200 Hz, and seizure onset occurred at the 8th minute (Litt et al., 2001). The raw band powers are always very noisy and need to be smoothed before further analysis. Figure 4(b) shows the reconstructions obtained from traditional smoothing splines and the proposed adaptive smoothing splines. We also tried the Loco-Spline, but the program exited due to a singular matrix error.

Traditional smoothing splines clearly undersmooth the pre- and post-seizure regions and oversmooth the seizure period, because a single smoothing parameter is insufficient to capture the abrupt change before the onset of the seizure. Our estimate smoothes out the noise at both ends but retains the details before the onset of seizure. In particular, we see a fluctuation in power starting from a minute or so before the onset of the seizure, which could be a meaningful predictor of seizure initiation. The band power then increases sharply at the beginning of the seizure. At the end of the seizure, around the 10th minute, the band power drops sharply to a level even lower than the pre-seizure level, indicating the suppression of neuronal activity after seizure. Afterwards, the band power starts to rise again, but it still fails to reach the pre-seizure level even at the end of the 15th minute. These findings agree with those of Liu & Guo (2010).

The proposed method took less than 10 minutes for the whole analysis, compared with 40–50 minutes for the method of Liu & Guo (2010). This is not surprising, since the latter not only needs a dense grid search to locate the jump points but also lacks good initial step sizes.

Acknowledgement

We are grateful to two referees and the associate editor for constructive and insightful comments. We also thank Wensheng Guo and Ziyue Liu for providing the electroencephalogram data, and Howard Bondell for help with the Loco-Spline program. Wang and Shen were supported by the U.S. National Science Foundation.

Supplementary material

Supplementary material available at *Biometrika* online includes the proofs of Theorems 1–3 and Corollary 1, and the detailed derivation of the Green's function.

Appendix

Here we outline the proofs of Theorems 1 and 2. For the full proofs of the three theorems and Corollary 1, we refer the reader to the Supplementary Material.

Outline of proof of Theorem 1. For any $f, g \in W_2^m$ and $\delta \in \mathbb{R}$,

$$\psi(f+\delta g) - \psi(f) = 2\delta\psi_1(f, g) + \delta^2 \left[\int_0^1 g^2(t)\, \mathrm{d}\omega_n(t) + \lambda \int_0^1 \rho(t)\{g^{(m)}(t)\}^2\, \mathrm{d}t \right],$$

where

$$\psi_1(f, g) = \int_0^1 \sigma^{-2}(t)\{f(t) - h(t)\} g(t)\, \mathrm{d}\omega_n(t) + \lambda \int_0^1 \rho(t) f^{(m)}(t) g^{(m)}(t)\, \mathrm{d}t. \tag{A1}$$

Lemma A1. *The function $f \in W_2^m$ minimizes $\psi(f)$ in (1) if and only if $\psi_1(f, g) = 0$ for all $g \in W_2^m$.*

Let $g(t) = t^k$ $(k = 0, \ldots, m-1)$ in (A1). By Lemma A1, if f minimizes $\psi(f)$, then

$$\int_0^1 \sigma^{-2}(t)\{f(t) - h(t)\}\, t^k\, \mathrm{d}\omega_n(t) = 0 \quad (k = 0, 1, \ldots, m-1).$$

First, we have

$$\check{l}_1(f, 1) - \check{l}_1(h, 1) = \int_0^1 \sigma^{-2}(t)\{f(t) - h(t)\}\, \mathrm{d}\omega_n(t) = 0.$$

Further,

$$\check{l}_2(f, 1) - \check{l}_2(h, 1) = \int_0^1 \int_0^s \sigma^{-2}(t)\{f(t) - h(t)\}\, \mathrm{d}\omega_n(t)\, \mathrm{d}s = \int_0^1 \sigma^{-2}(t)\{f(t) - h(t)\}\, \mathrm{d}\omega_n(t) = 0.$$

Similarly, $\check{l}_k(f, 1) = \check{l}_k(h, 1)$ for $k = 1, \ldots, m$.

Lemma A2. *If $f \in W_2^m$ satisfies $\check{l}_k(f, 1) = \check{l}_k(h, 1)$ for $k = 1, \ldots, m$, then for all $g \in W_2^m$,*

$$\psi_1(f, g) = \int_0^1 \psi_2(f)\, g^{(m)}(t)\, \mathrm{d}t,$$

where

$$\psi_2(f) = \lambda\, \rho(t) f^{(m)}(t) + (-1)^m \{\check{l}_m(f, t) - \check{l}_m(h, t)\}.$$

Let $B^+ = \{t \in [0, 1] : \psi_2(f) > 0\}$ and $B^- = \{t \in [0, 1] : \psi_2(f) < 0\}$. Define $g_+^{(m)}(t) = -I_{B^+}(t)$ and $g_-^{(m)}(t) = I_{B^-}(t)$, where I denotes the indicator function. Since $\psi_1(f, g) = 0$ for all $g \in W_2^m$, we have $\psi_1(f, g_+) < 0$ and $\psi_1(f, g_-) < 0$ unless B^+ and B^- have measure zero. This shows that $\psi_2(f) = 0$ almost everywhere.

Outline of proof of Theorem 2. It follows from (7) that $r^{-1}(t)\hat{f}(t) = V_1(t) + V_2(t) + V_3(t) + V_4(t)$, where

$$V_1(t) = \frac{\mathrm{d}^m}{\mathrm{d}t^m}\int_0^1 P(t, s) l_m(f_0, s)\,\mathrm{d}s, \quad V_2(t) = \frac{\mathrm{d}^m}{\mathrm{d}t^m}\int_0^1 P(t, s)\{\check{l}_m(h, s) - \check{l}_m(f_0, s)\}\,\mathrm{d}s,$$

$$V_3(t) = \frac{\mathrm{d}^m}{\mathrm{d}t^m}\int_0^1 P(t, s)\{l_m(\hat{f} - f_0, s) - \check{l}_m(\hat{f} - f_0, s)\}\,\mathrm{d}s, \quad V_4(t) = \sum_{k=1}^{2m} a_k C_k^{(m)}(t).$$

Let $\bar{f}$ minimize the functional

$$\int_0^1 r^{-1}(s)\{f(s) - f_0(s)\}^2\,\mathrm{d}s + \lambda\int_0^1 \rho(t) f^{(m)}(s)^2\,\mathrm{d}s.$$

Similar to Theorem 1, we have

$$(-1)^m\lambda\rho(t)\bar{f}^{(m)}(t) + l_m(\bar{f}, t) = l_m(f_0, t) \tag{A2}$$

and

$$l_m(\bar{f}, t) = \int_0^1 P(t, s) l_m(f_0, s)\,\mathrm{d}s. \tag{A3}$$

Hence $V_1(t) = r^{-1}(t)\bar{f}(t)$. Taking the mth derivative of both sides of (A2), we get

$$(-1)^m\lambda\{\rho(t)\bar{f}^{(m)}(t)\}^{(m)} + r^{-1}(t)\bar{f}(t) = r^{-1}(t) f_0(t).$$

Recall that f_0 is $2m$ times continuously differentiable and $\beta = \lambda^{-1/(2m)}$. By combining this with (A3), it is easy to show that $\bar{f}^{(k)}(t) \to f_0^{(k)}(t)$ as $\beta \to \infty$ for $k = 1, \ldots, 2m$. Therefore

$$V_1(t) = r^{-1}(t) f_0(t) + (-1)^{m-1}\lambda\{\rho(t) f_0^{(m)}(t)\}^{(m)} + o(\lambda).$$

PROPOSITION A1. *Assume that a function* $\tilde{J}(t, s)$ *satisfies* $(-1)^m \frac{\partial^m}{\partial s^m}\tilde{J}(t, s) = \frac{\partial^m}{\partial t^m} P(t, s)$ *for* $t, s \in [0, 1]$. *Then* $\tilde{J}(t, s) + \sum_{k=0}^{m-1}(-1)^k \zeta_{k+1}(s)\tilde{J}_k(t) = \{r(s)/r(t)\}J(t, s)$, *where*

$$\zeta_k(s) = \int_s^1 \cdots \int_{s_{k-3}}^1 \int_{s_{k-2}}^1 \mathrm{d}s_{k-1}\mathrm{d}s_{k-2}\cdots \mathrm{d}s_1, \quad \tilde{J}_k(t) = \frac{\partial^k}{\partial s^k}\tilde{J}(t, s)\Big|_{s=1},$$

and $J(t, s)$ *is the Green's function for*

$$(-1)^m\lambda r(t)\{\rho(t) u^{(m)}(t)\}^{(m)} + u(t) = 0.$$

By Proposition A1, for any $t \in (0, 1)$ we have

$$\begin{aligned} V_2(t) &= \int_0^1 (-1)^m \frac{\partial^m}{\partial s^m}\tilde{J}(t, s)\check{l}_m(h - f_0, s)\,\mathrm{d}s \\ &= \int_0^1 \tilde{J}(t, s)\,\mathrm{d}\{\check{l}_1(h - f_0, s)\} + (-1)^m\sum_{k=1}^{m-1}(-1)^{k-1}\tilde{J}_{m-k}(t)\check{l}_{m-k+1}(h - f_0, 1) \\ &= \frac{1}{n}\sum_{i=1}^n \frac{r(t_i)}{r(t)} J(t, t_i)\sigma^{-1}(t_i)\epsilon_i + \text{ higher order terms.} \end{aligned}$$

Eggermont & LaRiccia (2006) established uniform error bounds for regular smoothing splines. We adopt the same approach for adaptive smoothing splines; the details are omitted here. For $\lambda \ll (n^{-1}\log n)^{2m/(1+4m)}$, we obtain

$$\|\hat{f} - f_0\| = O\left(\left[\frac{\max\{\log(1/\lambda), \log\log n\}}{n\lambda^{1/(2m)}}\right]^{1/2}\right).$$

Therefore, $\|V_3\| \leqslant O(\beta^m)D_n\|\hat{f} - f_0\|$. Finally, it is shown in detail in the Supplementary Material that $\|V_4\|$ is of order $O(\beta^m)\exp[-\beta Q_\beta(t)\{Q_\beta(1) - Q_\beta(t)\}]$ and is thus a negligible term in the asymptotic expansion of $r^{-1}(t)\hat{f}(t)$. This completes the representation for $\hat{f}$.

References

Abramovich, F. & Grinshtein, V. (1999). Derivation of an equivalent kernel for general spline smoothing: A systematic approach. *Bernoulli* **5**, 359–79.

Berlinet, A. & Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. New York: Springer.

Cai, T. & Wang, L. (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression. *Ann. Statist.* **36**, 2025–54.

Cox, D. D. (1983). Asymptotics of M-type smoothing splines. *Ann. Statist.* **11**, 530–51.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377–403.

DiMatteo, I., Genovese, C. R. & Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88**, 1055–71.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–55.

Donoho, D. L. & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Assoc.* **90**, 1200–24.

Donoho, D. L. & Johnstone, I. M.(1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879–921.

Eggermont, P. P. B. & LaRiccia, V. N. (2006). Uniform error bounds for smoothing splines. *IMS Lecture Notes Monogr. Ser. High Dim. Prob.* **51**, 220–37.

Eggermont, P. P. B. & LaRiccia, V. N. (2009). *Maximum Penalized Likelihood Estimation. Volume II: Regression.* New York: Springer.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Boca Raton, Florida: Chapman & Hall.

Fan, J. & Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645–60.

Friedman, J. & Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with Discussion). *Technometrics* **31**, 3–39.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer.

Hall, P. & Carroll, R. J.(1989). Variance function estimation in regression: The effect of estimating the mean. *J. R. Statist. Soc.* B **51**, 3–14.

Hansen, M. H. & Kooperberg, C. (2002). Spline adaptation in extended linear models. *Statist. Sci.* **17**, 2–51.

Kimeldorf, G. S. & Wahba, G.(1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33**, 82–95.

Kohn, R. & Ansley, C. F. (1983). On the smoothness properties of the best linear unbiased estimate of a stochastic process observed with noise. *Ann. Statist.* **11**, 1011–17.

Litt, B., Esteller R., Echauz, J., D'Alessandro, M., Shor, R., Henry, T., Pennell, P., Epstein, C., Bakay, R., Dichter, M. & Vachtsevanose, G. (2001). Epileptic seizures may begin hours in advance of clinical onset: A report of five patients. *Neuron* **30**, 51–64.

Liu, Z. & Guo, W. (2010). Data driven adaptive spline smoothing. *Statist. Sinica* **20**, 1143–63.

Luo, Z. & Wahba, G. (1997). Hybrid adaptive splines. *J. Am. Statist. Assoc.* **92**, 107–16.

Mao, W. & Zhao, L. (2003). Free-knot polynomial splines with confidence intervals. *J. R. Statist. Soc.* B **65**, 901–19.

Messer, K. (1991). A comparison of a spline estimate to its equivalent kernel estimate. *Ann. Statist.* **19**, 817–29.

Müller, H.-G. & Stadtmüller, U. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.* **15**, 282–301.

Nychka, D. (1995). Splines as local smoothers. *Ann. Statist.* **23**, 1175–97.

Pintore, A., Speckman, P. & Holmes, C. C. (2006). Spatially adaptive smoothing splines. *Biometrika* **93**, 113–25.

Qin, L., Guo, W. & Litt, B. (2009). A time-frequency functional model for locally stationary time series data. *J. Comp. Graph. Statist.*, **18**, 675–93.

R DEVELOPMENT CORE TEAM (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

RICE, J. & ROSENBLATT, M. (1983). Smoothing splines: Regression, derivatives and deconvolution. *Ann. Statist.* **11**, 141–56.

RUPPERT, D. & CARROLL, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Aust. New Zeal. J. Statist.* **42**, 205–23.

RUPPERT, D., WAND, M. P., HOLST, U. & HÖSJER, O. (1997). Local polynomial variance function estimation. *Technometrics* **39**, 262–73.

SILVERMAN, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Ann. Statist.* **12**, 898–916.

SMITH, M. & KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Economet.* **75**, 317–44.

STONE, C. J. (1982). Optimal rate of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–53.

STONE, C. J., HANSEN, M., KOOPERBERG, C. & TRUONG, Y. K. (1997). Polynomial splines and their tensor products in extended linear models. *Ann. Statist.* **25**, 1371–425.

STORLIE, C. B., BONDELL, H. D. & REICH, B. J. (2010). A locally adaptive penalty for estimation of functions with varying roughness. *J. Comp. Graph. Statist.* **19**, 569–89.

WAHBA, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13**, 1378–402.

WAHBA, G. (1990). *Spline Models for Observation Data*. Philadelphia: Society for Industrial and Applied Mathematics.

WAHBA, G. (1995). Discussion of ‘Wavelet shrinkage: asymptopia?’ by D. L. Donoho, I. M. Johnstone, G. Kerkyacharian & D. Picard. *J. R. Statist. Soc.* B **57**, 360–1.

WANG, X., SHEN, J. & RUPPERT, D. (2010). Local asymptotics of P-splines. *Electron. J. Statist.* **5**, 1–17.

WOOD, S. A., JIANG, W. & TANNER, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* **89**, 513–21.

[*Received January* 2011. *Revised April* 2013]