Journal of Statistical
Theory & Methods
1998, Vol. 9, No. 1, pp. 77 ~ 88

Boundary Corrected Smoothing Splines [1]

Jong Tae Kim [2]

Abstract

Smoothing spline estimators are modified to remove boundary bias effects using the technique proposed in Eubank and Speckman (1991). An $O(n)$ algorithm is developed for the computation of the resulting estimator as well as associated generalized cross-validation criteria, etc. The asymptotic properties of the estimator are studied for the case of a linear smoothing spline and the upper bound for the average mean squared error of the estimator given in Eubank and Speckman (1991) is shown to be asymptotically sharp in this case.

Key Words and Phrases: Smoothing spline, boundary correction.

1. Introduction

Smoothing splines provide a popular tool for nonparametric regression. However, these estimators are known to have certain "boundary bias" problems that result from the rather peculiar way they handle estimation in the boundary regions. Methods for removing these boundary effects have been proposed in Eubank and Speckman (1989). In this paper we further explore the properties of the Eubank/Speckman boundary correction. In particular, we develop a $O(n)$ algorithm for computing a boundary corrected smoothing spline and other related quantities such as the generalized cross-validation criterion. We also examine the asymptotic properties of the boundary corrected estimator in the special instance of a linear smoothing spline.

Consider now the *nonparametric regression* problem where responses $y_1, \ldots, y_n$ are obtained at non-coincident design points $t_1, \ldots, t_n$ from the model

$$y_i = \mu(t_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

[1] This paper was supported by research fund, Taegu University, 1997

[2] Associate Professor, Department of Statistics, Taegu University, Kyungpook, 712-714

Here μ is an unknown regression function and $\epsilon, \ldots, \epsilon_n$ are zero mean uncorrelated random errors with common variance σ^2.

There are a number of estimators that can be used for μ in (1). Our interest is in the mth order smoothing spline μ_λ that is obtained by minimizing

$$n^{-1}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda\int_{t_1}^{t_n} f^{(m)}(t)^2 dt, \quad \lambda > 0, \tag{2}$$

over all functions f with $(m-1)$ absolutely continuous derivatives and a square integrable mth derivative. If $n \geq m$, (2) has a unique minimizer that is a natural spline of order $2m$ with knots at the design points (See, e.g., Wahba 1990). The quantity λ in (2) is called the smoothing parameter and it controls the level of smoothing that the estimator performs on the data. The value of λ can be selected using data-driven techniques such as generalized cross-validation that will be discussed in more detail subsequently.

Suppose that we assess the performance of μ_λ through its average mean squared error or risk

$$\mathrm{R}_n(\lambda) = n^{-1}\sum_{i=1}^{n}\mathrm{E}\ (\mu_\lambda(t_i) - \mu(t_i))^2. \tag{3}$$

Then it is known (Rice and Rosenblatt 1983, Eubank 1988, Chpt. 6) that

$$\inf_\lambda \mathrm{R}_n(\lambda) = O(n^{-\frac{2m}{2m+1}}), \tag{4}$$

i.e., μ_λ provides an mth order estimator of μ. However, this rate can improve substantially and we can have

$$\inf_\lambda \mathrm{R}_n(\lambda) = O(n^{-\frac{4m}{4m+1}}) \tag{5}$$

if μ has $2m$ derivatives and satisfies the natural boundary coditions

$$\mu^{(m+j-1)}(0) = \mu^{(m+j-1)}(1) = 0, \quad j = 1, \ldots, m. \tag{6}$$

This states that μ_λ actually provides a $2m$th order estimator if μ is sufficiently smooth and satisfies (6).

In practice we will not generally know how many derivatives μ might possess. Thus (5) is a good quality in that it suggests that μ_λ may be able to utilize extra, possibly unexpected, smoothness in μ to produce a more efficient estimator. However, the need for condition (6) makes the actual utility of this result circumspect since it is unlikely that one would be fortunate enough to have the underlying regression function satisfy the natural boundary conditions in practice.

Eubank and Speckman (1989) and Oehlert (1992) have developed methods of altering the smoothing spline estimator so that (5) holds regardless of whether or not

(6) is true. Oehlert's approach is to modify the smoothness criterion $\int_{t_1}^{t_n} f^{(m)}(t)^2dt$ in a way that removes the boundary effects. This approach is quite effective but the resulting estimator appears to be somewhat difficult to compute. Thus, we focus insted on the Eubank and Speckman (1989) method for boundary adjustment.

To describe the Eubank/Speckman approach let $\mathbf{S}_\lambda$ be the $n \times n$ matrix which transformations the response vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ to the vector of fitted values for the smoothing spline estimator, i.e.,

$$\mu_\lambda = (\mu_\lambda(t_1), \ldots, \mu(t_n))^T = \mathbf{S}_\lambda \mathbf{y}. \tag{7}$$

Now define $3m$th order polynomials (q_{0i}, q_{1i}), $i = 1, \ldots, m$ such that for $k = 1, \ldots, m$

$$q_{0i}^{(m+k-1)}(0) = \delta_{ik}, \quad q_{0i}^{(m+k-1)}(1) = 0, \tag{8}$$

$$q_{1i}^{(m+k-1)}(0) = 0, \quad q_{1i}^{(m+k-1)}(1) = \delta_{ik}, \tag{9}$$

and set $q_{ij} = (q_{ij}(t_1), \ldots, q_{ij}(t_n))^T$ for $i = 1, \ldots, m$ and $j = 1, 2$. We then take

$$\mathbf{Q} = [q_{01}, q_{11}, q_{02}, q_{12}, \ldots, q_{0m}, q_{1m}]$$

and define the boundary adjucted estimator to be

$$\tilde{\mu}_\lambda = \mu_\lambda + \tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T(\mathbf{y} - \mu_\lambda) \tag{10}$$

for

$$\tilde{\mathbf{Q}} = (\mathbb{I} - \mathbf{S}_\lambda)\mathbf{Q} \tag{11}$$

Eubank and Speckaman (1989) give an upper bound on the risk for μ_λ which ensures that if μ has $4m$ derivatives then

$$\inf_\lambda n^{-1}\mathrm{E}\,(\mu - \tilde{\mu}_\lambda)^T(\mu - \tilde{\mu}_\lambda) = O(n^{-\frac{4m}{4m+1}}), \tag{12}$$

even if (6) does not hold.

It follows from Eubank and Speckman (1989) that one may interpret $\mathbf{b} = (\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T(\mathbf{y} - \mu_\lambda)$ as an estimator of the vector $(\mu^{(m)}(0), \mu^{(m)}(1), \ldots, \mu^{(2m-1)}(0), \mu^{(2m-1)}(1))^T$. Thus, (10) has the interpretation that the boundary behavior of μ_λ is being adjusted using $2m$ transformed by $(\mathbf{I} - \mathbf{S}_\lambda)$ polynomials that are in one-to-one correspondence with each of the $2m$ natural boundary conditions in (6). This fact can actually be used to develop statistical tests for the whether or not boundary corrections are needed. We will discuss this point further in the sequel.

In the next section we develop an order n algorithm for the computation of μ_λ in (10) as well as other related quantities such as the generalized cross-validation criterion that can be used for selecting λ. Then, in Section 3 we study the asymptotic properties of $\tilde{\mu}_\lambda$ in the special case of $m = 1$. In this instance we are able to develop large sample expression for the bias of $\tilde{\mu}_\lambda$ that allows us to characterize asymptotically the effect of the bias correction and show that the Eubank/Speckman upper bound for the risk is sparp in this case. Several technical lemmas that are needed for the work in Section 3 are collected in Section 4.

2. Computation of the estimator

In this section we describe a $O(n)$ algorithms for the computation of $\tilde{\mu}_\lambda$ in (10). This algorithm can be used in conjuction with any order n method for conducting the tranformation (7) such a those given in Reinsch (1967, 1971), de Boor (1978) and Kohn and Ansley (1987). In particular, the simulation results discussed at the end of the section were obtained using a modification of the code in de Boor (1978).

Oberve from (10) that

$$\tilde{\mu}_\lambda = \mu_\lambda + \tilde{\mathbf{Q}}\mathbf{b} \tag{13}$$

for $\mathbf{b}$ any solution of

$$\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}}\mathbf{b} = \tilde{\mathbf{Q}}^T\tilde{\mathbf{y}} \tag{14}$$

with $\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}$. Therefore, to compute $\tilde{\mu}_\lambda$ we can transform $\mathbf{Q}$ and $\mathbf{y}$ to $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{y}}$ in order n operations and then obtain $\mathbf{b}$ through ordinary least-squares regresion of $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{Q}}$. Thus, $\mathbf{b}$ can be computed using statndard statistical software given an efficient method for computing $\tilde{y}$ and $\tilde{\mathbf{Q}}$.

Once b has been computed $\tilde{\mu}_\lambda$ and the residual sum-of-squares

$$\mathrm{RSS}(\lambda) = (\mathbf{y} - \tilde{\mu}_\lambda)^T(\mathbf{y} - \mu_\lambda)$$

can then be computed in $O(n)$ operations using (13). It will also generally be necessary to use a data-driven choice for the smoothing parameter of the modified estimator. One way to accomplish this is to use the value of λ that minimizes the generalized cross-validation criterion

$$\mathrm{GCV}(\lambda) = \frac{n\mathrm{RSS}(\lambda)}{\mathrm{tr}(\mathbf{I} - \mathbf{H}_\lambda)^2} \tag{15}$$

for

$$\mathbf{H}_\lambda = \mathbf{S}_\lambda + \tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T(\mathbf{I} - \mathbf{S}) \tag{16}$$

the hat or smoother matrix for the $\tilde{\mu}_\lambda$. This requires the computation of tr $\mathbf{H}_\lambda$ and we now show how this can be accomplished.

First note that there are several $O(n)$ algorithms for computing tr $\mathbf{S}_\lambda$; see, e.g., Hutchinson and de Hoog (1985). Given any such algorithm, the problem then reduces to the computation of

$$\tau = \mathrm{tr}\ \tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T(\mathbf{I} - \mathbf{S}) = \mathrm{tr}\ (\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}}$$

for $\check{\mathbf{Q}} = (\mathbf{I} - \mathbf{S})\tilde{\mathbf{Q}}$. Set $\check{\mathbf{Q}} = [\check{\mathbf{q}}_1, \ldots, \check{\mathbf{q}}_{2m}]$. Then, we can compute τ by solving the $2m$ normal equation systems

$$\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}}\mathbf{a}_i = \tilde{\mathbf{Q}}^T\check{q}_i, \quad i = 1, \ldots, 2m. \tag{17}$$

Again, this can be done with standard least squares software and the resulting trace is

$$\tau = \sum_{i=1}^{n} a_{ii} \tag{18}$$

for a_{ii} the ith element of $\mathbf{a}_i$ in (17).

As noted in Section 1, there is a one-to-one relationship between the polynomials used in the boundary correction and the boundary conditions in (6). If $\mu^{(n+j-1)}(0) = 0$ then the q_{0j} term in the estimator is not necessary and, similarly, q_{1j} is not needed if $\mu^{(n+j-1)}(1) = 0$ statistical assessment of whether or not a particular boundary conditions holds or, equivalently, whether a particular q_{ij} is needed in the estimator can be obtained by comparing the corresponding element of $\mathbf{b}$ in (14) to its estimated standard error. Thus, it may be useful to also compute an estimator of the varian-covariance matrix for $\mathbf{b}$ in some cases.

For fixed λ, the variance-covariance matrix of $\mathbf{b}$ is

$$\mathbf{V} = \sigma^2(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T(\mathbf{I} - \mathbf{S})^2\tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}$$

assuming that $\tilde{\mathbf{Q}}$ is full rank, which is generally the case. The variance σ^2 in V can be estimated using methods such as those in Gasser, et al (1986). To compute the elements of $(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T(\mathbf{I} - \mathbf{S})^2\tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}$ we solve the linear systems

$$\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}}\mathbf{a}_i = \mathbf{c}_i, \quad i = 1, \ldots, n, \tag{19}$$

for $\mathbf{c}_1, \ldots, \mathbf{c}_n$ the column of the $2m \times n$ matrix $\tilde{\mathbf{Q}}^T(\mathbf{I} - \mathbf{S})$. Note that although there are n systems, each system can be solved in order m^3 calculations so that all n of the $\mathbf{a}_i$ can be obtained in a total of $O(n)$ operations. If $\mathbf{a}_i = (a_{1,i}, \ldots, a_{2m,i})^T$ then the ijth element of $\mathbf{V}$ is

$$\sum_{r=1}^{n} a_{ir}a_{jr}$$

apart from the factor σ^2. These quantities can be accumulated to avoid storage of the $\mathbf{a}_i$.

We have implemented the algorithm described above for thre cubic smoothing spline case of $m = 2$ in (2). In this instance the polynomials in (8) can be chosen to be

$$q_{01}(t) = \frac{1}{2}t^2 - \frac{1}{4}t^4 + \frac{1}{10}t^5,$$

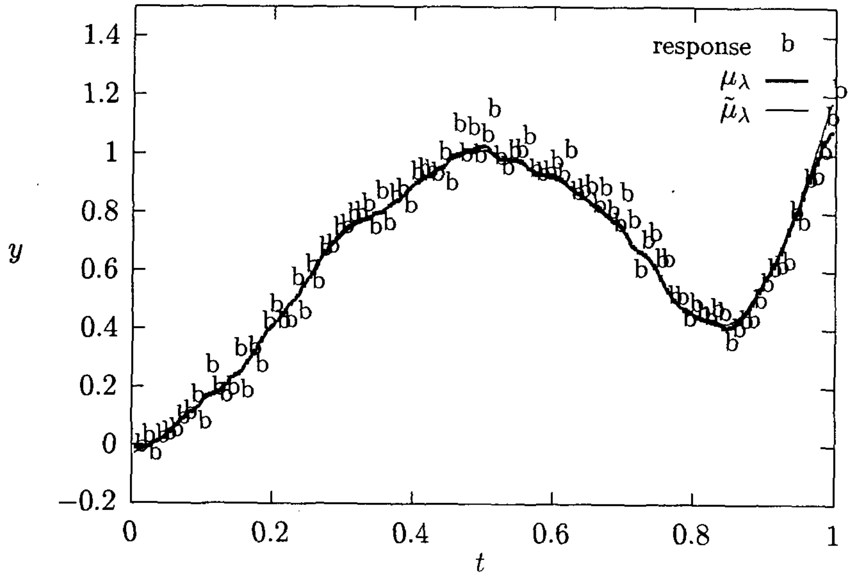

Figure 1: Smoothing spline fits to a simulated data set.

$$q_{11}(t) = \frac{1}{4}t^4 - \frac{1}{10}t^5,$$
$$q_{02}(t) = \frac{1}{6}t^3 - \frac{1}{6}t^4 + \frac{1}{20}t^5,$$

and

$$q_{12}(t) = -\frac{1}{12}t^4 + \frac{1}{20}t^5.$$

The smoothing spline transformation (7) is then carried out using a modified version of the FORTRAN code in Section of de Boor (1978) with $\mathrm{tr}\mathbf{S}_\lambda$ computed by the Hutchinson and de Hoog (1985) algorithm. For each value of λ we then construct a Cholesky factorization of the 4×4 matrix $\tilde{\mathbf{Q}}^T\tilde{\mathbf{q}}$ as $\mathbf{T}^T\mathbf{T}$ for $\mathbf{T}$ upper triangular. The same matrix $\mathbf{T}$ is then used repeadedly to solve the systems (14), (17) and (18) by back substitution.

To test our code and also see the practical effects of boundary correction for the cubic case we conducted a small simulation. Data was generated from model (1) using normal random errors with $\sigma^2 = .1, n = 50, t_i = (2i-1)/2n,\ i = 1, \ldots, n$ and

$$\mu(t) = \gamma g(t) + (\frac{27}{4})^4\ [t^4(1-t)^8 + t^8(1-t^4)]\ /\ 2 \tag{20}$$

for

$$g(t) = \mathrm{I}\ (t > .9)\ (10t - 9)^3 \tag{21}$$

with I (A) the indicator function for the set A. Note that μ satisfies the lower boundary conditions in (6) for $m = 2$ but the upper boundary conditions are not met unless $\gamma = 0$ in (19). By increasing γ a way from zero we move the regression function further away from the situation where (5) can be expected to hold.

For each data set both regular and boundary corrected cubic smoothing spline were computed with their repective smoothing parameters selected by generalized cross validation. The results for a typical data set with $\gamma = 0.05$ in (19) is shown in Figure 1. Notice that the two estimators are virtually identical at the lower boundary but differ near 1 where the natural boundary conditions do not hold.

3. Asymptotics for $m = 1$

While (12) ensures that the risk will be improved by boundary correction when (6) fails to hold, it would be useful to have more precise information concerning the asymptotic form of the risk. The problem with obtaining such a result is that there is not an easily manipulated closed form for the smoothing spline estimator in general. This makes it quite difficult to charactize the asymptotic behavior of the risk even for the non boundary corrected estimator (cf Rice and Rosenblatt, 1983).

Eubank (1997) has shown that in the special case of $m = 1$ and a uniform design it is possible to the obtain uniform approximations for the pointwise variance and bias of a smoothing spline. Thus we now specialize to the case of $m = 1$ with $t_i = (2i-1)/2n,\ i = 1, \ldots, n$, and employ these approximation to the analyze the boundary corrected linear smoothing spline.

For the case of $m = 1$ we can choose the polynomials (8) - (9) to be

$$q_0(t) = t - \frac{1}{2}t^2 \tag{22}$$

and

$$q_1(t) = \frac{1}{2}t^2. \tag{23}$$

Now let $q_{0\lambda}$ and $q_{1\lambda}$ be the linear smoothing spline approximations to q_0 and q_1. More precisely $q_{0\lambda}$ and $q_{1\lambda}$ are the functions obtained by minimizing the criterion

$$n^{-1}\sum_{i=1}^{n}(g(t_i) - f(t_i))^2 + \lambda\int_0^1 f'(t)^2 dt \tag{24}$$

with respect to f using $g(t_i) = q_0(t_i),\ i = 1, \ldots, n$ and $g(t_i) = q_1(t_i),\ i = 1, \ldots, n$, respectively. If we now define

$$\tilde{q}_i(t) = q_i(t) - q_{i\lambda}(t), \quad i = 0, 1 \tag{25}$$

then the boundary corrected linear smoothing spline is

$$\tilde{\mu}_\lambda = \mu_\lambda(t) + b_0\tilde{q}_0(t) + b_1\tilde{q}_1(t), \tag{26}$$

where

$$b_0 = (\tilde{q}_{11}\sum_{i=1}^{n}\tilde{q}_0(t_i)y_i - \tilde{q}_{01}\sum_{i=1}^{n}\tilde{q}_1(t_i)y_i)/(\tilde{q}_{00}\tilde{q}_{11} - \tilde{q}_{01}^2) \tag{27}$$

and

$$b_1 = (\tilde{q}_{00}\sum_{i=1}^{n}\tilde{q}_1(t_i)y_i - \tilde{q}_{01}\sum_{i=1}^{n}\tilde{q}_0(t_i)y_i)/(\tilde{q}_{00}\tilde{q}_{11} - \tilde{q}_{01}^2) \tag{28}$$

for

$$\tilde{q}_{ij} = \sum_{k=1}^{n}\tilde{q}_i(t_k)\tilde{q}_j(t_k), \quad i,j = 1,2. \tag{29}$$

We can use expressions (25) - (28) along with the approximation lemmas in Section 4 to obtain large sample expressions for the bias risk of the estimator. In what follows we will impose the restriction that $\lambda \to 0$, as $n \to \infty$ in such a way that $\log n/n\lambda^2 \to 0$. Under this condition we have from Lemma 1 in Section 4 that

$$\tilde{q}_0(t) = \sqrt{\lambda}e^{-t/\sqrt{\lambda}} + O(\lambda)$$

and

$$\tilde{q}_1(t) = \sqrt{\lambda}e^{-(t-1)/\sqrt{\lambda}} + O(\lambda)$$

Lemma 3 then gives that

$$\mathrm{E}\ b_0 = \mu'(0) + O(\sqrt{\lambda})$$

and

$$\mathrm{E}\ b_1 = \mu'(1) + O(\sqrt{\lambda})$$

Thus, the bias of $\tilde{\mu}_\lambda$ is

$$\begin{aligned} \mu(t) - \mathrm{E}\ \tilde{\mu}_\lambda(t) &= \mu(t) - \mathrm{E}\ \mu_\lambda(t) \\ &\quad -\mu'(0)\sqrt{\lambda}e^{-t/\sqrt{\lambda}} - \mu'(1)\sqrt{\lambda}e^{(t-1)/\sqrt{\lambda}} \\ &\quad +O(\lambda(e^{-t/\sqrt{\lambda}} + e^{(t-1)/\sqrt{\lambda}}) + \lambda^{5/2}). \end{aligned} \tag{30}$$

Expression (30) clearly shows that $\tilde{\mu}_\lambda$ gives an adjustment to the bias of the original estimator μ_λ by subtracting off terms corresponding to each of the boundaries. Note that the correction is localized to the boundaries in Lemma 1 combined with (30) gives

$$\mu(t) - \mathrm{E}\ \tilde{\mu}_\lambda(t) = \lambda\mu''(t) + O(\lambda^{5/2})$$

for any fixed $t \in (0, 1)$. Thus $\tilde{\mu}_\lambda$ how the bias properties of a second order estimator in the interior on [0, 1] as does μ_λ. However, for boundary points such as a $t = \sqrt{\lambda}v$

or $t = (1 - \sqrt{\lambda})v$ we still have $\mu(t) - \mathrm{E}\,\tilde{\mu}_\lambda(t) = O(\lambda)$ while $\mu(t) - \mathrm{E}\,\mu_\lambda(t)$ is of exact order $\sqrt{\lambda}$. Thus, $\tilde{\mu}_\lambda$ has effectively reduced the pointwise bias of μ_λ to that of a second order estimator at the boundaries.

Theorem 1. Assume that $\lambda \to 0$ as $n \to \infty$ with $\log n/n\lambda^2 \to 0$ and that $\mu \in C^3[0,1]$. Define $\mu_0(t) = \mu(t) - \mu'(0)q_0(t) - \mu'(1)q_1(t)$ for q_0 and q_1 in (21) - (22). Then

$$\begin{aligned}\tilde{\mathrm{R}}_n(\lambda) &= n^{-1}\sum_{i=1}^{n}\mathrm{E}\,(\tilde{\mu}_\lambda(t_i) - \mu(t_i))^2 \\ &= \lambda^2\int_0^1 \mu_0''(t)^2dt + \frac{\sigma^2}{4n\sqrt{\lambda}}(1 + o(1)) + O(\lambda^{5/2} + n^{-1}).\end{aligned}$$

Proof. Define $\mathbf{Q}$ to be the $n \times 2$ matrix with ij element $q_{j-1}(t_i)$, $i = 1, \ldots, n$, $j = 1, 2$. If we now take $\mu_0 = (\mu_0(t_1), \ldots, \mu_0(t_n))^T$ we have

$$\mu - \mathrm{E}\,\tilde{\mu}_\lambda = \tilde{\mu}_0 - \tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T\tilde{\mu}_0$$

with $\tilde{\mu}_0 = (\mathbf{I} - \mathbf{S}_\lambda)\mu_0$, $\tilde{\mathbf{Q}} = (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Q}$ and $\mathbf{S}_\lambda$ the $\mathbf{S}_\lambda$ smoother matrix in (7) for the linear smoothing spline.

Using Lemma 3 in Section 4 we obtain

$$n^{-1}\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}} = \frac{\lambda^{3/2}}{2}\begin{bmatrix} 1 + O(\sqrt{\lambda}) & O(\sqrt{\lambda}) \\ O(\sqrt{\lambda}) & 1 + O(\sqrt{\lambda}) \end{bmatrix}$$

and

$$n^{-1}\tilde{\mu}_0^T\tilde{\mathbf{Q}} = (O(\lambda^2), O(\lambda^2))$$

since $\mu_0'(0) = \mu_0'(1) = q_0'(1) = q_1'(0) = 0$ and $q_0'(0) = q_1'(1) = 1$. Thus,

$$\begin{aligned}n^{-1}(\mu - \mathrm{E}\tilde{\mu}_\lambda)^T(\mu - \mathrm{E}\tilde{\mu}_\lambda) &= n^{-1}\tilde{\mu}_0^T\tilde{\mu}_0 - n^{-1}\tilde{\mu}_0^T(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T\tilde{\mu}_0 \\ &= \lambda^2\int_0^1 \mu_0''(t)^2dt + O(\lambda^{5/2} + n^{-1}).\end{aligned}$$

For the variance part of $\tilde{\mathrm{R}}_n(\lambda)$ we can write and observe that

$$\mathrm{tr}\mathbf{H}_\lambda = \mathrm{tr}\mathbf{S}_\lambda^2 + 2 - \mathrm{tr}\mathbf{S}_\lambda{}^2\tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T.$$

The trace term is a most 2 in magnitude since the eigenvalues of $\mathbf{S}_\lambda$ are all bounded by 1. The proof is then completed using Lemma 4.

It follows from the Theorem that an asymptotically optimal choice of the smoothing parameter for $\tilde{\mu}_\lambda$ is provided by

$$\lambda_n^* = \left(\sigma^2/4n\int_0^1 \mu_0''(t)^2dt\right)^{2/5}$$

which produces the risk

$$\mathrm{R}_n(\lambda_n^*) = 1.25\left(\frac{\sigma^2}{4n}\right)^{4/5}\left(\int_0^1 \mu_0''(t)^2 dt\right)^{1/5} + o(n^{-4/5}).$$

Thus, $\tilde{\mu}_\lambda$ behaves essentially like a boundary corrected, second order kernel estimator in terms of its risk asymptotics. However, there is an important difference in that (30) involves $\int_0^1 \mu_0''(t)^2 dt$ rather than $\int_0^1 \mu''(t)^2 dt$ which is that would be expected from a kernel estimator. Note that

$$\int_0^1 \mu_0''(t)^2 dt = \int_0^1 \mu''(t)^2 - \left(\int_0^1 \mu''(t)^2\right)^2 \leq \int_0^1 \mu''(t)^2$$

with strict inequality unless $\mu'(0) = \mu'(1) = 0$. Thus, the boundary correction that is being made to μ_λ actually has a global impact on the estimation risk.

4. Lemmas

In this section we prove several lemmas that are needed for the proof of Theorem in Section 3. All the results that follow pertain only to the case of a linear smoothing spline with a uniform design $t_i = (2i-1)/2n$, $i = 1, \ldots, n$.

For any function g let g_λ be its linear smoothing spline approximation obtained by minimizing

$$n^{-1}\sum_{i=1}^{n}(g(t_i) - f(t_i))^2 + \lambda\int_0^1 f'(t)^2 dt$$

over all absolutely continuous functions f with square integrable derivatives. For convenience we will suppress the λ subscript and use the notation

$$\tilde{g} = g - g_\lambda$$

to denote the error in approximating g by g_λ. The function $\tilde{g}$ is also the bias from a linear smoothing spline fit to a data set with regression function g.

Our first three lemmas provide asymptotic approximations to the linear smoothing spline bias and bias inner products.

Lemma 1. If $g \in C^3[0,1]$ and $\lambda \asymp n^{-8}$ for $\gamma \in (0,1)$, then

$$\tilde{g}(t) = \sqrt{\lambda}g'(t)h_{1\lambda}(t) - \lambda g''(t) - \lambda g''(t)h_{2\lambda}(t) + O(\frac{\log n}{n\lambda} + \lambda^{5/2})$$

uniformly in $t \in [0,1]$ for

$$h_{1\lambda}(t) = e^{\frac{t-1}{\sqrt{\lambda}}} - e^{\frac{t}{\sqrt{\lambda}}}$$

and

$$h_{2\lambda}(t) = \frac{t-1}{\sqrt{\lambda}} e^{\frac{t-1}{\sqrt{\lambda}}} - \frac{t}{\sqrt{\lambda}} e^{\frac{t}{\sqrt{\lambda}}}.$$

Lemma 2. If $\lambda \asymp n^{-8}$ for some $\gamma \in (0,1)$ and $g \in C^2[0,1]$, then

$$n^{-1} \sum_{i=1}^{n} g(t_i) h_{1\lambda}^{k}(t_i) = \left(\begin{array}{l} \frac{\sqrt{\lambda}}{k} \left(g(1) + (-1)^j g(0)\right) + O(\lambda + \frac{1}{n\sqrt{\lambda}}) \\ \\ O(\lambda^{3/2} + \frac{1}{n\sqrt{\lambda}}), \text{ if } g'(0) = g'(1) = 0. \end{array} \right.$$

Also,

$$n^{-1} \sum_{i=1}^{n} g(t_i) h_{2\lambda}(t_i) h_{r\lambda}(t_i) = O(\lambda^{j/2} + \frac{1}{n\sqrt{\lambda}})$$

for $r = 1, 2$ with $j = 2$ when $g(0) = g(1) = 0$ and $j = 1$ otherwise.

Lemma 3. If $\lambda \to 0$ as $n \to \infty$ with $\log n / n\lambda^2 \to 0$ and $f, g \in C^3[0,1]$, then

$$n^{-1} \sum_{i=1}^{n} \tilde{g}(t_i) \tilde{f}(t_i) = \frac{\lambda^{3/2}}{2} \left(g'(1) f'(1) + g'(0) f'(0)\right) + O(\lambda^2).$$

If $g'(0) = g'(1) = f'(0) = f'(1)$, then

$$n^{-1} \sum_{i=1}^{n} \tilde{g}(t_i) \tilde{f}(t_i) = \lambda^2 \int_0^1 g''(t) f''(t) dt + O(\lambda^{5/2}).$$

Lemma 4. If $\lambda \to 0$ as $n \to \infty$ with $\log n / n\lambda^2 \to 0$, then $n^{-1} \mathrm{tr} \mathbf{S}_\lambda = \frac{1}{4n\sqrt{\lambda}}(1 + o(1))$.

References

1. de Boor, C. (1978). *A Practical Guide to Splines.* New York: Springer-Verlag.

2. Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, Inc.

3. Eubank, R. L. (1997). Regression. Marcel Dekker, Inc.

4. Eubank, R. L. and Speckman, P. L. (1991). A bias reduction theorem with Applications in nonparameteric regression. *Scandinavian Journal of Statistics, 18,* 211 - 222

5. Gasser, Th., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika, 73,* 625-633.

6. Hutchinson, M. F. anf de Hoog, F. R. (1985). Smoothing noisy data with spline functions. *Numerical Mathematics, 47,* 99-106.

7. Kohn, R. and Ansely, C. F. (1987). A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM Journal on Scientific and Statistical Computing, 8,* 33-48.

8. Oehlet, G. W. (1992). Relaxed boundary smoothing spline. *Annals of Statistics, 20,* 146-160.

9. Reinsch, C. (1971). Smoothing by spline functions, II. *Numerical Mathematics, 16,* 451-454.

10. Rice, J. and Rosenblatt, M. (1983). Smoothing by spline functions. *Numerical Mathematics, 10,* 177-183.

11. Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, B, 50,* 413-436.

12. Wahba, G. (1990). *Spline Models for Observational Data.* SIAM:Philadelphia.