Econometric Theory, **15**, 1999, 114–138. Printed in the United States of America.

CONSTRAINED SMOOTHING SPLINES

JUAN M. RODRIGUEZ PÓO
Universidad de Cantabria

We use smoothing splines to introduce prior information in nonparametric models. The type of information we consider is based on the belief that the regression curve is similar in shape to a parametric model. The resulting estimator is a convex sum of a fit to data and the parametric model, and it can be seen as shrinkage of the smoothing spline toward the parametric model. We analyze its rates of convergence and we provide some asymptotic distribution theory. Because the asymptotic distribution is intractable, we propose to carry out inference with the estimator by using the method proposed by Politis and Romano (1994, *Annals of Statistics* 22, 2031–2050). We also propose a data-driven technique to compute the smoothing parameters that provides asymptotically optimal estimates. Finally, we apply our results to the estimation of a model of investment behavior of the U.S. telephone industry and we present some Monte Carlo results.

1. INTRODUCTION

In this paper we address the problem of incorporating prior information in nonparametric regression estimates. The type of information we introduce is based on the belief that the regression curve is similar in shape to a parametric curve. If we assume directly a parametric form for the regression curve, the unknown parameters can be estimated by several procedures. Parametric estimation methods provide good efficiency properties under very general conditions (see Bickel and Doksum, 1977), but they often impose too much structure on data resulting in estimators with non-negligible asymptotic bias. Instead of constraining the regression curve to belong to a family of parametric functions, we can estimate it by nonparametric regression methods. These methods allow us to extend the class of structures under which the chosen procedures give valid inference. Among these methods, the most popular are kernel regression (Härdle, 1990), smoothing splines (Wahba, 1990), and local polynomial regression (Fan and Gijbels, 1996).

Preliminary parametric components traditionally have been incorporated in the nonparametric setup through partial linear models (see Heckman, 1986; Green, 1987; Speckman, 1988). In fact, this kind of model introduces the parametric

The author thanks W. Härdle, M. Delecroix, one co-editor, and two anonymous referees for their very helpful comments and suggestions. Research on this paper was done within the Sonderforshungsbereich 373 at Humboldt University Berlin. Financial support from the Dirección General de Enseñanza Superíor, research project PB96-1469-C05-03, is gratefully acknowledged. Address correspondence to Juan M. Rodriguez Póo, Universidad de Cantabria, Avda. de los Castros s/n, 39005 Santander, Spain; e-mail: rodrigjm@ccaix3.unican.es.

 © 1999 Cambridge University Press 0266-4666/99 $9.50

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

component in the belief that at least part of the mean response can be approximated by this parametric structure. However, as has been remarked by Rice (1986), the statistical properties of both parametric and nonparametric components are interrelated, and incorrect specification of the parametric part of the model leads to asymptotic bias in the estimators of both components.

We propose a nonparametric estimator that introduces the preliminary parametric component through a constrained optimization technique. This procedure has at least two advantages with respect to the previous developments. First, it introduces the parametric model as a restriction, and, therefore, it is possible to relax this constraint with a penalty parameter that is data driven. In the second place, the statistical properties of the estimator remain intact even if the parametric model is incorrectly specified. In such cases, the estimator still achieves its proper rates of convergence.

The proposed constrained nonparametric estimator is based on smoothing splines. Because they are computed as a solution to a constrained optimization problem, they are well suited to imposing additional restrictions in nonparametric models. The resulting estimator is a convex sum of a fit to data and the parametric model, and it can be viewed as "shrinking the smoothing spline towards the parametric model," in the Stein shrinkage style (see James and Stein, 1961). The attractiveness of this procedure is that when the parametric model is correctly specified then the nonparametric smoother can borrow strength from the low dimensional model by shrinkage toward it. We also establish connections with other works combining parametric and nonparametric estimators, such as Olkin and Spiegelman (1987) and Burman and Chaudhuri (1992).

In Section 2 we introduce the model and we propose the estimator. In Section 3 we compute the asymptotic bounds and the optimal rates of convergence and derive the asymptotic distribution of our estimator. Section 4 provides a data-driven method to compute the penalty parameters. Section 5 contains an application in the estimation of a model of investment behavior in the U.S. telephone industry and a simulation study. Finally, in Section 6 we give some conclusions and suggest areas for future research. The main results are proved in the Appendix.

2. STATISTICAL MODEL AND ESTIMATION PROCEDURE

We consider a model of n independent observations $y_1, y_2, \ldots, y_n$ with expectation $m(x)$:

$$y_i = m(x_i) + \epsilon_i \qquad (i = 1, \ldots, n). \tag{1}$$

The design variables x_i are nonrandom and real valued, and, for simplicity, they are assumed to lie in $[0,1]$. The function $m(\cdot)$ is unknown and needs to be estimated. The parametric curve that represents the prior belief is $g(x;\theta)$, where θ is a p-dimensional unknown parameter vector. From now on, we introduce the following notation $y = (y_1, \ldots, y_n)^T$, $m = (m(x_1), \ldots, m(x_n))^T$, and $g(\theta) =$

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

$(g(x_1;\theta),\ldots,g(x_n;\theta))^T$. For any vector h in R^n, we define $\|h\|_n^2$ as the Euclidean norm of h in R^n. We shall make the following assumptions.

A.1. The ϵ_i are independent and identically distributed (i.i.d.) random variables with zero mean and variance $\sigma^2 > 0$. They also have finite absolute $(2+\alpha)$ moments, $\alpha \geq 8$.

A.2. Define $W_2^{(\nu)}[0,1]$ as a νth order Sobolev space:

$$W_2^{(\nu)}[0,1] = \left\{ m \mid m,\ldots,m^{(\nu-1)} \quad \text{are absolutely continuous} \quad \text{and} \right.$$
$$\left. \int (m^{(\nu)}(t))^2\, dt < \infty \right\}.$$

We assume that $m(x) \in W_2^{(\nu)}[0,1]$.

A.3. We also assume that the function $g(x;\theta)$ and their $\tau - 1$ first derivatives are absolutely continuous and that its τth derivative is a bounded square integrable function in $[0,1]\times\Theta$. The functions $\partial g(x;\theta)/\partial\theta_r$ and $\partial^2 g(x;\theta)/\partial\theta_r\partial\theta_s$ $(r,s = 1,2,\ldots,p)$ are continuous on $[0,1]\times\Theta$.

A.4. Θ is a closed, bounded (compact) subset of R^p.

A.5. The observations x_i are such that $F_n(x) \to F(x)$, where $F_n(x)$ is the empirical distribution function and $F(x)$ is a distribution function.

A.6. Define θ_0 as

$$\theta_0 = \arg\inf_{\theta\in\Theta} \|m - g(\theta)\|_n^2. \tag{2}$$

We assume that if $g(z;\theta) = g(z;\theta_0)$ then $\theta = \theta_0$. Moreover, θ_0 is an interior point of Θ.

A.7. The matrix function $A(\theta_0) = [a_{rs}(\theta_0)]$, where

$$a_{rs}(\theta_0) = \int \frac{\partial g(x;\theta)}{\partial\theta_r}\,\frac{\partial g(x;\theta)}{\partial\theta_s}\, dF(x) \tag{3}$$

is nonsingular.

We also define the distance between the unknown regression function and the parametric family of functions under consideration as

$$\delta_n = \|m - g(\theta_0)\|_n^2. \tag{4}$$

Note that if the parametric model is correct, $\delta_n = 0$, and if the parametric model is incorrectly specified, $\delta_n > 0$. The distance δ_n will determine the optimal amount of shrinkage of the smoothing spline toward the parametric model.

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

To introduce the belief that $m(\cdot)$ is similar in shape to $g(\cdot,\theta)$ we propose to derive an estimator that is based on a modified version of the following penalized residual sum of squares:

$$L(m) = \|y - m\|_n^2 + \lambda_1 \int_0^1 (m^{(\nu)}(u))^2 \, du. \tag{5}$$

The first term of the sum accounts for the degree of fitness, and the second term is the roughness penalty. It is well known (see Eubank, 1988) that minimizing $L(m)$ over the class of all (ν-differentiable) functions $m(\cdot)$, at grid points $\{x_i\}_{i=1}^n$, yields an estimate $\hat{m}_{\lambda_1} = (\hat{m}_{\lambda_1}(x_1), \hat{m}_{\lambda_1}(x_2), \ldots, \hat{m}_{\lambda_1}(x_n))^T$, which for given values of λ_1 is the best compromise between smoothness and goodness of fit. It can be also shown that the curve estimate $\hat{m}_{\lambda_1}$ has the following properties:

(i) $\hat{m}_{\lambda_1}(\cdot)$ is a polynomial of degree $2\nu - 1$ on any subinterval $[x_i, x_{i+1})$ for $i = 1, \ldots, n-1$.
(ii) $\hat{m}_{\lambda_1}(\cdot)$ and its $2\nu - 2$ first derivatives are continuous at the observation points $\{x_i\}_{i=1}^n$.
(iii) The $(2\nu - 1)$th derivative is a step function with jumps at the x_i's.
(iv) $\hat{m}_{\lambda_1}^{(\nu)}(t) = 0$ outside of $[x_1, x_n]$.

As a result of (iv), $\hat{m}_{\lambda_1}(x)$ satisfies the so-called natural boundary conditions

$$\hat{m}_{\lambda_1}^{(j)}(0) = 0 \qquad j = \nu, \ldots, 2\nu - 1 \tag{6}$$

and

$$\hat{m}_{\lambda_1}^{(j)}(1) = 0 \qquad j = \nu, \ldots, 2\nu - 1. \tag{7}$$

The additional restriction imposed by the parametric function $g(x;\theta)$ is introduced by adding a new term in (5):

$$L(m;\theta) = \|y - m\|_n^2 + \lambda_1 \int_0^1 (m^{(\nu)}(u))^2 \, du + \lambda_2 \|m - g(\theta)\|_n^2. \tag{8}$$

The new term reflects the distance between the unknown regression function $m(\cdot)$ and the prespecified parametric function $g(\cdot;\theta)$. The two parameters λ_1 and λ_2 represent a trade-off in the requirements about fidelity to data, degree of smoothness, and closeness to a prespecified function. Ansley, Kohn, and Wong (1993) proposed to introduce the parametric constraint by considering a roughness penalty defined by a differential equation instead of introducing a new term in the criterion function.

We estimate the unknown regression curve $m(\cdot)$ and the parameter vector θ using a two-step procedure. First, the unknown vector of parameters θ is estimated by nonlinear least squares techniques, $\hat{\theta}_n$, i.e.,

$$\hat{\theta}_n = \arg\inf_{\theta \in \Theta} \|y - g(\theta)\|_n^2. \tag{9}$$

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

Then, the nonlinear estimate is plugged into the criterion function $L(m;\theta)$,

$$L_n(m;\hat{\theta}_n) = \|y - m\|_n^2 + \lambda_1 \int_0^1 (m^{(\nu)}(u))^2\, du + \lambda_2 \|m - g(\hat{\theta}_n)\|_n^2, \tag{10}$$

in this way finding the value of m that minimizes the preceding criterion function,

$$\hat{m}_\lambda = \underset{m \in W_2^{(\nu)}[0,1]}{\arg\inf}\; L(m;\hat{\theta}_n), \tag{11}$$

where $\hat{m}_\lambda = (\hat{m}_{\lambda_1,\lambda_2}(x_1), \hat{m}_{\lambda_1,\lambda_2}(x_2), \ldots, \hat{m}_{\lambda_1,\lambda_2}(x_n))^T$ and $g(\hat{\theta}_n) = (g(x_1;\hat{\theta}_n), \ldots, g(x_n;\hat{\theta}_n))^T$. Before giving a closed form for the resulting estimator, $\hat{m}_\lambda$, let us introduce some more notation. Denote by $NS^{2\nu}(x_1, x_2, \ldots, x_n)$ the linear space of all real valued functions defined in [0,1] satisfying (i)–(iv). A closed form for the estimator $\hat{m}_\lambda$ is given by the following result.

THEOREM 2.1. *Let $X_1, \ldots, X_n$ be any basis for $NS^{2\nu}(x_1, \ldots, x_n)$. Assume that* A.1 *and* A.7 *hold. Let $\hat{\theta}_n$ be the solution to the following problem*:

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\arg\inf} \|y - g(\theta)\|_n^2. \tag{12}$$

For fixed $0 < \lambda_1 < \infty$ and $0 < \lambda_2 < \infty$, there is a unique minimizer $\hat{m}_\lambda$ of $L_n(m;\hat{\theta}_n)$ in $m(x) \in W_2^{(\nu)}[0,1]$. Moreover, $\hat{m}_\lambda \in NS^{2\nu}(x_1, x_2, \ldots, x_n)$ and $\hat{m}_\lambda = \sum_{j=1}^n \beta_{\lambda j} X_j$. The coefficients $\beta_\lambda = (\beta_{\lambda 1}, \beta_{\lambda 2}, \ldots, \beta_{\lambda n})^T$ are the solution to

$$\left(X^T X + \frac{\lambda_1}{1+\lambda_2}\,\Omega_n\right)\beta_\lambda = \frac{1}{1+\lambda_2}\,X^T y + \frac{\lambda_2}{1+\lambda_2}\,X^T g(\hat{\theta}_n), \tag{13}$$

where

$$X = \{X_j(x_i)\}_{i=1,\ldots,n}^{j=1,\ldots,n} \tag{14}$$

and

$$\Omega_n = \left\{ \int_0^1 X_i^{(\nu)}(u) X_j^{(\nu)}(u)\, du \right\}_{i=1,\ldots,n}^{j=1,\ldots,n}. \tag{15}$$

The resulting nonparametric estimator has a very simple form in terms of smoothing splines. Because

$$H_n\left(\frac{\lambda_1}{1+\lambda_2}\right) = X\left(X^T X + \frac{\lambda_1}{1+\lambda_2}\,\Omega_n\right)^{-1} X^T \tag{16}$$

is the smoothing matrix for the $(\nu - 1)$th order smoothing spline with penalty parameter $\lambda_1/1 + \lambda_2$ (see Eubank, 1988, p. 206), we can rewrite (13) as

$$\hat{m}_\lambda = \frac{1}{1+\lambda_2} H_n\left(\frac{\lambda_1}{1+\lambda_2}\right) y + \frac{\lambda_2}{1+\lambda_2} H_n\left(\frac{\lambda_1}{1+\lambda_2}\right) g(\hat{\theta}_n). \tag{17}$$

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

This expression is crucial both in terms of understanding and of analyzing the estimator. In fact, it is a convex sum of a fit to y and to $g(\hat{\theta}_n)$, and it encompasses a great variety of models. If $\lambda_2 = 0$, then we have the natural smoothing spline estimator. If $\lambda_2 \to \infty$ and $\lambda_1 = 0$ then we obtain the parametric fit $g(\hat{\theta}_n)$. This estimator is also related to the one proposed by Burman and Chaudhuri (1992), where their estimator is also a convex combination of both a nonparametric and a parametric estimator. Unfortunately, they were unable to justify the expression obtained for their estimator. Expression (17) also makes it very easy to derive the asymptotic properties for this estimator.

3. ASYMPTOTIC PROPERTIES

As a measure of discrepancy between the estimator $\hat{m}_\lambda(\cdot)$ and the true regression function $m(\cdot)$, we define the following functions:

$$L_n(\lambda_1, \lambda_2) = \|\hat{m}_\lambda - m\|_n^2 \tag{18}$$

and

$$R_n(\lambda_1, \lambda_2) = E\|\hat{m}_\lambda - m\|_n^2. \tag{19}$$

Let us introduce some more notation. The expression $\alpha_n = O(\beta_n)$ means that there exists a constant M such that $|\alpha_n| \leq M|\beta_n|$, as $n \to \infty$. The expression $\alpha_n \sim \beta_n$ means that $\alpha_n/\beta_n \to c$ as n tends to infinity.

The following theorem describes the asymptotic behavior of $\hat{m}_\lambda$.

THEOREM 3.1 (Asymptotic bounds). *Assume conditions* A.1–A.7 *hold. Then if* $\lambda_1/1 + \lambda_2 \to 0$ *and* $n(\lambda_1/1 + \lambda_2) \to \infty$,

$$L_n(\lambda_1, \lambda_2) = \left(\frac{1}{1+\lambda_2}\right)^2 \left[O\left(\frac{\lambda_1}{1+\lambda_2}\right) + O_p\left(n^{-1}\left(\frac{\lambda_1}{1+\lambda_2}\right)^{-(1/2\nu)}\right)\right] + \left(\frac{\lambda_2}{1+\lambda_2}\right)^2 \left[O\left(\frac{\lambda_1}{1+\lambda_2}\right) + O_p\left(\frac{1}{n}\right) + \delta_n\right]$$

as n goes to infinity.

In the case that $g(x;\theta)$ is incorrectly specified ($\delta_n > 0$) then the quantity $L_n(\lambda_1, \lambda_2)$ is bounded away from zero and we do not get the desired consistency result for the estimator $\hat{m}_\lambda$. One solution to this problem is to choose a sequence of values for λ_2 that tends to zero when the sample size n increases. Therefore, the consistency of $\hat{m}_\lambda$ requires that both penalty parameters, λ_1 and λ_2, tend to zero. We can also observe that if the parametric model $g(x;\theta)$ is correctly specified, the value of λ_2 is negligible for the consistency of $\hat{m}_\lambda$. The following theorem gives the achievable rates of convergence for this estimator.

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

THEOREM 3.2 (Rates of convergence). *Assume conditions* A.1–A.7 *hold. If* $\lambda_1 \sim n^{-2\nu/(2\nu+1)}$ *and* $\lambda_2 = O(n^{-\nu/(2\nu+1)})$ *then*

$$L_n(\lambda_1, \lambda_2) = O_p(n^{-2\nu/(2\nu+1)}), \tag{20}$$

as n tends to infinity.

Moreover, if the parametric model is correctly specified ($\delta_n = 0$), *when* $\lambda_1 \sim n^{-2\nu/(2\nu+1)}$ *and* $\lambda_2 \sim n^{1/2(2\nu+1)}$ *then*

$$L_n(\lambda_1, \lambda_2) = O_p(n^{-1}) \tag{21}$$

as n tends to infinity.

The constrained smoothing spline estimator achieves, at worst, a rate of convergence that was shown by Stone (1982) to be optimal for this class of nonparametric regression models. However, if the parametric model is correctly specified then, under a predetermined sequence of penalty parameters, $\hat{m}_\lambda$ achieves parametric rates. We can also derive a convergence in distribution result for $\hat{m}_\lambda$.

THEOREM 3.3 (Asymptotic normality). *Assume conditions* A.1–A.7 *hold and let* $x \in (0,1)$. *If* $\lambda_1 \sim n^{-2\nu/(2\nu+1)}$ *and* $n^{2\nu(2\nu+1)}\lambda_2 \to 0$ *then*

$$\frac{\hat{m}_\lambda(x) - m(x)}{\sigma(x)} \to_d N(0,1) \tag{22}$$

as n tends to infinity, where

$$\sigma^2(x) = E(\hat{m}_\lambda(x) - m(x))^2.$$

Note that under the conditions established in Theorem 3.3, the rate of convergence is the optimal $n^{-2\nu/(2\nu+1)}$. However, we needed to impose stronger conditions for the speed of convergence of the penalty parameter λ_2 toward zero. The reason is that, although under the rates of decrease for λ_1 and λ_2 established in Theorem 3.2 the limiting distribution is still normal, the task of estimating the bias is awkward. It is therefore preferable to eliminate the mean of the limiting normal distribution by increasing the rate at which the penalty parameter λ_2 must tend to zero.

The variance $\sigma^2(x)$ is intractable, and, therefore, it is very difficult to make inferences based directly on the asymptotic distribution of the estimator. However, the method provided by Politis and Romano (1994) to construct confidence intervals provides a very nice solution to this problem. Our goal is to construct a confidence region for $m(x)$.

Let $(x^i, y^i) = \{(x_{(1)}, y_{(1)}), \ldots, (x_{(N_n)}, y_{(N_n)})\}$ be one of the $N_n = \binom{n}{b}$ ordered subsets of $(x_1, y_1), \ldots, (x_n, y_n)$. In typical situations it will be assumed that $b/n \to \infty$ and $b \to \infty$ as n tends to infinity. Now, let $Z_{n,i}(x)$ be $\hat{m}_\lambda(x)$ evaluated at the data set (x^i, y^i). Define also as J_n the sampling distribution of $\gamma_n(\hat{m}_\lambda(x) - m(x))$ based

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

on a sample of size n with corresponding c.d.f. denoted by $J_n(\cdot)$. The term γ_n is a rate of convergence that will be defined precisely in the next theorem. The c.d.f. of the Gaussian distribution is $\Phi(\cdot)$, and the approximation to $J_n(\cdot)$ we study is defined by

$$T_n(u) = \frac{1}{N_n} \sum_{i=1}^{N_n} 1\{\gamma_b(Z_{n,i}(x) - \hat{m}_\lambda(x)) \leq u\}. \tag{23}$$

The following result is based on Theorem 2.1 from Politis and Romano (1994).

THEOREM 3.4. *Assume conditions* A.1–A.7 *hold and let u be a continuity point of* $J_n(\cdot)$. *Then if* $\lambda_1 \sim n^{-2\nu/(2\nu+1)}$ *and* $n^{2\nu/(2\nu+1)}\lambda_2 \to 0$

(i) $T_n(u) \to \Phi(u)$ *in probability.*
(ii) $\sup_u |T_n(u) \to J_n(u)| \to 0$ *in probability.*
(iii) *Let* $c_n(1-\alpha) = \inf\{u : T_n(u) \geq 1-\alpha\}$. *Correspondingly, define* $c(1-\alpha) = \inf\{u : \Phi(u) \geq 1-\alpha\}$. *Then*

$$\Pr\{n^{\nu/(2\nu+1)}(\hat{m}_\lambda(x) - m(x)) \leq c_n(1-\alpha)\} \to 1-\alpha, \tag{24}$$

and the asymptotic coverage probability of the interval $[\hat{m}_\lambda(x) - n^{-\nu/(2\nu+1)} c_n(1-\alpha), \infty)$ *is* $1-\alpha$.

4. THE CHOICE OF THE PENALTY PARAMETERS

In Section 3, we have shown some asymptotic properties of the constrained smoothing spline estimator. However, all of these properties have been derived under a predetermined sequence of values for the penalty parameters λ_1 and λ_2. In this section, we provide an automatic method to compute both parameters, and then we show how this method of selection gives the proper rates of convergence.

In order to estimate the penalty parameters λ_1 and λ_2 we propose to use the generalized cross validation method. This procedure computes the penalty parameters λ_1 and λ_2 that minimize the following criterion function:

$$GCV_n(\lambda_1, \lambda_2) = \frac{n^{-1} \sum_{i=1}^{n} (y_i - \hat{m}_\lambda(x_i))^2}{\left[1 - n^{-1} \operatorname{tr}\left(H_n\left(\frac{\lambda_1}{1+\lambda_2}\right)\right)\right]^2}. \tag{25}$$

The generalized cross validation criterion was originally introduced by Craven and Wahba (1979) in the context of smoothing spline functions, and its performance with respect to other estimation criteria has been studied in Wahba (1985) and Kohn and Ansley (1991).

We will first investigate the relationship between the $GCV_n(\lambda_1, \lambda_2)$ and $R_n(\lambda_1, \lambda_2)$. In order to do this we introduce a modified version of the generalized cross validation theorem.

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

THEOREM 4.1. *Let* $\tau_j(\lambda_1,\lambda_2) = n^{-1}\operatorname{tr}(H_n(\lambda_1/1+\lambda_2)^j)$, $j=1,2$, *and assume that* $\tau_1(\lambda_1,\lambda_2) \le 1$. *Then*,

$$\left| \frac{E[GCV_n(\lambda_1,\lambda_2)] - \sigma^2 - R_n(\lambda_1,\lambda_2)}{R_n(\lambda_1,\lambda_2)} \right| \le h(\lambda_1,\lambda_2), \tag{26}$$

where

$$\begin{aligned} h(\lambda_1,\lambda_2) = {} & \frac{2\tau_1^2(\lambda_1,\lambda_2)}{(1-\tau_1(\lambda_1,\lambda_2))^2} + \frac{1+\lambda_2}{(1-\tau_1(\lambda_1,\lambda_2))^2} \\ & \times \left| 2\lambda_2 \frac{\tau_1(\lambda_1,\lambda_2)}{\tau_2(\lambda_1,\lambda_2)} - (1+\lambda_2)\frac{\tau_1^2(\lambda_1,\lambda_2)}{\tau_2(\lambda_1,\lambda_2)} \right| \\ & + \frac{2n^{-1}\lambda_2}{(1+\lambda_2)(1-\tau_1(\lambda_1,\lambda_2))^2} \\ & \times \left| E\left[\epsilon^T H_n\left(\frac{\lambda_1}{1+\lambda_2} \right)(g(\hat{\theta}_n) - g(\theta_0)) \right] \right|. \end{aligned}$$

If $h(\cdot)$ is small, Theorem 4.1 implies that the distance between the risk $R_n(\lambda_1,\lambda_2)$ and the expected value of the generalized cross validation function is small relative to the intrinsic accuracy measure of the risk. This can be roughly interpreted as if $GCV_n(\lambda_1,\lambda_2)$ is an unbiased estimator of $\sigma^2 + R_n(\lambda_1,\lambda_2)$, and, therefore, the values of λ_1 and λ_2 that minimize both criteria will be asymptotically the same.

We rigorize this conjecture in the following theorem.

THEOREM 4.2. *Let* λ_1^* *and* λ_2^* *be minimizers of*

$$R_n(\lambda_1,\lambda_2) = n^{-1}\sum_{i=1}^{n} E(\hat{m}_\lambda(x_i) - m(x_i))^2 \tag{27}$$

and $\tilde{\lambda}_1$ *and* $\tilde{\lambda}_2$ *the minimizers of* $E[GCV_n(\lambda_1,\lambda_2)]$. *Then under Assumptions* A.1–A.7

$$\frac{R_n(\tilde{\lambda}_1,\tilde{\lambda}_2)}{R_n(\lambda_1^*,\lambda_2^*)} \to_p 1 \tag{28}$$

as n tends to infinity.

From Theorems 4.1 and 4.2 we realize that to get asymptotic optimality, it is necessary that the smoothing parameter λ_2 tend to zero as the sample size increases. This avoids the case when our estimator achieves the parametric optimal rate of convergence (this rate is achieved when the parametric model is correctly specified and among other assumptions λ_2 tends to infinity). This result is already known, and it basically means that the optimal rate of convergence of the minimal expected error $\min_{\lambda_1,\lambda_2} R_n(\lambda_1,\lambda_2)$, must be slower than $1/n$. Li (1986) pointed out that without this requirement it seems that no selection procedure can be

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

asymptotically optimal. For many problems this condition is equivalent to the condition that $\hat{m}_\lambda$ is not infinitely smooth.

Another question that remains open is whether the asymptotic results claimed in Theorems 3.2 and 3.3 remain valid when the penalty parameters λ_1 and λ_2 are estimated by the generalized cross validation method. From Theorem 4.3 it follows that the constrained smoothing spline remains consistent and has the same rates of convergence as in the theoretical case. However, it is not clear whether the asymptotic normality goes through.

THEOREM 4.3. *Let $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ be the minimizers of $E[GCV_n(\lambda_1, \lambda_2)]$. Under Assumptions* A.1–A.7,

$$L_n(\tilde{\lambda}_1, \tilde{\lambda}_2) = O_p(n^{-2\nu/(2\nu+1)})$$

as n tends to infinity.

5. APPLICATION AND SIMULATION STUDY

As an application, we have considered the estimation of a model of investment behavior in the U.S. telephone industry over the period 1949 to 1968 (see Sankar, 1973). To estimate the appropriate relationship between the desired capital stock and other variables, it is necessary to assume a specific production function, e.f. Cobb–Douglas, C.E.S., or others. If we assume that the production in the industry is characterized by the Cobb–Douglas production function the desired stock of capital at year t, K_t^+, can be expressed as

$$K_t^+ = \alpha \left(\frac{PQ}{c} \right)_t \qquad t = 1, \ldots, T, \tag{29}$$

where α is the elasticity of output with respect to capital, P is the price of final output, Q is the production, and c is the rental price of capital. Assuming the observed capital stock K_t is a function of the desired capital stock, the following model is estimated:

$$K_t = \alpha_0 + \alpha_1 \left(\frac{PQ}{c} \right)_t + \epsilon_t \qquad t = 1, \ldots, T. \tag{30}$$

The linear relationship between K_t and $(PQ/c)_t$ is recommended by the underlying specification of the Cobb–Douglas production function. However, it would be much more robust against misspecifications in the functional form to estimate nonparametrically the preceding relationship. We combine the two approaches by estimating nonparametrically the relationship between the observed capital stock and (PQ/c) and by keeping the underlying Cobb–Douglas approach by imposing the shape constraint of linearity. Following the procedure introduced in Section 2, we have first computed the ordinary least squares (OLS) estimation of α_0 and α_1, $\hat{\alpha}_0 = -14{,}695.71$ and $\hat{\alpha}_1 = 0.43387$. Second, we have computed the

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

estimator $\hat{m}_\lambda$ with values for λ_1 and λ_2 estimated by generalized cross validation. Figure 1 plots the data together with the linear OLS fit, the smoothing spline constrained to the linear parametric function, and the confidence bands constructed by the method proposed in Section 3. The results suggest that another specification different to the linear form should be adopted.

Finally, we carried out a simulation study to analyze the small sample behavior of the constrained smoothing spline estimator and the estimated values of the penalty parameters. The following errors have been empirically evaluated:

$$E_{1n} = \frac{1}{n}\sum_{i=1}^{n}[\hat{m}_{\lambda_1,\lambda_2}(x_i;\hat{\theta}_n) - m(x_i)]^2, \tag{31}$$

$$E_{2n} = \frac{1}{n}\sum_{i=1}^{n}[g(x_i;\hat{\theta}_n) - m(x_i)]^2, \tag{32}$$

and

$$E_{3n} = \frac{1}{n}\sum_{i=1}^{n}[\hat{m}_\lambda(x_i;\hat{\theta}_n) - m(x_i)]^2. \tag{33}$$

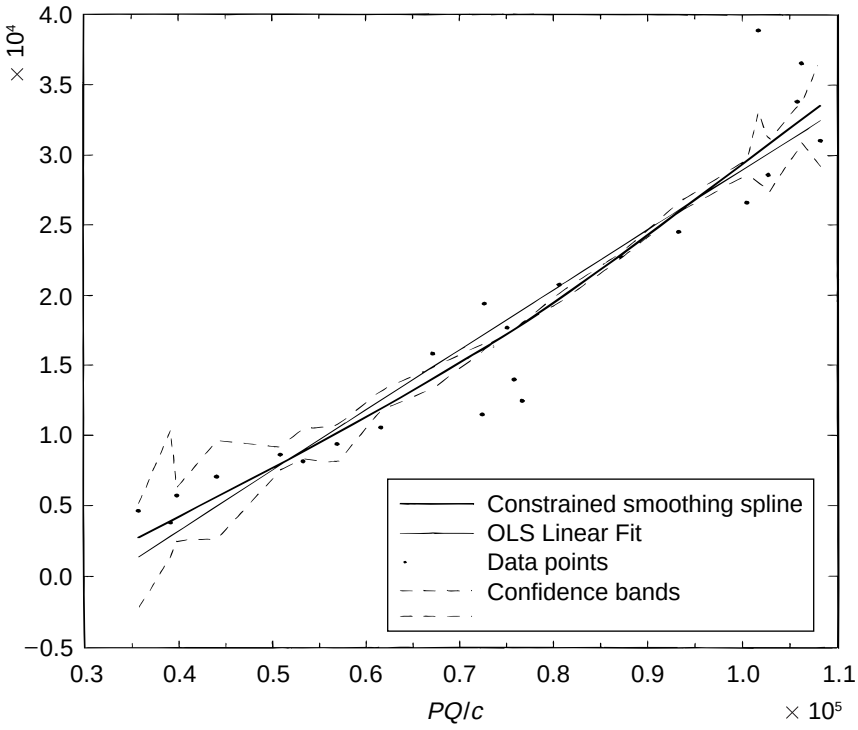


FIGURE 1. Constrained smoothing splines and confidence bands.

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

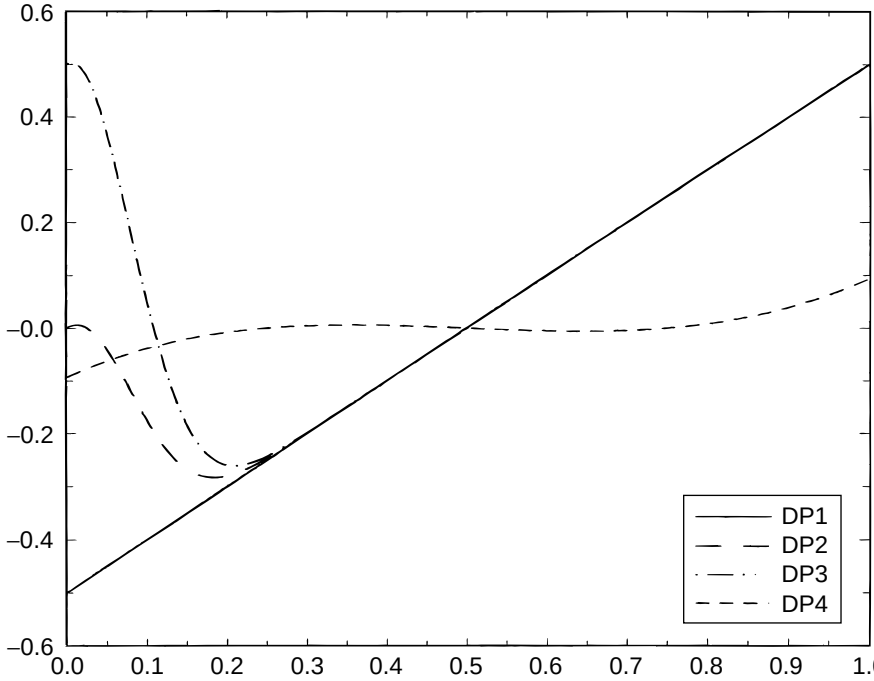


FIGURE 2. Data generating processes.

The term E_{1n} represents the risk function for the constrained smoothing spline estimator of penalty parameters λ_1 and λ_2, E_{2n} is the risk for the nonlinear least squares estimator, and, finally, E_{3n} represents the risk for the natural cubic smoothing spline with penalty parameter λ. We have considered the following data generating processes.

TABLE 1. Empirical errors under DP1

	E_{1n}			E_{2n}			E_{3n}		
Sample size	10%	50%	90%	10%	50%	90%	10%	50%	90%
$n = 50$	0.021	0.049	0.068	0.043	0.072	0.164	0.017	0.058	0.096
$n = 100$	0.003	0.017	0.035	0.006	0.015	0.102	0.005	0.019	0.054
$n = 200$	—	0.007	0.012	—	0.005	0.096	—	0.009	0.023
$n = 500$	—	0.003	0.016	—	0.002	0.083	—	0.006	0.026
$n = 1{,}000$	—	0.002	0.011	—	0.001	0.062	—	0.003	0.018

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

TABLE 2. Empirical errors under DP2

	E_{1n}			E_{2n}			E_{3n}		
Sample size	10%	50%	90%	10%	50%	90%	10%	50%	90%
$n = 50$	0.014	0.056	0.081	0.047	0.079	0.180	0.037	0.067	0.089
$n = 100$	0.021	0.052	0.068	0.007	0.016	0.112	0.040	0.051	0.078
$n = 200$	0.018	0.044	0.051	0.005	0.009	0.144	0.031	0.040	0.105
$n = 500$	0.018	0.032	0.043	—	0.009	0.035	0.021	0.027	0.058
$n = 1{,}000$	0.001	0.019	0.028	—	0.001	0.072	0.007	0.013	0.020

DP1. $m(x) = x - 0.5$.

DP2. $m(x) = x - 0.5 + 0.25\exp(-80x^2)$.

DP3. $m(x) = x - 0.5 + \exp(-80x^2)$.

DP4. $m(x) = (x - 0.25)(x - 0.5)(x - 0.75)$.

DP1 is a standard linear model, and DP2 and DP3 are used in Gasser, Sroka, and Jennen-Steinmetz (1986). This family of functions represents a unimodal departure from linearity, where the magnitude of this departure is controlled by the coefficient associated to the exponential function (0.25 in DP2 and 1.0 in DP3). Finally, DP4 was proposed by Rice and Rosenblatt (1981). The observations are generated

$$y_i = m(x_i) + \epsilon_i \tag{34}$$

with $m(\cdot)$ being either DP1, DP2, DP3, or DP4 and where the errors are independent $N(0,\frac{1}{4})$. In Figure 2, we show the functions related to the different data generating processes. For each sample size ($n = 50$, 100, 200, 500, 1,000) the observations x_i are equally spaced on the interval (0,1). We make 1,000 replica-

TABLE 3. Empirical errors under DP3

	E_{1n}			E_{2n}			E_{3n}		
Sample size	10%	50%	90%	10%	50%	90%	10%	50%	90%
$n = 50$	0.016	0.063	0.091	0.053	0.081	0.114	0.047	0.084	0.112
$n = 100$	0.024	0.058	0.076	0.036	0.061	0.106	0.051	0.064	0.098
$n = 200$	0.020	0.049	0.057	0.011	0.047	0.062	0.039	0.051	0.132
$n = 500$	0.012	0.036	0.045	0.009	0.024	0.051	0.027	0.034	0.073
$n = 1{,}000$	0.003	0.020	0.031	0.023	0.054	0.088	0.009	0.017	0.026

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

TABLE 4. Empirical errors under DP4

	E_{1n}			E_{2n}			E_{3n}		
Sample size	10%	50%	90%	10%	50%	90%	10%	50%	90%
$n = 50$	0.076	0.186	0.233	0.271	0.393	0.422	0.116	0.201	0.299
$n = 100$	0.101	0.183	0.221	0.199	0.226	0.301	0.221	0.170	0.098
$n = 200$	0.099	0.179	0.200	0.368	0.300	0.246	0.122	0.172	0.212
$n = 500$	0.013	0.084	0.111	0.152	0.214	0.299	0.003	0.098	0.128
$n = 1{,}000$	—	0.036	0.089	0.175	0.199	0.214	—	0.041	0.083

tions. The penalty parameters for E_{1n} and E_{3n} have been calculated by generalized cross validation, and the parametric function was

$$g(x,\theta) = x - 0.5. \tag{35}$$

The results of estimating the empirical errors under the different data generating processes are shown in Tables 1–4. We show the median, the 10th, and the 90th percentiles of the empirical distribution of the different errors from the simulation.

We now investigate the size of the empirical errors under the different data generating processes. The case DP1 represents a shrinkage of $\hat{m}_{\lambda_1\lambda_2}$ toward the linear model when this is in fact the true model. As could be expected, the parametric estimator presents the best performance in terms of empirical errors, followed by the constrained smoothing spline and finally the natural cubic spline. The same results hold for slight departures from linearity. In DP2, when we shrink our estimator toward a linear function being the true model slightly apart from linearity, the constrained smoothing spline estimator presents the best performance for almost all sample sizes. However, when the departure from linearity is stronger as in DP3 and DP4 the natural smoothing spline dominates in terms of empirical errors. These results can be related to shrinkage estimators in paramet-

TABLE 5. Estimated penalty parameters under DP1

	$\hat{\lambda}_1$			$\hat{\lambda}_2$			$\hat{\lambda}$		
Sample size	10%	50%	90%	10%	50%	90%	10%	50%	90%
$n = 50$	2.901	3.882	4.231	3.023	4.237	5.001	5.994	5.121	3.913
$n = 100$	0.937	1.495	2.014	3.002	3.642	4.146	3.999	4.217	5.526
$n = 200$	0.925	1.251	1.988	2.842	3.726	4.273	2.487	3.911	4.003
$n = 500$	0.017	0.232	0.745	3.978	4.452	5.001	2.449	3.016	3.649
$n = 1{,}000$	0.001	0.221	0.392	3.593	4.831	5.103	2.132	2.561	3.221

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

TABLE 6. Estimated penalty parameters under DP2

	$\hat{\lambda}_1$			$\hat{\lambda}_2$			$\hat{\lambda}$		
Sample size	10%	50%	90%	10%	50%	90%	10%	50%	90%
$n = 50$	3.145	3.634	3.924	1.729	2.328	2.834	4.311	4.731	5.001
$n = 100$	1.572	1.829	2.441	2.000	1.669	1.387	2.599	2.989	3.998
$n = 200$	1.201	1.659	2.001	1.101	1.226	1.661	1.535	2.173	2.888
$n = 500$	0.286	0.669	0.997	0.105	0.274	0.302	0.583	1.209	1.799
$n = 1{,}000$	0.113	0.412	0.585	0.026	0.101	0.384	0.009	0.389	0.804

ric models. In this context, James and Stein (1961) exhibited some slightly biased estimators that might have superior mean squared errors when compared to unbiased least squares ones. This result was true only when the model one is shrinking toward is close to the true specification; otherwise, the mean squared error of these shrinkage estimators is inferior to the least squares type.

The adaptation of our shrinkage estimator according to the specification of the model is also shown in the simulations. When the true model is close to the model that we are shrinking toward, then the error of the constrained smoothing spline is close to the one in the parametric estimator, whereas if the true model is far from the shrinkage, then the error becomes closer to the one in the natural smoothing spline.

In Tables 5–8 we present the median, the 10th, and the 90th percentiles of the empirical distribution of the estimated penalty parameters from the simulation. When estimating under DP1, the descriptive statistics for $\hat{\lambda}_1$ and $\hat{\lambda}_2$ show large values, and as could be expected from the theoretical results shown in Theorem 3.2, the penalty parameter$\hat{\lambda}_2$ seems to increase with the sample size. The parameter $\hat{\lambda}_1$ tends to zero. In the natural cubic smoothing spline, because the true model is linear, we get large values for $\hat{\lambda}$, although as could be expected they decrease in average with the sample size. Under DP2 and DP3, the descriptive

TABLE 7. Estimated penalty parameters under DP3

	$\hat{\lambda}_1$			$\hat{\lambda}_2$			$\hat{\lambda}$		
Sample size	10%	50%	90%	10%	50%	90%	10%	50%	90%
$n = 50$	3.011	3.421	3.732	1.831	2.338	2.783	4.212	4.621	4.943
$n = 100$	1.638	1.972	2.243	2.123	1.722	1.312	2.691	3.170	4.092
$n = 200$	1.199	1.643	1.999	0.945	1.220	1.646	1.572	2.001	2.833
$n = 500$	0.232	0.726	1.001	0.099	0.252	0.498	0.444	1.102	1.841
$n = 1{,}000$	0.097	0.334	0.699	0.011	0.232	0.491	0.011	0.419	0.899

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

TABLE 8. Estimated penalty parameters under DP

	$\hat{\lambda}_1$			$\hat{\lambda}_2$			$\hat{\lambda}$		
Sample size	10%	50%	90%	10%	50%	90%	10%	50%	90%
$n = 50$	2.202	2.538	2.841	2.231	1.729	1.401	3.101	3.772	4.227
$n = 100$	1.701	2.043	2.299	1.927	1.488	1.002	1.851	2.372	2.971
$n = 200$	1.103	1.419	1.871	0.422	0.941	1.392	1.461	1.901	2.319
$n = 500$	0.513	0.818	1.132	0.473	0.736	1.354	0.295	0.612	0.889
$n = 1{,}000$	0.483	0.681	0.999	0.001	0.131	0.200	0.215	0.307	0.377

statistics for $\hat{\lambda}_1$ and $\hat{\lambda}_2$ show smaller and decreasing values, whereas the statistics for $\hat{\lambda}$ present the expected behavior. It is interesting to remark that as $\hat{\lambda}_2$ tends to zero with the sample size, $\hat{\lambda}_1$ and $\hat{\lambda}$ become closer. This effect is confirmed under DP4, where again $\hat{\lambda}_2$ tends to zero and $\hat{\lambda}_1$ and $\hat{\lambda}$ present the same performance as in DP3.

6. CONCLUSION

The method presented in this paper provides a link between parametric and nonparametric regression models. It allows us to shrink the nonparametric smoothing spline estimator toward a prespecified parametric model. The resulting hybrid estimator is of great interest in situations when there is uncertainty about model specification but there is some prior knowledge that could be included in the estimation procedure.

There remains some work to be done. Although the two-step procedure has nice theoretical properties, it would be more interesting to have a procedure that simultaneously estimates the parametric and the nonparametric components.

REFERENCES

Ansley, C.F., R. Kohn, & C. Wong (1993) Nonparametric spline regression with prior information. *Biometrika* 80, 75–88.

Bickel, P. & K.A. Doksum (1977) *Mathematical Statistics: Basic Ideas and Selected Topics.* San Francisco: Holden-Day.

Burman, P. & P. Chaudhuri (1992) A Hybrid Approach to Parametric and Nonparametric Regression. Manuscript, Department of Statistics, University of California, Davis.

Cox, D.D. (1984) Gaussian Approximation of Smoothing Splines. Technical report 743, Department of Statistics, University of California-Berkeley.

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

Eubank, R. (1988) *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker.

Fan, J. & I. Gijbels (1996) *Local Polynomial Modeling and Its Applications.* New York: Chapman and Hall.

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

Gasser, T., L. Sroka, & C. Jennen-Steinmetz (1986) Residual variance and residual pattern in nonlinear regression. *Biometrika* 73, 625–633.
Green, P. (1987) Penalized likelihood for general semi-parametric regression models. *International Statistical Review* 55, 245–259.
Härdle, W. (1990) *Applied Nonparametric Regression.* New York: Cambridge University Press.
Heckman, N. (1986) Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society, Series B.* 48, 244–248.
James, W. & C. Stein (1961) Estimation with quadratic loss. In J. Neyman (ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 361–379. Berkeley: University of California Press.
Kohn, R. & C.F. Ansley (1991) The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the American Statistical Association* 86, 1042–1050.
Li, K.C. (1986) Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Annals of Statistics* 14, 1101–1112.
Olkin, I. & C. Spiegelman (1987) A semiparametric approach to density estimation. *Journal of the American Statistical Association* 82, 858–865.
Politis, D.N. & J.P. Romano (1994) Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics* 22, 2031–2050.
Rice, J. (1986) Convergence rates for partially splined models. *Statistics and Probability Letters* 4, 203–208.
Rice, J. & M. Rosenblatt (1981) Integrated mean square error of a smoothing spline. *Journal of Approximation Theory* 33, 353–369.
Rodriguez Poo, J.M. (1992) *Constrained Nonparametric Regression.* Ph.D. Thesis, Catholic University of Louvain, Louvain-la-Neuve, Belgium.
Sankar, U. (1973) Investment behavior in the U.S. Telephone Industry—1949 to 1968. *Bell Journal of Economic and Management Science* 4, 665–678.
Speckman, P. (1985) Spline smoothing and optimal rates of convergence in nonparametric regression models. *Annals of Statistics* 13, 970–983.
Speckman, P. (1988) Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B* 50, 413–436.
Stone, C.J. (1982) Optimal global rates of convergence for nonparametric estimation. *Annals of Statistics* 10, 1040–1053.
Wahba, G. (1985) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics* 11, 1378–1402.
Wahba, G. (1990) *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics 59.

APPENDIX A: PROOF OF THEOREM 2.1

Previous to the proof of Theorem 2.1 we need the following result.

LEMMA 1. *Under the following equalities and definitions*:

$$L_n(m;\theta) = n^{-1}(y-m)^T W_1 (y-m) + n^{-1}(m-g(\theta))^T W_2 (m-g(\theta)) + \lambda_1 \int_0^1 (m^{(\nu)}(t))^2\, dt,$$

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

$$L_n^*(m) = n^{-1}(y_* - m)^T W_*(y_* - m) + \lambda_1 \int_0^1 (m^{(\nu)}(t))^2\, dt,$$

with

$$y = (y_1, y_2, \ldots, y_n)^T,$$

$$m = (m(x_1), m(x_2), \ldots, m(x_n))^T,$$

and

$$g(\theta) = (g(x_1;\theta), g(x_2;\theta), \ldots, g(x_n;\theta))^T,$$

$$W_* = W_1 + W_2,$$

$$y_* = W_*^{-1}(W_1 y + W_2 g(\theta)). \tag{A.1}$$

And for fixed $\lambda_1 > 0$ *and* $\lambda_2 > 0$ *then the problem*

$$m = \arg \min_{m \in W_2^{\nu}[0,1]} L_n(m) \tag{A.2}$$

can be calculated via

$$m = \arg \min_{m \in W_2^{\nu}[0,1]} L_n^*(m). \tag{A.3}$$

Proof. Let us redefine

$$L_n(m;\theta) = n^{-1}(y - m)^T W_1 (y - m) + n^{-1}(m - g(\theta))^T W_2 (m - g(\theta)) + \lambda_1 \int_0^1 (m^{(\nu)}(t))^2\, dt$$

as

$$L_n(m;\theta) = S_n(m, g(\theta)) + \lambda_1 \int_0^1 (m^{(\nu)}(t))^2\, dt. \tag{A.4}$$

We develop the squared terms in $S_n(m, g(\theta))$ obtaining

$$S_n(m, g(\theta)) = n^{-1} m^T (W_1 + W_2) m - 2n^{-1} m^T (W_1 + W_2)(W_1 + W_2)^{-1}(W_1 y + W_2 g(\theta)) + n^{-1}(W_1 y + W_2 g(\theta))^T (W_1 + W_2)^{-1}(W_1 y + W_2 g(\theta)) + C_n, \tag{A.5}$$

where

$$C_n = n^{-1} y^T W_1 y + n^{-1} g(\theta)^T W_2 g(\theta) - n^{-1}(W_1 y + W_2 g(\theta))^T (W_1 + W_2)^{-1} \times (W_1 y + W_2 g(\theta))$$

does not depend on $m(\cdot)$. Now if we use that

$$W_* = W_1 + W_2 \tag{A.6}$$

$$y_* = W_*^{-1}(W_1 y + W_2 g(\theta)), \tag{A.7}$$

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

we obtain

$$S_n(m, g(\theta)) = n^{-1}(y_* - m)^T W_*(y_* - m) + C_n \tag{A.8}$$

and

$$L_n(m;\theta) = L_n^*(m;\theta) + C_n, \tag{A.9}$$

where

$$L_n^*(m;\theta) = n^{-1}(y_* - m)^T W_*(y_* - m) + \lambda_1 \int_0^1 (m^{(\nu)}(t))^2\, dt. \tag{A.10}$$

Because C_n does not depend on $m(\cdot)$ the proof is finished. ■

Let $W_1 = I_n$ and $W_2 = \lambda_2 I_n$. The proof of Theorem 2.1 follows from Lemma 1 and Theorem 5.3 from Eubank (1988). ■

APPENDIX B: PROOF OF THEOREM 3.1

Note that

$$\begin{aligned}
\|\hat{m}_\lambda - m\|_n^2 &= \left\| \frac{1}{1+\lambda_2}(H_n y - m) + \frac{\lambda_2}{1+\lambda_2}(H_n g(\hat{\theta}_n) - m) \right\|_n^2 \qquad \text{(B.1)} \\
&= \left\| \frac{1}{1+\lambda_2}(H_n y - m) + \frac{\lambda_2}{1+\lambda_2} \right. \\
&\quad \left. \times \{(H_n g(\hat{\theta}_n) - H_n g(\theta_0)) + (H_n g(\theta_0) - g(\theta_0)) + (g(\theta_0) - m)\} \right\|_n^2 \\
&\le 2\left(\frac{1}{1+\lambda_2}\right)^2 \|H_n y - m\|_n^2 \\
&\quad + 2\left(\frac{\lambda_2}{1+\lambda_2}\right)^2 \\
&\quad \times \{\|H_n g(\hat{\theta}_n) - H_n g(\theta_0)\|_n^2 + \|H_n g(\theta_0) - g(\theta_0)\|_n^2 + \|g(\theta_0) - m\|_n^2\} \\
&\equiv 2\left(\frac{1}{1+\lambda_2}\right)^2 I_1 + 2\left(\frac{1}{1+\lambda_2}\right)^2 (I_2 + I_3 + I_4). \qquad \text{(B.2)}
\end{aligned}$$

From Assumptions A.1, A.2, and A.5 it already has been shown that (see Speckman, 1985) if $\lambda_1/(1+\lambda_2) \to 0$, and $n(\lambda_1/1+\lambda_2) \to \infty$ then

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

$$I_1 = O_p\left(\frac{\lambda_1}{1+\lambda_2}\right) + O_p\left(n^{-1}\left(\frac{\lambda_1}{1+\lambda_2}\right)^{-1/2\nu}\right) \tag{B.3}$$

as n tends to infinity.

Assumptions A.3–A.7 imply that

$$\|g(\hat{\theta}_n) - g(\theta_0)\|_n^2 = O_p(n^{-1}). \tag{B.4}$$

Let $\lambda_i(H_n^T H_n)$ be the ith eigenvalue of the matrix $H_n^T H_n$. Given that

$$\lambda_i(H_n^T H_n) = \begin{cases} 1 & i = 1,\ldots,\nu \\ \dfrac{1}{\left(1 + \dfrac{\lambda_1}{1+\lambda_2}\, n\eta_{in}\right)^2} & i = \nu+1,\ldots,n \end{cases} \tag{B.5}$$

(see Eubank, 1988) and

$$n\eta_{in} = (\pi i)^{2\nu}\left(\int_0^1 p(t)^{1/2\nu}\, dt\right)^{-2\nu}(1 + o(1)), \tag{B.6}$$

which is found to hold uniformly over $i = o(n^{2/(2\nu+1)})$ (see Speckman, 1985), then $\lambda_{\max}(H_n^T H_n) = 1$ and

$$I_2 \leq \lambda_{\max}(H_n^T H_n)\|g(\hat{\theta}_n) - g(\theta_0)\|_n^2 = O_p\left(\frac{1}{n}\right). \tag{B.7}$$

Assumption A.3 together with Lemma 4.3 from Craven and Wahba (1979) is enough to show that

$$I_3 = O\left(\frac{\lambda_1}{1+\lambda_2}\right) \tag{B.8}$$

as n tends to infinity, and finally $I_4 = \delta_n$. This completes the proof. ■

APPENDIX C: PROOF OF THEOREM 3.3

For the proof of this theorem, we rely on the convergence in distribution results of the natural smoothing spline estimator shown by Cox (1984) and Eubank (1988). Recall that $\hat{m}_\lambda(x)$ can be written as

$$\hat{m}_\lambda(x) = \frac{1}{1+\lambda_2}\,\hat{S}_\lambda(x) + \frac{\lambda_2}{1+\lambda_2}\,\hat{r}_\lambda(x) \tag{C.1}$$

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

for any $x \in [0,1]$. Here $\hat{S}_\lambda(x)$ and $\hat{r}_\lambda(x)$ are, respectively, the natural smoothing spline fits to the y and the $g(\hat{\theta}_n)$ values at grid point x. Let us write

$$\hat{m}_\lambda(x) - m(x) = \frac{1}{1+\lambda_2}(\hat{S}_\lambda(x) - E[\hat{S}_\lambda(x)]) + \frac{1}{1+\lambda_2}(E[\hat{S}_\lambda(x)] - m(x))$$
$$+ \frac{\lambda_2}{1+\lambda_2}(\hat{r}_\lambda(x) - E[\hat{r}_\lambda(x)]) + \frac{\lambda_2}{1+\lambda_2}(E[\hat{r}_\lambda(x)] - m(x)) \qquad \textbf{(C.2)}$$

and

$$E[\hat{r}_\lambda(x)] - m(x) = E[\hat{r}_\lambda(x)] - r_\lambda(x) + r_\lambda(x) - g(x;\theta_0) + g(x;\theta_0) - m(x). \qquad \textbf{(C.3)}$$

Note also that according to the notation, $E[\hat{S}_\lambda(x)]$, $E[\hat{r}_\lambda(x)]$, and $r_\lambda(x)$ are the natural smoothing spline fits to the values of m, $E[g(x;\hat{\theta}_n)]$, and $g(\theta_0)$, respectively, all evaluated at a grid point x.

Assuming A.1–A.7 and relying on the results that we have already shown in Theorem 3.2, if $\lambda_1/(1+\lambda_2) \to 0$ and $n(\lambda_1/(1+\lambda_2)) \to \infty$ then as n tends to infinity

$$E(\hat{m}_\lambda(x) - E[\hat{m}_\lambda(x)])^2 = O\left(n^{-1}\left(\frac{\lambda_1}{1+\lambda_2}\right)^{-1/2\nu}\right),$$
$$E[\hat{S}_\lambda(x)] - m(x) = O\left(\frac{\lambda_1}{1+\lambda_2}\right),$$
$$\hat{r}_\lambda(x) - E[\hat{r}_\lambda(x)] = O_p(1), \qquad \textbf{(C.4)}$$
$$E[\hat{r}_\lambda(x)] - r_\lambda(x) = O_p(1),$$
$$r_\lambda(x) - g(x;\theta_0) = O\left(\frac{\lambda_1}{1+\lambda_2}\right),$$
$$\sup_x |g(x;\theta_0) - m(x)| < \delta_n.$$

Dividing all terms by $\sigma(x) = \sqrt{E(\hat{m}_\lambda(x) - E[\hat{m}_\lambda(x)])^2}$, then under the conditions previously imposed we have that

$$\frac{E[\hat{S}_\lambda(x)] - m(x)}{\sigma(x)} = O\left(n^{1/2}\left(\frac{\lambda_1}{1+\lambda_2}\right)^{(4\nu+1)/4\nu}\right),$$
$$\frac{\hat{r}_\lambda(x) - E[\hat{r}_\lambda(x)]}{\sigma(x)} = O\left(n^{1/2}\left(\frac{\lambda_1}{1+\lambda_2}\right)^{1/4\nu}\right), \qquad \textbf{(C.5)}$$
$$\frac{E[\hat{r}_\lambda(x)] - r_\lambda(x)}{\sigma(x)} = O\left(n^{1/2}\left(\frac{\lambda_1}{1+\lambda_2}\right)^{1/4\nu}\right),$$
$$\frac{r_\lambda(x) - g(x;\theta_0)}{\sigma(x)} = O\left(n^{1/2}\left(\frac{\lambda_1}{1+\lambda_2}\right)^{(4\nu+1)/4\nu}\right),$$

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

as n tends to infinity. Under the rates of decrease established in the theorem for λ_1 and λ_2 then

$$\frac{E[\hat{S}_\lambda(x)] - m(x)}{\sigma(x)} = O(n^{-\nu/(2\nu+1)}),$$

$$\frac{\hat{r}_\lambda(x) - E[\hat{r}_\lambda(x)]}{\sigma(x)} = \frac{E[\hat{r}_\lambda(x)] - r_\lambda(x)}{\sigma(x)} = O_p(n^{\nu/(2\nu+1)}), \tag{C.6}$$

and

$$\frac{r_\lambda(x) - g(x;\theta_0)}{\sigma(x)} = O(n^{\nu/(2\nu+1)}). \tag{C.7}$$

It is also easy to check that

$$\frac{g(x;\theta_0) - m(x)}{\sigma(x)} = O(n^{\nu/(2\nu+1)}). \tag{C.8}$$

The proof is completed because under the preceding conditions on the rate of decrease for λ_1 and λ_2, as n tends to infinity,

$$\frac{n^{4\nu/5}}{\log n}\left(\frac{\lambda_1}{1+\lambda_2}\right) \to \infty.$$

Therefore Theorem O.1 from Cox (1984, p. 6) holds; then

$$\frac{\hat{S}_\lambda(x) - E[\hat{S}_\lambda(x)]}{\sigma(x)} \to_d N(0,1), \tag{C.9}$$

and the result is proved. ■

APPENDIX D: PROOF OF THEOREM 3.4

Under conditions A.1–A.7, from Theorem 3.1 $\sigma^2(x) = O(n^{-1}(\lambda_1/(1+\lambda_2))^{-1/2\nu})$ as n tends to infinity. Then, if $\lambda_1 \sim n^{-2\nu/(2\nu+1)}$ and $n^{2\nu/(2\nu+1)}\lambda_2 \to 0$ we obtain $\gamma_n = n^{\nu/(2\nu+1)}$. Theorem 3.3 implies Assumption A of Politis and Romano (1994). Moreover, if $b \to \infty$ such that $b/n \to 0$, as n tends to infinity then $\gamma_b/\gamma_n \to 0$. Theorem 2.1 from Politis and Romano (1994) applies, and the proof is done. ■

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

APPENDIX E: PROOF OF THEOREM 4.1

Using (18), (19), and (25) we show that

$$E[GCV_n(\lambda_1,\lambda_2)] = \frac{R_n(\lambda_1,\lambda_2)}{(1-\tau_1(\lambda_1,\lambda_2))^2} + \frac{\sigma^2}{(1-\tau_1(\lambda_1,\lambda_2))^2} - \frac{2n^{-1}E[\epsilon^T(\hat{m}_\lambda - m)]}{(1-\tau_1(\lambda_1,\lambda_2))^2}.$$

Thus in view of

$$n^{-1}E[\epsilon^T(\hat{m}_\lambda - m)] = \frac{1}{1+\lambda_2}\left\{\sigma^2\tau_1(\lambda_1,\lambda_2) + \lambda_2 E\left[\epsilon^T H_n\left(\frac{\lambda_1}{1+\lambda_2}\right)(g(\hat{\theta}_n) - g(\theta))\right]\right\}$$

then

$$\begin{aligned}
&E[GCV_n(\lambda_1,\lambda_2)] - \sigma^2 - R_n(\lambda_1,\lambda_2) \\
&\quad = \frac{\tau_1(\lambda_1,\lambda_2)[2-\tau_1(\lambda_1,\lambda_2)]}{(1-\tau_1(\lambda_1,\lambda_2))^2} R_n(\lambda_1,\lambda_2) \\
&\qquad + \frac{\sigma^2}{(1+\lambda_2)(1-\tau_1(\lambda_1,\lambda_2))^2}[2\lambda_2\tau_1(\lambda_1,\lambda_2) - (1+\lambda_2)\tau_1(\lambda_1,\lambda_2)^2] \\
&\qquad - \frac{2n^{-1}\lambda_2}{(1+\lambda_2)(1-\tau_1(\lambda_2,\lambda_2))^2} E\left[\epsilon^T H_n\left(\frac{\lambda_1}{1+\lambda_2}\right)(g(\hat{\theta}_n) - g(\theta))\right]. \qquad \textbf{(E.1)}
\end{aligned}$$

This implies that

$$\begin{aligned}
&\left|\frac{E[GCV_n(\lambda_1,\lambda_2)] - \sigma^2 - R_n(\lambda_1,\lambda_2)}{R_n(\lambda_1,\lambda_2)}\right| \\
&\quad \le \left|\frac{\tau_1(\lambda_1,\lambda_2)[2-\tau_1(\lambda_1,\lambda_2)]}{(1-\tau_1(\lambda_1,\lambda_2))^2}\right| + \frac{\sigma^2}{R_n(\lambda_1,\lambda_2)(1+\lambda_2)(1-\tau_1(\lambda_1,\lambda_2))^2} \\
&\qquad \times |2\lambda_2\tau_1(\lambda_1,\lambda_2) - (1+\lambda_2)\tau_1(\lambda_1,\lambda_2)^2| + \frac{2n^{-1}\lambda_2}{(1+\lambda_2)(1-\tau_1(\lambda_1,\lambda_2))^2} \\
&\qquad \times \left|E\left[\epsilon^T H_n\left(\frac{\lambda_1}{1+\lambda_2}\right)(g(\hat{\theta}_n) - g(\theta))\right]\right|. \qquad \textbf{(E.2)}
\end{aligned}$$

Then, using the following two inequalities,

$$1 \le 2 - \tau_1(\lambda_1,\lambda_2) \le 2,$$

$$R_n(\lambda_1,\lambda_2) \ge \left(\frac{1}{1+\lambda_2}\right)^2 \sigma^2\tau_2(\lambda_1,\lambda_2), \qquad \textbf{(E.3)}$$

it is possible to show that

$$\left|\frac{E[GCV_n(\lambda_1,\lambda_2)] - \sigma^2 - R_n(\lambda_1,\lambda_2)}{R_n(\lambda_1,\lambda_2)}\right| \le h(\lambda_1,\lambda_2). \qquad \textbf{(E.4)}$$

■

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

APPENDIX F: PROOF OF THEOREM 4.2

If λ_1^* and λ_2^* are the minimizers of $R_n(\lambda_1, \lambda_2)$ and $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ are the minimizers of $E[GCV_n(\lambda_1, \lambda_2)]$ then as a result of Theorem 4.1

$$E[GCV_n(\lambda_1^*, \lambda_2^*)] - \sigma^2 \leq R_n(\lambda_1^*, \lambda_2^*)(1 + h_n(\lambda_1^*, \lambda_2^*)) \tag{F.1}$$

and

$$E[GCV_n(\tilde{\lambda}_1, \tilde{\lambda}_2)] - \sigma^2 \leq R_n(\tilde{\lambda}_1, \tilde{\lambda}_2)(1 + h_n(\tilde{\lambda}_1, \tilde{\lambda}_2)). \tag{F.2}$$

Combining these inequalities with the fact that $E[GCV_n(\tilde{\lambda}_1, \tilde{\lambda}_2)] \leq E[GCV_n(\lambda_1^*, \lambda_2^*)]$ gives

$$R_n(\tilde{\lambda}_1, \tilde{\lambda}_2)(1 - h_n(\tilde{\lambda}_1, \tilde{\lambda}_2)) \leq R_n(\lambda_1^*, \lambda_2^*)(1 + h_n(\lambda_1^*, \lambda_2^*)) \tag{F.3}$$

and rearranging terms,

$$1 \leq \frac{R_n(\tilde{\lambda}_1, \tilde{\lambda}_2)}{R_n(\lambda_1^*, \lambda_2^*)} \leq \frac{1 + h_n(\lambda_1^*, \lambda_2^*)}{1 + h_n(\tilde{\lambda}_1, \tilde{\lambda}_2)}. \tag{F.4}$$

We will have proved the theorem if we can show that $h_n(\lambda_1^*, \lambda_2^*)$ converges to zero as n tends to infinity.

From Theorem 4.1 we have the following equality:

$$\begin{aligned} h(\lambda_1, \lambda_2) = {} & \frac{2\tau_1^2(\lambda_1, \lambda_2)}{(1 - \tau_1(\lambda_1, \lambda_2))^2} + \frac{1 + \lambda_2}{(1 - \tau_1(\lambda_1, \lambda_2))^2} \\ & \times \left| 2\lambda_2 \frac{\tau_1(\lambda_1, \lambda_2)}{\tau_2(\lambda_1, \lambda_2)} - (1 + \lambda_2) \frac{\tau_1^2(\lambda_1, \lambda_2)}{\tau_2(\lambda_1, \lambda_2)} \right| \\ & + \frac{2n^{-1}\lambda_2}{(1 + \lambda_2)(1 - \tau_1(\lambda_1, \lambda_2))^2} \left| E\left[\epsilon^T H_n\left(\frac{\lambda_1}{1 + \lambda_2} \right)(g(\hat{\theta}_n) - g(\theta_0)) \right] \right| \\ \equiv {} & S_1 + S_2 + S_3. \end{aligned} \tag{F.5}$$

Using some results from Speckman (1985), if $n(\lambda_1/(1 + \lambda_2)) \to \infty$

$$\tau_j(\lambda_1, \lambda_2) \sim \frac{\ell_m}{n\left(\frac{\lambda_1}{1 + \lambda_2} \right)^{1/2\nu}} \pi^{-1} \int_0^1 p(u)^{1/2\nu}\, du, \qquad j = 1, 2 \tag{F.6}$$

and

$$\ell_m = \int_0^1 \frac{1}{(1 + x^{2\nu})^2}\, dx. \tag{F.7}$$

Then

$$S_1 = O\left(n^{-2} \left(\frac{\lambda_1}{1 + \lambda_2} \right)^{-1/\nu} \right) \tag{F.8}$$

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press

and

$$S_2 = O(\lambda_2) \tag{F.9}$$

as n tends to infinity. We now show that under Assumptions A.1–A.7

$$S_3 = O\left(n^{-3/2}\left(\frac{\lambda_2}{1+\lambda_2}\right)\left(\frac{\lambda_1}{1+\lambda_2}\right)^{-1/2\nu}\right). \tag{F.10}$$

This is proved by showing that

$$E\left[\epsilon^T H_n\left(\frac{\lambda_1}{1+\lambda_2}\right)(g(\hat{\theta}_n) - g(\theta_0))\right] = O\left(n^{-1/2}\left(\frac{\lambda_1}{1+\lambda_2}\right)^{-1/2\nu}\right) \tag{F.11}$$

as n tends to infinity.

The following equalities hold:

$$E\left[\epsilon^T H_n\left(\frac{\lambda_1}{1+\lambda_2}\right)\epsilon\right] = n\sigma^2\tau_1(\lambda_1,\lambda_2) = O\left(\left(\frac{\lambda_1}{1+\lambda_2}\right)^{-1/2\nu}\right) \tag{F.12}$$

and

$$E\left[(g(\hat{\theta}_n) - g(\theta_0))^T H_n\left(\frac{\lambda_1}{1+\lambda_2}\right)(g(\hat{\theta}_n) - g(\theta_0))\right] = \mathrm{tr}\left(H_n\left(\frac{\lambda_1}{1+\lambda_2}\right)V_n(g(\hat{\theta}_n))\right)$$

as n tends to infinity. Then using the Cauchy–Schwarz inequality result, (F.9) follows.

Thus, if $n \to \infty$, $\lambda_2 \to 0$, and $\lambda_1 \to 0$ in such a manner that $n(\lambda_1/(1+\lambda_2))^{1/2\nu} \to \infty$, then, $h_n(\lambda_1,\lambda_2)$ tends to zero.

If either λ_1, λ_2, or $n^{-1}(\lambda_1/(1+\lambda_2))^{-1/2\nu}$ is bounded away from zero, $R_n(\lambda_1,\lambda_2)$ does not tend to zero. Thus to minimize $R_n(\lambda_1,\lambda_2)$ we must have a sequence such that $\lambda_1^* \to 0$, $\lambda_2^* \to 0$, and $n(\lambda_1^*/(1+\lambda_2^*))^{1/2\nu} \to \infty$, and then $h_n(\lambda_1^*,\lambda_2^*) \to 0$. Furthermore $E[GCV_n(\tilde{\lambda}_1,\tilde{\lambda}_2)] \to \sigma^2$ (equations (F.2) and (F.3)). But to make this possible, $\tilde{\lambda}_1 \to 0$, $\tilde{\lambda}_2 \to 0$, and $n(\tilde{\lambda}_1/(1+\tilde{\lambda}_2))^{1/2\nu} \to \infty$, and so it can be concluded that $h_n(\tilde{\lambda}_1,\tilde{\lambda}_2) \to 0$ as n goes to infinity, and then the proof is finished. ■

APPENDIX G: PROOF OF THEOREM 4.3

Under conditions A.1–A.7, from Theorem 4.2,

$$R_n(\tilde{\lambda}_1,\tilde{\lambda}_2) = R_n(\lambda_1^*,\lambda_2^*) + o_p(R_n(\lambda_1^*,\lambda_2^*))$$

as n tends to infinity. If either λ_1, λ_2, or $n^{-1}(\lambda_1/(1+\lambda_2))^{-1/2\nu}$ is bounded away from zero, $R_n(\lambda_1,\lambda_2)$ does not tend to zero. Therefore, the minimizers of $R_n(\lambda_1,\lambda_2)$ must fulfill that $\lambda_1^* \to 0$, $\lambda_2^* \to 0$, and $n(\lambda_1^*/(1+\lambda_2^*))^{1/2\nu} \to \infty$. Then, if we apply Theorem 3.2

$$R_n(\tilde{\lambda}_1,\tilde{\lambda}_2) = O(n^{-2\nu/(2\nu+1)}).$$

The proof is finished using the Markov inequality. ■

https://doi.org/10.1017/S0266466699151065 Published online by Cambridge University Press