JOURNAL OF MULTIVARIATE ANALYSIS 32, 230–255 (1990)

Estimation of Generalized Additive Models

PRABIR BURMAN

Rutgers University

Communicated by the Editors

Spline estimation of generalized additive models is considered here. Cross-validation is used as a criterion of model estimation. Some computationally simpler approximations to cross-validation are given.

1. INTRODUCTION

In a recent paper, Stone [16] proved some general results concerning the optimal rates of convergence for generalized additive models. In another recent paper, Hastie and Tibshirani [10] discussed various computational methods for such models. For a thorough treatment of generalized additive models, see Mccullagh and Nelder [14].

Let $(X, Y) \in [0, 1]^p \times R$ be random variables; Y is the response variable and $\mathbf{X} = (X_1, ..., X_p)$ is the vector of covariates. The function of interest here is a response function f, i.e., f is a function of the conditional distribution of Y given $X = x$. We seek additive approximation to f, i.e., the functions of the form $\phi(x) = \phi_0 + \phi_1(x_1) + \cdots + \phi_p(x_p)$, ϕ_0 is a constant. The motivation and the reasons for considering additive approximations are given in the above-mentioned papers. Let $f^* = f_0^* + f_1^* + \cdots + f_p^*$ be the "closest" additive approximation to f (defined in Section 2). Stone considered the problem of estimating f^* based on the data (i.e., n independent copies of (X, Y)) with B-splines. Under the assumption that for each $j = 1, ..., p$, the qth derivative $f_j^{*(q)}$ exists and $|f_j^{*(q)}(x) - f_j^{*(q)}(y)| \leqslant M\,|x - y|^{\nu}$ for some $0 < \nu \leqslant 1$ and $M > 0$, Stone showed that one can construct an estimate $\hat{f}_n$ of f^* such that $\int (\hat{f}_n - f^*)^2\, dG = O_p(n^{-2r})$, where $r = (q + \nu)/(2q + 2\nu + 1)$ and G is the marginal distribution of X. This tells us that the optimal rate of convergence for estimating f^* is independent of p.

Received June 25, 1985; revised September 1988, June 1989.

AMS 1980 subject classifications: Primary 62G05; Secondary 62G25.

Key words and phrases: B-spline, quasi-exponential family, Newton–Raphson, cross-validation.

0047-259X/90 $3.00

However an important practical aspect of this method remains unsolved since the amount of smoothness of f^* is never known in practice. If one considers the estimation of f^* by splines, the question remains "how many knots". Stone's result tells us that the optimal number of knots depend directly on the smoothness assumption of f^*. Since the amount of smoothness is never known, choosing the number of knots based only on the data is an important task. As it will be clear later on, by choosing the number of knots we will provide a parametric model for f^*.

Here we use cross-validation as a criterion. However, in some cases, like binomial logit, closed forms of the estimates are not available and iterations are needed and hence cross-validation can be computationally heavy (see [10]). We give here one- and two-step Newton–Raphson approximations to the cross-validation and show that these approximations are as good as the cross-validation itself. As will be obvious in Section 3, these approximations save a good deal of computing. Some recent papers including Li [13] and Hardle and Marron [9] discuss cross-validation in different contexts.

The main technical difficulty here is the very frequent uses of Newton–Raphson approximation.

The organization of the paper is as follows. In Section 2, we introduce quasi-exponential families and additive spline approximations. In Section 3, we give the cross-validated criterion along with one- and two-step approximations to it and the main result. We present some simulation results in Section 3b. Some technical lemmas are given in Section 4. In Section 5, the proofs of the main remsults are given and in Section 6 the proofs of the technical lemmas of Section 4 are given.

The author is indebted to Professor Charles J. Stone for suggesting the problem. The author wishes to thank the referees for their helpful comments.

2. Quasi-Exponential Families and Additive Splines

Let (X, Y) be a r.v. on $[0, 1]^p \times R$. We assume that the distribution of Y given $X = x$ has a density with respect to some σ-finite measure $\bar{v}$. Following Stone [16] we say that (X, Y) has a quasi-exponential distribution if the expected log-likelihood has the form

$$\Lambda(a) = E\{Yb_1(a(X)) + b_2(a(X))\}$$

$$= \int [b_3(f(x))b_1(a(x)) + b_2(a(x))]\, dG(x), \tag{2.1}$$

where G is the marginal distribution of X; b_1, b_2, and b_3 are known functions;

$$m(x) = E(Y \mid X = x) = b_3(f(x)), \qquad b_3 = -b_2'/b_1'. \tag{2.2}$$

The function of interest here is the response function f. Let A be the class of additive functions ϕ of the form

$$\phi(x) = \phi_0 + \phi_1(x_1) + \phi_2(x_2) + \cdots + \phi_p(x_p), \tag{2.3}$$

where ϕ_0 is a constant and $\int \phi_i(x)\, dG(x) = 0$, $i = 1, ..., p$.

It can be shown that the ϕ_i's are uniquely determined except on a set of lebesgue measure zero. Stone proved that under certain conditions (conditions 1, 2, and 3) supremum of Λ on A is attained at some f^* in A. It also turns out that this f^* is unique and essentially bounded with respect to the lebesgue measure. Let us first state some of the conditions we need.

CONDITION 1. (a) *There is a subinterval S of R such that $\bar{v}$ is concentrated on S (i.e., $\bar{v}(S^c) = 0$) and*

$$b_1''(u)\, y + b_2''(u) < 0 \qquad \textit{for all } u \in R \textit{ and } y \in S.$$

(b) *It follows from part* (a) *that*

$$b_1''(u) b_3(u_0) + b_2''(u) < 0 \qquad \textit{for all } u, u_0 \in R.$$

CONDITION 2. *Marginal distribution G of X has a density g with respect to the lebesgue measure and there exists constants $0 < c_1 < c_2$ such that $c_1 \leqslant g(x) \leqslant c_2$ for all x.*

CONDITION 3. *There exists a positive constant t_0 such that $E(e^{tY} \mid X = x) < \infty$ for all $|t| \leqslant t_0$ and $x \in [0, 1]^p$.*

We note that for this paper Condition 3 could be weakened to $\sup_x E(Y^{2u} \mid X = X) < \infty$ for all $u \geqslant 1$.

CONDITION 4. *b_1 and b_2 have continuous third derivatives.*

Now let us consider a few examples

EXAMPLE 1. Normal regression. The model is: $Y = m(X) + \varepsilon$, where $\varepsilon \sim \text{Normal}(0, \sigma^2)$. Then m is the ordinary regression function and we seek additive approximation to m. Here $b_1(u) = u$, $b_2(u) = -u^2/2$, and $b_3(u) = u$. The response function f here is m and $S = R$.

EXAMPLE 2. Binomial-logit. The model is: Y given $X=x$ is Binomial$(n_0, \pi(x))$. Here, the response function $f(x)=\log\{\pi(x)/(1-\pi(x))\}$, $b_1(u)=u$, $b_2(u)=-n_0\log(1+e^u)$, $b_3(u)=n_0e^u/(1+e^u)$, and $S=[0, n_0]$.

EXAMPLE 3. Poisson. The model is: Y given $X=x$ is Poisson$(\lambda(x))$. Here, the response function $f(x)=\log\lambda(x)$, $b_1(u)=u$, $b_2(u)=-e^u$, $b_3(u)=e^u$, and $S=[0, \infty)$.

For other examples like binomial-probit, geometric models, see Stone [16].

Now let us discuss spline approximation to the aditive function f^*. Let $a(x)=a_0+a_1(x_1)+\cdots+a_p(x_p)$ be such that a_0 is a constant and each a_i is a spline of degree d on $[0, 1]$ with k_i equispaced knots $i=1, ..., p$; i.e., for each $i=1, ..., p$, (i) a_i is a polynimial of degree d on $[(j-1)k_i^{-1}, jk_i^{-1}]$ for $j=1, ..., k_i$; (ii) a_i is $(d-1)$ times continuously differentiable on $[0, 1]$; and (iii) $\int a_i(x)\, dG(x)=0$.

For $\mathbf{k}=(k_1, ..., k_p)$, let $S_{\mathbf{k}}$ be the class of functions $a(\cdot)$, as given above, satisfying (i), (ii), and (iii). Then $S_{\mathbf{k}}$ is called the class of additive splines. For $i=1, ..., p$, let $B_{k_i,t}$, $t=1, ..., k_i+d$, be the normalized B-splines of degree d on $[0, 1]$ with k_i equispaced knots. Then $\{B_{k_i,t},\ t=1, ..., k_i+d\}$ is the basis of the class of all splines of degree d on $[0, 1]$ with k_i equispaced knots. Since $B_{k_i,t}$'s are normalized $\sum_t B_{k_i,t}\equiv 1$ on $[0, 1]$. It is very easy to see that $\{B_{k_i,t},\ t=1, ..., k_i+d,\ i=1, ..., p\}$ spans S_k but is not a basis of S_k, since 1 can be written as $p-1$ independent ways. The dimension of S_k is $\lambda_k=k_1+\cdots+k_p+p(d-1)+1$.

CONDITION 5. *$f^*\neq a$ a.e., for any $a\in S_k$ for any k.*

Let us now find a basis for S_k. We will do it only for the case $k_1=k_2=\cdots=k_p=k$. It will be clear how to handle the general case. Let $\mathbf{e}_1, ..., \mathbf{e}_{k+d-1}$ be $(k+d)$-dimensional vectors which are orthogonal to each other, orthogonal to $\mathbf{1}$, the vector of 1's, and $\mathbf{e}_i'\mathbf{e}_i=1$ for all i. For $2\leqslant t\leqslant p$ and $1\leqslant i\leqslant k+d-1$, let

$$\psi_{kti}(x_t)=\sum_j e_{ij}B_{kj}(x_t)/(kb_{ktj}), \tag{2.4}$$

where $b_{ktj}=\int B_{kj}(x_t)\, dx_t$.

It is easy to check that $\int \psi_{kti}(x_t)\, dx_t=0$. For $t=1$, $\psi_{k1i}(x_1)=B_{ki}(x_1)$, $i=1, ..., k+d$. It is also easy to see that ψ_{kti}'s form a basis for S_k. Let ψ_k be the vector of all the ψ_{kti}'s. Let us note that ψ_k can be written as $\psi_k=D_kB_k$, where D_k is $\lambda_k\times(kp+dp)$ matrix and B_k is the vector all the B-splines of the X-variables.

3. Estimates of f^* and Cross-Validated Estimates

a. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. from the quasi-exponential distribution described in (2.1). Let $l_n(a)$ be the empirical estimate of $\Lambda(a)$; i.e.,

$$l_n(a) = n^{-1} \sum \{Y_j b_1(a(X_j)) + b_2(a(X_j))\}. \tag{3.1}$$

Let us recall $\lambda_k = k_1 + \cdots + k_p + p(d-1) + 1$ is the dimension of S_k (defined in the last section). In (3.1) we consider functions $a(\cdot)$ of the form $\theta'\psi_k$, $\theta \in R^{\lambda_k}$ (where ψ_k is defined in (2.4)) and maximize l_n with respect to θ. Let the maximum be attained at $\hat{\theta}_k$ (since l_n is concave because of Condition 1) and we will call $\hat{s}_k = \hat{\beta}'_k B_k = \hat{\theta}'_k \psi_k$. It is easy to see that we get the usual likelihood equation,

$$l_{1n}(\hat{s}_k) = 0, \tag{3.2}$$

where

$$l_{1n}(a) = n^{-1} \sum \{Y_j b'_1(a(X_j)) + b'_2(a(X_j))\} B_k(X_j). \tag{3.3}$$

Let us point out that b'_1 and b'_2 denote the derivatives of b_1 and b_2. Otherwise "$'$" denotes the transpose of vectors or matrices. For a matrix A, $\|A\|$ is the matrix norm of A. Let us denote

$$\begin{aligned} w(y, u) &= yb_1(u) + b_2(u), & w_1(y, u) &= yb'_1(u) + b'_2(u), \\ w_2(y, u) &= yb''_1(u) + b''_2(u), & \Lambda_n(a) &= \int w(m(x), a(x))\, dG_n(x) \end{aligned} \tag{3.4}$$

$$\begin{aligned} \Lambda_1(a) &= \int w_1(m(x), a(x)) B_k(x)\, dG(x), \\ \Lambda_2(a) &= \int w_2(m(x), a(x)) B_k(x) B'_k(x)\, dG(x), \\ \Lambda_{1n}(a) &= \int w_1(m(x), a(x)) B_k(x)\, dG_n(x), \\ \Lambda_{2n}(a) &= \int w_2(m(x), a(x)) B_k(x) B'_k(x)\, dG_n(x), \\ l_{2n}(a) &= n^{-1} \sum w_2(Y_j, a(X_j)) B_k(X_j) B'_k(X_j). \end{aligned} \tag{3.5}$$

Let $L_n(k) = \Lambda(\hat{s}_k) - \Lambda(f^*)$, where f^* is the closest additive approximation to the response function f (Section 2). We would like to choose $\hat{s}_k$ so that $L_n(k)$ is maximized over a broad class of possible candidates. The class of possible candidates could be:

$$\text{(i)} \qquad \hat{s}_k : k_i \leqslant K_n = n^{1-\alpha}, \qquad i = 1, ..., p \tag{3.6a}$$

or

$$\text{((ii)} \qquad \hat{s}_k : k_1 = \cdots = k_p = k \leqslant K_n = n^{1-\alpha} \tag{3.6b}$$

for some small $\alpha > 0$ ($\lambda_k < n$ in both cases). The first class of condidates has more flexibility than the second, whereas the latter is computationally more advantageous. For notational convenience, we will present the methods and the results for the second class of candidates. However, we would like to emphasize that each and everything in this paper is valid for the first class of alternatives with the same conditions and the same proofs.

Let $L_n(\tilde{k}) = \sup\{L_n(k) : k \leqslant K_n\}$, then $\hat{s}_{\tilde{k}}$ is the best estimate of f^*. Unfortunately, $\tilde{k}$ depends, among other things, on the smoothness of f^* and hence $\tilde{k}$ is unknown. One way out is to estimate $L_n(k)$ by the method of cross-validation (see Duin [5], Habbema, Hermans, and Van den Broek [8], Stone [17]). Let $\hat{s}_{kj}$ be the estimate of f^* (as in (3.2)) with all the $(n-1)$ observations except (X_j, Y_j) and then an estimate of $L_n(k)$ is

$$\hat{L}_n(k) = n^{-1} \sum \{ Y_j b_1(\hat{s}_{kj}(X_j)) + b_2(\hat{s}_{kj}(X_j))\} - \Lambda(f^*). \tag{3.7}$$

Let $\hat{L}_n(\hat{k}) = \max\{\hat{L}_n(k) : k \leqslant K_n\}$; then we declare $\hat{s}_{\hat{k}}$ as our estimate of f^*. The fact that $\hat{s}_{\hat{k}}$ is a good choice is proved by the result $L_n(\hat{k})/L_n(\tilde{k}) \to^P 1$ (Theorem 3.4).

As we have mentioned earlier, calculation of $\hat{L}_n(k)$ can be computationally very heavy. So we will give the following approximations to $\hat{L}_n(k)$ and these approximations are as good as $\hat{L}_n(k)$ itself (Theorem 3.2). Lemma 4.8(a) tells us that one-step Newton–Raphson approximation to $\hat{s}_{kj}(x)$ is given by

$$\hat{s}_k(X_j) + n^{-1}\{ Y_j b_1'(\hat{s}_k(X_j)) + b_2'(\hat{s}_k(X_j))\} v_j, \tag{3.8}$$

where $v_j = B_k'(X_j) \hat{U}_n(\hat{s}_k) B_k(X_j)$, $\hat{U}_n(a) = D_k'(D_k l_{2n}(a) D_k')^{-1} D_k$. Substituting this in (3.7) and denoting $\hat{\varepsilon}_{kj} = Y_j - b_3(\hat{s}_k(X_j))$, we get the one-step approximation to $\hat{L}_n(k)$,

$$\hat{L}_n^{(1)}(k) = l_n(\hat{s}_k) + n^{-2} \sum \{ Y_j b_1'(\hat{s}_k(X_j)) + b_2'(\hat{s}_k(X_j))\}^2 v_j - \Lambda(f^*) \tag{3.9a}$$

$$= l_n(\hat{s}_k) + n^{-2} \sum \hat{\varepsilon}_{kj}^2 b_1'^2(\hat{s}_k(X_j)) v_j - \Lambda(f^*) \tag{3.9b}$$

(since $b_3(u) b_1'(u) + b_2'(u) \equiv 0$ for all n by definition).

A two-step Newton–Raphson method (along the line of Lemma 4.8(b)) will give us the following two-step approximation $\hat{L}_n^{(2)}(k)$ to $\hat{L}_n(k)$:

$$l_n(\hat{s}_k) + n^{-2} \sum \hat{\varepsilon}_{kj}^2 b_1'^2(\hat{s}_k(X_j)) v_j [1 + (3/2) n^{-1} v_j] - \Lambda(f^*). \tag{3.10}$$

A brief discussion on the relationship between our estimators of L_n, $\hat{L}_n^{(1)}$, and $\hat{L}_n^{(2)}$ and the estimator proposed by O'Sullivan, Yandall, and Raynor [15] is given at the end of this section.

Now let us present the main results of this paper. Let $U(a) = D_k'(D_k \Lambda_2(a) D_k')^{-1} D_k$ and $U_n(a) = D_k'(D_k \Lambda_{2n}(a) D_k')^{-1} D_k$. Let us note that Condition 1 guarantees the existence of U, U_n, and $\hat{U}_n$. If functions of the form $\theta'\psi_k$, $\theta \in R^{\lambda_k}$, are used to maximize Λ and Λ_n (see (3.3)), then let the maxima be attained at θ_k and $\bar{\theta}_k$, respectively. Let $s_k = \theta_k'\psi_k = \beta_k' B_k$ and $\bar{s}_k = \bar{\theta}_k'\psi_k = \bar{\beta}_k' B_k$. Let

$$V_n(k) = (2n)^{-1} \int \sigma^2(x) w_1^2(m(x), \bar{s}_k(x)) B_k'(x) U_n(\bar{s}_k) \times B_k(x)\, dG_n(x) + \Lambda(\bar{s}_k) - \Lambda(f^*), \tag{3.11}$$

where $\sigma^2(x) = \operatorname{Var}(Y \mid X = x)$.

We get the following important lemma,

Lemma 3.1. $\sup_{k \leqslant K_n} |(L_n(k) - V_n(k))/V_n(k)| \to^P 0$.

The next result tells us that the approximations (3.9) and (3.10) to $\hat{L}_n$ are as good as $\hat{L}_n$ itself.

Theorem 3.2. (a) $\sup_{k \leqslant K_n} |(\hat{L}_n(k) - \hat{L}_n^{(1)}(k))/V_n(k)| \to^P 0$,
(b) $\sup_{k \leqslant K_n} |(\hat{L}_n(k) - \hat{L}_n^{(2)}(k)))/V_n(k)| \to^P 0$.

The next lemma along with Lemma 3.1 will give us the main results of this paper.

Lemma 3.3. $\sup_{k \leqslant K_n} |[\{\hat{L}_n(k) - \hat{L}_n(\tilde{k})\} - \{L_n(k) - L_n(\tilde{k})\}]/L_n(k)| \to^P 0$.

Lemmas 3.1 and 3.3 tell us that for any $\varepsilon > 0$, the following inequalities hold with probability approaching one:

$$\hat{L}_n(\hat{k}) - \hat{L}_n(\tilde{k}) \leqslant L_n(\hat{k}) - L_n(\tilde{k}) - \varepsilon L_n(\hat{k}).$$

Since $\hat{L}_n(\hat{k}) - \hat{L}_n(\tilde{k}) \geqslant 0$, we obtain $L_n(\tilde{k}) \leqslant (1-\varepsilon) L_n(\hat{k})$. Since $\varepsilon > 0$ is arbitrary we obtain

Theorem 3.4. $L_n(\hat{k})/L_n(\tilde{k})| \to^P 1$.

Let, $\hat{L}_n^{(1)}(\hat{k}^{(1)}) = \max\{\hat{L}_n^{(1)}(k): k \leqslant K_n\}$ and $\hat{L}_n^{(2)}(\hat{k}^{(2)}) = \max\{\hat{L}_n^{(2)}(k):$

$k \leqslant K_n\}$. Then Lemmas 3.1, 3.3 and Theorem 3.2 give us the following result.

THEOREM 3.5. (a) $L_n(\hat{k}^{(1)})/L_n(\tilde{k}) \to^P 1$,
(b) $L_n(\hat{k}^{(2)})/L_n(\tilde{k}) \to^P 1$.

The following results will not be proved, but a reading of this paper will reveal that arguments very similar to the ones used in this paper are enough to prove them. Let $E_n(k) = E(L_n(k) \mid \mathbf{X}_n)$ and $\bar{E}_n(k) = EL_n(k)$, where $\mathbf{X}_n = (X_1, ..., X_n)$. Let

$$\bar{V}_n(k) = (2n)^{-1} \int \sigma^2(x) w_1^2(m(x), s_k(x)) B_k'(x) U(s_k) B_k(x)\, dG(x)$$

$$+ (2n)^{-1} \int w_1^2(m(x), s_k(x)) B_k'(x) U(s_k) B_k(x)\, dG(x)$$

$$- (2n)^{-1} \Lambda_1'(s_k) U(s_k) \Lambda_1(s_k) + \Lambda(s_k) - \Lambda(f^*).$$

LEMMA 3.6. $\sup_{k \leqslant K_n} |(V_n(k) - \bar{V}_n(k))/\bar{V}_n(k)| \to^P 0$.

LEMMA 3.7. (a) $\sup_{k \leqslant K_n} |(E_n(k) - V_n(k))/V_n(k)| \to^P 0$,
(b) $\sup_{k \leqslant K_n} |(V_n(k) - \bar{V}_n(k))/\bar{V}_n(k)| \to^P 0$.

The next results tell us that results analogous to Theorem 3.4 hold for $E_n(k)$ and $\bar{E}_n(k)$.

THEOREM 3.8. (a) $E_n(\hat{k})/\max_{k \leqslant K_n} E_n(k) \to^P 1$, $E_n(\hat{k}^{(1)})/\max_{k \leqslant K_n} E_n(k) \to^P 1$, $E_n(\hat{k}^{(2)})/\max_{k \leqslant K_n} E_n(k) \to^P 1$,
(b) $\bar{E}_n(\hat{k})/\max_{k \leqslant K_n} \bar{E}_n(k) \to^P 1$, $\bar{E}_n(\hat{k}^{(1)})/\max_{k \leqslant K_n} \bar{E}_n(k) \to^P 1$, $\bar{E}_n(\hat{k}^{(2)})/\max_{k \leqslant K_n} \bar{E}_n(k) \to^P 1$.

Remarks. 1. In practice, it is desirable to place the knots at the sample quantiles. All the results of this paper would remain valid in that case. Also, it may be sometimes desirable to have some linear constraints on the splines near the endpoints (see Koo and Stone [12]) and place some smoothness constraints (see O'Sullivan *et al.* [15]). Recently, Von Golitschek and Schumaker [6] have shown that the use of a penalized least squares method leads to an improvement over ordinary least squares.

2. O'Sullivan, Yandell, and Raynor [15] have discussed the estimation of the generalized additive model using the generalized cross-validation (gcv) method. Their method is a generalization of the work by Golub, Heath, and Wahba [7]. Their method is a little different from ours as will be obvious in the discussion here. Let us examine this when the

response function f is an additive function, i.e., $f=f^*$. Let $E(Y|X=x)=b_3(f^*(x))$ and $v(f^*(x))=\mathrm{Var}(Y|X=x)$. It can be shown that

$$\begin{aligned} &E[\{Y-b_3(\hat{s}_k(X))\}^2 v^{-1}(f^*(X))\,|\,\text{data}] \\ &\quad = 1+E[\{b_3(\hat{s}_k(X))-b_3(f^*(X))\}^2 v^{-1}(f^*(X))\,|\,\text{data}] \\ &\quad \simeq 1+E[\{\hat{s}_k(X)-f^*(X)\}^2 \\ &\qquad\qquad \times b_3'^2(f^*(X))v^{-1}(f^*(X))\,|\,\text{data}] \simeq 1-L_n(k). \end{aligned}$$

(Here data refers to (X_i, Y_i), $i=1, ..., n$). Since f^* is unknown, O'Sullivan *et al.* estimate the quantity $E[\{Y-b_3(\hat{s}_k(X))\}^2 v^{-1}(\hat{s}_k(X))\,|\,\text{data}]$ using the gcv method and that is certainly quite reasonable. However, $E[\{Y-b_3(\hat{s}_k(X))\}^2\, v^{-1}(\hat{s}_k(X))\,|\,\text{data}] \simeq E[v(f^*(X))\, v^{-1}(\hat{s}_k(X))\,|\,\text{data}]-L_n(k)$. Since $\hat{L}_n^{(1)}$ and $\hat{L}_n^{(2)}$ estimate L_n, this shows that our method could be different from the gcv method.

b. *A Simulation Study*. The model we consider here is a logistic model with $\log\{\pi(x)/(1-\pi(x))\}=\sin(2\pi x)$, where $\pi(x)=P(Y=1\,|\,X=x)$. Here, we take $X \sim \mathrm{Uniform}[0, 1]$. Obviously, $f^*(x)=\sin(2\pi x)$ here. We fit splines of degree 1 (i.e., piecewise linear functions) with the knots at the sample quantiles. Let the maxima of $L_n(k)$, $\hat{L}_n^{(1)}$, and $\hat{L}_n^{(2)}$ be attained at $\tilde{k}$, $\hat{k}^{(1)}$, and $\hat{k}^{(2)}$, respectively. In the table below, we calculated the mean, s.d., and the median of $r_1=L_n(\tilde{k})/L_n(\hat{k}^{(1)})$ and $r_2=L_n(\tilde{k})/L_n(\hat{k}^{(2)})$ for sample sizes $n=50$ and $n=100$. All the estimates are based on 200 repeats.

n		Mean	s.d.	Median
50	r_1	0.94	0.08	0.96
	r_2	0.95	0.07	0.98
100	r_1	0.97	0.03	0.98
	r_2	0.97	0.03	0.98

This study suggests that $\hat{L}_n^{(1)}$ does the job of picking the model very well. $\hat{L}_n^{(2)}$ might be slightly better than $L_n^{(1)}$, but the difference seems to be negligible.

4. Technical Results

We begin this section with two important lemmas which will be used in proving subsequent lemmas. The first one is an extension of the inequality by Hoeffding [11] and the second one contains the moment inequalities by Whittle [18]. For two classes of random variables $\{\xi_{n,u}\}$ and $\{\mu_{n,u}\}$

indexed by u, by $\xi_{nu} = O_p(\mu_{nu})$ we mean $\sup_u |\xi_{nu}/\mu_{nu}| = O_p(1)$. Similarly, $\xi_{nu} = o_p(\mu_{nu})$ is defined.

LEMMA 4.1. *Let $\{Z_{\alpha nj}: j = 1, ..., n\}$ be an array of independent random variables with zero means, where α varies over an index set D_n, the cardinality of which is less than n^t for some $t > 0$. We will assume that the moment generating functions of $Z_{\alpha nj}$'s are uniformly bounded by a constant in an interval about zero.*

Let $S_{\alpha n} = n^{-1} \sum_{j=1}^n Z_{\alpha nj}$, $\sigma_{\alpha nj}^2 = \mathrm{Var}(Z_{\alpha nj})$, and $v_{\alpha n}^2 = n^{-1} \sum_{j=1}^n \sigma_{\alpha nj}^2$. Let $\delta_n = (\log n)^{1/2+\delta}/\sqrt{n}$ for some $\delta > 0$. Then $\sup_\alpha |v_{\alpha n}^{-1} S_{\alpha n}| = o_p(\delta_n)$, if the following hold:

(i) $\sup_\alpha \delta_n v_{\alpha n}^{-1} = O(1)$,

(ii) $\sup_{\alpha, j} (\delta_n v_{\alpha n}^{-1})^2 \sigma_{\alpha nj}^2 = o(1)$, *and*

(iii) $\sup_{\alpha, j} (\delta_n v_{\alpha n}^{-1})\ r_{\alpha nj} \sigma_{\alpha nj}^{-2} = o(1)$, *where* $r_{\alpha nj} = \sup_{|t| \leqslant 1} |EZ_{\alpha nj}^3 \exp\{t\delta_n v_{\alpha n}^{-1} Z_{\alpha nj}\}|$.

THEOREM 4.2. *Let Z_j, $j = 1, ..., n$, be independent random variables with zero means and $EZ_j^{4s} < \infty$. Let $\gamma_j(u) = \{E|Z_j|^u\}^{1/u}$. Then there exist constants $c_1(s) > 0$ and $c_2(s) > 0$ such that*

(a) $E(\sum b_j Z_j)^{2s} \leqslant c_1(s)(\sum b_j^2 \gamma_j^2(2s))^s$ and

(b) $E(\sum a_{ij} Z_i Z_j - E \sum a_{ij} Z_i Z_j)^{2s} \leqslant c_2(s)(\sum a_{ij}^2 \gamma_i^2(2s) \gamma_j^2(2s))^s$,

for any sequences of numbers $\{b_j: j = 1, ..., n\}$ and $\{a_{ij}: i, j = 1, ..., n\}$.

The following Lemma tells us about β_k, $\bar{\beta}_k$, and $\hat{\beta}_k$ defined in Section 3.

LEMMA 4.3. *Let $\|\beta_k\|_\infty$ denote the maximum of the absolute values of the elements of β_k. $\|\bar{\beta}_k\|_\infty$ and $\|\hat{\beta}_k\|_\infty$ are also similarly defined. Then,*

(a) $\sup_k \|\beta_k\|_\infty < \infty$,

(b) $\sup_k \|\bar{\beta}_k\|_\infty$ *and* $\sup_k \|\hat{\beta}_k\|_\infty$ *are bounded in probability.*

The following lemmas tell us about the behaviour of l_{2n}, Λ_{2n}, and Λ_2. Let $\delta_{nk} = \max(\delta_n, (\lambda_k \log n)^\delta/\sqrt{n})$ and $\delta < \alpha/(6(1-\alpha))$.

LEMMA 4.4. (a) $\|\int B_k B_k' \, dG\| = O(\lambda_k^{-1})$,

(b) $\|\int B_k B_k' (G_n - G)\| = o_p(\lambda_k^{-1/2} \delta_{nk})$

(c) $\|\int B_k B_k' \, dG_n\| = O_p(\lambda_k^{-1})$.

(d) *There exists constant $0 < c_1 < c_2$ such that all the eigenvalues of $\int \psi_k \psi_k' \, dG$ lie between $c_1 \lambda_k^{-1}$* and *$c_2 \lambda_k^{-1}$.*

(e) *All the eigenvalues of $\int \psi_k \psi_k' \, dG_n$ lie between $c_3 k^{-1}$ and $c_4 k^{-1}$, for some $0 < c_3 < c_4$, with probability approaching one.*

LEMMA 4.5. *Let A_n be the class of all functions h on $[0, 1]^p$ such that $\|h\| = (\int h^2)^{1/2} \leqslant \varepsilon_n \lambda_k^{1/2} \delta_{nk}$ and $\|h\|_\infty \leqslant \varepsilon_n \lambda_k^{1/2} \delta_{nk}$ for some $\varepsilon_n \to^P 0$. Then we obtain*:

(a) $\sup\{\|\Lambda_{2n}(s_k + h) - \Lambda_2(s_k + h)\| : h \in A_n\} = o_p(\lambda_k^{-1/2} \delta_{nk})$,

(b) $\sup\{\|l_{2n}(s_k + h) - \Lambda_{2n}(s_k + h)\| : h \in A_n\} = o_p(\lambda_k^{-1/2} \delta_{nk})$,

(c) *all the three quantities* $\sup\{\|\Lambda_2(s_k + h)\| : h \in A_n\}$, $\sup\{\|\Lambda_{2n}(s_k + h)\| : h \in A_n\}$, *and* $\sup\{\|l_{2n}(s_k + h)\| : h \in A_n\}$ *are* $O_p(\lambda_k^{-1})$.

(d) $\sup\{\|\Lambda_2(s_k + h) - \Lambda_2(s_k)\| : h \in A_n\} = o_p(\lambda_k^{-1/2} \delta_{nk})$.

LEMMA 4.6. *Let $U(a)$, $U_n(a)$, and $\hat{U}_n(a)$ be as in Section 3 and let A_n be the same as in Lemma 4.5*:

(a) *All the following three quantities,* $\sup\{\|U(s_k + h) - U(s_k)\| : h \in A_n\}$, $\sup\{\|U_n(s_k + h) - U_n(s_k)\| : h \in A_n\}$, *and* $\sup\{\|\hat{U}_n(s_k + h) - \hat{U}_n(s_k)\| : h \in A_n\}$ *are* $o_p(\lambda_k^{3/2} \delta_{nk})$.

(b) *Both* $\sup\{\|U_n(s_k + h) - U(s_k + h)\| : h \in A_n\}$ *and* $\sup\{\|\hat{U}_n(s_k + h) - U_n(s_k + h)\| : h \in A_n\}$ *are* $o_p(\lambda_k^{3/2} \delta_{nk})$.

(c) *All the eigenvalues of $U(s_k + h)$, $U_n(s_k + h)$, and $\hat{U}_n(s_k + h)$, $h \in A_n$, are between $c_5 \lambda_k$ and $c_6 \lambda_k$ (for some $0 > c_5 > c_6$) with probability approaching one.*

The following two lemmas tell us about the behaviour of β_k, $\bar{\beta}_k$, $\hat{\beta}_k$, and the cross-validated estimates $\hat{\beta}_{kj}$'s of $\hat{\beta}_k$.

LEMMA 4.7. (a) $\|l_{1n}(\bar{s}_k)\|$ *and* $\|\Lambda_{1n}(s_k)\|$ *are* $o_p(\delta_n)$.

(b) (i) $\bar{\beta}_k - \beta_k = -U_n(s_k) \Lambda_{1n}(s_k) + r_{1n}(k)$,

(ii) $\hat{\beta}_k - \bar{\beta}_k = -U_n(\bar{s}_k) l_{1n}(\bar{s}_k) + r_{2n}(k)$, *where both* $\|r_{1n}(k)\|$ *and* $\|r_{2n}(k)\|$ *are* $o_p(\lambda_k^{3/2} \delta_n^2)$.

(c) *Both* $\|\bar{\beta}_k - \beta_k\|$ *and* $\|\hat{\beta}_k - \bar{\beta}_k\|$ *are* $o_p(\lambda_k \delta_n)$.

(d) *Both* $\|\bar{\beta}_k - \beta_k\|_\infty$ *and* $\|\hat{\beta}_k - \bar{\beta}_k\|_\infty$ *are* $o_p(\lambda_k^{1/2} \delta_n)$.

LEMMA 4.8. (a) $\hat{\beta}_{kj} = \hat{\beta}_k + n^{-1} \hat{U}_n(\hat{s}_k) B_k(X_j) w_1(Y_j, \hat{s}_k(X_j)) + r_{3n}(k)$, *where* $\|r_{3n}(k)\| = o_p(\lambda_k^2 n^{-2} (\log n)^{1+2\delta})$.

(b) *A two-step approximation is given by* $\hat{\beta}_{kj} = \hat{\beta}_k + n^{-1} \hat{U}_n(\hat{s}_k) B_k(X_j) w_1(Y_j, \hat{s}_k(X_j)) + n^{-2} \hat{U}_n(\hat{s}_k) B_k(X_j) B_k'(X_j) \hat{U}_n(\hat{s}_k) B_k(X_j) w_1^2(Y_j, \hat{s}_k(X_j)) + r_{4n}(k)$, *where* $\|r_{4n}(k)\| = o_p(\lambda_k^3 n^{-3} (\log n)^{3/2 + 3\delta})$.

We conclude this section with two important technical lemmas.

LEMMA 4.9. (a) $\inf_{k \leqslant K_n} |n V_n(k)| \to^P \infty$,

(b) $\sum_{k \leqslant K_n} (n V_n(k))^{-2} \to^P 0$, *where $V_n(k)$ is defined in* (3.11).

The first part of the following lemma is due to Stone [16].

LEMMA 4.10. *There exists a constant* $c_7 > 0$, $c_8 > 0$ *such that*

(a) $-c_7 \|s_k - f^*\|_G^2 \leqslant \Lambda(s_k) - \Lambda(f^*) \leqslant -c_8 \|s_k - f^*\|_G^2$

(b) $-c_7 \|\bar{s}_k - s_k\|_G^2 \leqslant \Lambda(\bar{s}_k) - \Lambda(s_k) \leqslant -c_8 \|\bar{s}_k - s_k\|_G^2$,

where for any function h *on* $[0, 1]^p$, $\|h\|_G^2 = \int h^2\, dG$.

5. PROOF OF THE MAIN RESULTS

Proof of Lemma 3.1. Let us note that $V_n(k)$ can be written as $(2n)^{-1} \int \sigma^2(x) b_1'^2(\bar{s}_k(x)) H_n(\bar{s}_k, x)\, dG_n(x) + \Lambda(\bar{s}_k) - \Lambda(s_k) + \Lambda(s_k) - \Lambda(f^*)$, where $H_n(\bar{s}_k, x) = B_k'(x) U_n(\bar{s}_k) B_k(x)$. Because of Lemmas 4.6(c) and 4.10, we conclude that with probability approaching one,

$$V_n(k) \geqslant -c_9[\lambda_k n^{-1} + \|\bar{s}_k - s_k\|_G^2 + \|s_k - f^*\|_G^2] \qquad \text{for some } c_9 > 0. \tag{5.1}$$

Throughout, we will use the facts that $\lambda_k^{3/2}\delta_{nk}^3 = o_p(V_n(k))$, $\lambda_k^2 \delta_{nk}^4 = o_p(V_n(k))$, and $\delta_n \leqslant \delta_{nk}$ (see Section 3). Now,

$$\begin{aligned}
\Lambda(\hat{s}_k) - \Lambda(\bar{s}_k) &= (\hat{\beta}_k - \beta_k)' \Lambda_1(\bar{s}_k) \\
&\quad + 2^{-1}(\hat{\beta}_k - \beta_k)' \Lambda_2(\bar{s}_k)(\hat{\beta}_k - \beta_k) + r_{5n}(k), \\
|r_{5n}(k)| &= \left| (\hat{\beta} - \beta)' \int_0^1 (1-t)\{\Lambda_2(t\hat{s}_k + (1-t)\bar{s}_k) - \Lambda_2(\bar{s}_k)\}\, dt (\hat{\beta} - \beta) \right| \\
&\leqslant \|\hat{\beta}_k - \beta_k\|^2 \left\| \int_0^1 (1-t)\{\Lambda_2(t\hat{s}_k + (1-t)\bar{s}_k) - \Lambda_2(\bar{s}_k)\}\, dt \right\|. \\
&= o_p(\lambda_k^2 \delta_n^2) o_p(\lambda_k^{-1/2}\delta_{nk}) \qquad \text{(by Lemmas 4.5(d) and 4.7(c))} \\
&= o_p(\lambda_k^{3/2}\delta_{nk}^3) = o_p(V_n(k)).
\end{aligned} \tag{5.2}$$

So, $L_n(k) - V_n(k)$ can be written as

$$\begin{aligned}
&\{(\hat{\beta}_k - \beta_k)' \Lambda_1(\bar{s}_k)\} + 2^{-1} \Big\{ (\hat{\beta}_k - \beta_k)' \Lambda_2(\bar{s}_k)(\hat{\beta}_k - \beta_k) \\
&\qquad - n^{-1} \int \sigma^2(x) b_1'(\bar{s}_k(x)) J_n(\bar{s}_k, x)\, dG_n(x) \Big\} \\
&\qquad + \Big\{ n^{-1} \int \sigma^2(x) b_1'(\bar{s}_k)) [J_n(\bar{s}_k, x) - H_n(\bar{s}_k, x)]\, dG_n(x) \Big\} + o_p(V_n(k)) \\
&= S_1(k) + 2^{-1} S_2(k) + S_3(k) + o_p(V_n(k)), \qquad \text{say},
\end{aligned}$$

where $J_n(\bar{s}_k, x) = B_k'(x) U_n(\bar{s}_k) \Lambda_2(\bar{s}_k) U_n(\bar{s}_k) B_k(x)$.

We will show that $S_i(k) = o_p(V_n(k))$ for $i = 1, 2, 3$ (i.e., $\sup_{k \leqslant K} |S_i(k)/V_n(k)| \to^P 0$ for $i = 1, 2, 3$). By Lemma 4.7(b) we obtain

$$S_1(k) = -l'_{1n}(\bar{s}_k) U_n(\bar{s}_k) A_{1n}(\bar{s}_k) + r'_{2n}(k) A_1(\bar{s}_k) = -S_{11}(k) + S_{12}(k), \qquad \text{say.}$$

Because of Lemma 4.3,

$$\sup_x |w_1(m(x), \bar{s}_k(x)) - w_1(m(x), s_k(x))| = O_p(1) \sup_x |\bar{s}_k(x) - s_k(x)|.$$

Since $A_1(s_k) = 0$, for any $u \in R^{\lambda_k}$, $\|u\| = 1$, $|u' A_1(\bar{s}_k)|^2 = O_p(1) \, \|\int B_k B'_k \, dG\| \, \|\bar{s}_k - s_k\|_G^2$. So using Lemma 4.4(a), we obtain

$$\|A_1(\bar{s}_k)\| = O_p(\lambda_k^{-1/2}) \, \|\bar{s}_k - s_k\|_G. \tag{5.3}$$

Consequently, by (5.1),

$$|S_{12}(k)| \leqslant o_p(\lambda_k^{3/2} \delta_{nk}^2) O_p(\lambda_k^{-1/2}) \, \|\bar{s}_k - s_k\|_G = o_p(V_n(k)). \tag{5.4}$$

Since $A_{1n}(\bar{s}_k) = 0$,

$$l_{1n}(\bar{s}_k) = l_{1n}(\bar{s}_k) - A_{1n}(\bar{s}_k).$$

So, we conclude

$$S_{11}(k) = n^{-1} \sum \varepsilon_j \xi_j,$$

where $\varepsilon_j = Y_j - m(X_j)$ (where $m(x) = E(Y \mid X = x)$) and

$$\xi_j = b'_1(\bar{s}_k(X_j)) B'_k(X_j) U_n(\bar{s}_k) A_1(\bar{s}_k).$$

By Theorem 4.2(a), (5.3), Lemmas 4.4(c) and 4.6(c) we obtain $E\{(nS_{11}(k)^4 \mid \mathbf{X}_n\} = O(1)(\sum \xi_j^2)^2$. Now,

$$\begin{aligned}
\sum \xi_j^2 &= O_p(n) A'_1(\bar{s}_k) U_n(\bar{s}_k) \left\{ \int B_k B'_k \, dG_n \right\} U_n(\bar{s}_k) A_1(\bar{s}_k) \\
&\leqslant O_p(n) \, \|A_1(\bar{s}_k)\|^2 \, \|U_n(\bar{s}_k)\|^2 \left\| \int B_k B'_k \, dG_n \right\| \\
&= O_p(n) \lambda_k^{-1} \, \|s_k - \bar{s}_k\|_G^2 \, O_p(\lambda_k^2) O_p(\lambda_k^{-1}) = O_p(1) n \, \|\bar{s}_k - s_k\|_G^2.
\end{aligned}$$

So,

$$\begin{aligned}
& P[\sup_{k \leqslant K_n} |S_{11}(k)/V_n(k)| > \varepsilon \mid \mathbf{X}_n] \\
&\leqslant \varepsilon^{-4} \sum_{k \leqslant K_n} E\{(nS_{11}(k))^4 \mid \mathbf{X}_n\}/(nV_n(k))^4 \\
&= O(1) \sum_{k \leqslant K_n} \left(\sum \xi_j^2 \right)^2 \Big/ (nV_n(k))^4 = O_p(1) \sum_{k \leqslant K_n} (nV_n(k))^{-2}.
\end{aligned}$$

Because of Lemma 4.9(b), the right-hand side of the last line converges to zero in probability and hence $S_{11}(k) = o_p(V_n(k))$. This together with (5.4) proves that $S_1(k) = o_p(V_n(k))$.

Using Lemma 4.7(b),

$$
\begin{aligned}
&|(\hat{\beta}_k - \bar{\beta}_k)' A_2(\bar{s}_k)(\hat{\beta}_k - \bar{\beta}_k) - l'_{1n}(\bar{s}_k) U_n(\bar{s}_k) A_2(\bar{s}_k) U_n(\bar{s}_k) l_{1n}(\bar{s}_k)| \\
&\quad \leqslant 2\, \|l_{1n}(\bar{s}_k)\|\, \|U_n(\bar{s}_k)\|\, \|A_2(\bar{s}_k) r_{2n}(k)\| + |r'_{2n}(k) A_2(\bar{s}_k) r_{2n}(k)| \\
&\quad \leqslant 2\, \|l_{1n}(\bar{s}_k)\|\, \|U_n(\bar{s}_k)\|\, \|A_2(\bar{s}_k)\|\, \|r_{2n}(k)\| + \|r_{2n}(k)\|^2\, \|A_2(\bar{s}_k)\|.
\end{aligned}
$$

Using Lemmas 4.5(c) and 4.6(c), the last line above is

$$
\begin{aligned}
&o_p(\delta_n) O_p(\lambda_k) O_p(\lambda_k^{-1}) o_p(\lambda_k^{3/2}\delta_n^2) + o_p(\lambda_k^3 \delta_n^4) O_p(\lambda_k^{-1}) \\
&\quad = o_p(\lambda_k^{3/2}\delta_n^3) + o_p(\lambda_k^2 \delta_n^4) = o_p(V_n(k)).
\end{aligned}
$$

This tells us $S_2(k)$ can be written as (by noting that $A_{1n}(\bar{s}_k) = 0$)

$$
\begin{aligned}
&l'_{1n}(\bar{s}_k) U_n(\bar{s}_k) A_2(\bar{s}_k) U_n(\bar{s}_k) l_{1n}(\bar{s}_k) \\
&\qquad - n^{-1} \int \sigma^2(x) b_1'^2(\bar{s}_k(x)) J_n(\bar{s}_k, x)\, dG_n(x) + o_p(V_n(k)) \\
&\quad = \left\{ n^{-2} \sum \varepsilon_i \varepsilon_j u(X_i, X_j) - n^{-2} \sum \sigma^2(X_j) u(X_j, X_j) \right\} + o_p(V_n(k)) \\
&\quad = \bar{S}_2(k) + o_p(V_n(k)),
\end{aligned}
$$

say, where

$$
u(x_1, x_2) = b'_1(\bar{s}_k(x_1)) b'_1(\bar{s}_k(x_2)) B'_k(x_1) U_n(\bar{s}_k)\, A_2(\bar{s}_k) U_n(\bar{s}_k) B_k(x_2).
$$

Using Theorem 4.2 we obtain

$$
\begin{aligned}
E\{(n\bar{S}_2(k))^4 \mid \mathbf{X}_n\} &= O(n^{-4}) \left(\sum u^2(X_i, X_j) \right)^2 \\
&= O(1) \left(\int u^2(x_1, x_2)\, dG_n(x_1)\, dG_n(x_2) \right)^2.
\end{aligned}
$$

Since $\sup_{k,x} |b'_1(\bar{s}_k, x))| = O_p(1)$,

$$
\begin{aligned}
&\int u^2(x_1, x_2)\, dG_n(x_1)\, dG_n(x_2) \\
&\qquad \leqslant O_p(1) \int \|B_k(x)\|^2\, dG_n(x)\, \|U_n(\bar{s}_k)\|^4\, \|A_2(\bar{s}_k)\|^2 \left\| \int B_k B'_k\, dG_n \right\|.
\end{aligned}
$$

Since $\sup_{k,x} \|B_k(x)\|^2 < \infty$, using Lemmas 4.4(c), 4.5(c), and 4.6(c), the last line is $O_p(1)0(1)\,O_p(\lambda_k^4)O_p(\lambda_k^{-2})O_p(\lambda_k^{-1}) = O_p(\lambda_k)$. This shows that $E\{(n\bar{S}_2(k))^4 \mid \mathbf{X}_n\} = O_p(1)\lambda_k^2$. This proves that $\bar{S}_2(k) = o_p(V_n(k))$ and consequently $S_2(k) = o_p(V_n(k))$.

Now it remains to be shown that $S_3(k) = o_p(V_n(k))$. Let us note that $U_n(\bar{s}_k)\Lambda_{2n}(\bar{s}_k)\,U_n(\bar{s}_k) = U_n(\bar{s}_k)$ and hence by Lemmas 4.5(a) and 4.6(c),

$$\begin{aligned}\|U_n(\bar{s}_k)\Lambda_2(\bar{s}_k)\,U_n(\bar{s}_k) - U_n(\bar{s}_k)\| &\leqslant \|U_n(\bar{s}_k)\|^2\ \|\Lambda_{2n}(\bar{s}_k) - \Lambda_2(\bar{s}_k)\|^2 \\ &= o_p(\lambda_k^2)\,o_p(\lambda_k^{-1/2}\delta_{nk}) = o_p(\lambda_k^{3/2}\delta_{nk}).\end{aligned}$$

So

$$\begin{aligned}|J_n(\bar{s}_k, x) - H_n(\bar{s}_k, x)| &\leqslant \|B_k(x)\|^2\ \|U_n(\bar{s}_k)\Lambda_2(\bar{s}_k)\,U_n(\bar{s}_k) - U_n(\bar{s}_k)\| \\ &= o_p(\lambda_k^{3/2}\delta_{nk}).\end{aligned}$$

Hence $|S_3(k)| = O_p(1)n^{-1}o_p(\lambda_k^{3/2}\delta_{nk}) = O_p(1)\,o_p(\lambda_k^{3/2}\delta_{nk}^3) = o_p(V_n(k))$. And this completes the proof of Lemma 3.1. ∎

Proof of Lemma 3.2. (a) By Lemma 4.8(a), $\hat{s}_{kj}(x) = \hat{s}_k(x) + n^{-1}B_k'(x)\,\hat{U}_n(\hat{s}_k)\,B_k(X_j)\,w_1(Y_j, \hat{s}_k(X_j)) + B_k'(x)r_{3n}(k)$. Since $\sup_x |\hat{s}_k(x)| = O_p(1)$ by Lemma 4.3(b), we obtain $\max_{1 \leqslant j \leqslant n} |w_1(Y_j, \hat{s}_k(X_j))| \leqslant O_p(1) \max_{1 \leqslant j \leqslant n} |Y_j| + O_p(1) = O_p((\log n)^{1/2+\delta})$, by Condition 3 in Section 2. Since $|r_{3n}(k)| = o_p(\lambda_k^2 n^{-2}(\log n)^{1+2\delta})$, and $\|\hat{U}_n(\hat{s}_k)\| = O_p(\lambda_k)$, we obtain $\max_{1 \leqslant j \leqslant n} \sup_x |\hat{s}_{kj}(x) - \hat{s}_k(x)| = O_p(\lambda_k n^{-1}(\log n)^{1/2+\delta})$.

Using the Taylor expansion and the fact that $n^{-1} \sum |Y_j| = O_p(1)$,

$$\begin{aligned}|\hat{L}_n(k) - \hat{L}_n^{(1)}(k)| &= \left| \int_0^1 (1-t)n^{-1} \sum w_2(Y_j, t\hat{s}_{kj}(X_j) + (1-t)\hat{s}_k(X_j)) \right. \\ &\qquad \left. \times (\hat{s}_{kj}(X_j) - \hat{s}_k(X_j))^2\, dt \right| \\ &\leqslant \left\{ O_p(1)n^{-1} \sum |Y_j| + O_p(1) \right\} \sup_j \sup_x |\hat{s}_{kj}(x) - \hat{s}_k(x)|^2 \\ &= O_p(\lambda_k^2 n^{-2}(\log n)^{1+2\delta}) = o_p(V_n(k)).\end{aligned}$$

(b) Proof of this part is almost the same as that of part (a).

In order to prove Lemma 3.3 we will need the following lemma, proof of which is given in Burman [2].

LEMMA 5.1. *Let ϕ be a function bounded in absolute value by a constant, say* 1. *Then*

$$E\left\{ \int \phi\, d(G_n - G) \right\}^{2s} \leqslant c_9 n^{-2s} \sum_{l=1}^{s} \left\{ n \int \phi^2\, dG \right\}^l$$

where c_9 depends only on s.

Proof of Lemma 3.3. Let us note that because of Lemmas 3.1 and 3.2 it is enough to prove $\sup_{k \leqslant K_n} |[\{\hat{L}_n^{(1)}(k) - \hat{L}_n^{(1)}(\tilde{k})\} - \{L_n(k) - L_n(\tilde{k})\}]/V_n(k)| \to^P 0$. Let $H_n(a, x)$ be as in the beginning of the proof of Lemma 3.1. Let us also define

$$H(a, x) = B_k'(x)\, U(a)\, B_k(x) \qquad \text{and} \qquad \hat{H}(a, x) = B_k'(x)\, \hat{U}_n(a)\, B_k(x). \quad (5.5)$$

Let us note that $\hat{L}_n^{(1)}(k) - L_n(k)$ can be written as

$$\begin{aligned}
&\Big\{(l_n - \Lambda_n)(\hat{s}_k) - (l_n - \Lambda_n)(\bar{s}_k) \\
&\qquad + n^{-1} \int \sigma^2(x)\, b_1'^2(\bar{s}_k(x))\, H_n(\bar{s}_k, x)\, dG_n(x)\Big\} \\
&\qquad + \{(l_n - \Lambda_n)(\bar{s}_k) - (l_n - \Lambda_n)(f^*)\} \\
&\qquad + \{(\Lambda_n - \Lambda)(\hat{s}_k) - (\Lambda_n - \Lambda)(\bar{s}_k)\} \\
&\qquad + \Big\{(\Lambda_n - \Lambda)(\bar{s}_k) - (\Lambda_n - \Lambda)(s_k) \\
&\qquad + n^{-1} \int w_1^2(m(x), s_k(x))\, H_n(s_k, x)\, dG_n(x)\Big\} \\
&\qquad + \Big\{n^{-2} \sum w_1^2(Y_j, \hat{s}_k(X_j))\, H_n(\hat{s}_k, X_j) \\
&\qquad - n^{-2} \int \sigma^2(x)\, b_1'^2(\bar{s}_k(x))\, H_n(\bar{s}_k, x)\, dG_n(x) \\
&\qquad - n^{-1} \int w_1^2(m(x), s_k(x))\, H_n(\bar{s}_k, x)\, dG_n(x)\Big\} + \{(l_n - \Lambda_n) f^*)\}. \\
&\qquad = T_1(k) + T_2(k) + T_3(k) + T_4(k) + T_5(k) + T_6(k), \qquad \text{say.} \quad (5.6)
\end{aligned}$$

Let us first note that $T_6(k)$ does not depend on k and hence the difference $\{\hat{L}_n^{(1)}(k) - L_n(k)\} - \{\hat{L}_n^{(1)}(\tilde{k}) - L_n(\tilde{k})\}$ does not contain $T_6(k)$. We will show that $T_i(k) = o_p(V_n(k))$, $i = 1, ..., 5$. Let

$$V_n^+(k) = -c_9[\lambda_k n^{-1} + \|s_k - f^*\|_G^2], \quad (5.7)$$

where c_9 is the same as in (5.2). Since $s_k \neq f^*$ a.e. for any k, an argument similar to one given for Lemma 4.9 will show that

$$\inf_k |nV_n^+(k)| \to \infty \qquad \text{and} \qquad \sum_{k \leqslant K_n} (nV_n^+(k))^{-2} \to 0 \qquad \text{as} \quad n \to \infty. \quad (5.8)$$

Now,

$$(l_n - \Lambda_n)(\hat{s}_k) - (l_n - \Lambda_n)(\bar{s}_k) = (\hat{\beta}_k - \bar{\beta}_k)' l_{1n}(\bar{s}_k) + r_{6n}(k), \tag{5.9}$$

where $r_{6n}(k) = o_p(V_n(k))$ by Lemmas 4.5(b) and 4.7(b) (as in (5.2)).

Because of Lemma 4.7(b), $(\hat{\beta}_k - \bar{\beta}_k)' l_{1n}(\bar{s}_k) + l_{1n}(\bar{s}_k) U_n(\bar{s}_k) l_{1n}(\bar{s}_k)$ can be bounded above in absolute value by $|r'_{2n}(k) l_{1n}(\bar{s}_k)| \leqslant \|r_{2n}(k)\| \, \|l_{1n}(\bar{s}_k)\| = o_p(\lambda_k^{3/2}\delta_n^2) o_p(\delta_n) = o_p(\lambda_k^{3/2}\delta_n^3) = o_p(V_n(k))$. So (5.9) tells us that

$$\begin{aligned} T_1(k) &= -l'_{1n}(\bar{s}_k) U_n(\bar{s}_k) l_{1n}(\bar{s}_k) \\ &\quad + n^{-2} \int \sigma^2(x) b'_1(\bar{s}_k(x)) H_n(\bar{s}_k, x) \, dG_n(x) + o_p(V_n(k)) \\ &= -\bar{T}_1(k) + o_p(V_n(k)), \qquad \text{say.} \end{aligned}$$

Now $\bar{T}_1(k)$ can be written as, $n^{-2} \sum \varepsilon_i \varepsilon_j u_{ij} - n^{-2} \sum \sigma^2(X_j) u_{jj}$, where $u_{ij} = b'_1(\bar{s}_k(X_i)) b'_1(\bar{s}_k(X_j)) B'_k(X_i) U_n(\bar{s}_k) B_k(X_j)$.

Because of Theorem 4.2(b), $E\{(n\bar{T}_1(k))^4 \mid \mathbf{X}_n\} = O(1) n^{-4} (\sum u_{ij}^2)^2 = O(1)(n^{-2} \sum u_{ij}^2)^2$. Now, $n^{-2} \sum u_{ij}^2 = O_p(1)\{\int \|B_k\|^2 \, dG_n\} \, \|U_n(\bar{s}_k)\|^2 \, \|\int B_k B'_k \, dG_n\| = O_p(1) O(1) O_p(\lambda_k^2) O_p(\lambda_k^{-1}) = o_p(\lambda_k)$. Hence, $\bar{T}_1(k) = o_p(V_n(k))$ and consequently $T_1(k) = o_p(V_n(k))$. Now,

$$T_2(k) = n^{-1} \sum \varepsilon_j [b_1(\bar{s}_k(X_j)) - b_1(f^*(X_j))]. \tag{5.10}$$

Using Theorem 4.2(a) we obtain

$$\begin{aligned} E\{(nT_2(k))^4 \mid \mathbf{X}_n\} &= O(1) \left(\sum [b_1(\bar{s}_k(X_j)) - b_1(f^*(X_j))]^2 \right)^2 \\ &= O(1) n^2 \left(\int (b_1(\bar{s}_k(x)) - b_1(f^*(x))^2 \, dG_n(x) \right)^2 \\ &= O_p(n^2) \, \|\bar{s}_k - f^*\|_{G_n}^4 \\ &= O_p(n^2) [\|\bar{s}_k - s_k\|_{G_n}^2 + \|s_k - f^*\|_{G_n}^2]^2, \end{aligned} \tag{5.11}$$

where for any g on $[0, 1]^p$, $\|g\|_{G_n}^2 = \int g^2 \, dG_n$. Now,

$$\begin{aligned} &| \, \|\bar{s}_k - s_k\|_{G_n}^2 - \|\bar{s}_k - s_k\|_G^2 \, | \\ &\quad \leqslant \|\bar{\beta}_k - \beta_k\|^2 \left\| \int B_k B'_k \, d(G_n - G) \right\| \\ &\quad = o_p(\lambda_k^2 \delta_n^2) o_p(\lambda_k^{-1/2} \delta_n) = o_p(\lambda_k^{3/2} \delta_n^3) = o_p(V_n(k)). \end{aligned} \tag{5.12}$$

By Lemma 5.1, $E[\|s_k - f^*\|^2_{G_n} - \|s_k - f^*\|^2_G]^4 = o(1)[n^{-3}\|s_k - f^*\|^2_G + n^{-2}\|s_k - f^*\|^2_G]$. So,

$$\|s_k - f^*\|^2_{G_n} - \|s_k - f^*\|^2_G = o_p(V_n^+(k)) = o_p(V_n(k)). \tag{5.13}$$

Combining (5.10) through (5.13) we obtain

$$E\{(nT_2(k))^4 \mid \mathbf{X}_n\} = O_p(1)[n^2 \|\bar{s}_k - s_k\|^4_G + n^2 \|s_k - f^*\|^4_G + o_p(n^2 V_n^2(k))].$$

Because of (5.1), this tells us $T_2(k) = o_p(V_n(k))$. Now, $T_3(k) = (\hat{\beta}_k - \bar{\beta}_k)'(A_{1n} - A_1)(\bar{s}_k) + r_{7n}(k)$.

As in (5.2), we can show that $r_{7n}(k) = o_p(V_n(k))$ (by Lemmas 4.5(a) and 4.7(b)). Since $A_{1n}(\bar{s}_k) = 0$, $T_3(k) = -(\hat{\beta}_k - \bar{\beta}_k)' A_1(\bar{s}_k) + o_p(V_n(k))$. The first term of $T_3(k)$ is the same as $S_1(k)$ in the proof of Lemma 3.1 and we proved that $S_1(k) = o_p(V_n(k))$ and so $T_3(k) = o_p(V_n(k))$. Now,

$$(A_n - A)(\bar{s}_k) - (A_n - A)(s_k) = (\hat{\beta}_k - \bar{\beta}_k)'(A_{1n} - A)(s_k) + r_{8n}(k),$$

where $r_{8n}(k) = o_p(V_n(k))$ because of Lemmas 4.5(a) and 4.7(b).

Using Lemma 4.7(b), $(\hat{\beta}_k - \bar{\beta}_k)' A_{1n}(s_k) + A'_{1n}(s_k) U(s_k) A_{1n}(s_k)$ can be bounded above in absolute value by

$$\|r_{1n}(k)\| \; \|A_{1n}(s_k)\| = o_p(\lambda_k^{3/2}\delta_n^2) o_p(\delta_n) = o_p(\lambda_k^{3/2}\delta_n^3) = o_p(V_n(k)).$$

This tells us that $T_4(k)$ can be written as

$$\begin{aligned} &-A'_{1n}(s_k) U(s_k) A_{1n}(s_k) \\ &\qquad + n^{-1} \int w_1^2(m(x), s_k(x)) H(s_k, x)\, dG(x) + o_p(V_n(k)) \\ &= -\bar{T}_4(k) + o_p(V_n(k)), \qquad \text{say;} \end{aligned}$$

$\bar{T}_4(k)$ can be written as

$$\int \gamma_k(x_1, x_2)\, d(G_n - G)(x_1)\, d(G_n - G)(x_2) - n^{-1} \int \gamma_k(x, x)\, dG(x),$$

where

$$\gamma_k(x_1, x_2) = w_1(m(x_1), s_k(x_1)) w_1(m(x_2), s_k(x_2))\, B'_k(x_1) U(s_k) B_k(x_2).$$

By an easy extension of Lemma 3.4 in Burman [2] we obtain

$$\begin{aligned} E[n\bar{T}_4(k)]^4 \leqslant C[&n^{-2} E\gamma_k^4(X, X) + n^{-1} E\gamma_k^4(X_1, X_2) \\ &+ \{E\gamma_k^2(X_1, X_2)\}^2 + \{E\gamma_k(X_1, X_2)\}^4] \end{aligned} \tag{5.14}$$

Let us note that, since $\sup_x \lambda_k^{-1} |\gamma_k(x, x)| = O(1)$, $E\gamma_k^4(X, X) = O(\lambda_k^4)$.

A fairly straightforward argument will show that $E\gamma_k^4(X_1, X_2) = O(\lambda_k^3)$, $E\gamma_k^2(X_1, X_2) = O(\lambda_k)$, and $|E\gamma_k(X_1, X_2)| = O(1)$. So (5.14) tells us

$$E[n\bar{T}_4(k)]^4 = O(1)[n^{-2}\lambda_k^4 + n^{-1}\lambda_k^3 + \lambda_k^2] + O(1). \tag{5.15}$$

This proves that $\bar{T}_4(k) = o_p(V_n(k))$.

Now it remains to prove that $T_5(k) = o_p(V_n(k))$. Let us recall that because of Lemmas 4.6 and 4.7,

$$\begin{aligned} \sup_x |\hat{s}_k(x) - \bar{s}_k(x)| &= o_p(\lambda_k^{1/2}\delta_{nk}), \\ \sup_x |\bar{s}_k(x) - s_k(x)| &= o_p(\lambda_k^{1/2}\delta_{nk}), \end{aligned} \tag{5.16}$$

$$\begin{aligned} \|\hat{U}_n(\hat{s}_k) - U_n(\bar{s}_k)\| &= o_p(\lambda_k^{3/2}\delta_{nk}), \\ \|U_n(\bar{s}_k) - U_n(s_k)\| &= o_p(\lambda_k^{3/2}\delta_{nk}). \end{aligned} \tag{5.17}$$

From (5.17) we obtain

$$\begin{aligned} \sup_x |\hat{H}_n(\hat{s}_k, x) - H_n(\bar{s}_k, x)| &= o_p(\lambda_k^{3/2}\delta_{nk}), \\ \sup_x |H_n(\bar{s}_k, x) - H_n(s_k, x)| &= o_p(\lambda_k^{3/2}\delta_{nk}). \end{aligned} \tag{5.18}$$

Noting that $|w_1(y, \hat{s}_k(x)) - w_1(y, \bar{s}_k(x))| = o_p(\lambda_k^{1/2}\delta_{nk})[|y| + 1]$ and by (5.16) and (5.18),

$$\begin{aligned} &\left| n^{-2} \sum w_1^2(Y_j, \hat{s}_k(X_j)) \hat{H}_n(\hat{s}_k, X_j) - n^{-2} \sum w_1^2(Y_j, \bar{s}_k(X_j)) H_n(\bar{s}_k, X_j) \right| \\ &\leqslant \left| n^{-2} \sum \{w_1^2(Y_j, \hat{s}_k(X_j)) - w_1^2(Y_j, \bar{s}_k(X_j))\} H_n(\bar{s}_k, X_j) \right| \\ &\quad + \left| n^{-2} \sum w_1^2(Y_j, \bar{s}_k(X_j)) \{\hat{H}_n(\hat{s}_k, X_j) - H_n(\bar{s}_k, X_j)\} \right| \\ &= o_p(\lambda_k^{3/2}\delta_{nk} n^{-1}) + o_p(\lambda_k^{3/2}\delta_{nk} n^{-1}) = o_p(V_n(k)) \end{aligned} \tag{5.19}$$

(since $\sup_x |H_n(\bar{s}_k, x)| = O_p(\lambda_k)$, $n^{-1} \sum |Y_j| = O_p(1)$, and $n^{-1} \sum w_1^2 (Y_j, \bar{s}_k(X_j)) = O_p(1)$.)

From (5.19) we obtain

$$\begin{aligned} T_5(k) = {} & n^{-2} \sum w_1^2(Y_j, \bar{s}_k(X_j)) H_n(\bar{s}_k, X_j) \\ & + n^{-1} \int \sigma^2(x) b_1'^2(\bar{s}_k(x)) H_n(\bar{s}_k, x)\, dG_n(x) \\ & - n^{-1} \int w_1^2(m(x), s_k(x)) H(s_k, x)\, dG(x) + o_p(V_n(k)). \end{aligned}$$

Since $w_1(Y_j, \bar{s}_k(x_j)) = \varepsilon_j b_1'(\bar{s}_k(x_j)) + w_1(m(X_j), s_k(X_j))$,

$$\begin{aligned}
T_5(k) &= \left\{n^{-2} \sum [\varepsilon_j^2 - \sigma^2(X_j)] b_1'^2(\bar{s}_k(X_j)) H_n(\bar{s}_k, X_j)\right\} \\
&\quad + 2\left\{n^{-2} \sum \varepsilon_j b_1'(\bar{s}_k(X_j)) w_1(m(X_j), \bar{s}_k(X_j)) H_n(\bar{s}_k, X_j)\right\} \\
&\quad + \left\{n^{-1} \int [w_1^2(m(x), \bar{s}_k(X)) H_n(\bar{s}_k, X)\right. \\
&\quad \left. - w_1^2(m(x), s_k(X)) H(s_k, X)]\, dG_n(x)\right\} \\
&\quad + \left\{n^{-1} \int w_1^2(m(x), s_k(X)) H(s_k, X) d(G_n - G)(x)\right\} \\
&= T_{51}(k) + 2T_{52}(k) + T_{53}(k) + T_{54}(k), \qquad \text{say.}
\end{aligned}$$

By Whittle's theorem (Theorem 4.2(a)), it can be proved that T_{51} and T_{52} are $o_p(V_n(k))$. Now,

$$\begin{aligned}
|T_{53}(k)| &\leqslant n^{-1} \int |(w_1^2(m(x), \bar{s}_k(x)) - w_1^2(m(x), s_k(x)) H_n(\bar{s}_k, x)|\, dG_n(x) \\
&\quad + n^{-1} \int w_1^2(m(x), s_k(x))\, |H_n(\bar{s}_k, x) - H(s_k, x)|\, dG_n(x) \\
&= n^{-1} o_p(\lambda_k^{1/2} \delta_{nk}) O_p(\lambda_k) + n^{-1} o_p(\lambda_k^{3/2} \delta_{nk}) = o_p(V_n(k)). \qquad (5.20)
\end{aligned}$$

Finally, since $\operatorname{Var}(nT_{54}(k)) = O(\lambda_k^2 n^{-1})$, using Lemma 5.1, we obtain $nT_{54}(k) = o_p(\delta_n)\lambda_k^{3/2}$, hence

$$T_{54}(k) = o_p(1) n^{-1} \delta_n \lambda_k^{3/2} = o_p(V_n(k)). \qquad (5.21)$$

This concludes the proof of Lemma 3.3.

6. Proof of the Results in Section 4

Proof of Lemma 4.1. We will only give a sketch of the proof. Let us note that for any $t > 0$,

$$\begin{aligned}
P[v_{\alpha n}^{-1} S_{\alpha n} > \delta_n] &= P[\exp(t v_{\alpha n}^{-1} S_{\alpha n}) > \exp(t\delta_n)] \\
&\leqslant \exp(-t\delta_n) \prod_{j=1}^{n} E \exp\{t n^{-1} v_{\alpha n}^{-1} Z_{\alpha nj}\}. \qquad (6.1)
\end{aligned}$$

Now if we take $t = n\delta_n$ and expand $E(\exp\{tn^{-1}v_{\alpha n}^{-1}Z_{\alpha nj}\})$ in a Taylor series up to the term involving t^3, a simple calculation will show that the sum of the terms in (6.1) over $\alpha \in D_n$ converges to zero.

Similarly, one can show that $\Sigma_{\alpha \in D_n} P[v_{\alpha n}^{-1} S_{\alpha n} < -\delta_n] \to 0$ and hence the proof.

Before we prove the rest of the results, the following lemma (Lemma 4.8 from Burman and Chen [3]) will be needed. For $l \leqslant t$, l and t positive integers, let $Z(t, l) = \{(i_1, ..., i_t)\colon 1 \leqslant i_1, ..., i_t \leqslant l,\ i_1, ..., i_t$ are integers and each $(i_1, ..., i_t)$ has l distinct indices$\}$ and for any $\xi \in Z(t, l)$, let $a(\xi)$ = the number of indices in ξ appearing only once.

LEMMA 6.1. *Let ϕ be a bounded function on I^t, $I = [0, 1]^p$. Then there exists a constant c_{11} (depending only on t) such that*

$$\left| E \int \phi(x_1, ..., x_t) \prod_{i=1}^{t} d(G_n - G)(x_i) \right|$$

$$\leqslant c_{11} \sum_{l=1}^{t} n^{-t+l} \sum_{\xi \in Z(t,l)} n^{-[(a(\xi)+1)/2]} \int |\phi(x_\xi)|\, dF(x_1) \cdots dF(x_l),$$

where $[v]$ is the largest integer not exceeding v.

Let B_{k1} be the vector of $(k+d)$ B-splines of the first coordinate of $x \in [0, 1]^p$; i.e., B_{k1} is the vector of $(k+d)$ B-splines of $[0, 1]$. Then using property (viii) on page 155 of deBoor [4] and Lemma 5.1 of Burman [2], we get the result given below,

LEMMA 6.2. *There exists $0 < c_{12} < c_{13}$ such that for any $u \in R^{k+d}$, $c_{12} k^{-1} \|u\|^2 \leqslant \int (u' B_{k1}(x))^2\, dG(x) \leqslant c_{13} k^{-1} \|u\|^2$.*

Proof of Lemma 4.3. Follows quite easily by noting that Λ, Λ_n, and l_n are convex functions (because of Condition 1), f^* is bounded, and by using Lemma 4.10 and the property (viii) from page 155 of deBoor [4].

Proof of Lemma 4.4. (a) For any $u \in R^{t_k}$, $\|u\| = 1$ $(t_k = pk + pd)$,

$$u' \left\{ \int B_k B_k'\, dG \right\} u = \int (u' B_k)^2\, dG = O(\lambda_k^{-1}), \tag{6.2}$$

because of Lemma 6.2 and the fact that the density of G is bounded.

(b) Let us note that it is enough to prove that

$$\text{(i)} \quad \left\| \int B_{ki} B_{ki}' d(G_n - G) \right\| = o_p(\lambda_k^{-1/2} \delta_{nk}), \qquad i = 1, ..., p \tag{6.3}$$

$$\text{(ii)} \quad \left\| \int B_{ki} B_{kj}' d(G_n - G) \right\| = o_p(\lambda_k^{-1/2} \delta_{nk}), \qquad 1 \leqslant i \neq j \leqslant p, \tag{6.4}$$

where B_{ki} is the vector of $(k+d)$ B-splines for the ith coordinate of $x \in [0,1]^p$.

To prove (i) let us note that

$$\left\| \int B_{ki} B'_{ki} d(G_n - G) \right\| \leqslant \sup_{1 \leqslant t_1 \leqslant k+d} \sum_{t_2} \left| \int B_{kit_1} B_{kit_2} d(G_n - G) \right|. \tag{6.5}$$

Let us first note that $B_{kit_1} B_{kit_2} \equiv 0$ for $|t_1 - t_2| > M$ some $M > 0$. Since $\mathrm{Var}(B_{kit_1}(X) B_{kit_2}(X)) = O(k^{-1})$, Lemma 4.1 gives us $|\int B_{kit_1} B_{kit_2} d(G_n - G)| = o_p(\lambda_k^{-1/2} \delta_n^2)$ for $|t_1 - t_2| \leqslant M$ and this proves the desired result.

Now let us prove (ii). We will show that $\|\int B_{k1} B'_{k2} d(G_n - G)\| = o_p(\lambda_k^{-1/2} \delta_{nk})$. Let $R_{k1}(x_1, x_2) = \sum_{t=1}^{k+d} B_{k1t}(x_1) B_{k1t}(x_2)$, $R_{k2}(x_1, x_2) = \sum_{t=1}^{k+d} B_{k2t}(x_1) B_{k2t}(x_2)$. Let $A_k = \int B_{k1} B'_{k2} d(G_n - G)$, $A'_k A_k = \int B_{k1}(x_1) B'_{k1}(x_2) R_{k2}(x_1, x_2) \prod_{i=1}^{2} d(G_n - G)(x_i)$. Let us note that

$$\begin{aligned} \|A_k\|^{2s} \leqslant \mathrm{Trace}((A'_k A_k)^s) &= \int R_{k1}(x_1, x_2) \prod_{t=1}^{s} R_{k2}(x_{2t-1}, x_{2t}) \\ &\quad \times \prod_{t=1}^{s-1} R_{k1}(x_{2t}, x_{2t+1}) \prod_{t=1}^{2s} d(G_n - G)(x_t) \\ &= \int R_k(x) \prod_{t=1}^{2s} d(G_n - G)(x_t), \qquad \text{say.} \end{aligned}$$

We will take $s > (2\delta)^{-1}$. Using Lemma 6.1,

$$\begin{aligned} &\left| E \int R_k(\mathbf{x}) \prod_{t=1}^{2s} d(G_n - G)(x_t) \right| \\ &\quad \leqslant c_{10} \sum_{l=1}^{2s} n^{-2s+l} \sum_{\xi \in Z(2s,l)} n^{-[(a(\xi)+1)/2]} \int |R_k(x_\xi)| \, dG(x_1) \cdots dG(x_l). \end{aligned}$$

A simple calculation will show that for $\xi \in Z(2s, l)$, $\int |R_k(\mathbf{x}_\xi)| \, dG(x_1) \cdots dG(x_l) \leqslant c_{13} k^{-l}$.

For $l > s$, $a(\xi) > 2l - 2s$. Since $\sup_{k \leqslant K_n} k/n = O(1)$, we obtain

$$\begin{aligned} &\left| E \int R_k(\mathbf{x}) \prod_{t=1}^{2s} d(G_n - G)(x_t) \right| \\ &\quad \leqslant O(1) \left\{ \sum_{l=1}^{s} n^{-2s+l} k^{-l} + \sum_{l=s+1}^{2s} n^{-s} k^{-l} \right\} = O(1) k^{-s} n^{-s}. \end{aligned}$$

Now, $P[\sup_{k \leqslant K_n} \lambda_k^{1/2} \delta_{nk}^{-1} \|A_k\| > \varepsilon] \leqslant \varepsilon^{-2s} \sum_{k \leqslant K_n} \lambda_k^s \delta_{nk}^{-2s} E \|A_k\|^{2s}$. Since $\delta_{nk} \geqslant (\lambda_k \log n)^\delta / \sqrt{n}$, the last term is $O(1)(\log n)^{-2s\delta} \sum_{k \leqslant K_n} \lambda_k^{-2s\delta} = o(1)$ and hence the proof.

(c) Follows easily from (a) and (b).

(d) Note that $\|\int \psi_k \psi_k' \, dG\| \leqslant \|D_k D_k'\| \, \|\int B_k B_k' \, dG\| = O(\lambda_k^{-1})$. Because of part (a) and the fact that $u'\{\int \psi_k \psi_k' \, dG\} u = \int (u'\psi_k)^2 \, dG \geqslant c_1 \int (u'\psi_k)^2 \, d\mu$, μ is the lebesgue measure on $[0, 1]^p$.

(e) Follows from part (b) by noting that $\|\int \psi_k \psi_k' d(G_n - G)\| \leqslant \|D_k D_k'\| \, \|\int B_k B_k' d(G_n - G)\|$.

Proof of Lemma 4.5. (a) First note that

$$\begin{aligned} \|\Lambda_{2n}(s_k + h) - \Lambda_2(s_k + h)\| &\leqslant \|\Lambda_{2n}(s_k + h) - \Lambda_{2n}(s_k)\| \\ &\quad + \|\Lambda_{2n}(s_k) - \Lambda_2(s_k)\| + \|\Lambda_2(s_k) - \Lambda_2(s_k + h)\|. \end{aligned} \tag{6.6}$$

Let $u_1, u_2 \varepsilon R^{t_k}$, $t_k = kp + kd$, $\|u_1\| = \|u_2\| = 1$. Then

$$\begin{aligned} &|u_1'(\Lambda_{2n}(s_k + h) - \Lambda_{2n}(s_k)) u_2| \\ &= \bigg| \int [w_2(m(x), s_k(x) + h(x)) \\ &\qquad - w_2(m(x), s_k(x))](u_1' B_k(x))(u_2' B_k(x)) \, dG_n(x) \bigg|. \end{aligned}$$

Since $\sup_x |w_2(m(x), s_k(x) + h(x)) - w_2(m(x), s_k(x))| \leqslant O(1) \varepsilon_n \lambda_k^{1/2} \delta_{nk}$, the last line is bounded above by $O(1) \varepsilon_n \lambda_k^{1/2} \delta_{nk} \, \|\int B_k B_k' \, dG_n\| = O(1) \varepsilon_n \lambda_k^{1/2} \delta_{nk} O_p(\lambda_k^{-1}) = o_p(\lambda_k^{-1/2} \delta_{nk})$.

A similar argument shows that the third term in (6.6) is $o_p(\lambda_k^{-1/2} \delta_{nk})$. A proof similar to that of part (c) of Lemma 4.4 shows that the second term in (6.6) is $o_p(\lambda_k^{-1/2} \delta_{nk})$.

(b) Let us first note that

$$\begin{aligned} &\|l_{2n}(s_k + h) - \Lambda_{2n}(s_k + h)\| \\ &\leqslant \|l_{2n}(s_k + h) - l_{2n}(s_k)\| + \|l_{2n}(s_k) - \Lambda_{2n}(s_k)\| \\ &\quad + \|\Lambda_{2n}(s_k) - \Lambda_{2n}(s_k + h)\|. \end{aligned}$$

Arguing the same way as in part (a) we can show that each of the terms above is $o_p(\lambda_k^{-1/2} \delta_{nk})$.

(c) $\|\Lambda_2(s_k + h) - \Lambda_2(s_k)\| = o_p(\lambda_k^{-1/2} \delta_{nk})$ and this shows the eigenvalues of $\Lambda_2(s_k + h)$, $h \in A_n$, are $O_p(\lambda_k^{-1})$. The other parts are similarly proved.

Proof of Lemma 4.6. (a) Since $\|\Lambda_2(s_k + h) - \Lambda_2(s_k)\| = o_p(\lambda_k^{-1/2} \delta_{nk})$ and $\|D_k\| = O(1)$, it follows easily that $\|U(s_k + h) - U(s_k)\| = o(\lambda_k^{3/2} \delta_{nk})$. The other results of this part are similarly proved.

(b) Let us first note that

$$\|U_n(s_k+h)-U(s_k+h)\| \leqslant \|U_n(s_k+h)-U_n(s_k)\| + \|U_n(s_k)-U(s_k)\| + \|U(s_k)-U(s_k+h)\|. \tag{6.7}$$

By part (a), the first and third terms in (6.7) are $o_p(\lambda_k^{3/2}\delta_{nk})$. Since $\|A_{2n}(s_k+h)-A_2(s_k)\| = o_p(\lambda_k^{-1/2}\delta_{nk})$, a simple calculation will show that $\|U_n(s_k)-U(s_k)\| = o_p(\lambda_k^{3/2}\delta_{nk})$. The other results are similarly proved.

(c) Note that $\|U(s_k+h)-U(s_k)\| = o_p(\lambda_k^{3/2}\delta_{nk})$. Since $\lambda_k^{1/2}\delta_{nk} = O(1)$, this proves the result for $U(s_k+h)$. The other results are similarly proved.

We will use the following lemma, the proof of which is postponed to the very end of this section.

LEMMA 6.3. *Let u be a function on $R^t \to R^t$. Let $u(\theta_0)=0$ and $u(\theta_1)=\varepsilon$. Let Du denote the positive definite matrix derivative of u. Let us also assume that for $\|\theta-\theta_1\| \leqslant \|\theta_1-\theta_0\| + \|(Du(\theta_1))^{-1}\varepsilon\|$, all the eigenvalues of $Du(\theta)$ are between μ_1 and μ_2 $(0<\mu_1<\mu_2)$. Then $\theta_1-\theta_0=(Du(\theta_1))^{-1}\varepsilon+r$, where $\|r\| \leqslant \mu_1^{-2}\,\|\varepsilon\|\,\sup\{\|Du(\theta_1-t(Du(\theta_1))^{-1}\varepsilon)-Du(\theta_1)\| : 0\leqslant t\leqslant 1\}$.*

Proof of Lemma 4.7. Let $s=\theta'\psi_k$ for $\theta\in R^{\lambda_k}$. Let us denote $\pi(\theta)=D_kA_1(s)$, $\pi_n(\theta)=D_kA_{1n}(s)$, and $\hat\pi_n(\theta)=D'_k l_{1n}(s))$. Let us recall that $\pi(\theta_k)=\pi_n(\bar\theta_k)=\hat\pi_n(\hat\theta_k)=0$. The proof of part (a) is simple.

(b)(i) In order to use Lemma 6.3, let $\theta_0=\bar\theta_k$, $\theta_1=\theta_k$, $\pi_n=-u$, then $\varepsilon=-\pi_n(\theta_k)+\pi(\theta_k)$, $D\pi_n(\theta)=-D_kA_{2n}(s)D'_k$. Because of Lemma 4.3 and 4.4(e), and Condition 4, all the conditions of Lemma 6.3 are satisfied with $\mu_1=c_{15}\lambda_k^{-1}$, $c_{15}>0$. Lemma 6.3 tells us

$$\theta_k-\bar\theta_k=(D_kA_{2n}(s_k)D'_k)^{-1}D_kA_{1n}(s_k)+r.$$

Multiplying both sides by D'_k we get, $\beta_k-\bar\beta_k=U_n(s_k)A_{1n}(s_k)+r_{1n}(k)$. By (6.2) and Lemma 4.6(c) we get $\|\varepsilon\|=o_p(\delta_{nk})$, $\|(Du(\theta_1))^{-1}\|=O_p(\lambda_k)$, and hence $\|(Du(\theta_1))^{-1}\varepsilon\|=o_p(\lambda_k\delta_{nk})$ and so, by Lemma 4.5(a),

$$\sup_{0\leqslant t\leqslant 1}\|(Du(\theta_1-t(Du(\theta_1))^{-1}\varepsilon))-Du(\theta_1)\|=o_p(\delta_{nk}).$$

This shows $\|r_{1n}(k)\|=o_p(\lambda_k^{3/2}\delta_{nk}^2)$.

(ii) can be similarly proved.

(c) Follows easily from part (a) and Lemma 4.6(c).

(d) The proof follows from arguments similar to those in part (b).

Proof of Lemma 4.8. Let us note that $\hat\theta_{kj}$ $(j=1, ..., n)$ is the solution of

$\sum_{i \neq j} w_1(Y_i, \hat{\theta}'_{kj}\psi(X_i))\psi_k(X_i) = 0$, or, $n^{-1}\sum_{i=1}^{n} w_1(Y_i, \hat{\theta}_{kj}\psi_k(X_i))\psi_k(X_i) = n^{-1}w_1(Y_j, \hat{\theta}'_{kj}\psi_k(X_j))\psi_k(X_j)$; $\hat{\theta}_k$ is the solution of $n^{-1}\sum_{i=1}^{n} w_1(Y_i, \hat{\theta}'_k\psi_k(X_i))\psi_k(X_i) = 0$.

Noting that $\max\{|Y_j|: j = 1, ..., n\} = o_p((\log n)^{1/2+\delta})$, the result follows by the use of Lemma 6.3 (uniformly in $j = 1, ..., n$) by taking

$$u(\theta_0) = -n^{-1}\sum_{i=1}^{n} w_1(Y_i, \hat{\theta}'_k\psi_k(X_i))\,\psi_k(X_i),$$

$$\varepsilon = -n^{-1}w_1(Y_j, \hat{\theta}'_{kj}\psi_k(X_j))\psi_k(X_j),$$

$$\theta_0 = \hat{\theta}_k, \qquad \theta_1 = \hat{\theta}_{kj}.$$

(b) Follows easily by using Lemma (6.3) twice.

Proof of Lemma 4.9. (a) Since $s_k \neq f$ for any k, then there exists a $k_n \leqslant K_n$ such that $\inf\{n\,\|s_k - f^*\|_G^2 : k \leqslant k_n\} \to \infty$. Using (5.1), we have

$$\inf\{|nV_n(k)|: k \leqslant K_n\} \geqslant c_9[\lambda_{k_n} + \inf\{n\,\|s_k - f^*\|_G^2 : k \leqslant k_n\}] \to \infty.$$

(b) Let $\max\{nV_n(k): k \leqslant K_n\} = nV_n(k^*)$. Because of part (a), $k^* \to^P \infty$. So, $\sum_{k \leqslant K_n}(nV_n(k))^{-2} = \sum_{k \leqslant k^*} + \sum_{k^* < k \leqslant K_n}$. The first term is $O(1)k^*/\lambda_k^* \to^P 0$ and the second term is $O(1)\sum_{k > k^*}\lambda_k^{-1} \to 0$.

Proof of Lemma 4.10. Part (b) is proved the same way as part (a) (see Stone [16]).

Proof of Lemma 6.3. The proof is quite simple and so we will skip the proof.

References

[1] Breiman and Friedman (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Asoc.* **80** 580–619.

[2] Burman, P. (1985). A data dependent approach to density estimation. *Z. Wahrsch. Verw. Gebiete* **69** 609–628.

[3] Burman, P., and Chen, W. K. (1989). Nonparametric estimation of a regression function. *Ann. Statist.*, December issue.

[4] deBoor, C. (1978). *A Practical Guide to Splines.* Springer-Verlag, New York.

[5] Duin, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. Comput.* **C-25** 1175–1179.

[6] Von Golitschek, M., and Schumaker, L. (1988). Data fitting by penalized least squares. Manuscript.

[7] Golub, G., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameters. *Technometrics* **21** 215–223.

[8] Habbema, J., Hermans, J., and Van den Broek, K. (1974). A stepwise discriminant analysis program using density estimation. In *COMPSTAT 1974, Proceedings in Computational Statistics* (G. Bruckmann, Ed.). Physica Verlag, Vienna.

[9] HARDLE, W., AND MARRON, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.

[10] HASTIE, T. J., AND TIBSHIRANI, R. J. (1984). *Generalized Additive Models.* Technical Report, Division of Biostatistics, Stanford University.

[11] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–20.

[12] KOO, C. Y. AND STONE, C. (1986). Log spline density estimation. Manuscript.

[13] LI, K. C. (1987). Asymptotic optimality for c_p, c_L, cross-validation and generalized cross-validation. *Ann. Statist.* **15** 958–975.

[14] MCCULLAGH, P., AND NELDER, J. A. (1983). *Generalized Linear Models.* Chapman & Hall, London.

[15] O'SULLIVAN, F., YANDELL, B., AND RAYNOR, W. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81** 96–103.

[16] STONE, C. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **10** 1040–1053.

[17] STONE, M. (1974). Cross-validation and multinomial prediction. *Biometrika* **61** 509–615.

[18] WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **3** 302–305.