The Annals of Statistics
1992, Vol. 20, No. 4, 1844–1874

EMPIRICAL SMOOTHING PARAMETER SELECTION IN ADAPTIVE ESTIMATION[1]

BY KUN JIN

Florida State University

We provide a solution to the smoothing parameter selection problem involved in the construction of adaptive estimates for the symmetric location model and the general linear model. Linear B-splines are used to give a simple form of the estimate of the score function of the underlying density. New empirical methods are proposed to locate the knots optimally and to select the number of knots. We also give asymptotic bounds for the empirical selection method and show that an estimate with an empirically selected smoothing parameter is adaptive. Our estimates are easy to compute and possess useful computational features. Simulation studies reveal that our estimates perform well in comparison with some well-known estimates.

1. Introduction. The development of adaptive estimation theory has had a rich history since Stein proposed the idea in 1956. The large sample theory based on deterministic bandwidths is now well established. However, the problem of selecting the smoothing parameter involved in the construction of adaptive estimates has continued to be a major obstacle to applying the technique in practical problems. In this paper, we provide a solution to this problem for the symmetric location model and the general linear regression model.

First, let us describe the models to be considered:

MODEL I (Symmetric location model)

$$Y_i = \theta + e_i,$$

where $\theta \in R$ and $\{e_i\}$ iid $\sim f \in \mathscr{F}$, where $\mathscr{F}$ is a class of density functions symmetric about zero.

MODEL II (the general linear model)

$$Y_i = \alpha + X_i^t\beta + e_i,$$

where $\alpha \in R$, $\beta, X_i \in R^p$, $\{X_i\}$ are iid bounded random vectors and $\{e_i\}$

Received September 1990; revised January 1992.

[1] Work supported in part by Office of Naval Research Grant ONR N00014-89-J-1563 to the Department of Statistics at the University of California, Berkeley.

AMS 1980 *subject classifications*. Primary 62F35; secondary 62F11, 62G20, 62J05.

Key words and phrases. Adaptation, efficient estimation, B-splines, cross-validation, linear regression.

iid $\sim f \in \mathscr{F}$, where $\mathscr{F}$ is a class of density functions. $\{X_i\}$ and $\{e_i\}$ are independent. The slope parameter β is identifiable and represents the parameter of prime interest.

Stone (1975) derived an adaptive estimate for Model I. Bickel (1982) established a rigorous definition and necessary condition for adaptation and constructed the adaptive estimates for several important models, including Models I and II. The basic idea in the construction of adaptive estimates is to replace a score function $\phi = (\log f)'$ in the one-step maximum likelihood estimation by an estimator $\hat{\phi}_\lambda$, where $\hat{\phi}_\lambda$ is obtained from a kernel density estimate $\hat{f}_\lambda$ and its derivative $\hat{f}'_\lambda$, and λ is a bandwidth (a smoothing parameter). The large-sample theory has been established with a deterministic sequence of λ_n that converges to 0 at a slow rate.

The practical problems often come with a sample of small or moderate size. The large-sample theory offers very little that is useful regarding selection of a smoothing parameter in these situations. Hsieh and Manski (1987) demonstrated that, in Monte Carlo simulation studies, the behavior of an adaptive estimate could be changed dramatically by using different smoothing parameters. Therefore, the method for selecting a smoothing parameter becomes a crucial issue in the implementation of adaptive estimation. This difficulty is the major reason that there are so few applications of adaptive estimation.

Smoothing parameter selection has been a difficult issue in density estimation problems, and empirical bandwidth selection in kernel density estimation has been the subject of considerable study. Stone (1984) and Hall and Marron (1987) proposed the asymptotic optimality theory of the empirical bandwidth selection based on least squares cross-validation, but it is well recognized that this method is subject to large sample variation. It offers us little help with our problem.

Hsieh and Manski (1987) first attempted smoothing parameter selection in the adaptive estimation problem for Model II. They used kernel estimation to estimate the density f and its derivative f'. Then, for a given choice of bandwidth λ, an estimate $\hat{v}_\lambda$ of the MSE of $\hat{\beta}$ was obtained by bootstrapping. The empirical choice $\hat{\lambda}$ of λ was chosen as the λ minimizing $\hat{v}_\lambda$ among a preselected set of bandwidths. Faraway (1992) worked on this problem for Model I and Model II by using the logsplines method instead of the kernel method and by estimating the MSE directly. The smoothing parameter was k, the number of knots. The empirical selection rule was to choose a $\hat{k}$ that minimized the estimate of MSE over all $k \le K$, where K is a preselected integer. Both of their works produced interesting simulation results and showed that the empirical selection method is worthwhile. However, three problems remain. One is showing that the estimates of Hsieh and Manski and of Faraway, based on empirical smoothing parameter selection, are indeed adaptive estimates. Since both studies involved selecting the smoothing parameter from a preselected range of smoothing parameters, this left a problem of determining the range. Finally, their methods require intensive computing time and are not easy to implement.

In this paper, we solve these problems. We develop new empirical smoothing parameter selection methods. Instead of estimating the density function first, we use the linear B-splines to estimate the score function directly and derive a simple form of the estimate $\hat{\phi}_k$ of ϕ, where k, the number of knots, is a smoothing parameter (see Section 4). The empirical selection rule based only on cross-validation does not work well in simulations. Therefore, we propose to form an interval of possible k based on cross-validation criteria and then introduce a new method, termed *stationary correction*, to select $\hat{k}_n$ (see Section 6). The key feature of this new approach is that it enables us to develop the asymptotic theory for the empirical selection rule as well as providing a flexibility to modify the selection rule to achieve better numerical results for small or moderate sample sizes. We also propose to pick the *first local minimizer* of k in the process od determining $\hat{k}_n$ (see Section 6). The first local minimization approach makes the selection method entirely empirical without the need for a preselected range of k. This approach also efficiently eliminates the large variance in small sample situations that, as Friedman and Silverman (1989) pointed out, is a drawback to using this type of regression spline. We also propose an *optimal forward approach* to locate the knots (see Section 5) that, although derived via different motivation, is in the same spirit as the method proposed by Friedman and Silverman (1989).

We establish asymptotic bounds on $\hat{k}_n$ and, for the first time, prove that an estimate with an empirically selected smoothing parameter is adaptive. The main theorems are presented in Section 8, and their proofs are given in Sections 10 and 11.

Our estimates are easy to compute and possess useful computational features, such as stable performance regardless of choice of initial estimate and fast convergence without iterative computation. Simulation studies reveal that our estimates perform well in comparison with some well-known estimates, such as Hampel's three-part redescending M-estimate in location case and perform better than Huber's M-estimates in regression case (see Section 9).

2. Basic setup. Let $\tilde{\theta}_n$, $\tilde{\alpha}_n$ and $\tilde{\beta}_n$ be $\sqrt{n}$-consistent estimates of θ, α and β. It will be convenient for proving theorems to take the "discretized" versions of $\tilde{\theta}_n$, $\tilde{\alpha}_n$ and $\tilde{\beta}_n$. The idea of discretization is due to Le Cam (1969) and was employed by Bickel (1982) in these contexts. Let $R_n = \{n^{-1/2}i$, where i is an arbitrary integer$\}$ and define the discretized estimate $\bar{\theta}_n =$ the point in R_n closest to $\tilde{\theta}_n$. Similarly define $\bar{\alpha}_n, \bar{\beta}_n$.

Take the residuals

$$\bar{e}_i = Y_i - \bar{\theta}_n \tag{1}$$

in Model I and

$$\bar{e}_i = Y_i - \bar{\alpha}_n - X_i^t\bar{\beta}_n \tag{2}$$

in Model II. Notice that $\bar{e}_i$ also depends on n. Let F_n be the empirical distribution of $\{\bar{e}_i\}$.

For a given $\phi(x)$, let $I(\phi) = \int \phi^2 f dx$. Then the one-step mle's of θ and β are

$$\hat{\theta}_{\text{mle}} = \bar{\theta}_n - \frac{\sum_{i=1}^n \phi(\bar{e}_i)}{nI(\phi)}, \tag{3}$$

$$\hat{\beta}_{\text{mle}} = \bar{\beta}_n - \frac{\sum_{i=1}^n (X_i - \bar{X})\left(\widehat{\text{Var}}(X)\right)^{-1} \phi(\bar{e}_i)}{nI(\phi)}, \tag{4}$$

where

$$\widehat{\text{Var}}(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^t.$$

In order to get the adaptive estimates, we need to find a good estimate $\hat{\phi}(x)$ of $\phi(x)$. Before doing that, we will introduce the notation to be used.

3. Relevant notation. On an interval (b_l, b_r), for any integer k, let the knots $\{\xi\}_k$ be $b_l = \xi_{k(0)} < \xi_{k(1)} < \cdots < \xi_{k(k)} = b_r$. Define the linear B-spline basis $B_{k(i)}(x)$, $i = 1, \dots, k$, as follows:

$$B_{k(i)}(x) = \begin{cases} \dfrac{x - \xi_{k(i-1)}}{\xi_{k(i)} - \xi_{k(i-1)}}, & \text{if } \xi_{k(i-1)} \le x \le \xi_{k(i)}, \\ \dfrac{\xi_{k(i+1)} - x}{\xi_{k(i+1)} - \xi_{k(i)}}, & \text{if } \xi_{k(i)} < x \le \xi_{k(i+1)} \text{ and } i < k, \\ 1, & \text{if } \xi_{k(k)} \le x \text{ and } i = k, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Let $D_{k(i)}(x)$, $i = 1, \dots, k$, be their piecewise derivatives. For Model I, we take $b_l = -b_r$, define $B_{k(i)}(x)$ only on $(0, b_r)$ and then extend it antisymmetrically to $(b_l, 0)$.

Denote $B_k(x) = (B_{k(1)}(x), \dots, B_{k(k)}(x))^t$, $D_k(x) = (D_{k(1)}(x), \dots, D_{k(k)}(x))^t$ and $A_k(x) = B_k \cdot B_k^t(x)$. Define

$$B_k(F) = \left(\int B_{k(1)}(x)\, dF(x), \dots, \int B_{k(k)}(x)\, dF(x) \right)^t.$$

Similarly define $D_k(F)$, $A_k(F)$, $B_k(F_n)$, $D_k(F_n)$ and $A_k(F_n)$.

Let $\Delta\xi_{k(i)} = \xi_{k(i)} - \xi_{k(i-1)}$, $i = 1, \dots, k$, and $|\xi_k| = \max_{1 \le i \le k} \Delta\xi_{k(i)}$. For each k, let $\mathscr{L}_{k,\xi}$ be the linear space generated by $\{B_{k(i)}(x)\colon i = 1, \dots, k\}$. For any function g, letting $\|g\| = \sup_x |g(x)|$, define the distance from g to $\mathscr{L}_{k,\xi}$ as $d(g, \mathscr{L}_{k,\xi}) = \min_{a \in R^k} \|g(x) - a^t B_k(x)\|$.

For any matrix $A = (a_{ij})_{n\times m}$, define the norms

$$\|A\|_1 = \max_{1\le i\le n} \sum_{j=1}^{m} |a_{ij}|, \qquad \|A\| = \sum_{i=1}^{n} \sum_{j=1}^{m} |a_{ij}| \quad \text{and} \quad \|A\|_\infty = \max_{i,j} |a_{i,j}|.$$

For any function g, denote $g_{a,b}(x) = g(x)$, if $a \le x \le b$, and 0 otherwise. Let $I_b(\phi) = \int \phi^2_{b_l,b_r} f\,dx = \int_{b_l}^{b_r} \phi^2 f\,dx$.

4. Smoothing on ϕ. Now we consider how to get an estimate of ϕ. The method dicussed here is the simplest one proposed so far, yet it does perform very well. To begin with, consider how to set the intervals on which the B-splines are built.

4.1. *Nested intervals.* The domain of $\phi(x)$ is often infinite. Usually, a sequence of finite intervals, on which B-splines are built, is taken to approach the real domain, but the situation becomes very complicated when we consider empirical smoothing parameter selection on such intervals. Here, a nested intervals method is proposed.

Let b_l and b_r be percentiles of the underlying distribution, say, the 5th and 95th percentiles. For smoothing parameter selection, we will only estimate the truncated score function ϕ_{b_l,b_r} instead of ϕ. It is reasonable to believe that, as long as the interval (b_l, b_r) is sufficiently wide, there should be no significant difference between estimating ϕ_{b_l,b_r} and estimating ϕ in terms of selecting the smoothing parameter. Practically, b_l and b_r are estimated by empirical quantiles of F_n, say, $\bar{e}_{(0.05n)}$ and $\bar{e}_{(0.95n)}$. For Model I, we restrict $b_l = -b_r$.

In constructing adaptive estimates, it is not enough to estimate ϕ only on (b_l, b_r), since the estimates would not be efficient. Let d_{rn}, d_{ln} be real sequences that converge to $+\infty$. Take $b_{l_n} = b_l - d_{ln}$ and $b_{rn} = b_r + d_{rn}$. Then $(b_l, b_r) \subset (b_{ln}, b_{rn})$. We will estimate ϕ on the enlarged interval (b_{ln}, b_{rn}) in order to construct adaptive estimates. For Model I, we take $d_{ln} = d_{rn}$. So $b_{ln} = -b_{rn}$.

Later on, we will discuss how to locate the knots on (b_l, b_r). Now we consider how to place knots in the intervals (b_{ln}, b_l) and (b_r, b_{rn}), given that k knots have been placed in (b_l, b_r). Let $|\xi_k| = \max_{1\le i\le k} \Delta\xi_{k(i)}$ in (b_l, b_r). Then, place the knots in (b_{ln}, b_l) at distances $|\xi_k|, 2|\xi_k|, \dots$ from b_l until the endpoint b_{ln} is reached. Do that on (b_r, b_{rn}) as well.

Denote the number of knots in (b_{ln}, b_{rn}) by k'. It is easy to see that for a given (b_{ln}, b_{rn}) and k knots placed in (b_l, b_r), there is a unique k' corresponding to k.

The reason for introducing (b_{ln}, b_{rn}) is only to ensure adaptiveness of the estimates when the sample size goes to ∞. Practically, if we have a small- or moderate-sized sample, there are only a few e_i's outside (b_l, b_r). That is, with a sample of small or moderate size, the computations can be restricted to (b_l, b_r) only.

The nested intervals method is useful for developing the large-sample theory and studying the small-sample performance of our approach.

4.2. *Spline interpolation and estimation of ϕ.* Now we discuss how to interpolate and estimate ϕ on (b_l, b_r). We take the expected squared error, $\int(g - \phi)^2 f dx$, known to be a key measure regarding the adaptiveness of estimates, as the loss function.

Given any k knots on (b_l, b_r), let $B_k(x)$ be the B-splines basis defined in Section 3. The interpolation of $\phi(x)$ is defined as $a_k^t(F)B_k(x)$, where $a_k(F)$ minimizes $\int_{b_l}^{b_r}(a_k^t B_K(x) - \phi(x))^2 f dx$ for all $a_k \in R^k$. By partial integration,

$$\int_{b_l}^{b_r}\left(a_k^t B_k(x) - \phi(x)\right)^2 f dx$$

$$(6) \qquad = a_k^t\left(\int_{b_l}^{b_r} B_k B_k^t(x) f dx\right)a_k - 2a_k^t \int_{b_l}^{b_r} B_k \phi f dx + \int_{b_l}^{b_r}\phi^2 f dx$$

$$= a_k^t A_k(F) a_k + 2a_k^t D_k(F) + I_b(\phi).$$

It is easy to see that minimizing (6) is equivalent to minimizing $a_k^t A_k(F)a_k + 2a_k^t D_k(F)$. So the $a_k(F)$ exists and $a_k(F) = -A_k^{-1}(F)D_k(F)$. Then ϕ_{b_l, b_r} is interpolated as $\phi_k(x) = a_k^t(F)B_k(x)$. Naturally, we take $\hat{a}_k = a_k(F_n) = -A_k^{-1}(F_n)D_k(F_n)$ as the estimate of $a_k(F)$, and we estimate ϕ_{b_l, b_r} by $\hat{\phi}_k(x) = a_k^t(F_n)B_k(x)$.

Similarly, for k' knots on (b_{l_n}, b_{rn}), let $B_{k'}(x)$ be the B-splines basis on (b_{ln}, b_{rn}). We have that the interpolation of $\phi_{b_{ln}, b_{rn}}$ is $\phi_{k'}(x) = a_{k'}^t(F)B_{k'}(x)$, and the estimator is $\hat{\phi}_{k'}(x) = a_{k'}^t(F_n)B_{k'}(x)$.

This approach also could be used in logsplines density estimation, where it also yields an easily computable algorithm. The partial integration in (6) was first used by Cox (1985) to interpolate ϕ by smoothing splines.

REMARK. The simplicity of linear B-splines is the main reason for employing them in the estimation of ϕ. However, it should be noted that using high-order B-splines (e.g., parabolic or cubic splines) does not give satisfactory simulation results with small sample sizes. In Section 9, we shall see that $k = 0$ or 1 is suitable for sample size $n = 40$ (which means taking a linear spline with at most one interior knot). A parabolic spline must start with at least one interior knot, and a cubic spline with two interior knots. For small sample size problems, those high-order splines seem to be "oversmoothing."

5. Knot placement. We now give two data-driven approaches for locating the knots.

5.1. *Equally spaced quantiles method.* For a given integer k, let $\{\bar{e}_{(i)}\}$, $i = 1, \dots, m(n)$, be the order statistics of the residuals $\{\bar{e}_i\}$ in (b_l, b_r). Then, the knots are $\xi_{k(i)} = \bar{e}_{(i \cdot m(n)/(k+1))}$, $i = 1, \dots, k$. This simple method is motivated by the idea that the data should be evenly distributed among the knots in the interval. Faraway (1992) took this approach.

Simulations indicate that this approach occasionally gives an inadequate fit. However, this approach behaves well when combined with the optimal selection rule $\hat{k}_n$ given in Section 6.

5.2. *Optimal forward method.* This approach, although motivated differently, is in the same spirit as the method proposed by Friedman and Silverman (1989).

We consider the nested knot sequence $\{\xi\}_k \subseteq \{\xi\}_{k+1}$, for $k = 0, 1, \dots$. Start with $k = 0$ and add selected knots consecutively to the knot set. Recall that ϕ is interpolated by $\phi_k = a_k^t(F)B_k(x)$. Denote the bias of the interpolation as

$$\text{Bias}(k, \phi) = \int_{b_l}^{b_r} \left(a_k^t(F)B_k(x) - \phi(x)\right)^2 f\,dx. \tag{7}$$

The idea is to reduce Bias(k, ϕ) as much as possible during the process of adding knots.

It is easy to see that if ϕ is not a piecewise linear function, Bias(k, ϕ) decreases monotonically to 0 as $k \to \infty$. Denote $\Delta(k, \xi_{k+1}, F) = \text{Bias}(k, \phi) - \text{Bias}(k + 1, \phi) > 0$, where ξ_{k+1} is the $(k + 1)$th knot to be added to $\{\xi\}_k$. The quantity $\Delta(k, \xi_{k+1}, F)$ can be thought of as a measure of the effect of adding ξ_{k+1} to $\{\xi\}_k$. With several possible candidates of ξ_{k+1}, the greater value of $\Delta(k, \xi_{k+1}, F)$ implies that Bias$(k + 1, \phi)$ becomes smaller after adding ξ_{k+1}. If possible, we wold like to pick ξ_{k+1} with the largest $\Delta(k, \xi_{k+1}, F)$. By the definition of $a_k(F)$, we can see that Bias$(k, \phi) = -D_k^t(F)A_k^{-1}(F)D_k(F) + I_b(\phi)$. Thus

$$\begin{aligned}\Delta(k, \xi_{k+1}, F) &= \text{Bias}(k, \phi) - \text{Bias}(k + 1, \phi)\\ &= D_{k+1}^t(F)A_{k+1}^{-1}(F)D_{k+1}(F) - D_k^t(F)A_k^{-1}(F)D_k(F).\end{aligned}$$

We propose to estimate Δ by

$$\begin{aligned}\hat{\Delta}(k, \xi_{k+1}, F_n) = {}& D_{k+1}^t(F_n)A_{k+1}^{-1}(F_n)D_{k+1}(F_n)\\ &- D_k^t(F_n)A_k^{-1}(F_n)D_k(F_n).\end{aligned} \tag{8}$$

Now we discuss how to choose the nested sequence of knots on (b_l, b_r). For $k = 2^i$, $i = 0, 1, \dots$, let $\{\xi\}_k$ be the 2^i equally spaced points on (b_l, b_r), with b_r being taken as a knot. Clearly, $\{\xi\}_{2^i} \subset \{\xi\}_{2^{i+1}}$. Next, we deal with the situation when $2^{i-1} < k < 2^i$. Since we have already dealt with $k = 2^{i-1}$, take 2^{i-1} middle points of the intervals divided by $\{\xi\}_{2^{i-1}}$ as candidates. For each candidate ξ_{k+1}, we compute $\hat{\Delta}(k, \xi_{k+1}, F_n)$ and then order all candidates according to the magnitude of $\hat{\Delta}(k, \xi_{k+1}, F_n)$: first the candidate with the largest $\hat{\Delta}(k, \xi_{k+1}, F_n)$, etc. Then, $\{\xi\}_k$ can be obtained by entering these ordered candidates into $\{\xi\}_{k-1}$ sequentially.

The method can be extended in several ways. Instead of dividing each interval into two subintervals, we can take m subintervals. In practice, the equally spaced points can be changed to equally spaced quantiles estimated by empirical quantiles.

6. Empirical selection of k.

To choose the number of knots, we would like to pick $\hat{k}$ to minimize

$$\int_{b_l}^{b_r}\left(a_k^t(F_n)B_k(x) - \phi(x)\right)^2 f\,dx. \tag{9}$$

This cannot be done since (9) depends on the unknown f. This is a typical situation in which cross-validation can be applied.

Minimizing (9) is equivalent to minimizing

$$L(k, F_n, F) = a_k^t(F_n)A_k(F)a_k(F_n) + 2a_k^t(F_n)D_k(F).$$

Split the residuals into $\bar{e}_1, \dots, \bar{e}_{n_1}, \bar{e}_{n_1+1}, \dots, \bar{e}_{n_1+n_2}$, where $n_1 + n_2 = n$. We can estimate $L(k, F_n, F)$ by

$$L(k, F_{n_1}, F_{n_2}) = a_k^t(F_{n_1})A_k(F_{n_2})a_k(F_{n_1}) + 2a_k^t(F_{n_1})D_k(F_{n_2}),$$

where F_{n_1} and F_{n_2} are the empirical distribution functions of $\{\bar{e}_1, \dots, \bar{e}_{n_1}\}$ and $\{\bar{e}_{n_1+1}, \dots, \bar{e}_{n_1+n_2}\}$, respectively. We find that a reasonable choice of n_1 and n_2 is $n_1 = \min\{3n^{1/2}, n/2\}$ and $n_2 = n - n_1$. We propose to select the *first local minimizer* $\hat{k}_{\text{cv}}$ of $L(k, F_{n_1}, F_{n_2})$ as a cross-validatory estimate of k, that is, to choose $\hat{k}_{\text{cv}}$ satisfying

$$L(1, F_{n_1}, F_{n_2}) \geq \cdots \geq L\left(\bar{k}_{\text{cv}}, F_{n_1}, F_{n_2}\right) < L\left(\hat{k}_{\text{cv}} + 1, F_{n_1}, F_{n_2}\right). \tag{10}$$

Simulations at small or moderate sample sizes do not give satisfactory results using this method. This is due to the splitting of the sample in constructing $L(k, F_{n_1}, F_{n_2})$, as well as the fact that the cross-validation method suffers a fairly large sample variation. We propose a new method, *stationary correction*, to overcome this difficulty.

Split the residuals again, this time into $\bar{e}_1, \dots, \bar{e}_{n_2}, \bar{e}_{n_2+n_1}$. Let F'_{n_2} and F'_{n_1} to be the empirical distribution functions of $\{\bar{e}_1, \dots, \bar{e}_{n_2}\}$ and $\{\bar{e}_{n_2+1}, \dots, \bar{e}_{n_2+n_1}\}$, respectively. These will give another estimate of $L(k, F_n, F)$, namely, $L(k, F'_{n_1}, F'_{n_2})$. Similarly, pick the first local minimizer $\hat{k}'_{\text{cv}}$ as defined in (10). Suppose $\hat{k}'_{\text{cv}} \leq \hat{k}_{\text{cv}}$. We create an interval of possible k's as $I(n) = \{k\colon \hat{k}'_{\text{cv}} \leq k \leq \hat{k}^2_{\text{cv}}\}$.

Every k in $I(n)$ is an empirical selection $\hat{k}_n$ of k. Theorems 4 and 5 in Section 10 reveal that, from an asymptotic point of view, such a $\hat{k}_n$ is suitable for constructing our adaptive estimate. The skey issue now is to select a $\hat{k}_n$ for which the adaptive estimate has adequate numerical performance for small or moderate sample sizes.

It is well recognized that the cross-validation method suffers from large sample variation. This variation appears to cause unstable behavior of $\hat{\phi}_k$; even consecutive $\hat{\phi}_k$ can be quite different from one another. We want to select a $\hat{k}_n$ that would produce a stable $\hat{\phi}_{\hat{k}_n}$. We begin by considering whether there is a k_n in $I(n)$ minimizing $d(k) = \int(\hat{\phi}_{k-1} - \hat{\phi}_k)^2\,dF(x)$ so that $\hat{\phi}_{k_n-1}$ and $\hat{\phi}_{k_n}$

are very close and can be considered as stable estimators of ϕ. This motivation is further supported by the following fact: If k_n is a local minimizer of $L(k) = \int(\hat{\phi}_k - \phi)^2\,dF(x)$, then k_n is a stationary point of the function $L(k)$ and $d(k_n)$ will have a small value. We estimate $d(k)$ by $\hat{d}(k) = \int(\hat{\phi}_{k-1} - \hat{\phi}_k)^2\,dF_n(x)$. Simulation shows that minimizing $\hat{d}(k)$ within $I(n)$ yields good results. Noticing that there is fairly large variation among $\{d(k)\}$, we introduce a smoothed version of $\hat{d}(k)$:

$$ST(k, F_n) = \frac{1}{k}\sum_{j=0}^{k-1}\int_{b_l}^{b_r}\left(a_j^t(F_n)B_j(x) - a_k^t(F_n)B_k(x)\right)^2 dF_n.$$

Define $\hat{k}_n$ to be the first local minimizer of $ST(k, F_n)$ over $k \in I(n)$. That is, choose $\hat{k}_n$ within $I(n)$ satisfying

$$ST(\hat{k}'_{\mathrm{cv}}, F_n) \geq \cdots \geq ST(\hat{k}_n, F_n) < ST(\hat{k}_n + 1, F_n).$$

We call $\hat{k}_n$ a *stationary correction* to $\hat{k}_{\mathrm{cv}}$. If there is no such $\hat{k}_n$ within $I(n)$, choose $\hat{k}_n = \hat{k}^2_{\mathrm{cv}}$. From the above motivation, we can see that both $\hat{k}_n$ and $\hat{k}_n - 1$ should be good choices of k. Let $\hat{I}_k(\phi) = D_k^t(F_n)A_k^{-1}(F_n)D_k(F_n)$. Then $\hat{I}_k(\phi)$ is an estimate of $I(\phi)$. The final selection $\hat{k}_n$ is defined as the one of $\hat{k}_n$ and $\hat{k}_n - 1$ giving the large $\hat{I}_k(\phi)$.

On the enlarged interval (b_{l_n}, b_{rn}), we use $\hat{k}'_n$ (defined in Section 4.1) as the empirical selection rule. Simulation studies show that stationary correction rule works very well.

To summarize, cross-validation is used to find $I(n)$ that possess a desirable asymptotic property (Theorems 4 and 5). The stationary correction rule gives good numerical results at small sample sizes. In Section 9, we present some graphics to show more of what is going on for $\hat{k}_n$ and $\hat{k}_{\mathrm{cv}}$.

The idea of using the first local minimizer has been consistently employed throughout our optimization process. This makes $\hat{k}_n$ easily computable and also produces good simulation results. Bickel initially suggested that the author consider the first local minimizer of k in this research and noted that a small number of knots corresponds to a large kernel bandwidth. The first local minimization method essentially follows the same practice used in kernel density estimation. In studying empirical bandwidth choice in kernel density estimation, Rudemo (1982) pointed out that very small bandwidth may cause irregular behavior. Park and Marron (1990) further noted that if multiple local minimia arise when using a cross-validation selection rule, the largest local minimizer should be used.

Remark 1. $\hat{k}'_{\mathrm{cv}}$ in $I(n)$ actually refers to the smaller of $\hat{k}'_{\mathrm{cv}}$ and $\hat{k}_{\mathrm{cv}}$. We keep the same notation for convenience. The computing procedure is as follows: Initially, inspect both $L(k, F_{n_1}, F_{n_2})$ and $L(k, F'_{n_1}, F'_{n_2})$ to find $\hat{k}'_{\mathrm{cv}}$. Once $\hat{k}'_{\mathrm{cv}}$ is found, say, from $L(k, F'_{n_1}, F'_{n_2})$, then start to inspect $ST(k, F_n)$ as well. As soon as $\hat{k}_n$ is found, stop the procedure. Simulation reveals that in almost all cases (99.8% in the location model) $\hat{k}_n$ is found before $\hat{k}_{\mathrm{cv}}$. The computation takes little computer time.

Remark 2. Choosing $\hat{k}^2_{\text{cv}}$ as the right endpoint of $I(n)$ gives the largest possible interval over which $ST(k, F_n)$ can range. This is permissible in terms of the asymptotic theory, and has no effect on the actual computation due to Remark 1.

Remark 3. Choosing $n_1 \sim O(n^{1/2})$ and $n_2 \sim O(n)$ is sufficient for our asymptotic results. However, for small n, $n^{1/2}$ is too small for practical use, so we take $n_1 = \min\{3n^{1/2}, n/2\}$. Such an n_1 is usable when n is small and has a rate of $O(n^{1/2})$ when $n \to \infty$. Further research on how to choose the optimal rate is needed.

7. Assumptions.

We assume the following:

(A-1) $I(\phi) = \int \phi^2(x) f(x)\, d < \infty$.

(A-2) In Model II, $\operatorname{Var}(X) > 0$.

(A-3) $\phi \in C^{(2)}(R)$.

(A-4) $\|f\| \le b(f) < +\infty$, $\|f'\| \le l(f') < +\infty$ and $f(x) > 0$, for all $x \in R$.

(A-5) There are constants c_1 and c_2 such that $c_1(b_r - b_l)/k \le \Delta\xi_i < c_2(b_r - b_l)/k$, $i = 1, \dots, k$, for all $k = 1, 2, 3, \dots$.

(A-6) $\operatorname{Bias}(k, \phi)$ decreases monotonically to 0 as $k \to \infty$.

(A-7) There exists $\alpha_L > 4$ such that $\liminf_{k \to \infty} k^{\alpha_L}(\operatorname{Bias}(k, \phi) - \operatorname{Bias}(k + 1, \phi)) > 0$.

(A-8) The initial estimates are $\sqrt{n}$-consistent.

(A-9) There exists a constant $c(f) > 0$ such that, for $|x| \ge c(f)$, $f(x)$ monotonically decreases as $|x|$ increases.

(A-10) There exists a constant $\gamma(\mathscr{F}) > 2$ such that $f(x)\exp\{x^{\gamma(\mathscr{F})}\} \to \infty$.

(A-11) d_{ln}, d_{rn} are of rate $O((\log\log n)^{1/\gamma(\mathscr{F})})$.

(A-12) For any $\varepsilon > 0$, $\|\phi_n^{(2)}(x)\|/n^{\varepsilon} \to 0$, where $\phi_n(x) = \phi_{b_{ln}, b_{rr}\mathrm{n}}(x)$.

Remarks. These conditions are not very restrictive:

(i) (A-5) holds under all of the knot placement methods discussed above.

(ii) A sufficient condition for (A-6) is that the knot sequence $\{\xi\}_k$ be nested and ϕ not be piecewise linear.

(iii) Let $b_n(f') = \max_{b_{ln} \le x \le b_{rn}} |f'|$, $f_{\min,n} = \min_{b_{ln} \le x \le b_{rn}} f(x)$, $f_{\max,n} = \max_{b_{ln} \le x \le b_{rn}} f(x)$. Under conditions (A-9), (A-10) and (A-11), we have that, for any $\varepsilon > 0$,

$$\frac{d_{ln} + d_{rn}}{n^{\varepsilon}} \to 0, \tag{11}$$

$$n^{\varepsilon} f^2_{\min,n} \to +\infty, \tag{12}$$

$$\frac{f^3_{\max,n} b_n(f')}{f^5_{\min,n} n^{\varepsilon}} \to 0. \tag{13}$$

These three conditions are what we really need in Section 11. However, they are not easily understood. Instead, we use conditions (A-9), (A-10) and (A-11) since they are less restrictive and more easily checked.

(iv) If we further assume that $|f^{(i)}| \le c$, $i = 1, 2, 3$ (which is often true because, for most commonly seen densities, we actually have $|f^{(i)}(x)| \to 0$, as $x \to \infty$), then (A-12) holds under (A-9)–(A-11).

8. Main theorems. Define the estimate of $I(\phi)$ as $\hat{I}_{k'}(\phi) = D^t_{k'}(F_n) A^{-1}_{k'}(F_n) D_{k'}(F_n)$. Substituting $\phi(x)$ and $I(\phi)$ in (3) and (4) by $\hat{\phi}_{k'}(x)$ and $\hat{I}_{k'}(\phi)$, we have

$$\hat{\theta}_{n,k'} = \bar{\theta}_n - \frac{a^t_{k'}(F_n) B_{k'}(F_n)}{\hat{I}_{k'}(\phi)},$$

$$\hat{\beta}_{n,k'} = \bar{\beta}_n - \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(\widehat{\mathrm{Var}}(X)\right)^{-1} a^t_{k'}(F_n) B_{k'}(\bar{e}_i)}{n \hat{I}_{k'}(\phi)},$$

where $\hat{\phi}_{k'}(x)$ is defined in Section 4. Recall that $A_{k'}(x)$, $D_{k'}(x)$, $B_{k'}(x)$ are defined on (b_{ln}, b_{rn}), and k' is the number of knots on (b_{ln}, b_{rn}).

Let $\hat{k}'_n$ be the empirical selection rule defined in Section 4.1. Then the adaptive estimates are defined as

$$\hat{\theta}_n = \hat{\theta}_{n,\hat{k}'_n} \tag{14}$$

for the location parameter θ in Model I and

$$\hat{\beta}_n = \hat{\beta}_{n,\hat{k}'_n} \tag{15}$$

for the slope parameter β in Model II.

We have the following theorem.

THEOREM 1. *Under the conditions of Section* 7, $\hat{\theta}_n$ *and* $\hat{\beta}_n$ *defined in* (14) *and* (15) *are adaptive estimates, that is,*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to_L N(0, I^{-1}(\phi));$$

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \to_L N\left(0, (\mathrm{Var}(X) I(\phi))^{-1}\right).$$

From Bickel (1982), Theorem 1 can be proved by showing

$$n^{-1/2}\sum_{i=1}^{n}\left(a_{\hat{k}_n'}(F_n)B_{\hat{k}_n'}(\bar{e}_i)-\phi(\bar{e}_i)\right)=o_P(1) \tag{16}$$

for Model I,

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}\Big\{&\left(X_i^t-\bar{X}^t\right)\widehat{\mathrm{Var}}(X)^{-1}\hat{I}_{\hat{k}_n'}^{-1}(\phi)a_{\hat{k}_n'}(F_n)B_{\hat{k}_n'}(\bar{e}_i)\\ &-\left(X_i^t-E(X)^t\right)\mathrm{Var}^{-1}(X)I^{-1}(\phi)\phi(\bar{e}_i)\Big\}=o_P(1) \end{aligned} \tag{17}$$

for Model II and

$$\hat{I}_{\hat{k}_n'}(\phi)-I(\phi)=o_P(1) \tag{18}$$

for both models.

From assumptions (A-3) and (A-4), we can see that f is absolutely continuous. This and (A-1) give that $\{f_\theta,\ \theta\in R\}$ is a regular parametrized family. Thus, for any such f, $\{P_{\theta_n,F}\}$ and $\{P_{\theta,F}\}$ are contiguous for any deterministic sequence $\{\theta_n\}$ satisfying $\sqrt{n}(\theta_n-\theta)=O(1)$. See Hájek and Šidák (1967) or Bickel (1982) for definitions of regularity and contiguity.

By applying contiguity back and forth, it can be seen that the discretized $\bar{\theta}_n$ in $\bar{e}_i$ (see Section 2) can be replaced by the real parameter θ in proving (16)–(18). See Bickel (1982) for detailed discussion. That is to say, we can replace the residuals $\{\bar{e}_i\}$ by the real errors $\{e_i\}$ in (16)–(18). Notice that $\hat{k}_n'$ in these formulas also becomes a function of $\{e_i\}$ after these replacements.

To prove (16)–(18), we need asymptotic bounds on $\hat{k}_n'$, with $\{\bar{e}_i\}$ replaced by $\{e_i\}$. Since $\hat{k}_n'$ depends on $\hat{k}_n$ (see Section 4), the real issue is to bound $\hat{k}_n$. We have the following theorem.

THEOREM 2 (Bound theorem). *For $\hat{k}_n$ defined in Section 6, suppose that the residuals $\{\bar{e}_i\}$, on which $\hat{k}_n$ depends, are replaced by the real errors $\{e_i\}$. Then, under conditions (A-1)–(A-12) of Section 7, there exist constants $0<\beta<1/32$ and $c_l,\ c_r>0$ such that*

$$P\left(c_l n^{\beta}\le\hat{k}_n\le c_r n^{1/7+\varepsilon}\right)\to 1\quad \text{as } n\to\infty,$$

for all $\varepsilon>0$.

REMARK. A referee has pointed out the following strengthening of Theorem 1. Let θ_n and (α_n,β_n) be deterministic sequences such that

$$\left|\sqrt{n}(\theta_n-\theta)-h\right|\to 0 \quad\text{and}\quad \left|\sqrt{n}\left[(\alpha_n,\beta_n)^t-(\alpha,\beta)^t\right]-r\right|\to 0,$$

where $h\in R$ and $r\in R^{p+1}$, and let f_n be a sequence of densities such that $\|\sqrt{n}(f_n^{1/2}-f^{1/2})-g\|_2\to 0$ for some $g\in L_2$ (it is necessary that $g\perp f^{1/2}$). Let L_n represent, respectively, $P_{(\theta_n,f_n)}$ and $P_{(\alpha_n,\beta_n,f_n)}$. Then

$$\sqrt{n}\left(\hat{\theta}_n-\theta_n\right)\to_{L_n}N(0,I^{-1}(\phi)),$$

$$\sqrt{n}\left(\hat{\beta}_n-\beta_n\right)\to_{L_n}N\left(0,(\mathrm{Var}(X)I(\phi))^{-1}\right).$$

From Begun, Hall, Huang and Wellner (1983), $P_{(\theta_n, f_n)}$ (or $P_{(\alpha_n, \beta_n, f_n)}$) and $P_{(\theta, f)}$ (or $P_{(\alpha, \beta, f)}$) are contiguous. Hence we still only have to prove (16)–(18).

9. Simulation studies.

A variety of distributions were chosen as the underlying distribution of the errors $\{e_i\}$. These were normal(0, 1), Cauchy(0, 1), beta(2, 2), t with three degrees of freedom, bimodal mixture of normals $0.5N(3, 1) + 0.5N(-3, 1)$, contaminated normal $0.9N(0, 1/9) + 0.1N(0, 9)$ and lognormal(0, 1). These distributions cover a fairly large range of distributions. All the distributions are standardized to have mean 0 and variance 1, except for the Cauchy distribution. The sample sizes were 40 and 100 for the location model, and 50 and 100 for the regression model. These choices of sample size allow us to compare our results with other existing methods. Although we use the discretized initial estimates in our asymptotic theory, we will only use the original initial estimates in simulation studies, as is the common practice.

9.1. *Distributions of $\hat{k}_n$ and $\hat{k}_{\mathrm{cv},n}$.*

The empirical selection rules $\hat{k}_n$ and $\hat{k}_{\mathrm{cv},n}$ are key issues in this research. It is of interest to look at their distributions graphically. In the process of computing the location estimate, we took the histograms of $\hat{k}_n$ and $\hat{k}_{\mathrm{cv},n}$ in 5000 replications, n to be 40 and 100, respectively.

Graphs of asymptotics of $\hat{k}_n$. Theorem 2 claims that $\hat{k}_n \to \infty$ at a slow rate. Figure 1 presents interesting graphs to show this. The light shading represents the distributions of $\hat{k}_{40}$, and the dark shading that of $\hat{k}_{100}$. We can see that in most cases, except for the normal, $\hat{k}_n$ increases slightly as n increases from 40 to 100. The maximum number of $\hat{k}_{100}$ is 4 or 5, instead of 3 for $\hat{k}_{40}$. Especially, look at Figure 1f, where the error distribution is the Cauchy distribution, and observe that the range of $\hat{k}_n$ moves from (0, 3) up to (1, 5).

The peculiar behavior of $\hat{k}_n$ in the normal case, in which $\hat{k}_n$ tends to go to 0 as n increases from 40 to 100, does not contradict our theorem. The score function ϕ of the normal distribution is simply a straight line. Thus, the condition $\phi_b \notin \mathscr{L}_{k,\xi}$, which is required by the theorem, fails to hold here. Now $\mathrm{Bias}(k, \phi) = 0$, for all $k \geq 0$. It is apparent that the optimal selection of k should be $\hat{k}_n = 0$. Figure 1a reveals that $\hat{k}_n$ tends to adapt to the different situations in approaching an optimal selection.

$\hat{k}'_{\mathrm{cv}}$ and $\hat{k}_{\mathrm{cv}}$ and their related distributions. We have made similar plots for $\hat{k}'_{\mathrm{cv}}$ and $\hat{k}_{\mathrm{cv}}$. These show that $\hat{k}'_{\mathrm{cv}}$ and $\hat{k}_{\mathrm{cv}}$ have the same tendency of moving up as $\hat{k}_n$ does. The upper bound $\hat{k}_{\mathrm{cv}}$ moves faster than the lower bound $\hat{k}'_{\mathrm{cv}}$. To understand better what is going on about the stationary correction rule and the cross-validation selection, we present the distributions of $\hat{k}_n - \hat{k}'_{\mathrm{cv}}$ (see Figure 2). There is considerably nonzero portion among these distribution. This indicates that $\hat{k}_n$ is substantially different from $\hat{k}'_{\mathrm{cv}}$. However, $\hat{k}_n$ is not far away from $\hat{k}'_{\mathrm{cv}}$. In most of the cases, the stationary correction $\hat{k}_n$ gives no more than a one step correction for $\hat{k}'_{\mathrm{cv}}$. This agrees with the Remark 1 in Section 6 that the computational procedure takes little time.

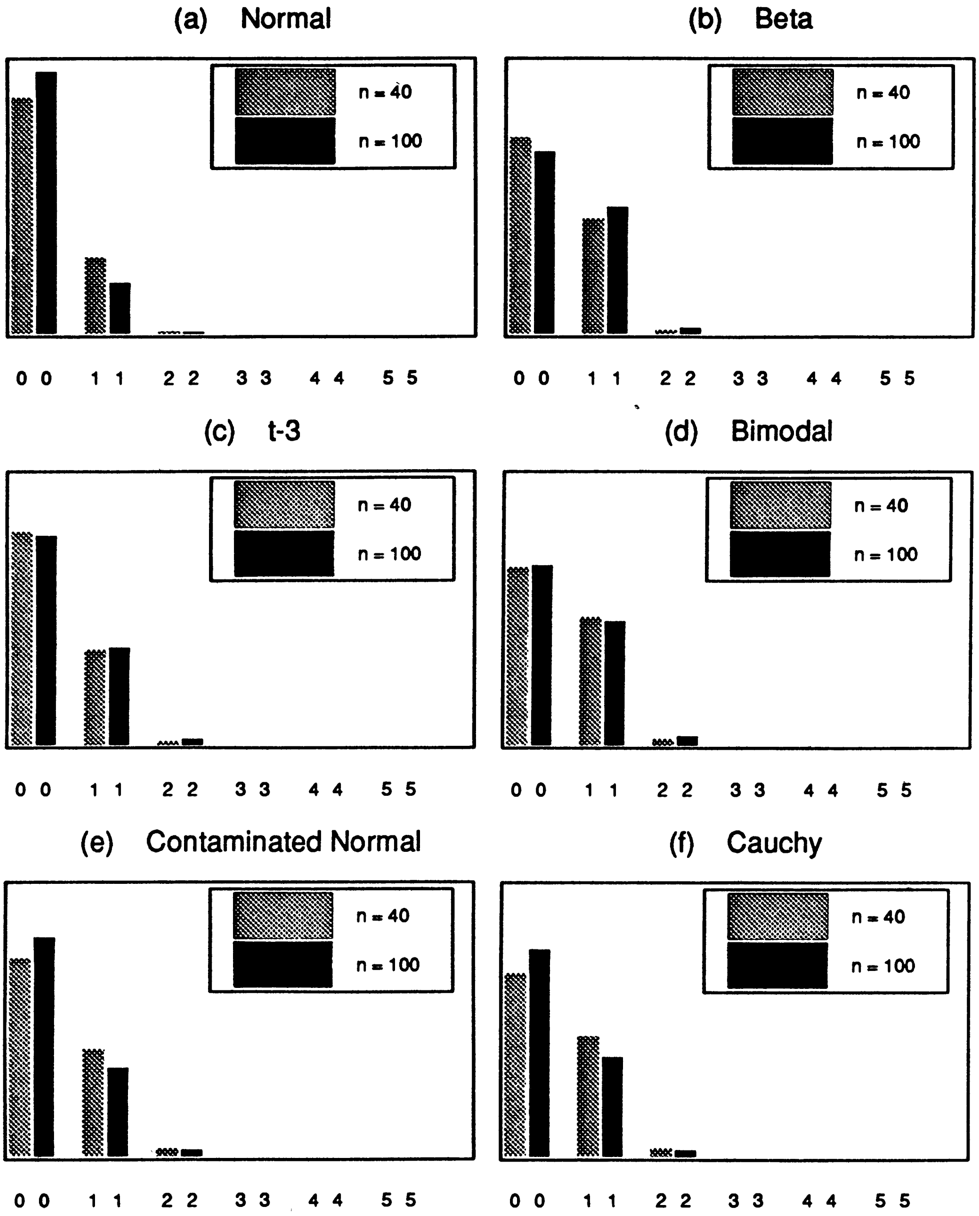

FIG. 1. *(a)–(f) compare histograms of* $\hat{k}_{40}$ *and* $\hat{k}_{100}$ *under the various error distributions. The light shading represents the histogram of* $\hat{k}_{40}$*, and the dark shading that of* $\hat{k}_{100}$*. It is clear that in most cases, except for the normal,* $\hat{k}_n$ *increases slightly as n increases from* 40 *to* 100. *We used* 5000 *replications.*

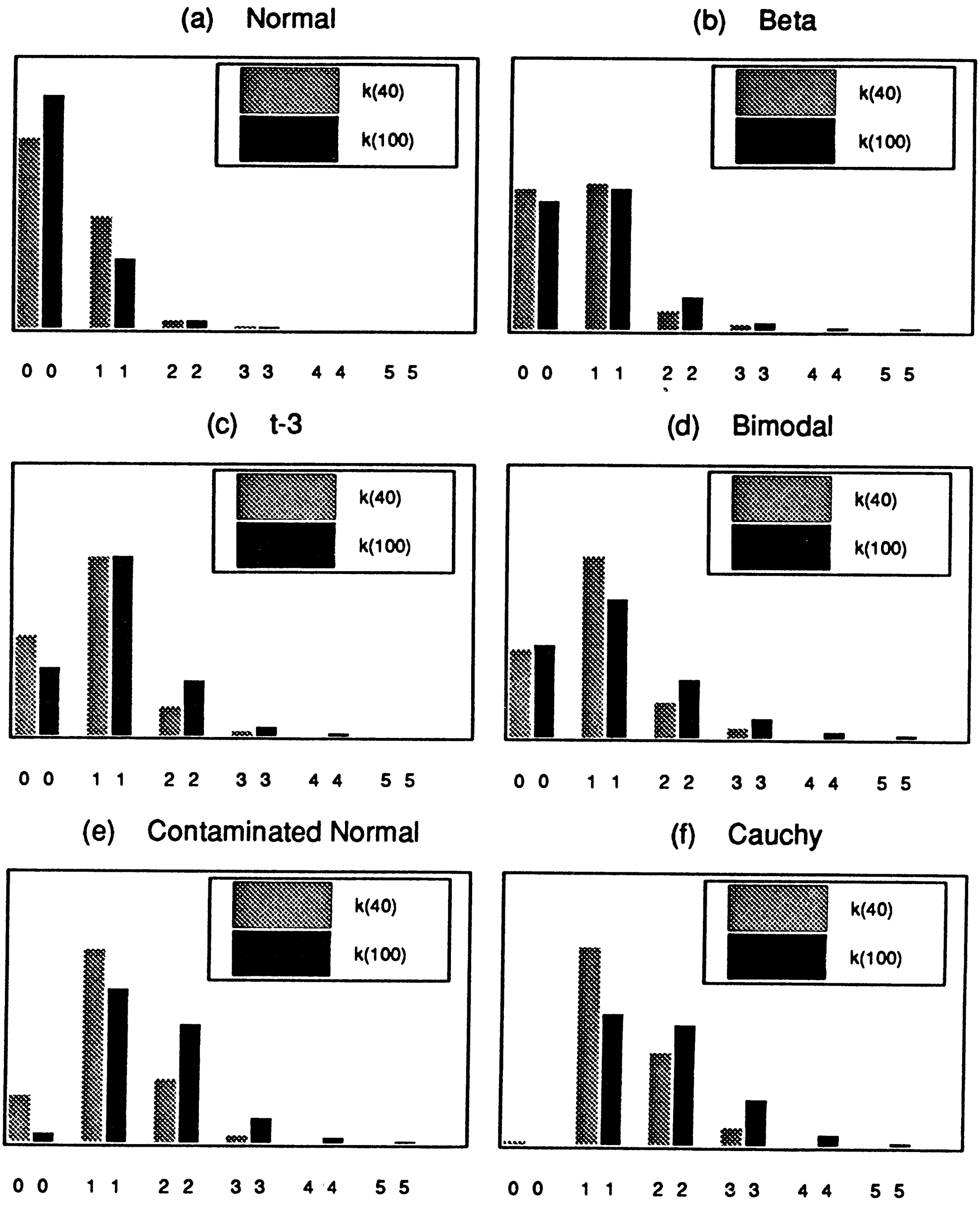

FIG. 2. *(a)–(f) present histograms of $\hat{k}_n - \hat{k}'_{\mathrm{cv},n}$ under the various error distributions. The light shading represents the histogram of $\hat{k}_{40} - \hat{k}'_{\mathrm{cv},40}$, and the dark shading that of $\hat{k}_{100} - \hat{k}'_{\mathrm{cv},100}$. The considerable nonzero portion of these histograms indicates that $\hat{k}_n$ is different from $\hat{k}'_{\mathrm{cv}}$. However, the stationary correction $\hat{k}_n$ gives no more than a one-step correction to the cross-validation selection $\hat{k}'_{\mathrm{cv},n}$ in most of the cases.*

TABLE 1
Location case (Model I); sample median is initial estimate

Distribution	Sample size	Mean	Median	10% trim	Hampel	Our estimate	SE of RMSE
Normal	$n = 40$	0.154	0.188	0.157	0.158	0.161	0.0016
	$n = 100$	0.100	0.121	0.102	0.101	0.101	0.0010
Cauchy	$n = 40$	∞	0.254	0.391	0.255	0.250	0.0028
	$n = 100$	∞	0.159	0.228	0.154	0.150	0.0015
t with three	$n = 40$	0.156	0.119	0.117	0.114	0.118	0.0012
degrees of freedom	$n = 100$	0.100	0.078	0.074	0.073	0.074	0.0007
Beta(2, 2)	$n = 40$	0.162	0.233	0.180	0.171	0.160	0.0017
	$n = 100$	0.101	0.147	0.113	0.104	0.096	0.0010
Bimodal	$n = 40$	0.158	0.552	0.187	0.320	0.306	0.0051
	$n = 100$	0.100	0.478	0.119	0.166	0.101	0.0032
Contaminated	$n = 40$	0.158	0.069	0.065	0.059	0.064	0.0006
normal	$n = 100$	0.099	0.045	0.040	0.037	0.040	0.0004

9.2. *Location case.* There exist many location estimates. The well-known Princeton project [Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1971)] investigated more than 68 different types. We have picked the commonly used sample mean, sample median and 10% trimmed mean, as well as Hampel's three-part redescending M-estimator, which is one of the best estimates recommended in the Princeton project.

The estimate's performance is measured by the Monte Carlo standard error

$$\text{RMSE} = \sqrt{\tfrac{1}{5000} \sum_{i=1}^{5000} \left(\hat{\theta}_i - \theta\right)^2},$$

where θ is the true location parameter and $\hat{\theta}_i$ is the estimate for the ith replication. Table 1 shows the results. We take the sample median as the initial estimate that is a $\sqrt{n}$-consistent estimate under the error distributions considered, and then take the first estimate as new initial estimate to get a two-step estimate as $\hat{\theta}_n$.

Our estimate shows consistent improvement over the initial estimate (sample median). Its performance compared well with Hampel's estimate. Overall, $\hat{\theta}_n$ stays within the range of good estimates for all cases. It is also worth noting that $\hat{\theta}_n$ improves the sample median in both the normal and the Cauchy cases, a rarity in the Princeton project.

We also carried out simulations with $n = 20$ and $n = 200$; $\hat{\theta}_n$ performed well consistently. For $n = 200$, with the bimodal error distribution, the RMSE of $\hat{\theta}_n$ (two-step) was 0.034, which significantly improved on the initial estimate, sample median (which had RMSE = 0.427). The RMSE of Hampel's estimate was 0.089.

REMARK. There is Monte Carlo variability in the simulation study. To assess the magnitude of such variability, we calculate the SE of the RMSE by

$$\text{SE} = \left(\sqrt{\text{RMSE}^2 + \hat{\sigma}/\sqrt{n}} - \sqrt{\text{RMSE}^2 - \hat{\sigma}/\sqrt{n}}\right)\Big/2,$$

where $\hat{\sigma}^2 = \sum_{i=1}^{5000}((\hat{\theta}_i - \theta)^2 - \mathrm{RMSE}^2)^2/5000$. The last column of Table 1 gives such SE's for the RMSE of $\hat{\theta}_n$. The SE's of the RMSE for the other estimates have approximately the same magnitude [see Jin (1990), page 39, for more details].

9.3. *Regression case.* The performance of the estimate $\hat{\beta}_n$ of the slope parameter in Model II is of more interest. We compare our estimate with least square estimate, least absolute deviation estimate, one-step maximum likelihood estimate based on knowing ϕ, Huber's M-estimate with c = SD of residuals, Hsieh and Manski's bootstrap adaptive estimate and Faraway's spline adaptive estimate.

Our estimate is calculated by taking a least square estimate $L2$ as an initial estimate that is a $\sqrt{n}$-consistent estimate under the error distributions considered. From a computational point of view, obtaining a least square estimate is the easiest task. Unlike the location case, computation of a least absolute deviation estimate $L1$ is time-consuming.

We only report the case of Bernoulli X_i (see Table 2). We obtained similar results when using $X_i \sim U(0,1)$. Further details can be found in Jin (1990).

From Table 2 we can see the following advantages of $\hat{\beta}_n$:

1. *Improvement over the initial estimate:* It is clear that $\hat{\beta}_n$ is a significant improvement over the initial estimate $L2$.
2. *Close to one-step mle:* Except for the bimodal case, $\hat{\beta}_n$ comes very close to matching the performance of the one-step mle. It performs much better than the one-step mle for the lognormal.
3. *Better performance than Huber's estimate:* For the normal, t with three degrees of freedom and the contaminated normal, $\hat{\beta}_n$ comes close to

TABLE 2
Regression case (Model II); $P(X = 0) = P(X = 1) = \frac{1}{2}$

Distribution	Sample size	$L2$	$L1$	Huber	Hsieh and Manski	Faraway	Our estimate	One-step mle
Normal	$n = 50$	0.287	0.357	0.30	0.31	0.29	0.294	0.287
	$n = 100$	0.200	0.251	0.21		0.20	0.203	0.200
t with three	$n = 50$	0.279	0.224	0.21	0.23	0.22	0.228	0.207
degrees of freedom	$n = 100$	0.200	0.159	0.15		0.15	0.155	0.144
Beta(2, 2)	$n = 50$	0.284	0.423	0.34	0.29	0.29	0.296	0.266
	$n = 100$	0.200	0.299	0.23		0.20	0.213	0.179
Bimodal	$n = 50$	0.285	0.871	0.39	0.16	0.14	0.217	0.141
	$n = 100$	0.204	0.775	0.28		0.08	0.155	0.083
Contaminated	$n = 50$	0.283	0.133	0.14	0.17	0.13	0.172	0.144
normal	$n = 100$	0.198	0.091	0.10		0.09	0.096	0.086
Lognormal	$n = 50$	0.288	0.173	0.16	0.17	0.19	0.102	0.239
	$n = 100$	0.201	0.119	0.11		0.13	0.059	0.178

matching the performance of Huber's estimate, while for beta(2, 2), the bimodal and the lognormal, it performs much better than Huber's.

4. *Better performance than the L1 estimate:* $\hat{\beta}_n$ consistently performs better than the least absolute deviation estimate ($L1$). There has been considerable effort toward finding a better algorithm to calculate $L1$ in the linear model; $\hat{\beta}_n$ starts with an easily computable $L2$, yet it performs better than $L1$.
5. *Comparable to other adaptive estimates:* Our estimate performs as well as those of Hsieh and Manski and Faraway for the normal, t with three degrees of freedom, beta(2, 2) and contaminated normal. Those of Hsieh and Manski and Faraway do better for the bimodal, while ours is superior for the lognormal. Taking ease of computation into account, $\hat{\beta}_n$ is certainly the better choice.

We have found several interesting features of our estimate that could substantially reduce computational requirements for multiple linear regression.

Effect of initial estimate. It is commonly accepted that one-step mle's are often severely affected by bad initial estimates. M-estimates also suffer from this problem. An important feature of $\hat{\beta}_n$ is that it is less prone to such behavior.

Table 3 compares $\hat{\beta}_n$ with one-step mle's having initial estimates $L1$ and $L2$. It is clear that one-step mle's suffer from bad initial estimates, while the performance of $\hat{\beta}_n$ is more stable. The SE's of our estimates come very close to each other, regardless of the initial estimate. Furthermore, our estimates tend to give more "correction" to a bad initial estimate, especially in bimodal and lognormal cases.

TABLE 3
Effect of the initial estimate in Model II; $P(X = 0) = P(X = 1) = \frac{1}{2}$

Distribution	Sample size	$L1$ as initial	One-step mle	Our estimate	$L2$ as initial	One-step mle	Our estimate
Normal	$n = 50$	0.357	0.292	0.294	0.287	0.287	0.294
	$n = 100$	0.251	0.201	0.204	0.200	0.200	0.203
t with three	$n = 50$	0.224	0.202	0.229	0.279	0.207	0.228
degrees of freedom	$n = 100$	0.159	0.142	0.155	0.200	0.144	0.155
Beta(2, 2)	$n = 50$	0.423	0.332	0.297	0.284	0.266	0.296
	$n = 100$	0.299	0.226	0.213	0.200	0.179	0.213
Bimodal	$n = 50$	0.871	0.697	0.249	0.285	0.141	0.218
	$n = 100$	0.775	0.590	0.168	0.204	0.083	0.155
Contaminated	$n = 50$	0.133	0.106	0.176	0.283	0.144	0.172
normal	$n = 100$	0.091	0.075	0.096	0.198	0.086	0.096
Lognormal	$n = 50$	0.173	0.149	0.085	0.288	0.239	0.102
	$n = 100$	0.119	0.106	0.054	0.201	0.178	0.059

TABLE 4
Iteration in regression case (Model II); $n = 50$, $P(X = 0) = P(X = 1) = \frac{1}{2}$

		Iterations of our estimate				Iterations of our estimate		
Distribution	**Initial $L2$**	**1st**	**2nd**	**6th**	**Initial $L1$**	**1st**	**2nd**	**6th**
Normal	0.287	0.294	0.295	0.296	0.357	0.294	0.295	0.296
t with three degrees of freedom	0.279	0.228	0.228	0.228	0.224	0.229	0.227	0.228
Beta(2.2)	0.284	0.296	0.297	0.298	0.423	0.297	0.298	0.298
Bimodal	0.285	0.218	0.210	0.206	0.871	0.249	0.216	0.206
Contaminated normal	0.283	0.172	0.173	0.173	0.133	0.176	0.171	0.172
Lognormal	0.288	0.102	0.082	0.079	0.173	0.085	0.079	0.079

One-step versus iteration. Bickel (1975) proposed a one-step M-estimate as an approximation to an M-estimate. Huber [(1981), Section 6.7] advocated an iterative algorithm to compute M-estimates more precisely. It turns out that the one-step M-estimate has essentially the same behavior as the fully iterated one.

Table 4 depicts the results with $L2$ and $L1$ as the initial estimates. We used six iterations. We can see that the fully iterative procedure is not necessary for computing $\hat{\beta}_n$. One-step iteration is usually sufficient; two-step iteration gives perfect results.

10. Asymptotics of $\hat{k}_{\mathrm{cv}}$.

To prove Theorem 2, notice that $\hat{k}'_{\mathrm{cv}} \le \hat{k}_n \le \hat{k}^2_{\mathrm{cv}}$, so we only have to give upper and lower bounds on $\hat{k}_{\mathrm{cv}}$.

Recall from Section 6 that $\hat{k}_{\mathrm{cv}}$ is defined as the first local minimizer of $L(k, F_{n_1}, F_{n_2})$. Let $CV(k, F_n) = L(k, F_{n_1}, F_{n_2}) + I_b(\phi)$. Then $\hat{k}_{\mathrm{cv}}$ is also the first local minimizer of $CV(k, F_n)$. As a first step, decompose $CV(k, F_n)$,

$$
\begin{aligned}
CV(k, F_n) &= L(k, F_{n_1}, F_{n_2}) - L(k, F_{n_1}, F) + \int_{b_l}^{b_r} \left(a_k^t(F_{n_1}) - \phi\right)^2 f\,dx \\
&= D_k^t(F_{n_1}) A_k^{-1}(F_{n_1}) \left(A_k(F_{n_2}) - A_k(F)\right) A_k^{-1}(F_{n_1}) D_k(F_{n_1}) \\
&\quad + 2 D_k^t(F_{n_1}) A_k^{-1}(F_{n_1}) \left(D_k(F_{n_2}) - D_k(F)\right) \\
&\quad + \left(D_k^t(F_{n_1}) - D_k^t(F)\right) A_k^{-1}(F) \left(D_k(F_{n_1}) - D_k(F)\right) \\
&\quad + 2\left(D_k^t(F_{n_1}) - D_k^t(F)\right)\left(A_k^{-1}(F_{n_1}) - A_k^{-1}(F)\right) D_k(F_{n_1}) \\
&\quad + D_k^t(F_{n_1}) \left(A_k^{-1}(F_{n_1}) - A_k^{-1}(F)\right) A_k(F) \\
&\qquad \times \left(A_k^{-1}(F_{n_1}) - A_k^{-1}(F)\right) D_k(F_{n_1}) + \mathrm{Bias}(k, \phi) \\
&\triangleq CV_1(k, n_1, n_2) + 2CV_2(k, n_1, n_2) + V(k, n_1) \\
&\quad + 2CV_3(k, n_1) + CV_4(k, n_1) + \mathrm{Bias}(k, \phi),
\end{aligned}
\tag{19}
$$

where $\mathrm{Bias}(k, \phi)$ is given in (7). We will repeatedly use this decomposition.

THEOREM 3. *Under condition* (A-6),

$$\hat{k}_{\text{cv}} \to +\infty \quad a.s.$$

as $n \to \infty$.

PROOF. For any fixed k, it is easy to check that $CV(k, F_n) \to \text{Bias}(k, \phi)$ a.s. Denoting $EAS(k) = \{\omega\colon \lim_{n\to\infty} CV(k, F_n) = \text{Bias}(k, \phi)\}$, we have

$$P\{EAS(k)\} = 1.$$

Suppose the theorem is not true. Then there must be a k_0 such that $P\{\omega\colon \liminf_{n\to\infty} \hat{k}_{\text{cv},n} < k_0\} > 0$. Letting $EAS = \bigcap_{k \le k_0} EAS(k)$, we have

$$P\{EAS\} = 1.$$

Denote $Ek_0 = \{\omega\colon \liminf_{n\to\infty} \hat{k}_{cv,n} < k_0\}$. Then $P\{Ek_0 \cap EAS\} > 0$. For any $\omega \in Ek_0 \cap EAS$, there must exist a sequence $\{n_i\}$ and an integer $k_1 \le k_0$ such that $\hat{k}_{n_i}(\omega) = k_1$. Also, since $\hat{k}_{n_i}(\omega)$ is the first local minimizer of $CV(k, F_n)$,

$$CV(k_1, F_{n_i}) < CV(k_1 + 1, F_{n_i}). \tag{20}$$

However, since $\omega \in EAS$, (20) implies that $\text{Bias}(k_1, \phi) \le \text{Bias}(k_1 + 1, \phi)$. This is contrary to condition (A-6). □

The rate of convergence of $\text{Bias}(k, \phi)$ is important for later proofs. We have the following lemma.

LEMMA 1. *Under conditions* (A-3) *and* (A-5) *in Section* 7, $\text{Bias}(k, \phi)$ *converges to* 0 *at rate* $O(1/k^4)$.

PROOF. See Jin (1990) for the detailed proof. □

The next three lemmas give upper bounds on $\|A_k^{-1}(F)\|_1$ and $\|A_k^{-1}(F_n)\|_1$. Let $f_{\max} = \max_{b_l \le x \le b_r} f(x)$ and $f_{\min} = \min_{b_l \le x \le b_r} f(x)$.

LEMMA 2. *Under condition* (A-5), *the following hold*:

(i) $\|A_k^{-1}(F)\|_1 \le ck(f_{\max})^{3/2}/((b_r - b_l)(f_{\min})^{5/2})$;

(ii) $\sum_{j=1}^{k} \alpha_{ij}^2 \le ck^2 f_{\max}^2/((b_r - b_l) f_{\min}^4)$ *holds for all* $1 \le i \le k$;

(iii) *if* $\lambda_1 \ge \cdots \ge \lambda_k$ *are the eigenvalues of* $A_k(F)$, *then* $c_1 f_{\min}(b_r - b_l)/3k \le \lambda_k$, $\lambda_1 \le c_2 f_{\max}(b_r - b_l)/k$.

PROOF. See Jin [(1990), Corollaries 1 and 2 and Lemma 4]. The difficult proof of (i) relies on the elegant work of de Boor (1976). □

For the bound on $\|A_k^{-1}(F_n)\|_1$, let $S_{n,k} = \{\|A_k(F_n) - A_k(F)\|_\infty \le c_1 f_{\min}(b_r - b_l)/18k\}$. We have similarly the following lemma.

LEMMA 3. *If* $k = O(n^{1/2-\varepsilon})$ *and condition* (A-5) *holds, then*

$$\|A_k^{-1}(F_n)\|_1 \le ck(f_{\max})^{3/2}\big/\big((b_r - b_l)(f_{\min})^{5/2}\big)$$

on $S_{n,k}$, *and* $P\{S_{n,k}^c\} \to \infty$ *as* $n \to \infty$.

The matrix norms define din Section 3 satisfy various inequalities (Lemma 4).

LEMMA 4. *For any matrices* $A_{k\times k}$, $B_{k\times k}$ *and* $C_{k\times k}$:

(i) $\|A\| \le k\|A\|_1$,
(ii) $\|ABC\| \le \|A\|\,\|B\|_\infty \|C\|$,
(iii) $\|ABC\| \le \|A\|\,\|B\|\,\|C\|$,
(iv) $\|ABC\|_1 \le \|A\|_1 \|B\|_1 \|C\|_1$.

PROOF. Follows directly from the definitions. □

Now we give an upper bound for $\hat{k}_{\mathrm{cv}}$.

THEOREM 4 (Upper bound). *Under conditions* (A-1)–(A-5) *in Section* 7, *there exists a constant* $c > 0$ *such that*

$$P\{\hat{k}_{\mathrm{cv}} \le c n_1^{1/7+\varepsilon}\} \to 1, \quad \text{as } n \to \infty,$$

for any $\varepsilon > 0$.

The key to proving Theorem 4 lies in showing that the variance $V(k, n_1)$ in (19) is the dominant term.

In the sequel, let c denote a generic positive constant which can change from place to place, and assume that the conditions of Section 7 are in force. Let

$$D_k^*(x) = D_k(x)/k,$$

$$SA_{n,k,\gamma} = \{\| A_k(F_n) - A_k(F)\|_\infty \le n^\gamma/n^{1/2}\},$$

$$SD^*_{n,k,\gamma} = \{\| D_k^*(F_n) - D_k^*(F)\|_\infty \le n^\gamma/n^{1/2}\},$$

$$SD_{n,k} = \{\| D_k(F_n) - D_n(F)\|_\infty \le 1/n^{1/4}\}.$$

LEMMA 5. (i) $\max_i \|D^*_{k_i}(x)\| \le c$;
(ii) $\|D_k(F)\|_\infty \le c/k$;
(iii) $\|D_k(F_n)\|_\infty \le c/k$ *holds on* $SD_{n,k}$ *for* $k = O(n^{1/4})$.

PROOF. Follows directly from the definitions. □

LEMMA 6. *If* $k = O(n_1^{1/4})$ *and* $0 < \gamma < 1/2$, *then the following hold*:

(i) *on* $S_{n_1,k} \cap SD_{n_1,k} \cap SA_{n_2,k,\gamma}$, $|CV_1(k, n_1, n_2)| \le c n_2^\gamma k/n_2^{1/2}$;
(ii) *on* $S_{n_1,k} \cap SD_{n_1,k} \cap SD^*_{n_2,k,\gamma}$, $|CV_2(k, n_1, n_2)| \le c n_2^\gamma k^2/n_2^{1/2}$;

(iii) *on* $S_{n_1,k} \cap SD_{n_1,k} \cap SA_{n_1,k,\gamma}$, $|CV_4(k,n_1)| \le cn_1^{2\gamma}/n_1$;
(iv) *on* $SD_{n_1,k} \cap S_{n_1,k}$, $|CV_3(k,n_1)| \le cV(k,n_1)/k^{1/2} + c'n_1^{2\gamma}k^{2.5}/n_1$.

PROOF. (i) By Lemmas 2–5,

$$\begin{aligned}|CV_1(k,n_1,n_2)| &\le k\left\| D_k(F_{n_1})\right\|_\infty^2 \left\| A_k^{-1}(F_{n_1})\right\|_1^2 3\left\| A_k(F_{n_2}) - A_k(F)\right\|_\infty \\ &\le 3k\left(\frac{c}{k}\right)^2 \left(\frac{ck(f_{\max})^{3/2}}{(b_r - b_l)(f_{\min})^{5/2}}\right)^2 n_2^{-1/2+\gamma} \le \frac{cn_2^\gamma k}{n_2^{1/2}}.\end{aligned}$$

The proofs of (ii) and (iii) are similar to (i).

$$\begin{aligned}\text{(iv)}\quad |CV_3(k,n_1)| \le{}& \left(D_k^t(F_{n_1}) - D_k^t(F)\right)A_k^{-1/2}(F)\left(D_k(F_{n_1}) - D_k(F)\right) \\ &+ D_k^t(F_{n_1})\left(A_k^{-1}(F_{n_1}) - A_k^{-1}(F)\right)A_k^{1/2}(F) \\ &\times\left(A_k^{-1}(F_{n_1}) - A_k^{-1}(F)\right)D_k(F_{n_1}) \\ \triangleq{}& CV_{3,1} + CV_{3,2}.\end{aligned}$$

Let λ_1 be the largest eigenvalue of $A_k(F)$, by Lemma 2(iii), $|CV_{3,1}| \le \lambda_1^{1/2}V(k,n_1) \le cV(k,n_1)/k^{1/2}$. As in (i), $|CV_{3,2}| \le c'n_1^{2\gamma}k^{2.5}/n_1$. □

LEMMA 7. *There are constants* $0 < c^1 < c^2$ *such that*

$$c^1k^3/n_1 - O(1/n_1) \le E(V(k,n_1)) \le c^2k^3/n_1.$$

PROOF. See Jin [(1990), Lemma 11] for the detailed proof. □

To show $V(k,n_1) = O_P(k^3/n_1)$, we need the variance of $V(k,n_1)$.

LEMMA 8. *If* $k = O(n_1^{1/4})$, *then*

$$\sigma^2(k,n_1) \triangleq E\left(V(k,n_1) - E(V(k,n_1))\right)^2 = O\left(k^5/n_1^2\right).$$

PROOF. The proof is a long and tedious computation. Interested readers can find it in Jin [(1990), pages 60–64]. □

Lemmas 7 and 8 give the following lemma.

LEMMA 9. *If* $k = O(n_1^{1/4})$, *then for any* $\gamma > 0$

$$n_1^\gamma \frac{n_1}{k^3} V(k,n_1) \to_P \infty \quad n \to \infty.$$

Now we prove Theorem 4.

PROOF OF THEOREM 4. Since $\hat{k}_{\mathrm{cv}}$ is the first local minimizer of $CV(k, F_{n_1}, F_{n_2})$,

$$\begin{aligned}
&\left\{\hat{k}_{\mathrm{cv}} > cn_1^{1/7+\varepsilon}\right\}\\
&\quad\subset \left\{CV\left(cn_1^{1/7}, F_{n_1}, F_{n_2}\right) \geq CV\left(cn_1^{1/7+\varepsilon}, F_{n_1}, F_{n_2}\right)\right\}\\
&\quad= \left\{CV_1\left(cn_1^{1/7}, n_1, n_2\right) + 2CV_2\left(cn_1^{1/7}, n_1, n_2\right) + V\left(cn_1^{1/7}, n_1\right)\right.\\
&\qquad\qquad + 2CV_3\left(cn_1^{1/7}, n_1\right) + CV_4\left(cn_1^{1/7}, n_1\right) + \mathrm{Bias}\left(cn_1^{1/7}, n_1\right)\\
&\qquad \geq CV_1\left(cn_1^{1/7+\varepsilon}, n_1, n_2\right) + 2CV_2\left(cn_1^{1/7+\varepsilon}, n_1, n_2\right) + V\left(cn_1^{1/7+\varepsilon}, n_1\right)\\
&\qquad\qquad \left.+ 2CV_3\left(cn_1^{1/7+\varepsilon}, n_1\right) + CV_4\left(cn_1^{1/7+\varepsilon}, n_1\right) + \mathrm{Bias}\left(cn_1^{1/7+\varepsilon}, n_1\right)\right\}.
\end{aligned} \tag{21}$$

Take $\gamma + 2\varepsilon < 1/7$. Recall that $n_1 = O(\sqrt{n_2})$. By Hoeffding's inequality, for $k_n = cn_1^{1/7}$ or $cn_1^{1/7+\varepsilon}$,

$$\begin{aligned}
P\left\{S^c_{n_1, k_n}\right\} &= P\left\{\left\| A_{k_n}(F_{n_1}) - A_{k_n}(F)\right\|_\infty > c_1 f_{\min}(b_r - b_l)/18k_n\right\}\\
&\leq 3k_n \exp\left\{-cn_1/k_n^2\right\} \to 0 \quad \text{as } n \to \infty.
\end{aligned} \tag{22}$$

Similarly, $P\{SD^c_{n_1, k_n}\}$, $P\{SA^c_{n_1, k_n, \gamma}\}$, $P\{SA^c_{n_2, k_n, \gamma}\}$ and $P\{SD^{*c}_{n_2, k_n, \gamma}\}$ tend to zero as $n \to \infty$.

Hence, by Lemma 6 with $k_n = cn_1^{1/7}$ or $cn_1^{1/7+\varepsilon}$,

$$CV_i(k_n, n_1, n_2) = o_P\left(n_1^{-4/7+\gamma}\right), \qquad i = 1, 2 \text{ and } 4. \tag{23}$$

By Lemma 6, on $SD_{n_1, k_n} \cap S_{n_1, k_n}$ with $k_n = cn_1^{1/7}$,

$$n_1^{4/7-\gamma} CV_3\left(cn_1^{1/7}, n_1\right) \leq cn_1^{4/7} V\left(cn_1^{1/7}, n_1\right)/n_1^{1/14+\gamma} + cn_1^{\gamma+6.5/7}/n_1,$$

Take $\gamma + 2.5\varepsilon < 1/14$. By Lemma 7

$$E\left(n_1^{4/7} V\left(cn_1^{1/7}, n_1\right)\right)/n_1^{1/14+\gamma} \leq cn_1^{4/7}\left(cn_1^{1/7}\right)^3/n_1^{1+1/14+\gamma} \to 0$$

and similarly for $k_n = cn_1^{1/7+\varepsilon}$. Hence, with $k_n = cn_1^{1/7}$ or $cn_1^{1/7+\varepsilon}$,

$$CV_3(k_n, n_1) = o_P\left(n_1^{-4/7+\gamma}\right). \tag{24}$$

Lemma 1 gives, for $k_n = cn_1^{1/7}$ or $cn_1^{1/7+\varepsilon}$,

$$\mathrm{Bias}(k_n, \phi) = o_P\left(n_1^{-4/7+\gamma}\right). \tag{25}$$

By Lemma 7,

$$E\left(n_1^{4/7-\gamma} V\left(cn_1^{1/7}, n_1\right)\right) \leq cn_1^{4/7-\gamma}\left(cn_1^{1/7}\right)^3/n_1 = c/n_1^\gamma \to 0.$$

Hence,

$$V\left(cn_1^{1/7}, n_1\right) = o_P\left(n_1^{-4/7+\gamma}\right). \tag{26}$$

Finally, let $\gamma = \varepsilon$ and denote $k_n = n_1^{1/7+\varepsilon}$. By Lemma 9,

$$n_1^{4/7-\gamma} V\left(cn_1^{1/7+\varepsilon}, n_1\right) = cn_1^{2\varepsilon}\frac{n_1}{k_n^3} V(k_n, n_1) \to \infty. \tag{27}$$

Applying (23)–(27) to (21),

$$P\{\hat{k}_{\mathrm{cv}} > cn_1^{1/7+\varepsilon}\} \to 0. \qquad \square$$

As for a lower bound of $\hat{k}_{\mathrm{cv}}$, we have the following theorem.

THEOREM 5 (Lower bound). *Under conditions* (A-1)–(A-7) *of Section* 7, *there exist constants* $0 < \beta < 1/16$ *and* $c > 0$ *such that*

$$P\{cn_1^{\beta} \le \hat{k}_{\mathrm{cv}}\} \to 1 \quad as\ n \to \infty.$$

The key to proving this theorem is to show that $\mathrm{Bias}(k, \phi)$ in the decomposition (19) becomes the dominant term when $\hat{k}_{\mathrm{cv}}$ converges to infinity at a slow rate.

First we find a uniform upper bound on $\|A_k^{-1}(F_n)\|_1$. For a given deterministic sequence $\{K_n\}$, define

$$SU_n = \Big\{\sup_{k \le K_n} \|A_k(F_n) - A_k(F)\|_\infty \le c_1 f_{\min}(b_r - b_l)/18K_n\Big\},$$

$$SAU_{n,\gamma} = \Big\{\sup_{k \le K_n} \|A_k(F_n) - A_k(F)\|_\infty \le n^{\gamma}/n^{1/2}\Big\},$$

$$SDU_{n,\gamma}^{*} = \Big\{\sup_{k \le K_n} \|D_k^{*}(F_n) - D_k^{*}(F)\|_\infty \le n^{\gamma}/n^{1/2}\Big\},$$

$$SDU_n = \Big\{\sup_{k \le K_n} \|D_k(F_n) - D_k(F)\|_\infty \le 1/n^{1/4}\Big\}.$$

We shall need the following analogues of Lemmas 3, 5 and 6.

LEMMA 10. *On* SU_n,

$$\sup_{k \le K_n} \|A_k^{-1}(F_n)\|_1 \le cK_n(f_{\max})^{3/2}\Big/\Big((b_r - b_l)(f_{\min})^{5/2}\Big).$$

LEMMA 11. (i) $\sup_{k \le K_n} \max_i \|D_{k_i}^{*}(x)\| \le c$;
(ii) $\sup_{k \le K_n} \|D_k(F)\|_\infty \le c$;
(iii) $\sup_{k \le K_n} \|D_k(F_n)\|_\infty \le c$ *holds on* SDU_n *for* $K_n = O(n^{1/4})$.

LEMMA 12. *If* $K_n = O(n_1^{1/4})$ *and* $0 < \gamma < 1/2$, *then the following hold*:

(i) *on* $SU_{n_1} \cap SDU_{n_1} \cap SAU_{n_2,\gamma}$, $\sup_{k \le K_n} |CV_1(k, n_1, n_2)| \le cn_2^{\gamma}K_n^3/n_2^{1/2}$;
(ii) *on* $SU_{n_1} \cap SDU_{n_1} \cap SDU_{n_2,\gamma}^{*}$, $\sup_{k \le K_n} |CV_2(k, n_1, n_2)| \le cn_2^{\gamma}K_n^3/n_2^{1/2}$;
(iii) *on* $SU_{n_1} \cap SDU_{n_1} \cap SAU_{n_1,\gamma}$, $\sup_{k \le K_n} |CV_4(k, n_1)| \le cn_1^{2\gamma}K_n^4/n_1$;
(iv) *on* $SU_{n_1} \cap SDU_{n_1} \cap SAU_{n_1,\gamma}$, $\sup_{k \le K_n} |CV_3(k, n_1)| \le cn_1^{\gamma}K_n^3/n_1^{3/4}$.

PROOF. See Jin [(1990), Lemmas 14–17] for the detailed proofs. □

Here the variance term $V(k, n_1)$ is no longer the dominant term. We do not need sharp bounds for $V(k, n_1)$. It is easy to obtain the following.

LEMMA 13. *On SDU_{n_1},*

$$\sup_{k \le K_n} |V(k, n_1)| \le cK_n^2/n_1^{1/2}.$$

Denote

$$\begin{aligned}\mathrm{Ran}(k, n_1, n_2) &\triangleq CV_1(k, n_1, n_2) + 2CV_2(k, n_1, n_2)\\ &\quad + V(k, n_1) + 2CV_3(k, n_1) + CV_4(k, n_1).\end{aligned}$$

Then,

$$CV(k, F_n) = \mathrm{Ran}(k, n_1, n_2) + \mathrm{Bias}(k, \phi).$$

From Lemmas 12 and 13, on the intersections of SU_{n_1}, SDU_{n_1}, $SDU^*_{n_2,\gamma}$, $SAU_{n_1,\gamma}$ and $SAU_{n_2,\gamma}$, with $K_n = O(n_1^{1/4})$ and $0 < \gamma < 1/2$, we have

$$\sup_{k \le K_n} |\mathrm{Ran}(k, n_1, n_2)| \le cn_1^{2\gamma}K_n^2/n_1^{1/2}. \tag{28}$$

PROOF OF THEOREM 5. Since $\hat{k}_{\mathrm{cv}}$ is the first local minimizer of $CV(k, n_1, n_2)$,

$$\begin{aligned}\{\hat{k}_{\mathrm{cv}} \le cn_1^{\beta}\} &\subset \{CV(\hat{k}_{\mathrm{cv}}, n_1, n_2) < CV(\hat{k}_{\mathrm{cv}} + 1, n_1, n_2)\}\\ &= \{\hat{k}_{\mathrm{cv}}^{\alpha_L} CV(\hat{k}_{\mathrm{cv}}, n_1, n_2) < \hat{k}_{\mathrm{cv}}^{\alpha_L} CV(\hat{k}_{\mathrm{cv}} + 1, n_1, n_2)\}\\ &= \{\hat{k}_{\mathrm{cv}}^{\alpha_L}(\mathrm{Bias}(\hat{k}_{\mathrm{cv}}, \phi) - \mathrm{Bias}(\hat{k}_{\mathrm{cv}} + 1, \phi))\\ &\qquad < \hat{k}_{\mathrm{cv}}^{\alpha_L}(\mathrm{Ran}(\hat{k}_{\mathrm{cv}} + 1, n_1, n_2) - \mathrm{Ran}(\hat{k}_{\mathrm{cv}}, n_1, n_2))\},\end{aligned}$$

where α_L is given in (A-7) of Section 7. Taking $K_{n_1} = cn_1^{\beta}$ and $\gamma = \beta$, we have from (28)

$$|\hat{k}_{\mathrm{cv}}^{\alpha_L} \mathrm{Ran}(\hat{k}_{\mathrm{cv}}, n_1, n_2)| \le cn_1^{(\alpha_L + 4)\beta}/n_1^{1/2}, \tag{29}$$

$$|\hat{k}_{\mathrm{cv}}^{\alpha_L} \mathrm{Ran}(\hat{k}_{\mathrm{cv}} + 1, n_1, n_2)| \le c' n_1^{(\alpha_L + 4)\beta}/n_1^{1/2}. \tag{30}$$

It is clear that there exists β, $0 < \beta < 1/(2\alpha_L + 8) < 1/16$, such that the right-hand sides of (29) and (30) converge to 0. Also, for this β, by Hoeffding's inequality, we have

$$\begin{aligned}P\{SU_{n_1}^c\} &= P\Big\{\sup_{k \le K_{n_1}} \|A_k(F_n) - A_k(F)\|_\infty > c_1 f_{\min}(b_r - b_l)/18K_{n_1}\Big\}\\ &\le cK_{n_1}^2 \exp\{-c' n_1/K_{n_1}^2\}\\ &= cn_1^{2\beta} \exp\{-c' n_1^{1-2\beta}\} \to 0 \quad \text{as } n \to \infty.\end{aligned}$$

Similarly, $P\{SDU_{n_1}^{*c}\}$, $P\{SDU_{n_2,\gamma}^{c}\}$, $P\{SAU_{n_1,\gamma}^{c}\}$ and $P\{SAU_{n_2,\gamma}^{c}\}$ tend to zero as $n \to \infty$. Hence,

$$\left|\hat{k}_{\text{cv}}^{\alpha_L} \operatorname{Ran}\left(\hat{k}_{\text{cv}}, n_1, n_2\right)\right| \to_P 0 \quad \text{as } n \to \infty,$$

$$\left|\hat{k}_{\text{cv}}^{\alpha_L} \operatorname{Ran}\left(\hat{k}_{\text{cv}} + 1, n_1, n_2\right)\right| \to_P 0 \quad \text{as } n \to \infty.$$

That is,

$$\hat{k}_{\text{cv}}^{\alpha_L}\left(\operatorname{Ran}\left(\hat{k}_{\text{cv}}, n_1, n_2\right) - \operatorname{Ran}\left(\hat{k}_{\text{cv}} + 1, n_1, n_2\right)\right) \to_P 0 \quad \text{as } n \to \infty.$$

On the other hand, by Lemma 3 and assumption (A-7),

$$\liminf_{n \to \infty} \hat{k}_{\text{cv}}\left(\operatorname{Bias}\left(\hat{k}_{\text{cv}}, \phi\right) - \operatorname{Bias}\left(\hat{k}_{\text{cv}} + 1, \phi\right)\right) > 0 \quad \text{a.s.}$$

This leads to $P\{\hat{k}_{\text{cv}} \le c n_1^{\beta}\} \to 0$ as $n \to \infty$. □

It is clear that Theorem 2 is a consequence of Theorems 4 and 5.

11. Adaptiveness of the estimates. In this section we prove Theorem 1. As mentioned in Section 8, we only have to prove (16)–(18).

Decompose (17) into

$$\begin{aligned} & n^{-1/2} \sum_{i=1}^{n} \Big\{\left(X_i^t - \bar{X}^t\right)\widehat{\operatorname{Var}}(X)^{-1} \hat{I}_{\hat{k}'_n}^{-1}(\phi) a_{\hat{k}'_n}^t(F_n) B_{\hat{k}'_n}(e_i) \\ & \qquad - \left(X_i^t - E(X)^t\right)\operatorname{Var}^{-1}(X) I^{-1}(\phi)\phi(e_i)\Big\} \\ & \quad = n^{-1/2} \sum_{i=1}^{n} \Big\{\left(X_i^t - E(X)^t\right)\operatorname{Var}^{-1}(X) I^{-1}(\phi)\left(a_{\hat{k}'_n}^t(F_n) B_{\hat{k}'_n}(e_i) - \phi(e_i)\right)\Big\} \\ & \qquad + n^{-1/2} \sum_{i=1}^{n} \left(X_i^t - E(X)^t\right) a_{\hat{k}'_n}^t(F_n) B_{\hat{k}'_n}(e_i) \\ & \qquad\qquad \times \left(\widehat{\operatorname{Var}}(X)^{-1} \hat{I}_{\hat{k}'_n}^{-1} - \operatorname{Var}^{-1}(X) I^{-1}\right) \\ & \qquad - \sqrt{n}\left(\bar{X}^t - E(X)^t\right)\widehat{\operatorname{Var}}(X)^{-1} \hat{I}_{\hat{k}'_n}^{-1}(\phi) \frac{1}{n} \sum_{i=1}^{n} a_{\hat{k}'_n}^t(F_n) B_{\hat{k}'_n}(e_i) \\ & \quad \triangleq \Delta_1 + \Delta_2 + \Delta_3. \end{aligned} \tag{31}$$

To show (17), we only have to prove

$$n^{-1/2} \sum_{i=1}^{n} \left\{\left(X_i^t - E(X)^t\right)\left(a_{\hat{k}'_n}^t(F_n) B_{\hat{k}'_n}(e_i) - \phi(e_i)\right)\right\} = o_P(1), \tag{32}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \left\{a_{\hat{k}'_n}^t(F_n) B_{\hat{k}'_n}(e_i) - \phi(e_i)\right\} = o_P(1). \tag{33}$$

To see this, notice that (33) implies

$$\frac{1}{n}\sum_{i=1}^{n} a^t_{\hat{k}'_n}(F_n)B_{\hat{k}'_n}(e_i) = o_P(1). \tag{34}$$

Also it is not hard to check that (32) and (34) imply $\Delta_i = o_P(1)$ for $i = 1, 2, 3$.

Since $\hat{k}'_n$ is the number of knots used in constructing the adaptive estimates (see Section 8), to show (16), (18), (32) and (33), we need to extend Theorem 2.

LEMMA 14. *Under* (A-5) *and* (A-11) *in Section* 7, *if there are constants* $c_1, c_2, \alpha, \beta > 0$ *such that*

$$c_1 n^{\alpha} \le k_n \le c_2 n^{\beta+\varepsilon},$$

for any $\varepsilon > 0$, *then there exists* c'_2 *such that*

$$c_1 n^{\alpha} \le k'_n \le c'_2 n^{\beta+\varepsilon},$$

for any $\varepsilon > 0$.

The proof of Lemma 14 is fairly straightforward and is omitted. From Theorem 2 and Lemma 14 we can see immediately that the following lemma holds.

LEMMA 15. *Under conditions* (A-1)–(A-12) *of Section* 7, *there exist constants* $0 < \beta < 1/32$ *and* $c_l, c_r > 0$ *such that*

$$P\big(c_l n^{\beta} \le \hat{k}'_n \le c_r n^{1/7+\varepsilon}\big) \to 1 \quad \textit{as } n \to \infty,$$

for all $\varepsilon > 0$.

Now (16), (18), (32) and (33) will follow from the following theorems.

THEOREM 6. *Under the conditions in Section* 7, *for any constants* $c_1, c_2, \beta > 0$ *and* $0 < \varepsilon < 1/42$, *the following hold*:

$$\text{(i)} \qquad \sup_{c_1 n^{\beta} \le k' \le c_2 n^{1/6-\varepsilon}} \left| n^{-1/2}\left(\sum_{i=1}^{n} \big(a^t_{k'}(F_n)B_{k'}(e_i) - \phi(e_i)\big)\right| \to_P 0$$

for Model I;

$$\text{(ii)} \qquad \sup_{c_1 n^{\beta} \le k' \le c_2 n^{1/6-\varepsilon}} \left| \frac{1}{n}\sum_{i=1}^{n} a^t_{k'}(F_n)D_{k'}(e_i) + I(\phi)\right| \to_P 0$$

for Models I and II;

(iii)

$$\sup_{c_1 n^{\beta} \le k' \le c_2 n^{1/6-\varepsilon}} \left| n^{-1/2}\sum_{i=1}^{n} \big(X_i^t - E(X)^t\big)\big(a^t_{k'}(F_n)B_{k'}(e_i) - \phi(e_i)\big)\right| \to_P 0$$

for Model II;

$$\text{(iv)} \qquad \sup_{c_1 n^{\beta} \le k' \le c_2 n^{1/6-\varepsilon}} \left| \frac{1}{n} \sum_{i=1}^{n} \left(a_{k'}^{t}(F_n) B_{k'}(e_i) - \phi(e_i) \right) \right| \to_P 0,$$

for Model II.

Before proving Theorem 6 (i)–(iv), we state some lemmas. The proofs of these lemmas are similar to those of Section 10. Let $K_n = O(n^{1/6-\varepsilon})$. Denote

$$SU_n = \left\{ \sup_{k' \le K_n} \| A_{k'}(F_n) - A_{k'}(F) \|_{\infty} \le c_1 f_{\min,n}(b_{rn} - b_{ln})/(18K_n) \right\},$$

$$SAU_{n,\gamma} = \left\{ \sup_{k' \le K_n} \| A_{k'}(F_n) - A_{k'}(F) \|_{\infty} \le n^{\gamma}/n^{1/2} \right\},$$

$$SDU_{n,\gamma}^{*} = \left\{ \sup_{k' \le K_n} \| D_{k'}^{*}(F_n) - D_{k'}^{*}(F) \|_{\infty} \le n^{\gamma}/n^{1/2} \right\},$$

$$SDU_n = \left\{ \sup_{k' \le K_n} \| D_{k'}(F_n) - D_{k'}(F) \|_{\infty} \le 1/n^{1/4} \right\},$$

$$SBU_{n,\gamma} = \left\{ \sup_{k' \le K_n} \| B_{k'}(F_n) \| \le n^{\gamma}/n^{1/2} \right\}.$$

By Hoeffding's inequality, it can be shown that $P\{SU_n^c\}$, $P\{SAU_{n,\gamma}^c\}$, $P\{SDU_{n,\gamma}\}$, $P\{(SDU_{n,\gamma}^{*})^c\}$, $P\{SDU_n^c\}$ and $P\{SBU_{n,\gamma}^c\}$ tend to zero as $n \to \infty$.

LEMMA 16.

$$\text{(i)} \qquad \sup_{k' \le K_n} \| A_{k'}^{-1}(F) \|_1 \le \frac{cK_n (f_{\max,n})^{3/2}}{(b_{rn} - b_{ln})(f_{\min,n})^{5/2}};$$

(ii) *On* SU_n,

$$\sup_{k' \le K_n} \| A_{k'}^{-1}(F_n) \|_1 \le \frac{cK_n (f_{\max,n})^{3/2}}{(b_{rn} - b_{ln})(f_{\min,n})^{5/2}}.$$

LEMMA 17. (i) $\sup_{k' \le K_n} \max_i \|D_{k'_i}^{*}(x)\| \le c$;
(ii) $\sup_{k' \le K_n} \|D_{k'}(F)\|_{\infty} \le c b_n(f')$;
(iii) $\sup_{k' \le K_n} \|D_{k'}(F_n)\|_{\infty} \le c b_n(f')$ *holds on* SDU_n *for* $K_n = O(n^{1/4})$.

Proof of Theorem 6. We only give the proof of (i). The proofs of (ii)–(iv) are similar. Interested readers can refer to Jin [(1990), Lemmas 24 and 25] for the details.

$$\begin{aligned}
&\left|n^{-1/2}\sum_{i=1}^{n}\left(a^{t}_{k'}(F_n)B_{k'}(e_i)-\phi(e_i)\right)\right| \\
&\quad=\left|n^{-1/2}\sum_{i=1}^{n}\left(-D^{t}_{k'}(F_n)A^{-1}_{k'}(F_n)B_{k'}(e_i)-\phi(e_i)\right|\right. \\
&\quad\le\left|n^{-1/2}\sum_{i=1}^{n}\left(D^{t}_{k'}(F)-D^{t}_{k'}(F_n)\right)A^{-1}_{k'}(F_n)B_{k'}(e_i)\right| \\
&\qquad+\left|n^{-1/2}\sum_{i=1}^{n}D^{t}_{k'}(F)\left(A^{-1}_{k'}(F)-A^{-1}_{k'}(F_n)\right)B_{k'}(e_i)\right| \\
&\qquad+\left|n^{-1/2}\sum_{i=1}^{n}\left(a^{t}_{k'}(F)B_{k'}(e_i)-\phi_n(e_i)\right)\right| \\
&\qquad+\left|n^{-1/2}\sum_{i=1}^{n}\left(\phi_n(e_i)-\phi(e_i)\right)\right| \\
&\quad\triangleq\Delta_1+\Delta_2+\Delta_3+\Delta_4.
\end{aligned}$$

Let $K_n = O(n^{1/6-\varepsilon})$ and $\gamma = \varepsilon$. Then, by Lemmas 16 and 17 and by (13), on the intersection of SU_n, $SDU^{*}_{n,\gamma}$, $SBU_{n,\gamma}$,

$$\begin{aligned}
\sup_{\substack{c_1 n^{\beta}\le k'\le\\ c_2 n^{1/6-\varepsilon}}}|\Delta_1| &\le \sup_{k'\le K_n}\left|n^{-1/2}\sum_{i=1}^{n}\left(D^{t}_{k'}(F)-D^{t}_{k'}(F_n)\right)A^{-1}_{k'}(F_n)B_{k'}(e_i)\right| \\
&\le \sup_{k'\le K_n} n^{1/2}k\|D^{*}_{k'}(F)-D^{*}_{k'}(F_n)\|_{\infty}k\|A^{-1}_{k'}(F_n)\|_{1}\|B_{k'}(F_n)\|_{\infty} \qquad (35)\\
&\le c\frac{f^{3/2}_{\max,n}}{(b_{rn}-b_{ln})n^{\varepsilon}f^{5/2}_{\min,n}} \to_P 0, \quad \text{by } (13).
\end{aligned}$$

On the intersection of SU_n, $SAU_{n,\gamma}$, $SBU_{n,\gamma}$,

$$\begin{aligned}
&\sup_{c_1 n^{\beta}\le k'\le c_2 n^{1/6-\varepsilon}}|\Delta| \\
&\quad\le \sup_{k'\le K_n}\left|n^{-1/2}\sum_{i=1}^{n}D^{t}_{k'}(F)\left(A^{-1}_{k'}(F)-A^{-1}_{k'}(F_n)\right)B_{k'}(e_i)\right| \\
&\quad\le n^{1/2}\sup_{k'\le K_n}\Big\{\|D_{k'}(F)\|_{\infty}k\|A^{-1}_{k'}(F_n)\|_{1}3\|A_{k'}(F_n)-A_{k'}(F)\|_{\infty} \qquad (36)\\
&\qquad\qquad\times\|A^{-1}_{k'}(F)\|_{1}\|B_{k'}(F_n)\|_{\infty}\Big\} \\
&\quad\le c\frac{f^{3}_{\max,n}b_n(f')}{(b_{rn}-b_{ln})^{2}f^{5}_{\min,n}n^{\varepsilon}} \to_P 0, \quad \text{by } (13).
\end{aligned}$$

Next, we work on Δ_3.

$$
\begin{aligned}
& E\Bigl(\sup_{c_1 n^{\beta} \le K' \le c_2 n^{1/6-\varepsilon}} |\Delta_3|^2\Bigr) \\
(37) \qquad & \le \sum_{c_1 n^{\beta} \le k' \le c_2 n^{1/6-\varepsilon}} E\biggl(n^{-1/2} \sum_{i=1}^{n} \bigl(a_{k'}^{t}(F) B_{k'}(e_i) - \phi_n(e_i)\bigr)\biggr)^2 \\
& = \sum_{c_1 n^{\beta} \le k' \le c_2 n^{1/6-\varepsilon}} \int_{b_{ln}}^{b_{rn}} \bigl(a_{k'}^{t}(F) B_{k'}(x) - \phi_n(x)\bigr)^2 f(x)\, dx.
\end{aligned}
$$

Under conditions (A-5) and (A-12),

$$
\begin{aligned}
(38) \qquad & \int_{b_{ln}}^{b_{rn}} \bigl(a_{k'}^{t}(F) B_{k'}(x) - \phi_n(x)\bigr)^2 f(x)\, dx \\
& \le \bigl(d(\phi_n, \mathscr{L}_{k',\xi})\bigr)^2 \le c \|\phi_n^{(2)}\|_{\infty}^2 / k'^4 \le c/k'^2.
\end{aligned}
$$

Combining (37) and (38) gives

$$
E\Bigl(\sup_{c_1 n^{\beta} \le k' \le c_2 n^{1/6-\varepsilon}} |\Delta_3|^2\Bigr) \le c \sum_{c_1 n^{\beta} \le k' \le c_2 n^{1/6-\varepsilon}} \frac{1}{k'^2} \to 0
$$

as $n \to 0$. Therefore, $\sup_{c_1 n^{\beta} \le k' \le c_2 n^{1/6-\varepsilon}} |\Delta_3| \to_P 0$ as $n \to \infty$.

Finally,

$$
\sup_{c_1 n^{\beta} \le k' \le c_2 n^{1/6-\varepsilon}} |\Delta_4| = |\Delta_4| = \left| n^{-1/2} \sum_{i=1}^{n} \bigl(\phi_n(e_i) - \phi(e_i)\bigr) \right|,
$$

and

$$
(39) \qquad E|\Delta_4|^2 = \int_{-\infty}^{b_{ln}} \phi^2 f dx + \int_{b_{rn}}^{+\infty} \phi^2 f dx \to 0
$$

as $b_{ln} \to -\infty$ and $b_{rn} \to \infty$. Therefore, $\sup_{c_1 n^{\beta} \le k' \le c_2 n^{1/6-\varepsilon}} |\Delta_4| = |\Delta_4| \to_P 0$. This completes the proof. □

Acknowledgments. This paper is adapted from the author's Ph.D. dissertation written at the University of California, Berkeley, under the supervision of Professor Peter Bickel, whose help is gratefully acknowledged. The author thanks an Associate Editor and the referees, whose comments helped to clarify several important points. The author also thanks Professor Ian McKeague for his help in polishing the final version of the paper.

REFERENCES

Andrews, D., Bickel, P. J., Hampel, F., Huber, P., Rogers, W. and Tukey, J. (1971). *Robust Estimates of Location: Survey and Advances*. Princeton Univ. Press.

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11** 432–452.

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434.

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1992). *Efficient and Adaptive Inference in Semiparametric Models.* Johns Hopkins Univ. Press. To appear.

COX, D. D. (1985). A penalty method for nonparametric estimation of the logarithmic derivative of a density function. *Ann. Inst. Statist. Math.* **37** 271–288.

DE BOOR, C. (1976). A bound on the L_∞-norm of L_2-approximation by splines in terms of a global mesh ratio. *Math. Comp.* **30** 756–771.

FARAWAY, J. J. (1992). Smoothing in adaptive estimation. *Ann. Statist.* **20** 414–427.

FRIEDMAN, J. H. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics* **31** 3–21.

HÁJEK, J. and ŠIDÁK, Z. (1967). *Theory of Rank Tests.* Academic, New York.

HALL, P. and MARRON, J. S. (1987). Extent to which least-square cross-validation minimizes integrated square error in nonparametric density estimation. *Probab. Theory Related Fields* **74** 567–581.

HSIEH, D. A. and MANSKI, C. F. (1987). Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. *Ann. Statist.* **15** 541–551.

HUBER, P. (1981). *Robust Statistics.* Wiley, New York.

JIN, K. (1990). Empirical smoothing parameter selection in adaptive estimation. Ph.D. dissertation. Dept. Statistics., Univ. California, Berkeley.

LE CAM, L. (1969). *Théorie Asymptotique de la Décision Statistique.* Les Presses de l'Univérsite de Montréal.

PARK, B. U. and MARRON, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85** 66–72.

RUDEMO, M. (1982). Empirical choice of histogram and kernel density estimators. *Scand. J. Statist.* **8** 65–78.

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 187–195. Univ. California Press, Berkeley.

STONE, C. J. (1975). Adaptive maximum likelihood estimation of a location parameter. *Ann. Statist.* **3** 267–284.

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

DEPARTMENT OF CLINICAL EPIDEMIOLOGY
AND FAMILY MEDICINE
M-200 SCAIFE HALL
UNIVERSITY OF PITTSBURGH SCHOOL OF MEDICINE
PITTSBURGH, PENNSLYVANIA 15261