The Annals of Statistics
1987, Vol. 15, No. 4, 1520–1535

ASYMPTOTICALLY OPTIMAL BANDWIDTH SELECTION FOR KERNEL DENSITY ESTIMATORS FROM RANDOMLY RIGHT-CENSORED SAMPLES

BY J. S. MARRON[1] AND W. J. PADGETT[2]

University of North Carolina, Chapel Hill and University of South Carolina

This paper makes two important contributions to the theory of bandwidth selection for kernel density estimators under right censorship. First, an asymptotic representation of the integrated squared error into easily understood variance and squared bias components is given. Second, it is shown that if the bandwidth is chosen by the data-based method of least-squares cross-validation, then it is asymptotically optimal in a compelling sense. A by-product of the first part is an interesting comparison of the two most popular kernel estimators.

1. Introduction. Kernel-type estimators of an unknown probability density function from right-censored data have been studied recently by several authors [e.g., Blum and Susarla (1980), Diehl and Stute (1985), Földes, Rejtö and Winter (1981), McNichols and Padgett (1986), Mielniczuk (1986) and Stute (1985)]. Padgett and McNichols (1984) gave a review of available results on kernel density estimation from censored data. The details of the forms of these estimators are in Section 2.

As in the complete sample (i.e., uncensored) case, the choice of the smoothing parameter, or bandwidth, is crucial to the effective performance of the estimator. Intuitively, if the bandwidth is too small, there is too much "variance" in the sense that features which belong only to the particular data set, and not to the underlying density, may be seen in the estimate. If the bandwidth is too large, there is too much "bias" in the sense that features of the density are smoothed away.

In the complete sample case, an elegant mathematical quantification of the preceding intuition may be found in Rosenblatt (1956), Parzen (1962), Watson and Leadbetter (1963) and Rosenblatt (1971). In particular, they show that the mean integrated squared error (MISE) has an asymptotic decomposition as a simple variance term, a simple squared bias term and some negligible terms. In Section 3, it is seen how this type of decomposition may be done in the case of randomly right-censored data. Along the way, approximations are found for the two most popular censored-data kernel estimators which give insight into exactly how they are related.

Received March 1986; revised February 1987.

[1]Research supported by National Science Foundation grant DMS-84-00602.

[2]Research supported by U.S. Air Force Office of Scientific Research grant AFOSR-84-0156 and U.S. Army Research Office grant MIPR ARO 139-85.

AMS 1980 *subject classifications*. Primary 62G05; secondary 62G20.

Key words and phrases. Nonparametric density estimation, optimal bandwidth, random censorship, smoothing parameter, cross-validation.

While this asymptotic representation of MISE provides considerable insight, it is not very useful for selecting the bandwidth because the minimizer of the two dominant terms contains quantities which are harder to estimate than f itself. As this is also true in the complete sample case, there has recently been considerable work done there on data-based bandwidth selectors. One of the most promising methods is least-squares cross-validation, introduced by Rudemo (1982) and Bowman (1984). The bandwidth selected in this way has been shown to be asymptotically optimal under various conditions by Hall (1983), Stone (1984), Burman (1985), Hall (1985) and Marron (1985). Deeper asymptotic properties are established in Hall and Marron (1987a, b).

In Section 4, it is shown that least-squares cross-validation is also effective in the case of right-censored data. In particular, asymptotic optimality, in the same sense as for the complete sample case, is established. Section 5 contains the proofs. Finally, a practical method for choosing between the two different common kernel estimators is suggested.

2. The estimators. The two best known kernel density estimators are based on estimates of distribution functions. In the censored-data case, a widely used distribution function estimator is defined as follows.

Let $X_1^0, \ldots, X_n^0$ denote the i.i.d. survival times of n items or individuals that are censored on the right by i.i.d. random variables $U_1, \ldots, U_n$ which are independent of the X_i^0's. Denote the common distribution function of the X_i^0's by F^0 and that of the U_i's by H. Let $H^* = 1 - H$. It is assumed that F^0 is absolutely continuous with density f^0 and that H is continuous.

The observed randomly right-censored data are denoted by the pairs (X_i, Δ_i), $i = 1, \ldots, n$, where

$$X_i = \min\{X_i^0, U_i\} \quad \text{and} \quad \Delta_i = 1_{[X_i^0 \le U_i]},$$

with $1_{[\cdot]}$ denoting the indicator random variable of the event $[\cdot]$.

Based on (X_i, Δ_i), $i = 1, \ldots, n$, a popular estimator of the survival function $1 - F^0(t)$ is the product-limit (PL) estimator, proposed by Kaplan and Meier (1958) and shown to be "self-consistent" by Efron (1967). Let (Z_i, Λ_i), $i = 1, \ldots, n$, denote the ordered X_i's along with their corresponding Δ_i's. The PL estimator of $1 - F^0(t)$ is defined by

$$\hat{P}_n(t) = \begin{cases} 1, & 0 \le t \le Z_1, \\ \displaystyle\prod_{i=1}^{k-1} \left(\frac{n-i}{n-i+1} \right)^{\Lambda_i}, & Z_{k-1} < t \le Z_k,\ k = 2, \ldots, n, \\ 0, & t > Z_n. \end{cases}$$

Denote the PL estimator of $F^0(t)$ by $\hat{F}_n(t) = 1 - \hat{P}_n(t)$ and let s_j denote the jump of $\hat{P}_n$ (or $\hat{F}_n$) at Z_j, that is,

$$s_j = \begin{cases} 1 - \hat{P}_n(Z_2), & j = 1, \\ \hat{P}_n(Z_j) - \hat{P}_n(Z_{j+1}), & j = 2, \ldots, n-1, \\ \hat{P}_n(Z_n), & j = n. \end{cases}$$

Then for $j < n$, $s_j = 0$ if and only if $\Lambda_j = 0$, that is, Z_j is a censored observation. For various properties of the PL estimator, see Breslow and Crowley (1974), Csörgő and Horváth (1983), Földes and Rejtö (1981), Földes, Rejtö and Winter (1980), Gill (1983) and Wellner (1982), among others.

The distribution function estimator $\hat{F}_n$ is very naturally used to construct a density estimator by defining

$$f_n(x) = h^{-1}\int_{-\infty}^{\infty} K\left(\frac{x-t}{h}\right) d\hat{F}_n(t)$$
$$= h^{-1}\sum_{j=1}^{n} s_j K\left(\frac{x-Z_j}{h}\right).$$

This estimator has been studied by Földes, Rejtö and Winter (1981), McNichols and Padgett (1986), Diehl and Stute (1985), Stute (1985) and Mielniczuk (1986).

An alternative kernel estimator has been proposed by Blum and Susarla (1980), extending the results of Rosenblatt (1976) to censored data. It is motivated by the fact that a reasonable (and technically easy to handle) estimate of $f^{0}(x)H^{*}(x)$ is given by

$$(f^{0}H^{*})_n(x) \equiv (nh)^{-1}\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)1_{[\Delta_j=1]}.$$

Hence, it makes sense to estimate $f^{0}(x)$ by $(f^{0}H^{*})_n(x)$ divided by an estimate of $H^{*}(x)$. If we reverse the intuitive roles played by X_i^0 and U_i, then the product-limit estimator for H^{*} is given by

$$\hat{H}_n(t) = \begin{cases} 1, & 0 \le t \le Z_1, \\ \prod_{i=1}^{k-1}\left(\frac{n-i}{n-i+1}\right)^{1-\Lambda_i}, & Z_{k-1} < t \le Z_k,\ k = 2,\ldots,n, \\ 0, & t > Z_n. \end{cases}$$

This does not make a good denominator because it takes on the value zero, so Blum and Susarla propose changing $\hat{H}_n$ slightly to

$$H_n^{*}(t) = \begin{cases} 1, & 0 \le t \le Z_1, \\ \prod_{i=1}^{k-1}\left(\frac{n-i+1}{n-i+2}\right)^{1-\Lambda_i}, & Z_{k-1} < t \le Z_k,\ k = 2,\ldots,n, \\ \prod_{i=1}^{n}\left(\frac{n-i+1}{n-i+2}\right)^{1-\Lambda_i}, & Z_n < t. \end{cases}$$

Hence, define

$$f_n^{*}(x) = [nhH_n^{*}(x)]^{-1}\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)1_{[\Delta_j=1]}.$$

To get some idea of what the relationship is between the estimators f_n and f_n^{*}, note that from Susarla, Tsai and Van Ryzin (1984), for each j, $s_j =$

$\Lambda_j[n\hat{H}_n(Z_j)]^{-1}$. Hence, we may write

$$f_n(x) = \sum_{j=1}^{n} \frac{\Delta_j}{n\hat{H}_n(X_j)h} K\left(\frac{x - X_j}{h}\right), \tag{2.1}$$

$$f_n^*(x) = \sum_{j=1}^{n} \frac{\Delta_j}{nH_n^*(x)h} K\left(\frac{x - X_j}{h}\right). \tag{2.2}$$

Since $\hat{H}_n$ and H_n^* are essentially the same, the only significant difference between the estimators is the argument of the estimate of H^*. It will be seen in the next section that the difference is typically not negligible.

It will be assumed throughout that K is a probability density with compact support and that K is Hölder continuous. Further, it is assumed that f^0H^* and f^0 are Hölder continuous of order $\alpha > 0$. In addition, $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

3. Asymptotic representation. The main idea of this section is that $f_n(x)$ and $f_n^*(x)$ are essentially the same as

$$\bar{f}_n(x) = \sum_{j=1}^{n} \frac{\Delta_j}{nH^*(X_j)h} K\left(\frac{x - X_j}{h}\right),$$

$$\bar{f}_n^*(x) = \sum_{j=1}^{n} \frac{\Delta_j}{nH^*(x)h} K\left(\frac{x - X_j}{h}\right), \tag{3.1}$$

respectively, because the convergence of $\hat{H}_n$ and H_n^* to H^* is faster ($\sim n^{-1/2}$) than that of the density estimators (often $\sim n^{-2/5}$). Essentially, the same idea has been used by Diehl and Stute (1985) and Stute (1985). For $\hat{f}$ equal to any of f_n, f_n^*, $\bar{f}_n$ or $\bar{f}_n^*$, we choose to analyze its performance by studying the integrated squared error $\mathrm{ISE}(\hat{f}) = \int_0^\infty [\hat{f}(x) - f^0(x)]^2 w(x)\,dx$, where $w(x)$ is a nonnegative weight function.

There are three major reasons for working with ISE instead of with its expected value MISE. First, MISE will typically be infinite for the estimators based on $\hat{H}_n$. Second, ISE is a more compelling error criterion because it assesses how well $\hat{f}$ is doing *for the data set at hand*, instead of only for the average over all possible data sets as is done by MISE. Third, ISE is more natural for the automatic bandwidth selection results of the next section. It should be pointed out that by using methods slightly easier than those used here, all of our results can be formulated in terms of MISE when it exists. Also, there is an obvious extension of the theorems of this section to the pointwise convergence of the estimators when it is assessed by the mean squared error.

The role of the weight function w is to eliminate endpoint effects. Assume in particular that w is bounded and supported on $[0, T]$, where $T < \min(T_H, T_{F_0})$, where $T_G \equiv \sup\{t: G(t) < 1\}$ for a distribution function G.

The statement of the theorem will be uniform over $h \in [n^{-1+\varepsilon}, n^{-\varepsilon}]$, some $\varepsilon > 0$. This is necessary for the automatic bandwidth selection results of Section 4.

THEOREM 3.1. *Under the conditions on w, K, f^0H^*, F^0, f^0 and H stated in Sections 2 and 3, for $h \in [n^{-1+\epsilon}, n^{-\epsilon}]$, we have*

$$\sup_h \left| \frac{\mathrm{ISE}(f_n) - [an^{-1}h^{-1} + b]}{an^{-1}h^{-1} + b} \right| \to 0 \quad a.s.,$$

$$\sup_h \left| \frac{\mathrm{ISE}(f_n^*) - [an^{-1}h^{-1} + b^*]}{an^{-1}h^{-1} + b^*} \right| \to 0 \quad a.s.,$$

(3.2)

where

$$a = \left(\int K^2\right)\left(\int \frac{f^0 w}{H^*}\right)$$

and where b, b^* are defined by

$$b = \int B(x,h)^2 w(x)\,dx, \qquad b^* = \int B^*(x,h)^2 \frac{w(x)}{[H^*(x)]^2}\,dx,$$

$$B(x,h) = \int K(u)[f^0(x-hu) - f^0(x)]\,du,$$

$$B^*(x,h) = \int K(u)[f^0(x-hu)H^*(x-hu) - f^0(x)H^*(x)]\,du.$$

REMARK 3.1. Note that with the Hölder continuity conditions on f^0 and f^0H^*, an immediate consequence of Theorem 3.1 is the ISE consistency of f_n and f_n^*.

REMARK 3.2. The only difference in the asymptotic representations of ISE shows up in the bias part. Note that for some choices of f^0 and H^*, b will be smaller, while for other choices, b^* will be smaller. Hence, the estimators f_n and f_n^* are really not comparable from this representation. However, note that, by an addition–subtraction,

$$\frac{B^*(x,h)}{H^*(x)} = \int K(u) f^0(x-hu)\left[\frac{H^*(x-hu) - H^*(x)}{H^*(x)}\right] dx + B(x,h).$$

So in a weak sense, f_n^* has an extra "noise term," which may make f_n slightly preferable.

Rates of convergence may be computed in the usual manner of Rosenblatt and Parzen. Further, Theorem 3.1 yields an asymptotic bandwidth which is optimal in the same sense as the bandwidths of Rosenblatt and Parzen except that the random error criterion ISE is used in place of its mean. This is given in the next remark.

REMARK 3.3. (i) It is well known in the complete sample case that by allowing K to take on negative values, a faster rate of convergence can be

obtained. Theorem 3.1 demonstrates that the same is true here. In particular, suppose

$$\int x^j K(x)\,dx = \begin{cases} 1, & j = 0, \\ 0, & j = 1, \ldots, k-1, \\ \kappa, & j = k \end{cases} \tag{3.3}$$

(for $k > 2$, this violates the assumptions of Theorem 3.1; however, it is straightforward, but space-consuming, to modify the proofs to allow for this). If we assume that f^0 and $f^0 H^*$ have k uniformly continuous derivatives, then

$$b = h^{2k}\left(\frac{\kappa}{k!}\right)^2 \int \left[(f^0)^{(k)}\right]^2 w\,dx + o(h^{2k}),$$

$$b^* = h^{2k}\left(\frac{\kappa}{k!}\right)^2 \int \left[(f^0 H^*)^{(k)}\right]^2 \frac{w}{(H^*)^2}\,dx + o(h^{2k}).$$

Hence, for the estimator f_n, the "classical optimal bandwidth" has the form

$$h_0 = \left\{ \frac{(\int K^2)(\int (f^0 w/H^*))}{2k(\kappa/k!)^2 \left[\int \left((f^0)^{(k)}\right)^2 w\right]} \right\}^{1/(2k+1)} n^{-1/(2k+1)}$$

and the rate of convergence is ISE $\sim n^{-2k/(2k+1)}$. Here and in the following remarks, there are obvious analogues for the estimator f_n^*.

To see how Theorem 3.1 implies that h_0 behaves like the optimal bandwidth of Rosenblatt and Parzen (the complete sample case), define

$$EI_0 = n^{-1}h^{-1}\left[\int K^2\right]\left[\int \frac{f^0 w}{H^*}\right] + h^{2k}\left(\frac{\kappa}{k!}\right)^2 \int \left[(f^0)^{(k)}\right]^2 w.$$

By (3.2), with obvious notation,

$$\sup_h \left| \frac{\mathrm{ISE}(f_n, h) - EI_0(h)}{EI_0(h)} \right| \to 0 \quad \text{a.s.}$$

Let h_M denote the minimizer of $\mathrm{ISE}(f_n, h)$ and recall that h_0 is the minimizer of $EI_0(h)$. Then from the inequalities $\mathrm{ISE}(f_n, h_0) \geq \mathrm{ISE}(f_n, h_M)$ and $EI_0(h_M) \geq EI_0(h_0)$, it follows that

$$\begin{aligned} \frac{|\mathrm{ISE}(f_n, h_0) - \mathrm{ISE}(f_n, h_M)|}{\mathrm{ISE}(f_n, h_0)} &\leq \left| \frac{\mathrm{ISE}(f_n, h_0) - EI_0(h_0)}{EI_0(h_0)} \right| \frac{EI_0(h_0)}{\mathrm{ISE}(f_n, h_0)} \\ &\quad + \left| \frac{\mathrm{ISE}(f_n, h_M) - EI_0(h_M)}{EI_0(h_M)} \right| \frac{EI_0(h_M)}{\mathrm{ISE}(f_n, h_M)} \\ &\to 0 \quad \text{a.s.} \end{aligned}$$

Hence,

$$\frac{\mathrm{ISE}(f_n, h_0)}{\inf_h \mathrm{ISE}(f_n, h)} \to 1 \quad \text{a.s.},$$

which shows that h_0 is optimal in the same sense as the bandwidths of Rosenblatt and Parzen, except for the fact that the random ISE criterion is used in place of its mean.

REMARK 3.3. (ii) If we keep the assumption (3.3), but suppose f^0 has $p < k$ derivatives (p need not be an integer by putting a Hölder condition of order $p - [p]$ on the $[p]$th derivative, where $[\cdot]$ denotes the greatest integer less than or equal to p), then it can be shown that $b^* \leq Ch^{2p}$ for some positive constant C. Hence, by taking $h \sim n^{-1/(2p+1)}$, the well known [see, for example, Bretagnolle and Huber (1979)] "optimal rate" ISE $\sim n^{-2p/(2p+1)}$ can be obtained for our censored-data problem.

4. Automatic bandwidth selection. For data-based bandwidth selection, we propose least-squares cross-validation, which was invented for complete sample density estimators by Rudemo (1982) and Bowman (1984). This is motivated as follows. Let $\hat{f}$ denote either f_n or f_n^*. Since the third term of

$$\mathrm{ISE}(\hat{f}) = \int \hat{f}^2 w - 2\int \hat{f} f^0 w + \int (f^0)^2 w$$

is independent of h, we would like to choose h to minimize the sum of the first two terms. The first term is known. To gain insight into how the second term may be estimated, note that by the type of argument given in Section 3, we can replace H_n^* by its expected value. In this sense, the integral of the second term can be nearly unbiasedly estimated by

$$n^{-1} \sum_{i=1}^{n} \hat{f}_i(X_i) \frac{w(X_i)}{H_n^*(X_i)} 1_{[\Delta_i = 1]}, \tag{4.1}$$

where $\hat{f}_i$ is the "leave-one-out" version of $\hat{f}$, given by

$$f_{n,i}(x) = \sum_{j \neq i} \frac{1}{(n-1)H_n^*(X_j)h} K\left(\frac{x - X_j}{h}\right) 1_{[\Delta_j = 1]}$$

when $\hat{f}$ is f_n and by

$$f_{n,i}^*(x) = \sum_{j \neq i} \frac{1}{(n-1)H_n^*(x)h} K\left(\frac{x - X_j}{h}\right) 1_{[\Delta_j = 1]}$$

when $\hat{f}$ is f_n^*. Thus, we define $\hat{h}_c$ to be the minimizer of the least-squares cross-validation criterion

$$\mathrm{CV}(h) = \int [\hat{f}(x)]^2 w(x)\, dx - 2n^{-1} \sum_{i=1}^{n} \hat{f}_i(X_i) \frac{w(X_i)}{H_n^*(X_i)} 1_{[\Delta_i = 1]}.$$

THEOREM 4.1. *Under the conditions of Theorem* 3.1, $\hat{h}_c$ *is asymptotically optimal in the sense that*

$$\frac{\mathrm{ISE}(\hat{f}, \hat{h}_c)}{\inf_h \mathrm{ISE}(\hat{f}, h)} \to 1 \quad a.s.$$

REMARK 4.1. Theorem 4.1 says that $\hat{h}_c$ is optimal under *either* of the assumptions stated in Remark 3.3(i) or (ii). This generalizes the important asymptotic optimality results of Hall (1983), Stone (1984), Burman (1985), Hall (1985) and Marron (1985) to the case of censored data.

REMARK 4.2. The fact that CV(h) essentially provides an estimate of ISE($\hat{f}$, h) suggests a practical method of choosing between f_n and f_n^*. In particular, if CV(h) for $\hat{f} = f_n$ is smaller than CV(h) for $\hat{f} = f_n^*$, then the estimator f_n should be used, as its ISE will probably be smaller.

REMARK 4.3. The hazard rate $r^0(x) \equiv f^0(x)/[1 - F^0(x)]$ can be estimated by using one of the density estimators f_n or f_n^* together with a reasonable estimator of $1 - F^0$. Thus, it is straightforward to use the results of Csörgő and Horváth (1983) to prove hazard rate analogues of all of the results of this paper.

5. Proofs of theorems. All proofs are given for the estimator $f_n(x)$, as it will be obvious how to adapt them to handle $f_n^*(x)$. The symbol C will be used for a generic constant. Note first that, using the notation (3.1), by adding and subtracting $\bar{f}_n(x)$,

$$\text{ISE}(f_n) = \text{ISE}(\bar{f}_n) + \text{II} + \text{III}, \tag{5.1}$$

where

$$\text{II} = 2\int_0^\infty [\bar{f}_n(x) - f^0(x)][f_n(x) - \bar{f}_n(x)]w(x)\,dx,$$

$$\text{III} = \int_0^\infty [f_n(x) - \bar{f}_n(x)]^2 w(x)\,dx.$$

PROOF OF THEOREM 3.1. We analyze each of the terms ISE($\bar{f}_n$), II and III separately. First, by a "variance-bias squared" decomposition and standard computations of the type in Rosenblatt (1971),

$$\text{MISE}(\bar{f}_n) \equiv E(\text{ISE}(\bar{f}_n)) = v + b, \tag{5.2}$$

where

$$v = n^{-1}h^{-1}\left(\int K^2\right)\left(\int \frac{f^0 w}{H^*}\right) + o(n^{-1}h^{-1}) \tag{5.3}$$

and where b is defined in Section 3. The fact that ISE($\bar{f}_n$) behaves like MISE($\bar{f}_n$) is contained in the following lemma.

LEMMA 1.

$$\sup_h \left| \frac{\text{ISE}(\bar{f}_n) - \text{MISE}(\bar{f}_n)}{\text{MISE}(\bar{f}_n)} \right| \to 0 \quad a.s.$$

The fact that term III is negligible is contained in

LEMMA 2.

$$\sup_h \left| \frac{\text{III}}{\text{MISE}(\bar{f}_n)} \right| \to 0.$$

It follows from the Schwarz inequality, Lemma 1 and Lemma 2 that III may be replaced by II in the statement of Lemma 2.

This last fact, together with (5.1), (5.2), (5.3), Lemma 1 and Lemma 2, completes the proof of Theorem 3.1. □

Before proving Theorem 4.1, we give the proofs of Lemmas 1 and 2.

PROOF OF LEMMA 1. Let $N = \#(\Delta_i = 1)$. For $\nu = 1, \ldots, n$, conditioning on $[N = \nu]$, $\{X_i\colon \Delta_i = 1\}$ is a set of ν i.i.d. random variables with density $f^0 H^*/p$, where

$$p = \int_0^\infty f^0(x) H^*(x)\, dx.$$

Let E_ν denote expectation under this conditional distribution. The method of the proof of Theorem 1 of Marron and Härdle (1986) shows that under the stated assumptions, for $k = 1, 2, \ldots$, there exist constants $C > 0$ and $\gamma > 0$ so that

$$\sup_h E_\nu \left[\frac{\text{ISE}(\bar{f}_n) - E_\nu(\text{ISE}(\bar{f}_n))}{E_\nu(\text{ISE}(\bar{f}_n))} \right]^{2k} \leq C\nu^{-\gamma k}. \tag{5.4}$$

To analyze $E_\nu(\text{ISE}(\bar{f}_n))$, note first that

$$\begin{aligned} E_\nu \bar{f}_n(x) - f^0(x) &= \frac{\nu}{n} \int \frac{1}{H^*(y)h} K\left(\frac{x-y}{h}\right) \frac{f^0(y)H^*(y)}{p}\, dy - f^0(x) \\ &= \int K(u) \left[\frac{\nu}{np} f^0(x - hu) - f^0(x) \right] du \\ &= \frac{\nu}{np} B(x, h) + \left(\frac{\nu}{np} - 1 \right) f^0(x), \end{aligned}$$

where $B(x, h)$ was defined in Section 3. Next note that

$$\begin{aligned} E_\nu \left[\bar{f}_n(x) - E_\nu \bar{f}_n(x) \right]^2 &= \text{var}_\nu \left[\sum_{i=1}^n \frac{1}{nH^*(X_i)h} K\left(\frac{x - X_i}{h}\right) 1_{[\Delta_i = 1]} \right] \\ &= \frac{\nu}{n^2} \text{var}_\nu \left[\frac{1}{H^*(X_i)h} K\left(\frac{x - X_i}{h}\right) \right] \\ &= \frac{\nu}{np} n^{-1}h^{-1} \left(\int K^2 \right) \frac{f^0(x)}{H^*(x)} + o\left(\frac{\nu}{np} n^{-1}h^{-1} \right). \end{aligned}$$

Thus, by a variance-bias squared decomposition,

$$E_\nu(\text{ISE}(\bar{f}_n)) = v_\nu + b_\nu,$$

where

$$v_\nu = \frac{\nu}{np} v + o\left(\frac{\nu}{np} n^{-1}h^{-1} \right)$$

for v defined in (5.3) and where

$$b_\nu = \left(\frac{\nu}{np}\right)^2 b + 2\left(\frac{\nu}{np}\right)\left(\frac{\nu}{np} - 1\right)\int_0^\infty B(x, h) f^0(x) w(x) \cdot dx + \left(\frac{\nu}{np} - 1\right)^2 \int_0^\infty f^0(x)^2 w(x)\, dx$$

for b and B as in Section 3. Hence,

$$E_\nu(\mathrm{ISE}(\bar{f}_n)) = \mathrm{MISE}(\bar{f}_n) + \left(\frac{\nu}{np} - 1\right)v + o\left(\frac{\nu}{np} n^{-1}h^{-1}\right) + \left(\left(\frac{\nu}{np}\right)^2 - 1\right)b + 2\frac{\nu}{np}\left(\frac{\nu}{np} - 1\right)\int_0^\infty B(x, h) f^0(x) w(x)\, dx + \left(\frac{\nu}{np} - 1\right)^2 \int_0^\infty f^0(x)^2 w(x)\, dx.$$

Now for small $\tau > 0$ and for $n = 1, 2, 3, \ldots$, restrict attention to ν between $np - n^{1/2+\tau}$ and $np + n^{1/2+\tau}$. For such ν, $\nu/np \le 2$ and

$$\left|\frac{\nu}{np} - 1\right| \le C_1 n^{-1/2+\tau}$$

for a constant C_1. It follows from (5.2) and (5.3) that, for a different value of C and for n sufficiently large,

$$\inf_h \mathrm{MISE}(\bar{f}_n) \ge C n^{-1+\varepsilon}. \tag{5.5}$$

Hence, for small τ, large n and another C,

$$\sup_h \left|\frac{E_\nu(\mathrm{ISE}(\bar{f}_n)) - \mathrm{MISE}(\bar{f}_n)}{\mathrm{MISE}(\bar{f}_n)}\right| \le C n^{-\varepsilon+2\tau}.$$

Thus, for such ν, from (5.4),

$$\sup_h E_\nu\left[\frac{\mathrm{ISE}(\bar{f}_n) - \mathrm{MISE}(\bar{f}_n)}{\mathrm{MISE}(\bar{f}_n)}\right]^{2k} \le C n^{-\gamma k}.$$

Now, let Γ_n be a subset of $[n^{-1+\varepsilon}, n^{-\varepsilon}]$ so that successive members of Γ_n are separated by a distance less than or equal to $n^{-\rho}$ and so that $\#(\Gamma_n) \le n^\rho$ for some $\rho > 0$. Then, using obvious notation and letting $M(\hat{f}, h) \equiv [\mathrm{ISE}(\hat{f}, h) - \mathrm{MISE}(\hat{f}, h)]/\mathrm{MISE}(\hat{f}, h)$,

$$P\left[\sup_h |M(\bar{f}_n, h)| \ge \varepsilon\right] \le P\left[\sup_{h\in\Gamma_n} |M(\bar{f}_n, h)| > \frac{\varepsilon}{2}\right] + P\left[\sup_{|h-h'|\le n^{-\rho}} |M(\bar{f}_n, h) - M(\bar{f}_n, h')| > \frac{\varepsilon}{2}\right] = \sum_{\nu=0}^{n} \binom{n}{\nu} p^\nu (1-p)^{n-\nu} P_\nu\left[\sup_{h\in\Gamma_n} |M(\bar{f}_n, h)| > \frac{\varepsilon}{2}\right],$$

where the last equality comes from a continuity argument and the assumptions that K is Hölder continuous and has compact support. Letting $A_{n,\tau} = [np - n^{1/2+\tau}, np + n^{1/2+\tau}]$,

$$
\begin{aligned}
&P\left[\sup_h \left|M(\bar{f}_n, h)\right| > \varepsilon\right] \\
&\quad \le \sum_{\nu \in A_{n,\tau}} \binom{n}{\nu} p^\nu (1-p)^{n-\nu} P_\nu\left[\sup_h \left|M(\bar{f}_n, h)\right| > \frac{\varepsilon}{2}\right] \\
&\qquad + \sum_{\nu \notin A_{n,\tau}} \binom{n}{\nu} p^\nu (1-p)^{n-\nu} \\
&\quad \le \sum_{\nu \in A_{n,\tau}} \binom{n}{\nu} p^\nu (1-p)^{n-\nu} \sum_{h \in \Gamma_n} P_\nu\left[\left|M(\bar{f}_n, h)\right| > \frac{\varepsilon}{2}\right] + 2\Phi(-n^\tau) \\
&\quad \le \sum_{\nu \in A_{n,\tau}} \binom{n}{\nu} p^\nu (1-p)^{n-\nu} n^\rho \left(\frac{2}{\varepsilon}\right)^{2k} \sup_h E_\nu\left[M(\bar{f}_n, h)\right]^{2k} + 2\Phi(-n^\tau) \\
&\quad \le Cn^\rho n^{-\gamma k} + 2\Phi(-n^\tau),
\end{aligned} \tag{5.6}
$$

where Φ denotes the standard normal c.d.f. But, for k sufficiently large, the first term on the right side of (5.6) is summable on n and, since the second term is also summable on n, the proof of Lemma 1 is complete. □

PROOF OF LEMMA 2. Using the assumption on the support of w and using the compactness of the support of K, observe that for n sufficiently large,

$$
\begin{aligned}
\sup_h \text{III} &= \sup_h \int_0^\infty \left[\sum_{i=1}^n \left(\frac{1}{\hat{H}_n(X_i)} - \frac{1}{H^*(X_i)}\right) \frac{1}{nh} K\left(\frac{x - X_i}{h}\right) 1_{[\Delta_i = 1]}\right]^2 w(x)\, dx \\
&\le \left(\sup_{t \in [0, T']} \left|\frac{1}{\hat{H}_n(t)} - \frac{1}{H^*(t)}\right|\right)\left(\sup_h \int_0^\infty \left[\left(f\, {}^0H^*\right)_n(x)\right]^2 w(x)\, dx\right),
\end{aligned}
$$

where $T' = (T + T_H)/2$ and where $(f\, {}^0H^*)_n$ was defined in Section 2. Lemma 2 is now a consequence of the results of Csörgő and Horváth (1983) together with (5.5) and the fact that there is a constant C so that

$$
\sup_h \int_0^\infty \left[\left(f\, {}^0H^*\right)_n(x)\right]^2 w(x)\, dx \le C \quad \text{a.s.} \tag{5.7}
$$

To verify (5.7), note that by adding and subtracting $f\, {}^0(x)H^*(x)$,

$$
\int_0^\infty \left[\left(f\, {}^0H^*\right)_n(x)\right]^2 w(x)\, dx = U + V + W,
$$

where

$$
U = \int_0^\infty \left[\left(f\, {}^0H^*\right)_n - f\, {}^0H^*\right]^2 w(x)\, dx,
$$

$$
V = 2\int \left[\left(f\, {}^0H^*\right)_n - f\, {}^0H^*\right]\left[f\, {}^0H^*\right] w(x)\, dx,
$$

$$
W = \int \left[f\, {}^0H^*\right]^2 w(x)\, dx.
$$

Now W is deterministic and independent of h. An argument similar to (but slightly easier than) that used previously on $\mathrm{ISE}(\bar{f}_n)$ gives

$$\sup_h U \to 0 \quad \text{a.s.}$$

An application of the Schwarz inequality to V yields (5.7), which completes the proof of Lemma 2. □

PROOF OF THEOREM 4.1. Here again, only the proof in the slightly harder case of $\hat{f} = f_n$ is given. We note that by a computation similar to that used to verify Remark 3.3(i), Theorem 4.1 follows from (3.2) and the result that

$$\sup_{h,\,h'} \frac{\left|\mathrm{CV}(h) - \mathrm{ISE}(f_n, h) - [\mathrm{CV}(h') - \mathrm{ISE}(f_n, h')]\right|}{\mathrm{MISE}(f_n, h) + \mathrm{MISE}(f_n, h')} \to 0 \quad \text{a.s.} \tag{5.8}$$

To prove (5.8), it is enough to show that

$$\sup_h \frac{\left|\mathrm{CV}(h) - \mathrm{ISE}(f_n, h) - 2\left[\frac{1}{n}\sum_{i=1}^{n} \frac{f^0(X_i)w(X_i)}{H_n^*(X_i)} 1_{[\Delta_i=1]} - \int (f^0(x))^2 w(x)\,dx\right]\right|}{\mathrm{MISE}(f_n, h)}$$

$$\to 0 \quad \text{a.s.}$$

This may be rewritten as

$$\sup_h \frac{\left|2n^{-1}(n-1)^{-1}\Sigma_{i=1}^{n}\Sigma_{j\neq i} U_{ij}\right|}{\mathrm{MISE}(f_n, h)} \to 0 \quad \text{a.s.},$$

where

$$\begin{aligned} U_{ij} &= h^{-1}K\left(\frac{X_i - X_j}{h}\right)\frac{w(X_i)}{H_n^*(X_i)H_n^*(X_j)} 1_{[\Delta_i=1,\,\Delta_j=1]} \\ &\quad - \int h^{-1}K\left(\frac{x - X_j}{h}\right)\frac{f^0(x)w(x)}{H_n^*(X_j)}\,dx\, 1_{[\Delta_j=1]} \\ &\quad - \frac{f^0(X_i)w(X_i)}{H_n^*(X_i)} 1_{[\Delta_i=1]} + \int [f^0(x)]^2 w(x)\,dx \\ &\equiv U_{ij}' + Z_{ij}, \end{aligned}$$

defining

$$\begin{aligned} U_{ij}' &= h^{-1}K\left(\frac{X_i - X_j}{h}\right)\frac{w(X_i)}{H_n^*(X_j)H^*(X_i)} 1_{[\Delta_i=1,\,\Delta_j=1]} \\ &\quad - \int h^{-1}K\left(\frac{x - X_j}{h}\right)\frac{f^0(x)w(x)}{H_n^*(X_j)}\,dx\, 1_{[\Delta_j=1]} \\ &\quad - \frac{f^0(X_i)w(X_i)}{H^*(X_i)} 1_{[\Delta_i=1]} + \int [f^0(x)]^2 w(x)\,dx \end{aligned}$$

and

$$Z_{ij} = \left[h^{-1}K\left(\frac{X_i - X_j}{h}\right)\frac{w(X_i)}{H_n^*(X_j)}1_{[\Delta_i=1,\,\Delta_j=1]} - f^{\,0}(X_i)w(X_i)1_{[\Delta_i=1]}\right] \times\left[\frac{1}{H_n^*(X_i)} - \frac{1}{H^*(X_i)}\right].$$

Theorem 4.1 then follows from the following two lemmas.

LEMMA 3.

$$\sup_h \frac{\left|n^{-1}(n-1)^{-1}\Sigma_{i=1}^n\Sigma_{j\neq i}U_{ij}'\right|}{\mathrm{MISE}(\hat{f}, h)} \to 0 \quad a.s.$$

LEMMA 4.

$$\sup_h \frac{\left|n^{-1}(n-1)^{-1}\Sigma_{i=1}^n\Sigma_{j\neq i}Z_{ij}\right|}{\mathrm{MISE}(\hat{f}, h)} \to 0 \quad a.s.$$

PROOF OF LEMMA 3. This proof combines the ideas of Lemma 2 of Marron (1985) with those of the earlier proof of Lemma 1. Recall that in the proof of Lemma 1, the notation E_ν meant expected value taken over $\{X_i\colon \Delta_i = 1\}$, conditioned on the event $\{N = \nu\}$. The censored observations $\{X_i\colon \Delta_i = 0\}$ were ignored in the definition of E_ν since they did not appear in the quantities being analyzed. The censored X_i's do appear in the following, so it will be understood that E_ν denotes expected value as before, only also conditioned on $\{X_i\colon \Delta_i = 0\}$ [or, equivalently, E_ν denotes integration over $\{X_i\colon \Delta_i = 1\}$, which are i.i.d. random variables with density $f^{\,0}H^*/p$].

For $\nu = 1, \ldots, n$, $U_{ij}' = U_{ij}'' + Z_{j\nu}'$, where

$$U_{ij}'' = h^{-1}K\left(\frac{X_i - X_j}{h}\right)\frac{w(X_i)}{H_n^*(X_j)H^*(X_i)}1_{[\Delta_i=1,\,\Delta_j=1]} - \frac{\nu}{np}\int h^{-1}K\left(\frac{x - X_i}{h}\right)\frac{w(x)}{H_n^*(x)}f^{\,0}(x)\,dx\,1_{[\Delta_j=1]} - \frac{f^{\,0}(X_i)w(X_i)}{H^*(X_i)} + \frac{\nu}{np}\int\left[f^{\,0}(x)\right]^2 w(x)\,dx$$

and

$$Z_{j\nu}' = \left(\frac{\nu}{np} - 1\right)\left[\int h^{-1}K\left(\frac{x - X_j}{h}\right)\frac{w(x)}{H_n^*(x)}f^{\,0}(x)\,dx\,1_{[\Delta_j=1]} - \int\left[f^{\,0}(x)\right]^2 w(x)\,dx\right].$$

Using the method of proof of Lemma 2 of Marron (1985), it can be shown that, for $k = 1, 2, \ldots$ and n sufficiently large,

$$\sup_h E_\nu \left[\frac{n^{-1}(n-1)^{-1}\Sigma_{i=1}^n \Sigma_{j \neq i} U_{ij}''}{\mathrm{MISE}(\hat{f}, h)} \right]^{2k} \leq Cn^{-\gamma k},$$

regardless of the realization of $\{X_i\colon \Delta_i = 0\}$. In a similar manner [i.e., approximate $H_n^*(x)$ by $H^*(x)$, including another $\nu/np - 1$ term and using the cumulant-style argument of Marron (1985)], we can obtain

$$\sup_h E_\nu \left[\frac{n^{-1}\Sigma_{j=1}^n Z_{j\nu}'}{\mathrm{MISE}(\hat{f}, h)} \right]^{2k} \leq Cn^{-\gamma k}.$$

These two inequalities may now be used in a computation similar to that yielding (5.6) in the proof of Lemma 1 to finish the proof of Lemma 3. □

PROOF OF LEMMA 4. Write

$$\begin{aligned} \left| n^{-1}(n-1)^{-1} \sum_i \sum_{j \neq i} Z_{ij} \right| &= \left| n^{-1} \sum_i \left[f_{ni}(X_i) - f^0(X_i) \right] \right. \\ &\qquad \left. \times \left[\frac{1}{H_n^*(X_i)} - \frac{1}{H^*(X_i)} \right] 1_{[\Delta_i = 1]} w(X_i) \right| \\ &\leq \left\{ n^{-1} \sum_i \left[f_{ni}(X_i) - f^0(X_i) \right]^2 1_{[\Delta_i = 1]} w(X_i) \right\}^{1/2} \\ &\qquad \times \left\{ n^{-1} \sum_i \left[\frac{1}{H_n^*(X_i)} - \frac{1}{H^*(X_i)} \right]^2 1_{[\Delta_i = 1]} w(X_i) \right\}^{1/2} \end{aligned} \tag{5.9}$$

The expression inside the first square root on the right-hand side of (5.9) is the leave-one-out version of the average squared error and will be denoted by $\mathrm{ASE}(f_{ni})$. Using the methods of Lemma 1 of Marron (1985) and Theorem 2 of Marron and Härdle (1986), it can be shown that, for $k = 1, 2, \ldots$, there is a constant C so that

$$E \left[\frac{\mathrm{ASE}(f_{ni}) - \mathrm{MISE}(f_n, h)}{\mathrm{MISE}(f_n, h)} \right]^{2k} \leq Cn^{-\gamma k}.$$

The proof of Lemma 4 is then completed by a computation like that leading to (5.6) in the proof of Lemma 1, which includes the uniform convergence result for the product-limit estimator H_n^* used in the proof of Lemma 2. □

Acknowledgment. The authors are grateful to a referee for the careful reading of the original manuscript and for several suggestions which improved the exposition.

REFERENCES

BLUM, J. R. and SUSARLA, V. (1980). Maximal deviation theory of density and failure estimates based on censored data. In *Multivariate Analysis V* (P. R. Krishnaiah, ed.) 213–222. North-Holland, New York.

BOWMAN, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360.

BRESLOW, N. and CROWLEY, J. (1974). A large sample study of the life table and product-limit estimates under random censorship. *Ann. Statist.* **2** 437–453.

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densites: Risque minimax. *Z. Wahrsch. verw. Gebiete* **47** 119–137.

BURMAN, P. (1985). A data dependent approach to density estimation. *Z. Wahrsch. verw. Gebiete* **69** 609–628.

CSÖRGŐ, S. and HORVÁTH, L. (1983). The rate of strong uniform consistency for the product-limit estimator. *Z. Wahrsch. verw. Gebiete* **62** 411–426.

DIEHL, S. and STUTE, W. (1986). Kernel density estimation in the presence of censoring. Technical Report, Univ. Giessen.

EFRON, B. (1967). The two-sample problem with censored data. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4** 831–853. Univ. California Press.

FÖLDES, A. and REJTÖ, L. (1981). A LIL type result for the product-limit estimator. *Z. Wahrsch. verw. Gebiete* **56** 75–86.

FÖLDES, A., REJTÖ, L. and WINTER, B. B. (1980). Strong consistency properties of nonparametric estimators for randomly censored data, I: The product-limit estimator. *Period. Math. Hungar.* **11** 233–250.

FÖLDES, A., REJTÖ, L. and WINTER, B. B. (1981). Strong consistency properties of nonparametric estimators for randomly censored data, II: Estimation of density and failure rate. *Period. Math. Hungar.* **12** 15–29.

GILL, R. (1983). Large sample behavior of the product-limit estimator on the whole line. *Ann. Statist.* **11** 49–58.

HALL, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156–1174.

HALL, P. (1985). Asymptotic theory of minimum integrated square error for multivariate density estimation. In *Multivariate Analysis VI* (P. R. Krishnaiah, ed.) 25–29. North-Holland, Amsterdam.

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

HALL, P. and MARRON, J. S. (1987b). On the amount of noise inherent in bandwidth selection for a kernel density estimator. *Ann. Statist.* **15** 163–181.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481.

MARRON, J. S. (1985). A comparison of cross-validation techniques in density estimation. North Carolina Institute of Statistics Mimeo Series 1568.

MARRON, J. S. and HÄRDLE, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. *J. Multivariate Anal.* **20** 91–113.

MCNICHOLS, D. T. and PADGETT, W. J. (1986). Mean and variance of a kernel density estimator under the Koziol–Green model of random censorship. *Sankhyā Ser. A* **48** 150–168.

MIELNICZUK, J. (1986). Some asymptotic properties of kernel estimators of a density function in case of censored data. *Ann. Statist.* **14** 766–773.

PADGETT, W. J. and MCNICHOLS, D. T. (1984). Nonparametric density estimation from censored data. *Comm. Statist. A—Theory Methods* **13** 1581–1611.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

ROSENBLATT, M. (1976). On the maximal deviation of k-dimensional density estimates. *Ann. Probab.* **4** 1009–1015.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

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

STUTE, W. (1985). Optimal bandwidth selection in pointwise density and hazard rate estimation when censoring is present. Technical Report, Univ. Giessen.

SUSARLA, V., TSAI, W. Y. and VAN RYZIN, J. (1984). A Buckley–James-type estimator for the mean with censored data. *Biometrika* **71** 624–625.

WATSON, G. S. and LEADBETTER, M. (1963). On the estimation of the probability density, I. *Ann. Math. Statist.* **34** 480–491.

WELLNER, J. (1982). Asymptotic optimality of the product-limit estimator. *Ann. Statist.* **10** 595–602.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27514

DEPARTMENT OF STATISTICS
UNIVERSITY OF SOUTH CAROLINA
COLUMBIA, SOUTH CAROLINA 29208