JOURNAL OF MULTIVARIATE ANALYSIS **61**, 38–60 (1997)
ARTICLE NO. MV971659

A Cross-Validation Bandwidth Choice for Kernel Density Estimates with Selection Biased Data

Colin O. Wu

John Hopkins University

This paper studies the risks and bandwidth choices of a kernel estimate of the underlying density when the data are obtained from s independent biased samples. The main results of this paper give the asymptotic representation of the integrated squared errors and the mean integrated squared errors of the estimate and establish

cally optimal in the sense of Stone (1984, *Ann. Statist.* **12**, 1285–1297). The finite sample properties of the cross-validation bandwidth are investigated through a Monte Carlo simulation.

1. INTRODUCTION

Suppose we observe s independent samples $X_{i1}, ..., X_{in_i}$ i.i.d. on $\mathbf{R}^d$ with distribution G_i and density g_i, $i = 1, ..., s$, with respect to Lebesgue measure. Here g_i depends on an underlying density function f with distribution function F such that

$$g_i(t) = \frac{w_i(t)}{W_i} f(t), \tag{1}$$

where w_i are known nonnegative weight functions and

$$W_i = \int w_i(t)\, f(t)\, dt < \infty.$$

In the literature, distributions satisfying (1) are usually referred as the *weighted distributions* or the *selection biased models*. In this paper, we

Received April 4, 1995; revised November 13, 1996.

AMS 1991 subject classifications: primary 62G07; secondary 62C20.

Key words and phrases: kernel density estimate, integrated squared error, bandwidth, nonparametric MLE, weighted distribution, biased sampling model, cross-validation.

0047-259X/97 $25.00

consider the estimation of the underlying density f based on the independent samples $X_{i1}, ..., X_{in_i}$, $i = 1, ..., s$. This type of data frequently arises in survey sampling, economics, epidemiology, reliability theory, and many other fields. Examples and applications of this type of data can be found in Cox (1969), Patil, Rao, and Zelen (1988), Patil and Taillie (1989), Morgenthaler and Vardi (1986), among others.

Theory and methods of nonparametric estimation with selection biased models have been mostly concentrated on the underlying distribution function F and the nonparametric maximum likelihood estimate (NPMLE) $\hat{F}_n$ of F introduced by Vardi (1985). Gill, Vardi, and Wellner (1988) further developed the necessary and sufficient conditions for the identifiability of F and the asymptotic efficiency of $\hat{F}_n$ in the sense described in Bickel, Klaassen, Ritov, and Wellner (1993). In density estimation, Jones (1991) obtained a kernel density estimate of f by smoothing the Vardi's NPMLE when $s = 1$, and Ahmad (1995) extended this estimate to a multivariate setting. Furthermore, Wu and Mao (1996) derived explicit forms of the asymptotic minimax kernels and bandwidths for the estimate of Jones (1991) and Ahmad (1995). Extending the estimation to the multi-sample case, Wu (1996a) studied the mean squared errors and the integrated mean squared errors of a kernel density estimate by smoothing the multisample version of the Vardi's NPMLE $\hat{F}_n$, and showed that this estimate is superior to any estimate obtained by a linear combination of $\tilde{f}_{(i)}$, where $\tilde{f}_{(i)}$ is the kernel density estimate of Jones (1991) and Ahmad (1995) constructed using the ith sample only.

The main theme of this paper is to establish the asymptotic representations of the integrated squared errors and the mean integrated squared errors for a multivariate generalization of the kernel NPMLE density estimate of Wu (1966a) and develop a natural cross-validation criterion for the corresponding data-driven bandwidth choices. Similar to the i.i.d. direct samples, the choice of the bandwidth plays a crucial role in the performance of the kernel NPMLE density estimate. Too small or too large of a bandwidth will lead to "undersmoothing" or "oversmoothing" of the estimate, respectively.

Under the i.i.d. direct samples, various bandwidth selection procedures in kernel density estimates have been discussed in the literature. A partial list of these results include Rudemo (1982), Bowman (1984), Hall (1983), Stone (1984), Marron (1985, 1987), and Marron and Härdle (1986), among others. Among all the popular bandwidth selection techniques, the least-squares cross-validation remains as a promising method. There are two main reasons for the popularity of the least-squares cross-validation. First, this method requires only the minimal condition that the kernels to be Hölder continuous; hence, it can be applied to a large class of kernel estimates. Second, it is asymptotically optimal in the sense of Stone (1984)

or Marron (1985), and such asymptotic optimality does not require excessive smoothness conditions on the density to be estimated as many other competing methods usually do. The bandwidth selection procedure of this paper is a natural extension of the least-squares cross-validation to the biased sampling case and is shown to be asymptotically optimal in the sense of Stone (1984) or Marron (1985). The finite sample properties of the kernel NPMLE density estimate and its cross-validation bandwidths are investigated through a Monte Carlo simulation study.

In Section 2, we first recall the identifiability conditions of (1) introduced by Vardi (1985) and Gill, Vardi and Wellner (1988), and then give a derivation of the kernel NPMLE density estimate. The asymptotic representations of the integrated squared errors (ISE) and the mean integrated squared errors (MISE) of the estimate are established in Section 3, while the asymptotic optimality of the cross-validation bandwidths is shown in Section 4. In Section 5, we present some simulation results for the estimate and its comparison with other intuitive estimation procedures. The proofs of the main results are deferred to Section 6.

2. KERNEL SMOOTHING OF NPMLE

In general, the distribution function F, hence f, is not identifiable nonparametrically, in the sense that there is no one-to-one mapping between the underlying density f and the weighted densities $g_1, ..., g_s$, if there is no restriction on f and the weight functions $w_1, ..., w_s$. Examples of nonidentifiable cases in stratified samples and case-control studies can be found in Vardi (1985) and Gill, Vardi, and Wellner (1988). For instance, if f has positive measure on the set $\{x: w_i(x) = 0 \text{ for all } i = 1, ..., s\}$, then f itself is not identifiable and it is only possible to estimate the conditional density $f(x \mid w_i(x) > 0 \text{ for some } i = 1, ..., s)$.

Let $\mathbf{S}$ be the support of f, that is, $\mathbf{S}$ is the smallest closet set which satisfies $\int_{\mathbf{S}} f(t)\, dt = 1$. Throughout this paper we assume that the following *support* and *graph connectedness* conditions of Gill, Vardi, and Wellner (1988) are satisfied:

A1: (support) $\mathbf{S}$ is a subset of the *of* $[x: w_i(x) > 0 \text{ for some } i = 1, ..., s]$

A2: (graph connectedness) For any $1 \leqslant i \leqslant s$ and $1 \leqslant j \leqslant s$, there exist $1 \leqslant i_1 \leqslant s, ..., 1 \leqslant i_k \leqslant s, 1 \leqslant k \leqslant s-2$, such that

$$\int 1_{[w_i(t)>0]} 1_{[w_{i_1}(t)>0]} f(t)\, dt > 0, \qquad \int 1_{[w_{i_k}(t)>0]} 1_{[w_j(t)>0]} f(t)\, dt > 0,$$

and

$$\int 1_{[w_{il}(t)>0]} 1_{[w_{il+1}(t)>0]} f(t)\, dt > 0 \qquad \text{for all} \quad l = 1, ..., k-1.$$

The following identifiability result is essentially Proposition 1.1 of Gill, Vardi, and Wellner (1988).

PROPOSITION 2.1 (Gill, Vardi, and Wellner). *The distribution F, hence f, is identifiable in the sense that there exists an one-to-one mapping between f and $(g_1, ..., g_s)$, if and only if A1 and A2 are satisfied.*

Now assume that the distributions given in (1) satisfy A1 and A2. Let $n = n_1 + \cdots + n_s$ be the overall sample size and $\lambda_{n_i} = n_i/n$ be the proportion of the observations in the ith sample relative to the overall sample. Let $F(A) = P(T \in A)$ for any random variable $T \in \mathbf{S}$ with distribution F and $A \subseteq \mathbf{S}$, it is straightforward to derive from (1) that

$$\begin{aligned}\bar{G}_n(A) &= \lambda_{n_1} G_1(A) + \cdots + \lambda_{n_s} G_s(A) \\ &= \int_A \left(\sum_{i=1}^{s} \frac{\lambda_{n_i} w_i(y)}{W_i} \right) f(y)\, dy\end{aligned}$$

and, consequently,

$$f(x) = \sum_{i=1}^{s} \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(x)}{W_r} \right)^{-1} \lambda_{n_i} g_i(x), \tag{2}$$

$$\begin{aligned}F(A) &= \int_A \sum_{i=1}^{s} \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(x)}{W_r} \right)^{-1} \lambda_{n_i}\, dG_i(x) \\ &= \int_A \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(x)}{W_r} \right)^{-1} d\bar{G}_n(x).\end{aligned} \tag{3}$$

Denote further that $V_i = W_i/W_s$ for all $i = 1, ..., s-1$ and $V_s = 1$. If W_i were known, then replacing G_i in (3) by the empirical measure $\hat{G}_{n_i}(A) = n_i^{-1} \sum_{j=1}^{n_i} 1_{[X_{ij} \in A]}$, an intuitive estimate of $F(A)$ is

$$F_n(A) = D_n^{-1} \int_A \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(y)}{V_r} \right)^{-1} d\hat{G}_n(y), \tag{4}$$

where $\hat{G}_n(A) = n^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n_i} 1_{[X_{ij} \in A]}$ and

$$D_n = \int_{\mathbf{S}} \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(y)}{V_r} \right)^{-1} d\hat{G}_n(y).$$

Here, in general, D_n does not necessarily equal 1. Equivalently, $F_n(A)$ is computed by summing up the point mass

$$\left(D_n \sum_{r=1}^{s} \frac{n_r w_r(X_{ij})}{V^r}\right)^{-1} = \left(nD_n \sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(X_{ij})}{V_r}\right)^{-1}$$

for all $X_{ij} \in A$. Applying kernel smoothing to this point mass, we obtain an intuitive estimate of $f(x)$,

$$f_n(x) = \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left\{\left(D_n \sum_{r=1}^{s} \frac{n_r w_r(X_{ij})}{V_r}\right)^{-1} h^{-d} K\left(\frac{x - X_{ij}}{h}\right)\right\}, \tag{5}$$

where $K(\cdot)\colon \mathbf{R}^d \to \mathbf{R}$ is a kernel function and $h \in \mathbf{R}^+$ is a bandwidth sequence.

Since W_i are generally unknown, a natural modification of $f_n(x)$ is to estimate V_i from the data and substitute the estimates into (5). Based on the obvious identity

$$H_{n_i}(W_1, ..., W_s) = W_i^{-1} \int_{\mathbf{S}} w_i(y) \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(y)}{W_r}\right)^{-1} d\bar{G}_n(y) = 1,$$

Vardi (1985) and Gill, Vardi, and Wellner (1988) showed that the equations

$$\begin{aligned} \hat{H}_{n_i}(\hat{V}_{n_1}, ..., \hat{V}_{n_{s-1}}, 1) &= \hat{V}_{n_i}^{-1} \int_{\mathbf{S}} w_i(y) \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(y)}{\hat{V}_{n_r}}\right)^{-1} d\hat{G}_n(y) \\ &= 1 \end{aligned} \tag{6}$$

for all $i = 1, ..., s$ have a unique solution $(\hat{V}_{n_1}, ..., \hat{V}_{n_{s-1}})$ with probability one when $n_1, ..., n_s$ are sufficiently large. Thus $(\hat{V}_{n_1}, ..., \hat{V}_{n_{s-1}})$ is a natural estimate of $(V_1, ..., V_{s-1})$.

The NPMLE of $F(A)$ proposed by Vardi (1985) is defined by

$$\hat{F}_n(A) = \hat{D}_n^{-1} \int_A \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(y)}{\hat{V}_{n_r}}\right)^{-1} d\hat{G}_n(y), \tag{7}$$

where $\hat{V}_{n_s} = 1$,

$$\hat{D}_n = \int_{\mathbf{S}} \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(y)}{\hat{V}_{n_r}}\right)^{-1} d\hat{G}_n(y). \tag{8}$$

Furthermore, an estimate of $\mathbf{W} = (W_i, ..., W_s)$ based on (7) is given by

$$\hat{\mathbf{W}} \equiv (\hat{W}_{n_1}, ..., \hat{W}_{n_s}), \tag{9}$$

where

$$\hat{W}_{n_i} = \int_{\mathbf{S}} w_i(t)\, d\hat{F}_n(t) = \hat{V}_{n_i} \hat{W}_{n_s}.$$

and

$$\hat{W}_{n_s} = \hat{D}_n^{-1} = \left(\int_{\mathbf{S}} \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(y)}{\hat{V}_{n_r}} \right)^{-1} d\hat{G}_n(y) \right)^{-1}.$$

When (6) has a unique solution, a natural estimate of f based on kernel smoothing of the NPMLE $\hat{F}_n$ can be defined by substituting $\hat{D}_n$ and $\hat{V}_{n_i}$ back to (5), so that

$$\hat{f}_n(x) = \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left\{ \left(\hat{D}_n \sum_{r=1}^{s} \frac{n_r w_r(X_{ij})}{\hat{V}_{n_r}} \right)^{-1} h^{-d} K\left(\frac{x - X_{ij}}{h} \right) \right\}. \tag{10}$$

When (6) does not have a unique solution, we do not have a computable NPMLE $\hat{F}_n$. In this case, we simply define $\hat{f}_n(x)$ to be $f_n(x)$. The sufficient and complete conditions of the existence of the unique solution of (6) can be found in Theorem 1.1 of Gill, Vardi, and Wellner (1988).

More generally, it is also possible to consider the bandwidth $\bar{h} = (h_1, ..., h_d)$ with $h_j \in \mathbf{R}^+$, $j = 1, ..., d$, and the kernel estimate of the form

$$\hat{f}_{n,\bar{h}}(x) = \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left\{ \left(\hat{D}_n \sum_{r=1}^{s} \frac{n_r w_r(X_{ij})}{\hat{V}_{n_r}} \right)^{-1} v_{\bar{h}}^{-1} K\left(\frac{x - X_{ij}}{\bar{h}} \right) \right\},$$

where $v_{\bar{h}} = \prod_{j=1}^{d} h_j$ and $x/\bar{h} = (x_1/h_1, ..., x_d/h_d)$. For simplicity, we only discuss here the asymptotic properties of $\hat{f}_n$ as defined in (10). With some more complicated notation, the techniques of this paper can also be extended to that of $\hat{f}_{n,\bar{h}}$.

3. ASYMPTOTIC RISKS

The aim of this section is to study the asymptotic properties of the integrated squared error (ISE) and the mean integrated squared error (MISE) of $\hat{f}_n$ defined by

$$\mathrm{ISE}(\hat{f}_n) = \int_{\mathbf{S}} (\hat{f}_n(x) - f(x))^2\, \pi(x)\, dx$$

and

$$\mathrm{MISE}(\hat{f}_n) = E \int_{\mathbf{S}} (\hat{f}_n(x) - f(x))^2 \pi(x)\, dx,$$

respectively, where $\pi(x)$ is a bounded and known nonnegative weight function. For simplicity, the support of π is assumed to be a compact subset of $\mathbf{S}$. This eliminates the complication of boundary effects which are well known in the i.i.d. direct samples case. The results here are essential for the development of the cross-validation criterion (Section 4). Our approach here is to first consider a "pseudo-estimate" f_n^* of f by replacing the random jumps of (10) with some deterministic quantities, and then to show that $\hat{f}_n$ is asymptotically equivalent to f_n^* in terms of ISE and MISE.

Intuitively, $\mathrm{ISE}(\hat{f}_n)$ measures the global L_2 risk of $\hat{f}_n$ for each given data set, while $\mathrm{MISE}(\hat{f}_n)$ measures the average effect of the global L_2 risk of $\hat{f}_n$. In density estimation with i.i.d. direct samples, both ISE and MISE defined above have been extensively studied in the literature, for example, Marron (1985, 1987), Marron and Härdle (1986), among others. But the cross-validation bandwidths have been mostly developed using ISE as a compelling risk criterion (cf. Stone, 1984, and Marron, 1985, 1987). Here we adopt the same framework and extend the approach of ISE and MISE to the more general biased sampling case. Analyses using other types of errors, such as the L_1 risks, in the i.i.d. direct sample case, can be found in Devroye and Györfi (1985), Devroye (1994), Fan and Hall (1994), among others.

By the obvious identity

$$\int_{\mathbf{S}} \left(\sum_{i=1}^{s} \frac{\lambda_{n_i} w_i(t)}{W_i} \right)^{-1} d\bar{G}_n(t) = 1, \tag{11}$$

we define $f_n^*(x)$ to be a "pseudo-estimate" of $f(x)$ such that

$$f_n^*(x) = \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left\{ \frac{\xi_n(X_{ij})}{nh^d} K\left(\frac{x - X_{ij}}{h} \right) \right\}, \tag{12}$$

where

$$\xi_n(t) = \left(\sum_{i=1}^{s} \frac{\lambda_{n_i} w_i(t)}{W_i} \right)^{-1}. \tag{13}$$

Notice here that f_n^* is different from f_n as defined in (5) since the jumps used in (5) still depend on the empirical measure $\hat{G}_n$, while the jumps used in (12) are totally deterministic.

Parallel to the usual conditions in the i.i.d. direct samples (cf. Marron, 1985), we also assume the following conditions:

A3: (a) the underlying density f is uniformly bounded and Hölder continuous on its support $\mathbf{S}$, that is, there exist constants $C_1 > 0$ and $0 < \alpha \leqslant 1$ so that

$$|f(x) - f(y)| \leqslant C_1 |x - y|^{\alpha} \qquad \text{for all} \quad x, y \in \mathbf{S},$$

where $|x| = (\sum_{i=1}^{d} x_i^2)^{1/2}$;

(b) the bandwidth h satisfies $\lim_{n \to \infty} h = 0$ and $\lim_{n \to \infty} nh^d = \infty$;

(c) the kernel function K has compact support on $\mathbf{R}^d$, satisfies $\int K(u)\, du = 1$ and is Hölder continuous in the sense that there exist constants $C_2 > 0$ and $0 < \beta \leqslant 1$ so that

$$|K(x) - K(y)| \leqslant C_2 |x - y|^{\beta} \qquad \text{for all} \quad x, y \in \mathbf{R}^d;$$

(d) there exist constants $0 < \lambda_i < 1$ so that $\sum_{i=1}^{s} \lambda_i = 1$ and $\lambda_{n_i} \to \lambda_i$ as $n \to \infty$ for all $i = 1, ..., s$;

(e) there exists a function $\phi: \mathbf{R}^d \to \mathbf{R}$ such that

$$\int_{\mathbf{R}^d} \xi_n(x - hu)\, K^2(u)\, du \mapsto \phi(x) \qquad \text{as} \quad n \to \infty,$$

where $x - hu = (x_1 - hu_1, ..., x_d - hu_d)$.

The last condition A3(e) is set to cover a general enough class of interesting situations. For the special case of $d = 1$, A3(e) is automatically satisfied when each w_i has only finite number of discontinuity points on its support. For the general case of $\mathbf{R}^d$, A3(e) is satisfied if, for each w_i, there are finite number of subsets $\mathbf{S}_{i1}, ..., \mathbf{S}_{iJ}$, $0 < J < \infty$, such that $\mathbf{S} = \bigcup_{j=1}^{J} \mathbf{S}_{ij}$ and w_i is continuous on each $\mathbf{S}_{ij}$, $1 \leqslant j \leqslant J$. In Wu and Mao (1996), it has been shown that, if our goal is to estimate the density value $f(x)$ at the point x when $s = d = 1$ and the corresponding weight function w is discontinuous at x, then a discontinuous kernel function should be used in order to obtain some asymptotic minimax properties for $\hat{f}_n(x)$. So far, no result in the literature has shown that whether discontinuous kernels are still asymptotically superior than continuous kernels in the minimax sense when a global measure, such as ISE or MISE, is used. Thus the Hölder continuity condition of A3(c) is assumed only because it is a natural condition for the i.i.d. direct samples and is technically convenient for the discussion of the next section.

Let $B(\hat{f}, f(x))$ and $V(\hat{f}, f(x))$ be the bias and variance of any estimate $\hat{f}(x)$ of $f(x)$, respectively. By the well-known variance-bias squares decomposition,

$$\mathrm{MISE}(f_n^*) = \int_S B^2(f_n^*, f(x))\,\pi(x)\,dx + \int_S V(f_n^*, f(x))\,\pi(x)\,dx, \tag{14}$$

where, by direct computation and A1 through A3,

$$B(f_n^*, f(x)) = \int_S K(u)(f(x-hu) - f(x))\,du, \tag{15}$$

$$V(f_n^*, f(x)) = n^{-1}h^{-d}f(x)\,\phi(x) + o(n^{-1}h^{-d}), \tag{16}$$

and $x - hu = (x_1 - hu_1, ..., x_d - hu_d)$.

To compare the risks between $\hat{f}_n$ and f_n^*, a simple decomposition shows that

$$\mathrm{ISE}(\hat{f}_n) = \mathrm{ISE}(f_n^*) + 2\mathrm{II}(h) + \mathrm{III}(h) \tag{17}$$

and

$$\mathrm{MISE}(\hat{f}_n) = \mathrm{MISE}(f_n^*) + 2E[\mathrm{II}(h)] + E[\mathrm{III}(h)], \tag{18}$$

where

$$\mathrm{II}(h) = \int_S (\hat{f}_n(x) - f_n^*(x))(f_n^*(x) - f(x))\,\pi(x)\,dx, \tag{19}$$

$$\mathrm{III}(h) = \int_S (\hat{f}_n(x) - f_n^*(x))^2\,\pi(x)\,dx. \tag{20}$$

We now give the main results of this section.

THEOREM 3.1. *Suppose that* $h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]$ *for some* $0 < \varepsilon < \frac{1}{2}$ *and A1 through A3 are satisfied. Then, as* $n \to \infty$,

$$\sup_{h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]} \left| \frac{\mathrm{ISE}(\hat{f}_n) - \mathrm{MISE}(f_n^*)}{\mathrm{MISE}(f_n^*)} \right| \to 0 \quad \textit{with probability one.} \tag{21}$$

Furthermore, replacing $\mathrm{ISE}(\hat{f}_n)$ *by* $\mathrm{MISE}(\hat{f}_n)$, *it follows that*

$$\sup_{h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]} \left| \frac{\mathrm{MISE}(\hat{f}_n) - \mathrm{MISE}(f_n^*)}{\mathrm{MISE}(f_n^*)} \right| \to 0 \quad \textit{as} \quad n \to \infty. \tag{22}$$

Proof. By A1 through A3 and the Cauchy–Schwartz inequality, (21) and (22) follow from the next two technical lemmas whose proofs are deferred to Section 6. ▮

LEMMA 3.1. *If the assumptions of Theorem* 3.1 *are satisfied, then*

$$\sup_{h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]} \left| \frac{\mathrm{ISE}(f_n^*) - \mathrm{MISE}(f_n^*)}{\mathrm{MISE}(f_n^*)} \right| \to 0 \qquad \text{with probability one as } n \to \infty. \tag{23}$$

LEMMA 3.2. *If the assumptions of Theorem* 3.1 *are satisfied, then*

$$\sup_{h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]} n^{\delta} \mathrm{III}(h) \to 0 \qquad \text{with probability one as } n \to \infty \tag{24}$$

and

$$\sup_{h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]} n^{\delta} E[\mathrm{III}(h)] \to 0 \qquad \text{as} \quad n \to \infty \tag{25}$$

hold for any constant δ *with* $0 < \delta < \frac{1}{2}$.

Similar to the i.i.d. direct samples case, the conclusions of Theorem 3.1 are uniform over $h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]$. This restriction should not bring any difficulty in practice since one has to select h^d within the interval $[n^{-1+\varepsilon}, n^{-\varepsilon}]$ so that $\hat{f}_n$ has the desired rate of convergence.

Remark 3.1. It immediately follows from A1 through A3 and (14) through (16) that $\hat{f}_n$ is consistent under both integrated squared error and mean integrated squared error in the sense that $\sup_{h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]} \mathrm{ISE}(\hat{f}_n) \to 0$ with probability one as $n \to \infty$ and $\sup_{h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]} \mathrm{MISE}(\hat{f}_n) \to 0$ as $n \to \infty$. In fact, it is a direct consequence of Stone (1980, 1982) and Bretagnolle and Huber (1979) that the optimal convergence rate of $\mathrm{ISE}(\hat{f}_n)$ is $n^{-2\alpha/(2\alpha+d)}$ which is exactly the same rate as in the i.i.d. direct samples case, and this rate is achieved by taking $h = n^{-1/(2\alpha+d)} h_0$ for some constant $h_0 > 0$.

Remark 3.2. Similar to the i.i.d. direct samples case, higher order kernels should also be used in order to obtain the best possible convergence rates when the underlying density is assumed to have more than 2 times derivatives. Specifically, if f is k times differentiable with $k > 2$ and the kth derivative $f^{(k)}$ is continuous, then the optimal rate of $\mathrm{ISE}(\hat{f}_n)$ is $n^{-2k/(2k+d)}$ in the sense of Stone (1980, 1982), and this rate is achieved by taking K such that $\int K(u)\, du = 1$, $\int u^j K(u)\, du = 0$ for all $1 \leqslant |j| \leqslant k-1$, where $j = (j_1, ..., j_d)$, $|j| = j_1 + \cdots + j_d$ and $u^j = u_1^{j_1} \cdots u_d^{j_d}$, and $h = n^{-1/(2k+d)} h_0$ for some constant $h_0 > 0$.

Remark 3.3. The special feature which distinguishes the biased sampling model (1) from the i.i.d. direct sampling model enters into the analysis only through the variance term given in (16). Denote

$$\xi(x) = \left(\sum_{i=1}^{s} \frac{\lambda_i w_i(x)}{W_i} \right)^{-1}. \tag{26}$$

If w_i are continuous functions on the support $\mathbf{S}$, then

$$V(f_n^*, f(x)) = n^{-1}h^{-d} \left(\int K^2(u)\, du \right) \left(\int_{\mathbf{S}} \xi(x)\, f(x)\, \pi(x)\, dx \right) + o(n^{-1}h^{-1}),$$

which differs from the variance of the usual kernel density estimate with i.i.d. direct data only through a factor $\xi(x)$. If w_i have discontinuity points, the explicit form of $\phi(x)$, or equivalently $V(f_n^*, f(x))$, depends on the particular choices of w_i.

4. CROSS-VALIDATION BANDWIDTH SELECTION

The effort here is to establish a cross-validation criterion for $\hat{f}_n$ and then show that, by minimizing this criterion, we can obtain an asymptotically optimal data-driven bandwidth. Define h to be an ideal bandwidth if it minimizes $\mathrm{ISE}(\hat{f}_n)$. An easy expansion shows that

$$\mathrm{ISE}(\hat{f}_{n,h}) = \int_{\mathbf{S}} \hat{f}_{n,h}^2(x)\, \pi(x)\, dx - 2 \int_{\mathbf{S}} \hat{f}_{n,h}(x)\, f(x)\, \pi(x)\, dx + \int_{\mathbf{S}} f^2(x)\, \pi(x)\, dx,$$

where $\hat{f}_{n,h}$ is used throughout this section to denote the kernel estimate of (10) when a specific bandwidth h is used.

Since the third term of $\mathrm{ISE}(\hat{f}_{n,h})$ does not depend on h, it suffices to minimize the first two terms of $\mathrm{ISE}(\hat{f}_{n,h})$ with respect to h. To estimate the second term of $\mathrm{ISE}(\hat{f}_{n,h})$, we first define a "leave-one-out" version of the kernel NPMLE density estimate of f,

$$\hat{f}_{-(i,j),h}(x) = (n-1)^{-1} \sum_{(i',j') \neq (i,j)} \times \left\{ \left(\hat{D}_n \sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(X_{i'j'})}{\hat{V}_{n_r}} \right)^{-1} h^{-d} K\left(\frac{x - X_{i'j'}}{h} \right) \right\}, \tag{27}$$

and then estimate $\int_{\mathbf{S}} \hat{f}_{n,h}(x)\, f(x)\, \pi(x)\, dx$ by

$$\frac{1}{n} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \hat{f}_{-(i,j),h}(X_{ij})\, \pi(X_{ij})\, \frac{\hat{W}_{n_i}}{w_i(X_{ij})}.$$

This leads to the biased sampling cross-validation criterion,

$$\mathrm{CV}(h) = \int_{\mathbf{S}} \hat{f}^2_{n,h}(x)\,\pi(x)\,dx - \frac{2}{n}\sum_{i=1}^{s}\sum_{j=1}^{n_i} \hat{f}_{-(i,j),h}(X_{ij})\,\pi(X_{ij})\,\frac{\hat{W}_{n_i}}{w_i(X_{ij})}, \qquad (28)$$

which is an estimate the first two terms of $\mathrm{ISE}(\hat{f}_{n,h})$. The cross-validation bandwidth $\hat{h}_c$ is then defined to be the minimizer of $\mathrm{CV}(h)$ over the interval $[n^{(-1+\varepsilon)/d}, n^{-\varepsilon/d}]$, that is, $\mathrm{CV}(\hat{h}_c) = \inf_{h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]} \mathrm{CV}(h)$.

We now give the main result of this section.

THEOREM 4.1; *If the assumptions A*1 *through A*3 *are satisfied, then* $\hat{h}_c$ *is asymptotically optimal in the sense that*

$$\frac{\mathrm{ISE}(\hat{f}_{n,\hat{h}_c})}{\inf_{h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]} \mathrm{ISE}(\hat{f}_{n,h})} \to 1 \quad \textit{with probability one as } n \to \infty. \qquad (29)$$

Remark 4.1. An alternative viewpoint is to define an ideal bandwidth $h_0(f)$ such that it minimizes the mean integrated squared error $\mathrm{MISE}(\hat{f}_{n,h})$. For i.i.d. direct samples, it has been shown that the least-squares cross-validation bandwidth $\hat{h}_c$ converges to $h_0(f)$ in probability with the very slow rate $n^{-1/10}$, and this rate can be improved by some alternative bandwidth selection methods (cf. Hall, Sheather, Jones, and Marron, 1991, or Fan and Marron, 1992). However, these methods suffer some drawbacks by requiring extra smoothness conditions on f and K. Obviously further study is worthwhile to determine whether these alternative methods can be extended to the current biased sampling model.

5. SIMULATION AND COMPARISON WITH OTHER ESTIMATES

In this section, we first compare the theoretical properties of $\hat{f}_n(x)$ with another class of kernel estimates of $f(x)$, and then investigate the finite sample performance of $\hat{f}_n(x)$ and the cross-validation bandwidth $\hat{h}_c$ through a Monte Carlo simulation study based on stratified samples. The main features of this section indicate that: (i) the kernel estimate $\hat{f}_n$ has superior statistical properties than many other intuitive kernel estimates of f; (ii) the cross-validation bandwidths give adequate kernel estimates in practice with moderate (a few hundred) to large sample sizes.

5.1. *Comparison with Other Kernel Estimates*

In many special situations, f can also be estimated by kernel estimates other than (10). By (1), if $w_i(x) > 0$, then $f(x) = W_i g_i(x)/w_i(x)$ is uniquely

defined at $x \in \mathbf{R}^s$. it is easy to verify that, if $w_i(x) > 0$ for all $i = 1, ..., s$ and $p_1, ..., p_s$ are any nonnegative weights which may depend on x and satisfy $\sum_{i=1}^{s} p_i = 1$, then $f(x) = \sum_{i=1}^{s} p_i W_i g_i(x)/w_i(x)$.

Suppose that the support of f is contained in the set $\{x: w_i(x) > 0 \text{ for all } 1 \leqslant i \leqslant s\}$. Then $f(x)$ can also be estimated by a linear combination of kernel estimates,

$$\tilde{f}_p(x) = \sum_{i=1}^{s} p_i \tilde{f}_{(i)}(x),$$

where

$$\tilde{f}_{(i)}(x) = (n_i \hat{v}_i)^{-1} h^{-d} \sum_{j=1}^{n_i} w_i^{-1}(X_{ij}) K\left(\frac{x - X_{ij}}{h}\right)$$

and $\hat{v}_i = n_i^{-1} \sum_{j=1}^{n_i} w_i^{-1}(X_{ij})$ are natural estimates of $f(x)$ and W_i^{-1}, respectively, based on the ith sample. Here, one may select $p_i = \lambda_{n_i}$ or any other nonnegative weights $p_1, ..., p_s$.

By the similar calculations as in Section 3, we can show that

$$\text{MISE}(\tilde{f}_p) = \int_{\mathbf{S}} \text{B}^2(\tilde{f}_p, f(x))\, \pi(x)\, dx + \int_{\mathbf{S}} \text{V}(\tilde{f}_p, f(x))\, \pi(x)\, dx,$$

where $\text{B}(\tilde{f}_p, f(x)) = \int K(u)\, (f(x - hu) - f(x))\, du$ and

$$\text{V}(\tilde{f}_p, f(x)) = n^{-1} h^{-d} \int \sum_{i=1}^{s} \left(\frac{p_i^2 W_i}{\lambda_i w_i(x - hu)}\right) \times K^2(u)\, f(x - hu)\, du + o(n^{-1} h^{-d}).$$

For the bias term, $\text{B}(\tilde{f}_p, f(x))$ is exactly the same as $\text{B}(f_n^*, f(x))$ of (15). But, by Jensen's inequality, we have

$$\sum_{i=1}^{s} p_i \left(\frac{p_i W_i}{\lambda_i w_i(t)}\right) \geqslant \left(\sum_{i=1}^{s} \frac{\lambda_i w_i(t)}{W_i}\right)^{-1} \qquad \text{for all} \quad p_1, ..., p_s,$$

and the equality sign may hold only for the trivial cases when $w_i(t)$ are constants. It is straightforward to verify from (16) that, when n is sufficiently large,

$$\text{V}(\tilde{f}_p f(x)) - \text{V}(f_n^*, f(x)) \geqslant 0 \qquad \text{for all} \quad p_1, ..., p_s.$$

Thus, in terms of MISE, $\hat{f}_n$ is asymptotically superior to $\tilde{f}_p$ for any selection of $p_1, ..., p_s$, and the two estimates may be equivalent only for the most trivial cases.

In some other sampling schemes, the support of f is not necessarily contained in the set $\{x: w_i(x) > 0 \text{ for all } 1 \leqslant i \leqslant s\}$, so that $\tilde{f}_p$ can not be applied directly. Without loss of generality, suppose that there exists a positive integer $s^* \leqslant s$ such that $w_i(x) > 0$ for $1 \leqslant i \leqslant s^*$, but $w_i(x) = 0$ for $s^* + 1 \leqslant i \leqslant s$. In order to ensure the identifiability of f, we assume further that $W_1, ..., W_{s^*}$ are known. These kind of samples can be found, for example, in Jewell (1985) and Jewell and Quesenberry (1986). Then, since the last $s - s^*$ samples yield no information about $f(x)$, $f(x)$ can be analogously estimated using the first s^* samples such that

$$\bar{f}_p(x) = \sum_{i=1}^{s^*} p_i \left(n_i^{-1} h^{-d} W_i \sum_{j=1}^{n_i} w_i^{-1}(X_{ij}) K\left(\frac{x - X_{ij}}{h}\right)\right),$$

where $p_1, ..., p_{s^*}$ are nonnegative weights satisfying $\sum_{i=1}^{s^*} p_i = 1$. By similar calculations as that for $\tilde{f}_p$, we can show that $\mathrm{MISE}(\bar{f}_p)$ is asymptotically the same as $\mathrm{MISE}(\tilde{f}_p)$ with s replaced by s^*. Thus $\hat{f}_n$ is asymptotically superior or equivalent to $\bar{f}_p$ for all possible choices of $p_1, ..., p_s$ with equivalence only holds for the same trivial cases as those for $\tilde{f}_p$.

5.2. *Simulation with Stratified Samples*

As a special case of (1), we consider the following stratified samples: Let $X_{i1}, ..., X_{in_i}$ be i.i.d. on the real line having densities g_i, $i = 1, ..., s+1$, satisfying (1) with $w_i(t) = 1_{[t \in D_i]}$ for $i = 1, ..., s$, and $w_{s+1}(t) = 1$, where $D_1, ..., D_s$ is a partition on $[-\infty, \infty]$ such that $D_i \cap D_j = \varnothing$ if $i \neq j$.

By A1 and A2, it is straightforward to verify that f is identifiable, and on the set of $y \in D_k$, $1 \leqslant k \leqslant s$,

$$\sum_{j=1}^{s} \frac{\lambda_{nj} 1_{[y \in Dj]}}{u_j} = \frac{\lambda_{nk} 1_{[y \in D_k]}}{u_k}.$$

Consequently, (6) is reduced to

$$(\lambda_{n_i} + \lambda_{n_{s+1}} \hat{V}_{n_i})^{-1} \hat{D}_n(D_i) = 1,$$

where $\hat{G}_n(D_i) = \int_{D_i} d\hat{G}_n(y) = \lambda_{n_i} + \lambda_{n_{s+1}} \hat{G}_{n_{s+1}}(D_i)$. Thus, $\hat{V}_{n_i} = \hat{G}_{n_{s+1}}(D_i)$ for $i = 1, ..., s$, and (10) is reduced to

$$\hat{f}_n(x) = \hat{D}_n^{-1} \sum_{i=1}^{s+1} \sum_{j=1}^{n_i} \left[\left(\sum_{r=1}^{s} \frac{n_r 1_{[X_{in} \in D_r]}}{\hat{G}_{n_{s+1}}(D_r)} + n_{s+1} \right)^{-1} h^{-1} K\left(\frac{x - X_{ij}}{h}\right)\right],$$

where

$$\hat{D}_n = \sum_{i=1}^{s+1} \sum_{j=1}^{n_i} \left(\sum_{r=1}^{s} \frac{n_r 1_{[X_{ij} \in D_r]}}{\hat{G}_{n_{s+1}}(D_r)} + n_{s+1} \right)^{-1}.$$

If $W_1, ..., W_s$ were known, then $f(x)$ can also be estimated by $\bar{f}_p(x)$. Since $w_j(x) = 0$ if $x \in D_i$ for all $j \neq i$ and $j \neq s+1$, only the ith and the $(s+1)$th the samples can be used in $\bar{f}_p(x)$ for the estimation of $f(x)$ when $x \in D_i$. So that a natural choice of $(p_1, ..., p_s)$ could be $p_i = n_i/(n_i + n_{s+1})$, $p_{s+1} = 1 - p_i$, and $p_j = 0$ for $j \neq i$ and $j \neq s+1$ when $x \in D_i$ and $i = 1, ..., s$.

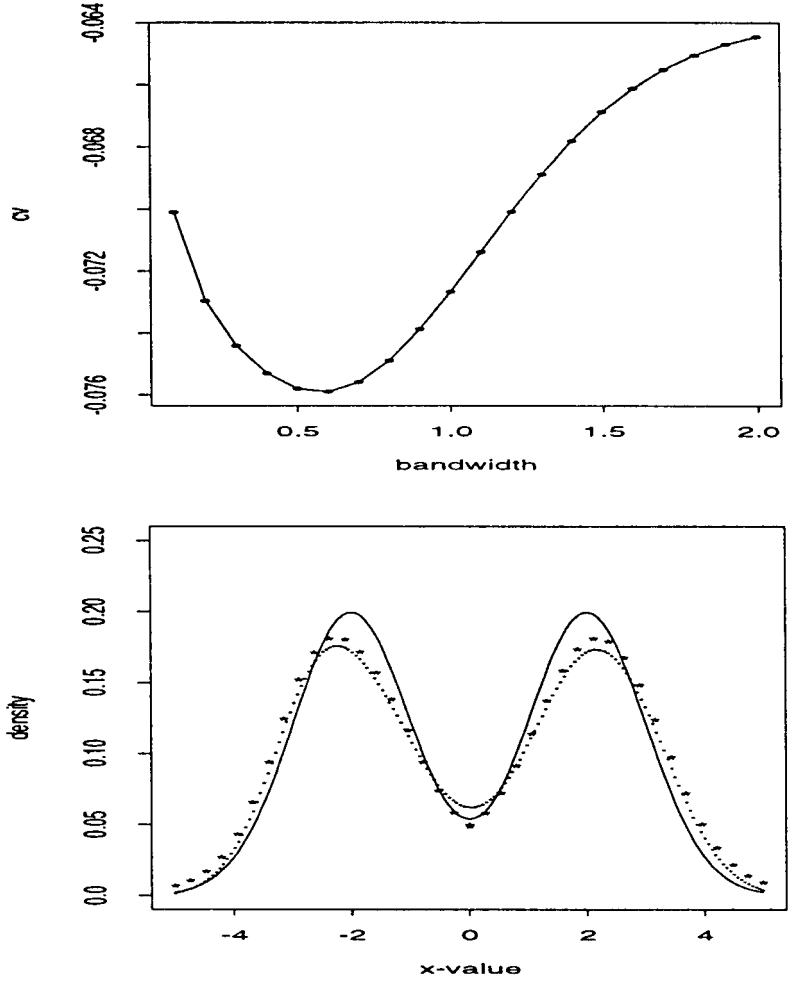

FIG. 1. (a) (Top): CV(h) versus h. (b) (Bottom): Comparison of the real and estimated density values when the cross-validation bandwidth 0.6 was used: (i) the solid curve gives the real density of the mixture normal distribution f; (ii) the dotted curve represents the estimated density values obtained from $\hat{f}_n$; (iii) "$*$" represents the selected estimated density values obtained from $\bar{f}_p$.

For the simulation results, we considered the estimation of a mixture density $f(x) = \frac{1}{2}\phi(x; -2, 1) + \frac{1}{2}\phi(x; 2, 1)$ where $\phi(x; \mu, \sigma^2)$ is the density value at x of a normal distribution with mean μ and variance σ^2. Here $\pi(x) \equiv 1$ and three independent samples of sizes $n_1 = n_2 = 150$ and $n_3 = 300$ were randomly generated based on (1) and $w_i(t) = 1_{[t \in D_i]}$, $i = 1, 2, 3$, where $D_1 = [-\infty, 0)$, $D_2 = [0, \infty]$ and $D_3 = [-\infty, \infty]$. For both $\hat{f}_n$ and $\bar{f}_p$, the kernels were chosen to be the standard Gaussian density. For the construction of $\bar{f}_p$, $W_1 = W_2 = \frac{1}{2}$, $W_3 = 1$ and (p_1, p_2, p_3) was chosen to be $(\frac{1}{3}, 0, \frac{2}{3})$ if $x < 0$ and $(0, \frac{1}{3}, \frac{2}{3})$ if $x \geqslant 0$.

Figure 1a shows the relationship of CV(h) versus h when h is within the range of $[0.1, 2.0]$. It is easy to see that CV(h) reaches the minimum when h is around 0.6 which was taken to be the cross-validation bandwidth. Figure 1b gives the actual mixture density $f(x)$ (solid curve), the estimated density values obtained from $\hat{f}_n$ (dotted curve) and the estimated density values obtained from $\bar{f}_p$. The cross-validation bandwidth $\hat{h}_c$ was used for both $\hat{f}_n$ and $\bar{f}_n$. Although slightly oversmoothed in this particular simulation, the general pattern of Fig. 1b shows that $\hat{f}_{n, \hat{h}_n}$ gives reasonable estimated values. It is also interesting to see from Fig. 1b that, in general, $\bar{f}_p(x)$ stays very close to $\hat{f}_n(x)$, and in some regions, $\bar{f}_p(x)$ even gives slightly better estimates than $\hat{f}_n(x)$. Thus, despite the theoretical advantage of $\hat{f}_n$ over $\bar{f}_p$, $\bar{f}_p$ remains to be a very competitive estimate of f at least for this sampling scheme. Other simulations with different samples sizes and bandwidths revealed the similar patterns in the statistical behavior of $\hat{f}_n$ and $\bar{f}_p$.

6. PROOFS

In this section, we sketch the proofs of the main results of Section 3 and Section 4. Further details and tedious computations used in the proofs can be found in the technical report of Wu (1996b).

First, the following technical lemma is useful to establish Lemma 3.1 and Lemma 3.2.

LEMMA 6.1. *Let ε be a constant satisfying $0 < \varepsilon < \frac{1}{2}$. If assumptions A1 through A3 are satisfied, then there exist constants c_1 and c_2 with $0 < c_1 < c_2 < \infty$ such that*

(a) $c_1 \leqslant V_r \leqslant c_2$ *for all* $r = 1, ..., s$, *and* $c_1 \leqslant \hat{V}_{n_r} \leqslant c_2$ *with probability one for all* $r = 1, ..., s$ *and sufficiently large* n;

(b) $c_1 \leqslant D_n \leqslant c_2$ *and* $c_1 \leqslant \hat{D}_n \leqslant c_2$ *with probability one for all sufficiently large* n;

(c) $n^{\varepsilon}(\hat{V}_{n_r} - V_r) \to 0$ *and* $n^{\varepsilon}(\hat{W}_{n_r} - W_r) \to 0$ *for all* $r = 1, ..., s$ *with probability one as* $n \to \infty$;

(d) $n^{\varepsilon}(W_s^{-1} - \hat{D}_n) \to 0$ *with probability one as* $n \to \infty$.

Proof. (a) By the definitions of W_i and V_i, $i = 1, ..., s$, it is obvious that the assertion for V_r holds. By Proposition 2.1 of Gill, Vardi, and Wellner (1988), we know that the unique $(\hat{V}_{n_1}, ..., \hat{V}_{n_s})$ exists almost surely and that $\hat{V}_{n_r} \to V_r$ with probability one as $n \to \infty$. This certainly implies that $c_1 \leqslant \hat{V}_{n_r} \leqslant c_2$ with probability one for sufficiently large n.

(b) Recall from (9) that $\hat{D}_n = \hat{W}_{n_s}$. Then Proposition 2.1 of Gill, Vardi, and Wellner (1988) shows that $\hat{W}_{n_s} \to W_s$ with probability one as $n \to \infty$. Since $c_1 \leqslant W_s \leqslant c_2$ for some positive c_1 and c_2, we know that the assertion for $\hat{D}_n$ holds.

By the definition given in Section 2, D_n can be written as

$$D_n = W_s^{-1} n^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(X_{ij})}{W_r} \right)^{-1}.$$

Thus, the almost sure boundness of $\hat{W}_{n_s}$, W_s, and $\hat{D}_n$ immediately shows that $c_1 \leqslant D_n \leqslant c_2$ almost surely for some positive c_1, c_2 and sufficiently large n.

(c) The assertions here are direct consequences of Proposition 2.2 and Proposition 2.3 of Gill, Vardi, and Wellner (1988).

(d) Since

$$n^{\varepsilon}(W_s^{-1} - \hat{D}_n) = \frac{n^{\varepsilon}(\hat{W}_{n_s} - W_s)}{\hat{W}_{n_s} W_s},$$

it immediately follows from Proposition 2.2 that $n^{\varepsilon}(\hat{W}_{n_s} - W_s) \to 0$ with probability one as $n \to \infty$. Hence, the assertion follows since $(\hat{W}_{n_s}, W_s)$ is bounded away from 0 almost surely. ∎

Proof of Lemma 3.1. The proof here is similar to the method used in the proof of Theorem 1 of Marron and Härdle (1986), with some slight modifications. To avoid repetition, we only sketch the main steps and refer to Marron and Härdle (1986) for the related computations.

Let H_n be any sequence of finite sets whose cardinality increases at most algebraically fast, that is,

$$\#(H_n) \leqslant \mathscr{C} n^{\rho} \qquad \text{for some constants } \mathscr{C} > 0 \text{ and } \rho > 0. \tag{30}$$

By the Hölder continuity condition A3(c) and a straightforward continuity argument (cf. Härdle and Marron, 1985; Marron, 1985; and Marron and Härdle, 1986), it suffices to prove (23) by showing

$$\sup_{h^d \in H_n} \left| \frac{\mathrm{ISE}(f_n^*) - \mathrm{MISE}(f_n^*)}{\mathrm{MISE}(f_n^*)} \right| \to 0 \qquad \text{with probability one as } n \to \infty. \quad (31)$$

Now (30) and the Chebyshev inequality imply that, for any $\delta > 0$ and $k = 1, 2, \ldots$,

$$P\left[\sup_{h^d \in H_n} \left| \frac{\mathrm{ISE}(f_n^*) - \mathrm{MISE}(f_n^*)}{\mathrm{MISE}(f_n^*)} \right| > \delta \right] \leqslant \mathscr{C} n^{\rho} \delta^{-2k} \sup_{h^d \in H_n} E\left[\frac{\mathrm{ISE}(f_n^*) - \mathrm{MISE}(f_n^*)}{\mathrm{MISE}(f_n^*)} \right]^{2k}.$$

It can be shown by the same method as in the proofs of (6.3) and (6.4) of Marron and Härdle (1986) that there is a constant $\gamma > 0$, so that for $k = 1, 2, \ldots$, there are constants $\mathscr{C}_k$ and

$$E\left[\frac{\mathrm{ISE}(f_n^*) - \mathrm{MISE}(f_n^*)}{\mathrm{MISE}(f_n^*)} \right]^{2k} \leqslant \mathscr{C}_k n^{-\gamma k}. \quad (32)$$

Then (32) and the Borel–Cantelli lemma imply that (31) holds. ∎

Proof of Lemma 3.2. First notice that, by (10) and (12),

$$|\hat{f}_n(x) - f_n^*(x)| \leqslant \sup_{t \in \mathbf{S}^*} |\Delta_n(t)| \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left| \frac{1}{nh^d} K\left(\frac{x - X_{ij}}{h} \right) \right|, \quad (33)$$

where

$$\Delta_n(t) = \left(\hat{D}_n \sum_{r=1}^{s} \frac{n_r w_r(t)}{\hat{V}_{n_r}} \right)^{-1} - \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(t)}{W_r} \right)^{-1}$$

$$\leqslant \left[\hat{D}_n W_s^{-1} \sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(t)}{\hat{V}_{n_r}} \right]^{-1} (\hat{D}_n^{-1} - W_s)$$

$$+ \frac{W_s \sum_{r=1}^{s} \left[\left(\frac{\lambda_{n_r} w_r(t)}{\hat{V}_{n_r} V_r} \right) (\hat{V}_{n_r} - V_r) \right]}{\left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(t)}{\hat{V}_{n_r}} \right) \left(\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(t)}{V_r} \right)}.$$

Applying Lemma 6.1 repeatedly, we have, for all $0<\varepsilon<\frac{1}{2}$,

$$\sup_{t\in \mathbf{S}^*} n^{\varepsilon}\,|\Delta_n(t)|\to 0 \qquad \text{with probability one as } n\to\infty. \tag{34}$$

Extending the method in the proof of Theorem 1 of Marron and Härdle (1986) to the current sampling case, we can show that

$$\sup_{h^d\in H_n}\int\left[\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left|\frac{1}{n_i h^d}K\left(\frac{x-X_{ij}}{h}\right)\right|\right]^2\pi(x)\,dx<c \qquad \text{with probability one,} \tag{35}$$

for some $c>0$ and all sufficiently large n. Thus (33) through (35) imply that (24) holds, while (25) is an easy consequence of (24). ∎

The following technical lemma is an essential component in the proof of Theorem 4.1.

LEMMA 6.2. *If the conditions of Theorem* 4.1 *hold, then*

$$\sup_{h^d\in[n^{-1+\varepsilon},\,n^{-\varepsilon}]}\left|\frac{\hat U_n-U_n}{\mathrm{MISE}(\hat f_{n,h})}\right|\to 0 \qquad \textit{with probability one as } n\to\infty, \tag{36}$$

where

$$\hat U_n=\frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{n_i}\hat f_{-(i,j),h}(X_{ij})\,\pi(X_{ij})\,\frac{\hat W_{n_i}}{w_i(X_{ij})}-\int_{\mathbf{S}}\hat f_{n,h}(x)\,f(x)\,\pi(x)\,dx$$

and

$$U_n=\frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{n_i}f(X_{ij})\,\pi(X_{ij})\,\frac{W_i}{w_i(X_{ij})}-\int_{\mathbf{S}}f^2(x)\,\pi(x)\,dx.$$

Proof. We first write $\delta_h(x,y)=h^{-d}K((x-y)/h)$, and observe that

$$\begin{aligned}\hat U_n=\frac{1}{n(n-1)}\sum_{i=1}^{s}\sum_{j=1}^{n_i}\sum_{(i',j')\neq(i,j)}\Bigg[&\left(\hat D_n\sum_{r=1}^{s}\frac{\lambda_{n_r}w_r(X_{i'j'})}{\hat V_{n_r}}\right)^{-1}\delta_h(X_{ij},X_{i'j'})\\ &\times\frac{\pi(X_{ij})\,\hat W_{n_i}}{w_i(X_{ij})}-\left(\hat D_n\sum_{r=1}^{s}\frac{\lambda_{n_r}w_r(X_{i'j'})}{\hat V_{n_r}}\right)^{-1}\\ &\times\int_{\mathbf{S}}\delta_h(x,X_{i'j'})\,f(x)\,\pi(x)\,dx\Bigg],\end{aligned}$$

$$U_n=\frac{1}{n(n-1)}\sum_{i=1}^{s}\sum_{j=1}^{n_i}\sum_{(i',j')\neq(i,j)}\left[f(X_{ij})\,\frac{\pi(X_{ij})\,W_i}{w_i(X_{ij})}-\int_{\mathbf{S}}f^2(x)\,\pi(x)\,dx\right].$$

It is straightforward to verify that

$$\hat{U}_n - U_n = \frac{1}{n(n-1)} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \sum_{(i', j') \neq (i, j)} [Z_{(i, j), (i', j')} - Z'_{(i, j), (i', j')}],$$

where

$$\begin{aligned} Z_{(i, j), (i', j')} = & \left(D_n \sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(X_{i'j'})}{V_r} \right)^{-1} \delta_h(X_{ij}, X_{i'j'})\, \pi(X_{ij})\, \frac{W_i}{w_i(X_{ij})} \\ & - \left(D_n \sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(X_{i'j'})}{V_r} \right)^{-1} \int_{\mathbf{S}} \delta_h(x, X_{i'j'})\, f(x)\, \pi(x)\, dx \\ & - f(X_{ij})\, \pi(X_{ij})\, \frac{W_i}{w_i(X_{ij})} + \int_{\mathbf{S}} f^2(x)\, \pi(x)\, dx, \end{aligned}$$

$$\begin{aligned} Z'_{(i, j), (i', j')} = & \left[\left(\hat{D}_n \sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(X_{i'j'})}{\hat{V}_{n_r}} \right)^{-1} \hat{W}_{n_i} - \left(D_n \sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(X_{i'j'})}{V_r} \right)^{-1} W_i \right] \\ & \times [\delta_h(X_{ij}, X_{i'j'})\, \pi(X_{ij})\, w_i^{-1}(X_{ij})] \\ & - \left[\left(\hat{D}_n \sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(X_{i'j'})}{\hat{V}_{n_r}} \right)^{-1} - \left(D_n \sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(X_{i'j'})}{V_r} \right)^{-1} \right] \\ & \times \int_{\mathbf{S}} \delta_h(x, X_{i'j'})\, f(x)\, \pi(x)\, dx. \end{aligned}$$

Now it can be shown by Lemma 6.1 and a direct extension of the proofs of (7.2) and (7.4) of Marron (1987) and Section 6 of Marron and Härdle (1986) from direct samples to the current biased sampling case that the following limits hold with probability one as $n \to \infty$,

$$\sup_{h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]} \left| \frac{n^{-1}(n-1)^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \sum_{(i', j') \neq (i, j)} Z_{(i, j), (i', j')}}{\mathrm{MISE}(\hat{f}_{n, h})} \right| \to 0 \quad (37)$$

and

$$\sup_{h^d \in [n^{-1+\varepsilon}, n^{\varepsilon\varepsilon}]} \left| \frac{n^{-1}(n-1)^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \sum_{(i', j') \neq (i, j)} Z'_{(i, j), (i', j')}}{\mathrm{MISE}(\hat{f}_{n, h})} \right| \to 0. \quad (38)$$

See also Wu (1986b) for further details in the derivations of (37) and (38). Thus the proof is completed since (36) follows from (37) and (38).

Proof of Theorem 4.1. Let h_I be the minimizer of $\mathrm{ISE}(\hat{f}_{n,h})$, so that

$$\mathrm{ISE}(\hat{f}_{n,h_I}) = \inf_{h^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]} \mathrm{ISE}(\hat{f}_{n,h}).$$

As a direct consequence of the inequalities

$$\mathrm{ISE}(\hat{f}_{n,\hat{h}_c}) \geqslant \mathrm{ISE}(\hat{f}_{n,h_I}), \qquad \mathrm{CV}(h_I) \geqslant \mathrm{CV}(\hat{h}_c),$$

we have

$$\left|\frac{\mathrm{ISE}(\hat{f}_{n,\hat{h}_c}) - \mathrm{ISE}(\hat{f}_{n,h_I})}{\mathrm{ISE}(\hat{f}_{n,\hat{h}_c})}\right| \leqslant \left|\frac{\mathrm{ISE}(\hat{f}_{n,\hat{h}_c}) - \mathrm{ISE}(\hat{f}_{n,h_I}) + \mathrm{CV}(h_I) - \mathrm{CV}(\hat{h}_c)}{\mathrm{MISE}(\hat{f}_{n,\hat{h}_c}) + \mathrm{MISE}(\hat{f}_{n,h_I})}\right| \times \frac{\mathrm{MISE}(\hat{f}_{n,\hat{h}_c}) + \mathrm{MISE}(\hat{f}_{n,h_I})}{\mathrm{ISE}(\hat{f}_{n,\hat{h}_c})}.$$

Thus, by Theorem 3.1, it is enough to show that

$$\sup_{h_1^d, h_2^d \in [n^{-1+\varepsilon}, n^{-\varepsilon}]} \left|\frac{\mathrm{ISE}(\hat{f}_{n,h_1}) - \mathrm{CV}(h_1) - [\mathrm{ISE}(\hat{f}_{n,h_2}) - \mathrm{CV}(h_2)]}{\mathrm{MISE}(\hat{f}_{n,h_1}) + \mathrm{MISE}(\hat{f}_{n,h_2})}\right| \to 0 \quad (39)$$

with probability one as $n \to \infty$.

Now since

$$\mathrm{CV}(h) = \mathrm{ISE}(\hat{f}_{n,h}) - \int_{\mathbf{S}} f^2(x)\,\pi(x)\,dx - 2\left[\frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{n_i} \hat{f}_{-(i,j),h}(X_{ij})\,\pi(X_{ij})\,\frac{\hat{W}_{n_i}}{w_i(X_{ij})} - \int_{\mathbf{S}} \hat{f}_{n,h}(x)\,f(x)\,\pi(x)\,dx\right]$$

and, by the triangular array central limit theorem,

$$\lim_{n\to\infty} n^{\delta} U_n = 0 \qquad \text{with probability one for all } 0 < \delta < \tfrac{1}{2},$$

where U_n is defined in Lemma 6.2, (29) is then an easy consequence of Lemma 6.2. ∎

ACKNOWLEDGMENT

The author is grateful to a referee for careful reading of the original manuscript and many interesting suggestions which greatly improved the presentation of the paper and led to the inclusion of Section 5.

REFERENCES

Ahmad, I. A. (1995). On multivariate kernel estimation for samples from weighted distributions. *Statist. Probab. Lett.* **22** 121–129.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, Baltimore.

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

Bretagnolle, J., and Huber, C. (1979). Estimation des densities: Risque minimax. *Z. Wahrsch. Verw. Gebiete* **47** 119–137.

Cox, D. R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling* (N. L. Johnson and H. Smith, Jr., Eds.), pp. 506–527. Wiley–Interscience, New York.

Devroye, L. (1994). On good deterministic smoothing sequences for kernel density estimates. *Ann. Statist.* **22** 886–889.

Devroye, L., and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York.

Fan, J. Q., and Hall, P. (1994). On curve estimation by minimizing mean absolute deviation and its implication. *Ann. Statist.* **22** 867–885.

Fan, J. Q., and Marron, J. S. (1992). Best possible constant for bandwidth selection. *Ann. Statist.* **20** 2057–2070.

Gill, R. D., Vardi, Y., and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16** 1069–1112.

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

Hall, P., Sheater, S. J., Jones, M. C., and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78** 263–269.

Härdle, W., and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression. *Ann. Statist.* **13** 1465–1481.

Jewell, N. (1985). Regression from stratified samples of dependent variable. *Biometrika* **72** 11–21.

Jewell, N., and Quesenberry, C. P. (1986). Regression analysis based on stratified samples. *Biometrika* **73** 605–614.

Jones, M. C. (1991). Kernel density estimation for length biased data. *Biometrika* **78** 511–519.

Marron, J. S. (1985). An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Ann. Statist.* **13** 1011–1023.

Marron, J. S. (1987). A comparison of cross-validation techniques in density estimation. *Ann. Statist.* **15** 152–162.

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

Morgenthaler, S., and Vardi, Y. (1986). Ranked set samples: A nonparametric approach. *J. Econometrics* **32** 109–125.

Patil, G. P., Rao, C. R., and Zelen, M. (1988). Weighted distributions. In *Encyclopedia of Statistical Sciences*, Vol. 9 (S. Kotz and N. L. Johnson, Eds.), pp. 565–571. Wiley, New York.

Patil, G. P., and Taillie, C. (1989). Probing encountered data, meta analysis and weighted distribution methods. In *Statistical Data Analysis and Inference* (Y. Dodge, Ed.), pp. 317–346. North-Holland, Amsterdam.

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

Stone, C. J. (1980). Optimal convergence rates for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence of nonparametric regression. *Ann. Statist.* **10** 1040–1053.

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

Vardi, Y. (1985). Empirical distributions in selection bias models (with discussions). *Ann. Statist.* **13** 178–205.

Wu, C. O. (1996a). Kernel smoothing of the nonparametric maximum likelihood estimates for biased sampling models. *Math. Methods Statist.* **3** 275–298.

Wu, C. O. (1996b). *Large Sample Properties of Cross-Validation bandwidths of Kernel Density Estimators with Biased Data*. Technical Report 542, Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD21218.

Wu, C. O., and Mao, A. Q. (1996). Minimax kernels for density estimation with biased data. *Ann. Inst. Statist. Math.* **48** 451–467.