JOURNAL OF MULTIVARIATE ANALYSIS **20**, 91–113 (1986)

Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation

JAMES STEPHEN MARRON*

Department of Statistics, University of North Carolina,
Chapel Hill, North Carolina 27514

AND

WOLFGANG HÄRDLE†

Universität Heidelberg, Sonderforschungsbereich 123, Heidelberg, West Germany

Communicated by M. Rosenblatt

This paper deals with a quite general nonparametric statistical curve estimation setting. Special cases include estimation or probability density functions, regression functions, and hazard functions. The class of "fractional delta sequence estimators" is defined and treated here. This class includes the familiar kernel, orthogonal series, and histogram methods. It is seen that, under some mild assumptions, both the *average square error* and *integrated square error* provide reasonable (random) approximations to the *mean integrated square error*. This is important for two reasons. First, it provides theoretical backing to a practice that has been employed in several simulation studies. Second, it provides a vital tool for proving theorems about selecting smoothing parameters for several different nonparametric curve estimators.

1. INTRODUCTION

Let $X, X_1, \ldots, X_n$ be a random sample of d-dimensional random vectors having density function $f(x)$ and cumulative distribution function $F(x)$.

Received December 6, 1984; revised December 15, 1985.

AMS 1980 subject classifications: Primary 62H12, 62G05.

Key words and phrases: Hazard functions, mean integrated square error, nonparametric estimation, regression function.

* Research partially supported by Office of Naval Research, Contract N00014-75-C-0809, and the Deutsche Forschungsgemeinschaft.

† Research partially supported by the Deutsche Forschungsgemeinschaft, SFB 123, "Stochastische Mathematische Modelle," and Air Force Office of Scientific Research Contract AFOSR-F49620 82 C009

0047-259X/86 $3.00

Suppose we are interested in a certain functional $g(x)$, $x \in \mathbb{R}^d$ of the distribution of X. The problem of estimating the curve $g(x)$ from the random sample is called *nonparametric curve estimation.*

Some special cases of nonparametric curve estimation are:

D—*density estimation*: where g is taken to be f.

H—*hazard function estimation*: where g is given by

$$g(x) = \frac{f(x)}{1 - F(x)}.$$

R—*Regression estimation*: where g is the regression curve of Y on Z',

$$g(x) = g(z) = E[Y \mid Z = z],$$

using the notation

$$\begin{aligned} d' &= d - 1, \\ z &= (z^{(1)}, \ldots, z^{(d')}), \\ x &= (z^{(1)}, \ldots, z^{(d')}, y), \\ X &= (Z^{(1)}, \ldots, Z^{(d')}, Y). \end{aligned} \tag{1.1}$$

This list of examples is meant to be representative, not exhaustive. See Prakasa-Rao [26] for other possibilities.

Quite a number of different estimators have been proposed for each of the curves given above. For comparison of these estimators, several measures of accuracy have been considered. A very common measure is the mean integrated square error,

$$\text{MISE} = E \int [\hat{g}(x) - g(x)]^2 \, w(x) \, dF(x),$$

with some nonnegative weight function $w(x)$ (depending only on z in the regression setting).

While MISE is theoretically pleasing as a distance between $\hat{g}$ and g, it is often hard to compute. The literature contains two different ways of overcoming this difficulty. The first is to study the asymptotic (as $n \to \infty$) behavior of MISE. The second is to consider Monte Carlo (and hence random) appoximations to MISE. In this paper it is seen that, for many estimators, these two approaches give quite similar results for large values of n.

Stochastic (i.e., random) distances that have been considered include the integrated square error (ISE) given by

$$\text{ISE} = \int [\hat{g}(x) - g(x)]^2 \, w(x) \, dF(x),$$

and the average square error (ASE) given by

$$\mathrm{ASE} = n^{-1} \sum_{i=1}^{n} [\hat{g}(X_i) - g(X_i)]^2 \, w(X_i).$$

Wegman [48] argued in the setting of density estimation that, for n large, ASE should be a good approximation of MISE.

He then used ASE as a distance measure for a Monte Carlo comparison of several density estimators. ASE has also been employed for this purpose by Fryer [11] and Wahba [42]. Breiman, Meisel, and Purcell [5] and Raatgever and Duin [27] used a "normalized version" of ASE in their Monte Carlo studies. The distance ISE also has been attractive to several authors, see, for example, Rust and Tsokos [32], Scott and Factor [33], Bean and Tsokos [1], and Bowman [4]. In the regression setting, Stone [36] has used a "leave-one-out" version of ASE and Engle, Granger, Rice, and Weiss [9] and Silverman [34] have used ASE to study cross-validated estimators. In the hazard function setting, Tanner and Wong [40] have compared two estimators by computing the difference of their ASEs.

The use of ASE and ISE as measures of accuracy was criticized by Steele [35], who gave an example in which, asymptotically as $n \to \infty$, ASE behaved very differently from ISE (hence, at least one is a poor approximation to MISE). In reply to this objection, Hall [13] showed that Steele's example was somewhat artificial by showing that, in the case $d = 1$, if $\hat{g}(x)$ is a kernel density estimator, then under some reasonable assumptions, as $n \to \infty$,

$$\mathrm{ASE} = \mathrm{MISE} + o_p(\mathrm{MISE}), \tag{1.2}$$

$$\mathrm{ISE} = \mathrm{MISE} + o_p(\mathrm{MISE}), \tag{1.3}$$

and if $\hat{g}(x)$ is a trigonometric series density estimator (1.3) holds.

The object of this paper is twofold. First, Hall,'s results are extended to a wider class of estimators and to a variety of nonparametric curve estimation settings. This demonstrates that the objections of Steele [35] need cause no concern in the case of many commonly considered estimators. Second, the results of this paper provide an important tool for use in analyzing curve estimators with data-based smoothing parameter selection. In particular, asymptotic optimality results can be derived from suitable uniform versions of (1.2) and (1.3). Special cases of this may be seen, either explicitly or implicitly, in the results of Hall [14], Stone [37, 38], Burman [2], and Marron [22, 23] in the density estimation setting, and in the results of Rice [28], Härdle and Marron [19, 20], and Burman and Chen [3] in the regression setting.

Section 2 introduces the class of "fractional delta sequence estimators" and makes evident that many of the most widely studied nonparametric estimators are contained in this class. Section 3 contains theorems which give sufficient conditions for (1.2) and (1.3) for a subset of these estimators. Section 4 contains theorems which extend the results of Sections 3 to all fractional delta sequence estimators. Section 5 contains examples for illustration of these theorems. The proofs of the theorems are in Section 6.

2. Fractional Delta Sequence Estimators

The class of *fractional delta sequence estimators* is defined to consist of all estimators of the form

$$\hat{g}_\lambda(x) = \frac{\sum_{i=1}^n \delta_\lambda(x, X_i)}{\sum_{i=1}^n \delta'_\lambda(x, X_i)}, \tag{2.1}$$

where δ_λ and δ'_λ are measurable functions on $\mathbb{R}^d \times \mathbb{R}^d$, which are indexed by a "smoothing parameter" $\lambda = \lambda(n) \in \mathbb{R}^+$. The special case $\delta'_\lambda(x, X_i) \equiv 1$ gives the delta sequence estimators studied by Watson and Leadbetter [47], Földes and Revesz [10], and Walter and Blum [44], among others.

In the setting of density estimation, some well-known estimators of this type are:

D-1. *Kernel estimators.* Introduced by Rosenblatt [29] and Parzen [25], given a "kernel function," $K\colon \mathbb{R}^d \to \mathbb{R}$, and the smoothing parameter, $\lambda \in \mathbb{R}^+$, define

$$\begin{aligned} \delta_\lambda(x, X_i) &= \lambda K(\lambda^{1/d}(x - X_i)), \\ \delta'_\lambda(x, X_i) &\equiv 1. \end{aligned} \tag{2.2}$$

D-2. *Histogram estimators.* Write $\mathbb{R}^d = \bigcup_{l=1}^\infty A_l$, were the "bins" A_l are disjoint with Lebesque measure λ^{-1} (where λ is the smoothing parameter). For $l = 1, 2,...$ let $1_l(x)$ denote the indicator of A_l. Define

$$\begin{aligned} \delta_\lambda(x, X_i) &= \sum_{l=1}^\infty \lambda 1_l(x)\, 1_l(X_i), \\ \delta'_\lambda(x, X_i) &\equiv 1. \end{aligned} \tag{2.3}$$

The extension to unequal bin sizes is straightforward, but requires more notation.

D-3. *Orthogonal series estimators.* Introduced by Cencov [6]. Suppose $\{\psi_l(x)\}$ is a sequence of functions which is orthonormal and complete with respect to the inner product

$$\int \psi_l(x)\, \psi_{l'}(x)\, w(x)\, dF(x). \tag{2.4}$$

Given the smoothing parameter $\lambda \in \mathbb{Z}^+$, define

$$\begin{aligned} \delta_\lambda(x, X_i) &= \sum_{l=1}^{\lambda} \psi_l(x)\, \psi_l(X_i)\, w(X_i), \\ \delta'_\lambda(x, X_i) &\equiv 1. \end{aligned} \tag{2.5}$$

Further examples of delta sequence density estimators may be found in Walter and Blum [44] and Susarla and Walter [39]. Some examples of fractional delta sequence estimators in the regression setting are:

R-1. *Kernel estimators.* Introduced by Nadaraya [24] and Watson [45]. Given a kernel function, $K(x')$ and a smoothing parameter, λ, using the notation (1.1), define

$$\begin{aligned} \delta_\lambda(x, X_i) &= \lambda K(\lambda^{1/d'}(z - Z_i))\, Y_i \\ \delta'_\lambda(x, X_i) &= \lambda K(\lambda^{1/d'}(z - Z_i)). \end{aligned}$$

Note that, $\hat{g}(x)$ is a weighted average of the Y_i.

R-2. *Known-marginal kernel estimators.* Studied by Johnston [21]. Let $f_M(z)$ denote the marginal density of Z_i and define

$$\begin{aligned} \delta_\lambda(x, X_i) &= \lambda K(\lambda^{1/d'}(z - Z_i))\, Y_i \\ \delta'_\lambda(x, X_i) &= f_M(z). \end{aligned}$$

To see the idea behind this estimator, note that when the denominator of R-1 is properly normalized, it becomes the estimate D-1 of the marginal density, $f_M(z)$.

R-3. *Delta sequence estimators.* A generalization of R-1, discussed in Collomb [7]; define $\tilde{\delta}_\lambda(z, Z_i)$ as for any of the density estimators and let

$$\begin{aligned} \delta_\lambda(x, X_i) &= \tilde{\delta}_\lambda(z, Z_i)\, Y_i, \\ \delta'_\lambda(x, X_i) &= \tilde{\delta}_\lambda(z, Z_i). \end{aligned}$$

Note that the regressogram of Tukey [41] is a special case where $\tilde{\delta}_\lambda$ is defined as for D-2.

In the setting of hazard function estimation, Watson and Leadbetter [46] have introduced the following fractional delta sequence estimators:

H-1. *Kernel estimators.* Given a kernel function, $K(x)$, and a smoothing parameter, λ, define

$$\begin{aligned} \delta_\lambda(x, X_i) &= \lambda K(\lambda(x - X_i)), \\ \delta'_\lambda(x, X_i) &= 1 - \int_{-\infty}^{x} \lambda K(\lambda(t - X_i))\, dt. \end{aligned} \tag{2.6}$$

H-2. *Delta sequence estimators.* A straightforward generalization of H-1; define $\delta_\lambda(x, X_i)$ as in any of the density estimators and let

$$\delta'_\lambda(x, X_i) = 1 - \int_{-\infty}^{x} \delta_\lambda(t, X_i)\, dt.$$

3. Approximation Theorems for Delta Sequence Estimators

This section gives sufficient conditions for (1.2) and (1.3) in the special case of delta sequence estimators, which are of the form

$$\hat{g}_\lambda(x) = n^{-1} \sum_{i=1}^{n} \delta_\lambda(x, X_i) = \int \delta_\lambda(x, x_1)\, dF_n(x_1). \tag{3.1}$$

Assume that λ ranges over a finite set Λ_n, whose cardinality is bounded by

$$\#(\Lambda_n) \leqslant \mathscr{C} n^\rho, \qquad \rho > 0 \tag{3.2}$$

(i.e., is increasing at most algebraically fast). For estimators with a continuous smoothing parameter, such as the kernel estimators, the result of this paper can be easily extended to Λ_n an interval, by a continuity argument (compare Marron [22] and Härdle and Marron [19]).

For ease of presentation, it will be assumed that there are constants $\mathscr{C}$ and $\varepsilon > 0$ so that, for each n, and for all $\lambda \in \Lambda_n$,

$$\mathscr{C}^{-1} n^\varepsilon \leqslant \lambda \leqslant \mathscr{C} n^{1-\varepsilon}. \tag{3.3}$$

The next assumptions are rather technical in nature, but are stated in this form because these are the common properties which make all of the diverse estimators of Section 2 satisfy (1.2) and (1.3). Implicit in these assumptions are conditions on w and f, e.g., boundedness of f or integrability of $w \cdot f$. Precise conditions (on w and f) depend on which estimator is being considered. These conditions are given in Section 5, where it is seen that quite different methods of verification of these assumptions are needed for different estimators. The assumption (3.4) represents the most important property of delta sequence estimators. Intuition can be gained by considering the kernel density estimation case and performing integration by substitution.

For $k = 1, 2, \ldots$ assume there is a constant $\mathscr{C}_k$ so that for any $m = 2, \ldots, 2k$ and $\lambda \geqslant 1$,

$$\left| \int \cdots \int \left[\prod_{i,i'=1}^{m} \delta_\lambda(x_i, x_{i'})^{\alpha_{ii'}} \right] \times \left[\prod_{i=1}^{m} w(x_i)^{\beta_i} \right] dF(x_1) \cdots dF(x_m) \right| \leqslant \mathscr{C}_k \lambda^{k - m/2}, \tag{3.4}$$

where $\alpha_{ii'} = 0, \dots, k$ subject to

$$\sum_{i,i'=1}^{m} \alpha_{ii'} = k$$

and the restriction that for each $i = 1, \dots, m$, there is an $i' \neq i$ so that either $\alpha_{ii'}$ or $\alpha_{i'i}$ is nonzero, and where $\beta_i = 0, 1$ with $\beta_i = 1$ any time an $\alpha_{ii'} \geqslant 1$ (with $w(x_i)^{\beta_i}$ taken to be 1 when $w(x_i) = \beta_i = 0$).

Assume that the quantity

$$\tilde{\delta}_\lambda(x_1, x_2) = \int \delta_\lambda(x_3, x_1)\, \delta_\lambda(x_3, x_2)\, w(x_3)\, dF(x_3) \tag{3.5}$$

satisfies the assumption (3.4), with each $\beta_i = 0$, and that there is a constant $\mathscr{C}$ so that

$$\iint \tilde{\delta}_\lambda(x_1, x_2)\, dF(x_1)\, dF(x_2) \leqslant \mathscr{C}. \tag{3.6}$$

Assume there is a constant $\mathscr{C}$ so that

$$\int \tilde{\delta}_\lambda(x, x)\, dF(x) \geqslant \mathscr{C}\lambda. \tag{3.7}$$

Another assumption is that there is a constant $\xi > 0$, so that for $k = 1, 2, \dots$ there is a constant $\mathscr{C}_k$ such that

$$\int B(x)^{2k}\, w(x)\, dF(x) \leqslant \mathscr{C}_k b(\lambda)\, \lambda^{(k-1)(1-\xi)}, \tag{3.8}$$

where $B(x)$ denotes the bias and $b(\lambda)$ denotes the integrated squared bias of the estimator $\hat{g}$ given by

$$\begin{aligned} B(x) &= E[\hat{g}(x)] - g(x) = \int \delta_\lambda(x, x_2)\, dF(x_2) - g(x), \\ b(\lambda) &= \int B(x)^2\, w(x)\, dF(x). \end{aligned} \tag{3.9}$$

Finally assume that for $k = 1, 2, \dots$ there is a constant $\mathscr{C}_k$ so that

$$\int [\delta_\lambda(x, x)]^{2k}\, w(x)\, dF(x) \leqslant \mathscr{C}_k \lambda^{2k}. \tag{3.10}$$

THEOREM 1. *Under the assumptions* (3.1)–(3.7),

$$\lim_{n \to \infty} \sup_{\lambda \in \Lambda_n} \left| \frac{\mathrm{ISE}(\lambda) - \mathrm{MISE}(\lambda)}{\mathrm{MISE}(\lambda)} \right| = 0 \quad \text{a.s.}$$

THEOREM 2. *Under the assumptions* (3.1)–(3.10), *and w bounded,*

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda_n} \left|\frac{\mathrm{ASE}(\lambda)-\mathrm{MISE}(\lambda)}{\mathrm{MISE}(\lambda)}\right| = 0 \qquad a.s.$$

Remark 1. We believe that the proofs of these approximations can be extended to the case of λ, a vector, or even a matrix, but additional messy notation and tedious work are required for this.

Remark 2. In this case of kernel density estimation, under stronger conditions than those given here, the strong law of large numbers in Theorem 1 has been extended to a central limit theorem by Hall [15].

Remark 3. The supremum over λ is essential for analyzing curve estimators with a data-dependent smoothing parameter. Such estimators are of the form

$$\hat{g}_L(x) = n^{-1} \sum_{i=1}^{n} \delta_L(x, X_i),$$

where $L = L(X_1,..., X_n)$. Note that as long as $L \in \Lambda$ a.s., we immediately have, under the above assumptions,

$$\lim_{n\to\infty} \left|\frac{\mathrm{ISE}(L)-\mathrm{MISE}(L)}{\mathrm{MISE}(L)}\right| \to 0 \qquad \text{a.s.}$$

and similarly for ASE.

4. APPROXIMATION THEOREMS FOR FRACTIONAL DELTA SEQUENCE ESTIMATORS

This section extends Theorems 1 and 2 to include fractional delta sequence estimators. Since these estimators have denominators containing random variables, they are technically more difficult to work with. In fact, for the estimator R-1, if the kernel function, K, is allowed to take on negative values, then the moments of $\hat{g}(x)$ may not exist (see Rosenblatt [30] and Härdle and Marron [18]) so MISE is not a reasonable distance. These difficulties are overcome using the same method as that employed in Chapter 6 of Cochran [8] for the study of ratio estimators. Assume there is a function $D(x)$ and a set $S \subset \mathbb{R}^d$ so that, uniformly over $x \in S$, $\lambda \in \Lambda_n$,

$$n^{-1} \sum \delta'_\lambda(x, X_i) \to D(x) \qquad \text{a.s.} \tag{4.1}$$

and assume that

$$\inf_{x \in S} D(x) > 0. \tag{4.2}$$

Then, uniformly over $x \in S$, $\lambda \in \Lambda_n$,

$$\begin{aligned} \hat{g}(x) - g(x) &= n^{-1} \sum_i \left[\frac{\delta_\lambda(x, X_i) - \delta'_\lambda(x, X_i)\, g(x)}{D(x)} \right] \\ &\quad + \frac{[D(x) - n^{-1} \sum_i \delta'_\lambda(x, X_i)]\, n^{-1} \sum_i [\delta_\lambda(x, X_i) - \delta'_\lambda(x, X_i)\, g(x)]}{D(x)\, n^{-1} \sum_i \delta'_\lambda(x, X_i)} \\ &= n^{-1} \sum_i \delta^*_\lambda(x, X_i) + o\left(n^{-1} \sum_i \delta^*_\lambda(x, X_i) \right), \end{aligned}$$

where

$$\delta^*_\lambda(x, X_i) = [\delta_\lambda(x, X_i) - \delta'_\lambda(x, X_i)\, g(x)]/D(x). \tag{4.3}$$

Thus, for $w(x)$ supported inside S, it makes sense to replace MISE by

$$\text{MISE}^* = E \int \left[n^{-1} \sum_{i=1}^{n} \delta^*_\lambda(x, X_i) \right]^2 w(x)\, dF(x). \tag{4.4}$$

Similarly, ISE and ASE may be replaced with

$$\begin{aligned} \text{ISE}^* &= \int \left[n^{-1} \sum_{i=1}^{n} \delta^*_\lambda(x, X_i) \right]^2 w(x)\, dF(x) \\ \text{ASE}^* &= n^{-1} \sum_{j=1}^{n} \left[n^{-1} \sum_{i=1}^{n} \delta^*_\lambda(X_j, X_i) \right]^2 w(X_j). \end{aligned} \tag{4.5}$$

Before the theorems are stated, note that MISE* may be considered to be an assessment of how accurately the delta sequence estimator $\hat{g}^*(x)$, defined by

$$\hat{g}^*(x) = n^{-1} \sum_{i=1}^{n} \delta^*_\lambda(x, X_i),$$

estimates the function $g^*(x)$, defined by

$$g^*(x) \equiv 0.$$

Similarly ISE* and ASE* are the ISE and ASE for this new estimation problem. This observation allows immediate application of Theorems 1 and 2.

THEOREM 3. *If δ_λ^* satisfies the assumptions* (3.1) – (3.7) *then*

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda_n} \left|\frac{\mathrm{ISE}^*(\lambda)-\mathrm{MISE}^*(\lambda)}{\mathrm{MISE}^*(\lambda)}\right| = 0 \qquad a.s.$$

COROLLARY. *If, in addition,* (4.1) *holds, then*

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda_n} \left|\frac{\mathrm{ISE}(\lambda)-\mathrm{MISE}^*(\lambda)}{\mathrm{MISE}^*(\lambda)}\right| = 0 \qquad a.s.$$

THEOREM 4. *If δ_λ^* satisfies the assumptions* (3.1)–(3.10) *and w is bounded, then*

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda_n} \left|\frac{\mathrm{ASE}^*(\lambda)-\mathrm{MISE}^*(\lambda)}{\mathrm{MISE}^*(\lambda)}\right| = 0 \qquad a.s.$$

COROLLARY. *If, in addition,* (4.1) *holds, then*

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda_n} \left|\frac{\mathrm{ASE}(\lambda)-\mathrm{MISE}^*(\lambda)}{\mathrm{MISE}^*(\lambda)}\right| = 0 \qquad a.s.$$

To see how Theorem 1 and 2 are intimately related to Theorems 3 and 4, note that in the special case where $\hat{g}(x)$ is a delta sequence estimator (i.e., $\delta'_\lambda(x, X_i) \equiv 1$), conditions (4.1) and (4.2) hold trivially and the quantities MISE*, ISE*, and ASE* are the same as their unasterisked counterparts. Thus Theorems 1 and 2 are special cases of Theorems 3 and 4. On the other hand, using the viewpoint given above, Theorems 3 and 4 are consequences of Theorems 1 and 2.

5. EXAMPLES

In this section it is seen how the fractional delta sequence estimators of Section 2 satisfy the conditions of Sections 3 and 4.

D-1. *Kernel estimators.* Conditions (3.4)–(3.7) follow easily from integration by substitution and the assumptions that f, $w\cdot f$, and K are bounded with $\int K(x)\,dx = 1$ and f, w not mutually singular. Condition (3.8) is also easily satisfied with $\xi = 1$. Condition (3.10) requires the additional assumption that $w\cdot f$ be integrable. Thus the results of Marron [23] and Theorems 1 and 2 of Hall [13] are special cases of the results of this paper.

D-2. *Histogram estimators.* Note that

$$\sup_{x_1,x_2} \delta_\lambda(x_1, x_2) = \lambda, \qquad \sup_{x_2} \int \delta_\lambda(x_1, x_2)\,dx_1 = 1.$$

Hence, (3.4), (3.8), and (3.10) follow easily when it is assumed that f and $w \cdot f$ are bounded and integrable. Next observe that

$$\tilde{\delta}_\lambda(x_1, x_2) = \sum_{l=1}^{\infty} \lambda 1_l(x_1)\, 1_l(x_2) \left(\lambda \int_{A_l} w(x)\, dF(x) \right),$$

and so (3.4) with δ_λ replaced by $\tilde{\delta}_\lambda$, (3.6), and (3.7) are satisfied under the above assumptions, together with (for (3.7)) the assumption that f and w are not mutually singular.

D-3. *Orthogonal series estimators.* The assumptions needed to verify (3.4) are summarized in

LEMMA 1. *If, for* $k = 1, 2,\ldots$ *there is a constant* $\mathscr{C}_k$ *so that for* $l_1,\ldots, l_k = 1, 2,\ldots$ *and for* $r = 1,\ldots, k$,

$$\int \psi_{l_1}^2(x) \cdots \psi_{l_k}^2(x)\, w(x)^r\, dF(x) \leqslant \mathscr{C}_k^2, \tag{5.1}$$

then (3.4) *holds.*

The proof of this lemma is in Section 7. Note that (5.1) is easily satisfied for either the familiar trigonometric or Hermite series. Next observe that

$$\iint \delta_\lambda(x_1, x_2)^2\, w(x_1)\, dF(x_1)\, dF(x_2) = \int \sum_{l=1}^{\lambda} \psi_l(x_2)^2\, w(x_2)^2\, dF(x_2),$$

so (3.7) is easily satisfied. Condition (3.5) follows from

$$\tilde{\delta}_\lambda(x_1, x_2) = \delta_\lambda(x_1, x_2)\, w(x_1), \tag{5.2}$$

and the assumption that w is bounded. Condition (3.6) follows from (5.2) together with the Schwartz inequality. Verifidation of (3.8) follows easily from

$$\sup_{x_1} \left[\int \delta_\lambda(x_1, x_2)\, dF(x_2) - f(x_1) \right]^2 w(x_1) \leqslant \mathscr{C} \lambda^{(1-\xi)},$$

which is easy to check in the Hermite series case, but, using the computations of Hall [12], requires the additional assumption of f'' bounded in the case of trigonometric series. Condition (3.10) is obvious under the above assumptions for either the trigonometric or Hermite series. Theorem 3 of Hall [13] is a special case of this.

R-1, *Kernel estimators.* Conditions (3.4)–(3.10) are easily verified under the same assumptions as D-1, above, together with the assumption that for $k = 1, 2,\ldots$ there is a constant $\mathscr{C}_k$ so that, for z in the support of w,

$$E[Y^k \mid Z = z] \leqslant \mathscr{C}_k.$$

The verification of (4.1) is easy, in view of Lemma 1 of Härdle and Marron [19], under the additional assumption that f_M is Hölder continuous.

R-2. *Known marginal kernel estimators.* This case is similar to R-1 except that (4.1) is not required (but (4.2) is still important). R-1 and R-2 contain the results of Hardle [17] and Hall [16] as special cases.

H-1. *Kernel estimators.* Conditions (3.4)–(3.10) are easily checked when it is assumed that

$$\int K(x)\,dx = 1,$$

and $K, f,$ and $w \cdot f$ are bounded, together with the assumption that $1 - F$ is bounded above 0 on the support of w.

6. Proofs of Theorems 1 and 2

Note that, by (3.2) and the Chebyshev inequality, for $\varepsilon > 0$, $k = 1, 2, \ldots,$

$$P\left[\sup_{\lambda \in \Lambda_n} \left|\frac{\mathrm{ISE}(\lambda) - \mathrm{MISE}(\lambda)}{\mathrm{MISE}(\lambda)}\right| > \varepsilon\right] \leqslant \mathscr{C} n^{\rho} \sup_{\lambda \in \Lambda_n} E\left[\frac{\mathrm{ISE}(\lambda) - \mathrm{MISE}(\lambda)}{\mathrm{MISE}(\lambda) \cdot \varepsilon}\right]^{2k}.$$

Thus, by the Borel–Cantelli lemma, the proof of Theorem 1 will be complete when it is seen that there is a constant $\gamma > 0$, so that for $k = 1, 2, \ldots,$ there are constants $\mathscr{C}_k$ so that

$$E\left[\frac{\mathrm{ISE}(\lambda) - \mathrm{MISE}(\lambda)}{\mathrm{MISE}(\lambda)}\right]^{2k} \leqslant \mathscr{C}_k n^{-\gamma k}. \tag{6.1}$$

Theorem 2 will be established by the same technique when it is shown that

$$E\left[\frac{\mathrm{ASE}(\lambda) - \mathrm{MISE}(\lambda)}{\mathrm{MISE}(\lambda)}\right]^{2k} \leqslant \mathscr{C}_k n^{-\gamma k}. \tag{6.2}$$

The distance ISE can be decomposed as

$$\mathrm{ISE} = R(\lambda) + 2S(\lambda) + b(\lambda),$$

where $b(\lambda)$ is defined in (3.9) and

$$R(\lambda) = \iiint \delta_\lambda(x_1, x_2)\, \delta_\lambda(x_1, x_3)\, w(x_1)\, dF(x_1) \times d(F_n - F)(x_2)\, d(F_n - F)(x_3),$$

$$S(\lambda) = \iint \delta_\lambda(x_1, x_2)\, B(x_1)\, w(x_1)\, dF(x_1)\, d(F_n - F)(x_2).$$

The first term may be further split into

$$R(\lambda) = R_1(\lambda) + R_2(\lambda) + R_3(\lambda),$$

where, using the notation (3.5),

$$R_1(\lambda) = \iint_{\{x_2 \neq x_3\}} \bar{\delta}_\lambda(x_2, x_3)\, d(F_n - F)(x_2)\, d(F_n - F)(x_3),$$

$$R_2(\lambda) = n^{-1} \int \bar{\delta}_\lambda(x_2, x_2)\, d(F_n - F)(x_2),$$

$$R_3(\lambda) = n^{-1} \int \bar{\delta}_\lambda(x_2, x_2)\, dF(x_2).$$

To finish the proof of (6.1) it is enough to show that

$$\left[\frac{R_3(\lambda) + b(\lambda) - \mathrm{MISE}(\lambda)}{\mathrm{MISE}(\lambda)}\right]^{2k} \leqslant \mathscr{C}_k n^{-\gamma k}, \tag{6.3}$$

and for "term" denoting R_1, R_2, or S,

$$E\left[\frac{\text{term}}{\mathrm{MISE}(\lambda)}\right]^{2k} \leqslant \mathscr{C}_k n^{-\gamma k}. \tag{6.4}$$

Write

$$\mathrm{ASE} = \mathrm{ISE} + T(\lambda).$$

As above, $T(\lambda)$ admits the decomposition

$$\begin{aligned} T = {} & T_1 + T_2 + T_3 + 2T_4 + 2T_5 + T_6 \\ & + T_7 + 2U_1 + 2U_2 + 2U_3 + V, \end{aligned}$$

where

$$\begin{aligned} T_1(\lambda) = {} & \iint_{\{x_1 \neq x_2 \neq x_3 \neq x_1\}} \delta_\lambda(x_1, x_2)\, \delta_\lambda(x_1, x_3)\, w(x_1)\, d(F_n - F)(x_1) \\ & \times d(F_n - F)(x_2)\, d(F_n - F)(x_3), \end{aligned}$$

$$T_2(\lambda) = n^{-1} \iint_{\{x_1 \neq x_2\}} \delta_\lambda(x_1, x_2)^2\, w(x_1)\, d(F_n - F)(x_1)\, d(F_n - F)(x_2),$$

$$T_3(\lambda) = n^{-1} \iint \delta_\lambda(x_1, x_2)^2\, w(x_1)\, d(F_n - F)(x_1)\, dF(x_2),$$

$$T_4(\lambda) = n^{-1} \iint_{\{x_1 \neq x_2\}} \delta_\lambda(x_1, x_2\} \, \delta_\lambda(x_1, x_1) \, w(x_1)$$
$$\times \, d(F_n - F)(x_1) \, d(F_n - F)(x_2),$$

$$T_5(\lambda) = n^{-1} \iint \delta_\lambda(x_1, x_2) \, \delta_\lambda(x_1, x_1) \, w(x_1) \, dF(x_1) \, d(F_n - F)(x_2),$$

$$T_6(\lambda) = n^{-2} \int \delta_\lambda(x, x)^2 \, w(x) \, d(F_n - F)(x),$$

$$T_7(\lambda) = n^{-2} \int \delta_\lambda(x, x)^2 \, w(x) \, dF(x),$$

$$U_1(\lambda) = \iint_{\{x_1 \neq x_2\}} \delta_\lambda(x_1, x_2) \, B(x_1) \, w(x_1) \, d(F_n - F)(x_1) \, d(F_n - F)(x_2),$$

$$U_2(\lambda) = n^{-1} \int \delta_\lambda(x_1, x_1) \, B(x_1) \, w(x_1) \, d(F_n - F)(x_1),$$

$$U_3(\lambda) = n^{-1} \int \delta_\lambda(x_1, x_1) \, B(x_1) \, w(x_1) \, dF(x_1),$$

$$V(\lambda) = \int B(x_1)^2 \, w(x_1) \, d(F_n - F)(x_1).$$

Thus, (6.2) will be established when (6.4) is verified for each of the above terms as well.

To check (6.3), note that by the familiar variance-bias squared decomposition (see, e.g., Rosenblatt [31]), using the notation (3.9),

$$\text{MISE} = R_3(\lambda) - r(\lambda) + b(\lambda),$$

where, using the notation (3.5),

$$r(\lambda) = n^{-1} \iint \tilde{\delta}_\lambda(x_2, x_3) \, dF(x_2) \, dF(x_3).$$

The inequality (6.3) follows from this and from (3.3), (3.6), and (3.7).

The verification of (6.4) will now be done term by term, starting with those which do not involve $d(F_n - F)$:

Term T_7. Using (3.10),

$$\left[\frac{n^{-2} \int \delta_\lambda(x, x)^2 \, w(x) \, dF(x)}{\text{MISE}(\lambda)} \right]^{2k} \leqslant \mathscr{C}_k \left[\frac{n^{-2}\lambda^2}{n^{-1}\lambda} \right]^{2k} \leqslant \mathscr{C}'_k n^{-2k\xi}.$$

Term U_3. As above, using the Schwartz inequality,

$$\left[\frac{n^{-1}\int \delta_\lambda(x_1, x_1)\, B(x_1)\, w(x_1)\, dF(x_1)}{\mathrm{MISE}(\lambda)}\right]^{2k}$$

$$\leqslant \left[\frac{n^{-1}[\int \delta_\lambda(x_1, x_1)^2\, w(x_1)\, dF(x_1)]^{1/2}\, b(\lambda)^{1/2}}{\mathrm{MISE}(\lambda)}\right]^{2k}$$

$$\leqslant \mathscr{C}_k \left[\frac{n^{-1}\lambda \cdot b(\lambda)^{1/2}}{(n^{-1}\lambda)^{1/2}\, b(\lambda)^{1/2}}\right]^{2k} \leqslant \mathscr{C}_k (n^{-1}\lambda)^k \leqslant \mathscr{C}'_k n^{-k\varepsilon}.$$

The remaining terms all have at least one $d(F_n - F)$, and so have mean 0. Thus to check (6.4), by the cumulant expansion of the $2k$th moment, it is enough to check that, for $k = 2, 3, \ldots$, there is a constant $\mathscr{C}_k$ so that

$$\left|\mathrm{cum}_k\left(\frac{\text{term}}{\mathrm{MISE}}\right)\right| \leqslant \mathscr{C}_k n^{-\gamma k}, \tag{6.5}$$

where $\mathrm{cum}_k(\cdot)$ denotes the kth order cumulant, for which each argument is the same.

To verify (6.5) in the case of those terms having only one $d(F_n - F)$, note that they may be written

$$n^{-1} \sum_{i=1}^{n} W(X_i).$$

Thus, using the independence property and linearity of cumulants, it is enough to show that

$$n^{-k+1}\, \mathrm{MISE}^{-k}\, |E[W(X_1)]^k| \leqslant \mathscr{C}_k\, n^{-\gamma k}.$$

Term R_2. Note that here

$$W(X_2) = n^{-1}\left[\int \delta_\lambda(x_1, X_2)^2\, w(x_1)\, dF(x_1)\right.$$
$$\left. - \iint \delta_\lambda(x_1, x_2)^2\, w(x_1)\, dF(x_1)\, dF(x_2)\right].$$

So by the binomial theorem and repeated application of (3.4),

$$n^{-k+1}\, \mathrm{MISE}^{-k}\, |E[W(X_2)]^k| \leqslant \mathscr{C}_k n^{-2k+1} (n^{-1}\lambda)^{-k}\, \lambda^{2k-(k+1)/2} \leqslant \mathscr{C}'_k n^{-k/4}.$$

Term T_3. Similar to R_2.

Term T_5. Similar to R_2.

Term T_6. Note that here

$$W(X_1)=n^{-2}\left[\delta_\lambda(X_1, X_1)^2\, w(X_1)-\int \delta_\lambda(x_1, x_1)^2\, w(x_1)\, dF(x_1)\right].$$

So by (3.10)

$$n^{-k+1}\,\mathrm{MISE}^{-k}\,|E[W(X_1)]^k| \leqslant \mathscr{C}_k n^{-3k+1}(n^{-1}\lambda)^{-k}\,\lambda^{2k} \leqslant \mathscr{C}'_k n^{-k/2}.$$

Term V. Note that here

$$W(X_1)=B(X_1)^2\, w(X_1)-b(\lambda).$$

Thus, by (3.8),

$$\begin{aligned} n^{-k+1}\,\mathrm{MISE}^{-k}\,|E[W(X_1)]^k| \leqslant{}& \mathscr{C}_k n^{-k+1}\Bigg(b(\lambda)^k[b(\lambda)]^{-k} \\ &+\sum_{j=1}^{k}\,[b(\lambda)\,\lambda^{(j-1)(1-\varepsilon)}]\, b(\lambda)^{k-j}[(n^{-1}\lambda)^{j-1}b(\lambda)^{k-j+1}]^{-1}\Bigg) \\ \leqslant{}& \mathscr{C}'_k n^{-\gamma k}. \end{aligned}$$

Term U_2. Note that here

$$W(X_1)=n^{-1}\left[\delta_\lambda(X_1, X_1)\, B(X_1)\, w(X_1)-\int \delta_\lambda(x_1, x_1)\, B(x_1)\, w(x_1)\, dF(x_1)\right].$$

By (3.8), (3.10), and the Schwartz inequality, for $j=1, 2,\ldots$, there is $\mathscr{C}_j$ so that

$$\begin{aligned} &\left|\int [\delta_\lambda(x_1, x_1)\, B(x_1)\, w(x_1)]^j\, dF(x_1)\right| \\ &\quad\leqslant \left[\int \delta_\lambda(x_1, x_1)^{2j}\, w(x_1)^{2j-1}\, dF(x_1)\right]^{1/2}\left[\int B(x_1)^{2j}\, w(x_1)\, dF(x_1)\right]^{1/2} \\ &\quad\leqslant \mathscr{C}_j\lambda^{(3j-1)/2}b(\lambda)^{1/2}. \end{aligned}$$

Hence,

$$\begin{aligned} &n^{-k+1}\,\mathrm{MISE}^{-k}\,|E[W(X_1)]^k| \\ &\quad\leqslant \mathscr{C}_k n^{-2k+1}[(n^{-1}\lambda)^{-k+1/2}b(\lambda)^{-1/2}]\,\lambda^{(3k-1)/2}b(\lambda)^{1/2} \\ &\quad\leqslant \mathscr{C}^1_k n^{-k/4}. \end{aligned}$$

Term S. Note that here

$$W(X_2)=\int \delta_\lambda(x_1, X_2)\, B(x_1)\, w(x_1)\, dF(x_1)$$

$$-\iint \delta_\lambda(x_1, x_2)\, B(x_1)\, w(x_1)\, dF(x_1)\, dF(x_2).$$

It follows from the Schwartz inequality that,

$$\left|\int \left[\int \delta_\lambda(x_1, x_2)\, B(x_1)\, w(x_1)\, dF(x_1)\right]^j dF(x_2)\right|$$

$$\leqslant \int \left[\int \delta_\lambda(x_1, x_2)^2\, w(x_1)\, dF(x_1)\right]^{j/2} b(\lambda)^{j/2}\, dF(x_2). \qquad (6.6)$$

So, by (3.4), for j even, there is a constant $\mathscr{C}_j$ such that (6.6) is bounded by

$$\mathscr{C}_j b(\lambda)^{j/2}\, \lambda^{j-(j/2+1)/2}=\mathscr{C}_j \lambda^{3j/4-1/2} b(\lambda)^{j/2}.$$

And by the moment inequality, for j odd, there is a constant $\mathscr{C}_j$ such that (6.6) is bounded by

$$b(\lambda)^{j/2}\left[\int \left(\int \delta_\lambda(x_1, x_2)^2\, w(x_1)\, dF(x_1)\right)^{(j+1)/2} dF(x_2)\right]^{j/(j+1)}$$

$$\leqslant \mathscr{C}_j b(\lambda)^{j/2}\, \lambda^{3j/4-j/2(j+1)}.$$

Thus, for $k=3, 4,...$

$$n^{-k+1}\,\mathrm{MISE}^{-k}\, |EW(X_2)^k| \leqslant \mathscr{C}_k n^{-k+1}(n^{-1}\lambda)^{-k/2}\, b(\lambda)^{-k/2}\, b(\lambda)^{k/2}\, \lambda^{3k/4-3/8}$$

$$=\mathscr{C}_k n^{-k/2+1}\lambda^{k/4-3/8}=\mathscr{C}_k(n^{-1}\lambda)^{k/2-1}\, \lambda^{-k/4+5/8}$$

$$\leqslant \mathscr{C}'_k n^{-\varepsilon k/4}.$$

More precise computations are required in the case $k=2$. By (3.5),

$$E(W(X_2)^2)\leqslant E\left[\int \delta_\lambda(x_1, X_2)\, B(x_1)\, w(x_1)\, dF(x_1)\right]^2$$

$$=\int \left[\int \left(\int \delta_\lambda(x_1, x_2)\, \delta_\lambda(x'_1, x_2)\, dF(x_2)\right)\right.$$

$$\left.\times\ B(x'_1)\, w(x'_1)\, dF(x'_1)\right] B(x_1)\, w(x_1)\, dF(x_1)$$

$$\leqslant \left(\int \left[\iint \delta_\lambda(x_1, x_2)\, \delta_\lambda(x'_1, x_2)\, dF(x_2)\right.\right.$$

$$\left.\left.\times\ B(x'_1)\, w(x'_1)\, dF(x'_1)\right]^2 w(x_1)\, dF(x_1)\right)^{1/2} b(\lambda)^{1/2}$$

$$\begin{aligned}&\leqslant \Bigg(\int\Bigg[\int\Bigg(\int\Bigg(\int \delta_\lambda(x_1, x_2)\,\delta_\lambda(x_1', x_2)\,dF(x_2)\Bigg)^2 w(x_1')\\ &\quad\times\, dF(x_1')\Bigg)^{1/2} b(\lambda)^{1/2}\Bigg]^2 w(x_1)\,dF(x_1)\Bigg)^{1/2} b(\lambda)^{1/2}\\ &= b(\lambda)\Bigg(\iint \tilde{\delta}_\lambda(x_2, x_2')^2\,dF(x_2)\,dF(x_2')\Bigg)^{1/2} \leqslant b(\lambda)\,\mathscr{C}(\lambda^{2-2/2})^{1/2}.\end{aligned}$$

Thus,

$$n^{-1}\,\mathrm{MISE}(\lambda)^{-2}\,EW(X_2)^2 \leqslant \mathscr{C} n^{-1}(n^{-1}\lambda b(\lambda))^{-1}\,b(\lambda)\,\lambda^{1/2} \leqslant \mathscr{C}' n^{-\varepsilon/2}.$$

It remains to verify (6.5) for the terms containing two or three $d(F_n - F)$'s. The terms containing 2 may all be written in the form

$$n^{-1}\sum_{\substack{i,i'=1\\ i\neq i'}}^{n} W(X_i, X_{i'}),$$

where

$$EW(X_i, X_{i'}) = 0, \qquad i \neq i'.$$

So, using the linearity property of cumulants, (6.5) will be established in this case when it is seen that there is a constant $\gamma > 0$, so that for $k = 2, 3,\ldots,$ there are constants $\mathscr{C}_k$ such that

$$\Bigg| n^{-2k}\,\mathrm{MISE}^{-k} \sum_{i_1, i_1',\ldots, i_k, i_k'} \mathrm{cum}_k(W(X_{i_1}, X_{i_1'}),\ldots, W(X_{i_k}, X_{i_k'}))\Bigg| \leqslant \mathscr{C}_k n^{-\gamma k},$$

where, by a moment expansion of cum_k, it may be assumed that each of $i_1, i_1',\ldots, i_k, i_k'$ appears at least twice. In each case, it will be convenient to let m denote the number of $i_1, i_1',\ldots, i_k, i_k'$ that are unique. Note that, for $m = 2, 3,\ldots, k$, the number of cum_k with m distinct indices is bounded by $\mathscr{C}_k n^m$.

Term T_2. Note that here

$$\begin{aligned}W(X_i, X_{i'}) = n^{-1}\Bigg[&\delta_\lambda(X_i, X_{i'})^2\,w(X_i) - \int \delta_\lambda(X_i, x_2)^2\,w(X_i)\,dF(x_2)\\ &- \int \delta_\lambda(x_1, X_{i'})^2\,w(x_1)\,dF(x_1)\\ &+ \iint \delta_\lambda(x_1, x_2)^2\,w(x_1)\,dF(x_1)\,dF(x_2)\Bigg].\end{aligned}$$

So, by (3.4)

$$\left| n^{-2k}\,\mathrm{MISE}^{-k} \sum \mathrm{cum}_k(W(X_{i_1}, X_{i'_1}),\ldots, W(X_{i_k}, X_{i'_k})) \right|$$

$$\leqslant n^{-2k}(n^{-1}\lambda)^{-k}\, n^{-k}\, \mathscr{C}_k \sum_{m=2}^{k} n^m \lambda^{2k-m/2} \leqslant \mathscr{C}'_k n^{-k/2}.$$

Term T_4. Similar to T_2.

Term R_1. Here

$$W(X_i, X'_i) = \tilde{\delta}_\lambda(X_i, X'_i) - \int \tilde{\delta}_\lambda(X_i, x_2)\, dF(x_2) - \int \tilde{\delta}_\lambda(x_1, X_{i'})\, dF(x_1)$$

$$+ \iint \tilde{\delta}_\lambda(x_1, x_2)\, dF(x_1)\, dF(x_2).$$

Thus,

$$\left| n^{-2k}\,\mathrm{MISE}^{-k} \sum \mathrm{cum}_k(W(X_{i_1}, X_{i'_1}),\ldots, W(X_{i_k}, X_{i'_k})) \right|$$

$$\leqslant n^{-2k}(n^{-1}\lambda)^{-k}\, \mathscr{C}_k \sum_{m=2}^{k} n^m \lambda^{k-m/2} \leqslant \mathscr{C}'_k n^{-\varepsilon k/2}.$$

Term U_1. Here

$$W(X_i, X_{i'}) = \delta_\lambda(X_i, X_{i'})\, B(X_i)\, w(X_i) - \int \delta_\lambda(X_i, x_2)\, B(X_i)\, w(X_i)\, dF(x_2)$$

$$- \int \delta_\lambda(x_1, x_{i'})\, B(X_1)\, w(x_1)\, dF(x_1)$$

$$+ \iint \delta_\lambda(x_1, x_2)\, B(x_1)\, w(x_1)\, dF(x_1)\, dF(x_2).$$

This term is handled by means quite similar to those used on Term T_2 above, except that (3.4) is augmented by the Schwartz inequality and (3.8). The result is, for $k = 2, 3,\ldots,$

$$\left| n^{-2k}\,\mathrm{MISE}^{-k} \sum \mathrm{cum}_k(W(X_{i_1}, X_{i'_1}),\ldots, W(X_{i_k}, X_{i'_k})) \right|$$

$$\leqslant n^{-2k}((n^{-1}\lambda)^{k-1/2}\, b(\lambda)^{1/2})^{-1}\, \mathscr{C}_k \sum_{m=2}^{k} n^m \lambda^{(2k-m)/2} b(\lambda)^{1/2}\, \lambda^{(k-1)(1-\varepsilon)/2}$$

$$\leqslant \mathscr{C}'_k n^{-\varepsilon^2 k/4}.$$

It remains to verify (6.5) for

Term T_1. This term may be handled by methods similar to those used on term T_2.

This completes the proof of Theorems 1 and 2.

7. Proof of Lemma 1

Using the definition of $\delta_\lambda(x, y)$, write

$$\left|\int \cdots \int \left[\prod_{i,i'} \delta_\lambda(x_i, x_{i'})^{\alpha_{ii'}}\right]\left[\prod_i w(x_i)^{\beta_i}\right] dF(x_1) \cdots dF(x_m)\right|$$

$$= | \sum_{l_1=1}^{\lambda} \cdots \sum_{l_k=1}^{\lambda} \int \cdots \int \psi_{l_1}\psi_{l_1} w \cdots \psi_{l_k}\psi_{l_k} w \left[\prod w^{\beta_i}\right] dF(x_1) \cdots dF(x_m) \Bigg|. \quad (7.1)$$

The multiple integral on the right-hand side may now be factored to give an expression of the form

$$\left|\sum_{l_1} \cdots \sum_{l_k} \left[\int \qquad dF(x_1)\right] \cdots \left[\int \qquad dF(x_m)\right]\right|. \quad (7.2)$$

Consider the set of $\alpha_{ii'}$ which have $i \neq i'$ and are positive. Find a subset, A, which has the property that each of 1,..., m appears at least once as an i or i', and suppose that this subset is minimal in the sense that if any $\alpha_{ii'}$ is removed, then 1,..., m no longer all appear as the index of an α.

Group 1,..., m into two subsets, I and I', by the following rules:

(1) Any of 1,..., m that appear twice (or more) as an index of an α in A goes into I.

(2) If i is in I, and $\alpha_{ii'}$ (or $\alpha_{i'i}$) is in A, put i' into I'.

(3) For the remaining $\alpha_{ii'}$ in A, put i in I and i' in I'.

The above rules partition $\{1,..., m\}$ into I and I'.

Observe that for each $\alpha_{ii'}$ in A, there is an l so that $\psi_l(x_i)\, \psi_l(x_{i'})$ appears in the integrand on the right side of (7.1). Suppose, without loss of generality, that $l_1,..., l_L$ each correspond in this manner to a different element of A, where L denotes the cardinality of A. Also assume, without loss of generality, that $l_1,..., l_b$ correspond to those α in A which have an index appearing more than once in A.

By the Schwartz inequality, (7.2) may be written as

$$\left|\sum_{l_1} \cdots \sum_{l_k} \prod_{i=1}^{m} \int [\quad]\, dF(x_i)\right|$$

$$= \left|\sum_{l_{L+1}} \cdots \sum_{l_k} \left[\sum_{l_1} \cdots \sum_{l_L} \left(\prod_{i \in I} \int [\quad]\, dF(x_i)\right)\left(\prod_{i \in I'} \int [\quad]\, dF(x_i)\right)\right]\right|$$

$$\leqslant \sum_{l_{L+1}} \cdots \sum_{l_k} \left[\sum_{l_1} \cdots \sum_{l_L} \left(\prod_{i \in I} \int [\quad]\, dF(x_i) \right)^2 \right]^{1/2}$$
$$\times \left[\sum_{l_1} \cdots \sum_{l_L} \left(\prod_{i \in I'} \int [\quad]\, dF(x_i) \right)^2 \right]^{1/2}.$$

Suppose, without loss of generality, that $I' = \{1,..., L\}$. Then,

$$\sum_{l_1} \cdots \sum_{l_L} \left(\prod_{i \in I'} \int [\quad]\, dF(x_i) \right)^2 = \prod_{i=1}^{L} \sum_{l_i} \left(\int [\quad]\, dF(x_i) \right)^2 \leqslant \prod_{i=1}^{L} \mathscr{C}_k,$$

where the last inequality follows from (5.1) and the Bessel inequality, because $\int [\quad]\, dF(x_i)$ is the l_ith Fourier coefficient of a function whose norm is bounded in (5.1). Similar techniques give

$$\sum_{l_1} \cdots \sum_{l_L} \left(\prod_{i \in I} \int [\quad]\, dF(x_i) \right)^2 \leqslant \lambda^{b-1} \mathscr{C}_k^{L-b+1}.$$

It follows from the above that there is a constant $\mathscr{C}_k$ so that (7.1) is bounded by

$$\mathscr{C}_k \lambda^{k-L} \lambda^{(b-1)/2}.$$

To put this in more useful terms, note that

$$2L - b \geqslant m - 1$$

and so

$$-L + b/2 - \tfrac{1}{2} \leqslant -m/2.$$

It follows that (7.1) is bounded by

$$\mathscr{C}_k \lambda^{k-m/2}.$$

This completes the proof of Lemma 1.

References

[1] Bean, S., and Tsokos, C. P. (1982). Bandwidth selection proceures for kernel density estimates. *Comm. Statist.* A **11** 1045–1069.

[2] Burman, P. (1984). A data dependent approach to density estimation. *Z. Wahrsch. Verw. Gebiete* **69** 609–628.

[3] Burman, P., and Chen, K. W. (1984). Nonparametric estimation of a regression function, unpublished.

[4] Bowman, A. W. (1982). A Comparative Study of Some Kernel-Based Non-Parametric Density Estimators. *J. Statist. Comput. Simulation* **21** 313–327.

[5] Breiman, L., Meisel, W., and Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics* **19** 135–144.

[6] CENCOV, N. N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.* **3** 1559–1562.

[7] COLLOMB, G. (1981). Estimation non parametrique de la regression: Revue Bibliographique. *Internat. Statist. Rev.* **49** 75–93.

[8] COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd. ed. Wiley, New York.

[9] ENGLE, R. F., GRANGER, C. W. J., RICE, J., AND WEISS, A. (1983). *Non-Parametric Estimates of the Relation Between Weather and Elasticity of Demand.* Discussion paper #83-17, Department of Economics, University of California, San Diego.

[10] FÖLDES, A., AND REVESZ, P. (1974). A general method for density estimation. *Studia Sci. Math. Appl. Hungar.* **9** 81–92.

[11] FRYER, M. J. (1977). Review of some non-parametric methods of density estimation. *J. Inst. Math. Its Appl.* **20** 335–354.

[12] HALL, P. (1981). On trigonometric series estimates of densities. *Ann. Statist.* **9** 683–685.

[13] HALL, P. (1982). Limit theorems for stochastic measures of the accuracy of density estimators. *Stochastic Process. Appl.* **13** 11–25.

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

[15] HALL, P. (1984a). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Anal.* **14** 1–16.

[16] HALL, P. (1984b). Asymptotic properties of integrated square error and cross-validation for kernel estimation of a regression function. *Z. Wahrsch. Verw. Gebiete* **67** 175–196.

[17] HÄRDLE, W. (1984). Approximations to the mean integrated squared error with applications to optimal bandwidth selection for nonparametric regression function estimators. *J. Multivariate Anal.* **18** 150–160.

[18] HÄRDLE, W., AND MARRON, J. S. (1983). *The Nonexistence of Moments of Some Kernel Regression Estimators.* Mimeo Series No. 1537. (North Carolina Institute of Statistics.)

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

[20] HÄRDLE, W., AND MARRON, J. S. (1985b). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika* **72** 481–484.

[21] JOHNSTON, G. J. (1982). Properties of maximal deviations for nonparametric regression function estimates. *J. Multivariate Anal.* **12** 402–414.

[22] MARRON, J. S. (1984). An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Ann. Statist.* **13** 1011–1023.

[23] MARRON, J. S. (1986). Convergence properties of an empirical error criterion for multivariate density estimation. *J. Multivariate Anal.* **19** 1–13.

[24] NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

[25] PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Statist.* **33** 1056–1076.

[26] PRAKASA RAO, B. L. S. (1983). *Nonparametric Functional Estimation.* Acadamic Press, New York.

[27] RAATGEVER, J. W., AND DUIN, R. P. W. (1978). On the variable kernel model for multivariate nonparametric density estimation. In *COMPSTAT 1978*: *Proceedings* (L. C. A. Corsten and J. Hermans, Eds.). Birkhäuser, Basel.

[28] RICE, J. (1982). Bandwidth choice for nonparametric kernel regression. *Ann. Statist.* **12** 1215–1230.

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

[30] ROSENBLATT, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis*-II (P. R. Krishnaiah, Ed.) pp. 25–31. Academic Press, New York.

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

[32] Rust, A. E., and Tsokos, C. P. (1981). On the convergence of kernel estimators of probability density functions. *Ann. Inst. Statist. Math.* **33** 233–246.

[33] Scott, D. W., and Factor, L. E. (1981). Monte Carlo study of three data-based nonparametric probability density estimators. *J. Amer. Statist. Assoc.* **76** 9–15.

[34] Silverman, B. W. (1984). A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *J. Amer. Statist. Assoc.*, in press.

[35] Steele, J. M. (1978). Invalidity of average squared error criterion in density estimation. *Canad. J. Statist.* **6** 193–200.

[36] Stone, C. J. (1976). Nearest neighbor estimators of a nonlinear regression function. In *Proceedings, Comput. Sci. Statist. 8th Annual Symposium on the Interface.* Health Sciences Computing Facility, U.C.L.A., pp. 413–418.

[37] Stone, C. J. (1984a). An asymptotically efficient histogram selection rule. *Proceedings of the Neyman–Kiefer Meeting*, in press.

[38] Stone, C. J. (1984b). An asymptotically optimal window selection rule for kernel density estimates *Ann. Statist.*, in press.

[39] Susarla, V., and Walter, G. (1981). Estimation of a multivariate density function using delta sequences. *Ann. Statist.* **9** 347–355.

[40] Tanner, M. A., and Wong W. H. (1982). Data based nonparametric estimation of the hazard function with applications to model diagnostics and exploratory analysis. *J. Amer. Statist. Assoc.* **79** 174–182.

[41] Tukey, J. W. (1961). Curves as parameters, and touch estimation. *Proceedings, 4th Berkely Sympos.* 681–694.

[42] Wahba, G. (1977). Optimal smoothing of density estimates. In *Classification and Clustering.* (J. van Ryzin, Ed.), pp. 423–458.

[43] Walter, G. (1977). Properties of Hermite series estimation of probability density. *Ann. Statist.* **5** 1258–1264.

[44] Walter, G., and Blum, J. (1976). Probability density estimation using delta sequences. *Ann. Statist.* **7** 328–340.

[45] Watson, G. S. (1964). Smooth regression analysis. *Sankhyā* Ser. *A.* **26** 359–372.

[46] Watson, G. S., and Leadbetter, M. R. (1964a). Hazard analysis, I. *Biometrika* **51** 175–184.

[47] Watson, G. S., and Leadbetter, M. R. (1964b). Hazard Analysis, II. *Sankhyā Ser.* A **26** 101–116.

[48] Wegman, E. J. [1972]. Nonparametric probability density estimation. II. A comparison of density estimation methods. *J. Statist. Comput. Simulation* **1** 225–245.