Probab. Th. Rel. Fields 80, 619–632 (1989)

Probability Theory and Related Fields

© Springer-Verlag 1989

# Uniform Consistency of Automatic and Location-adaptive Delta-sequence Estimators

Deborah Nolan[1]* and J. Stephen Marron[2]**

[1] University of California, Department of Statistics, Berkeley CA 94720

[2] University of North Carolina, Department of Statistics, Chapel Hill, NC 27514

**Summary.** The class of delta-sequence estimators for a probability density includes the kernel, histogram and orthogonal series types, because each can be characterized as a collection of averages of some function that is indexed by a smoothing parameter. There are two important extensions of this class. The first allows a random smoothing parameter, for example that specified by a cross-validation method. The second allows the smoothing parameter to be a function of location, for example an estimator based on nearest-neighbor distance. In this paper a general method is presented which establishes uniform consistency for all of these estimators.

## 1. Introduction

Kernel density estimators, histograms, and orthogonal series estimators are well known methods for estimating a density. All three can be represented as an average over independent observations from the unknown distribution. The kernel estimator averages kernel or density functions centered on the observations; the histogram is an average of indicator functions; and the orthogonal series estimator averages products of pairs of functions belonging to a finite subset of a complete orthonormal system. In Sect. 2, notation is introduced to represent all of these averages in the general framework of delta sequence estimators, as introduced by Földes and Revesz (1974) and Walter and Blum (1979).

All three averages are essentially local in character. The effective width of the local average is crucial to the performance of these estimators (see Tapia and Thompson, 1978; Prakasa Rao, 1983; Devroye and Györfi, 1984; and Silverman, 1986). This window width is called the bandwidth for the kernel and the binwidth for the histogram, and for the orthogonal series estimator, it is

* Research partially supported by AFOSR Grant No. S-49620-82-C-0144, and by NSF Grant DMS-850-3347

** Research supported by NSF Grant DMS-8400602

typically controlled by the cardinality of the subset (see Wahba, 1981 for an interesting variant of this). In general, it is called the smoothing parameter.

To establish uniform consistency of these estimators, the most convenient assumption placed on the smoothing parameter is that, for each $n$, it is deterministic and constant with respect to location. See Bertrand-Retali (1974) and Silverman (1978) in the kernel case, Revesz (1972) and Kim and Van Ryzin (1975) in the histogram case, Bleuez and Bosq (1976) in the orthogonal series case, and for general results of this type, see Földes and Revesz (1974).

For practical applications, the assumption of a deterministic smoothing parameter is not realistic, because any reasonable choice of the smoothing parameter must, at least implicitly, be estimated. References to data-based or automatic choices for the smoothing parameter, most of which assume that it is constant with respect to location, may be found in Stone (1984), Marron (1985), Hall and Marron (1987a, b) in the kernel case, Rudemo (1982) and Stone (1985) in the histogram case, Hall (1986, 1987) in the orthogonal series case, and Burman (1985) and Marron (1987) in the general case.

An intuitively appealing variant of the constant smoothing parameter is the location-adaptive parameter. For example, where the data are relatively dense, some improvement in the bias can be made with a small window width, but in locations where the data are sparse, a larger window width reduces the variance. Unfortunately, this flexibility complicates the estimator, because the amount of smoothing is indexed by an entire function instead of simply a constant. The nearest-neighbor kernel estimator (Loftsgaarden and Quesenberry, 1965; Mack and Rosenblatt, 1979) is one approach to this problem. Section 2 describes several other approaches.

Note that it is possible to combine the above two extensions of the usual estimators. In particular, one could consider using a random smoothing parameter in a location adaptive estimator. This case is not treated here.

While there is a large literature on consistency of density estimators, nearly all of it is in the deterministic and location-constant case. Exceptions to this are in Devroye and Wagner (1980) and Devroye and Penrod (1984a) where kernel estimators with a random location-constant bandwidth are treated, and in Devroye (1985) where one particular location-adaptive kernel estimator is examined. This paper presents a general method for the simultaneous treatment of all of the above estimators within the framework of delta-sequence estimators. Our results contain many previous theorems as special cases.

The method of proof consists of separating the difference between the estimator and the target density into stochastic (i.e., variance) and nonrandom (i.e., bias) components. Empirical process techniques are employed to handle the variance component, whereas the bias is handled by an appropriate adhoc method for the specific example. Both components are shown to converge to zero uniformly over location and over a wide range of possible smoothing parameters.

Conditions for the uniform convergence of the variance component are presented in Sect. 3. They are no more restrictive than those commonly imposed on a deterministic sequence of window widths. Section 4 includes a variety of examples of delta-sequence estimators with automatic and location-adaptive parameters. In this section, the conditions from Sect. 3 for handling the variance

component are checked; also of interest, are the techniques for showing the bias component is negligible. The examples are first formally introduced in Sect. 2. Section 5 contains the proof of the result stated in Sect. 3.

## 2. Examples of Delta-sequence Estimators

Let $\xi_1, \ldots, \xi_n$ be independent observations from an unknown distribution $P$ on $\mathbf{R}^d$ with density $p$. A delta-sequence estimator of $p(x)$ can be written in the form

$$\hat{p}_\lambda(x) = n^{-1} \sum_{i=1}^{n} \delta_{\lambda,x}(\xi_i)$$

where $\delta_{\lambda,x}$ is chosen so that the expected value of $\hat{p}_\lambda(x)$ converges to $p(x)$ uniformly in $x$ and $\lambda$, for $0 < \lambda \leqq \beta_n$ and $\beta_n \to 0$.

Examples of delta-sequence estimators include:

(2.1) Kernel estimators. Define for some kernel $K$: $\mathbf{R} \to \mathbf{R}$,

$$\delta_{\lambda,x}(\xi_i) = \lambda^{-1} K_{\lambda,x}(\xi_i)$$

where $K_{\lambda,x}(\xi_i) = K\left(\frac{|x-\xi_i|}{\lambda^{1/d}}\right)$ and $\int K = 1$.

(2.2) Histogram estimators. Histograms provide simple examples of delta-sequence estimators. Consider the histogram with equal binwidths, in one dimension. Let $I_j(\cdot)$ be the indicator function for the bin $[(j-1)\lambda, j\lambda)$, $j \in N$ and $\lambda > 0$. Then

$$\delta_{\lambda,x}(\xi_i) = \sum_{-\infty}^{\infty} \lambda^{-1} I_j(x) I_j(\xi_i).$$

The extension to higher dimensions is straightforward.

Two estimators closely related to the histogram are: the histospline estimator of Boneva, Kendall and Stefanov (1971), Wahba (1971, 1975), Van Ryzin (1973), and Scott (1985a), and the average shifted histogram of Scott (1985b). These are not explicitly treated here.

(2.3) Orthogonal series estimators. Consider the orthogonal series estimator for the continuous density $p$ on $[a, b]$. In this case,

$$\delta_{m,x}(\xi_i) = \sum_{j=1}^{m} \Phi_j(\xi_i) \Phi_j(x),$$

where $\{\Phi_j\}$ is a complete orthonormal system on $[a, b]$, consisting of eigenfunctions of a compact operator on $L^2[a, b]$. Notice, for convenience we allow the slight abuse of notation $\delta_{m,x}$ rather than $\delta_{m^{-1},x}$. This estimator is an approxima-

tion for $\sum_{j=1}^{\infty} c_j \Phi_j(x)$ where the infinite sum over $\{\Phi_j(x)\}$ is approximated by the sum over the first $m$ eigenfunctions, and where $c_j$, the inner product of $p$ and $\Phi_j$, is approximated by $n^{-1} \sum_{i=1}^{n} \Phi_j(\xi_i)$.

Whether $\lambda$ is automatic or not, these three estimators have the intuitively unappealing feature of doing the same amount of smoothing at each location. There are several ways to relax this assumption which we illustrate through modifications of the kernel estimator.

(2.4) Location-adaptive estimators. Define

$$\delta_{\hat{\lambda},x}(\xi_i) = \hat{\lambda}(x)^{-1} K_{\hat{\lambda}(x),x}(\xi_i)$$

where $\hat{\lambda}(x)$ is a random function of location. One example treated here is where $\hat{\lambda}(x)$ is the distance from $x$ to its $k^{\text{th}}$ nearest neighbor among $\xi_1, \ldots, \xi_n$ (Fix and Hodges, 1951; Loftsgaarden and Quesenberry, 1965; Mack and Rossenblatt, 1979). Another example is a plug-in estimate of the pointwise optimal bandwidth; see Woodroofe (1970), Krieger and Pickands (1981), Hall (1983b), and Muller and Stadtmuller (1987). One more possibility, not explicitly treated here, is local cross-validation, see Hall and Shucany (1988), Mielniczuc, Sarda and Vieu (1988) and Vieu (1988). More specifically, note that for a particular location $x$, the minimizing bandwidth for mean square errors, provided $p$ is twice differentiable, is asymptotic to

$$\left[\frac{c_K p(x) n^{-1}}{V^2 p(x)}\right]^{1/d+4}$$

where $c_K = d \int K^2 / \int y^2 K$. This motivates plugging in a pilot estimate of $p(x)$ and $V^2 p(x)$ to produce a random location-adaptive estimator.

The location-adaptive version of (2.2), the histogram of equal bin counts, is also examined in Sect. 4.

(2.5) Location-adaptive estimators indexed by the observations. Define

$$\delta_{\hat{\lambda},x}(\xi_i) = \hat{\lambda}(\xi_i)^{-1} K_{\hat{\lambda}(\xi_i),x}(\xi_i),$$

where the random scale parameter is a function of the location of the observations. For example, $\hat{\lambda}(\xi_i)$ may be either the distance from $\xi_i$ to its $k^{\text{th}}$ nearest neighbor among the observations (Breiman et al., 1977), or an estimate of $n^{\alpha} p(\xi_i)^{-1/2}$ for some $\alpha > 0$ (Hall and Marron, 1988).

In addition, let $\hat{\lambda}(x, \xi_i) = n^{\alpha} \tilde{p}(x)^{1/d - 1/2} \tilde{p}(\xi_i)^{1/2}$, where $\tilde{p}$ is a pilot estimate of $p$, to produce a hybrid of the two location-adaptive estimators (Abramson, 1982a,b).

## 3. Statement of Results

We will prove that for $\Delta_n = \{\delta_{\lambda,x} : \alpha_n \leqq \lambda \leqq \beta_n, x \in R^d\}$

$$\sup_{\Delta_n} |\hat{p}_{\lambda}(x) - p(x)| \to 0 \quad \textit{almost surely.}$$

With the supremum over both $\lambda$ and $x$ we can establish consistency for location-adaptive estimators, such as (2.4) and (2.5). Convergence for random parameters, such as automatic or data-based choices for $\lambda$, also easily follow from this result.

Split the difference above into the bias $|p_\lambda(x)-p(x)|$ and the variance $|\hat{p}_\lambda(x)-p_\lambda(x)|$, where $p_\lambda(x)=P\delta_{\lambda,x}$. We use linear functional notation to express $\int\delta_{\lambda,x}(y)P(dy)$ as $P\delta_{\lambda,x}$ and $1/n\sum_i\delta_{\lambda,x}(\xi_i)$ as $P_n\delta_{\lambda,x}$. Theorem 1, below, provides the uniform convergence result for the variance term. Its proof relies on finding an approximating class for $\Delta=\{\delta_{\lambda,x}\colon \lambda>0 \text{ and } x\in\mathbf{R}^d\}$ that replaces the supremum over $\lambda$ and $x$ by a maximum over a smaller collection of strategically chosen functions. To handle the wide variety of estimators introduced in Sect. 2, we present two techniques for constructing this approximation: bracketing and covering.

The bracketing technique takes advantage of smoothness assumptions on $\delta$. If $\delta_{\lambda,x}$ is Lipschitz in $\lambda$ and $x$ then brackets can be easily found. Orthogonal series estimators provide examples in the next section. The bracketing technique bounds each $\lambda\delta_{\lambda,x}$ above and below by a pair of functions which are close in an $L^1(P)$ sense. That is, each $\lambda\delta_{\lambda,x}$ must have a bracket $f^l\leqq\lambda\delta_{\lambda,x}\leqq f^u$ where $P|f^u-f^l|$ is at most $\varepsilon$, say. For given $\varepsilon$, the collection of brackets is denoted $B(\varepsilon)$. Theorem 1 imposes conditions on the cardinality of the bracketing class, $\#B(\varepsilon)$, called the bracketing number.

The second technique uses combinatorial methods to approximate $\Delta$ (Dudley 1978; Pollard 1984). Here, a collection of functions $\Delta(\varepsilon, P_n)$ is found where each $\lambda\delta_{\lambda,x}$ is within $\varepsilon$ in an $L^1(P_n)$ sense of a member of $\Delta(\varepsilon, P_n)$. Notice $\Delta(\varepsilon, P_n)$ is a random subclass because $P_n$ depends on $\xi_1, \ldots, \xi_n$. However, for the delta-sequence families of interest, there is a constant bound on $\#\Delta(\varepsilon, P_n)$, regardless of $n$.

**(1) Theorem.** *Let $p$ be a bounded density on $\mathbf{R}^d$, and let $\Delta_n$ be a collection of delta-sequence functions. If*

(i) $n\alpha_n/\log n\to\infty$,

(ii) $\sup_\Delta P|\delta_{\lambda,x}|$ *is bounded,*

(iii) $\sup_\Delta\sup_y|\lambda\delta_{\lambda,x}(y)|$ *is bounded,*

(iv) *either* $\log\#B(\varepsilon)=O(\log\varepsilon^{-1})$ *or* $\log\#\Delta(\varepsilon, P_n)=O(\log\varepsilon^{-1})$, *as* $\varepsilon\to 0$ *then*

$$\sup_{\Delta_n}|\hat{p}_\lambda(x)-p_\lambda(x)|\to 0 \quad \textit{almost surely.}$$

The conditions on $p$ for the convergence of the variance component are quite weak, but the bias component requires more from the density. These additional requirements appear in Sect. 4 with the examples. Condition (iv) places a bound on either the bracketing or covering number. The classes $B(\varepsilon)$ and $\Delta(\varepsilon, P_n)$ are approximations to the entire class $\Delta$, not $\Delta_n$. However, uniform convergence of $\hat{p}_\lambda(x)$ to $p(x)$ is on the restricted subset $\Delta_n$.

In the examples that use the covering technique, the class $\Delta$ has the property that

$$\Delta(\varepsilon, Q)\leqq A\varepsilon^{-V}$$

for constants $A$ and $V$ that depend only on $\sup_{\Delta,y}|\lambda\delta_{\lambda,x}(y)|$, not on $\varepsilon$ or the measure $Q$. This property was dubbed "Euclidean" in Nolan and Pollard (1987). Clearly, if $\{\lambda\delta_{\lambda,x}\}$ is Euclidean then it meets the covering number condition on its cardinality. The covering number condition could be weakened, but our applications do not require it. Techniques for showing specific collections of delta functions are Euclidean are delayed until Sect. 4.

To prove Theorem 1, we treat the variance term as an empirical process indexed by $\Delta$. To see this, reexpress $\hat{p}_\lambda(x)-p_\lambda(x)$ as $(P_n-P)\delta_{\lambda,x}$ where $P_n$ represents the empirical measure which places mass $n^{-1}$ on each of the observations $\xi_1, \ldots, \xi_n$. Empirical process techniques are employed in Sect. 5 to prove this result.

The following corollaries are immediate consequences of Theorem 1. Corollary 2 contains results for the random location-adaptive smoothing parameter. The convergence of the bias and variance are incorporated into the corollary. To do this, we impose the extra conditions that $\beta_n$, the upper bound on $\lambda$, tends to 0 and that $p$ is uniformly continuous. Corollary 3 is needed for the location-adaptive estimator where $\lambda$ is a function of the observations. In this case, the estimator no longer belongs to the delta-sequence family, but this presents no problem given the continuity-like conditions on $\delta_{\lambda(\cdot)}$.

**(2) Corollary.** *Let $p$ be a uniformly continuous bounded density on $R^d$. Let $\hat{\lambda}$ be a random scale parameter, possibly a function of $x$. If $\beta_n \geqq \hat{\lambda}(x) \geqq \alpha_n$ for all $x$, eventually, almost surely; if $\beta_n \to 0$; and if the conditions of Theorem 1 hold, then*

$$\sup_x |\hat{p}_{\hat{\lambda}(x)}(x)-p(x)| \to 0 \quad \textit{almost surely}$$

**(3) Corollary.** *Suppose $p_{\sigma(x)}$ is a uniformly consistent delta-sequence estimator for $p$. Let $H(\gamma)=\{x: p(x)\geqq\gamma\}$ and let $S_{x,r}$ be the indicator function for the sphere centered at $x$ with radius $r$. Consider the function $\delta_{\lambda(\cdot),x}(\cdot)$. If, given $\varepsilon$, there exist $\gamma$ and $r$ such that*

(i) $\sup_{x\in H(\gamma),y} |\delta_{\lambda(y),x}(y)-\delta_{\sigma(x),x}(y)S_{x,r}|<\varepsilon$ *a.s.*

(ii) $\sup_{H(\gamma)^c} |n^{-1}\sum_{i=1}^{n} \delta_{\lambda(\xi_i),x}(\xi_i)|<\varepsilon$ *a.s.*

*then*

$$\sup_x |n^{-1}\sum_{i=1}^{n} \delta_{\hat{\lambda}(\xi_i),x}(\xi_i)-p(x)| \to 0 \quad a.s.$$

## 4. Examples

To establish uniform consistency for a specific estimator, show the bias is negligible and bound the size of the approximating subclass (condition (iv) of Theo-

rem 1). As mentioned earlier, we handle the bias by whatever adhoc method is appropriate for the example. As for the approximation problem, we rely on either the bracketing or the covering approach outlined in Sect. 3. For the covering approach, we bound $\{\delta_{\lambda,x}\}$ by showing it is a Euclidean class; techniques presented in Nolan and Pollard (1987) for establishing the Euclidean property for classes of functions are referred to as needed. Alternatively, the bracketing approach takes advantage of smoothness assumptions on $\delta$.

### *4.1. Kernel Estimators*

Recall the kernel estimator defined in (2.1). According to Lemma 22 of Nolan and Pollard (1987), the collection of translations and rescalings of a function of bounded variation is Euclidean. Any bounded, absolutely integrable kernel function of bounded variation, meets all the conditions of Theorem 1. Then $\hat{p}$ is consistent for a uniformly continuous $p$, if we restrict $\lambda$ to lie between $\alpha_n$ and $\beta_n$ with $n\alpha_n/\log n \to \infty$ and $\beta_n \to 0$. Corollary 2 implies that $\hat{p}_{\hat{\lambda}}$ is consistent for random $\hat{\lambda}$, such as that chosen by cross-validation. See Hall (1983a) and Marron (1985, 1987) for access to the literature on this topic.

### *4.2. Histogram Estimators*

Here, we change notation slightly from the definition of the histogram estimator in (2.2). Let $I_{\lambda,x}$ be the indicator function for the bin $[(j-1)\lambda, j\lambda)$ that contains $x$, and recast the histogram as: $\lambda^{-1} P_n I_{\lambda,x}$. For any uniformly continuous $p$, the bias converges to 0 uniformly over $\beta_n \geqq \lambda > 0$ if $\beta_n \to 0$. With the additional restriction that $\lambda \geqq \alpha_n$ and $n\alpha_n/\log n \to \infty$, we only need to check either the bracketing or the covering number bound to obtain consistency uniformly over $x$ and $\alpha_n \leqq \lambda \leqq \beta_n$.

It is easy to bracket this collection of indicator functions. Given $\varepsilon$, the indicator function for $\{|y| > c\}$, where $c$ is chosen such that $\int_{-c}^{c} p > 1 - \frac{1}{2}\varepsilon$, can bracket $I_{\lambda,x}$ for locations $x$ in the tails of the density. Next, find $\lambda^*$ small enough to bound $\sup_x P I_{\lambda^*,x}$ by some fraction of $\varepsilon$. These indicators together with a subset of their unions can bracket the remaining $I_{\lambda,x}$. Altogether, there are at most a constant multiple of $\varepsilon^{-2}$ brackets.

The covering number approach also provides a simple solution. Consider the collection $\mathbf{A}$ of intervals on the real line. This collection has the special property that there exists a polynomial $\rho(\cdot)$ for which

$$\#\{A \cap F : A \in \mathbf{A}\} \leqq \rho(\#F)$$

for any finite subset $F$ of $\mathbf{R}$. That is, the intersections of $F$ with intervals in $\mathbf{A}$ produce at most $\rho(\#F)$ distinct subsets of $F$. The indicator functions for a collection of sets with this property is a Euclidean class (Dudley, 1978;

Nolan and Pollard, 1987, Lemma 19). Because the class of indicator functions for the half open intervals on $\mathbf{R}$ contains $\{I_{\lambda,x}\}$, it follows that $\Delta$ is Euclidean, and therefore, the covering number bound is met.

The class of indicator functions for $\mathbf{A}$ also contains the indicator functions for histograms with equal bin counts. First, we formally define this histogram estimator. Restrict $p$ to the interval $[-1, 1]$. Choose $k(n)$ to divide $n$ and define $\lambda$ as $k(n)/n$. Let $\xi_{(i)}$ be the $i^{\text{th}}$ order statistic from the sample $\xi_1, \ldots, \xi_n$. Consider the intervals: $(\xi_{(kj-k)}, \xi_{(kj)}]$ for $j=2, 3, \ldots, \lambda^{-1}$, and the end intervals $[0, \xi_{(1)})$, $[\xi_{(1)}, \xi_{(k)}]$, $(\xi_{(\lambda^{-1})}, 1]$ for $j=0, 1, \lambda^{-1}+1$, respectively. Define $B_{x,\lambda}$ as the indicator function for the interval that contains $x$. Then our estimator is $n^{-1}\sum_j B_{x,\lambda} V(B_{x,\lambda})^{-1}$ where $V(B_{x,\lambda})$ is the length of the corresponding interval, $\int B_{x,\lambda}(y)\,dy$. This collection of indicator functions for intervals of random widths meets the covering number condition of Theorem 1 because it is a subset of the collection of indicator functions for half open intervals. We postpone the verification that $\inf_{2 \leqq j \leqq \lambda^{-1}} (\xi_{(kj)} - \xi_{(kj-k)})\, n/\log n \to \infty$ until the tools developed for the nearest neighbor estimator are available.

### 4.3. *Orthogonal Series Estimators*

For simplicity of presentation, we treat only the cosine series where $\Phi_j = \cos \pi j(\cdot)$ and $p$ is continuous, symmetric on $[-1, 1]$ and of bounded variation. Then, for $m \geqq 1$,

$$\delta_{m,x}(y) = \sin \pi(m+\tfrac{1}{2})(x-y)/2 \sin \frac{\pi}{2}(x-y).$$

By construction, the bias is negligible for the collection of orthogonal series estimators: $\Delta_n = \{\delta_{m,x}: x \in [-1, 1], L_n \leqq m \leqq U_n\}$, if $L_n \to \infty$. Also, if $n/U_n \log n \to \infty$ then, because $\Delta_n$ is a collection of smooth functions, all that needs to be checked is the bracketing number condition.

For large $m$, we can bracket $m^{-1}\delta_{m,0}$ above by

$$f^u = 1 \qquad \text{for } |y| < 1/N$$
$$\frac{1}{m}\left[\sin \frac{\pi}{2} y\right]^{-1} \text{ otherwise.}$$

The constants $M$ and $N$ are chosen to make $\int f^u p$ small. This bracket works for all $\delta_{m,x}$ with $x$ near 0. Similarly, brackets can be found for the remaining locations. As for $m$ small, the differentiability of cosine offers a crude bracket in $x$ for each $m$:

$$m^{-1} \sum_{j=1}^{m} \cos \pi j(x+h-y) < m^{-1} \sum_{j=1}^{m} \cos \pi j(x-y) + Mh < m^{-1}\delta_{m,x} + \varepsilon.$$

where $h \leqq \varepsilon M^{-1}$. Combine the brackets in $x$ for large and small $m$ to construct $B(\varepsilon)$ with cardinality at most a multiple of $\varepsilon^{-5}$. The bracketing condition is met; we have the desired consistency result.

By Corollary 2, any random choice for $m$ that falls between $L_n$ and $U_n$, almost surely, provides a consistent estimator for $p$. See Hall (1987) for possible examples.

### 4.4. *Location-Adaptive Estimators*

Theorem 1, or Corollary 2, is applicable to the nearest neighbor estimator, which is a version of (2.4). First consider the uniform kernel with unit support. That is, let $S_{x,k}$ be the indicator function for the sphere centered at $x$ that contains $k(n)$ observations, and let $r_k(x)$ be the radius of this sphere and $V(S_{x,k})$ be its volume. Then $V(S_{x,k})^{-1} P_n S_{x,k}$ is a kernel estimator with a random location-adaptive scale parameter; call it $\hat{p}(x)$. Note, $P_n S_{x,k} = k/n$ for all $x$. Consistency of this estimator is not a direct consequence of Theorem 1, because in regions of low density the distance $r_k(x)$ need not shrink as $n$ increases, even when we require $\frac{k}{n} \to 0$. However, Theorem 1 does provide the consistency result indirectly. See Devroye and Wagner (1980) for another version of this proof.

Uniform continuity of the density is necessary to ensure $\hat{p}(x)$ remains small in low density regions. Without loss of generality suppose $\sup_x p(x) \leqq 1$. Consider the event $\{\hat{p}(x) - p(x) > \varepsilon \text{ for some } x\}$, or equivalently, the event $\{V(S_{x,k}) < \frac{k}{n}[p(x)+\varepsilon]^{-1} \text{ for some } x\}$. It is a subset of

$$\left\{V(S_{x,k})[p(x)+\tfrac{1}{2}\varepsilon] < \frac{k}{n}\left[1 - \frac{1}{2}\varepsilon(p(x)+\varepsilon)\right],\ V(S_{x,k}) < \frac{k}{n}\varepsilon^{-1} \text{ for some } x\right\}$$

$$\subset \left\{V(S_{x,k})[p(x)+\tfrac{1}{2}\varepsilon] < P_n S_{x,k} - \frac{1}{4}\varepsilon\frac{k}{n},\ V(S_{x,k}) < \frac{k}{n}\varepsilon^{-1} \text{ for some } x\right\}$$

$$\subset \left\{(P_n - P) S_{x,k} > \frac{1}{4}\varepsilon\frac{k}{n},\ V(S_{x,k}) < \frac{k}{n}\varepsilon^{-1} \text{ for some } x\right\}.$$

The last inclusion follows from the uniform continuity of $p$, and the shrinking of the spheres $\{S_{x,k}\}$. Therefore $\{\hat{p}(x) - p(x) > \varepsilon \text{ for some } x\}$ is contained in

$$\left\{(P_n - P) S_{x,k} > \frac{\varepsilon k}{4n} \text{ for some } x \text{ with } V(S_{x,k}) < \varepsilon\frac{k}{n}\right\}.$$

Now apply Theorem 1 with $\delta_{\lambda,x} = S_{x,k}/\lambda_n$ and $\lambda_n(x) = \frac{k}{n} + V(S_{x,k})$. (Theorem 1 still applies even though $\hat{p}$ is not quite a density estimate.) Check the nonobvious conditions of the theorem. If $k(n)/\log n \to \infty$ then condition (i) is met. That leaves condition (iv). In (4.2) we saw that Lemma 22 of Nolan and Pollard (1987)

implied the collection of indicator functions for the intervals in $\mathbf{R}$ is Euclidean. Similarly, the collection of indicator functions for spheres in $\mathbf{R}^d$ is also Euclidean, and so the subset of indicator functions for the spheres with volume less than $\varepsilon \frac{k}{n}$ is Euclidean. This convergence result implies $\sup_{x,k(n)} \hat{p}(x) - p(x) < \varepsilon$ almost surely. A similar argument works for showing $\inf_{x,k(n)} \hat{p}(x) - p(x) > -\varepsilon$ almost surely. Together they give uniform consistency.

A sketch of the proof for more general kernels now follows. We require that $K$ be bounded and that: $K(|x|) = 0$ for $|x| > 1$. Without loss of generality, take $\sup_x K(x) \leqq 1$. Restrict attention to a compact region $H_\varepsilon = \{x: p(x) \geqq \varepsilon\}$. The previous result claims that $\hat{\lambda}(x) = V(S_{x,k})$ meets the conditions of Theorem 1 on $H_\varepsilon$, eventually, almost surely. As for $x$ in the complement of $H_\varepsilon$, use the crude upper bound of $2\varepsilon$ for $\hat{p}(x)$.

Abramson (1984) proposed using a metric other than Euclidean distance to determine $V(S_{x,k})$. Let $\rho$ be a metric on $\mathbf{R}^d$ and let $\eta_k(x)$ be the $\rho$-radius of the smallest $\rho$-sphere centered at $x$ that contains $k$ observations. Then the kernel estimate becomes

$$n^{-1} \sum_{i=1}^{n} V(S_{x,\eta_k(x)})^{-1} K\left(\frac{\rho(x-\xi_i)}{\eta_k(x)}\right).$$

Abramson (1984) considered a metric of the form $\rho(x) = \rho_1(x_1) \ldots \rho_d(x_d)$ where the $x_j$ are the components of $x$ and $\rho_j$ is an invertible distortion of the data. Review the previous proof for dependences on Euclidean distance: place the additional assumptions on $p$ that it be uniformly continuous with respect to $\rho$, and insist $\rho(x-y)$ shrink to 0 continuously as $|x-y| \to 0$. This estimator is uniformly consistent as well.

Hall's (1983b) version of the plug-in estimator (2.4), based on nearest neighbor distances, is a refinement of the nearest neighbor estimator. With a slight modification that protects the pilot estimates of $p(x)$ and $V^2 P(x)$ employed in $\hat{\lambda}(x)$ from exploding in some locations, the methods developed here will show that this estimator is uniformly consistent for all uniformly continuous densities regardless of differentiability conditions.

### *4.5. Location-Adaptive Estimators Indexed by the Observations*

The Breiman, Meisel, and Purcell (1977) estimator is defined as follows:

$$\hat{p}(x) = n^{-1} \sum_{i=1}^{n} V(S_{i,k})^{-1} K\left(\frac{|x-\xi_i|}{r_k(\xi_i)}\right).$$

The symbol $S_{i,k}$ represents the indicator of the sphere centered at $\xi_i$ that contains $k$ observations, and $r_k(\xi_i)$ is the radius of this sphere. Unlike the nearest neighbor estimator, this estimator is itself a density. But, because the scale parameter

varies with each observation, we need Corollary 3 for consistency. Condition (i) permits the replacement of $r_k(\xi_i)$ by $r_k(x)$ for observations near $x$ and it prohibits a large contribution from the remaining observations not near $x$. Condition (ii) ensures a crude upper bound on the estimator in the tails of the density. The following proof appears in Nolan (1984). See also Devroye and Penrod (1984b) for another version of this proof.

To check these two conditions, we again rely on the previous result for the $k^{\text{th}}$-nearest neighbor estimate with uniform kernel. Given $\varepsilon$, recall $H_\varepsilon$ is the compact region $\{p(x) \geqq \varepsilon\}$. Choose $\delta$ such that if $|x-y|<\delta$ then $|p(x)-p(y)|<\varepsilon^2$. For (i), if $n$ is large enough then, uniformly over the buffer region $H_\varepsilon - H_{2\varepsilon}$, the radii of any $k^{\text{th}}$-nearest neighbor ball is less than $\frac{1}{4}\delta$, almost surely. This implies that, uniformly over $H_\varepsilon^c$, all $k^{\text{th}}$-nearest neighbor balls do not intersect $H_{2\varepsilon}$. Also, on $H_\varepsilon$, for $n$ large enough,

$$\gamma_n[p(\xi_i)+\varepsilon^2]^{-1} < V(S_{i,k}) < \gamma_n[p(\xi_i)-\varepsilon^2]^{-1}$$

where $\gamma_n = \dfrac{k}{n}$. Combine these two results with the uniform continuity of $p$ to bound $\hat{p}(x)$ on $H_{2\varepsilon}$ above by

$$(1+4\varepsilon)\sum_{i=1}^{n} \gamma_n^{-1}[p(x)+2\varepsilon^2]\, K\left(\frac{|x-\xi_i|}{\gamma_n^{1/d}[p(x)+2\varepsilon^2]^{-1/d}}\right) S_{x,\delta}(\xi_i).$$

A similar lower bound is available for $\hat{p}(x)$. Apply Theorem 1 with $\sigma_n(x) = \gamma_n[p(x)+2\varepsilon^2]^{-1}$ and $x \varepsilon H_{2\varepsilon}$.

Finish the argument by checking (ii) of Corollary 3. That is, bound the estimator for $x$ on $H_{2\varepsilon}^c$. Argue as above to conclude that eventually no observation in $H_{4\varepsilon}$ will contribute to $\hat{p}(x)$ for $x \varepsilon H_{2\varepsilon}^c$. Now invoke a property of nearest neighbors which holds regardless of the unknown distribution: any $x$ can belong to at most $c_d k$ of the $k^{\text{th}}$-nearest neighbor spheres centered on the $n$ observations. In one dimension $c_1 = 2$, and in two dimensions $c_2 = 6$. Bound $\hat{p}(x)$ by $2c_d\varepsilon$ for $x$ in $H_{2\varepsilon}^c$ to complete the proof.

Finally, with Corollary 3, we have the tools to evaluate the consistency of the hybrid estimator of Abramson (1982a, b) defined in (2.5) as well.

## 5. Proof of Theorem 1

Consider the two one-sided inequalities separately,

$$\mathbf{P}\{\sup_{A_n}(P_n - P)\,\delta_{\lambda,x} > \varepsilon\},$$

$$\mathbf{P}\{\sup_{A_n}(P_n - P)\,\delta_{\lambda,x} < -\varepsilon\}.$$

Use (ii) to bound the first inequality above by

$$\mathbf{P}\{\sup_{\Delta_n}(P_n-P)\lambda\delta_{\lambda,x}/C_1[\alpha_n+P|\lambda\delta_{\lambda,x}|]>\varepsilon\} \tag{1}$$

for some constant $C_1$.

Suppose the bracketing condition in (vi) holds. Then the collection $B(\frac{1}{4}\varepsilon\alpha_n)$ of upper and lower brackets $g^l$ and $g^u$ for $\{\lambda\delta_{\lambda,x}\}$ exists, and in addition, (iii) allows us to assume the elements of $B(\frac{1}{4}\varepsilon\alpha_n)$ are bounded by some constant $C_2$, say. Then

$$(P_n-P)\lambda\delta_{\lambda,x}\leqq(P_n-P)g^u+\tfrac{1}{4}\varepsilon\alpha_n$$

and

$$P|\lambda\delta_{\lambda,x}|\geqq P|g^u|-\tfrac{1}{4}\varepsilon\alpha_n.$$

Apply these two inequalities to the one sided probability above,

$$\begin{aligned}&\mathbf{P}\{\sup_{\Delta_n}(P_n-P)\lambda\delta_{\lambda,x}/[\alpha_n+P|\lambda\delta_{\lambda,x}|]>C_1\varepsilon\} \\ &\leqq\mathbf{P}\{\max_{B(\varepsilon\alpha_n)}[(P_n-P)g^u+\tfrac{1}{4}\varepsilon\alpha_n]/[(1-\tfrac{1}{4}\varepsilon)\alpha_n+P|g^u|]>C_1\varepsilon\} \\ &\leqq\#B(\varepsilon\alpha_n)\max_{B(\varepsilon\alpha_n)}\mathbf{P}\{(P_n-P)g^u>\tfrac{1}{2}C_1\varepsilon\alpha_n+C_1\varepsilon P|g^u|\}.\end{aligned} \tag{2}$$

An application of Bernstein's inequality (see B.4 of Pollard, 1984) for the bounded random variables $g^u(\xi_i)-Pg^u$ gives, for some constant $C_3$,

$$\begin{aligned}&\mathbf{P}\left\{\sum_{i=1}^{n}(g^u(\xi_i)-Pg^u)>\tfrac{1}{2}C_1\varepsilon n(\alpha_n+P|g^u|)\right\} \\ &\leqq\exp[-\tfrac{1}{8}C_1n^2\varepsilon^2(\alpha_n+P|g^u|^2+\tfrac{1}{6}C_2n\varepsilon(\alpha_n+P|g^u|))] \\ &\leqq\exp[-nC_3\alpha_n].\end{aligned} \tag{3}$$

Finally, the condition on the cardinality of $B(\frac{1}{4}\varepsilon\alpha_n)$ gives a finite upper bound for the sum

$$\sum_{n=1}^{\infty}P\{\sup_{\Delta_n}(P_n-P)\delta_{\lambda,x}>\varepsilon\}\leqq\sum_{n=1}^{\infty}\#B(\varepsilon\alpha_n)\exp(-C_3n\alpha_n).$$

The Borel-Cantelli lemma completes the argument for one side of the inequality. The other side follows by symmetry, for $(P_n-P)\lambda\delta_{\lambda,x}>(P_n-P)g^l-\frac{1}{4}\varepsilon\alpha_n$.

The proof under the alternative covering number condition is similar. The differences follow from the empirical process techniques used to approximate $\Delta$; for an exposition of these techniques see Pollard (1984). We briefly outline the proof here; it is a special case of Theorem 2 in Pollard (1986). For simplicity, assume $\delta_{\lambda,x}$ is nonnegative. Bound the tail probability in (1) by a tail probability of the form:

$$\mathbf{P}\{\sup_{\Delta}|(P_n-P_n')\lambda\delta_{\lambda,x}|/[\alpha_n+(P_n+P_n')\lambda\delta_{\lambda,x}]>\varepsilon\}.$$

The $P_n'$ represents the empirical measure constructed from a second sample $\eta_1,\ldots,\eta_n$ on $P$, independent of the first sample $\xi_1,\ldots,\xi_n$. Next replace the

supremum over $\Delta$ by a maximum over $\Delta(1/4\varepsilon\alpha_n, P_n + P_n')$. In this case, the approximating class is random, so the rest of the argument is worked conditionally. The quantity $(P_n - P_n')\delta_{\lambda,x}$ has the same distribution as $1/n\Sigma\sigma_i[\delta_{\lambda,x}(x_{2i}) - \delta_{\lambda,x}(x_{2i-1})]$ where $\sigma_i = \pm 1$ with probability $\frac{1}{2}$ and is independent of the sample $x_1, \ldots, x_{2n}$ from $P$. Condition on the double sample. The approximating class is no longer random, and as in (2), the maximum can be moved outside the probability. Follow the steps outlined in (3) with Hoeffding's inequality applied to $\{a_i\sigma_i\}$ instead of Bernstein's inequality applied to $\{g^u(\xi_i) - Pg^u\}$; the $a_i$ are the constants $\lambda[\delta_{\lambda,x}(x_{2i}) - \delta_{\lambda,x}(x_{2i-1})]$. The proof finishes with the Borel-Cantelli lemma as above.

*Acknowledgements.* The authors are grateful to the referee for a number of helpful suggestions and new references.

## References

Abramson, I.S.: Arbitrariness of the pilot estimator in adaptive kernel methods. J. Multivariate Anal. **12**, 562–567 (1982a)

Abramson, I.S.: On bandwidth variation in kernel estimates – a square root law. Ann. Stat. **10**, 1217–1223 (1982b)

Abramson, I.S.: Adaptive density flattening – a metric distortion principle for combating bias in nearest neighbor methods. Ann. Stat. **12**, 880–886 (1984)

Bertrand-Retali, M.: Convergence uniforme stochastique d'un estimateur d'une densité de probabilité dans $R^S$. C.R. Acad. Sci. Paris, Ser. A. **278**, 1449–1452 (1974)

Bleuez, J., Bosq, D.: Conditions nécessaire et suffisantes de convergence de l'estimateur de la densité par la méthode des fonctions orthogonales. C.R. Acad. Sc. Paris, Ser. A. **282**, 1023–1026 (1976)

Boneva, L., Kendall, D., Stefanov, I.: Spline transformations: Three new diagnostic aids for the statistical data analyst. J. R. Stat. Soc. B. **33**, 1–70 (1971)

Breiman, L., Meisel, W., Purcell, E.: Variable kernel estimates of probability densities. Technometrics **19**, 135–144 (1977)

Burman, P.: A data dependent approach to density estimation. Z. Wahrscheinlichkeitstheor. Verw. Geb. **69**, 609–628 (1985)

Devroye, L.: A note on the $L^1$ consistency of variable kernel estimates. Ann. Stat. **13**, 1041–1049 (1985)

Devroye, L., Györfi, L.: Nonparametric density estimation: The $L^1$ view. New York: Wiley 1984

Devroye, L., Penrod, C.S.: The consistency of automatic kernel density estimates. Ann. Stat. **12**, 1231–1249 (1984a)

Devroye, L., Penrod, C.S.: The strong uniform convergence of multivariate variable kernel estimates. Preprint 1984b

Devroye, L., Wagner, T.J.: The strong uniform consistency of kernel density estimates. In: Krishndiah, P.R. (ed.) Multivariate analysis V, pp. 59–77. New York Amsterdam: North Holland 1980

Dudley, R.M.: Central limit theorems for empirical measures. Ann. Probab. **6**, 899–929 (1978)

Fix, E., Hodges, J.L.: Discriminatory analysis, nonparametric discrimination, consistency properties. Randolph Field, Texas, Project 21-49-004, Report No. 4

Földes, A., Revez, P.: A general method for density estimation. Studia Sci. Math. Hungar. **9**, 81–92 (1974)

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

Hall, P.: On near neighbor estimates of a multivariate density. J. Multivariate Anal. **13**, 24–39 (1983b)

Hall, P.: On the rate of convergence of orthogonal series density estimators. J. R. Stat. Soc. B**48**, 115–122 (1986)

Hall, P.: Cross-validation and the smoothing of orthogonal series density estimators. J. Multivariate Anal. **21**, 189–206 (1987)

Hall, P., Marron, J.S.: Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation, Probab. Th. Rel. Fields **74**, 567–581 (1987a)

Hall, P., Marron, J.S.: On the amount of noise inherent in bandwidth selection for a kernel density estimator. Ann. Stat. **15**, 163–181 (1987b)
Hall, P., Marron, J.S.: Variable window width kernel estimates of probability densities. Probab. Th. Rel. Fields (Ms. 372)
Hall, P., Shucany, W.: A local cross-validation algorithm. Preprint 1988
Kim, B.K., Van Ryzin, J.: Uniform consistency of a histogram density estimator and model estimation. Comm. Stat. **4**, 303–315 (1975)
Krieger, A.M., Pickands, J. III: Weak convergence and efficient density estimation at a point. Ann. Statist. **9**, 1066–1078 (1981)
Loftsgaarden, D.O., Quesenberry, C.P.: A nonparametric estimate of a multivariate density function. Ann. Math. Stat. **36**, 1049–1051 (1965)
Mack, Y.P., Rosenblatt, M.: Multivariate $k$-nearest neighbor density estimates. J. Multivariate Anal. **9**, 1–15 (1979)
Marron, J.S.: An asymptotically efficient solution to the bandwidth problem of kernel density estimation. Ann. Stat. **13**, 1011–1023 (1985)
Marron, J.S.: A comparison of cross-validation techniques in density estimation. Ann. Stat. **15**, 152–162 (1987)
Mielniczuk, J., Sarda, P., Vieu, P.: Local data-driven bandwidth choice for density estimation. Preprint 1988
Muller, H.G., Stadtmuller, U.: Variable bandwidth kernel estimates of regression curves. Ann. Stat. **15**, 182–201 (1987)
Nolan, D.: Uniform convergence of variable kernel estimates. Unpublished prospectus. Yale University 1984
Nolan, D., Pollard, D.: U-processes: rates of convergence. Ann. Stat. **15**, 780–789 (1987)
Pollard, D.: Convergence of stochastic processes. New York Berlin Heidelberg: Springer 1984
Pollard, D.: Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions. Preprint 1987
Prakasa Rao, B.L.S.: Nonparametric functional estimation. New York: Academic Press 1983
Revesz, P.: On empirical density function. Period. Math. Hungar. **2**, 85–110 (1972)
Rudemo, M.: Empirical choice of histograms and kernel density estimators. Scand. J. Stat. **9**, 65–78 (1982)
Scott, D.W.: Frequency Polygons: theory and application, J. Am. Stat. Soc. **80**, 348–354 (1985a)
Scott, D.W.: Averaged shifted histograms: effective nonparametric density estimators in several dimensions. Ann. Stat. **13**, 1024–1040 (1985b)
Silverman, B.W.: Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. Ann. Stat. **6**, 177–184 (1978)
Silverman, B.W.: Density estimation for statistics and data analysis. New York: Chapman and Hall 1986
Stone, C.J.: An asymptotically optimal window selection rule for kernel density estimates. Ann. Stat. **12**, 1285–1297 (1984)
Stone, C.J.: An asymptotically optimal histogram selection rule. In: LeCam, L., Olshen, R.A. (eds.) Conference in honor of Jerzy Neyman and Jack Kiefer. Proceedings, Berkeley 1985. Vol. II, pp. 513–520. Wadsworth 1985
Tapia, R.A., Thompson, J.R.: Nonparametric probability density estimation. Baltimore: Johns Hopkins University Press 1978
Van Ryzin, J.: A histogram method of density estimation. Comm. Stat. **2**, 493–506 (1973)
Vieu, P.: Nonpaprametric regression: Optimal global bandwidth. Preprint (1988)
Wahba, G.: A polynomial algorithm for density estimation. Ann. Math. Stat. **42**, 1870–1886 (1971)
Wahba, G.: Interpolating spline methods for density estimation I. Equi-spaced knots. Ann. Stat. **3**, 30–48 (1975)
Wahba, G.: Data-based optimal smoothing of orthogonal series density estimates. Ann. Stat. **9**, 146–156 (1981)
Walter, G.G., Blum, J.: Probability density estimation using delta sequences. Ann. Stat. **7**, 328–340 (1979)
Woodroofe, M.: On choosing a delta sequence. Ann. Math. Stat. **41**, 1665–1671 (1970)

Received March 13, 1987; in revised form July 20, 1988