The Annals of Statistics
1983, Vol. 11, No. 1. 25–38

CONSISTENT CROSS-VALIDATED DENSITY ESTIMATION

BY Y.-S. CHOW, S. GEMAN, L.-D. WU

Academia Sinica, Brown University and Fudan University

Application of nonparametric density estimators generally requires the specification of a "smoothing parameter." The kernel estimator, for example, is not fully defined until a window width, or scaling, for the kernels has been chosen. Many "data-driven" techniques have been suggested for the practical choice of smoothing parameter. Of these, the most widely studied is the method of cross-validation. Our own simulations, as well as those of many other investigators, indicate that cross-validated smoothing can be an extremely effective practical solution. However, many of the most basic properties of cross-validated estimators are unknown. Indeed, recent results show that cross-validated estimators can fail even to be consistent for seemingly well-behaved problems. In this paper we will review the application of cross-validation to the smoothing problem, and establish L_1 consistency for certain cross-validated kernels and histograms.

1. The smoothing problem. Application of the histogram estimator of a density to a collection of observations requires first the specification of a "bin width". Consistent and efficient estimation is achieved by allowing the bin width to shrink appropriately to zero as the number of observations increases to infinity. But a knowledge of the unknown target density is necessary to fully specify an optimal rate of shrinkage. Very little is known about the proper choice of bin width when faced with a fixed and finite sample from a distribution with unknown density. An analogous situation prevails for virtually all estimators of infinite dimensional target parameters; application of the estimator to a real set of observations requires the specification of a (usually one-dimensional) "smoothing parameter", and very little is known about how this is most effectively and practically done.

Given a random sample $x_1, x_2, \cdots x_n$ from a distribution with unknown density f, the time honored Rosenblatt-Parzen kernel estimator is

$$f_{\lambda,n}(x) = \frac{1}{n}\sum_{i=1}^{n} \lambda K(\lambda(x - x_i)), \tag{1.1}$$

where K is a fixed probability density (typically the zero-mean Gaussian), and $1/\lambda$ is the "window width"—the smoothing parameter for this estimator. Much is known about the relation between the rate of convergence of $f_{\lambda,n}$ to f, and the asymptotic growth of the parameter λ as a function of sample size ($\lambda = \lambda_n$). But for fixed n, $f_{\lambda,n}$ is sensitive to λ, and there is no generally agreed upon method for choosing this critical parameter. It has been repeatedly observed that all of the commonly studied nonparametric density estimators suffer this same limitation: the generally well-understood relation between the asymptotics of the smoothing parameter and the convergence of the estimator do not provide a practical guide for the implementation of the estimator to real data. Regarding kernel estimators, Silverman (1978) observes that "there seems to be considerable need for objective methods of determining the window width appropriate to a given sample." Speaking more generally, Wahba (1981) remarks: "A major problem in density estimation is to choose the smoothing parameter(s), which are part of every density estimate. . . ." Thus the maximum penalized likelihood estimator (Good and Gaskins, 1972) requires that

Received December 1981; revised August 1982.

[1] This work was partially supported by the Department of the Army contract DAAG29-80-K-0006 and the U.S. Air Force grant AFOSR 78-3514.

AMS 1980 *subject classifications.* Primary, 62G05; secondary, 62A10.

Key words and phrases. Cross-validation, consistency, nonparametric density estimation.

a weight be assigned the penalty term; orthogonal series estimators must be suitably truncated (Kronmal and Tarter, 1968), or "band limited" (Wahba, 1981); and sieve estimators (Geman and Hwang, 1982, Grenander, 1981) must be assigned a sieve size. And the problem is not peculiar to density estimation. Splines, kernels, and the newer "recursive partitions" (see, for example, Gordon and Olshen, 1980) for nonparametric regression, and sieves (Grenander, 1981) as a general approach to estimating parameters in abstract spaces, all require first a version of smoothing to be fully defined. Indeed, the problem seems to be inherent in the non-Bayesian approach to infinite dimensional estimation.

It is natural to try to use the observations themselves to determine an appropriate degree of smoothing, and this general approach is known as "data-driven smoothing". Of the many forms of data-driven smoothing that have been proposed in the literature, probably the most versatile and widely studied is the method of cross-validation. We will review here the method of cross-validation, focusing particularly on its application to nonparametric density estimation. Then we will present results establishing the strong (almost sure) L_1 consistency of certain cross-validated kernels and histograms.

2. Data-driven smoothing by cross-validation. The general idea behind data-driven smoothing is to measure, as a function of the smoothing parameter, the ability of the estimator to "explain", or to "fit" the observed data. The smoothing parameter, denoted here by λ, is then chosen to maximize this measure of explanation. When smoothing by cross-validation, in particular, the measure of explanation is obtained by deleting single observations, computing the estimator from the remaining observations, and then applying the estimator to the deleted observation. The details are most easily illustrated with specific examples. For this purpose, let us return to the kernel estimator, $f_{\lambda,n}$, defined in (1.1). We will denote by $f^i_{\lambda,n-1}$ the estimator computed after deleting the ith observation, i.e.

$$f^i_{\lambda,n-1}(x) = \frac{1}{n-1}\sum_{j\neq i}\lambda K(\lambda(x - x_j)).$$

Now $f^i_{\lambda,n-1}$ is not dependent on x_i, and $f^i_{\lambda,n-1}(x_i)$ may be taken as a measure of the appropriateness of λ as a value for the smoothing parameter: If $f^i_{\lambda,n-1}(x_i)$ is large, then it might be said that $f^i_{\lambda,n-1}$ "anticipated" the observation x_i, and that λ is an appropriate degree of smoothing (at least for samples of size $n-1$); small values of $f^i_{\lambda,n-1}(x_i)$ suggest that the observation x_i was unlikely (under the density $f^i_{\lambda,n-1}$), and may be interpreted as evidence against the appropriateness of λ. As i ranges through the full sample we obtain n such measures of fit, and these may be combined into the likelihood-like expression

$$L_\lambda = \prod_{i=1}^n f^i_{\lambda,n-1}(x_i). \tag{2.1}$$

One version of cross-validated density estimation, first proposed by Habbema et al. (1977) and separately by Duin (1976), chooses λ to maximize L_λ (call this value λ^*, the "cross-validated smoothing parameter"), and then forms the corresponding estimator, $f_{\lambda^*,n}$ (the "cross-validated kernel estimator").

If, instead of the kernel estimator, $f_{\lambda,n}$ is defined by a histogram with bin width $1/\lambda$, i.e.

$$f_{\lambda,n}(x) = \frac{\lambda}{n}\sum_{j=-\infty}^{\infty}\chi_{\left[\frac{j-1}{\lambda},\frac{j}{\lambda}\right)}(x)\left\{\sum_{i=1}^n\chi_{\left[\frac{j-1}{\lambda},\frac{j}{\lambda}\right]}(x_i)\right\},$$

where χ_A denotes the indicator function of the set A, then exactly the same procedure defines a cross-validated smoothing parameter, and a resulting cross-validated estimator. $f_{\lambda,n}$ could in fact be any density estimator in which λ represents the degree of smoothing (possibly, λ is vector-valued); in each example, the same cross-validation method provides a natural and completely data-defined value for λ.

In Section 3 we will provide a more formal justification for the smoothing criterion L_λ.

And, in Section 4 we will state conditions under which, for kernels and histograms,

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} |f_{\lambda^*,n}(x) - f(x)|\, dx \to 0 \quad \text{a.s.}$$

Cross-validation, and more generally data-driven smoothing, has analogous applications to other estimation problems. Applications to ridge and nonparametric regression (see Craven and Wahba, 1979; Egerton and Laycock, 1981; Golub et al., 1979; Lawless, 1981; Utreras, 1979; Wahba and Wold, 1975), are probably the most widely studied and best understood. For these problems, as well as the density estimation problem, numerous simulations have repeatedly demonstrated good performance of estimators smoothed by cross-validation and closely related techniques. Some examples are in Duin (1976), Golub et al. (1979), Scott and Factor (1981), Utreras (1979), Wahba (1977), and Wahba and Wold (1975). However, support for the technique is qualified; many authors have expressed reservations based on various heuristic and analytic grounds, and some unsuccessful simulations (see Dawid, 1974; Egerton and Laycock, 1981; Hall, 1982; Lawless, 1981; and Stone, 1977). There is surprisingly little known *analytically* about the behavior of these estimators. For example, with the modest exception of the results presented in our Section 4, there is not a single instance in which even the consistency of a cross-validated estimator has been established for an infinite dimensional target parameter. (In fact, as discussed below, consistency is *not* guaranteed.) (If one places a priori bounds on the allowed range of the smoothing parameter λ, then consistency results are available. But for these estimators the range itself plays the role of smoothing parameter, and it cannot be said that they are fully data-defined.) There are interesting and important analytic results, mostly by Wahba and coauthors (see, for example, Craven and Wahba, 1979; Golub et al. 1979; and Wahba, 1977), which suggest that certain cross-validated estimators will perform well, but the fundamental properties of consistency, efficiency, and asymptotic distribution remain to be established.

There appears to be a real practical need for a better analytic understanding of estimators smoothed by data-driven techniques. Success or failure of these estimators can hinge on a subtle relation between the estimator being smoothed and the particular (unknown) target parameter. For example, it can be demonstrated that the cross-validated histogram is not consistent for any density in which the difference between the last two order statistics diverges (call this difference Δ_n). The reason is this: If $x_i = \max(x_1, x_2, \ldots x_n)$, then $f^i_{\lambda,n-1}(x_i)$ (and hence L_λ) is 0 whenever $1/\lambda < \Delta_n$. Consequently, the cross-validated bin width, $1/\lambda^*$, is no smaller than Δ_n, and must therefore *diverge*.

The same argument demonstrates the inconsistency of cross-validated kernel estimators, when the kernel has compact support and when the target density has the property $\Delta_n \to \infty$ (because then the cross-validated window width also diverges). Schuster and Gregory (1981) have derived precise tail conditions on the target density under which compact kernel density estimators are not consistent. The circumstances can be surprisingly innocent: as they point out, the tail of the exponential density is "too heavy" for consistent cross-validated kernel estimation, when using compact kernels. (Schuster and Gregory suggest replacing the ordinary kernel estimator (1.1) by the Breiman et al. (1977) variable-width kernel. In their simulations, a cross-validated version of the latter nicely estimated the Cauchy density, whereas the cross-validated ordinary kernel estimator drastically oversmoothed, even when employing a kernel with infinite support.) It is a corollary, of sorts, that cross-validated kernel estimators employing kernels with "light tails" will oversmooth densities with heavy tails. And, although we have no proof, our analysis suggests an obverse: kernels with heavy tails will *undersmooth* densities with light tails (such as densities with compact support). We believe the latter because in these situations we have been unable to effectively upper bound the rate of growth of λ^*; it appears that the window width, $1/\lambda^*$, may converge very rapidly to 0. Hall's (1982) recent paper suggests

a similar conclusion: cross-validation tends to undersmooth when restricted to finite intervals on which the target density is smooth and bounded away from zero.

3. Some heuristics in favor of smoothing with cross-validation. We will prove that certain cross-validated histograms and kernels are consistent. Our proof is long and technical, and our approach is, in a sense, by "brute force". One might describe the plan of the proof, very loosely, as follows: (1) Show that the cross-validated smoothing parameter diverges to ∞; (2) establish an upper bound on the rate of divergence; and (3) demonstrate that consistent estimation obtains from *any* sequence of smoothing parameters (whether random or deterministic) respecting (1) and (2). What the proof does not tell us is why, in the intuitive sense, cross-validation is a reasonable mechanism for selecting smoothing parameters. Before formally stating our consistency results, we will take some time here to briefly argue, in a heuristic manner, for L_λ (defined in (2.1)) as an appropriate data-defined criterion for selecting smoothing parameters. (There is an entirely analogous argument for cross-validated regression estimators, with L_λ replaced by a least squares criterion.) Similar motivation has been offered before, by Stone (1974) and by Wong (1979).

Our heuristics are based on the following two observations: Let $f(x)$ be a probability density function satisfying $|\int f(x)\log f(x)\,dx| < \infty$ (integrals without limits are taken over the whole real line). Then

(i) $\int f(x)\log g(x)\,dx < \int f(x)\log f(x)\,dx$ whenever g is not equivalent to f, and
(ii) the Kullback-Leibler information, $\int f(x)\log\{f(x)/g(x)\}\,dx$, is a meaningful measure of how well a density g approximates f. In fact, if $\{f_n\}_{n=1}^{\infty}$ is any sequence of density functions for which $\int f(x)\log\{f(x)/f_n(x)\}\,dx \to 0$, then $\int |f(x) - f_n(x)|\,dx \to 0$ as well. (i) is an easy consequence of Jensen's inequality (see proof of Theorem 1), and (ii) is proved in Geman (1981). Now, we might reason that

$$\frac{1}{n}\log L_\lambda = \frac{1}{n}\sum_{i=1}^n \log f^i_{\lambda,n-1}(x_i) \approx \frac{1}{n}\sum_{i=1}^n \log f_{\lambda,n}(x_i) \approx \int f(x)\log f_{\lambda,n}(x)\,dx, \tag{3.1}$$

by a Law of Large Numbers. Let us denote the cross-validated λ, given n observations, by λ_n^*, and let γ_n be any deterministic sequence such that $f_{\gamma_n,n}$ is consistent in the sense that

$$\int f(x)\log\{f(x)/f_{\gamma_n,n}(x)\}\,dx \to 0 \quad \text{a.s.} \tag{3.2}$$

Recall that λ_n^* maximizes L_λ, and use this in (3.1): since $L_{\gamma_n} \le L_{\lambda_n^*}$, we expect (at least approximately)

$$\int f(x)\log f_{\gamma_n,n}(x)\,dx \le \int f(x)\log f_{\lambda^*_n,n}(x)\,dx \le \int f(x)\log f(x)dx$$

by (i). But then

$$\int f(x)\log\{f(x)/f_{\gamma_n,n}(x)\}\,dx \ge \int f(x)\log\{f(x)/f_{\lambda^*_n,n}(x)\}\,dx \ge 0.$$

Equation (3.2), together with (ii), would then lead us to conclude not only that $f_{\lambda^*_n,n}$ is consistent, but also that $f_{\lambda^*_n,n}$ will compare favorably with $f_{\gamma_n,n}$, for any deterministic sequence γ_n.

4. Two consistency results. (a) *Kernel estimators.* Given probability density functions f and K, and $x_1, x_2, \ldots$ a random (iid) sample from a distribution with density f, we define:

1. $f_{\lambda,n}(x) = (1/n)\sum_{i=1}^n \lambda K(\lambda(x - x_i))$, the kernel estimator of f with window size $1/\lambda$, $\lambda \ge 0$;

2. $f^{i}_{\lambda,n-1}(x) = (1/(n-1)) \sum_{j \neq i} \lambda K(\lambda(x - x_j))$, the kernel estimator based on the random sample excluding the ith observation;

3. $L_\lambda = \prod_{i=1}^{n} f^{i}_{\lambda,n-1}(x_i)$, a suitability measure for λ;

4. for a fixed number $0 < \pi < 1$, a family of cross-validated smoothing parameters $\Lambda_n = \{\lambda : L_\gamma \geq \pi \sup_{\gamma \geq 0} L_\gamma\}$, which is a function of $x_1, x_2, \cdots, x_n$. (Extra restrictions on K would be needed to insure that L_λ attains its supremum. Choosing $\pi < 1$ rather than $\pi = 1$ both generalizes the statement of the theorem and avoids unnecessary assumptions about K.)

If we assume:

A1. f is bounded and has compact support,

A2. K is bounded, has compact support, and (i) K is nondecreasing on $(-\infty, 0]$ (ii) K is nonincreasing on $[0, \infty)$ (iii) for some $\delta > 0$, $\min(K(-\delta), K(\delta)) > 0$,

then the cross-validated kernel is consistent:

THEOREM 1. *For each $n \geq 2$, Λ_n is almost surely nonempty, and*

$$\sup_{\lambda \in \Lambda_n} \int_{-\infty}^{\infty} |f_{\lambda,n}(x) - f(x)|\, dx \to 0 \quad \text{a.s.}$$

REMARKS.

1. Functions of $x_1, x_2, \ldots$ such as

$$\sup_{\lambda \in \Lambda_n} \int |f_{\lambda,n}(x) - f(x)|\, dx$$

may not be measurable. In the statements and proofs of our theorems, probabilities of events associated with such functions will always be 0 or 1. We interpret these in terms of the completed probability measure.

2. Michael Perlman suggested the following generalization: Even if f does not have compact support, for suitable transformations $g : (-\infty, \infty) \to (0, 1)$ the distribution of $g(x_1)$, $g(x_2)$, ... will satisfy the conditions of the theorem. (Suppose, for example, that f is bounded and has tails $o(1/x^2)$. If $g^{-1}(x) = -\cot(\pi x)$, then $g(x_1)$ has bounded density with compact support.) Apply Theorem 1 to the transformed data, and then transform back:

$$\sup_{\lambda \in \Lambda_n} \int |g'(x) f_{\lambda,n}(g(x)) - f(x)|\, dx \to 0 \quad \text{a.s.},$$

where Λ_n and $f_{\lambda,n}$ are constructed from the transformed observations $g(x_1), g(x_2), \ldots$.

3. Let $\lambda_n \to \infty$ be a deterministic sequence satisfying $\lambda_n = o(n/\log\log n)$. By techniques similar to those used in proving the theorem, it can be shown that, for a.e. $x(dx)$, $f_{\lambda_n,n}(x) \to f(x)$ a.s., and that $\int |f_{\lambda_n,n}(x) - f(x)|\, dx \to 0$ a.s. (cf. Devroye and Wagner, 1979). An analogous statement is true for the histogram estimator (see Theorem 2 below) as well. The details will be presented in a forthcoming article by Chow.

(b) *Histograms.* Again let f be a probability density function, and let $x_1, x_2, \ldots$ be a random sample from a distribution with density f. For each $\lambda \geq 0$, each $j = 0, \pm 1, \pm 2, \ldots$, and each $n = 2, 3, \ldots$, define:

1. $f_{j,\lambda,n} = \dfrac{\lambda}{n} \sum_{i=1}^{n} \chi_{\left[\frac{j-1}{\lambda}, \frac{j}{\lambda}\right)}(x_i) = \dfrac{\lambda}{n} \#\left\{x_i : \dfrac{j-1}{\lambda} \leq x_i < \dfrac{j}{\lambda}\right\}$,

the histogram estimator of f on the interval $\left[\dfrac{j-1}{\lambda}, \dfrac{j}{\lambda}\right)$;

2. $f_{\lambda,n}(x) = f_{j,\lambda,n}$ for $x \in \left[\dfrac{j-1}{\lambda}, \dfrac{j}{\lambda}\right)$, the histogram estimator of f;

3. $f^i_{j,\lambda,n-1} = \dfrac{\lambda}{n-1}\sum_{k\neq i} \chi_{\left[\frac{j-1}{\lambda}, \frac{j}{\lambda}\right)}(x_k)$;

4. $f^i_{\lambda,n-1}(x) = f^i_{j,\lambda,n-1}$ for $x \in \left[\dfrac{j-1}{\lambda}, \dfrac{j}{\lambda}\right)$, the histogram estimator based on the random sample excluding the ith observation.

Λ_n, the family of cross-validated smoothing parameters, is defined exactly as it was for the kernel estimator. Then, the cross-validated histogram is consistent for bounded densities with compact support:

THEOREM 2. *If f is bounded and has compact support, then for each $n \geq 2$, Λ_n is almost surely nonempty, and*

$$\sup_{\lambda\in\Lambda_n} \int_{-\infty}^{\infty} |f_{\lambda,n}(x) - f(x)|\, dx \to 0 \quad \text{a.s.}$$

We will include here only the proof for Theorem 1. The proof for Theorem 2 follows along similar lines. See Chow et al. (1981) for details.

5. Proofs.

PROOF OF THEOREM 1. Let us assume, for the time being, the following lemmas:

LEMMA 1.1. *For each $n \geq 2$, Λ_n is almost surely nonempty. Furthermore* (i) $P(\inf_{\lambda\in\Lambda_n}\lambda \to \infty) = 1$, *and* (ii) *there exists $k > 0$ such that $P\{\sup_{\lambda\in\Lambda_n}\lambda > (kn/\log n)$ i.o.$\} = 0$.*

LEMMA 1.2. *Let $\tilde{K}(x) = \chi_{[0,1]}(x)$, $\tilde{f}_{\lambda,n}(x) = \dfrac{1}{n}\sum_{i=1}^n \lambda\tilde{K}(\lambda(x - x_i))$, and*

$$\tilde{f}_\lambda(x) = \int f(y)\lambda\tilde{K}(\lambda(x-y))\, dy.$$

Define $n_0 = 0$, and $n_k = [e^k]$, $k = 1, 2, \cdots$ ($[x]$ is the greatest integer less than or equal to x.) If $\lambda_n \geq 0$ is any (deterministic) sequence satisfying (i) $\lambda_n \to \infty$, (ii) λ_n *constant on* $(n_{k-1}, n_k]$ $k = 1, 2, \cdots$, *and* (iii) $\lambda_n = o(n/\log\log n)$,
then

$$\int |\tilde{f}_{\lambda_n,n}(x) - \tilde{f}_{\lambda_n}(x)|\, dx \to 0 \quad \text{a.s.}$$

LEMMA 1.3. *Define $\tilde{K}$, $\hat{f}_{\lambda,n}$, and $\tilde{f}_\lambda$ as in Lemma* 1.2. *Then, with Λ_n defined as in the theorem (using K, not $\tilde{K}$),*

$$\sup_{\lambda\in\Lambda_n} \int |\tilde{f}_{\lambda,n}(x) - \tilde{f}_\lambda(x)|\, dx \to 0 \quad \text{a.s.}$$

LEMMA 1.4. *Let $\hat{K}(x) = \sum_{i=1}^m \alpha_i\chi_{[a_i,b_i]}(x)$ for some $\alpha_1, \cdots \alpha_m$ and $a_1 < b_1, \cdots a_m < b_m$. Use $\hat{K}$ to define $\hat{f}_{\lambda,n}$ and $\hat{f}_\lambda$ (see Lemma* 1.2). *Then, with Λ_n defined as in the theorem (using K, not $\hat{K}$),*

$$\sup_{\lambda\in\Lambda_n} \int |\hat{f}_{\lambda,n}(x) - \hat{f}_\lambda(x)|\, dx \to 0 \quad \text{a.s.}$$

Lemma 1.1 asserts that Λ_n is (a.s.) nonempty. Define $f_\lambda(x) = \int f(y)\lambda K(\lambda(x-y))\, dy$.

Then

$$\sup_{\lambda\in\Lambda_n}\int |f_{\lambda,n}(x)-f(x)|\,dx \le \sup_{\lambda\in\Lambda_n}\int |f_{\lambda,n}(x)-f_\lambda(x)|\,dx + \sup_{\lambda\in\Lambda_n}\int |f_\lambda(x)-f(x)|\,dx.$$

Since $f_\lambda \to f$ in L_1 as $\lambda\to\infty$ (see Stein, 1970, theorem 2, page 62), Lemma 1.1 implies that the latter term above goes to 0, almost surely. It is enough, then, to show that

$$\sup_{\lambda\in\Lambda_n}\int |f_{\lambda,n}(x)-f_\lambda(x)|\,dx \to 0 \text{ a.s.} \tag{5.1}$$

Since the continuous functions are dense in L_1, so are the functions of the form $\hat{K}$ defined in Lemma 1.4. Fix $\varepsilon>0$ and choose $\hat{K}$, of the form defined in Lemma 1.4, such that $\|K-\hat{K}\|_{L_1}<\varepsilon$. Then

$$\begin{aligned}\sup_{\lambda\in\Lambda_n}\int |f_{\lambda,n}(x)-f_\lambda(x)|\,dx \le{}& \sup_{\lambda\in\Lambda_n}\int |f_{\lambda,n}(x)-\hat{f}_{\lambda,n}(x)|\,dx\\ &+\sup_{\lambda\in\Lambda_n}\int |f_\lambda(x)-\hat{f}_\lambda(x)|\,dx\\ &+\sup_{\lambda\in\Lambda_n}\int |\hat{f}_{\lambda,n}(x)-\hat{f}_\lambda(x)|\,dx,\end{aligned}$$

and Lemma 1.4 asserts that the last term goes to 0, a.s., as $n\to\infty$. For the first term:

$$\begin{aligned}\int |f_{\lambda,n}(x)-\hat{f}_{\lambda,n}(x)|\,dx &= \int \left|\frac{1}{n}\textstyle\sum_{i=1}^n \{\lambda K(\lambda(x-x_i))-\lambda\hat{K}(\lambda(x-x_i))\}\right| dx\\ &\le \frac{1}{n}\textstyle\sum_{i=1}^n \int |\lambda K(\lambda(x-x_i))-\lambda\hat{K}(\lambda(x-x_i))|\,dx = \|K-\hat{K}\|_{L_1}<\varepsilon.\end{aligned}$$

Similarly (replace $n^{-1}\Sigma$ by $\int$)

$$\int |f_\lambda(x)-\hat{f}_\lambda(x)|\,dx<\varepsilon,$$

which proves (5.1). (The approach is similar to the one used by Bertrand-Retali (1978): realize K as a limit of approximating step functions.) So it is enough to prove the lemmas.

PROOF OF LEMMA 1.1. For each $n\ge 2$, the points $x_1,\cdots x_n$ are a.s. distinct. Since K has compact support, $L_\lambda=0$ for all λ sufficiently large, provided $x_1,\cdots x_n$ are distinct. On the other hand, for $\lambda>0$ sufficiently small, $L_\lambda>0$, and it follows that λ_n is a.s. nonempty.

PART (i). The proof of (i) is based on the following four lemmas:

LEMMA 1.a. *For any* $\lambda_1>\lambda_0>0$

$$\sup_{\lambda_0\le\lambda\le\lambda_1}\left|\frac{1}{n}\log L_\lambda-\frac{1}{n}\textstyle\sum_1^n \log f_\lambda(x_i)\right|\to 0 \text{ a.s.}$$

LEMMA 1.b. *For any* $\lambda_1>\lambda_0>0$

$$\sup_{\lambda_0\le\lambda\le\lambda_1}\left|\frac{1}{n}\textstyle\sum_1^n \log f_\lambda(x_i)-\int f(x)\log f_\lambda(x)\,dx\right|\to 0 \text{ a.s.}$$

LEMMA 1.c.

$$\lim_{\lambda\to\infty}\int f(x)\log f_\lambda(x)\,dx=\int f(x)\log f(x)\,dx$$

(*finite or infinite*).

LEMMA 1.d. *$\int f(x)\log f_\lambda(x)\,dx$ is a continuous function of λ on $(0,\infty)$.*

For now, let us postpone the proofs of these lemmas. Assuming that they are true, we will show that for any $\lambda_1 > 0$

$$P(\liminf_{n\to\infty}\inf_{\lambda\in\Lambda_n}\lambda \geq \lambda_1) = 1.$$

Now fix $\lambda_1 > 0$, and observe that it is enough to show that there exists $\lambda_2 > \lambda_1$ such that

$$\liminf_{n\to\infty}\{\pi L_{\lambda_2} - \sup_{0\leq\lambda\leq\lambda_1} L_\lambda\} > 0 \text{ a.s.} \tag{5.2}$$

Because, if (5.2) is true, then for a.e. ω, when n is large enough:

$$\gamma \in [0,\lambda_1] \Rightarrow L_\gamma \leq \sup_{0\leq\lambda\leq\lambda_1} L_\lambda < \pi L_{\lambda_2} \leq \pi \sup_{\lambda\geq 0} L_\lambda,$$

i.e. $\gamma \in [0,\lambda_1] \Rightarrow \gamma \notin \Lambda_n$, and hence $\inf_{\lambda\in\Lambda_n}\lambda \geq \lambda_1$. ($\omega$ is a sample point in the probability space underlying the observations $x_1, x_2, \cdots$.) For (5.2), in turn, it is sufficient that (almost surely):

$$\limsup_{n\to\infty}\sup_{0\leq\lambda\leq\lambda_1}\frac{1}{n}\log L_\lambda < \liminf_{n\to\infty}\frac{1}{n}\log \pi L_{\lambda_2} = \liminf_{n\to\infty}\frac{1}{n}\log L_{\lambda_2}. \tag{5.3}$$

Let F be the distribution function of f, and let S be the support of the associated probability measure (assumed compact, see A.1). For any $\lambda > 0$, $\inf_{x\in S} f_\lambda(x) > 0$ (a stronger statement is demonstrated in the proof of Lemma 1.a), and consequently

$$\int f(x)\log f_1(x)\,dx > -\infty,$$

where $f_1(x)$ is $f_\lambda(x)$ at $\lambda = 1$. Therefore, there exists λ_0 sufficiently small, such that $0 < \lambda_0 < \lambda_1$ and

$$\begin{aligned}\limsup_{n\to\infty}\sup_{0\leq\lambda\leq\lambda_0}\frac{1}{n}\log L_\lambda &= \limsup_{n\to\infty}\sup_{0\leq\lambda\leq\lambda_0}\frac{1}{n}\sum_{i=1}^n \log f^i_{\lambda,n-1}(x_i)\\ &\leq \limsup_{n\to\infty}\sup_{0\leq\lambda\leq\lambda_0}\frac{1}{n}\sum_{i=1}^n \log \lambda K(0) \qquad (5.4)\\ &= \log\lambda_0 K(0) < \int f(x)\log f_1(x)\,dx.\end{aligned}$$

Since $f(x)\log f(x) > -1/e$, $\int f(x)\log f(x)\,dx > -\infty$ whenever f has compact support. If $\int f(x)\log f(x)\,dx < +\infty$ as well, then by Jensen's inequality, for any $\lambda > 0$

$$\int f(x)\log f_\lambda(x)\,dx - \int f(x)\log f(x)\,dx = \int f(x)\log\{f_\lambda(x)/f(x)\}\,dx < \log\int_S f_\lambda(x)\,dx < 0.$$

(The support of f_λ is strictly larger than the support of f, hence the strict inequalities.) Since $f_\lambda(x) \leq \lambda K(0)$, $\int f(x)\log f_\lambda(x)\,dx < \infty$. Therefore, whether or not $\int f(x)\log f(x)\,dx < \infty$,

$$\int f(x)\log f_\lambda(x)\,dx < \int f(x)\log f(x)\,dx \tag{5.5}$$

for every $\lambda > 0$.

Because of (5.5), and Lemmas 1.c and 1.d, we can find $\lambda_2 > \lambda_1$ such that

$$\int f(x)\log f_{\lambda_2}(x)\,dx > \int f(x)\log f_1(x)\,dx$$

and

$$\int f(x)\log f_{\lambda_2}(x)\,dx > \sup_{\lambda_0\leq\lambda\leq\lambda_1}\int \mathrm{f}(\mathrm{x})\log f_\lambda(x)\,dx,$$

with λ_0 as in (5.4). Then, using (5.4) and Lemmas 1.a and 1.b,

$$\limsup_{n\to\infty} \sup_{0\le\lambda\le\lambda_0} \frac{1}{n}\log L_\lambda < \int f(x)\log f_{\lambda_2}(x)\,dx = \liminf_{n\to\infty} \frac{1}{n}\log L_{\lambda_2} \quad \text{a.s.},$$

and

$$\limsup_{n\to\infty} \sup_{\lambda_0\le\lambda\le\lambda_1} \frac{1}{n}\log L_\lambda < \int f(x)\log f_{\lambda_2}(x)\,dx = \liminf_{n\to\infty} \frac{1}{n}\log L_{\lambda_2} \quad \text{a.s.},$$

and (5.3) follows.

What remains (for part (i) of Lemma 1.1) are the proofs of Lemmas 1.a, 1.b, 1.c, and 1.d. For brevity these are deleted, but are available in Chow et al. (1981).

PART (ii). There are results on spacings of random variables, such as those by Devroye (1981), that can be adapted to the proof of ii. But for the present application it is no more work to prove the result directly. Let $y_1 \le y_2 \cdots \le y_n$ be the order statistics of $x_1, \cdots x_n$. Recalling that K has compact support, choose $s > 0$ such that $|x| \ge s \Rightarrow K(x) = 0$. Observe that if

$$\lambda \ge s/\min(y_i - y_{i-1}, y_{i+1} - y_i)$$

for some $i = 2, \cdots n-1$, then

$$f^i_{\lambda,n-1}(y_i) = \frac{1}{n-1}\sum_{j\ne i} \lambda K(\lambda(y_i - y_j)) = 0.$$

Hence $L_\lambda = 0$ for all

$$\lambda \ge s/\max_{2\le i\le n-1}\min(y_i - y_{i-1}, y_{i+1} - y_i).$$

Since $L_\lambda > 0$ for all λ sufficiently small, $\lambda \in \Lambda_n \Rightarrow$

$$\lambda < s/\max_{2\le i\le n-1}\min(y_i - y_{i-1}, y_{i+1} - y_i).$$

Let α_n be a sequence of positive numbers. Since $f(x)$ is bounded, we can choose $\delta > 0$ such that $|F(x) - F(y)| \le \delta|x - y|$ for all x and y. Then

$$\{\sup_{\lambda\in\Lambda_n}\lambda \ge \alpha_n\} \Rightarrow \left\{\max_{2\le i\le n-1}\min(y_i - y_{i-1}, y_{i+1} - y_i) < \frac{s}{\alpha_n}\right\}$$

$$\Rightarrow \left\{\max_{2\le i\le n-1}\min(F(y_i) - F(y_{i-1}), F(y_{i+1}) - F(y_i)) < \frac{s\delta}{\alpha_n}\right\}.$$

Let us call this latter event A_n. We want to show that for $\alpha_n = kn/\log n$, k sufficiently large, $P(A_n \text{ i.o.}) = 0$.

It is well known that $\{F(y_i) - F(y_{i-1})\}$ $i = 2, 3, \cdots, n$ have the same joint distribution as $(U_1/S_n, U_2/S_n, \cdots U_{n-1}/S_n)$, where $S_n = \sum_{k=0}^n U_k$, and $U_0, \cdots U_n$ are iid unit mean exponential random variables. Hence

$$P(A_n \text{ i.o.}) = P\{\max_{2\le i\le n-1}\min(U_{i-1}/S_n, U_i/S_n) \le s\delta/\alpha_n \text{ i.o.}\}.$$

Observe that

$$\{\max_{2\le i\le n-1}\min(U_{i-1}/S_n, U_i/S_n) \le s\delta/\alpha_n \text{ i.o., and } S_n/n \to 1\}$$

$$\Rightarrow \{\max_{2\le i\le n-1}\min(U_{i-1}, U_i) \le 2ns\delta/\alpha_n \text{ i.o.}\}$$

$$\Rightarrow \{\max_{2\le i\le n-1, i\text{even}}\min(U_{i-1}, U_i) \le 2ns\delta/\alpha_n \text{ i.o.}\}$$

Since $S_n/n \to 1$ a.s.,

$$P(A_n \text{ i.o.}) \le P\{\max_{2\le i\le n-1, i\text{even}}\min(U_{i-1}, U_i) \le 2ns\delta/\alpha_n \text{ i.o.}\}.$$

But

$$P\{\max_{2\le i\le n-1,\, i \text{ even}} \min(U_{i-1}, U_i) \le 2ns\delta/\alpha_n\}$$
$$= P\{\min(U_{i-1}, U_i) \le 2ns\delta/\alpha_n\}^{[(n-1)/2]}$$
$$= \{1 - e^{-4ns\delta/\alpha_n}\}^{[(n-1)/2]} \le \exp\left\{-\left[\frac{n-1}{2}\right] e^{-4ns\delta/\alpha_n}\right\}.$$

If $\alpha_n = kn/\log n$, then

$$P\{\max_{2\le i\le n-1,\, i \text{ even}} \min(U_{i-1}, U_i) \le 2ns\delta/\alpha_n\} \le \exp\left\{-\left[\frac{n-1}{2}\right] \Big/ n^{4s\delta/k}\right\},$$

which is summable for k sufficiently large. Hence, by the Borel-Cantelli lemma, $P(A_n \text{ i.o.}) = 0$ for k sufficiently large.

PROOF OF LEMMA 1.2. (There is a large literature on the law of the iterated logarithm, some of which is devoted to "triangular arrays" of random variables. But we have been unable to find results sufficient for our present purposes. The most closely related work appears to be in a recent paper by Hall (1981), but the results there are not sufficient for Lemma 1.2.)

Suppose we can show that, for a.e. x,

$$\tilde{f}_{\lambda_n,n}(x) - \tilde{f}_{\lambda_n}(x) \to 0 \quad \text{a.s.} \tag{5.6}$$

Then, also, for a.e., ω, $\tilde{f}_{\lambda_n,n}(x) - \tilde{f}_{\lambda_n}(x) \to 0$ a.s. (dx). Again by the theorem in Stein (1970, page 62), $\tilde{f}_{\lambda_n}(x) \to f(x)$ for a.e. x. Using this, (5.6), and Fatou's lemma with $A_n = \{x : \tilde{f}_{\lambda_n}(x) \ge \tilde{f}_{\lambda_n,n}(x)\}$

$$\limsup_{n\to\infty} \int |\tilde{f}_{\lambda_n}(x) - \tilde{f}_{\lambda_n}(x)|\, dx$$
$$= 2 - 2 \liminf_{n\to\infty} \int [\tilde{f}_{\lambda_n}(x) - \chi_{A_n}(x)\{\tilde{f}_{\lambda_n}(x) - \tilde{f}_{\lambda_n,n}(x)\}]\, dx$$
$$\le 2 - 2\int f(x)\, dx = 0 \quad \text{a.s.}$$

It is enough, therefore, to show (5.6).

Fix an x for which $\tilde{f}_{\lambda_n}(x) \to f(x)$. For each λ and i define $Z_i(\lambda) = \lambda \tilde{K}(\lambda(x - x_i)) - \tilde{f}_\lambda(x)$, and for each λ and n define $S_n(\lambda) = \sum_{i=1}^n Z_i(\lambda)$.

In this notation, (5.6) is written as $(1/n)S_n(\lambda_n) \to 0$ a.s.

Fix $\varepsilon > 0$. We will show that

$$P\left(\frac{1}{n} S_n(\lambda_n) > \varepsilon \text{ i.o.}\right) = 0.$$

The other inequality, $n^{-1}S_n(\lambda_n) < -\varepsilon$, is handled in the same way.

For $k \ge 2$, $n_{k-1} > n_k/9$. Hence

$$P\left\{\frac{1}{n} S_n(\lambda_n) > \varepsilon \text{ i.o.}\right\} = P\{\max_{n_{k-1}<n\le n_k} (S_n(\lambda_n) - n\varepsilon) > 0 \text{ i.o. } (k)\}$$
$$\le P\{\max_{n_{k-1}<n\le n_k} (S_n(\lambda_n) - n_{k-1}\varepsilon) > 0 \text{ i.o. } (k)\}$$
$$\le P\left\{\max_{n_{k-1}<n\le n_k} \left(S_n(\lambda_n) - \frac{n_k\varepsilon}{9}\right) > 0 \text{ i.o. } (k)\right\}$$
$$= P\left\{\max_{n_{k-1}<n\le n_k} S_n(\lambda_n) > \frac{n_k\varepsilon}{9} \text{ i.o. } (k)\right\}.$$

So, it is enough to show that, for every $\varepsilon > 0$,

$$P\{\max_{n_{k-1}<n\le n_k} S_n(\lambda_n) > n_k\varepsilon \text{ i.o.}\} = 0.$$

Observe that $E\{Z_i(\lambda_n)\} = 0$,

$$E\{Z_i(\lambda_n)\}^2 = \int f(y)\lambda_n^2 \tilde{K}(\lambda_n(x-y))^2\,dy - \{\tilde{f}_{\lambda_n}(x)\}^2 \le \lambda_n \tilde{f}_{\lambda_n}(x) = O(\lambda_n),$$

and $|Z_i(\lambda_n)| \le 2\lambda_n$. When $\lambda_n < \varepsilon/2$, then we have, already, $S_n(\lambda_n) < n\varepsilon/2 < n_k\varepsilon$ for $n_k \ge n$. So we can assume, without loss of generality, that $\lambda_n \ge \varepsilon/2$ for each n. Now consider $\phi_n(t) \equiv E[e^{t\{Z_i(\lambda_n)-\varepsilon\}}]$. Following a familiar argument (see, for example, Chernoff, 1952), we can find $t_0 > 0$ (which depends on n and ε) such that $\phi_n(t_0) \le 1 - \delta/\lambda_n$ for some sufficiently small $\delta > 0$ (which depends on ε, but not on n).

Because λ_n is constant on $(n_{k-1}, n_k]$, $S_n(\lambda_n)$ is a martingale on $(n_{k-1}, n_k]$ for each k. Apply the martingale inequality

$$P\{\max_{n_{k-1}<n\le n_k} S_n(\lambda_n) > n_k\varepsilon\} \le e^{-tn_k\varepsilon} E\{e^{tS_{n_k}(\lambda_{n_k})}\} = \{\phi_{n_k}(t)\}^{n_k}$$

for every $t \ge 0$. In particular,

$$P\{\max_{n_{k-1}<n\le n_k} S_n(\lambda_n) > n_k\varepsilon\} \le \{\phi_{n_k}(t_0)\}^{n_k} \le (1 - \delta/\lambda_{n_k})_{n_r}.$$

Thus Lemma 1.2 can be proved by demonstrating that $a_k \equiv (1 - \delta/\lambda_{n_k})^{n_k}$ is a summable sequence, and then applying the Borel-Cantelli lemma. But

$$a_k = \exp\{n_k \log(1 - \delta/\lambda_{n_k})\} < \exp(-\delta n_k/\lambda_{n_k}),$$

and the latter is summable since $\lambda_{n_k}/n_k = o(1/\log k)$.

PROOF OF LEMMA 1.3. According to Lemma 1.1, we can find a sequence $\{\lambda_n\}$ satisfying the conditions for Lemma 1.2, and such that

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda_n} \lambda/\lambda_n = 0 \quad \text{a.s.} \tag{5.7}$$

Fix such a sequence $\{\lambda_n\}$. Lemma 1.3 is proved by comparing

$$\sup_{\lambda\in\Lambda_n} \int |\tilde{f}_{\lambda,n}(x) - \tilde{f}_{\lambda}(x)|\,dx$$

to

$$\int |\tilde{f}_{\lambda_n,n}(x) - \tilde{f}_{\lambda_n}(x)|\,dx.$$

The latter goes to zero, as was established in Lemma 1.2.

Define, for each n and each $\lambda \in \Lambda_n$, $N_n(\lambda) = \inf\{n : n\lambda \ge \lambda_n\}$.

Because of (5.7), and because $\lambda_n \to \infty$,

$$\sup_{\lambda\in\Lambda_n} \left| \frac{N_n(\lambda)\lambda}{\lambda_n} - 1 \right| \to 0 \quad \text{a.s.}$$

An easy consequence is that

$$\sup_{\lambda\in\Lambda_n} \int |N_n(\lambda)\lambda \chi_{[0,1]}(N_n(\lambda)\lambda x) - \lambda_n \chi_{[0,1]}(\lambda_n x)|\,dx \to 0 \quad \text{a.s.},$$

and from this

$$\begin{aligned} &\sup_{\lambda\in\Lambda_n} \int |\tilde{f}_{N_n(\lambda)\lambda,n}(x) - \tilde{f}_{\lambda_n,n}(x)|\,dx \to 0 \quad \text{a.s.} \\ &\sup_{\lambda\in\Lambda_n} \int |\tilde{f}_{N_n(\lambda)\lambda}(x) - \tilde{f}_{\lambda_n}(x)|\,dx \to 0 \quad \text{a.s.} \end{aligned} \tag{5.8}$$

Notice that for any integer $N > 0$ and any $\lambda > 0$

$$\lambda\chi_{[0,1]}(\lambda x) = \frac{1}{N}\sum_{j=1}^{N} N\lambda\chi_{[0,1]}\left(N\lambda\left(x - \frac{j-1}{N\lambda}\right)\right) \quad \text{a.e.} \quad (dx).$$

Using this in the definitions of $\tilde{f}_{\lambda,n}$ and $\tilde{f}_{\lambda}$ we can rewrite

$$\tilde{f}_{\lambda,n}(x) = \frac{1}{N}\sum_{j=1}^{N} \tilde{f}_{N\lambda,n}\left(x - \frac{j-1}{N\lambda}\right)$$

and

$$\tilde{f}_{\lambda}(x) = \frac{1}{N}\sum_{j=1}^{N} \tilde{f}_{N\lambda}\left(x - \frac{j-1}{N\lambda}\right) \quad \text{a.e.} \quad (dx).$$

Now apply (5.8):

$$\begin{aligned}
\sup_{\lambda\in\Lambda_n}\int |\tilde{f}_{\lambda,n}(x) - \tilde{f}_{\lambda}(x)|\,dx &\le \sup_{\lambda\in\Lambda_n}\frac{1}{N_n(\lambda)}\sum_{j=1}^{N_n(\lambda)}\int |\tilde{f}_{N_n(\lambda)\lambda,n}\left(x - \frac{j-1}{N_n(\lambda)\lambda}\right) \\
&\quad - \tilde{f}_{N_n(\lambda)\lambda}\left(x - \frac{j-1}{N_n(\lambda)\lambda}\right)|\,dx \\
&= \sup_{\lambda\in\Lambda_n}\int |\tilde{f}_{N_n(\lambda)\lambda,n}(x) - \tilde{f}_{N_n(\lambda)\lambda}(x)|\,dx \\
&\le \sup_{\lambda\in\Lambda_n}\int |\tilde{f}_{\lambda_n,n}(x) - \tilde{f}_{\lambda_n}(x)|\,dx \\
&\quad + \sup_{\lambda\in\Lambda_n}\int |\tilde{f}_{N_n(\lambda)\lambda,n}(x) - \tilde{f}_{\lambda_n,n}(x)|\,dx \\
&\quad + \sup_{\lambda\in\Lambda_n}\int |\tilde{f}_{N_n(\lambda)\lambda}(x) - \tilde{f}_{\lambda_n}(x)|\,dx \to 0 \quad \text{a.s.}
\end{aligned}$$

PROOF OF LEMMA 1.4. Observe that

$$\hat{K}(x) = \sum_{i=1}^{m}\alpha_i\chi_{[a_i,b_i]}(x) = \sum_{i=1}^{m}\alpha_i\chi_{[0,1]}\left(\frac{x-a_i}{b_i-a_i}\right).$$

Use this to rewrite $\hat{f}_{\lambda,n}$ and $\hat{f}_{\lambda}$ in terms of $\tilde{f}_{\lambda,n}$ and $\tilde{f}_{\lambda}$, defined in Lemma 1.2,

$$\hat{f}_{\lambda,n}(x) = \sum_{i=1}^{m}\alpha_i(b_i-a_i)\tilde{f}_{(\lambda/(b_i-a_i)),n}(x - a_i/\lambda)$$

and

$$\hat{f}_{\lambda}(x) = \sum_{i=1}^{m}\alpha_i(b_i-a_i)\,\tilde{f}_{(\lambda/(b_i-a_i))}(x - a_i/\lambda).$$

Hence

$$\begin{aligned}
&\sup_{\lambda\in\Lambda_n}\int |\hat{f}_{\lambda,n}(x) - \hat{f}_{\lambda}(x)|\,dx \\
&\quad \le \sum_{i=1}^{m}\alpha_i(b_i-a_i)\sup_{\lambda\in\Lambda_n}\int |\tilde{f}_{(\lambda/(b_i-a_i)),n}(x - a_i/\lambda) - \tilde{f}_{(\lambda/(b_i-a_i))}(x - a_i/\lambda)|\,dx \\
&\quad = \sum_{i=1}^{m}\alpha_i(b_i-a_i)\sup_{\lambda\in\hat{\Lambda}_{n,i}}\int |\tilde{f}_{\lambda,n}(x) - \tilde{f}_{\lambda}(x)|\,dx
\end{aligned}$$

where $\hat{\Lambda}_{n,i} = \{\lambda : (b_i - a_i)\lambda \in \Lambda_n\}$. For each i, $\hat{\Lambda}_{n,i}$ has the same properties established in Lemma 1.1 for Λ_n. Since the proof of Lemma 1.3 made use of these properties only,

Lemma 1.3 applies with Λ_n replaced by $\hat{\Lambda}_{n,i}$:

$$\sup_{\lambda \in \hat{\Lambda}_{n,i}} \int |\tilde{f}_{\lambda,n}(x) - \tilde{f}_{\lambda}(x)|\, dx \to 0 \quad \text{a.s.}$$

for each i.

REFERENCES

BERTRAND-RETALI, M. (1978). Convergence uniforme d'un estimateur de la densité par la méthode du noyau. *Rev. Roumaine Math. Pures et Appl.* **23** 361–385.

BREIMAN, L., MEISEL, W., and PURCELL, E. (1977). Variable kernel estimates of multivariate densities and their calibration. *Technometrics* **19** 135–144.

CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23** 493–507.

CHOW, Y. S., GEMAN, S., and WU, L.-D. (1981). Consistent cross-validated density estimation. *Reports in Pattern Analysis* No. 110. Div. Appl. Math., Brown University.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.

DAWID, A. P. (1974). Discussion of: Cross-validatory choice and assessment of statistical predictions (by M. Stone). *J. Roy. Statist. Soc. Ser. B* **36** 136–138.

DEVROYE, L. (1981). Laws of the iterated logarithm for order statistics of uniform spacings. *Ann. Probability* **9** 860–867.

DEVROYE, L. P. and WAGNER, T. J. (1979). The L_1 convergence of kernel density estimates. *Ann. Statist.* **7** 1136–1139.

DUIN, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. Comput.* **C-25** 1175–1179.

EGERTON, M. F. and LAYCOCK, P. J. (1981). Some criticisms of stochastic shrinkage and ridge regression, with counterexamples. *Technometrics* **23** 155–159.

GEMAN, S. (1981). Sieves for nonparametric estimation of densities and regressions. *Reports in Pattern Analysis* No. 99, Div. of Appl. Math., Brown University.

GEMAN, S. and HWANG, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414.

GOLUB, G. H., HEATH, M., and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223.

GOOD, I. J. and GASKINS, R. A. (1972). Global nonparametric estimation of probability densities. *Virginia J. Sci.* **23** 171–193.

GORDON, L. and OLSHEN, R. A. (1980). Consistent nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **10** 611–627.

GRENANDER, U. (1981). *Abstract Inference.* Wiley, New York.

HABBEMA, J. D. F., HERMANS, J. and VAN DEN BROCK, K. (1977). Selection of variables in discriminant analysis by F-statistic and error rate. *Technometrics* **19** 487–493.

HALL, P. (1981). Laws of the iterated logarithm for nonparametric density estimators. *Z. Wahrsch. verw. Gebiete* **56** 47–61.

HALL, P. (1982). Cross-validation in density estimation. *Biometrika* **69** 1–12.

KRONMAL, R. and TARTER, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Statist. Assoc.* **63** 925–952.

LAWLESS, J. F. (1981). Mean squared error properties of generalized ridge estimators. *J. Amer. Statist. Assoc.* **76** 462–466.

SCHUSTER, E. F. and GREGORY, G. G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. In: *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface.* Ed. W. F. Eddy. Springer-Verlag, New York, 295–298.

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.

SILVERMAN, B. W. (1978). Choosing the window width when estimating a density. *Biometrika* **65** 1–11.

STEIN, E. M. (1970). Singular integrals and differentiability properties of functions. Princeton University Press, Princeton, N.J.

STONE, M. (1974). Cross-validation and multinomial prediction. *Biometrika* **61** 509–515.

STONE, M. (1977). Asymptotics for and against cross-validation. *Biometrika* **64** 29–35.

UTRERAS, F. (1979). Cross-validation techniques for smoothing spline functions in one or two dimension. In: *Smoothing Techniques for Curve Estimation.* Ed. T. Gasser and M. Rosenblatt. Springer-Verlag, Berlin, 196–232.

WAHBA, G. (1977). Optimal smoothing of density estimates. In: *Classification and Clustering*. Ed. J. Van Ryzin, Academic, New York. 423–458.

WAHBA, G. (1981). Data-based optimal smoothing of orthogonal series density estimates. *Ann. Statist.* **9** 146–156.

WAHBA, G. and WOLD, S. (1975). A completely automatic French curve: Fitting spline functions by cross validation. *Comm. Statist.* **4** 1–17.

WONG, W. H. (1979). Expected information criterion for the smoothing parameter of density estimates: an elucidation of the modified likelihood. Technical Report No. 589, Dept. of Statistics, Univ. of Wisconsin, Madison, Wisconsin.

YUN-SHYONG CHOW
INSTITUTE OF MATHEMATICS
ACADEMIA SINICA
TAIPEI, TAIWAN, R.O.C.

STUART GEMAN
DIVISION OF APPLIED MATHEMATICS
BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND 02912

LI-DE WU
DEPT. OF COMPUTER SCIENCE
FUDAN UNIVERSITY
SHANGHAI, PEOPLE'S REPUBLIC OF CHINA