*The Annals of Statistics*
1988, Vol. 16, No. 4, 1421–1427

# ADMISSIBLE KERNEL ESTIMATORS OF A MULTIVARIATE DENSITY

By Daren B. H. Cline

*Texas A & M University*

A kernel density estimator is defined to be admissible if no other kernel estimator has (among all densities and sample sizes) uniformly smaller mean integrated squared error. Admissible kernel density estimators are precisely those using kernels with nonnegative Fourier transforms bounded by 1. Several examples are given.

**Introduction.** The kernel estimator of a multivariate density $f$ on $\mathbb{R}^d$ is given by

$$\hat{f}_{n,\kappa}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\kappa(\mathbf{x}-\mathbf{X}_i),$$

where the kernel $\kappa$ is integrable and $X_1,\dots,X_n$ is a random sample from $f$. A commonly used measure of the global efficiency of this estimator is its mean integrated squared error

$$\operatorname{MISE}(\hat{f}_{n,\kappa}) = E\int(\hat{f}_{n,\kappa}(\mathbf{x}) - f(\mathbf{x}))^2\,d\mathbf{x}.$$

This of course requires that both $f$ and $\kappa$ are square integrable. MISE may be expressed as the sum of two portions, the integrated variance

$$\operatorname{IV}(\hat{f}_{n,\kappa}) = \int \operatorname{Var}(\hat{f}_{n,\kappa}(\mathbf{x}))\,d\mathbf{x}$$

and the integrated squared bias

$$\operatorname{ISB}(\hat{f}_{n,\kappa}) = \int(E\hat{f}_{n,\kappa}(\mathbf{x}) - f(\mathbf{x}))^2\,d\mathbf{x}.$$

An early paper by Watson and Leadbetter (1963) determined the kernel which minimizes MISE. (Their argument is valid for multivariate densities.) This kernel, however, is specific to the density and thus cannot have practical use. Numerous authors have suggested choices for sequences of kernels which at least obtain the optimal convergence rate for MISE within a class of densities. Perhaps the most familiar is the sequence generated by the kernel of Epanechnikov (1969) and its multivariate extension [Sacks and Ylvisaker (1981)] which are asymptotically optimal among nonnegative kernels for twice differentiable densities.

Presumably, the practitioner also would appreciate knowing whether his choice of kernel can be improved upon uniformly among all densities and sample

Received May 1987; revised March 1988.
*AMS* 1980 *subject classifications.* Primary 62G05; secondary 62C15.
*Key words and phrases.* Kernel density estimation, admissibility.

sizes. If it cannot, then at least some optimality is assured for finite sample size, without restricting the class of densities. With this in mind, we say a kernel $\kappa$ is *admissible* if the only kernel $\kappa_1$ satisfying

$$\mathrm{MISE}\big(\hat{f}_{n,\kappa_1}\big) \le \mathrm{MISE}\big(\hat{f}_{n,\kappa}\big), \quad \text{for all } n \text{ and } f,$$

is $\kappa_1 = \kappa$. A kernel estimator is admissible if it uses an admissible kernel.

In Theorem 1, we will characterize admissible kernels as those which have nonnegative Fourier transforms bounded by 1. Admissibility thus implies symmetry, a traditional requirement. Another property of an admissible kernel is that any rescaling of the kernel is again admissible. This allows the practitioner to choose an admissible kernel shape without knowing the best choice of bandwidth in advance. Thus, by choosing an admissible kernel shape with appropriate other properties, one may devise a sequence of admissible estimators with asymptotically optimal convergence rate.

Strictly speaking, our definition does not require the kernel to integrate to 1, a property that is necessary for consistency. However, any nonzero admissible kernel can be normalized to integrate to 1 and it will remain admissible. Requiring the kernel to integrate to 1 is perhaps analogous to requiring invariance in parametric statistics and at least avoids nonsensical, if admissible, estimators. An admissible estimator which does not integrate to 1 is a kind of shrinkage estimator which apparently attempts to estimate extremely flat densities and densities with mass near infinity.

On the other hand, Professor Jeff Hart points out to us that *all* nonparametric density estimators are shrinkage estimators which shrink towards smoothness. Avoiding the large variance (roughness) term due to the high dimensional parameter space, like James–Stein estimation, actually entails a biased estimator. The distinction, apparently, between parametric and nonparametric estimation is that in the former problem one chooses between unbiased estimators and accepting a bias in return for a reduction in variance, while in the latter problem one views bias as a necessity. In both, however, admissibility provides one criterion for choosing among shrinkage estimators. In fact, as Theorem 1 will also show, one can choose an admissible estimator over an inadmissible estimator and thereby reduce *both* integrated squared bias and integrated variance.

Other standard properties frequently required of kernels include unimodality and nonnegativity or a specific number of nonzero moments, depending on the uses for the estimate. Although admissible kernels which are unimodal probability densities form a rich class, an issue is the fact that the Epanechnikov kernel is not admissible among all kernels. That is, there exist kernels which have uniformly smaller MISE. Clearly such kernels must have some negative values since otherwise the Epanechnikov kernel would not be the unique asymptotically optimal kernel (for twice differentiable densities) among *nonnegative* kernels. Thus, this kernel can be said to be admissible among nonnegative kernels.

Similar other restrictions on a kernel, such as the number of sign changes or the number of vanishing moments, can also result in asymptotically optimal choices which are not admissible among the class of all kernels. When those restrictions are desired and the class of acceptable kernels is thus smaller, the admissibility criterion accordingly can be weakened. These evidently weaker forms of admissibility are not characterized here.

The next section proves our main result and corollaries. The final section provides examples.

**Results.** We start with a characterization of admissible kernels.

THEOREM 1. *Admissible kernels are precisely those whose Fourier transforms are nonnegative and bounded by* 1. *Furthermore, for any kernel* $\kappa$ *with transform* $\psi$, *let the kernel* $\kappa_1$ *have Fourier transform*

$$\psi_1(\mathbf{t}) = \max[0, \min[1, \mathrm{Re}(\psi(\mathbf{t}))]]. \tag{1}$$

*Then* $\kappa_1$ *is admissible with*

$$\mathrm{IV}\big(\hat{f}_{n,\kappa_1}\big) \le \mathrm{IV}\big(\hat{f}_{n,\kappa}\big) \tag{2}$$

*and*

$$\mathrm{ISB}\big(\hat{f}_{n,\kappa_1}\big) \le \mathrm{ISB}\big(\hat{f}_{n,\kappa}\big). \tag{3}$$

PROOF. Let $\varphi(\mathbf{t})$ be the characteristic function for $f$ and $\psi(\mathbf{t})$ be the Fourier transform for $\kappa$. From Plancherel's identity [cf. Watson and Leadbetter, (1963)],

$$\mathrm{IV}\big(\hat{f}_{n,\kappa}\big) = \frac{(2\pi)^{-d}}{n} \int |\psi(\mathbf{t})|^2 (1 - |\varphi(\mathbf{t})|^2)\, d\mathbf{t} \tag{4}$$

and

$$\mathrm{ISB}\big(\hat{f}_{n,\kappa}\big) = (2\pi)^{-d} \int |1 - \psi(\mathbf{t})|^2 |\varphi(\mathbf{t})|^2\, d\mathbf{t}. \tag{5}$$

Let $\kappa$ be any square integrable kernel and let $\kappa_1$ satisfy (1). Then

$$|\psi|^2 - |\psi_1|^2 = \begin{cases} \mathrm{Re}(\psi)^2 + \mathrm{Im}(\psi)^2, & \text{if } \mathrm{Re}(\psi) < 0, \\ \mathrm{Im}(\psi)^2, & \text{if } 0 \le \mathrm{Re}(\psi) \le 1, \\ \big(\mathrm{Re}(\psi)^2 - 1\big) + \mathrm{Im}(\psi)^2, & \text{if } 1 < \mathrm{Re}(\psi) \end{cases} \tag{6}$$

$$\ge 0.$$

From (4) and (6) we see that (2) holds, with equality for all $f$ and $n$ only if $\psi = \psi_1$. Also,

$$|1 - \psi|^2 - |1 - \psi_1|^2 = \begin{cases} (1 - \mathrm{Re}(\psi))^2 - 1 + \mathrm{Im}(\psi)^2, & \text{if } \mathrm{Re}(\psi) < 0, \\ \mathrm{Im}(\psi)^2, & \text{if } 0 \le \mathrm{Re}(\psi) \le 1, \\ (\mathrm{Re}(\psi) - 1)^2 + \mathrm{Im}(\psi)^2, & \text{if } 1 < \mathrm{Re}(\psi) \end{cases} \tag{7}$$

$$\ge 0.$$

From (5) and (7), we obtain (3). Again we have equality for all $f$ and $n$ only if $\psi = \psi_1$.

Furthermore, if $\kappa$ is admissible then (2) and (3) must be equalities and $\kappa = \kappa_1$. That is, $\kappa$ must have a nonnegative Fourier transform bounded by 1.

Conversely, suppose $\kappa$ has transform $\psi$, with $0 \le \psi \le 1$. Let $\kappa_1$ be any kernel satisfying

$$\mathrm{MISE}\big(\hat{f}_{n,\kappa_1}\big) \le \mathrm{MISE}\big(\hat{f}_{n,\kappa}\big), \quad \text{for all } f \text{ and } n,$$

and having transform $\psi_1$. By the argument above, we may assume $0 \le \psi_1 \le 1$. Thus,

$$0 \le \frac{1}{n}\int(\psi_1^2 - \psi^2)(1 - |\varphi|^2) + \int\big((1-\psi_1)^2 - (1-\psi)^2\big)|\varphi|^2, \tag{8}$$

for all $f$ and $n$. Let $n \to \infty$. Then

$$0 \le \int(\psi - \psi_1)(2 - \psi - \psi_1)|\varphi|^2.$$

Since the class of normal characteristic functions spans the class of symmetric bounded continuous functions, it follows that $\psi \ge \psi_1$. Now multiply both sides of (8) by $n$, choose $|\varphi(\mathbf{t})|^2 = \exp(-\|n\mathbf{t}\|^2)$ and again let $n \to \infty$. Then

$$0 \le \int(\psi_1^2 - \psi^2).$$

Since $\psi \ge \psi_1$, this clearly implies $\psi_1 = \psi$. Thus $\kappa$ is admissible. □

Several corollaries are readily available from the theorem or its proof. The first refers to the fact mentioned earlier that admissibility is independent of the bandwidth choice.

Corollary 1. *Any rescaling of an admissible kernel is admissible.*

Proof. If $\psi$ is nonnegative and bounded by 1, then $\psi(h_1 t_1, \dots, h_d t_d)$ certainly is for all positive bandwidths $h_j$. □

Convolving a kernel with its reflection across 0 is equivalent to squaring the modulus of its Fourier transform. This leads to the following observation.

Corollary 2. *Let $\kappa$ be any square integrable kernel which integrates to no more than* 1 *and define*

$$\kappa_1(\mathbf{x}) = \int\kappa(\mathbf{y} + \mathbf{x})\kappa(\mathbf{y})\,d\mathbf{y}.$$

*Then $\kappa_1$ is admissible.*

The theorem supports the traditional practice of using symmetric kernels. In fact, a similar proof shows that any asymmetric kernel may always be uniformly improved by a simple reflection, as is stated next.

COROLLARY 3. *Let $\kappa$ be any square integrable kernel and let*

$$\kappa_1(\mathbf{x}) = (\kappa(\mathbf{x}) + \kappa(-\mathbf{x}))/2.$$

*Then* (2) *and* (3) *hold.*

The *order* of a kernel with transform $\psi$ is the largest $\nu$ such that $\|\mathbf{t}\|^{-\nu}(1 - \psi(\mathbf{t}))$ is bounded near 0. This is a measure of the smoothness of the Fourier transform at 0 and hence is related to the number of vanishing moments of the kernel. Higher order kernels are used to obtain faster convergence rates for MISE, if the density is sufficiently smooth and if the sequence of bandwidths is chosen appropriately. That is, the possible rate of convergence depends on the order of the kernel. The next corollary will emphasize, however, that although the conversion in (1) will improve efficiency, it will not improve the convergence rate.

COROLLARY 4. *Suppose $\kappa$ is a symmetric kernel which integrates to* 1 *and $\kappa_1$ is given by* (1). *Then $\kappa$ and $\kappa_1$ are of the same order.*

PROOF. Under the assumptions, formula (1) does not alter $\psi$ near 0. □

**Examples.** Apart from the entirely useless kernel which is identically zero, the simplest admissible kernel is the pyramid,

$$\kappa(\mathbf{x}) = \prod_{j=1}^{d} \max(0, 1 - |x_j|).$$

This kernel is nonnegative, but has an infinity of modes. Examples of admissible nonnegative unimodal kernels include the centered normal, Laplace and logistic densities, as well as convolutions of these.

The optimal kernel estimator of Watson and Leadbetter (1963) is admissible but, again, it depends on the unknown density. The asymptotically optimal kernels of Parzen (1958), Watson and Leadbetter (1963) and Cline (1987) are also admissible. Their multivariate versions have transforms

$$\psi(\mathbf{t}) = (1 + \|\mathbf{t}\|^{\rho})^{-1}, \quad \text{for some } \rho > 1.$$

These kernels are nonnegative and unimodal if $\rho \leq 2$ [Lukacs (1970)]. The Laplace densities correspond to the case $\rho = 2$ and are asymptotically optimal for bounded, but discontinuous densities [van Eeden (1985) and Cline (1987)]. Similarly, $\rho = 4$ and $\rho = 6$, respectively, are optimal for densities with bounded but discontinuous first and second derivatives [Cline (1987), see also Silverman (1984), page 910, for their use in nonparametric regression]. These kernels have only slightly negative sidelobes and damp exponentially.

The Fourier integral estimator uses an infinite order kernel whose transform is nonnegative,

$$\psi(\mathbf{t}) = 1_{\|\mathbf{t}\| \leq 1}. \tag{9}$$

This could still be considered admissible under the criterion in the theorem, even though the kernel is not integrable. Davis (1977) demonstrated that the Fourier integral estimator will achieve optimal convergence rates under a wide variety of smoothness assumptions on the (univariate) density. In fact this property relies on the flatness of the transform of (9) at 0. [See Devroye and Györfi (1985), page 135.] Thus an example of an integrable and admissible kernel with the same property is the kernel with transform

$$\psi(\mathbf{t}) = \begin{cases} 1, & \text{if } \|\mathbf{t}\| \le \alpha - 1, \\ \alpha - \|\mathbf{t}\|, & \text{if } \alpha - 1 < \|\mathbf{t}\| \le \alpha, \\ 0, & \text{if } \alpha < \|\mathbf{t}\|. \end{cases}$$

In one dimension, this kernel is the difference of two Bartlett kernels,

$$\kappa(x) = \frac{1 - \cos \alpha x}{\pi x^2} - \frac{1 - \cos(\alpha - 1)x}{\pi x^2}.$$

Some familiar kernels are not admissible. For example, the parabolic kernel [Epanechnikov (1969)] and its extension for multivariate density estimators,

$$\kappa(\mathbf{x}) = \frac{\Gamma(2 + d/2)}{\pi^{d/2}} \max(0, 1 - \|\mathbf{x}\|^2),$$

given by Sacks and Ylvisacker (1981), clearly are not admissible. Of course, these were chosen to optimize the *asymptotic* MISE under the restrictions that the kernel be nonnegative and that the density be twice continuously differentiable. If using a nonnegative kernel is of primary concern, then the Epanechnikov kernel has the asymptotic edge and no other nonnegative kernel is uniformly better for all finite sample sizes. On the other hand, there are often good reasons for using kernels which take negative values. For example, one can achieve a convergence rate of $n^{-5/6}$, even without assuming a continuous second derivative, by using a kernel of order 3 or greater [Cline and Hart (1986) and Cline (1987)]. Devroye and Györfi [(1985), page 247] point out that nonnegative kernels can lead to estimates with spurious "bumps," even if the density has only moderate tails, such as the normal density. Various authors have pointed out that appropriate negative valued kernels will reduces bias at modes or help locate discontinuities. [For a short discussion, see Silverman (1986), pages 69–70.]

Similar to the Epanechnikov kernel, the optimal kernels obtained by Gasser, Müller and Mammitzsch (1985) and by Müller (1984) are not admissible among all kernels. These kernels were required to minimize the asymptotic MISE under restrictions which effectively limit the number of sign changes the kernel makes (and thus are more appealing), even though the density may be smooth enough to allow a higher order kernel. Like the Epanechnikov kernel, they are admissible among their respective restricted classes of kernels.

**Acknowledgments.** The author expresses his gratitude to Professors Jeffrey D. Hart and Raymond J. Carroll for the benefit of frequent fruitful discussions and to the referees for points of clarification. Professor J. Steven Marron provided the location of the Devroye and Györfi references and additional comments.

## REFERENCES

CLINE, D. B. H. (1987). Optimal kernel estimation of densities. Technical Report 5, Dept. Statistics, Texas A & M Univ.

CLINE, D. B. H. and HART, J. D. (1986). Kernel estimation of densities with discontinuities or discontinuous derivatives. Technical Report 9, Dept. Statistics, Texas A & M Univ.

DAVIS, K. B. (1977). Mean integrated square error properties of density estimates. *Ann. Statist.* **5** 530–535.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The $L_1$ View.* Wiley, New York.

EPANECHNIKOV, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl.* **14** 153–158.

GASSER, T., MÜLLER, H.-G. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238–252.

LUKACS, E. (1970). *Characteristic Functions.* Griffin, London.

MÜLLER, H.-G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Statist.* **12** 766–774.

PARZEN, E. (1958). On asymptotically efficient consistent estimates of the spectral density of a stationary time series. *J. Roy. Statist. Soc. Ser. B* **20** 303–322.

SACKS, J. and YLVISAKER, D. (1981). Asymptotically optimum kernels for density estimation at a point. *Ann. Statist.* **9** 334–346.

SILVERMAN, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Ann. Statist.* **12** 898–916.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

VAN EEDEN, C. (1985). Mean integrated squared error of kernel estimators when the density and its derivative are not necessarily continuous. *Ann. Inst. Statist. Math.* **37** 461–472.

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

DEPARTMENT OF STATISTICS
TEXAS A & M UNIVERSITY
COLLEGE STATION, TEXAS 77843