ON A SMOOTHING KERNEL OF INFINITE ORDER

by

D. N. Politis and Joseph P. Romano
Purdue University Stanford University

Technical Report #92-53

Department of Statistics
Purdue University

November 1992

On a smoothing kernel of infinite order

Dimitris N. Politis
Department of Statistics
Purdue University
W. Lafayette, IN 47907

Joseph P. Romano
Department of Statistics
Stanford University
Stanford, CA 94305

Abstract

In this note, a family of kernels of 'infinite order' is introduced. The resulting nonparametric density estimators have bias of order $O(1/M^r)$, where r can be intuitively interpreted as the number of continuous derivatives the unknown probability density f possesses, and M^{-1} is the bandwidth. These kernels may be obtained as a convex combination of kernels without moments. This may appear paradoxical because a convex combination of kernels possessing moments may be utilized to reduce bias by only a factor of M^{-2}.

Keywords. Bias reduction, kernel smoothing, mean squared error, nonparametric estimation, probability density.

1. Introduction.

Suppose $X_1, \ldots, X_N$ are real-valued, independent identically distributed observations, from a population with absolutely continuous distribution function F, and probability density function f. The density f is known to possess some smoothness, but is otherwise unknown and should be estimated using the data. In particular, it will be assumed that the characteristic function $\phi(s) = \int_{-\infty}^{\infty} e^{is2\pi x} f(x)dx$, satisfies $\int_{-\infty}^{\infty} |s|^r |\phi(s)| ds < \infty$, for some positive integer r; this implies that f has r bounded and continuous derivatives $f^{(1)}, \ldots, f^{(r)}$.

The nonparametric kernel smoothed estimator of $f(x)$, for some $x \in R$, is given by

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \Lambda\left(x - X_i\right) = \int_{-\infty}^{\infty} \lambda(s) \phi_N(s) e^{-is2\pi x} ds, \tag{1}$$

where $\Lambda(\cdot)$ is the smoothing kernel, satisfying $\int_{-\infty}^{\infty} \Lambda(x)dx = 1$, $\phi_N(s) = \frac{1}{N}\sum_{i=1}^{N} e^{is2\pi X_i}$ is the sample characteristic function, and $\lambda(s) = \int_{-\infty}^{\infty} \Lambda(x) e^{is2\pi x} dx$ is the Fourier transform of the kernel. In general, $\Lambda(\cdot)$ and $\lambda(\cdot)$ both depend on a parameter M although it will not be explicitly denoted, ($1/M$ is usually called the 'bandwidth' of the kernel); in particular, it will be assumed that $\Lambda(x) = M\Psi(Mx)$, where $\Psi(\cdot)$ is some fixed (not depending on M) given function. In the asymptotic results below, M depends on N and $M \to \infty$ as $N \to \infty$.

It is well known (cf., for example, Rosenblatt (1991)) that in this case

$$Var(\hat{f}(x)) = \frac{M}{N} f(x) \int_{-\infty}^{\infty} \Psi^2(x)dx + o(M/N), \tag{2}$$

as $N \to \infty$ and $M \to \infty$, but with $M/N \to 0$. The 'order' of the kernel is the maximum integer q, such that $\int_{-\infty}^{\infty} x^l \Lambda(x)dx = 0$, for $l = 1, \ldots, q-1$; it then follows, provided $\int_{-\infty}^{\infty} |x^l \Lambda(x)| dx < \infty$, for $l = 1, \ldots, q$, that

$$Bias(\hat{f}(x)) = \frac{1}{M^k k!} f^{(k)}(x) \int_{-\infty}^{\infty} x^k \Psi(x)dx + o(1/M^k), \tag{3}$$

where $k = \min(q, r)$. If, in fact, $r \geq q + 2$, then the error term in (3) can be replaced by $O(M^{-(k+2)})$; see Theorem 1 of Schucany and Sommers (1977). This idea of choosing a kernel of order q in order to get the $Bias(\hat{f}(x))$ to be $O(1/M^k)$ seems to date back to Parzen (1962) and Bartlett (1963).

Note that the asymptotic order of the bias is limited by the order of the kernel if the true spectral density is very smooth, i.e., r is large. It would be reasonable to say that a kernel has *infinite order* if it results in an estimator with bias of order $O(1/M^r)$, for *any* given r. In this correspondence a family of such kernels will be constructed.

2. Construction of kernels of order $q+2$ from kernels of order q.

Typically, kernels of order two are utilized since they correspond to probability densities. We now show how, given a symmetric kernel of order q, a symmetric kernel of order $q+2$ may be constructed, thus leading to a density estimator with smaller bias; see equation (3). The method has been described in Schucany and Sommers (1977) to construct kernels of order six, where it is suggested the method is more general. We discuss such a generalization explicitly. By starting with a symmetric kernel of order two and iterating the construction below, a symmetric kernel of arbitrary high (but finite) order may be determined.

Suppose Ψ is a symmetric kernel of order q. Let $I_j(\Psi) = \int x^j \Psi(x)dx$, which is assumed finite for $j = q+2$. Also, assume $r \geq q+2$. Consider a linear combination of kernels:

$$\Psi_{\alpha,c}(x) = \alpha\Psi(x) + (1-\alpha)c\Psi(cx) \tag{4}$$

for some constants α and c with $c > 0$. The estimator $\hat{f}$ defined in (1) with $\Lambda(x) = M\Psi_{\alpha,c}(Mx)$ has bias

$$\frac{f^{(q)}(x)I_q(\Psi)}{M^q q!} \cdot [\alpha + (1-\alpha)c^{-q}] + O(M^{-(q+2)}),$$

by equation (3). Hence, for a given choice of c, the choice of α satisfying $\alpha + (1-\alpha)c^{-q} = 0$ results in a kernel so that the corresponding density estimator has bias of order $M^{-(q+2)}$; specifically, $\alpha = c^{-q}/[c^{-q}-1]$.

3. A family of smoothing kernels of infinite order.

Suppose now you start with the kernel $\Psi(x) = \sin^2(\pi x)/(\pi^2 x^2)$, a symmetric kernel of order two. Let $c = 1/2$, for example, and apply the construction of the previous section to yield a kernel of order four by taking $\Psi_{\alpha,c}$ with $\alpha = 4/3$. In fact, the argument based on (3) leading to $\Psi_{4/3,1/2}$ having bias of order M^{-4} breaks down because Ψ does not have any positive moments. Paradoxically, we show that, by starting with such a Ψ, the kernel $\Psi_{2,1/2}$ has bias $o(M^{-r})$; that is, an appropriate linear combination of $\Psi(x)$ and $\Psi(x/2)/2$ already yields a kernel with a full bias correction! Notice that this choice of $\alpha = 2$ for the fixed choice $c = 1/2$ is different than the choice $\alpha = 4/3$ that eliminates the dominant term in the bias expansion when the initial kernel does have moments.

More generally, for some choice of the parameter $h > 0$, define the Fourier pair $\Lambda_h(x)$ and $\lambda_h(s)$ satisfying $\Lambda_h(x) = \int_{-\infty}^{\infty} \lambda_h(s) e^{-is2\pi x} ds$, where

$$\lambda_h(s) = \begin{cases} 1 & \text{for } |s| \le m \\ 1 - \frac{|s|-m}{M-m} & \text{for } m < |s| \le M \\ 0 & \text{for } |s| > M; \end{cases}$$

here $m = Mh/(h+1)$, and

$$\Lambda_h(x) = \frac{\sin^2(\pi x M) - \sin^2(\pi x m)}{\pi^2 x^2 (M-m)}. \tag{5}$$

Some motivation for the introduction of kernel $\Lambda_h(x)$, for $h > 0$, is given in Politis and Romano (1992).

It is obvious that $\Lambda_0(x) = \sin^2(\pi x M)/(\pi^2 x^2 M)$, is just the Fejér kernel, and that $\Lambda_h(x) = (h+1)\sin^2(\pi x M)/(\pi^2 x^2 M) - h \sin^2(\pi x m)/(\pi^2 x^2 m)$, i.e., $\Lambda_h(x)$ is a linear combination of Fejér kernels with different bandwidths. From the fact (cf. Papoulis (1962)) that the Fejér kernel integrates to unity, it follows that $\int_{-\infty}^{\infty} \Lambda_h(x) dx = 1$ as well; similarly, using the properties of the Fejér kernel, $\int_{-\epsilon}^{\epsilon} \Lambda_h(x) dx \to 1$, as $M \to \infty$, for any $\epsilon > 0$.

Note that, for any $h > 0$, the function $\lambda_h(s)$ is of trapezoidal shape, with all derivatives existing and equal to zero at the origin $s = 0$. It follows (cf. Papoulis (1962)) that $\int_{-\infty}^{\infty} x^l \Lambda_h(x) dx = 0$, for *any* positive integer l; note however that these integrals should be interpreted as Cauchy principal values because the integrals $\int_{-\infty}^{\infty} |x^l||\Lambda_h(x)| dx$ are not finite. For this reason, equation (3) can not be applied to claim that the bias of $\hat{f}_h(x) = \frac{1}{N}\sum_{i=1}^{N} \Lambda_h\left(x - X_i\right)$ is $O(1/M^r)$, for any given r. Nevertheless, this is a true statement as demonstrated by the following theorem.

Theorem 1 *Let x be a real number, and $h > 0$. Assume $\int_{-\infty}^{\infty} |s|^r|\phi(s)|ds < \infty$, for some positive integer r; also assume that $m = Mh/(h+1)$, and that $M \to \infty$, as $N \to \infty$, but with $M/N \to 0$. Then*

$$Bias(\hat{f}_h(x)) = o(1/M^r). \tag{6}$$

Proof. Observe that

$$Bias(\hat{f}_h(x)) \equiv E\hat{f}_h(x) - f(x) = A_1 + A_2$$

where

$$A_1 = \int_{-M}^{M} \left(\lambda_h(s) - 1\right)\phi(s)e^{-is2\pi x}ds$$

$$A_2 = -\int_{|s|\geq M} \phi(s)e^{-is2\pi x}ds.$$

But $|A_2| \leq \int_{|s|\geq M} |\phi(s)|ds \leq M^{-r}\int_{|s|\geq M} |s|^r|\phi(s)|ds = o(1/M^r)$, since $\int_{-\infty}^{\infty} |s|^r|\phi(s)|ds < \infty$. To complete the proof of equation (6), note that A_1 can be split into two terms, $A_1 = a_1 + a_2$, where

$$a_1 = \int_{|s|\leq m} \left(\lambda(s) - 1\right)\phi(s)e^{-is2\pi x}ds$$

$$a_2 = \int_{m<|s|\leq M} \left(\lambda(s) - 1\right)\phi(s)e^{-is2\pi x}ds.$$

First observe that $a_1 = 0$, because $\lambda(s) = 1$ for $|s| \leq m$. Now

$$|a_2| \leq \int_{m<|s|\leq M} |\lambda(s) - 1||\phi(s)|ds.$$

But $\lambda(s) = 1 - \frac{|s|-m}{M-m}$ for $m < |s| \leq M$. Thus,

$$|a_2| \leq \int_{m<|s|\leq M} \frac{|s| - m}{M - m}|\phi(s)|ds.$$

It is obvious that if $r = 1$, then $a_2 = o(1/M)$. On the other hand, if $r > 1$, we have

$$|a_2| \leq \frac{1}{m^{r-1}} \int_{m<|s|\leq M} |s|^{r-1} \frac{|s|-m}{M-m} |\phi(s)| ds = o(1/M^r),$$

where it was used that $\int_{-\infty}^{\infty} |s|^r |\phi(s)| ds < \infty$, and that both m and $M - m$ are asymptotically proportional to M. **QED.**

Remark 1. That the bias of $\hat{f}_h(x)$ turns out to be $o(1/M^r)$ instead of just $O(1/M^r)$ should not be surprising as it was mentioned that the assumption $\int_{-\infty}^{\infty} |s|^r |\phi(s)| ds < \infty$ is stronger than assuming f has r bounded and continuous derivatives $f^{(1)}, \ldots, f^{(r)}$. However, it is not much stronger; for example, it is satisfied if it is assumed that f has r absolutely integrable derivatives, and the the rth derivative $f^{(r)}$ satisfies a uniform Lipschitz condition of order $\alpha > 1/2$. Note that since f is absolutely integrable, the characteristic function $\phi(s)$ is continuous, and satisfies $\phi(s) \to 0$, as $|s| \to \infty$ (by the Riemann-Lebesgue lemma). Hence, the assumption $\int_{-\infty}^{\infty} |s|^r |\phi(s)| ds < \infty$ is actually equivalent to $\int_{-\infty}^{\infty} |s|^l |\phi(s)| ds < \infty$, for $l = 0, 1, \ldots, r$.

Remark 2. The asymptotic variance of $\hat{f}_h(x)$ can be calculated from equation (2). To compute $\int_{-\infty}^{\infty} \Lambda_h^2(x) dx$ in the case $M = 1$, it is easier to use the isometric properties of the Fourier transform, i.e., Parseval's theorem; one then finds that

$$Var(\hat{f}_h(x)) = \frac{2M}{3N} \left(\frac{3h+1}{h+1} \right) f(x) + o(M/N). \tag{7}$$

It follows from Theorem 1 that if the unknown density is quite smooth, that is, if r in the assumptions of the theorem is large, then the bias of $\hat{f}_h(x)$ will be of very small order, even when the bandwidth $1/M$ is not small. Note that due to equation (2) the rate of convergence of any kernel smoothed estimator $\hat{f}(x)$ is $\sqrt{N/M} << \sqrt{N}$, since one generally has to let $M \to \infty$ to make the asymptotic bias of $\hat{f}(x)$ vanish. In this connection, it is interesting to point out that the estimator $\hat{f}_h(x)$ outperforms any estimator corresponding to a kernel of finite order in case the true density is extremely smooth, satisfying $\phi(s) = 0$, for all $s >$ some s_0. Consistency of $\hat{f}_h(x)$ is then achieved with a *fixed* bandwidth, and $\hat{f}_h(x)$ is *finite-sample* unbiased;

therefore, $\hat{f}_h(x)$ is actually $\sqrt{N}$-consistent in this case.

Theorem 2 *Let x be a real number, and $h > 0$. Assume $\phi(s) = 0$, for all $|s| >$ some s_0; also assume that M is a fixed* constant *satisfying $s_0 \leq m = Mh/(h+1)$. Then $Bias(\hat{f}_h(x)) = 0$, for all $N > M$.*

Proof. Consider the decomposition of the $Bias(\hat{f}_h(x))$ in the proof of Theorem 1, and note that now the terms A_2 and a_2 are both exactly equal to zero, because $\phi(s) = 0$ for all $|s| \geq M > m \geq s_0$. **QED.**

Remark 3. It should be noted that the class of densities satisfying $\phi(s) = 0$, for all $|s| >$ some s_0 is quite large. For example, densities that are mixtures of Fejér kernels, satisfy this requirement; in fact, by taking a mixture of Fejér kernels with different bandwidths, a density can be constructed that approximates arbitrarily closely any given density with characteristic function that is symmetric, and convex on the positive half-axis. This is known in the literature as Pólya's construction of convex characteristic functions.

Of course, such densities have no moments; To construct densities possessing finite moments that satisfy $\phi(s) = 0$, for $|s| >$ some s_0, one can take mixtures of integer powers of Fejér kernels with different bandwidths. For example, the density given by the square of the Fejér kernel normalized, i.e., $\Lambda_0^2(x)/\int_{-\infty}^{\infty} \Lambda_0^2(t)dt$, has finite second moments, and a characteristic function that vanishes outside the interval $[-2M, 2M]$; this characteristic function is usually called Parzen's window in the time series literature, and is given by a convolution of the triangular window $\lambda_0(\cdot)$ with itself. Convolving $\lambda_0(\cdot)$ with itself k times, where k is a positive integer, yields a valid characteristic function that vanishes outside the interval $[-kM, kM]$ and corresponds to a probability density possessing at least k finite moments; if k is large, this k-fold convolution of triangles will approximate to a Gaussian density, by the Central Limit Theorem.

4. Some concluding remarks.

Because of the simple trapezoidal shape of $\lambda_h(s)$, the actual computation of $\hat{f}_h(x)$ might be performed more easily by a Fourier transform of the sample characteristic function $\phi_N(s)$ after it is multiplied ('tapered') by $\lambda_h(s)$, that is, using the right-hand-side of equation (1); see Silverman (1986, p. 61) for details.

As a matter of course, the kernel $\Lambda_h(x)$ is not everywhere nonnegative. To ensure a strictly nonnegative estimator one would take $\hat{f}_h^+(x) = \max(\hat{f}_h(x), 0)$, which has the additional advantage of possessing smaller (or equal) mean squared error as compared to $\hat{f}_h(x)$, (cf., for example, Politis and Romano (1992)). Note that, if $f(x) > 0$, $\hat{f}_h(x)$ and $\hat{f}_h^+(x)$ have the same large-sample variance (by the δ-method), while if $f(x) = 0$, the asymptotic distribution of either $\sqrt{N/M}\hat{f}_h(x)$ or of $\sqrt{N/M}\hat{f}_h^+(x)$ degenerates to a point mass at zero.

An interesting question concerns the choice of h; the whole family $\mathcal{L} = \{\Lambda_h(\cdot), h > 0\}$ consists of kernels of infinite order, but in any given situation only one of these kernels will be used. Now it was mentioned that $\Lambda_0(x)$ is the Fejér kernel, and it can also be verified that $\Lambda_\infty(x)$ is the Dirichlet kernel which is equal to $\sin(2\pi xM)/\pi x$. The extreme points $\Lambda_h(\cdot)$ for $h = 0$ and $h = \infty$ are not considered to be in the family $\mathcal{L}$. In particular, the Fejér kernel has no finite moments and, regardless of the degree of smoothness of f, the $Bias(\hat{f}_0(x)) \sim c_{x,f}/M$, for some constant $c_{x,f} \neq 0$; on the other hand, although an argument similar to the proof of Theorem 1 actually goes through for the extreme case $h = \infty$, the Dirichlet kernel is well known to have many prominent positive and negative side-lobes which is a most undesirable feature. This observation indicates that taking h very small (close to 0) or very large should be avoided. A choice of h in the neighborhood of 1 seems about right; taking $h = 1$ results in a kernel $\Lambda_1(x)$ that has no significant positive side-lobes, and only one negative side-lobe of much smaller area than the main lobe at the origin.

In the case it is known that f has finite support, say the interval $[0,1]$, it follows that f can be reconstructed from the values of ϕ on the integers by the Fourier series $f(x) = \sum_{s=-\infty}^{\infty} \phi(s)e^{-is2\pi x}$, and can be estimated by $\hat{f}_h^{[0,1]}(x) = \sum_{s=-\infty}^{\infty} \lambda_h(s)\phi_N(s)e^{-is2\pi x}$. Since $\phi(s)$ is a continuous function of the real variable s, the assumption $\int_{-\infty}^{\infty} |s|^r|\phi(s)|ds < \infty$ implies that

$\sum_{s=-\infty}^{\infty} |s|^r |\phi(s)| < \infty$; hence, Theorems 1 and 2 are true *verbatim* for the estimator $\hat{f}_h^{[0,1]}(x)$ as well.

Last but not least it should be mentioned that choosing the bandwidth of a kernel estimator in practical applications is a most important and difficult problem. Adaptive (data-dependent) optimal bandwidth selection works asymptotically (see Woodroofe (1970), Hall (1983), Stone (1984)), but not much can be said regarding finite sample sizes. Nevertheless, Theorem 2 suggests a simple procedure for choosing the bandwidth of the estimator $\hat{f}_h(x)$ in practice; this procedure parallels ideas from time series analysis, namely examination of the correlogram, and can be described as follows.

Suppose that by looking at a plot of the magnitude of the sample characteristic function it is observed that $|\phi_N(s)| \simeq 0$, for all $s >$ some $\hat{s}_0$. Then it might be inferred that $\hat{s}_0$ is an estimate of s_0 appearing in the assumptions of Theorem 2. It would then follow that m may be taken equal to $\hat{s}_0$, M may be taken equal to $\hat{s}_0(h+1)/h$, and the resulting estimators $\hat{f}_h(x)$ and $\hat{f}_h^+(x)$ will both have bias of very small order, and variance approximately as given by equation (7).

References

[1] Bartlett,M.S. (1963), Statistical Estimation of Density Functions, *Sankhya, Ser. A*, 25, 245-54.

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

[3] Papoulis, A. (1962), *The Fourier Integral and its Applications*, McGraw-Hill, New York.

[4] Politis,D.N., and Romano,J.P. (1992), Bias-Corrected Nonparametric Spectral Estimation, Technical Report No. 92-50, Department of Statistics, Purdue University.

[5] Parzen, E. (1962), On Estimation of a Probability Density Function and its Mode, *Ann. Math. Statist.*, 33, 1065-1076.

[6] Rosenblatt, M. (1991), *Stochastic Curve Estimation*, NSF-CBMS Regional Conference Series vol. 3, Institute of Mathematical Statistics, Hayward.

[7] Schucany, W.R. and Sommers, J.P. (1977), Improvement of kernel type density estimators, *J. Amer. Statist. Assoc.*, vol. 72, 420-423.

[8] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

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

[10] Woodroofe, M. (1970), On choosing a delta-sequence, *Ann. Math. Statist.*, 41, 1665-1671.