The Canadian Journal of Statistics
Vol. 18, No. 3, 1990, Pages 197–204
La Revue Canadienne de Statistique

Gaussian-based kernels

Matthew P. WAND and William R. SCHUCANY

Texas A&M University and *Southern Methodist University*

Key words and phrases: Bias reduction, density derivative, density estimation, Fourier-transform methods, Hermite polynomials, mean square efficiency, window-width selection.
AMS 1985 subject classifications: Primary 62G05; secondary 62G20, 65D10.

ABSTRACT

We derive a class of higher-order kernels for estimation of densities and their derivatives, which can be viewed as an extension of the second-order Gaussian kernel. These kernels have some attractive properties such as smoothness, manageable convolution formulae, and Fourier transforms. One important application is the higher-order extension of exact calculations of the mean integrated squared error. The proposed kernels also have the advantage of simplifying computations of common window-width selection algorithms such as least-squares cross-validation. Efficiency calculations indicate that the Gaussian-based kernels perform almost as well as the optimal polynomial kernels when the order of the derivative being estimated is low.

RÉSUMÉ

On obtient une classe de noyaux d'ordre supérieur pour l'estimation des densités et de leurs dérivées. Cette classe peut être considérée comme une généralisation de la classe des noyaux gaussiens de deuxième ordre. Ces noyaux possèdent des propriétés attrayantes, ils sont réguliers et se prêtent bient à des calculs de convolutions et de transformées de Fourier. Le calcul exact de l'erreur quadratique intégrée moyenne peut être effectué lorsque ces noyaux sont utilisés. Ils simplifient l'application des algorithmes pour déterminer la largeur de bande. Des calculs d'efficacité montrent que les noyaux gaussiens considérés se comportent presque aussi bien que les noyaux polynômiaux optimaux, lorsque l'ordre de la dérivée estimée est petit.

1. INTRODUCTION

Kernel estimators are a widely accepted means of estimating probability densities without parametric assumptions; see Silverman (1986) for a thorough account of this methodoogy. Density derivatives can also be estimated by straightforward extension of kernel estimators.

Let $X_1, \ldots, X_n$ be a random sample having density f, and assume that f has $\nu + 2r$ continuous derivatives, where $\nu \geq 0$ and $r \geq 1$ are integers. A class of kernel estimators for $f^{(\nu)}$ is generated by taking the νth derivatives of the usual kernel density estimator and is given by

$$f_n^{(\nu)}(x) = n^{-1} h^{-(\nu+1)} \sum_{i=1}^{n} K_{2r}^{(\nu)}\left(\frac{x - X_i}{h}\right), \tag{1.1}$$

where K_{2r} is a $2r$th-order kernel, that is, K_{2r} has $2r - 1$ vanishing moments. We also assume that K_{2r} is bounded, is ν times differentiable, and satisfies $\int K_{2r} = 1$. We restrict attention to symmetric functions; then the odd moments are zero, and only even-order kernels are considered. It is well known that the large-sample performance of (1.1) is

enhanced by increasing the value of r. This is addressed in work by Parzen (1962), Barlett (1963), and Schucany and Sommers (1977). Assuming f has enough smoothness to allow one to use a higher-order kernel, the question arises whether one should actually do that. Hall and Marron (1988) study the general problem of selecting r based on minimum mean integrated squared error (MISE) and using least squares cross-validation.

The parameter $h = h(n)$, often referred to as the window width or bandwidth, is of fundamental importance to the performance of $f_n^{(\nu)}$, since it controls the tradeoff between bias and variance. In an attempt to reduce the subjectivity in choosing h, there recently have been several proposals for automatic selection of h. Many of these are reviewed in Marron (1988) and Park and Marron (1990). A feature of some of these selection rules is that they require the use of additional kernels, sometimes of higher order than the original kernel.

In the case where $r = 1$ (that is, $f^{(\nu)}$ has two continuous derivatives) a popular choice of kernel in (1.1) is the Gaussian kernel $\phi(z) = (2\pi)^{-\frac{1}{2}} e^{-z^2/2}$. This kernel has a number of attractive features: it is infinitely smooth, it is well suited to Fourier-transform techniques for rapid computation of the estimator, and it has simple convolution properties. Each of these features is particularly relevant to certain window-width selection procedures such as least-squares cross-validation, as discussed in Silverman (1986, pp. 61–66).

Another attractive feature of the Gaussian kernel is that it permits exact calculation of the MISE of the kernel estimator if one is estimating a normal density. This was shown by Fryer (1976) and Deheuvels (1977) and is also discussed by Silverman (1986, p. 37).

In this note our main objective is to extend the Gaussian second-order kernel to a class of kernels of order $2r$ for general $r \geq 1$ with the intention of preserving the smoothness and convolution properties of ϕ. We show that the appropriate $2r$th-order kernel is of the form $G_{2r} \equiv Q_{2r-2}\phi$, where Q_{2r-2} is a polynomial of degree $2r - 2$. These kernels can be interpreted in terms of the generalized jackknife as discussed by Schucany and Sommers (1977). The kernel G_{2r} also has a convenient representation in terms of higher derivatives of ϕ which is very useful for Fourier transforms and convolution formulae.

A general class of kernel estimators of $f^{(\nu)}$ was studied by Müller (1984) and Gasser, Müller, and Mammitzsch (1985). This class has the form

$$f_{n,\nu}(x) = n^{-1}h^{-(\nu+1)} \sum_{i=1}^{n} W_{\nu,k}\left(\frac{x - X_i}{h}\right), \tag{1.2}$$

where $k > \nu + 1$ is an integer and ν and k are either both even or both odd. The function $W_{\nu,k}$ satisfies

$$\int x^j W_{\nu,k}(x)\,dx = \begin{cases} 0 & 0 \leq j \leq k-1, \quad j \neq \nu, \\ (-1^{\nu}\nu! & j = \nu, \\ \beta_k \neq 0 & j = k. \end{cases}$$

The estimator at (1.1) is a special case of that in (1.2) with $W_{\nu,k} = K_{k-\nu}^{(\nu)}$. By considering the asymptotic mean integrated squared error of $f_{n,\nu}$, Gasser *et al.* derive optimal kernels for varying values of (ν, k) with the restriction that the kernel has compact support and a miminal number of sign changes. In the work by Müller an additional parameter μ, indicating the number of continuous derivatives of $W_{\nu,k}$, is taken into account in the minimization. These optimal kernels are polynomials on the interval $[-1, 1]$. However, polynomial kernels suffer from the fact that they do not, in general, have a reducible convolution representation or Fourier transform. This causes difficulties if one decides to use

the Fourier-transform methods of computation. Also, Härdle, Marron, and Wand (1990) report that the use of compact-support kernels sometimes led to numerical instabilities when applying their cross-validation algorithm for window-width selection. On the other hand, consideration of computational speed for direct evaluation of the estimator and, in regression problems, concerns about boundary bias argue for compact support.

A further class of higher-order kernels was proposed by Hall and Marron (1988), who, for theoretical convenience, considered the class of kernels given by

$$T_{2r}(x) = \pi^{-1} \int_0^\infty (\cos tx) e^{-t^{2r}}\, dt.$$

Note that T_{2r} is simply the inverse Fourier transform of $\kappa_{2r}(t) = e^{-t^{2r}}$. These kernels correspond to the Gaussian kernel in the second-order case but have no closed form for higher orders.

Section 2 covers the derivation of the class of higher-order Gaussian-based kernels. In Section 3 we present some efficiency calculations which indicate that there is only a small loss in efficiency when a Gaussian kernel is used, provided that the value of ν is low.

2. GAUSSIAN-BASED KERNELS

The motivation for using higher-order kernels is the reduction in the order of magnitude of the bias of the curve estimator, leading to a faster rate of convergence of the mean integrated squared error. In the context of density estimation this principle is discussed by Schucany and Sommers (1977). As an illustration of the generalized jackknife, they introduced a class of fourth-order kernels $\{K_{4,c}, c > 0\}$ that can be constructed from a second-order kernel K_2 via the formula

$$K_{4,c}(x) = \frac{K_2(x) - c^3 K_2(cx)}{1 - c^2}.$$

Therefore, there is a class of fourth-order kernels based on the Gaussian kernel having the form

$$G_{4,c} = \frac{\phi(x) - c^3 \phi(cx)}{1 - c^2}$$

for positive values of c.

Following the approach in Section 4 of Schucany (1989), it can be shown that the asymptotically optimal MISE of $f_n^{(0)}$ is proportional to the product of powers of the integrated square of $G_{4,c}$ and its fourth moment. An equivalent objective with respect to c can be shown to be

$$\frac{(1 + c^5) - 2^{2/3} c^3 (1 + c^2)^{-\frac{1}{2}}}{(1 - c^2)^2 c^{\frac{1}{2}}}.$$

This has a unique minimum at $c = 1$. The expression for $G_{4,1}$ is indeterminate; however, application of L'Hospital's rule yields the kernel

$$G_4(x) = \frac{1}{2}(3 - x^2)\phi(x).$$

It is straightforward to show that G_4 is the only fourth-order kernel of the form $Q_2\phi$, where Q_2 is a quadratic polynomial. It therefore seems reasonable that a $2r$th-order kernel would be of the form $G_{2r} \equiv Q_{2r-2}\phi$, where Q_{2r-2} is a polynomial of degree $2r - 2$. In

TABLE 1: Gaussian-based kernels of orders 2, 4, 6, 8, and 10.

$2r$	$G_{2r}(x)$
2	$\phi(x)$
4	$\frac{1}{2}(3 - x^2)\phi(x)$
6	$\frac{1}{8}(15 - 10x^2 + x^4)\phi(x)$
8	$\frac{1}{48}(105 - 105x^2 + 21x^4 - x^6)\phi(x)$
10	$\frac{1}{384}(945 - 1260x^2 + 378x^4 - 36x^6 + x^8)\phi(x)$

fact, there is only one such kernel, and this is provided by the following theorem proved in the Appendix.

THEOREM 2.1. *For $r \geq 1$ let Q_{2r-2} be the polynomial given by $Q_{2r-2}(x) = \sum_{i=0}^{r-1} c_{2i}x^{2i}$, where*

$$c_{2i} = \frac{(-1)^i 2^{i-2r+1}(2r)!}{r!(2i+1)!(r-i-1)!}, \qquad i = 0, \ldots, r-1. \tag{2.1}$$

Then Q_{2r-2} is the unique polynomial of degree less than or equal to $2r-2$ for which $G_{2r} \equiv Q_{2r-2}\phi$ is a $2r$th-order kernel.

Table 1 contains the first five Gaussian-based even-order kernels. Note that some of these kernels have made previous appearances in the literature. References include Nadaraya (1974), Singh (1981), Deheuvels (1977), and Silverman (1986, p. 69).

The proof of Theorem 2.1 is in the Appendix. A key observation in this proof is that G_{2r} can be represented as

$$G_{2r}(x) = \frac{(-1)^r \phi^{(2r-1)}(x)}{2^{r-1}(r-1)!x}.$$

Consequently $Q_{2r-2}(x) = \{2^{r-1}(r-1)!\}^{-1}H_{2r-1}(x)/x$, where H_j denotes the jth normalized Hermite polynomial defined by $H_j(x) = (-1)^j\phi(x)^{-1}\phi^{(j)}(x)$, $j \geq 0$. Kendall, Stuart, and Ord (1983, p. 221) list the first ten such polynomials. These polynomials also satisfy the following recurrence formula for $j \geq 2$:

$$H_j(x) - xH_{j-1}(x) + (j-1)H_{j-2}(x) = 0, \tag{2.2}$$

which can be used to establish that

$$G_{2r} = \sum_{s=0}^{r-1} \frac{(-1)^s}{2^s s!}\phi^{(2s)}. \tag{2.3}$$

This representation is very useful for implementation of the Fourier-transform methods of computation, since it follows from (2.3) that the Fourier transform of G_{2r} is simply

$$\tilde{G}_{2r}(t) = \tilde{\phi}(t)\sum_{s=0}^{r-1} \frac{t^{2s}}{2^s s!},$$

where $\tilde{\phi}$ is the Fourier transform of ϕ.

We can also use (2.3) to find closed-form convolution formulae. This is particularly useful for least-squares cross-validatory choice of h, as discussed in Härdle *et al.* (1990),

becaue one needs to minimize an expression involving $G_{2r}^{(\nu)} * G_{2r}^{(\nu)}$ (where $*$ denotes convolution). Appealing to the convolution result

$$(\phi^{(s)} * \phi^{(t)})(x) = 2^{-\frac{1}{2}(s+t+1)}\phi^{(s+t)}(x/2^{\frac{1}{2}}),$$

we arrive at

$$(G_{2r}^{(\nu)} * G_{2r}^{(\nu)})(x) = 2^{-\nu-\frac{1}{2}}\phi(x/2^{\frac{1}{2}})\sum_{s=0}^{r-1}\sum_{t=0}^{r-1}\frac{(-1)^{s+t}}{4^{s+t}s!t!}H_{2(s+t+\nu)}(x/2^{\frac{1}{2}}).$$

The extension of exact MISE calculations for esitmation of normal densities (à la Fryer 1976) to higher-order kernels can be accomplished by employing Gaussian-based kernels. Once again, the representation of G_{2r} given by (2.3) proves to be very useful.

3. EFFICIENCY CALCULATIONS

A price that one pays for using a Gaussian-based kernel to estimate $f^{(\nu)}$ is the loss in efficiency compared to the optimal polynomial-based kernels of Müller (1984) and Gasser *et al.* (1985). If we only require that our estimators be continuous functions, then the appropriate class of optimal kernels is the family of kernels $W_{\nu,k}$ with $\mu = 1$ in the notation of Müller (1984). These kernels correspond to the optimal kernels of Gasser *et al.* (1985) and include the Epanechnikov kernel when $\nu = 0$ and $k = 2r = 2$. For a general comparison of kernels of the same order we extend Silverman's definition of efficiency (Silverman 1986) for estimating f with a second-order kernel. The efficiency of G_{2r} with respect to $W_{\nu,k}(2r = k - \nu)$, denoted by $\text{eff}(2r, \nu)$, is defined below. The motivation for this definition is that, for large n, the MISE of the estimate of $f^{(\nu)}$ is the same using n observations and the kernel G_{2r} as it is using $\text{eff}(2r, \nu)n$ observations and the kernel $W_{\nu,\nu+2r}$. Let

$$C_\nu(K_{2r,\nu}) \equiv \left(\int K_{2r,\nu}^2\right)^{4r/(4r+2\nu+1)}\left(\int x^{2r+\nu}K_{2r,\nu}\right)^{(4\nu+2)/(4r+2\nu+1)}. \tag{3.1}$$

Then because $\text{MISE}(K_{2r,\nu})$ is proportional to $C_\nu(K_{2r,\nu})n^{-4r/(4r+2\nu+1)}$, we define

$$eff(2r, \nu) \equiv \left(\frac{C_\nu(W_{\nu,\nu+2r})}{C_\nu(G_{2r}^{(\nu)})}\right)^{(4r+2\nu+1)/(4r)}. \tag{3.2}$$

Note that in (3.2) we are taking $K_{2r,\nu}$ to be $W_{\nu,\nu+2r}$ in the numerator and $G_{2r}^{(\nu)}$ in the denominator when applying the definition at (3.1). Marron and Nolan (1988) demonstrated that for $\nu = 0$ (3.1) can be considered as a measure of effect of the kernel on MISE regardless of the choice of window width.

Our efficiency calculations are limited to the case of second-, fourth-, and sixth-order kernels. Using the formulae on p. 241 of Gasser *et al.* (1985) and the result

$$\int (\phi^{(\nu)})^2 = \frac{(2\nu)!}{\pi^{\frac{1}{2}}2^{2\nu+1}\nu!},$$

it can be shown that

$$eff(2, \nu) = \frac{(2\nu+3)!\pi^{\frac{1}{2}}}{(2\nu+5)^{(2\nu+3)/2}(\nu+1)!},$$

$$eff(4, \nu) = \frac{2(2\nu+5)!\pi^{\frac{1}{2}}}{(2\nu+9)^{(2\nu+9)/4}(2\nu+7)^{(2\nu+1)/4}(\nu+2)!},$$

$$eff(6, \nu) = \frac{4(2\nu+7)!\pi^{\frac{1}{2}}}{(4\nu^2+48\nu+151)(2\nu+13)^{(2\nu+7)/6}\{(2\nu+11)(2\nu+9)\}^{(2\nu+1)/6}(\nu+3)!}.$$

TABLE 2: Efficiencies of Gaussian-based kernels compared to optimal kernels ($2r = 2, 4, 6$; $\nu = 0, 1, 2$).

$2r$	ν	eff$(2r, \nu)$
2	0	0.9512
	1	0.8203
	2	0.6808
4	0	0.9320
	1	0.7841
	2	0.6414
6	0	0.9200
	1	0.7601
	2	0.6144

Values of these for $\nu = 0, 1, 2$ are listed in Table 2. Note that for $\nu = 0$, the important special case of density estimation, there is only a slight loss in efficiency incurred by a Gaussian-based kernel compared to the optimal polynomial kernel. These results represent a higher-order extension of the well-known result concerning the efficiency of the second-order Gaussian kernel compared to the Epanechnikov kernel. For ν as great as 2, however, the efficiencies are considerably lower. It appears that the advantages of Gaussian-based kernels may not be as attractive for larger values of ν, since they must be weighed against a sizable loss in efficiency. Nevertheless, one should keep in mind that in Table 2 the Gaussian-based kernels are being compared with the least-smooth polynomial kernels. Higher values of the smoothness index μ (see Müller 1984) would result in an improvement in the relative efficiencies of the infinitely smooth Gaussian-based kernels.

APPENDIX

We give here the proof of Theorem 2.1. Define

$$G_{2r}(x) = \frac{(-1)^r \phi^{(2r-1)}(x)}{2^{r-1}(r-1)!x}.$$

We first show that G_{2r} is a $2r$th-order kernel. To prove that G_{2r} integrates to unity we need to show that

$$\int x^{-1}\phi^{(2r-1)}(x)\,dx = (-1)^r 2^{r-1}(r-1)!, \qquad r \geq 1.$$

Writing the left-hand side as $-\int x^{-1}H_{2r-1}(x)\phi(x)\,dx$ and applying the recurrence formula at (2.2) yields this result. It is easily established by induction that

$$\int x^{2p}\phi^{(2q)}(x)\,dx = \begin{cases} 0, & p < q, \\ 2^{q-p}(2p)!/(p-1)!, & p \geq q. \end{cases} \tag{A.1}$$

Let $1 \leq j \leq r-1$, and observe that

$$\int x^{2j}G_{2r}(x) = \frac{(-1)^r}{2^{r-1}(r-1)!}\int x^{2j-1}\phi^{(2r-1)}(x)\,dx$$

$$= \frac{(-1)^{r+1}}{2^r j(r-1)!}\int x^{2j}\phi^{(2r)}(x)\,dx = 0$$

from integration by parts and (A.1). Clearly $\int x^{2j-1}G_{2r}(x)\,dx = 0$ for all $j \geq 1$. The first nonvanishing moment of G_{2r} is given by

$$\int x^{2r}G_{2r}(x)\,dx = \frac{(-1)^{r+1}}{2^r r!}\int x^{2r}\phi^{(2r)}(x)\,dx = \frac{(-1)^{r+1}(2r)!}{2^r r!}.$$

Therefore G_{2r} is a $2r$th-order kernel. Notice that $G_{2r}(x) = Q_{2r-2}(x)\phi(x)$, where Q_{2r-2} is the $(2r-2)$th-degree polynomial given by $Q_{2r-2}(x) = \{2^{r-1}(r-1)!\}^{-1}H_{2r-1}(x)/x$. The expression for $Q_{2r-2}(x)$ with coefficients given by (2.1) can be derived using the explicit formula for normalized Hermite polynomials. Abramowitz and Stegun (1972, p. 775) present a version of this formula.

The uniqueness of Q_{2r-2} follows from the invertibility of the matrix $\mathcal{E}(\mathbf{M}_r\mathbf{M}_r^{\mathsf{T}})$, where $\mathbf{M}_r = (1, Z^2, \ldots, Z^{2r-2})^{\mathsf{T}}$ and Z is a standard normal random variable. This matrix arises in the system of equations when one solves for the coefficients of Q_{2r-2}. Let $\mathbf{b}$ be an arbitrary nonzero r-vector, and observe that

$$\mathbf{b}^{\mathsf{T}}\mathcal{E}(\mathbf{M}_r\mathbf{M}_r^{\mathsf{T}})\mathbf{b} = \mathcal{E}\{(\mathbf{b}^{\mathsf{T}}\mathbf{M}_r)^2\} \geq \mathcal{V}ar(\mathbf{b}^{\mathsf{T}}\mathbf{M}_r).$$

Clearly $\mathcal{V}ar(\mathbf{b}^{\mathsf{T}}\mathbf{M}_r) > 0$ unless $\mathbf{b} = (k, 0, \ldots, 0)$ for some $k \neq 0$, in which case $\mathcal{E}\{(\mathbf{b}^{\mathsf{T}}\mathbf{M}_r)^2\} = k^2 > 0$. Therefore $\mathcal{E}(\mathbf{M}_r\mathbf{M}_r^{\mathsf{T}})$ is positive definite and hence invertible.

ACKNOWLEDGEMENT

The authors wish to express their gratitude to Professors D.B.H. Cline, R.J. Carroll, and J.D. Hart for helpful comments. Suggestions of three anonymous referees are also greatly appreciated. This research was begun while both authors were at the Australian National University. The research of the second author was partially supported by DARPA Contract No. F19628-88-K-0042.

REFERENCES

Abramowitz, M., and Stegun, I.A. (1972). *Handbook of Mathematical Functions*. U.S. government Printing Office, Washington.

Bartlett, M.S. (1963). Statistical estimation of density functions. *Sankhyā Ser. A*, 25, 245–254.

Deheuvels, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés. *Rev. Statist. Appl.*, 25, 5–42.

Fryer, M.J. (1976). Some errors associated with the nonparametric estimation of density functions. *J. Inst. Math. Appl.*, 18, 371–380.

Gasser, T., Müller, H.G., and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B*, 47, 238–252.

Hall, P., and Marron, J.S. (1988). Choice of kernel order in density estimation. *Ann. Statist.*, 16, 161–173.

Härdle, W., Marron, J.S., and Wand, M.P. (1990). Bandwidth choice for density derivatives. *J. Roy. Statist. Soc. Ser. B*, 52, 223–232.

Kendall, M., Stuart, A., and Ord, J. K. 91987). *Kendall's Advanced Theory of Statistics*. Oxford Univ. Press, New York.

Marron, J.S. (1988). Automatic smoothing parameter selection: A survey. *Empirical Econom.*, 13, 187–208.

Marron, J.S., and Nolan, D. (1988). Canonical kernels for density estimation. *Statist. Probab. Lett.*, 7, 195–199.

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

Nadaraya (1974). On the integral mean square error of some nonparametric estimates for the density function. *Theory Probab. Appl.*, 19, 133–141.

Park, B., and Marron, J.S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.*, 85, 66–72.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33, 1065–1076.

Schucany, W.R. (1989). On nonparametric regression with higher-order kernels. *J. Statist. Plann. Inferences*, 23, 145–151.

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

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

Singh, R.S. (1981). Speed of convergence in nonparametric estimation of a multivariate μ-density and its mixed partial derivatives. *J. Statist. Plann. Inference*, 5, 287–298.

Received 4 July 1989
Revised 1 December 1989
Accepted 8 March 1990

Department of Statistical Science
Southern Methodist University
Dallas, Texas 75275-0332