GAUSSIAN-BASED KERNELS FOR CURVE ESTIMATION AND WINDOW WIDTH SELECTION

Matthew P. Wand and William R. Schucany

Texas A&M Univ. and Southern Methodist Univ.

SMU/DS/TR/226

June 1989

Research sponsored by DARPA
Contract F19628-88-K-0042

GAUSSIAN-BASED KERNELS FOR CURVE ESTIMATION AND WINDOW WIDTH SELECTION

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.

AMS 1980 *subject classifications:* Primary 62G05; Secondary 62G20, 65D10.

ABSTRACT

We derive a class of higher-order kernels for curve estimation and window width selection 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. 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.

Running title: Gaussian-based kernels.

1. INTRODUCTION

Kernel estimators are a widely accepted means of estimating curves such as densities, regression functions and failure rates without parametric assumptions. Derivatives of these functions can also be estimated by straightforward extensions of kernel estimators. In this note we shall confine our attention to estimation of densities and their derivatives, however the essential theme applies to other settings.

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 derivative of the usual kernel density estimator and is given by

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

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, ν-times differentiable and satisfies $\int K_{2r} = 1$. 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 Singh (1977) and Schucany and Sommers (1977). By restricting attention to symmetric functions, the odd moments are zero and only even-order kernels are considered.

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 trade-off 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 (1989). A feature of these selection rules is that they also require the use of kernels. Some of these procedures, including those of Devroye (1989) and Hall, Sheather, Jones and Marron (1989), require kernels of orders higher than the one used in the actual curve estimator itself.

In the case where $r = 1$ 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).

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

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

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

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

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 integrated mean squared error of $f_{n,\nu}$ these authors derive optimal kernels for varying values of (ν, k) with the restriction that the kernel has compact support. 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 $[-1, 1]$. This can sometimes lead to problems, especially at the window width selection stage, if the kernel is not sufficiently smooth. Polynomial kernels also 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 least-squares cross validation to select a window width or the Fourier transform methods of computation. On the other hand, in regression problems concerns about boundary bias argue for compact support.

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 integrated mean squared error. This is demonstrated in Singh (1977) in the context of density derivative estimation. In the context of density estimation this principle is also 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 could be constructed from a second order kernel K_2 via the formula

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

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

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

for positive values of c. Calculations along the lines of those performed by Schucany (1989) yield $c = 1$ as the value of c that minimizes the asymptotic integrated mean squared error of $f_n^{(\nu)}$. The expression for $G_{4,1}$ is indeterminant, however application of L'Hospital's rule yields the kernel

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

This kernel is briefly discussed by Silverman (1986 p.69). It is straightforward to show that G_4 is the only 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 fact there is only one such kernel, and this is provided by

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 2.1 contains the first five Gaussian-based even-order kernels.

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 the jth normalized Hermite polynomial defined by $H_j(x) = (-1)^j \phi(x)^{-1}\phi^{(j)}(x)$, $j \geq 0$. 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, Marron and Wand (1989), because 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

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

3. EFFICIENCY CALCULATIONS

The price we pay for using a Gaussian-based kernel G_{2r} to estimate $f^{(\nu)}$ is the loss in efficiency compared to the optimal polynomial-based kernels of Müller (1984) and Gasser, Müller and Mammitzsch (1985). If we only require that our estimators are 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, the efficiency of G_{2r}, with respect to $W_{\nu,k}$ $(2r = k - \nu)$, is denoted by $\text{eff}(2r, \nu)$, where

$$\text{eff}(2r, \nu) \equiv \{C_\nu(W_{\nu,\nu+2r}) / C_\nu(G_{2r}^{(\nu)})\}^{(4r+2\nu+1)/(4r)} \tag{3.1}$$

and

$$C_\nu(K_{2r,\nu}) \equiv \left\{ \int K_{2r,\nu}^2 \right\}^{4r/(4r+2\nu+1)} \left\{ \int x^{2r+\nu} K_{2r,\nu} \right\}^{(4\nu+2)/(4r+2\nu+1)}. \tag{3.2}$$

Note that in (3.1) 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 (3.2). This is the generalization of Silverman's

(1986) definition of efficiency for estimating f with a second order kernel. The motivation for such a definition is that, for large n, the integrated mean squared error of the estimate of $f^{(\nu)}$ will be the same using n observations and the kernel G_{2r} as it will using $\text{eff}(2r,\nu)n$ observations and the kernel $W_{\nu,\nu+2r}$.

Our efficiency calculations are limited to the case of second-, fourth- and sixth-order kernels. Using the formulae on page 241 of Gasser *et al* (1985) it can be shown that

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

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

and

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

Values of these for $\nu = 0,1,2$ are listed in Table 3.1. 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 as attractive for larger values of ν since they must be reconciled against a sizeable loss in efficiency. Nevertheless, one should keep in mind that in Table 3.1 the Gaussian-based kernels are being compared to 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

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

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

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

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

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 non-vanishing 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 second-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 $E(\mathbf{M}_r\mathbf{M}_r')$ where $\mathbf{M}_r = (1, Z^2, \ldots, Z^{2r-2})'$ 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 non-zero r-vector and observe that

$$\mathbf{b}'E(\mathbf{M}_r\mathbf{M}_r')\mathbf{b} = E\{(\mathbf{b}'\mathbf{M}_r)^2\} \geq \mathrm{Var}(\mathbf{b}'\mathbf{M}_r).$$

Clearly $\mathrm{Var}(\mathbf{b}'\mathbf{M}_r) > 0$ unless $\mathbf{b} = (k, 0, \ldots, 0)$ for some $k \neq 0$ in which case $E\{(\mathbf{b}'\mathbf{M}_r)^2\} = k^2 > 0$. Therefore $E(\mathbf{M}_r\mathbf{M}_r')$ is positive definite and hence invertible. ∎

ACKNOWLEDGEMENT

The authors wish to express their gratitude to Professor D. B. H. Cline and Professor R. J. Carroll for helpful comments. 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.* Washington: U.S. Government Printing Office.

Devroye, L. (1989). An L1 asymptotically optimal kernel estimator, unpublished manuscript.

Gasser, T., Müller, H.G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society, Series B*, **47**, 238–252.

Hall, P., Sheather, S.J., Jones, M.C. and Marron, J.S. (1989). Optimal data-based bandwidth selection in kernel density estimation, unpublished manuscript.

Härdle, W., Marron, J.S. and Wand, M.P. (1989). Bandwidth choice for density derivatives, *Journal of the Royal Statistical Society, Series B*, to appear.

Marron, J.S. (1988). Automatic smoothing parameter selection: a survey. *Empirical Economics*, to appear.

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

Park, B. and Marron, J.S. (1989). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, to appear.

Schucany, W.R. (1989). On nonparametric regression with higher-order kernels. *Journal of Statistical Planning and Inference*, to appear.

Schucany, W.R. and Sommers, J.P. (1977). Improvement of kernel-type density estimators. *Journal of the American Statistical Association*, **72**, 420–423.

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

Chapman and Hall.

Singh, R.S. (1977). Improvement of some known nonparametric uniformly consistent estimates of derivatives of a density. *Annals of Statistics*, **5**, 394–399.

Stuart, A. and Ord, K.J. (1983). *Kendall's Advanced Theory of Statistics.* London: Macmillian.

Table 2.1

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	$\frac{1}{384}(945-1260x^2+378x^4-36x^6+x^8)\phi(x)$

Table 3.1

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

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