Nonparametric Statistics, Vol. 0, pp. 000–000
Reprints available directly from the publisher
Photocopying permitted by license only

ERROR ANALYSIS FOR GENERAL MULTIVARIATE KERNEL ESTIMATORS

M. P. WAND

Department of Statistics, Rice University, Houston, TX 77251, U.S.A.

(Received 9 September 1991, Revised 3 February 1992, Accepted 13 April 1992)

Kernel estimators for d dimensional data are usually parameterized by either a single smoothing parameter, or d smoothing parameters corresponding to each of the coordinate directions. A generalization of each of these parameterizations is to use a $d \times d$ matrix which allows smoothing in arbitrary directions. We demonstrate that, at this level of generality, the usual error approximations can be done quite simply using matrix algebra. Particular attention is paid to the special case of kernel estimation of multivariate normal mixture densities where it is shown that the minimization of both asymptotic and exact mean integrated squared error can be set up in a matrix algebraic formulation. This provides a flexible family of multivariate smoothing problems for which error analyses can be performed in a computationally simple manner.

KEYWORDS: Density estimation, matrix differential calculus, mean integrated squared error, nonparametric regression, vec and vech operators, window width matrix.

1. INTRODUCTION

Multivariate kernel estimators provide an intuitively and theoretically attractive way of recovering features in the higher-dimensional surface of interest without the imposition of parametric assumptions. While it is clear that the performance of kernel estimators (and, in fact, all nonparametric function estimators) eventually becomes unacceptably poor for increasingly higher dimensional data they appear to be a viable tool for recovering important features for up to about five dimensions. This has been demonstrated in work by Scott (1983), Silverman (1986), Müller (1988) and Härdle (1990) and is studied by Scott & Wand (1991).

A very important component of the kernel estimator is its smoothing parameterization and, once this parameterization has been settled upon, the choice of the smoothing parameters themselves. In the univariate case the kernel estimator has a single smoothing parameter h, often called the *window width* or *bandwidth* and there have been several recent proposals for high-performance selection of h from the data (e.g. Park & Marron, 1990, Sheather & Jones, 1991, Gasser, Kneip & Kohler, 1991). For the d-dimensional kernel estimator there are several levels of options for the smoothing parameterization. The simplest of these involves having a single smoothing parameter h as in the one-dimensional case. A more sophisticated parameterization requires that one have a distinct window width h_i for the ith coordinate direction. However, each of these approaches may be viewed as subparameterizations of what we will call the *general kernel estimator* which is a parameterized by a $d \times d$ positive definite

symmetric "window width" matrix H. Having a full matrix allows for smoothing in arbitrary directions and, depending on the shape of the underlying surface, it may be beneficial to have this flexibility. The general kernel estimator was first studied by Deheuvels (1977) in the density estimation setting. More recently, Terrell & Scott (1991) and Staniswalis, Messer & Finston (1991) have considered the same generality in the density estimation and regression contexts respectively.

The main purpose of this article is to provide methodology for error analyses of general kernel estimators. An appealing feature of working with the full window width matrix is that virtually all calculations can be done using matrix algebraic techniques which turn out to be simpler than error analysis for the single smoothing parameter case. We also discuss the important problem of minimising global error criteria. The directness of our approach means that such minimisation is unconstrained and can be done using straightforward application of Newton's method. For illustrative purposes we derive asymptotic error approximations for the general kernel density estimator. Adaptations of this derivation to other kernel estimators is shown to be straightforward. For the case of estimating a normal mixture density we are able to derive expressions for the mean integrated squared error (MISE) and its asymptotic approximation which involves only matrix algebra. The numerical minimisation of these error criteria over H can be set-up in matrix algebraic terms. This provides a flexible class of multivariate smoothing problems where error analyses can be performed computationally simply. Wand & Jones (1991) recently used this family to compare different levels of smoothing parameterizations on the performance of the bivariate kernel density estimator and the results presented here allow the possibility of similar comparisons in higher dimensional settings. Such comparisons have important implications for data-driven smoothing parameter selection since, for increasingly higher dimensions, the number of independent smoothing parameters in the full window width matrix grows quadratically and only linearly for the diagonal window width matrix.

Another important problem closely related to data-driven smoothing parameter selection is the estimation of integrals of products of density derivatives. Our methodology is easily adaptable to this setting and we are able to present multivariate asymptotic approximations for kernel estimators of these functionals.

The techniques used to find asymptotic approximations for kernel density estimators are adaptable to other settings as well. As examples, we illustrate this in the regression and hazard rate contexts.

Section 2 deals with the general kernel density estimator. The minimisation of the asymptotic MISE is discussed in Section 3 and of exact MISE in Section 4. Section 5 gives an example of these calculations. In Section 6 we treat estimation of integrated products of density derivatives and in Section 7 we briefly describe the adaptation of our techniques to other settings.

2. MULTIVARIATE DENSITY ESTIMATION

Consider the problem of estimating a d-variate probability density function f based on a sample $X_1, \ldots, X_n$ of independent $\mathbb{R}^d$-valued independent random

variables having common density f. The general multivariate kernel density estimator of f is

$$\hat{f}(x; H) = n^{-1} \sum_{i=1}^{n} K_H(x - X_i) \tag{2.1}$$

where x is a vector in $\mathbb{R}^d$, H is a symmetric positive definite $d \times d$ matrix which we call the *window width matrix*, and $K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$ and K is a d-variate function satisfying $\int K(x)\,dx = 1$. Here, and throughout this article, $\int$ is shorthand for $\int \cdots \int_{\mathbb{R}^d}$ and dx is shorthand for $dx_1 \cdots dx_d$.

The kernel function is often taken to be a spherically symmetric probability density function such as the standard d-variate normal density

$$K(x) = (2\pi)^{-d/2} \exp(-\tfrac{1}{2}x^T x)$$

in which case $K_H(x - X_i)$ is the $N(X_i, H)$ density in the vector x. Other possible spherically symmetric kernels are those of the form

$$K(x) = \begin{cases} c_{p,d}(1 - x^T x)^p, & x^T x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

where $c_{p,d}^{-1} = \int (1 - x^T x)^p\,dx$ and $p > 0$.

The choice of H which is a very important one. In the univariate setting, in which case H is a scalar, there now exist several reliable methods for choosing H from the data. The general d-variate kernel estimator has $\frac{1}{2}d(d+1)$ independent parameters and for even moderate dimensions this number can be considerable, especially if they need to be chosen by the data. One way around this is to restrict H to certain subclasses of the family of all symmetric positive definite $d \times d$ matrices. The simplest such restriction is $H \in \{h^2 I : h > 0\}$, which means that we have a single smoothing parameter h and the kernel estimator is of the form

$$\hat{f}(x; h) = n^{-1} h^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\}. \tag{2.2}$$

A less stringent restriction is $H \in \{\mathrm{diag}(h_1^2, \ldots, h_d^2) : h_1, \ldots, h_d > 0\}$ which gives us

$$\hat{f}(x; h_1, \ldots, h_d) = n^{-1} h_1^{-1} \cdots h_d^{-1} \sum_{i=1}^{n} K\{(x_1 - X_{i1})/h_1, \ldots, (x_d - X_{id})/h_d\}. \tag{2.3}$$

In this case h_i can be thought of as being the smoothing parameter associated with the ith coordinate direction.

The general kernel density estimator (2.1) is discussed extensively by Deheuvels (1977). However, virtually all other literature on the topic deals with the simpler forms (2.2) and (2.3). While these simple forms may be appropriate for some densities there are others where one can do much better using a full window width matrix (Wand & Jones, 1991).

We will first present asymptotic theory for (2.1) for general smooth densities. This will be followed by a treatment of the important special case of normal mixture densities.

2.1. General Densities

For asymptotic approximations to the bias and variance of $\hat{f}(\cdot; H)$ we will make the following assumptions on f, H and K. The first of these involves the *Hessian matrix* of f which is the $d \times d$ matrix having (i, j) entry equal to $\partial^2/(\partial x_i\, \partial x_j) f(x)$ and we denote by $G_f(x)$.

(i) Each entry of $G_f(x)$ is bounded, continuous and square integrable for all $x \in \mathbb{R}^d$.

(ii) $H = H_n$ is a sequence of window width matrices for which all entries approach zero as $n \to \infty$. Also,

$$\lim_{n\to\infty} n^{-1} |H|^{-1/2} = 0.$$

(iii) K is a spherically symmetric d-variate kernel. That is, $K(x) = c_\kappa \kappa\{(x^T x)\}$ where κ is a symmetric univariate density and $c_\kappa^{-1} = \int \kappa\{(x^T x)\}$. We also that $\mu_2(K) = \int x_i^2 K(x)\, dx < \infty$.

We also define $D_f(x)$ to be the $d \times 1$ vector for which the ith entry is $(\partial/\partial x_i) f(x)$ and J to be the $d \times d$ matrix having each entry equal to 1.

By the multivariate version of Taylor's theorem

$$\begin{aligned} E\hat{f}(x; H) &= \int K(z) f(x - H^{1/2} z)\, dz \\ &= \int K(z)\{f(x) - (H^{1/2} z)^T D_f(x) + \tfrac{1}{2}(H^{1/2} z)^T G_f(x)(H^{1/2} z)\}\, dz + o\{\mathrm{tr}(JH)\} \\ &= f(x) - \int z^T H^{1/2} D_f(x) K(z)\, dz \\ &\quad + \frac{1}{2} \int z^T H^{1/2} G_f(x) H^{1/2} z K(z)\, dz + o\{\mathrm{tr}(JH)\}. \end{aligned} \tag{2.4}$$

Let Z be a random vector having density K. by symmetry of K, $E(Z) = 0$ and by spherical symmetry, $E(ZZ^T) = \mu_2(K) I$. Then the second term of (2.4) is

$$-E\{Z^T H^{1/2} D_f(x)\} = -E(Z)^T H^{1/2} D_f(x) = 0.$$

The leading bias term is therefore equal to $\frac{1}{2} E\{Z^T H^{1/2} G_f(x) H^{1/2} Z\}$ and, noting that the argument is scalar, this expectation can be written

$$E[\mathrm{tr}\{Z^T H^{1/2} G_f(x) H^{1/2} Z\}] = \mathrm{tr}\{H^{1/2} G_f(x) H^{1/2} E(ZZ^T)\} = \mu_2(K)\, \mathrm{tr}\{H G_f(x)\}.$$

Therefore, the leading bias term is

$$E\hat{f}(x; H) - f(x) \sim \tfrac{1}{2} \mu_2(K)\, \mathrm{tr}\{H G_f(x)\}. \tag{2.5}$$

This simple derivation of the bias uses some ideas in Staniswalis *et al.* (1991).

The variance of $\hat{f}(x; H)$ if given by

$$\begin{aligned} \mathrm{Var}\, \hat{f}(x; H) &= n^{-1}\left[|H|^{-1/2} \int K(z)^2 f(x - H^{1/2} z)\, dz - \left\{ \int K(z) f(x - H^{1/2} z) \right\}^2 \right] \\ &= n^{-1} |H|^{-1/2} R(K) f(x) + o(n^{-1} |H|^{-1/2}). \end{aligned} \tag{2.6}$$

where $R(K) = \int K(z)^2\,dz$. The mean integrated squared error (MISE) of $\hat{f}(\cdot; H)$ is given by

$$\text{MISE}\{\hat{f}(\cdot; H)\} = E \int \{\hat{f}(x; H) - f(x)\}^2\,dx.$$

The asymptotic approximation to $\text{MISE}\{\hat{f}(x; H)\}$, denoted by $\text{AMISE}\{\hat{f}(x; H)\}$, can be obtained by combining (2.5) and (2.6) to give

$$\text{AMISE}\{\hat{f}(\cdot; H)\} = n^{-1}\,|H|^{-1/2}\,R(K) + \tfrac{1}{4}\mu_2(K)^2 \int \text{tr}^2\{HG_f(x)\}\,dx. \tag{2.7}$$

For the important case of K being the normal kernel we have $R(K) = (4\pi)^{-d/2}$ and $\mu_2(K) = 1$.

2.2. *Normal Mixture Densities*

We now demonstrate that further simplification of the above AMISE expression is possible for the special case of normal mixture densities. The gist of this simplification is the eradication of the integral sign in the integrated squared bias resulting in an AMISE expression that involves only matrix algebra.

Let $\phi_\Sigma(x) = (2\pi)^{-d/2}\,|\Sigma|^{-1/2} \exp(-\frac{1}{2}x^T\Sigma^{-1}x)$ denote the $N(0, \Sigma)$ density. Then multivariate normal mixture densities are those of the form

$$f(x) = \sum_{l=1}^{k} w_l \phi_{\Sigma_l}(x - \mu_l) \tag{2.8}$$

where $w = (w_1, \ldots, w_k)^T$ is a vector of positive numbers summing to unity, and for each $1 \le l \le k$, μ_l is a $d \times 1$ vector and Σ_l is a $d \times d$ symmetric positive definite matrix. One appealing feature of this family is that any density can be approximated arbitrarily well by one of its members.

For each pair (l, l') let $A_{ll'} = (\Sigma_l + \Sigma_{l'})^{-1}$,

$$B_{ll'} = A_{ll'}\{I - 2(\mu_l - \mu_{l'})(\mu_l - \mu_{l'})^T A_{ll'}\}$$

and

$$C_{ll'} = A_{ll'}\{I - (\mu_l - \mu_{l'})(\mu_l - \mu_{l'})^T A_{ll'}\}.$$

Then we have

THEOREM 1. If f is the d-variate normal mixture density given by (2.8) then

$$\text{AMISE}\{\hat{f}(\cdot; H)\} = n^{-1}R(K)\,|H|^{-1/2} + \tfrac{1}{4}\mu_2(K)^2 \sum_{l=1}^{k}\sum_{l'=1}^{k} w_l w_{l'} \phi_{\Sigma_l + \Sigma_{l'}}(\mu_l - \mu_{l'})$$
$$\times \{2\,\text{tr}(HA_{ll'}HB_{ll'}) + \text{tr}^2(HC_{ll'})\}.$$

The proof is given in the appendix.

In the special case of the multivariate normal density considerable simplification is possible. We have for estimation of the $N(\mu, \Sigma)$ density and K the $N(0, I)$ density that

$$(4\pi)^{d/2}\,\text{AMISE}\{\hat{f}(\cdot; H)\} = n^{-1}\,|H|^{-1/2} + \tfrac{1}{16}\,|\Sigma|^{-1/2}\,[2\,\text{tr}(H\Sigma^{-1}H\Sigma^{-1}) + \text{tr}^2(H\Sigma^{-1})].$$

Using matrix calculus results (which are given in the Section 3) we can show that the AMISE-optimal window width matrix for this density is

$$H_{\text{AMISE}} = \{4/(d+2)\}^{2/(d+4)} \Sigma n^{-2/(d+4)}$$

and

$$\inf_H \text{AMISE}\{\hat{f}(\cdot; H)\} = (4\pi)^{-d/2}\{(d+4)/4\}\{(d+2)/4\}^{d/(d+4)} |\Sigma|^{-1/2} n^{-4/(d+4)}.$$

The first of these results says that, for multivariate normal data, use of the optimal H means that the kernel has same shape as the density being estimated.

3. MINIMIZATION OF AMISE

The multivariate normal example given above is rare in the sense that the AMISE-optimal window width matrix can be found in closed form. In general, minimization of AMISE can only be performed numerically. The purpose of this section is to provide the details for numerical minimization of $\text{AMISE}\{\hat{f}(\cdot; H)\}$ using Newton's method.

3.1. General Densities

We start with some matrix notation which allows easier handling of matrix derivatives required for the numerical minimization of $\text{AMISE}\{\hat{f}(\cdot; H)\}$. For a $d \times d$ matrix A the *vector* of A, denoted by $\text{vec}\, A$ is the $d^2 \times 1$ vector obtained by stacking the columns of A underneath each other in order from left to right. The *vector-half* of A, denoted by $\text{vech}\, A$ is the $\frac{1}{2}d(d+1) \times 1$ vector obtained from $\text{vec}\, A$ by eliminating each of the above-diagonal entries of A (Henderson & Searle, 1979). Thus, if A is symmetric then $\text{vech}\, A$ contains each of the distinct entries of A. Since, for symmetric A, $\text{vec}\, A$ contains the entries of $\text{vech}\, A$ with some repetitions there is a unique $d^2 \times \frac{1}{2}d(d+1)$ matrix D_d of zeroes and ones such that

$$D_d \,\text{vech}\, A = \text{vec}\, A \quad (A = A^T)$$

(see e.g. Magnus & Neudecker, 1988 p. 49) and is called the *duplication matrix* of order d. For example, the duplication matrix of order 2 is given by

$$D_2^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Since all window width matrices H are symmetric it suffices to obtain the vector $\text{vech}(H_{\text{AMISE}})$. This vector satisfies

$$\frac{\partial \,\text{AMISE}\{\hat{f}(\cdot; H)\}}{\partial \,\text{vech}\, H} = 0$$

and, for a starting value sufficiently close to the minimum, the solution of this equation is the limit of the sequence $H_1, H_2, \ldots$ where

$$\text{vech}\, H_{i+1} = \text{vech}\, H_i - \left[\frac{\partial^2 \,\text{AMISE}\{\hat{f}(\cdot; H)\}}{(\partial \,\text{vech}\, H)(\partial \,\text{vech}\, H)^T}\right]^{-1}_{H=H_i} \left[\frac{\partial \,\text{AMISE}\{\hat{f}(\cdot; H)\}}{\partial \,\text{vech}\, H}\right]_{H=H_i}. \tag{3.1}$$

We are now faced with the task of obtaining expressions for the derivative vector and Hessian matrix of $\mathrm{AMISE}\{\hat{f}(\cdot; H)\}$ with respect to $\mathrm{vech}\, H$. From the identity $\mathrm{tr}(A^T B) = (\mathrm{vec}\, A)^T(\mathrm{vec}\, B)$ it follows from (2.7) that

$$\mathrm{AMISE}\{\hat{f}(\cdot; H)\} = n^{-1}R(K)\,|H|^{-1/2} + \tfrac{1}{4}\mu_2(K)^2 \int (\mathrm{vec}\, H)^T \{\mathrm{vec}\, G_f(x)\}\{\mathrm{vec}\, G_f(x)\}^T (\mathrm{vec}\, H)\, dx.$$

Using (3.1) we can write this as

$$\mathrm{AMISE}\{\hat{f}(\cdot; H)\} = n^{-1}R(K)\,|H|^{-1/2} + \tfrac{1}{4}\mu_2(K)^2 (\mathrm{vech}\, H)^T \mathscr{G}_f (\mathrm{vech}\, H),$$

where $\mathscr{G}_f$ is the $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$ vector having (i, j) entry equal to

$$\int [D_d^T \mathrm{vec}(G_f(x))\{\mathrm{vec}(G_f(x))\}^T D_d]_{ij}\, dx.$$

The second term is now a quadratic form in $\mathrm{vech}\, H$ and can be diffentiated using standard matrix calculus results (see below). However, differentiation of the first term is non-trivial and will require the following technical result.

THEOREM 2. Let X be an invertible $d \times d$ matrix. Then

(a) $$\frac{\partial\, |X|^{-1/2}}{\partial\, \mathrm{vech}\, X} = -\tfrac{1}{2}\,|X|^{-1/2} D_d^T \mathrm{vec}(X^{-1}),$$

and

(b) $$\frac{\partial\, |X|^{-1/2}}{(\partial\, \mathrm{vech}\, X)(\partial\, \mathrm{vech}\, X)^T} = \tfrac{1}{4}\,|X|^{-1/2} D_d^T (X^{-1} \otimes I_d)^T \times \{(\mathrm{vec}\, I_d)(\mathrm{vec}\, I_d)^T + 2I_{d^2}\}(I_d \otimes X^{-1}) D_d$$

where $\otimes$ denotes Kronecker product and I_d is the $d \times d$ identity matrix.

The proof of this result is outlined in the appendix.

We can now apply Theorem 2 and standard results for quadratic forms to give

$$\frac{\partial\, \mathrm{AMISE}\{\hat{f}(\cdot; H)\}}{\partial\, \mathrm{vech}\, H} = -\tfrac{1}{2}n^{-1}R(K)\,|H|^{-1/2} D_d^T \mathrm{vec}(H^{-1}) + \tfrac{1}{2}\mu_2(K)^2 \mathscr{G}_f (\mathrm{vech}\, H) \tag{3.2}$$

and

$$\frac{\partial^2\, \mathrm{AMISE}\{\hat{f}(\cdot; H)\}}{(\partial\, \mathrm{vech}\, H)(\partial\, \mathrm{vech}\, H)^T} = \tfrac{1}{4}n^{-1}R(K)\,|H|^{-1/2} D_d^T (H^{-1} \otimes I_d)^T \times \{(\mathrm{vec}\, I_d)(\mathrm{vec}\, I_d)^T + 2I_{d^2}\}(I_d \otimes H^{-1}) D_d + \tfrac{1}{2}\mu_2(K)^2 \mathscr{G}_f. \tag{3.3}$$

3.2. Normal Mixture Densities

While (3.1), (3.2) and (3.3) provide full details of the numerical minimisation of $\mathrm{AMISE}\{\hat{f}(\cdot; H)\}$ via Newton's method, the matrix $\mathscr{G}_f$ still requires the evaluation

of several multivariate integrals which, in many cases, will not have explicit representations. However, in Section 2.2 we showed that for normal mixture densities it is possible to eliminate these integrals and obtain an explicit expression for AMISE$\{\hat{f}(\cdot; H)\}$. We now demonstrate that explicit derivative expressions for the AMISE$\{\hat{f}(\cdot; H)\}$ can be given for normal mixture densities.

The asymptotic integrated variance is independent of the density and can be handled using Result 1 above. However, the asymptotic integrated squared bias can be written

$$\sum_{l=1}^{k}\sum_{l'=1}^{k} w_l w_{l'} \phi_{\Sigma_l+\Sigma_{l'}}(\mu_l-\mu_{l'})(\text{vech } H)^T D_d^T \times \{2(A_{ll'}\otimes B_{ll'}) + (\text{vec } C_{ll'})(\text{vec } C_{ll'})^T\} D_d(\text{vech } H)$$

(see e.g. Magnus & Neudecker 1988, p. 31). Now that this is a quadratic form in vech H it is a simple matter to obtain the required derivatives with respect to this vector.

4. MINIMIZATION OF MISE

A further important property of normal mixture densities is that they admit a closed form expression for the finite sample MISE of $\hat{f}(\cdot; H)$ when the kernel is normal. For f given by (2.8) and $K = \phi_{I_d}$ it can be shown that

$$\text{MISE}\{\hat{f}(\cdot; H)\} = n^{-1}(4\pi)^{-d/2}|H|^{-1/2} + w^T\{(1-n^{-1})\Omega_2 - 2\Omega_1 + \Omega_0\}$$

where Ω_a is the $k \times k$ matrix having (l, l') entry equal to $\phi_{aH+\Sigma_l+\Sigma_{l'}}(\mu_l-\mu_{l'})$ (Wand & Jones, 1991, Theorem 1).

Numerical minimization of MISE$\{\hat{f}(\cdot; H)\}$ via Newton's method requires expressions for first and second order derivatives of expressions of the form $\phi_{aH+M}(x)$, where M is a symmetric $d \times d$ matrix, since the second term of (4.1) is a linear combination of such functions of H. The first term can be handled using Theorem 2 as in the AMISE case. Methods similar to those used to derive Theorem 2 (see the Appendix and Magnus & Neudecker, 1988) can be used to show that

$$\frac{\partial \phi_{aH+m}(x)}{\partial \text{ vech } H} = \tfrac{1}{2}a\phi_{aH+M}(x)D_d^T \text{vec}[(aH+M)^{-1}\{xx^T(aH+M)^{-1} - I\}]$$

and

$$\frac{\partial^2 \phi_{aH+M}(x)}{(\partial \text{ vech } H)(\partial \text{ vech } H)^T} = \tfrac{1}{4}a^2\phi_{aH+M}(x)$$

$$D_d^T\{\text{vec}[\{xx^T(aH+M)^{-1} - I\}(aH+M)^{-1}](\text{vec}[\{xx^T(aH+M)^{-1} - I\} \times (aH+M)^{-1}])^T + 2\{I\otimes(aH+M)^{-2}\} - 4\{(aH+M)^{-1}\otimes(aH+M)^{-1}\}xx^T(aH+M)^{-1}\}D_d$$

for all H such that $(aH+M)^{-1}$ exists.

5. EXAMPLE

We programmed the formulae for AMISE$\{\hat{f}(\cdot; H)\}$ for normal mixture densities, as well as the Newton's method algorithm for minimizing this quantity, using the matrix-oriented programming language GAUSS$_{TM}$. In the examples tried, rapid convergence to the minimum was always achieved, provided that the starting value was not too far from the minimum.

In this section we present an example which illustrates how these formulae can be used to gain a better understanding of multivariate kernel smoothing. Consider the trivariate normal mixture distribution

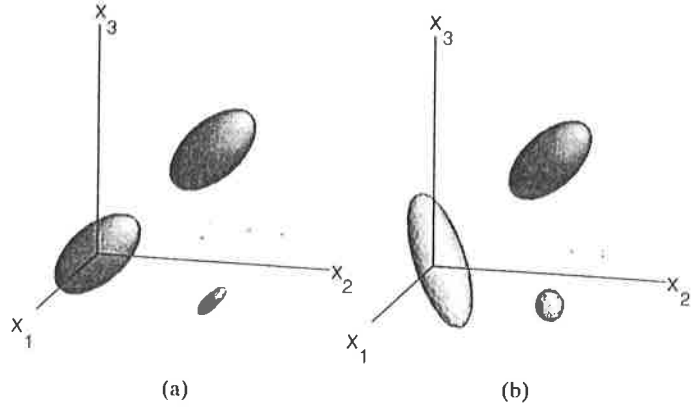
$$\tfrac{1}{2}N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.7 & 0.7\\0.7 & 1 & 0.7\\0.7 & 0.7 & 1\end{bmatrix}\right) + \tfrac{1}{2}N\left(\begin{bmatrix}2\\2\\2\end{bmatrix}, \begin{bmatrix}1 & 0.7 & 0.7\\0.7 & 1 & 0.7\\0.7 & 0.7 & 1\end{bmatrix}\right). \tag{5.1}$$

For the Gaussian kernel and a sample size of $n = 1000$ the optimal window width matrix for this example was found to be

$$(\text{vech } H_{\text{AMISE}})^T = [0.22\ \ 0.18\ \ 0.18\ \ 0.22\ \ 0.18\ \ 0.22]$$

which, because of the high correlation, is far from diagonal. The main ellipsoids in Figure 1b are typical contours of probability density for (5.1) while the small ellipsoid is a typical contour for the Gaussian kernel scaled by H_{AMISE} and has almost the same shape as the normal mixture components. To compare the performance of a kernel estimator based on this matrix to the optimal diagonal one we will use the Asymptotic Relative Efficiency (ARE) given by

$$\text{ARE}(\mathcal{F}:\mathcal{D}) = \left[\inf_{H\in\mathcal{F}} \text{AMISE}\{\hat{f}(\cdot; H)\} \Big/ \inf_{H\in\mathcal{D}} \text{AMISE}\{\hat{f}(\cdot; H)\}\right]^{7/4}$$



(a) (b)

Figure 1. Typical ellipsoidal contours of each of the trivariate normal components of the densities in the example. Figure 1a is for density (5.1) and Figure 1b is for the density obtained by rotating the lower component about the line $x_1 = -x_2$, $x_3 = 0$ by $\theta = \pi/2$. The smaller ellipsoids are typical contours of Gaussian kernels rescaled by the corresponding H_{AMISE}.

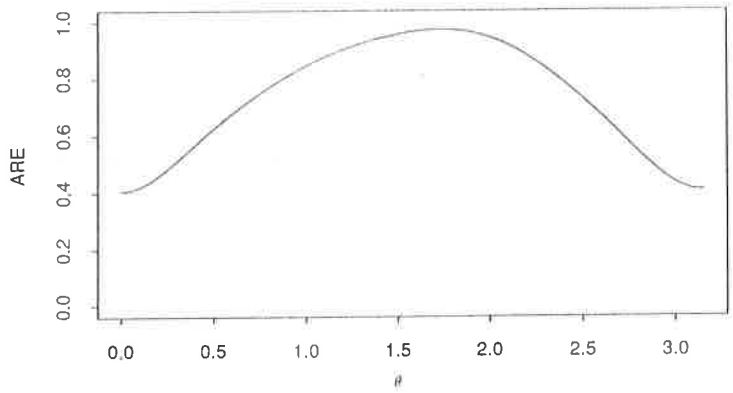


Figure 2. Values of ARE($\mathscr{F}:\mathscr{D}$) versus the rotation angle θ, $0 \leq \theta \leq \pi$, for the example in Section 5.

where $\mathscr{F}$ is the class of all 3×3 window width matrices and $\mathscr{D}$ is the subclass of symmetric ones. For density (5.1) we obtained ARE($\mathscr{F}:\mathscr{D}$) $= 0.40$ which says that, for large n, only about 40% of the data is needed to achieve the same performance with the optimal full window width matrix than that using the optimal diagonal matrix (this interpretation of the ARE is the reason for the 7/4 power since the minimum AMISE converges at the rate $n^{-4/7}$ in each case).

We next investigated how the ARE changes as we rotate the probability mass in the first component of (5.1) about the line $x_2 = -x_1$, $x_3 = 0$. This line is perpendicular to the major axis of the ellipsoidal contours of both of the components of (5.1). Figure 2 shows this ARE as a function of θ for $0 \leq \theta \leq \pi$. Notice that when $\theta = \pi/2$ the major axes of the ellipsoidal contours of each component are mutually perpendicular so the correlations have a cancelling effect and a diagonal window width matrix is close to being optimal since ARE($\mathscr{F}:\mathscr{D}$) $= 0.96$ in this case. Figure 1b shows the corresponding ellipsoids for this case and it is seen that the optimally scaled kernel is close to being spherical.

While this is a brief example it demonstrates the possibility of further error analyses for multivariate kernel estimators. Finite sample analyses using the results from Section 4 for exact MISE can also be performed.

6. INTEGRATED PRODUCTS OF DENSITY DERIVATIVES

An important companion problem to density estimation is the estimation of integrals of products of density derivatives. This is because such integrals are entries of the matrix $\mathscr{G}_f$ and since this is the only unknown in the AMISE expression, automatic selection of H via "plug-in" ideas would require estimation of $\mathscr{G}_f$.

For a d-variate function g and vector $m = (m_1, \ldots, m_d)$ of non-negative entries we will use the notation

$$g^{(m)}(x) = \frac{\partial^{|m|}}{\partial x_1^{m_1} \cdots \partial x_d^{m_d}} g(x).$$

to denote partial derivatives of g. Here, and throughout this section, $|m| = \sum_{i=1}^{d} |m_i|$. Using integration by parts it can be shown that for all m and m' such that $f^{(m+m')}$ exists integrated products of density derivatives satisfy

$$\int f^{(m)}(x) f^{(m')}(x)\,dx = \begin{cases} (-1)^{|m|} \int f^{(m+m')}(x) f(x)\,dx & |m-m'| \text{ even} \\ 0 & |m-m'| \text{ odd.} \end{cases}$$

Consequently, it suffices to study the estimation of integrals of the form

$$\psi_m = \int f^{(m)}(x) f(x)\,dx.$$

Notice that integrated squared density derivatives are related to ψ_m through

$$\int f^{(m)}(x)^2\,dx = (-1)^{|m|}\psi_{2m}.$$

Kernel estimators of this functional in the univariate setting have been analysed by Hall & Marron (1987) and Jones & Sheather (1991).

Writing $\psi_m = Ef^{(m)}(X)$ suggests the kernel estimator

$$\hat{\psi}_m(H) = n^{-1} \sum_{i=1}^{n} \hat{f}^{(m)}(X_i; H) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_H^{(m)}(X_i - X_j).$$

The bias and variance calculations for $\hat{\psi}_m(H)$ can be performed by combining the ideas presented in the appendix of Hall & Marron (1987) and Section 2 of the present paper. For the bias we obtain

$$E\hat{\psi}_m(H) - \psi_m \sim n^{-1} K_H^{(m)}(0) + \tfrac{1}{2}(-1)^{|m|}(1 - n^{-1})\mu_2(K) \int \operatorname{tr}\{HG_f(x)\} f^{(m)}(x)\,dx$$

and

$$\operatorname{Var}\{\hat{\psi}_m(H)\} \sim 2n^{-2}\psi_0 \int K_H^{(m)}(x)^2\,dx + 4n^{-1}\left\{\int f^{(m)}(x)^2 f(x)\,dx - \psi_m^2\right\}.$$

In the important special case where $H = \operatorname{diag}(h_1^2, \ldots, h_d^2)$ these expressions reduce to

$$E\hat{\psi}_m(h_1, \ldots, h_d) - \psi_m \sim n^{-1}\left(\prod_{i=1}^{d} h_i^{-m_i-1}\right) K^{(m)}(0) + \tfrac{1}{2}(-1)^{|m|}(1 - n^{-1})\mu_2(K) \sum_{i=1}^{d} h_i^2 \psi_{m+2\iota(i)},$$

where $\iota(i)$ is the d-tuple having ith entry equal to 1 and all other entries zero, and

$$\operatorname{Var}\{\hat{\psi}_m(h_1, \ldots, h_d)\} \sim 2n^{-2}\left(\prod_{i=1}^{d} h_i^{-2m_i-1}\right)\psi_0 \int R(K^{(m)}) + 4n^{-1}\left\{\int f^{(m)}(x)^2 f(x)\,dx - \psi_m^2\right\}.$$

7. OTHER KERNEL ESTIMATORS

We now briefly discuss some other settings where error calculations can be done in the same way as in Section 2. These will be versions of the uniform design kernel regression estimator and the hazard rate function kernel estimator.

7.1. *Nonparametric Regression*

Consider the regression model

$$Y_i = m(x_i) + \varepsilon_i, \quad 1 \leq i \leq n,$$

where $x_1, \ldots, x_n$ is an equally-spaced grid in $[0, 1]^d$, m is a real-valued regression function on $[0, 1]^d$ and the ε_i are independent real-valued random variables satisfying $E(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. This is essentially the model considered by Staniswalis *et al.* (1991). The Priestley–Chao type general kernel estimator for m is

$$\hat{m}(x; H) = n^{-1} \sum_{i=1}^{n} K_H(x - x_i) Y_i.$$

and, under conditions on m and K similar to those imposed on f and K in Section 2, we can show that for x in the interior of $[0, 1]^d$,

$$E\hat{m}(x; H) - m(x) = \tfrac{1}{2}\mu_2(K)\,\mathrm{tr}\{HG_m(x)\}\{1 + o(1)\} + O(n^{-1/d})$$

and

$$\mathrm{Var}\{\hat{m}(x; H)\} = n^{-1}|H|^{-1/2} R(K)\sigma^2\{1 + o(1)\} + O(n^{-2/d}).$$

The $O(n^{-1/d})$ and $O(n^{-2/d})$ terms are due to Reimann sum approximations to integrals. It is interesting to note that these terms dominate the asymptotically optimal error for $d \geq 5$ (Müller 1988, Staniswalis *et al.*, 1991).

7.2. *Hazard Rate Function Estimation*

Another problem where kernel methods provide an attractive technique is for estimation of the hazard rate function

$$\lambda(x) = f(x)/\{1 - F(x)\}, \quad F(x) < 1,$$

where the setting is the same as in Section 2 and F is the common distribution function of the sample. Patil, Wells & Marron (1991) give compelling reasons for using the kernel estimator

$$\hat{\lambda}(x; H) = n^{-1} \sum_{i=1}^{n} K_H(x - X_i)/\{1 - \hat{F}(X_i)\}$$

(Watson & Leadbetter, 1964) where $\hat{F}(x) = n^{-1}\{\sum_{i=1}^{n} I(x \leq X_i) - 1\}$ is a minor adjustment of the empirical distribution function. Because $\hat{F}$ converges uniformly to F at the rate $n^{-1/2}$ it follows that $\hat{\lambda}(\cdot; H)$ is asymptotically equivalent to $\bar{\lambda}(\cdot; H)$ where

$$\bar{\lambda}(x; H) = n^{-1} \sum_{i=1}^{n} K_H(x - X_i)/\{1 - F(X_i)\}.$$

Under the usual assumptions on λ and K we obtain

$$E\bar{\lambda}(x; H) - \lambda(x) \sim \tfrac{1}{2}\mu_2(K)\,\mathrm{tr}\{HG_\lambda(x)\}$$

and

$$\mathrm{Var}\{\bar{\lambda}(x; H)\} \sim n^{-1}|H|^{-1/2} R(K)\lambda(x)/\{1 - F(x)\}.$$

ACKNOWLEDGEMENTS

This research was performed while the author was visiting the National University of Singapore and the Open University, Milton Keynes, England. The support and hospitality of these two institutions is gratefully acknowledged. The author also thanks Dr M. C. Jones for his helpful suggestions and comments and Dr Doug Moore for his production of Figure 1.

APPENDIX

Proof of Theorem 1

In view of (2.7) we need only deal with the integrated squared bias component. The proof relies upon the following readily established results for multivariate normal distributions $N(\mu, \Sigma)$ and $N(\mu', \Sigma')$:

$$G_{\phi_\Sigma(\cdot-\mu)}(x) = \phi_\Sigma(x-\mu)\{\Sigma^{-1}(x-\mu)(x-\mu)^T - I\}\Sigma^{-1}, \tag{A.1}$$

$$\phi_\Sigma(x-\mu)\phi_{\Sigma'}(x-\mu') = \phi_{\Sigma+\Sigma'}(\mu-\mu')\phi_{\Sigma(\Sigma+\Sigma')^{-1}\Sigma'}(x-\mu^*) \tag{A.2}$$

where

$$\mu^* = \Sigma'(\Sigma+\Sigma')^{-1}\mu + \Sigma(\Sigma+\Sigma')^{-1}\mu',$$

and

$$\mathrm{Cov}(X^TAX, (X-c)^TB(X-c)) = 2\,\mathrm{tr}[A\Sigma B\{\Sigma + 2(\mu-c)\mu^T\}], \tag{A.3}$$

where X is a $N(\mu, \Sigma)$ random vector, A and B are constant $d \times d$ symmetric matrices and c is a constant $d \times 1$ vector. Result (A.3) can be derived using results given in Seber (1977).

From (2.7), (A.1) and (A.2),

$$\begin{aligned}\int \mathrm{tr}^2\{HG_f(x)\} = \sum_{l=1}^{k}\sum_{l'=1}^{k} & w_l w_{l'}\phi_{\Sigma_l+\Sigma_{l'}}(\mu_l - \mu_{l'}) \\ & \times E(\mathrm{tr}[H\Sigma_l^{-1}\{(Y-\mu_l)(Y-\mu_l)^T\Sigma_l^{-1} - I\}] \\ & \times \mathrm{tr}[H\Sigma_{l'}^{-1}\{(Y-\mu_{l'})(Y-\mu_{l'})^T\Sigma_{l'}^{-1} - I\}])\end{aligned} \tag{A.4}$$

where Y is a $N(\mu^*_{l,l'}, \Sigma_l(\Sigma_l+\Sigma_{l'})^{-1}\Sigma_{l'})$ random vector and

$$\mu^*_{l,l'} = \Sigma_{l'}(\Sigma_l+\Sigma_{l'})^{-1}\mu_l + \Sigma_l(\Sigma_l+\Sigma_{l'})^{-1}\mu_{l'}.$$

Since $E(UV) = \mathrm{Cov}(U, V) + E(U)E(V)$ for two random variables U and V the above expectation can be written

$$\begin{aligned}&\mathrm{Cov}\{(Y-\mu_l)^T\Sigma_l^{-1}H\Sigma_l^{-1}(Y-\mu_l), (Y-\mu_{l'})^T\Sigma_{l'}^{-1}H\Sigma_{l'}^{-1}(Y-\mu_{l'})\} \\ &\quad + \mathrm{tr}(H\Sigma_l^{-1}[E\{(Y-\mu_l)(Y-\mu_l)^T\}\Sigma_l^{-1} - I]) \\ &\quad \times \mathrm{tr}(H\Sigma_{l'}^{-1}[E\{(Y-\mu_{l'})(Y-\mu_{l'})^T\}\Sigma_{l'}^{-1} - I]).\end{aligned}$$

Noting that $\mu^*_{l,l'} - \mu_l = \Sigma_l(\Sigma_l + \Sigma_{l'})^{-1}(\mu_{l'} - \mu_l)$, (A.3) and direct matrix algebra can be used that the covariance term is

$$2\,\mathrm{tr}[H(\Sigma_l + \Sigma_{l'})^{-1}H(\Sigma_l + \Sigma_{l'})^{-1}\{I - 2(\mu_l - \mu_{l'})(\mu_l - \mu_{l'})^T(\Sigma_l + \Sigma_{l'})^{-1}\}].$$

Using $E\{(Y - \mu_l)(Y - \mu_l)^T\} = \Sigma_l(\Sigma_l + \Sigma_{l'})^{-1}\Sigma_{l'} + (\mu^*_{l,l'} - \mu_l)(\mu^*_{l,l'} - \mu_l)^T$ we can show that each of the factors in the second term is equal to

$$-\mathrm{tr}[H(\Sigma_l + \Sigma_{l'})^{-1}\{I - (\mu_l - \mu_{l'})(\mu_l - \mu_{l'})^T(\Sigma_l + \Sigma_{l'})^{-1}\}].$$

Combining these with (A.4) and applying the definitions of $A_{l,l'}$, $B_{l,l'}$ and $C_{l,l'}$ leads to the required result.

Proof of Theorem 2

Our proof uses the matrix differential calculus techniques of Magnus & Neudecker (1988). The notation and references in this appendix all pertain to this source.

Using a standard result for the differential of $|X|$ (p. 178) we have

$$\begin{aligned}
\mathrm{d}\,|X|^{-1/2} &= -\tfrac{1}{2}\,|X|^{-3/2}\,\mathrm{d}\,|X| \\
&= -\tfrac{1}{2}\,|X|^{-1/2}\,\mathrm{tr}(X^{-1}\,\mathrm{d}X) \\
&= -\tfrac{1}{2}\,|X|^{-1/2}\,(\mathrm{vec}\,X^{-1'})^T\,\mathrm{d}\,\mathrm{vec}\,X \\
&= -\tfrac{1}{2}\,|X|^{-1/2}\,(\mathrm{vec}\,X^{-1'})^T\,\mathrm{d}\,\mathrm{vech}\,X \\
&= -\tfrac{1}{2}\,|X|^{-1/2}\,(D_d^T\,\mathrm{vec}\,X^{-1'})^T\,\mathrm{d}\,\mathrm{vech}\,X.
\end{aligned}$$

Result (a) follows from the first identification theorem (p. 175) and (3.1).

For (b), the second differential of $|X|^{-1/2}$ is

$$\begin{aligned}
\mathrm{d}^2\,|X|^{-1/2} &= -\tfrac{1}{2}\,\mathrm{d}(|X|^{-3/2}\,\mathrm{d}\,|X|) \\
&= -\tfrac{1}{2}\{(d\,|X|^{-3/2})\,\mathrm{d}\,|X| + |X|^{-3/2}\,\mathrm{d}^2\,|X|\} \\
&= \tfrac{1}{4}\,|X|^{-5/2}\,\{3(\mathrm{d}\,|X|)^2 - 2\,|X|\,\mathrm{d}^2\,|X|\}.
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{d}^2\,|X| &= \mathrm{d}\{|X|\,\mathrm{tr}(X^{-1}\,\mathrm{d}X)\} \\
&= \mathrm{d}\,|X|\,\mathrm{tr}(X^{-1}\,\mathrm{d}X) + |X|\,\mathrm{tr}(X^{-1}\,\mathrm{d}X) \\
&= |X|\,[\mathrm{tr}^2(X^{-1}\,\mathrm{d}X) - \mathrm{tr}\{X^{-1}(\mathrm{d}X)X^{-1}(\mathrm{d}X)\}]
\end{aligned}$$

which leads to

$$\mathrm{d}^2\,|X|^{-1/2} = \tfrac{1}{4}\,|X|^{-1/2}\,[\mathrm{tr}^2(X^{-1}\,\mathrm{d}X) - \mathrm{tr}\{X^{-1}(\mathrm{d}X)X^{-1}(\mathrm{d}X)\}].$$

After some straightforward matrix algebra (using, for example, $\mathrm{vec}(ABC) = (C^T \otimes A)(\mathrm{vec}\,B)$) we arrive at

$$\begin{aligned}
\mathrm{d}^2\,|X|^{-1/2} = \tfrac{1}{4}\,|X|^{-1/2}\,&(\mathrm{d}\,\mathrm{vec}\,X)^T K_d(X^{-1} \otimes I)^T \\
&\times [(\mathrm{vec}\,I)(\mathrm{vec}\,I)^T + 2I](I \otimes X^{-1})(\mathrm{d}\,\mathrm{vec}\,X)
\end{aligned}$$

where K_d is the commutation matrix of order d (p. 47). Replacing of $\mathrm{vec}\,X$ by $D_d\,\mathrm{vech}\,X$, noting the results $K_d D_d = D_d$ (Theorem 3.12) and $K_d^T = K_d$ (p. 47) and applying the second identification theorem (Table 10.1, p. 190) we obtain the desired result.

REFERENCES

Deheuvels, P. (1977). Estimation non parametrique de la densité par histogrammes generalisés (II), *Publ. l'Inst. Statist. l'Univ. Paris,* **22,** 1–23.

Gasser, T., Kneip, A. & Köhler, W. (1991). A fast and flexible method for automatic smoothing, *J. Amer. Statist. Assoc.,* **86,** 643–652.

Hall, P. & Marron, J. S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.,* **6,** 109–115.

Härdle, W. (1990). *Applied Nonparametric Regression.* Boston: Cambridge University Press.

Henderson, H. V. & Searle, S. R. (1979). Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Can. J. Statist.,* **7,** 65–81.

Jones, M. C. & Sheather, S. J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statist. Probab. Lett.,* **11,** 511–514.

Magnus, J. R. & Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Chichester: Wiley.

Müller, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data.* Berlin: Springer-Verlag.

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

Patil, P. N., Wells, M. T. & Marron, J. S. (1991). Kernel based estimators of ratio functions. *J. Nonpar. Statist.,* to appear.

Scott, D. W. (1986). Data analysis in 3 and 4 dimensions with nonparametric density estimation. In *Statistical Image Processing and Graphics,* E. J. Wegman and D. De Priest, Eds., New York: Marcel Dekker, pp. 291–305.

Scott, D. W. & Wand, M. P. (1991). Feasibility of multivariate density estimates. *Biometrika,* **78,** 197–206.

Seber, G. A. F. (1977). *Linear Regression Analysis.* New York: Wiley.

Sheather, S. J. & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Statist. Soc. B,* **53,** 683–690.

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

Staniswalis, J. G., Messer, K. & Finston, D. R. (1991). Kernel estimators for multivariate smoothing. Unpublished manuscript.

Terrell, G. R. & Scott, D. W. (1991). Variable kernel density estimation. *Ann. Statist.,* to appear.

Wand, M. P. & Jones, M. C. (1991). Comparison of smoothing parameterizations in bivariate kernel density estimation. Unpublished manuscript.

Watson, G. S. & Leadbetter, M. R. (1964). Hazard analysis II. *Sankyā Ser. A,* **26,** 101–116.