Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation

M. P. Wand; M. C. Jones

Journal of the American Statistical Association, Vol. 88, No. 422. (Jun., 1993), pp. 520-528.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199306%2988%3A422%3C520%3ACOSPIB%3E2.0.CO%3B2-X

Journal of the American Statistical Association is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org
Mon Aug 6 14:22:58 2007

Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation

M. P. WAND and M. C. JONES*

The basic kernel density estimator in one dimension has a single smoothing parameter, usually referred to as the bandwidth. For higher dimensions, however, there are several options for smoothing parameterization of the kernel estimator. For the bivariate case, there can be between one and three independent smoothing parameters in the estimator, which leads to a flexibility versus complexity trade-off when using this estimator in practice. In this article the performances of the different possible smoothing parameterizations are compared, using both the asymptotic and exact mean integrated squared error. Our results show that it is important to have independent smoothing parameters for each of the coordinate directions. Although this is enough for many situations, for densities with high amounts of curvature in directions different to those of the coordinate axes, substantial gains can be made by allowing the kernel mass to have arbitrary orientations. The "sphering" approaches to choosing this orientation are shown to be detrimental in general, however.

KEY WORDS: Bandwidth selection; Exact mean integrated squared error; Kernel estimator; Normal mixture density.

1. INTRODUCTION

Recently, there has been a renewed interest in nonparametric density estimation. This is partly an attempt to bring this important data analytic tool closer to practical implementation and partly because density estimation provides a simple setting to understand important aspects of nonparametric curve estimation in general, a field that has many applications. Härdle (1990a) has provided many such examples in the nonparametric regression context.

One of the driving forces behind this renewed interest is the monograph of Silverman (1986), which covered most of the main density estimation methodology up to the time of its publication. However, since then there have been many important refinements and additions to this methodology in an effort to make density estimation even more practical. Most of these have been for the intuitively simple kernel estimator and include improved data-based procedures for selecting the smoothing parameter (e.g., Sheather and Jones 1991), highly efficient computing algorithms (Härdle 1990b; Scott 1985) and adaptations to the kernel method to enhance its flexibility (Hall and Marron 1988; Wand, Marron, and Ruppert 1991). This research has led to the availability of rapidly computed, automatically generated and high-quality density estimates, even for challenging density shapes.

Many of these refinements have been studied in the univariate setting for simplicity's sake. But Silverman (1986, chap. 4) and Scott (1992), for example, have demonstrated that density estimation can also be a useful practical tool in moderate higher dimensions as well. For just how many more dimensions direct generalizations of the type considered here remain viable is still a moot point (see, for example, Scott and Wand 1991), but there certainly is much potential for applicability to two and three dimensions in which direct visualization of surfaces is possible. We thus believe that such multivariate methods represent a major growth area; witness recent work such as Müller and Prewitt (1991) and Terrell and Scott (1992).

Thus it seems well worthwhile to explore the extension of recent density estimation methodology to higher-dimensional settings. An important starting point is the parameterization of the kernel estimator in more than one dimension. In its basic form the univariate kernel estimator has one smoothing parameter, usually called the bandwidth or window width. For higher dimensions there are several levels of options. The simplest multivariate kernel estimator has just one bandwidth (Cacoullos 1966). This means that the amount of smoothing is the same in every direction. An obvious extension is to have a different bandwidth for each of the coordinate directions (Epanechnikov 1969). One can go even further than this, however, and have a bandwidth matrix that permits smoothing in orientations different from those of the coordinate directions (Deheuvels 1977). Along with completely general utilizations of these strategies, we also consider versions involving variances and covariances associated with the density. Of course, adding extra smoothing parameters to the estimator increases its flexibility, but for implementation this means that more parameters must be selected. This is especially undesirable if the choice of bandwidth is subjective. Extensions of univariate automatic bandwidth selection methodology to the multivariate case could also become difficult and computationally expensive. We will address the question of automatic bivariate bandwidth choice in a forthcoming paper.

The purpose of this article is to investigate the effect of the various possible parameterizations of the bivariate kernel density estimator. This estimator is described in the next section. Besides the fact that the bivariate case is quite important in its own right, we have chosen to concentrate on this setting to keep our presentation simple and informative and in the hope that the lessons learned there are applicable to higher dimensions and other nonparametric curve estimation settings.

* M. P. Wand is Visiting Assistant Professor, Department of Statistics, Rice University, Houston, TX 77251. M. C. Jones is Lecturer, Department of Statistics, The Open University, Milton Keynes, MK7 6AA, U.K. The research of M. P. Wand was partially supported by Office of Naval Research Grant N00014-90-J-1176. The authors thank J. S. Marron and the referees for their helpful comments and D. W. Scott for a useful suggestion.

© 1993 American Statistical Association
Journal of the American Statistical Association
June 1993, Vol. 88, No. 422, Theory and Methods

Our comparison between parameterizations is done in two ways. First, in Section 3 we present some asymptotic theory for the mean integrated squared error (MISE), which gives important insights into the issue. Second, in Section 4 we perform calculations for a set of example densities, which are mixtures of normal densities. We concentrate on this wide class of densities because they admit explicit expressions for both the exact MISE and its asymptotic approximation, which greatly simplifies the computations. Finally, in Section 5 we summarize our findings and make recommendations for practice.

2. MULTIVARIATE KERNEL DENSITY ESTIMATION

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a $\mathbb{R}^d$-valued random sample with density f. In its most general form, the global bandwidth kernel estimator of f is $\hat{f}(\mathbf{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$, where K is a d-variate function that we take to be a probability density itself, $\mathbf{H}$ is a symmetric positive definite $d \times d$ matrix, and for $\mathbf{x} \in \mathbb{R}^d$, $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$. For technical convenience we will measure the global error incurred by using $\hat{f}(\mathbf{x}; \mathbf{H})$ using $\mathrm{MISE}(\mathbf{H}) = E \int \{\hat{f}(\mathbf{x}; \mathbf{H}) - f(\mathbf{x})\}^2 \, d\mathbf{x}$, where an unqualified integral is taken to mean integration over $\mathbb{R}^d$.

We now turn our attention to the bivariate case ($d = 2$). The first important classes of permissible values of $\mathbf{H}$ we consider are (with $\mathbf{I}$ standing for the identity matrix)

$$\mathscr{H}_1 = \{h_1^2 \mathbf{I}: h_1 > 0\}, \qquad \mathscr{H}_2 = \{\mathrm{diag}(h_1^2, h_2^2): h_1, h_2 > 0\}$$

and

$$\mathscr{H}_3 = \left\{ \begin{bmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{bmatrix} : h_1, h_2 > 0, \, |h_{12}| < h_1 h_2 \right\}.$$

Note that $\mathscr{H}_1 \subseteq \mathscr{H}_2 \subseteq \mathscr{H}_3$ and that each of these classes represent estimators with one, two, and three independent smoothing parameters, as the subscripts suggest.

It is easiest to understand the types of estimation performed within each of these classes by taking K to be the Gaussian kernel

$$K(\mathbf{x}) = (2\pi)^{-1} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right). \tag{1}$$

In this case $K_{\mathbf{H}}(\mathbf{x}) = (2\pi)^{-1}|\mathbf{H}|^{-1/2}\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{H}^{-1}\mathbf{x})$, which is simply the density of the bivariate normal distribution with a mean vector of $\mathbf{0}$ and a covariance matrix equal to $\mathbf{H}$. The restriction $\mathbf{H} \in \mathscr{H}_1$ means that the kernel mass will always be spherically symmetric (with circular contours). The addition of an extra smoothing parameter for the second dimension, so that $\mathbf{H} \in \mathscr{H}_2$, means that the Gaussian kernel contours may be elliptical but with elliptical axes parallel to the coordinate axes. In the full matrix situation, $\mathbf{H} \in \mathscr{H}_3$, the kernel density contours are elliptical but with arbitrary orientation.

A common practical approach to multivariate smoothing is to first rescale the data so that the sample variances are equal for each dimension. We will refer to this approach as *scaling*. An extension of this idea, attributed to Fukunaga (1972), involves linearly transforming the data so that the sample covariance matrix is the identity. This operation is sometimes referred to as *sphering* the data. For these reasons we shall also consider the classes

$$\mathscr{C}_2 = \{h^2\mathbf{D}: h > 0\} \quad \text{and} \quad \mathscr{C}_3 = \{h^2\mathbf{C}: h > 0\},$$

where $\mathbf{C}$ is the covariance matrix corresponding to the density f having (i, j) entry c_{ij} and $\mathbf{D} = \mathrm{diag}(c_{11}, c_{22})$. Of course, $\mathscr{C}_i \subseteq \mathscr{H}_i$ for $i = 2, 3$. Asymptotic theory for the bivariate normal density presented in Section 3 provides further motivation for considering the class $\mathscr{C}_3$.

We consider one other class that combines the notions of sphering and having independent smoothing parameters in each direction. This will be called the *hybrid* parameterization; the corresponding bandwidth matrix class is

$$\mathscr{Y} = \left\{ \begin{bmatrix} h_1^2 & \rho_{12}h_1h_2 \\ \rho_{12}h_1h_2 & h_2^2 \end{bmatrix} : h_1, h_2 > 0 \right\},$$

where $\rho_{12} = c_{12}/(c_{11}c_{22})^{1/2}$ is the correlation coefficient of the density f.

In the next two sections we compare the performances of kernel estimators when constrained to lie within each of these classes, to determine if and when significant gains can be made by more sophisticated smoothing parameterizations.

3. ASYMPTOTIC THEORY AND COMPARISON

In this section we take K to be the standard bivariate normal kernel, as defined by (1), and take f to be a bivariate density. We are concerned with the approximations

$$\mathbf{H}_{\mathscr{A},\mathrm{AMISE}} = \arg \inf_{\mathbf{H} \in \mathscr{A}} \mathrm{AMISE}(\mathbf{H}) \quad \text{and} \quad \inf_{\mathbf{H} \in \mathscr{A}} \mathrm{AMISE}(\mathbf{H}), \tag{2}$$

where AMISE is the first order asymptotic approximation of MISE and $\mathscr{A}$ can be any one of the bandwidth matrix classes discussed in Section 2. These asymptotics hold under the assumptions that f has all second-order partial derivatives bounded, continuous and square integrable and that all entries of $\mathbf{H}$ as well as $n^{-1}|\mathbf{H}|^{-1/2}$ tend to 0 as $n \to \infty$.

For a bivariate function g and $\mathbf{r} = (r_1, r_2)$, define $g^{(\mathbf{r})}(\mathbf{x}) = \partial^{r_1+r_2} g(\mathbf{x}) / \partial x_1^{r_1} \partial x_2^{r_2}$ and set $\psi_{r_1,r_2} = \int f^{(r_1,r_2)}(\mathbf{x}) f(\mathbf{x}) \, d\mathbf{x}$, assuming that f is sufficiently smooth for this quantity to exist. The formulas that follow in this section are not entirely novel individually (see, for example, Deheuvels 1977), but we believe this to be an unprecedented collection and comparison of particular cases.

For $\mathbf{H} \in \mathscr{H}_3$, the asymptotic approximation of $\mathrm{MISE}(\mathbf{H})$ is

$$\mathrm{AMISE}(\mathbf{H}) = (4\pi n)^{-1}|\mathbf{H}|^{-1/2} + \frac{1}{4}\psi_{4,0}h_1^4 + \psi_{3,1}h_1^2h_{12} + \frac{1}{2}\psi_{2,2}(h_1^2h_2^2 + 2h_{12}^2) + \psi_{1,3}h_2^2h_{12} + \frac{1}{4}\psi_{0,4}h_2^4.$$

We do not believe that it is possible, in general, to obtain explicit formulas for the quantities in (2) when $\mathscr{A}$ is $\mathscr{H}_3$ or $\mathscr{Y}$ and have had to resort to numerical computations to per-

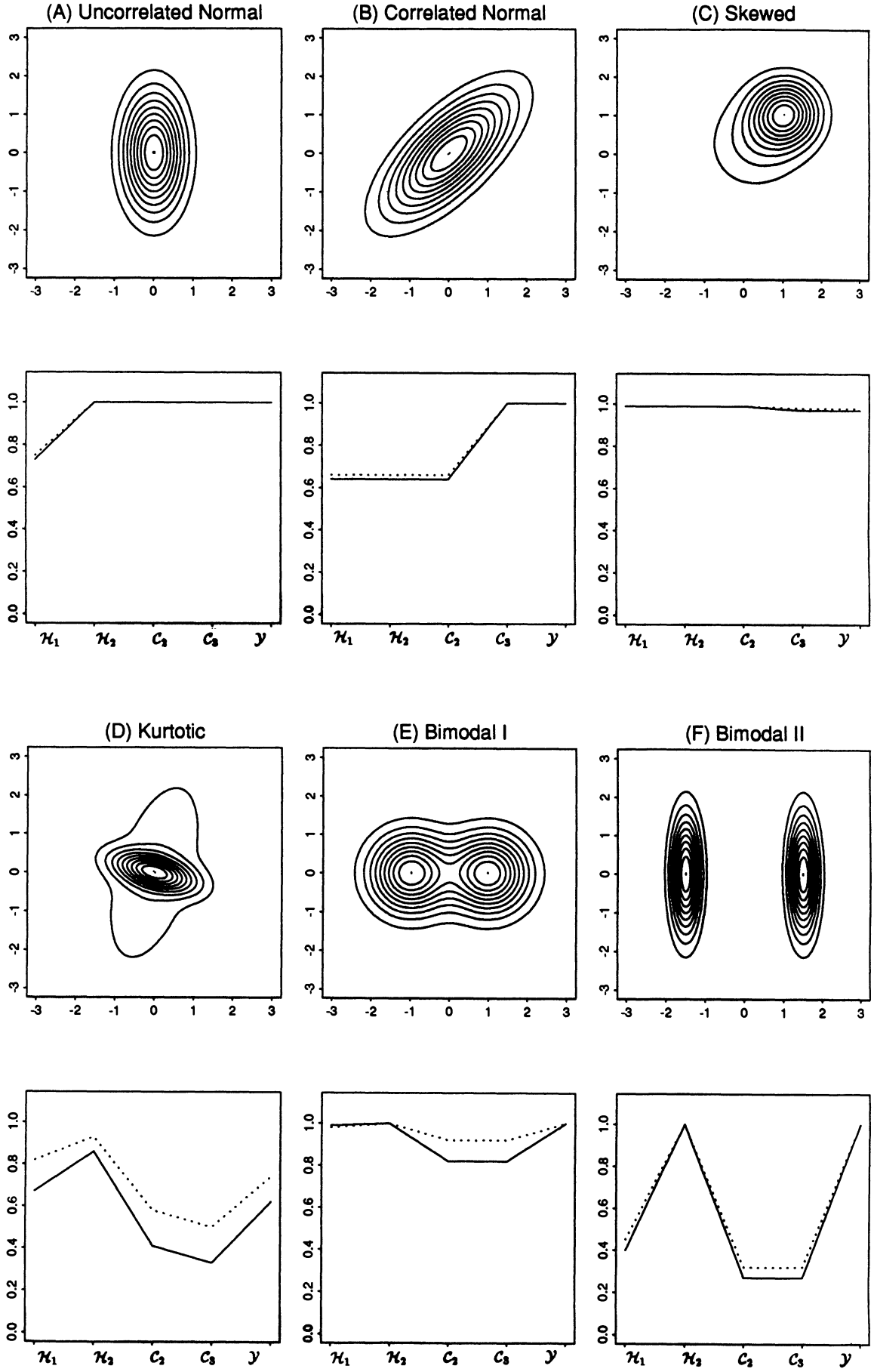

Figure 1. Contour Plots for the 12 Example Densities and Line Graphs (A–L) Showing Their Relative Efficiencies Compared to $\mathcal{H}_3$. Solid lines are used for $ARE_f(\mathcal{H}_3 : \mathcal{B})$ while dotted lines are used for $RE_f(\mathcal{H}_3 : \mathcal{B})$.

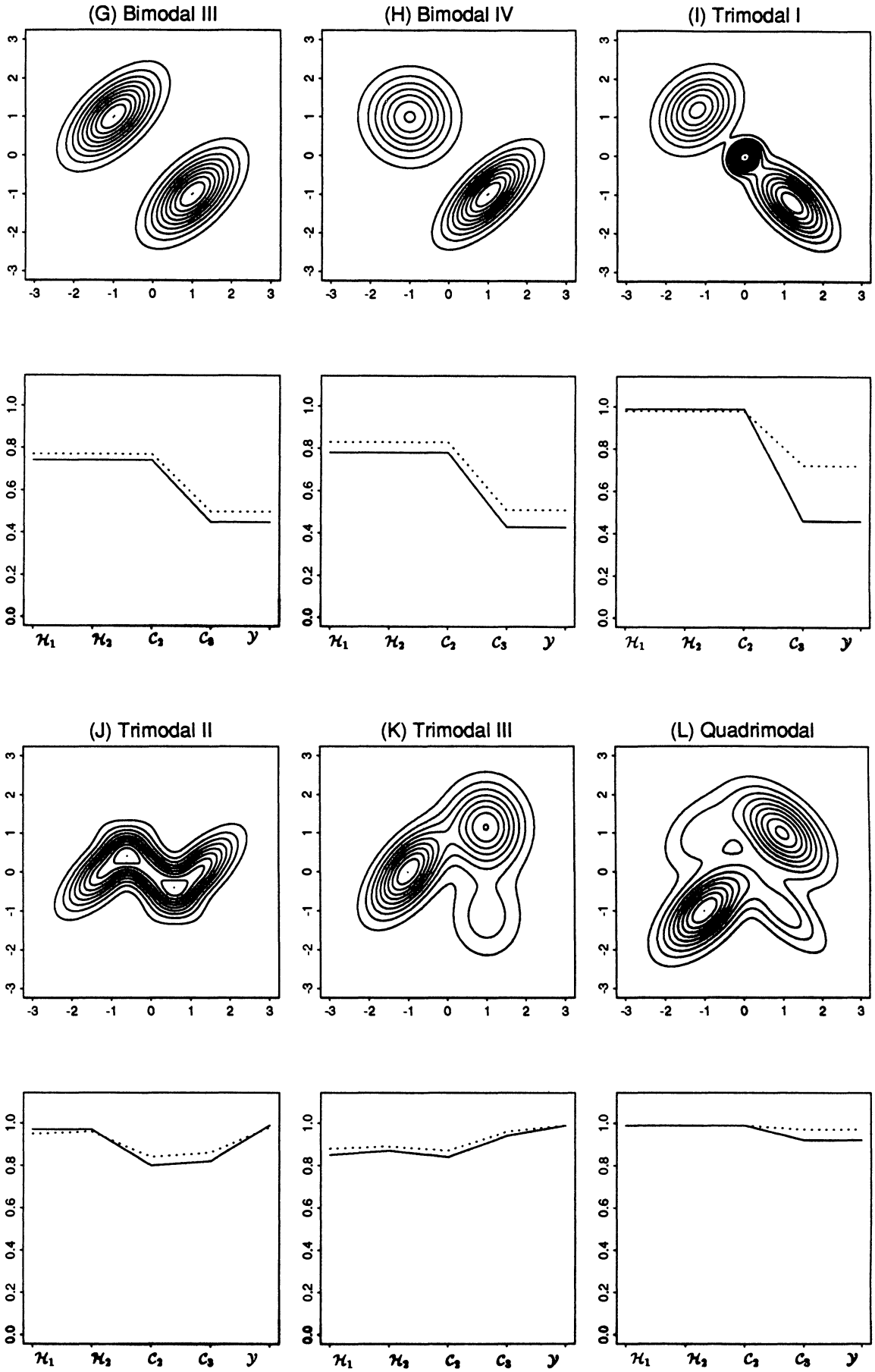

Figure 1. (Continued)

form the minimization in the examples of Section 4. For $\mathcal{A}$ being each of $\mathscr{H}_1$, $\mathscr{H}_2$, $\mathscr{C}_2$, and $\mathscr{C}_3$, however, elegant expressions for the quantities in (2) are available. Each minimum AMISE can be written in the form $\inf_{\mathbf{H}\in\mathcal{A}}\text{AMISE}(\mathbf{H}) = \frac{3}{2}(4\pi n)^{-2/3} \times \mathcal{T}_{\mathcal{A}}(f)^{1/3}$. In particular, $\mathcal{T}_{\mathscr{H}_1}(f) = \psi_{2,2} + \frac{1}{2}(\psi_{4,0} + \psi_{0,4})$ and $\mathcal{T}_{\mathscr{H}_2}(f) = \psi_{2,2} + (\psi_{4,0}\psi_{0,4})^{1/2}$. Of course, $\mathcal{T}_{\mathscr{H}_1}(f) \geq \mathcal{T}_{\mathscr{H}_2}(f)$, with the amount of improvement due to using $\mathscr{H}_2$ dependent on the amount by which the arithmetic mean of $\psi_{4,0}$ and $\psi_{0,4}$ exceeds the geometric mean of the two (and their sizes relative to $\psi_{2,2}$). Clearly, a single bandwidth does as well as two bandwidths if and only if $\psi_{4,0} = \psi_{0,4}$. But the fact that $\psi_{4,0}$ and $\psi_{0,4}$ can be made arbitrarily different by using different units in one of the coordinate directions means that one can always do arbitrarily better using two bandwidths by simply rescaling the data. It might be useful to note that $\mathbf{H}_{\mathscr{H}_2,\text{AMISE}} = \{4\pi n\mathcal{T}_{\mathscr{H}_2}(f)\}^{-1/6}\text{diag}(R, R^{-1})$, where $R = (\psi_{0,4}/\psi_{4,0})^{1/8}$.

For $\mathscr{C}_2$ we have $\mathcal{T}_{\mathscr{C}_2}(f) = \psi_{2,2} + \frac{1}{2}|\mathbf{D}|^{-1}(\psi_{4,0}c_{11}^2 + \psi_{0,4}c_{22}^2)$. Again, performance depends on combinations of $\psi_{4,0}$ and $\psi_{0,4}$; the weights attached to these are $\frac{1}{2}c_{11}/c_{22}$ and $\frac{1}{2}c_{22}/c_{11}$. It is easy to show that $\mathcal{T}_{\mathscr{C}_2}(f) \geq \mathcal{T}_{\mathscr{H}_2}(f)$, as intuition demands—but also note that $\mathbf{H} \in \mathscr{C}_2$ is by no means always better than $\mathbf{H} \in \mathscr{H}_1$.

Finally, the expression for $\mathcal{T}_{\mathscr{C}_2}(f)$ is a special case of $\mathcal{T}_{\mathscr{C}_3}(f)$ given by

$$\mathcal{T}_{\mathscr{C}_3}(f) = \psi_{2,2} + \frac{1}{2}|\mathbf{C}|^{-1}(\psi_{4,0}c_{11}^2 + 4\psi_{3,1}c_{11}c_{12} + 6\psi_{2,2}c_{12}^2 + 4\psi_{1,3}c_{22}c_{12} + \psi_{0,4}c_{22}^2).$$

Because the quantities $\psi_{3,1}$, $\psi_{1,3}$, and c_{12} can be either positive or negative, the class $\mathscr{C}_3$ can be either better or worse than $\mathscr{C}_2$, depending on the density.

As an important special case we consider the bivariate $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density having correlation coefficient ρ. Then, if K is the bivariate Gaussian kernel, it can be shown (Wand 1992) that

$$\mathbf{H}_{\mathscr{H}_3,\text{AMISE}} = \boldsymbol{\Sigma}n^{-1/3} \quad \text{and} \quad \inf_{\mathbf{H}\in\mathscr{H}_3}\text{AMISE}(\mathbf{H}) = \{3/(8\pi)\}|\boldsymbol{\Sigma}|^{-1/2}n^{-2/3}. \quad (3)$$

This result for the asymptotically optimal bandwidth matrix is very pleasing from an intuitive viewpoint. It simply says that to optimally estimate a bivariate normal density, one should have kernel mass with the same covariance structure as the density itself. This result also provides one motivation for considering Fukunaga's (1972) sphering approach, and hence the class $\mathscr{C}_3$, because $\mathbf{H}_{\mathscr{H}_3,\text{AMISE}} = \mathbf{H}_{\mathscr{C}_3,\text{AMISE}}$ in this case. Furthermore, this result might be used as the basis for bivariate (and, more generally, multivariate) extensions of the so-called "normal scale" rule (Silverman 1986, p. 45), which has often been suggested as a useful starting point for bandwidth selection. Such a bandwidth selection rule is $\hat{\mathbf{H}} = \mathbf{S}n^{-1/3}$, where $\mathbf{S}$ is the sample covariance matrix; as such, it can do no better than a sample-based version of $\mathbf{H}_{\mathscr{C}_3,\text{AMISE}}$. Therefore, results presented in Section 4 suggest that this rule is especially problematic in higher dimensions.

It can be shown that for this bivariate normal density,

$$\inf_{\mathbf{H}\in\mathscr{H}_2}\text{AMISE}(\mathbf{H}) = \{3/(8\pi)\}|\boldsymbol{\Sigma}|^{-1/2}[(2+\rho^2)/\{2(1-\rho^2)\}]^{1/3}n^{-2/3}.$$

So the asymptotic relative efficiency of $\mathscr{H}_3$ compared to $\mathscr{H}_2$ is the following function of ρ:

$$\begin{aligned}\text{ARE}(\mathscr{H}_3 : \mathscr{H}_2) &\equiv \{\inf_{\mathbf{H}\in\mathscr{H}_3}\text{AMISE}(\mathbf{H})/\inf_{\mathbf{H}\in\mathscr{H}_2}\text{AMISE}(\mathbf{H})\}^{3/2} \\ &= \{2(1-\rho^2)/(2+\rho^2)\}^{1/2}.\end{aligned}$$

The interpretation of $\text{ARE}(\mathscr{H}_3 : \mathscr{H}_2)$ is that, for large n, the minimum error using n observations and $\mathbf{H} \in \mathscr{H}_2$ can be achieved using only $\text{ARE}(\mathscr{H}_3 : \mathscr{H}_2)n$ observations if $\mathbf{H} \in \mathscr{H}_3$. For $|\rho|$ taking the values .3, .6, and .9, $\text{ARE}(\mathscr{H}_3 : \mathscr{H}_2)$ takes the values .93, .74, and .37 and $\text{ARE}_f(\mathscr{H}_3 : \mathscr{H}_2) \to 0$ as $|\rho| \to 1$, indicating that there can be appreciable costs from not using the full bandwidth matrix in this case.

4. EXACT MISE CALCULATIONS AND COMPARISON BY EXAMPLES

Although the asymptotic theory discussed in Section 3 provides important insight into the comparison of the bandwidth matrix classes, it is also desirable to investigate what happens for certain important and interesting cases. This can be done relatively easily for both the asymptotic and finite-sample MISE by taking our example densities from the class of general normal mixture densities. In d dimensions, these are of the form

$$f(\mathbf{x}) = \sum_{l=1}^{k} w_l\phi_{\boldsymbol{\Sigma}_l}(\mathbf{x} - \boldsymbol{\mu}_l), \quad (4)$$

where $\phi_{\boldsymbol{\Sigma}}(\mathbf{x}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp(-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})$ is the $N(0, \boldsymbol{\Sigma})$ density, $\mathbf{w} = (w_1, \ldots, w_k)^T$ is a vector of positive weights summing to 1, and, for each l, $\boldsymbol{\mu}_l$ is a $d \times 1$ mean vector and $\boldsymbol{\Sigma}_l$ is a $d \times d$ covariance matrix. The normal mixture densities form a very rich class that includes virtually any density shape. For exact (nonasymptotic) MISE calculations, we appeal to Theorem 1.

Theorem 1. If K is the multivariate Gaussian kernel ϕ_I and f is a d-variate normal mixture density as in (4), then

$$\text{MISE}(\mathbf{H}) = n^{-1}(4\pi)^{-d/2}|\mathbf{H}|^{-1/2} + \mathbf{w}^T\{(1-n^{-1})\boldsymbol{\Omega}_2 - 2\boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_0\}\mathbf{w}$$

where $\boldsymbol{\Omega}_a$ is the $k \times k$ matrix having (l, l') entry equal to $\phi_{a\mathbf{H}+\boldsymbol{\Sigma}_l+\boldsymbol{\Sigma}_{l'}}(\boldsymbol{\mu}_l - \boldsymbol{\mu}_{l'})$.

The proof of Theorem 1 is given in the Appendix. This result is the multivariate extension of Theorem 2.1 of Marron and Wand (1992). These authors exploit the flexibility of normal mixture densities and explicit MISE representation in the univariate setting.

Write $\boldsymbol{\Lambda}_{\mathbf{r}}$ for the $k \times k$ matrix having (l, l') entry equal to $\phi^{(\mathbf{r})}_{\boldsymbol{\Sigma}_l+\boldsymbol{\Sigma}_{l'}}(\boldsymbol{\mu}_l - \boldsymbol{\mu}_{l'})$. Here we have used the notation $g_A^{(\mathbf{r})}(\mathbf{x}) = \partial^{r_1+\cdots+r_d}g_A(\mathbf{x})/\partial x_1^{r_1}\ldots\partial x_d^{r_d}$. Then, the AMISE for bi-

variate normal mixture densities is given by the following theorem.

Theorem 2. If K is the bivariate Gaussian kernel $\phi_{\mathbf{I}}$ and f is a bivariate normal mixture density as in (4), ($d = 2$), then for $\mathbf{H} \in \mathscr{H}_3$,

$$\mathrm{AMISE}(\mathbf{H}) = (4\pi n)^{-1}|\mathbf{H}|^{-1/2} + \frac{1}{4}\mathbf{w}^T\{\boldsymbol{\Lambda}_{(4,0)}h_1^4 + 4\boldsymbol{\Lambda}_{(3,1)}h_1^2h_{12} + 2\boldsymbol{\Lambda}_{(2,2)}(h_1^2h_2^2 + 2h_{12}^2) + 4\boldsymbol{\Lambda}_{(1,3)}h_2^2h_{12} + \boldsymbol{\Lambda}_{(0,4)}h_2^4\}\mathbf{w}.$$

We also outline the proof of Theorem 2 in the Appendix.

These results imply that within the class of normal mixture densities, MISE($\mathbf{H}$) and AMISE($\mathbf{H}$) can be computed by direct algebraic calculations. To keep the computational burden to a minimum, we selected 12 example bivariate densities from this class. They were chosen to encompass many interesting features, but at the same time most of them are densities that one would usually hope to resolve reasonably well using a kernel estimator and a moderately sized sample. These are displayed in contour form in Figure 1; the parameters are listed in Table 1. For ease of presentation, Table 1 uses the bivariate normal notation $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, where the marginal means and variances are μ_i and σ_i^2, $i = 1, 2$ and the correlation coefficient is ρ.

To investigate both the finite sample and large sample situations, we computed values of $\inf_{\mathbf{H}\in\mathscr{A}}\mathrm{MISE}(\mathbf{H})$ when $n = 100$ and $\inf_{\mathbf{H}\in\mathscr{A}}\mathrm{AMISE}(\mathbf{H})$ for $\mathscr{A}$ as each of the classes discussed in Section 3. The numerical minimization of MISE($\mathbf{H}$) was done using an initial grid search over the space of possible smoothing parameter values within the particular class, followed by Newton's method improvement. The grid was logarithmically equally spaced and centered around the asymptotically optimal values of $\mathbf{H}$. Provided the initial grid was not too coarse, rapid convergence to the minimum was always achieved. Formula sheets detailing the minimization strategy are available on request from the first author. Because we are interested in comparing the minimum errors between classes, it is appropriate to consider values of the relative efficiency (RE): $\mathrm{RE}_f(\mathscr{A} : \mathscr{B}) \equiv \{\inf_{\mathbf{H}\in\mathscr{A}}\mathrm{MISE}(\mathbf{H})/\inf_{\mathbf{H}\in\mathscr{B}}\mathrm{MISE}(\mathbf{H})\}^{3/2}$ and $\mathrm{ARE}_f(\mathscr{A} : \mathscr{B})$ for classes $\mathscr{A}$ and $\mathscr{B}$. The RE also has a relative sample size interpretation similar to ARE, although the power $\frac{3}{2}$ is based on asymptotic considerations.

Although our comparisons are performed at the minimum MISE and AMISE, it should be noted that optimization of these error criteria is difficult in practice when using data-based bandwidth selectors. But because we are only concerned with the relative merits of each smoothing parameterization, comparison of their performances at the minima means that our results are not confounded with the bandwidth selection problem.

A visual impression of the variation in minimum errors is given in Figure 1, where line graphs connecting values of $\mathrm{RE}_f(\mathscr{H}_3 : \mathscr{B})$ and $\mathrm{ARE}_f(\mathscr{H}_3 : \mathscr{B})$ for each class $\mathscr{B}$ are displayed underneath the corresponding contour plot. An initial impression from Figure 1 is that the asymptotics are generally very good for this comparison, with RE and ARE telling much the same story. This is interesting; although the asymptotic integrated squared bias approximation need not be all that good (as discussed for the univariate case by Marron and Wand, in press), it is clear that taking ratios alleviates the problem.

The class $\mathscr{H}_1$ is included only for completeness. As the AMISE results given in Section 3 indicate, one should not blindly use a single bandwidth for unscaled multivariate data, and so class $\mathscr{H}_1$ is not a serious competitor.

The most striking outcome of the comparison study in Figure 1 is how poorly classes $\mathscr{C}_2$ and $\mathscr{C}_3$ perform. The sphering approach works perfectly for the bivariate normal densities, which, in view of (3), is not surprising. But for a density such as (F), both scaling and sphering can be very detrimental. For this density the variance of the horizontal coordinate variable is a very poor surrogate for measuring the optimal amount of smoothing in the horizontal direction, because it does not take into account the "within modes" curvature. One could make $\mathscr{C}_2$ and $\mathscr{C}_3$ perform arbitrarily

Table 1. Parameters for 12 Example Bivariate Normal Mixture Densities

Density	$w_1N(\mu_{11}, \mu_{12}, \sigma_{11}^2, \sigma_{12}^2, \rho_1) + \cdots + w_kN(\mu_{k1}, \mu_{k2}, \sigma_{k1}^2, \sigma_{k2}^2, \rho_k)$
(A) Uncorrelated normal	$N(0, 0, \frac{1}{4}, 1, 0)$
(B) Correlated normal	$N(0, 0, 1, 1, \frac{7}{10})$
(C) Skewed	$\frac{1}{5}N(0, 0, 1, 1, 0) + \frac{1}{5}N(\frac{1}{2}, \frac{1}{2}, (\frac{2}{3})^2, (\frac{2}{3})^2, 0) + \frac{3}{5}N(\frac{13}{12}, \frac{13}{12}, (\frac{5}{9})^2, (\frac{5}{9})^2, 0)$
(D) Kurtotic	$\frac{2}{3}N(0, 0, 1, 4, \frac{1}{2}) + \frac{1}{3}N(0, 0, (\frac{2}{3})^2, (\frac{1}{3})^2, -\frac{1}{2})$
(E) Bimodal I	$\frac{1}{2}N(-1, 0, (\frac{2}{5})^2, (\frac{2}{5})^2, 0) + \frac{1}{2}N(1, 0, (\frac{2}{3})^2, (\frac{2}{3})^2, 0)$
(F) Bimodal II	$\frac{1}{2}N(-\frac{3}{2}, 0, (\frac{1}{4})^2, 1, 0) + \frac{1}{2}N(\frac{3}{2}, 0, (\frac{1}{4})^2, 1, 0)$
(G) Bimodal III	$\frac{1}{2}N(-1, 1, (\frac{2}{3})^2, (\frac{2}{3})^2, \frac{3}{5}) + \frac{1}{2}N(1, -1, (\frac{2}{3})^2, (\frac{2}{3})^2, \frac{3}{5})$
(H) Bimodal IV	$\frac{1}{2}N(1, -1, (\frac{2}{3})^2, (\frac{2}{3})^2, \frac{7}{10}) + \frac{1}{2}N(-1, 1, (\frac{2}{3})^2, (\frac{2}{3})^2, 0)$
(I) Trimodal I	$\frac{9}{20}N(-\frac{6}{5}, \frac{6}{5}, (\frac{3}{5})^2, (\frac{3}{5})^2, \frac{3}{10}) + \frac{9}{20}N(\frac{6}{5}, -\frac{6}{5}, (\frac{3}{5})^2, (\frac{3}{5})^2, -\frac{3}{5}) + \frac{1}{10}N(0, 0, (\frac{1}{4})^2, (\frac{1}{4})^2, \frac{1}{5})$
(J) Trimodal II	$\frac{1}{3}N(-\frac{6}{5}, 0, (\frac{3}{5})^2, (\frac{3}{5})^2, \frac{7}{10}) + \frac{1}{3}N(\frac{6}{5}, 0, (\frac{3}{5})^2, (\frac{3}{5})^2, \frac{7}{10}) + \frac{1}{3}N(0, 0, (\frac{3}{5})^2, (\frac{3}{5})^2, -\frac{7}{10})$
(K) Trimodal III	$\frac{3}{7}N(-1, 0, (\frac{3}{5})^2, (\frac{7}{10})^2, \frac{3}{5}) + \frac{3}{7}N(1, \frac{2\sqrt{3}}{3}, (\frac{3}{5})^2, (\frac{7}{10})^2, 0) + \frac{1}{7}N(1, -\frac{2\sqrt{3}}{3}, (\frac{3}{5})^2, (\frac{7}{10})^2, 0)$
(L) Quadrimodal	$\frac{1}{8}N(-1, 1, (\frac{2}{3})^2, (\frac{2}{3})^2, \frac{2}{5}) + \frac{3}{8}N(-1, -1, (\frac{2}{3})^2, (\frac{2}{3})^2, \frac{3}{5}) + \frac{1}{8}N(1, -1, (\frac{2}{3})^2, (\frac{2}{3})^2, -\frac{7}{10}) + \frac{3}{8}N(1, 1, (\frac{2}{3})^2, (\frac{2}{3})^2, -\frac{1}{2})$

poorly by taking the modes to be sufficiently far apart, because this increases the variance but not the optimal amount of smoothing. Similar comments apply to other nonnormal densities. Intuitively, in multimodal circumstances one then expects that different elliptically oriented kernels would be appropriate within modes, but taking a single overall orientation based on sphering is inappropriate. We conclude that sphering and scaling of the data is usually inadvisable for general density shapes if using a global bandwidth matrix.

The hybrid class $\mathcal{Y}$ is also motivated by the optimal bandwidth matrix result for bivariate normal data, and, as expected, it gives good performance for densities close to normal. Moreover, for densities such as (F), the flexibility of having an extra smoothing parameter in the horizontal direction overcomes one of the problems of scaling and sphering by allowing for the correct amount in each of the coordinate directions. But for nonnormal densities with different orientations to the coordinate axes, such as (G), (H), and (I), the hybrid class can also be ineffective, because the correlation coefficient is not an appropriate measure of orientation for these densities.

Class $\mathcal{H}_2$ does not have the idiosyncrasies that arise when the smoothing parameterization is based on the covariance matrix. Therefore, it performs very well when the curvature tends to be in the same direction as the coordinate axes and does not fail as badly as $\mathcal{Y}$ in the nonnormal cases when the orientation is different from the axes. But $\mathcal{H}_2$ is still not allowing for different orientations of the data; as we saw at the end of Section 3, it can be made to do arbitrarily poorly.

A deficiency of the comparison study in Figure 1 is that the relative efficiencies of $\mathcal{H}_2$, $\mathcal{C}_2$, and $\mathcal{Y}$ depend on the orientation of the coordinate axes with respect to the probability mass. This orientation is of course subject to a degree of arbitrariness. The perfect relative efficiencies of $\mathcal{H}_2$ for densities (E) and (F), for example, are due to the fact that the coordinate axes are parallel to the axes of the components of f. It is, therefore, important to see what effect the orientation has for the relative efficiencies of these classes.

Figure 2 shows plots of $\mathrm{ARE}_f(\mathcal{H}_3 : \mathcal{B})$ for various classes $\mathcal{B}$ as a function of rotation angle when the probability mass is rotated about the origin. Because each curve is a periodic function with period $\pi/2$, the plots are only for angles in the range $(0, \pi/2)$. The densities represented are (A), (D), (F), and (H). We believe that these plots allow a more complete understanding of the cost due to omitting or misspecifying the orientation of the kernel mass. The plot for density (A) shows $\mathcal{C}_3$ and $\mathcal{Y}$ with perfect ARE values. The solid curve in this plot corresponds to the identical ARE values of $\mathcal{H}_2$ and $\mathcal{C}_2$. We see that for densities with shapes (A), (D), and (H), the value of $\mathrm{ARE}_f(\mathcal{H}_3 : \mathcal{H}_2)$ is never less than about .65, but for (F) this value can be as low as about .40. It is also interesting to note that for the nonnormal densities presented here, classes $\mathcal{C}_2$, $\mathcal{C}_3$, and $\mathcal{Y}$ are uniformly worse than $\mathcal{H}_2$. In these cases $\mathcal{C}_2$ performs better for orientations where $\mathcal{H}_2$ and $\mathcal{Y}$ perform worse, and vice versa. Finally, note that for the nonnormal densities depicted in Figure 2, class $\mathcal{C}_2$ performs uniformly better than $\mathcal{C}_3$. For these densities we have the surprising result that the orientation selected by

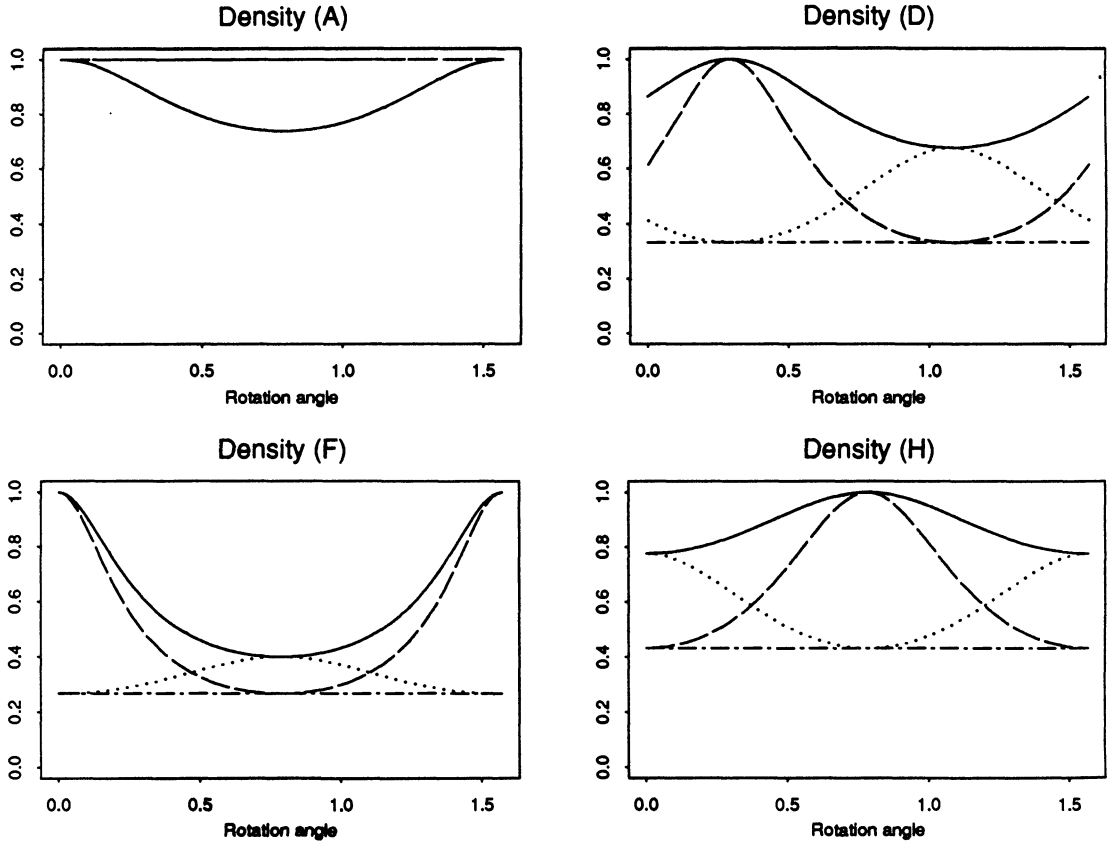

Figure 2. Plots of $\mathrm{ARE}_f(\mathcal{H}_3 : \mathcal{B})$ for Various Classes $\mathcal{B}$: $\mathcal{H}_2$ (solid line), $\mathcal{C}_2$ (dotted line), $\mathcal{C}_3$ (dot-dashed line) and $\mathcal{Y}$ (dashed line) for Densities (A), (D), (F), and (H) Against the Rotation Angle When Each Density Has Its Probability Mass Rotated About the Origin.

sphering is never better than any arbitrary orientation. At this stage we do not have a clear-cut explanation for this phenomenon, except to say that for nonnormal data, orientations based on the covariance matrix seem to be far from optimal.

5. CONCLUSIONS

The bivariate kernel density estimation problem, despite being the simplest multivariate curve estimation problem, presents many challenges when it comes to selecting the correct amount of smoothing. One of the main outcomes of this study is that smoothing strategies based on the covariance matrix are inappropriate in general and can be very detrimental for nonnormal data. This is because the entries of the covariance matrix are usually not able to take into account the curvature in f and its orientation.

It is clear that the kernel estimator should have the flexibility to smooth by different amounts independently in each direction, and that taking $\mathbf{H} \in \mathscr{H}_2$ will often be adequate. For a subjective choice of smoothing parameters this is good news, because in these cases one has to deal only with a single parameter for each coordinate direction. But as seen by the values of $\mathrm{ARE}_f(\mathscr{H}_3 : \mathscr{H}_2)$, there is also much to be gained by including an orientation parameter for other cases. This is, of course, equivalent to prerotating the data by the optimal amount and then using a diagonal bandwidth matrix. The "automation" of this rotation by a sphering approach is not appropriate. For bivariate data sets, one could presumably choose the optimal rotation fairly adequately by eye and do better than sphering. If automatic choice of the rotation amount is desired, then estimation of the optimal full bandwidth matrix may have to be considered.

APPENDIX: PROOFS

Proof of Theorem 1

Standard techniques show that for any two multivariate normal distributions $N(\mu, \Sigma)$ and $N(\mu', \Sigma')$, we have

$$\phi_{\Sigma}(\mathbf{x}-\mu)\phi_{\Sigma'}(\mathbf{x}-\mu') = \phi_{\Sigma+\Sigma'}(\mu-\mu')\phi_{\Sigma(\Sigma+\Sigma')^{-1}\Sigma'}(\mathbf{x}-\mu^*), \quad \text{(A.1)}$$

where $\mu^* = \Sigma'(\Sigma+\Sigma')^{-1}\mu + \Sigma(\Sigma+\Sigma')^{-1}\mu'$. A direct consequence of this is

$$\int \phi_{\Sigma}(\mathbf{x}-\mu)\phi_{\Sigma'}(\mathbf{x}-\mu')\,d\mathbf{x} = \phi_{\Sigma+\Sigma'}(\mu-\mu'). \quad \text{(A.2)}$$

We can now derive the formula for $\mathrm{MISE}(\mathbf{H}) = \int \mathrm{var}\{\hat{f}(\mathbf{x};\mathbf{H})\} \times d\mathbf{x} + \int \{E\hat{f}(\mathbf{x};\mathbf{H}) - f(\mathbf{x})\}^2\,d\mathbf{x}$ very simply by repeated application of this result. First, note that

$$\mathrm{var}\{\hat{f}(\mathbf{x};\mathbf{H})\} = n^{-1}[E\phi_{\mathbf{H}}(\mathbf{x}-\mathbf{X})^2 - \{E\phi_{\mathbf{H}}(\mathbf{x}-\mathbf{X})\}^2].$$

We have

$$\int E\phi_{\mathbf{H}}(\mathbf{x}-\mathbf{X})^2\,d\mathbf{x} = \int\int \phi_{\mathbf{H}}(\mathbf{x}-\mathbf{y})^2\,d\mathbf{x} f(\mathbf{y})\,d\mathbf{y} = \int \phi_{\mathbf{H}}(x)^2\,dx = \phi_{2\mathbf{H}}(0) = (2\pi)^{-d/2}|2\mathbf{H}|^{-1/2}.$$

Also

$$E\phi_{\mathbf{H}}(\mathbf{x}-\mathbf{X}) = \sum_{l=1}^{k} w_l \int \phi_{\mathbf{H}}(\mathbf{y}-\mathbf{x})\phi_{\Sigma_l}(\mathbf{y}-\mu_l)\,d\mathbf{y} = \sum_{l=1}^{k} w_l \phi_{\mathbf{H}+\Sigma_l}(\mathbf{x}-\mu_l),$$

which implies that

$$\int \{E\phi_{\mathbf{H}}(\mathbf{x}-\mathbf{X})\}^2\,dx = \sum_{l=1}^{k}\sum_{l'=1}^{k} w_l w_{l'} \int \phi\mathbf{H}+\Sigma_l(\mathbf{x}-\mu_l)\phi_{\mathbf{H}+\Sigma_{l'}}(\mathbf{x}-\mu_{l'})\,dx = \sum_{l=1}^{k}\sum_{l'=1}^{k} w_l w_{l'} \phi_{2\mathbf{H}+\Sigma_l+\Sigma_{l'}}(\mu_l-\mu_{l'}).$$

The integrated squared bias can be handled in a similar fashion by expanding the square and using (A.2). Introducing the Ω_a notation, we obtain the required result.

Proof of Theorem 2.

We will first prove the following extension of (A.2):

$$(-1)^{\sum_{i=1}^{d} r_i} \int \phi_{\Sigma}^{(\mathbf{r})}(\mathbf{x}-\mu)\phi_{\Sigma'}^{(\mathbf{r}')}(\mathbf{x}-\mu')\,dx = \phi_{\Sigma+\Sigma'}^{(\mathbf{r}+\mathbf{r}')}(\mu-\mu'). \quad \text{(A.3)}$$

Let the scaled convolution of two d-variate real valued functions f and g be denoted by $(f*g)(\mathbf{x}) = (2\pi)^{-d/2}\int f(\mathbf{u})g(\mathbf{x}-\mathbf{u})\,d\mathbf{u}$ and let the scaled Fourier transform of f be denoted by $\mathrm{FT}_f(\mathbf{t}) = (2\pi)^{-d/2}\int f(\mathbf{x})e^{-i\mathbf{t}^T\mathbf{x}}\,d\mathbf{x}$. Then, using (A.1), standard results for multivariate Fourier transforms (see, for example, Rudin 1973), and the notation $\mathbf{c}^{\mathbf{r}} = c_1^{r_1}\ldots c_d^{r_d}$ for a d-dimensional complex vector $\mathbf{c}$, we have

$$\begin{aligned}
&\mathrm{FT}_{\phi_{\Sigma}^{(\mathbf{r})}(\cdot-\mu)*\phi_{\Sigma'}^{(\mathbf{r}')}(\cdot-\mu')}(\mathbf{t}) \\
&= \mathrm{FT}_{\phi_{\Sigma}^{(\mathbf{r})}(\cdot-\mu)}(\mathbf{t})\mathrm{FT}_{\phi_{\Sigma'}^{(\mathbf{r}')}(\cdot-\mu')}(\mathbf{t}) \\
&= (i\mathbf{t})^{\mathbf{r}+\mathbf{r}'}e^{-i\mathbf{t}^T(\mu+\mu')}\phi_{\Sigma^{-1}}(\mathbf{t})\phi_{\Sigma'^{-1}}(\mathbf{t})|\Sigma|^{-1/2}|\Sigma'|^{-1/2} \\
&= (i\mathbf{t})^{\mathbf{r}+\mathbf{r}'}e^{-i\mathbf{t}^T(\mu+\mu')}(2\pi)^{-d/2}\phi_{(\Sigma+\Sigma')^{-1}}(\mathbf{t})|\Sigma+\Sigma'|^{-1/2} \\
&= (2\pi)^{-d/2}\mathrm{FT}_{\phi_{\Sigma+\Sigma'}^{(\mathbf{r}+\mathbf{r}')}(\cdot-\mu-\mu')}(\mathbf{t}).
\end{aligned}$$

So, by the Fourier Inversion Theorem,

$$\phi_{\Sigma}^{(\mathbf{r})}(\cdot-\mu)*\phi_{\Sigma'}^{(\mathbf{r}')}(\cdot-\mu')(\mathbf{x}) = \phi_{\Sigma+\Sigma'}^{(\mathbf{r}+\mathbf{r}')}(\mathbf{x}-\mu-\mu').$$

Setting $\mathbf{x} = \mathbf{0}$ and replacing μ by $-\mu$, we obtain (A.3). Theorem 2 now follows directly from (A.3) and the introduction of the $\Lambda_{\mathbf{r}}$ notation.

[*Received April 1991. Revised April 1992.*]

REFERENCES

Cacoullos, T. (1966), "Estimation of a Multivariate Density," *Annals of the Institute of Statistical Mathematics,* 18, 179–189.

Deheuvels, P. (1977), "Estimation Non Parametrique de la Densité par Histogrammes Generalisés (II)," *Publications de l'Institut Statistique de l'Université de Paris,* 22, 1–23.

Epanechnikov, V. A. (1969), "Non-Parametric Estimation of a Multivariate Probability Density," *Theory of Probability and Its Applications,* 14, 153–158.

Fukunaga, K. (1972), *Introduction to Statistical Pattern Recognition,* New York: Academic Press.

Hall, P., and Marron, J. S. (1988), "Variable Window Width Kernel Estimates of Probability Densities," *Probability Theory and Related Fields,* 80, 37–49.

Härdle, W. (1990a), *Applied Nonparametric Regression,* Cambridge, U.K.: Cambridge University Press.

——— (1990b), *Smoothing Techniques With Implementation in S,* New York: Springer-Verlag.

Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics,* 20, 712–736.

Müller, H.-G., and Prewitt, K. A. (1991), "Applications of Multiparameter Weak Convergence for Adaptive Nonparametric Curve Estimation," in *Nonparametric Functional Estimation and Related Topics,* ed. G. G. Roussas, Dordrecht: Klüwer, pp. 141–166.

Rudin, W. (1973), *Functional Analysis,* New York: McGraw-Hill.

Scott, D. W. (1985), "Average Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions," *The Annals of Statistics,* 13, 1024–1040.

——— (1992), *Multivariate Density Estimation,* New York: John Wiley.

Scott, D. W., and Wand, M. P. (1991), "Feasibility of Multivariate Density Estimates," *Biometrika,* 78, 197–206.

Sheather, S. J., and Jones, M. C. (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society,* Ser. B, 53, 683–690.

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

Terrell, G. R., and Scott, D. W. (1992), "Variable Kernel Density Estimates," *The Annals of Statistics,* 20, 1236–1265.

Wand, M. P. (1992), "Error Analysis for General Multivariate Kernel Estimators," *Journal of Nonparametric Statistics,* 2, 1–15.

Wand, M. P., Marron, J. S., and Ruppert, D. (1991), "Transformations in Density Estimation" with comments, *Journal of the American Statistical Association,* 86, 343–361.