Computational Statistics (1994) 9:97-116

Computational Statistics

Multivariate Plug-in Bandwidth Selection

M. P. Wand
Australian Graduate School of Management, University of New South Wales, Kensington, NSW 2033, Australia

M. C. Jones
Department of Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, England

Keywords: Bandwidth matrix, bivariate data, kernel estimator, mean squared relative error, nonparametric density estimation.

Abstract

Recently several high performance bandwidth selectors for univariate kernel density estimation have been proposed in the literature. In the multivariate case there has been relatively little work. Currently available selectors for multivariate data include cross-validation methods, such as least squares cross-validation, and oversmoothing. However, least squares cross-validation suffers from a high degree of sample variability while oversmoothing will often smooth out important features of the data. In the univariate context bandwidth selection strategies based on "plug-in" ideas have been shown to exhibit good theoretical and practical performance. In this article we investigate the multivariate extensions of plug-in bandwidth selection. It is seen through analysis and a simulation study that the good properties of the plug-in approach carry over to moderate dimensional settings.

1. Introduction

One of the most important issues in nonparametric curve estimation is the choice of the smoothing parameters. In the past few years there have been several significant breakthroughs in data-based smoothing parameter selection. These include important theoretical results on the problem itself such as bounds on the rates of convergence of any selector (Hall and Marron, 1991) as well as several methods which exhibit good asymptotic properties (e.g. Park and Marron, 1990, Hall, Sheather, Jones and Marron, 1991, Jones, Marron and Park, 1991, Sheather and Jones, 1991, Hall and Johnstone, 1992, Hall, Marron and Park, 1992, Chiu 1991, 1992). Of these, some transfer their asymptotic performance well to finite sample situations, while others are less successful. Virtually all of this work

0943-4062/94/2/97-116 $2.50

has been done in the simplest setting of univariate kernel density estimation where there is a single smoothing parameter, usually called the bandwidth or window width.

While a fast rate of convergence to the optimum is a desirable property of a bandwidth selector it is reliable finite sample performance that is of most importance to practitioners. Simulation studies have been carried out to assess the practical performance of the currently available selectors (see e.g. Chiu, 1992, Jones, Marron and Sheather, 1992, Park and Turlach, 1992) and while some seem to perform well in many important situations there does not yet seem to be a selector which does well for absolutely all density shapes. Nevertheless, some of these have been seen to perform very well in a wide variety of cases, particularly those based on the plug-in approach with non-negative kernels. See Jones, Marron and Sheather (1992) for a review of the univariate bandwidth selection problem.

The restriction to the univariate setting has simplified the analysis in understanding the smoothing parameter selection problem. However, there are many important applications of nonparametric curve estimation in higher dimensions. It may be argued that curve estimation is more profitable in multivariate settings since detection of structure in higher dimensional point clouds is more difficult. Scott (1992a) gives several excellent examples of finding structure in multivariate data sets using kernel density estimation. In most cases the smoothing parameters of a multivariate density estimator are chosen by the data analyst. However, as in the univariate setting, there are many good reasons for having an automatic data-driven technique for choosing them. Currently available selection procedures are least-squares cross-validation (see e.g. Stone, 1984), biased cross-validation (Sain, Baggerly and Scott, 1992) and oversmoothing (Terrell, 1990). However, the former of these suffers from a high degree of sample variability (see e.g. Hall and Marron, 1987a) while estimates based on oversmoothing tend to mask important features in the data and do not converge to the optimum. It is therefore desirable to pursue more stable and consistent bandwidth selectors for multivariate data.

Virtually all of the current univariate bandwidth selectors can be extended to the multivariate case in some fashion. Because of its good theoretical and practical performance in the univariate setting we have chosen to concentrate on the multivariate extensions of the plug-in selector of Sheather and Jones (1991). We give arguments that suggest that the multivariate plug-in selector also has good theoretical properties for moderate dimensional data and a small simulation study shows it to be reliable in practice in the bivariate case. It is hoped that this study will also illuminate the problems associated with smoothing parameter selection in higher dimensions for other settings such as nonparametric regression.

There are some important differences that arise when extending the Sheather and Jones plug-in selector from the univariate to the multivariate case. Firstly, there are several levels of options for parameterizing multivariate kernel estimators as discussed by Wand and Jones (1993). The more sophisticated smoothing parameterizations allow more flexibility for the size and direction of smoothing, but their automatic choice using

plug-in methods is more computationally expensive. However, an apparently simplistic parameterization has a useful role at one stage in affording a desirable bias cancellation. Secondly, except in the simple case where the multivariate kernel estimator is parameterized by a single smoothing parameter, there is no direct analogue of the "solve the equation" approach which has been used in univariate plug-in strategies (see Park and Marron 1990, Sheather and Jones 1991).

Section 2 contains relevant theory for multivariate kernel density estimation. In Section 3 we provide theory for the estimation of unknown functionals which arise in the asymptotic expressions of Section 2. Multivariate plug-in bandwidth selection strategies are discussed in Section 4. Theoretical backup for our selection procedure is provided by Section 5 where convergence rates to the optimum are obtained. We conclude with a small simulation study and a real data example in Section 6 and some remarks in Section 7.

2. Multivariate Kernel Density Estimation

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample of $\mathbb{R}^d$-valued random vectors having density f. The general kernel estimator of f is

$$\hat{f}(x; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i) \tag{2.1}$$

where $\mathbf{H}$ is a $d \times d$ symmetric positive definite matrix, $K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{u})$ and K is a d-variate spherically symmetric density function.

The choice of the matrix $\mathbf{H}$, which we will call the bandwidth matrix, is very important. This is most easily understood by taking K to be the $N(\mathbf{0}, \mathbf{I}_d)$ density which implies that $K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$ is the $N(\mathbf{X}_i, \mathbf{H})$ density in the vector $\mathbf{x}$. In this case (2.1) is constructed by placement of kernel mass about each data point and the contours of this mass are d-dimensional ellipsoids with equation of the form $\mathbf{x}^T\mathbf{H}^{-1}\mathbf{x} = c$ for constants $c > 0$. Therefore $\mathbf{H}$ controls both the size and orientation of these ellipsoids and, hence, the amount and direction of the local averaging performed by the kernel estimator.

Let $\mathcal{F}$ denote the class of symmetric, positive definite $d \times d$ matrices. If $\mathbf{H} \in \mathcal{F}$ then $\mathbf{H}$ has $\frac{1}{2}d(d+1)$ independent entries which, even for moderate d, can be a substantial number of smoothing parameters to have to select. A simplification of (2.1) can be obtained by imposing the restriction $\mathbf{H} \in \mathcal{D}$, where $\mathcal{D} \subseteq \mathcal{F}$ is the subclass of diagonal positive definite $d \times d$ matrices. For $\mathbf{H} \in \mathcal{D}$, suppose that $\mathbf{H} = \operatorname{diag}\{h_1^2, \ldots, h_d^2\}$. Let $\mathbf{h} = (h_1, \ldots, h_d)^T$ and $\mathbf{h}^{-1} = (h_1^{-1}, \ldots, h_d^{-1})^T$. The kernel estimator can then be written

$$\hat{f}(x; \mathbf{h}) = n^{-1} \left(\prod_{\ell=1}^{d} h_\ell \right)^{-1} \sum_{i=1}^{n} K\{(\mathbf{h}^{-1}) \odot (\mathbf{x} - \mathbf{X}_i)\} \tag{2.2}$$

where $\mathbf{A} \odot \mathbf{B}$ denotes the Hadamard or "element-wise" product of equally sized matrices $\mathbf{A}$ and $\mathbf{B}$. A further simplification follows from the restriction $\mathbf{H} \in \mathcal{S}$, where $\mathcal{S} = \{h^2\mathbf{I} : h > 0\}$ and leads to the single bandwidth kernel estimator

$$\hat{f}(x; h) = n^{-1}h^{-d}\sum_{i=1}^{n} K\{(\mathbf{x} - \mathbf{X}_i)/h\}. \tag{2.3}$$

While the estimators (2.2) and (2.3) are simpler than (2.1) in terms of having less smoothing parameters to deal with, there is often a price to be paid in terms of flexibility. A comparison of these bandwidth classes in the bivariate case is given by Wand and Jones (1993). There it was shown that the restriction $\mathbf{H} \in \mathcal{S}$ is usually too severe: one should at least have an independent bandwidth for each coordinate direction which corresponds to $\mathbf{H} \in \mathcal{D}$. However, there are some occasions where it would be advantageous to use a full $\mathbf{H}$ matrix. Another noteworthy point is that use of a full $\mathbf{H}$ matrix is equivalent to allowing for arbitrary pre-rotations of the data and then using a diagonal $\mathbf{H}$ matrix. Other "intermediate" approaches to choosing the bandwidth matrix based on covariance structure of the data are also shown to be inappropriate in general by Wand and Jones (1993).

The choice of $\mathbf{H}$ is usually based on minimisation of some global error criterion. The simplest criterion to work with is mean integrated squared error (MISE) given by

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

Let $\mathrm{vech}\,\mathbf{H}$ be the $\frac{1}{2}d(d+1) \times 1$ vector containing the entries of $\mathbf{H}$ that are on or below the main diagonal of $\mathbf{H}$, listed column by column. Since $\mathbf{H}$ is symmetric, $\mathrm{vech}\,\mathbf{H}$ contains all the distinct entries of $\mathbf{H}$. A useful approximation to $\mathrm{MISE}\{\hat{f}(\cdot; \mathbf{H})\}$ is the asymptotic MISE (AMISE) of $\hat{f}(\cdot; \mathbf{H})$ given by

$$\mathrm{AMISE}\{\hat{f}(\cdot; \mathbf{H})\} = n^{-1}|\mathbf{H}|^{-1/2}R(K) + \tfrac{1}{4}\mu_2(K)^2(\mathrm{vech}\,\mathbf{H})^T\Psi_{\mathcal{F}}(\mathrm{vech}\,\mathbf{H}) \tag{2.4}$$

(Wand, 1992). Here $R(K) = \int K(\mathbf{x})^2 \, d\mathbf{x}$ and $\mu_2(K) = \int x_1^2 K(\mathbf{x}) \, d\mathbf{x} < \infty$. Also, $\Psi_{\mathcal{F}}$ denotes the $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$ matrix given by

$$\Psi_{\mathcal{F}} = \int \mathrm{vech}\,\{2\mathcal{H}_f(\mathbf{x}) - \mathrm{dg}\mathcal{H}_f(\mathbf{x})\}[\mathrm{vech}\,\{2\mathcal{H}_f(\mathbf{x}) - \mathrm{dg}\mathcal{H}_f(\mathbf{x})\}]^T \, d\mathbf{x}$$

where $\mathcal{H}_f(\mathbf{x})$ is the Hessian matrix of f. The notation dg denotes the diagonal matrix formed by replacing all off-diagonal entries by zeroes. Wand (1992) showed that if the entries of $\mathcal{H}_f(\mathbf{x})$ are continuous and square integrable and all entries of $\mathbf{H}$ as well as $n^{-1}|\mathbf{H}|^{-1/2}$ tend to zero as $n \to \infty$ then

$$\mathrm{MISE}\{\hat{f}(\cdot; \mathbf{H})\} = \mathrm{AMISE}\{\hat{f}(\cdot; \mathbf{H})\} + o\{n^{-1}|\mathbf{H}|^{-1/2} + \mathrm{tr}^2(\mathbf{H})\}.$$

For a d-tuple $\mathbf{m} = (m_1, \ldots, m_d)$ of non-negative integers we define $|\mathbf{m}| = \sum_{i=1}^{d} m_i$ and for a generic real-valued d-variate function g and matrix $\mathbf{M} \in \mathcal{F}$,

$$g_{\mathbf{M}}^{(\mathbf{m})}(\mathbf{x}) = \frac{\partial^{|\mathbf{m}|}}{\partial^{m_1} x_1 \cdots \partial^{m_d} x_d} g_{\mathbf{M}}(\mathbf{x}),$$

where $g_{\mathbf{M}}(\mathbf{x}) = |\mathbf{M}|^{-1/2} g(\mathbf{M}^{-1/2}\mathbf{x})$. An important observation is that each entry of $\mathbf{\Psi}_{\mathcal{F}}$ can be written in terms of expressions of the form $\psi_{\mathbf{m}}$ where

$$\psi_{\mathbf{m}} = \int f^{(\mathbf{m})}(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}$$

provided each $f^{(\mathbf{m})}$ exists. Using integration by parts it may also be established that $\psi_{\mathbf{m}}$ is non-zero only if $|\mathbf{m}|$ is even. For example, if $d = 2$ it is easily shown that

$$\mathbf{\Psi}_{\mathcal{F}} = \begin{bmatrix} \psi_{40} & 2\psi_{31} & \psi_{22} \\ 2\psi_{31} & 4\psi_{22} & 2\psi_{13} \\ \psi_{22} & 2\psi_{13} & \psi_{04} \end{bmatrix}.$$

Notice that the AMISE expression depends only on f through the $\psi_{\mathbf{m}}$ functionals. We investigate their estimation in the next section.

In general the minimisation of $\mathrm{AMISE}\{\hat{f}(\cdot;\mathbf{H})\}$ can only be performed numerically. Wand (1992) derives the necessary formulae for minimisation of $\mathrm{AMISE}\{\hat{f}(\cdot;\mathbf{H})\}$ when $\mathbf{H} \in \mathcal{F}$ via Newton-Raphson iteration.

If the data analyst believes that the density can be adequately estimated using a bandwidth for each direction then significant simplification can be made by taking $\mathbf{H} \in \mathcal{D}$. This involves $\mathbf{H} = \mathrm{diag}(\mathbf{h} \odot \mathbf{h})$ as mentioned above. In this case the derivative formulae for Newton-Raphson need to be adjusted. Let $\mathbf{\Psi}_{\mathcal{D}}$ be the $d \times d$ matrix having (i,j) entry equal to $\psi_{2\mathbf{e}_i + 2\mathbf{e}_j}$ where $\mathbf{e}_i$ is the d-tuple having 1 in the ith position and zero elsewhere. Then the AMISE of $\hat{f}(\cdot;\mathbf{h})$ is given by

$$\mathrm{AMISE}\{\hat{f}(\cdot;\mathbf{h})\} = n^{-1} R(K) \left(\prod_{j=1}^{d} h_j \right)^{-1} + \tfrac{1}{4} \mu_2(K)^2 (\mathbf{h}^2)^T \mathbf{\Psi}_{\mathcal{D}} \mathbf{h}^2 \qquad (2.5)$$

where $\mathbf{h}^2 = \mathbf{h} \odot \mathbf{h}$. This is a simplification of (2.4) for diagonal bandwidth matrices. For $d > 2$, (2.5) also can only be minimised numerically. If $\mathbf{h}_1$ is a reasonable approximation to the optimal $\mathbf{h}$ vector $\mathbf{h}_{\mathrm{AMISE}}$, then $\mathbf{h}_{\mathrm{AMISE}}$ is the limit of the sequence $\mathbf{h}_1, \mathbf{h}_2, \ldots$ where

$$\mathbf{h}_{i+1} = \mathbf{h}_i - \left[\frac{\partial^2 \mathrm{AMISE}\{\hat{f}(\cdot;\mathbf{h})\}}{(\partial \mathbf{h})(\partial \mathbf{h})^T} \right]_{\mathbf{h}=\mathbf{h}_i}^{-1} \left[\frac{\partial \mathrm{AMISE}\{\hat{f}(\cdot;\mathbf{h})\}}{\partial \mathbf{h}} \right]_{\mathbf{h}=\mathbf{h}_i},$$

$$\frac{\partial \mathrm{AMISE}\{\hat{f}(\cdot;\mathbf{h})\}}{\partial \mathbf{h}} = -n^{-1} R(K) \mathbf{h}^{-1} \left(\prod_{j=1}^{d} h_j \right)^{-1} + \mu_2(K)^2 \mathbf{h} \odot (\mathbf{\Psi}_{\mathcal{D}} \mathbf{h}^2)$$

and

$$\frac{\partial^2 \mathrm{AMISE}\{\hat{f}(\cdot;\mathbf{h})\}}{(\partial\mathbf{h})(\partial\mathbf{h})^T} = n^{-1}R(K)\left(\prod_{j=1}^{d} h_j\right)^{-1}\{(\mathbf{h}^{-1})(\mathbf{h}^{-1})^T + \mathrm{diag}(\mathbf{h}^{-2})\} + \mu_2(K)^2\{2(\mathbf{h}\mathbf{h}^T)\odot\boldsymbol{\Psi}_{\mathcal{D}} + \mathrm{diag}(\boldsymbol{\Psi}_{\mathcal{D}}\mathbf{h}^2)\}.$$

In the bivariate case (2.5) can be minimised analytically to give

$$\mathbf{h}_{\mathrm{AMISE}} = (h_{1,\mathrm{AMISE}}, h_{2,\mathrm{AMISE}})^T$$

where

$$h_{1,\mathrm{AMISE}} = \left[\frac{\psi_{04}^{3/4}R(K)}{\mu_2(K)^2\psi_{40}^{3/4}(\psi_{40}^{1/2}\psi_{04}^{1/2}+\psi_{22})n}\right]^{1/6} \text{ and } h_{2,\mathrm{AMISE}} = (\psi_{40}/\psi_{04})^{1/4}h_{1,\mathrm{AMISE}}.$$

3. Kernel Estimation of $\psi_{\mathbf{m}}$ Functionals

The only unknowns in the AMISE expressions of the previous section are the $\psi_{\mathbf{m}}$ integrals. Wand (1992) considers kernel estimators of the form

$$\hat{\psi}_{\mathbf{m}}(\mathbf{A}) = n^{-1}\sum_{i=1}^{n}\hat{f}^{(\mathbf{m})}(\mathbf{X}_i;\mathbf{A}) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_{\mathbf{A}}^{(\mathbf{m})}(\mathbf{X}_i-\mathbf{X}_j) \tag{3.1}$$

where the matrix $\mathbf{A} \in \mathcal{F}$. This estimator is the natural extension of the estimators of integrated squared density derivatives of Jones and Sheather (1991) and a slight, but important, adjustment to those considered by Hall and Marron (1987b). While one could, in principle, work with the full $\mathbf{A}$ matrix, the asymptotically optimal $\mathbf{A}$, which is central to the plug-in ideas, is not easy to obtain. The same is true for the restriction $\mathbf{A} \in \mathcal{D}$. We will therefore sacrifice some flexibility at the functional estimation stage and work with $\mathbf{A} \in \mathcal{S}$. Fortuitously, the full benefit of a bias reduction afforded by estimators of the form (3.1) is still realised when $\mathbf{A} \in \mathcal{S}$, at least in terms of convergence rates. Because a single smoothing parameter is being used for each coordinate direction it is recommended that each variable be rescaled so that the chosen scale measurement is the same for each coordinate direction. In this case $\mathbf{A} = a^2\mathbf{I}$ for some $a > 0$ and the leading terms of the MSE are

$$\mathrm{AMSE}\{\hat{\psi}_{\mathbf{m}}(a)\} = 2n^{-2}\psi_{\mathbf{0}}R(K^{(\mathbf{m})})a^{-2|\mathbf{m}|-d} + \left[n^{-1}a^{-|\mathbf{m}|-d}K^{(\mathbf{m})}(\mathbf{0}) + \tfrac{1}{2}a^2\mu_2(K)\left(\sum_{i=1}^{d}\psi_{\mathbf{m}+2\mathbf{e}_i}\right)\right]^2 \tag{3.2}$$

Before derivation of the AMSE-optimal a it will be convenient to make some additional assumptions on K. If all m_ℓ are even then we will assume that $\mathrm{sgn}\, K^{(\mathbf{m})}(\mathbf{0}) = (-1)^{|\mathbf{m}|/2}$

while $K^{(\mathbf{m})}(\mathbf{0}) = 0$ if at least one m_ℓ is odd. These conditions are satisfied by most of the common kernels including the Gaussian. If all the m_ℓ are even then $K^{(\mathbf{m})}(\mathbf{0})$ and $\psi_{\mathbf{m}+2\mathbf{e}_i}$ will be of opposite sign, so the leading bias term can be eliminated by taking

$$a_{\mathrm{AMSE}} = \left[\frac{-2K^{(\mathbf{m})}(\mathbf{0})}{\mu_2(K) \left(\sum_{i=1}^{d} \psi_{\mathbf{m}+2\mathbf{e}_i} \right) n} \right]^{1/(2+d+|\mathbf{m}|)} . \tag{3.3}$$

This results in a minimised AMSE of order $n^{-\min\{8,(d+4)\}/(2+d+|\mathbf{m}|)}$. However, if at least one m_ℓ is odd, $K^{(\mathbf{m})}(\mathbf{0}) = 0$ and the asymptotically optimal a is

$$a_{\mathrm{AMSE}} = \left[\frac{2\psi_{\mathbf{0}}(2|\mathbf{m}| + d) R(K^{(\mathbf{m})})}{\mu_2(K)^2 \left(\sum_{i=1}^{d} \psi_{\mathbf{m}+2\mathbf{e}_i} \right)^2 n^2} \right]^{1/(2|\mathbf{m}|+d+4)} . \tag{3.4}$$

In this case, the minimised AMSE is of order $n^{-8/(2|\mathbf{m}|+d+4)}$; here, we have had to resort to the usual kind of bias/variance tradeoff and cannot attain the slightly better performance achieved by eliminating the leading bias term.

4. Plug-in Bandwidth Selection Strategies

We showed in the previous section that the formulas for the AMSE-optimal bandwidth for estimation of $\psi_{\mathbf{m}}$ depend on f only through other $\psi_{\mathbf{m}}$ functionals. Therefore, one can obtain kernel estimates for those functionals and plug them into (3.3) and (3.4). This process of plugging in kernel estimates of $\psi_{\mathbf{m}}$ functionals can be repeated successively for higher order stages, but eventually a simpler approximation to these functionals will have to be used. In univariate plug-in strategies considerable success has been obtained by replacing f by the $N(0, \sigma^2)$ density and estimating σ^2 by the sample variance or interquartile range (Park and Marron 1990, Sheather and Jones 1991). However, Janssen, Marron, Veraverbeke and Sarle (1992) show that there are some simple situations where common scale estimators such as these are inappropriate and suggest a more general class of scale estimators which remedies this problem.

In the multivariate setting one can, at an appropriate stage, estimate $\psi_{\mathbf{m}}$ in the same fashion by replacing f with the $N(\mathbf{0}, \mathbf{\Sigma})$ density and then choosing an estimator for $\mathbf{\Sigma}$. Let $\psi_{\mathbf{m}}^N(\mathbf{\Sigma})$ be the value of $\psi_{\mathbf{m}}$ when f is multivariate normal with covariance matrix $\mathbf{\Sigma}$. For general d it is difficult to give succinct expressions for $\psi_{\mathbf{m}}^N(\mathbf{\Sigma})$ so we will restrict attention to the bivariate case. Denote the entries of $\mathbf{\Sigma}$ by

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} .$$

Then using Parseval's Identity, results from the appendix of Wand and Jones (1993), and Kendall and Stuart (1966, p.91) we can show that for $\mathbf{m} = (m_1, m_2)$, such that $|\mathbf{m}|$ is even,

$$\psi_{\mathbf{m}}^N(\Sigma) = \frac{(-1)^{|\mathbf{m}|/2}\lambda_{m_1 m_2}}{2^{(|\mathbf{m}|+4)/2}\pi\sigma_1^{m_1+1}\sigma_2^{m_2+1}(1-\rho^2)^{(|\mathbf{m}|+1)/2}}. \tag{4.1}$$

Here

$$\lambda_{2r,2s} = \frac{(2r)!(2s)!}{2^{r+s}} \sum_{j=0}^{\min(r,s)} \frac{(2\rho)^{2j}}{(r-j)!(s-j)!(2j)!},$$

$$\lambda_{2r+1,2s+1} = \frac{\rho(2r+1)!(2s+1)!}{2^{r+s}} \sum_{j=0}^{\min(r,s)} \frac{(2\rho)^{2j}}{(r-j)!(s-j)!(2j+1)!}$$

and $\lambda_{2r,2s+1} = \lambda_{2r+1,2s} = 0$. Formula (4.1) is also useful for the evaluation of the expressions involving K in the asymptotics when K is the standard d-variate normal density. Specifically, if $K = \phi_{\mathbf{I}}$, then

$$K^{(\mathbf{m})}(\mathbf{0}) = (-1)^{|\mathbf{m}|}\psi_{\mathbf{m}}^N(\tfrac{1}{2}\mathbf{I}) \quad \text{and} \quad R(K^{(\mathbf{m})}) = (-1)^{|\mathbf{m}|}\psi_{2\mathbf{m}}^N(\mathbf{I}).$$

Given the different levels of sophistication for the matrix $\mathbf{H}$ used to estimate f and the matrices $\mathbf{A}$ used to estimate the $\psi_{\mathbf{m}}$, combined with the number of stages k of kernel functional estimation before using a scale approximation, there are numerous strategies for plug-in selection of $\mathbf{H}$. Each of these strategies can be expressed using the notation

$$\mathcal{A} - (\mathcal{B}_1, \ldots, \mathcal{B}_k) - \mathcal{C} \tag{4.2}$$

where $\mathcal{A}$ is the class to which $\mathbf{H}$ is restricted for estimation of f, $\mathcal{B}_i$ is the class to which $\mathbf{A}$ is restricted at the ith stage of functional estimation and $\mathcal{C}$ denotes the type of scale approximation of the $\psi_{\mathbf{m}}$. The strategy where $\psi_{\mathbf{m}}$ is replaced by $\psi_{\mathbf{m}}^N(\mathbf{S})$, where $\mathbf{S}$ is the sample covariance matrix, will be denoted by $\mathcal{N}_{\mathbf{S}}$. This is the only kind of multivariate scale estimation we have yet utilised. For an example of the notation at (4.2), suppose that we plan to estimate f using the full $\mathbf{H}$ matrix, but will estimate the $\psi_{\mathbf{m}}$ with single parameter $\mathbf{A}$ matrices. The normal scale rule, using $\mathbf{S}$, is plugged in after one stage of kernel functional estimation. Then this strategy is denoted by $\mathcal{F} - (\mathcal{S}, \mathcal{S}) - \mathcal{N}_{\mathbf{S}}$.

TABLE 1

Number of $\psi_{\mathbf{m}}$ kernel estimators required for given strategy

Dimension (d)	2	3	4	5
Strategy				
$\mathcal{D} - (\mathcal{S}) - \mathcal{N}_{\mathbf{S}}$	3	6	10	15
$\mathcal{D} - (\mathcal{S}, \mathcal{S}) - \mathcal{N}_{\mathbf{S}}$	7	16	30	50
$\mathcal{F} - (\mathcal{S}) - \mathcal{N}_{\mathbf{S}}$	5	15	36	70
$\mathcal{F} - (\mathcal{S}, \mathcal{S}) - \mathcal{N}_{\mathbf{S}}$	13	44	120	281

A cost of more sophisticated strategies is the number of kernel functional estimations of the type (3.4) which need to be performed. Let $N^{\mathcal{A}}(d, k)$ denote the number of kernel

estimations required for the k level strategy $\mathcal{A} - (\mathcal{S}, \ldots, \mathcal{S}) - \mathcal{N}_S$ for d dimensions. Then it can be shown that

$$N^{\mathcal{D}}(d,k) = \sum_{r=1}^{k} \sum_{\ell=0}^{\min(r,d-1)} \binom{r}{\ell}\binom{d}{\ell+1}$$

and

$$N^{\mathcal{F}}(d,k) = \nu_k + \sum_{r=1}^{k} \sum_{\ell=0}^{\min(2r+1,d-1)} \binom{2r+1}{\ell}\binom{d}{\ell+1}.$$

where $\nu_1 = 0$, $\nu_2 = 1$, $\nu_3 = 3$ and for $k = 4, 5, \ldots$,

$$\nu_k = \sum_{r=1}^{k-3} \sum_{\ell=0}^{\min(r,d-1)} \binom{r}{\ell}\binom{d}{\ell+1}.$$

For $d = 2, 3, 4$ and 5 Table 1 lists the number of these estimations that are required for certain strategies. Table 1 shows that, for higher dimensions, the number of functional estimations can become ridiculously high. However, for the more practical lower dimensions these numbers are fairly reasonable, especially for $\mathbf{H} \in \mathcal{D}$.

5. Theoretical Performance

The explicit nature of the bivariate, $\mathbf{H} \in \mathcal{D}$, $\mathbf{A} = a^2\mathbf{I}$, algorithms allow a straightforward analysis of their properties. Writing $\hat{h}_{1,\mathrm{AMISE}}$ for $h_{1,\mathrm{AMISE}}$ given at the end of Section 2 with all $\psi_{\mathbf{m}}$'s replaced by $\hat{\psi}_{\mathbf{m}}(\mathbf{A})$'s as in (3.1), we have that

$$\hat{h}_{1,\mathrm{AMISE}} = h_{1,\mathrm{AMISE}}(\psi_{40}^{1/2}\psi_{04}^{1/2} + \psi_{22})^{1/6} \left(1 + \frac{\hat{\psi}_{04} - \psi_{04}}{\psi_{04}}\right)^{1/8} \left(1 + \frac{\hat{\psi}_{40} - \psi_{40}}{\psi_{40}}\right)^{-1/8}$$
$$\times \left\{\psi_{40}^{1/2}\psi_{04}^{1/2}\left(1 + \frac{\hat{\psi}_{40} - \psi_{40}}{\psi_{40}}\right)^{1/2}\left(1 + \frac{\hat{\psi}_{04} - \psi_{04}}{\psi_{04}}\right)^{1/2} + \psi_{22}\left(1 + \frac{\hat{\psi}_{22} - \psi_{22}}{\psi_{22}}\right)\right\}^{-1/6}.$$

Taylor series expansion then affords

$$(\hat{h}_{1,\mathrm{AMISE}} - h_{1,\mathrm{AMISE}})/h_{1,\mathrm{AMISE}} \simeq \{24(\psi_{40}^{1/2}\psi_{04}^{1/2} + \psi_{22})\}^{-1}\left\{\left(\psi_{40}^{1/2}\psi_{04}^{1/2} + 3\psi_{22}\right)\right.$$
$$\left.\times\left(\frac{\hat{\psi}_{04} - \psi_{04}}{\psi_{04}}\right) - \left(5\psi_{40}^{1/2}\psi_{04}^{1/2} + 3\psi_{22}\right)\left(\frac{\hat{\psi}_{40} - \psi_{40}}{\psi_{40}}\right) - 2\psi_{22}\left(\frac{\hat{\psi}_{22} - \psi_{22}}{\psi_{22}}\right)\right\}.$$

While this expression can be used as a basis for a full mean squared error analysis of $\hat{h}_{1,\mathrm{AMISE}}$ including explicit constants, the resulting formulae are unedifying, and we shall limit ourselves to the rate of convergence only. Using the rates given in Section 3 (especially that connected with use of (3.3)) it is now easy to see that

$$\{(\hat{h}_{1,\mathrm{AMISE}} - h_{1,\mathrm{AMISE}})/h_{1,\mathrm{AMISE}}\}^2 = O_p(n^{-3/4})$$

since $d = 2$ and $|\mathbf{m}| = 4$ for all $\hat{\psi}_{\mathbf{m}}$'s of interest here.

The above is the only practically useful case for which explicit formulae for estimated bandwidths are readily available. However, more can be said about rates of convergence. First, we will look at another explicit situation, that of $\mathbf{H} \in \mathcal{S}$. Again, it is easy to see how $(\hat{h}_{\mathrm{AMISE}} - h_{\mathrm{AMISE}})/h_{\mathrm{AMISE}}$ relates to estimation of ψ_{40}, ψ_{04} and ψ_{22} (using formulae in Wand and Jones 1993, for example). Jones (1992) argues that this results in a mean squared relative error of order $n^{-\min\{8,(d+4)\}/(d+6)}$ which corresponds, again, to the rate given just below (3.3). However, a qualification of the above convergence rates is required. One should note that with regard to h_{MISE} rather than h_{AMISE}, the relationship

$$(h_{\mathrm{AMISE}} - h_{\mathrm{MISE}})/h_{\mathrm{MISE}} = O(n^{-2/(d+4)})$$

holds. Therefore, while $(\hat{h}_{\mathrm{AMISE}} - h_{\mathrm{MISE}})/h_{\mathrm{MISE}}$ remains of order $n^{-5/14}$ when $d = 1$, for $d \geq 2$, the rate $n^{-2/(d+4)}$ pertains. This still compares favourably with the relative error rate of convergence of least squares cross-validation, which is $n^{-d/2(d+4)}$ (Marron, 1986), for $d \leq 3$, is equal to it when $d = 4$, and suffers by comparison thereafter. The rates given by Jones (1992), and the $n^{-3/8}$ rate above in the bivariate case, are correct only for comparison of $\hat{h}_{\mathrm{AMISE}}$ with h_{AMISE} rather than with h_{MISE}. Those rates can, however, be reinstated if we modify our algorithm to additionally estimate the next term in the asymptotic expansion of h_{AMISE}, a method utilised in the univariate case by Hall, Sheather, Jones and Marron (1991), but this is not pursued further here.

The rates considered here continue to hold in the more general $\mathbf{H} \in \mathcal{D}$ and $\mathbf{H} \in \mathcal{F}$ cases. The following heuristic argument indicates how this can be established in the most general case. Write

$$Q(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}R(K) + \tfrac{1}{4}\mu_2(K)^2(\mathrm{vech}\,\mathbf{H})^T\hat{\mathbf{\Psi}}_{\mathcal{F}}(\mathrm{vech}\,\mathbf{H})$$

where $\hat{\mathbf{\Psi}}_{\mathcal{F}}$ is obtained from $\mathbf{\Psi}_{\mathcal{F}}$ by replacing each $\psi_{\mathbf{m}}$ by $\hat{\psi}_{\mathbf{m}}(\mathbf{A})$. Also let

$$\nabla Q(\mathbf{H}) = \partial Q(\mathbf{H})/(\partial \mathrm{vech}\ \mathbf{H}) \quad \text{and} \quad \nabla^2 Q(\mathbf{H}) = \partial^2 Q(\mathbf{H})/(\partial \mathrm{vech}\ \mathbf{H})(\partial \mathrm{vech}\ \mathbf{H})^T.$$

Formal expansion of $\nabla Q(\mathbf{H})$ leads to

$$\nabla Q(\mathbf{H}) \simeq \nabla Q(\mathbf{H}_{\mathrm{AMISE}}) + \{\nabla^2 Q(\mathbf{H}_{\mathrm{AMISE}})\}\mathrm{vech}\,(\mathbf{H} - \mathbf{H}_{\mathrm{AMISE}}).$$

Noting that the plug-in bandwidth selector $\hat{\mathbf{H}}$ satisfies $\nabla Q(\hat{\mathbf{H}}) = \mathbf{0}$ we obtain

$$\mathrm{vech}\,(\hat{\mathbf{H}} - \mathbf{H}_{\mathrm{AMISE}}) \simeq -\{\nabla^2 Q(\mathbf{H}_{\mathrm{AMISE}})\}^{-1}\nabla Q(\mathbf{H}_{\mathrm{AMISE}}).$$

From (3.3) of Wand (1992),

$$\nabla Q(\mathbf{H}_{\mathrm{AMISE}}) = \tfrac{1}{2}\mu_2(K)^2(\hat{\Psi}_{\mathcal{F}} - \Psi_{\mathcal{F}})(\mathrm{vech}\ \mathbf{H}_{\mathrm{AMISE}}).$$

Also, assuming that $\mathbf{H} = \mathbf{C}n^{-2/(d+4)}$ for a constant positive definite matrix $\mathbf{C}$, it follows from (3.4) of Wand (1992) that $\nabla^2 Q(\mathbf{H}_{\mathrm{AMISE}})$ converges to a constant non-singular matrix. Therefore,

$$\mathrm{vech}\,(\hat{\mathbf{H}} - \mathbf{H}_{\mathrm{AMISE}}) \simeq \mathbf{\Omega}(\hat{\Psi}_{\mathcal{F}} - \Psi_{\mathcal{F}})(\mathrm{vech}\ \mathbf{H}_{\mathrm{AMISE}})$$

for a constant matrix $\mathbf{\Omega}$, which indicates explicitly that the asymptotic performance of $\hat{\mathbf{H}}$ is driven by the goodness of $\hat{\Psi}_{\mathcal{F}}$ as an estimate of $\Psi_{\mathcal{F}}$.

6. Practical Performance

We conducted a small simulation study to test the efficacy of the $\mathcal{D} - (\mathcal{S}, \mathcal{S}) - \mathcal{N}_{\mathbf{S}}$ strategy for bivariate samples. The bivariate Gaussian kernel was used for both functional estimation and density estimation. Because a single bandwidth is used for functional estimation the data were rescaled to have the same sample variance in each direction before applying the bandwidth selection rule, and the selected bandwidths transformed back to the scale of the original data.

Direct computation of kernel functional estimates can be very computationally expensive for larger values of n, so all functional estimates were computed using a fast linear binned approximation over $[-3, 3]^2$ with an 80×80 grid. See Section 4 of Wand (1993) for a description of this approach.

We simulated 100 samples of size $n = 1000$ from each of five normal mixture densities, corresponding to densities (A), (E), (F), (H) and (K) in Wand and Jones (1993). See Table 1 of that article for specification of their parameters. Figures 1–5 show (a) a contour plot of the true density; (b) a scatterplot of

$$(\log_{10}(\hat{h}_1/h_{1,\mathrm{MISE}}), \log_{10}(\hat{h}_2/h_{2,\mathrm{MISE}}))$$

values, where $(\hat{h}_1, \hat{h}_2)^T$ is the selected bandwidth pair and $\mathbf{h}_{\mathrm{MISE}} = (h_{1,\mathrm{MISE}}, h_{2,\mathrm{MISE}})^T$ is the finite sample optimal bandwidth when $\mathbf{H} \in \mathcal{D}$; (c) the "median performance" sample and (d) the corresponding density estimate based on $(\hat{h}_1, \hat{h}_2)^T$ for that sample. For each density, the median performance density estimate was chosen to be the one that had the 50th lowest approximate integrated squared error out of the 100 replications, and is plotted to give an indication of "average case" behaviour of the bandwidth selection strategy. Values of $\mathbf{h}_{\mathrm{MISE}}$ were computed by numerical minimisation of the MISE expression for normal mixture densities given by Theorem 1 of Wand and Jones (1993). The contours in each contour plot are at increments of one tenth of the maximum height of the true density.

The distributions of the selected bandwidths are each quite tight, reflecting the low variability of plug-in rules. However, there is a slight bias towards oversmoothing, espe-

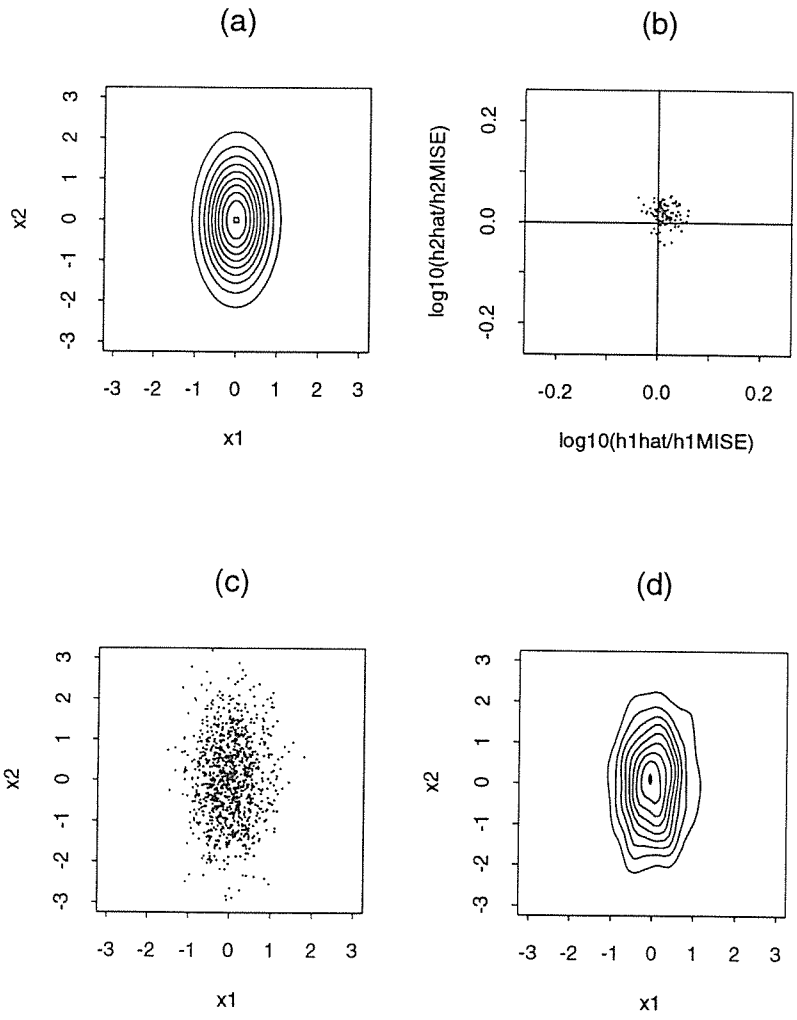

Figure 1. Simulation results for Density (A). A description of each plot is given in the text.

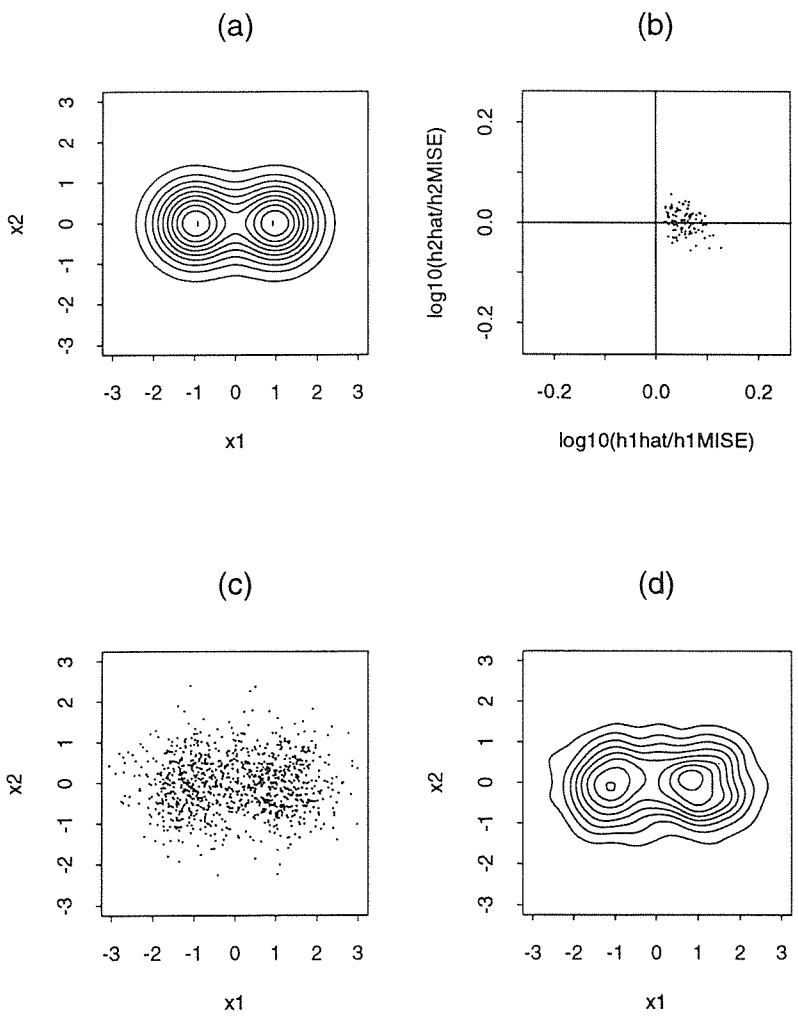

Figure 2. Simulation results for Density (E). A description of each plot is given in the text.

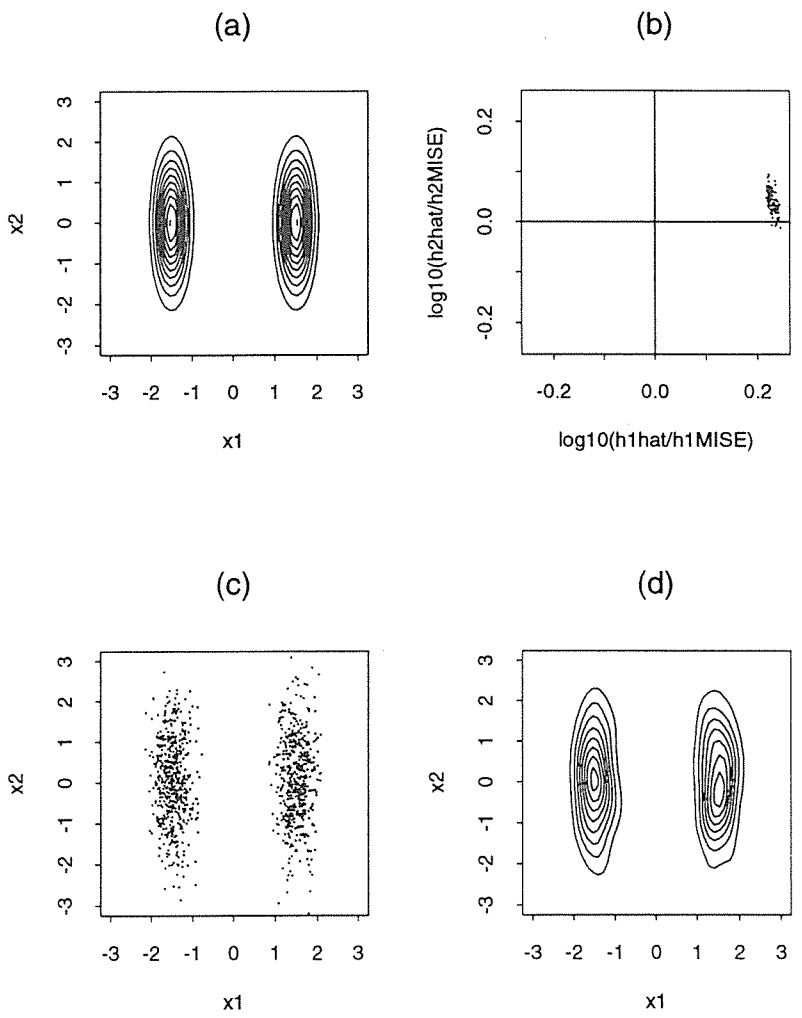

Figure 3. Simulation results for Density (F). A description of each plot is given in the text.

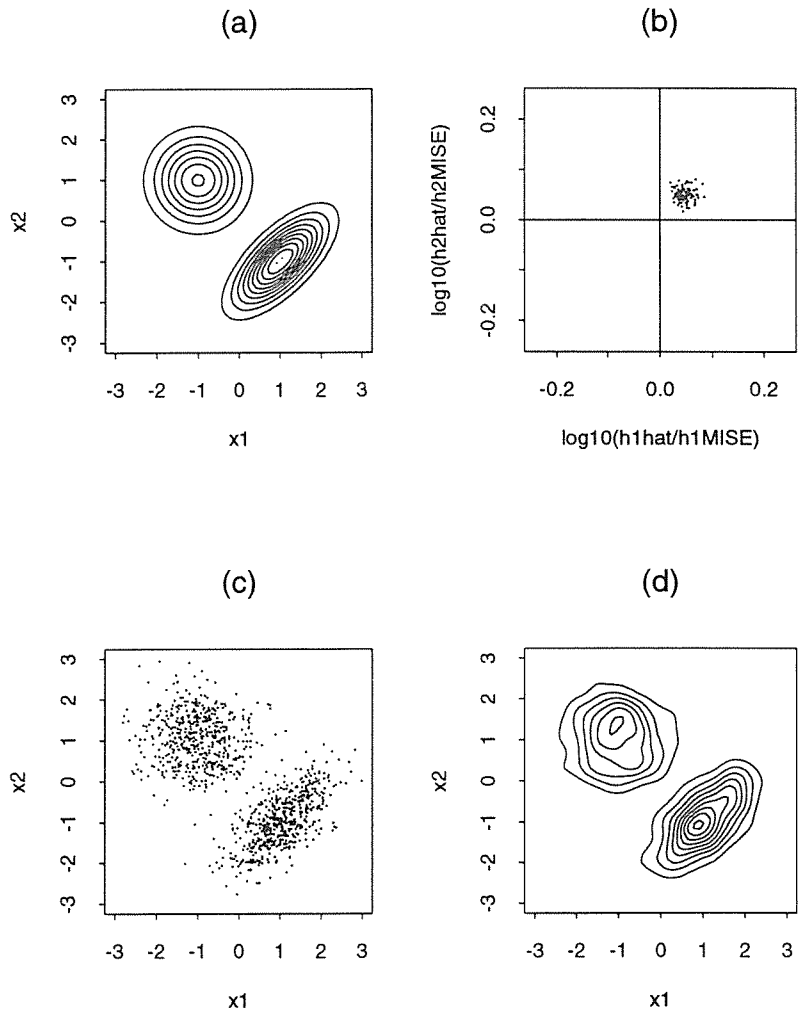

Figure 4. Simulation results for Density (H). A description of each plot is given in the text.

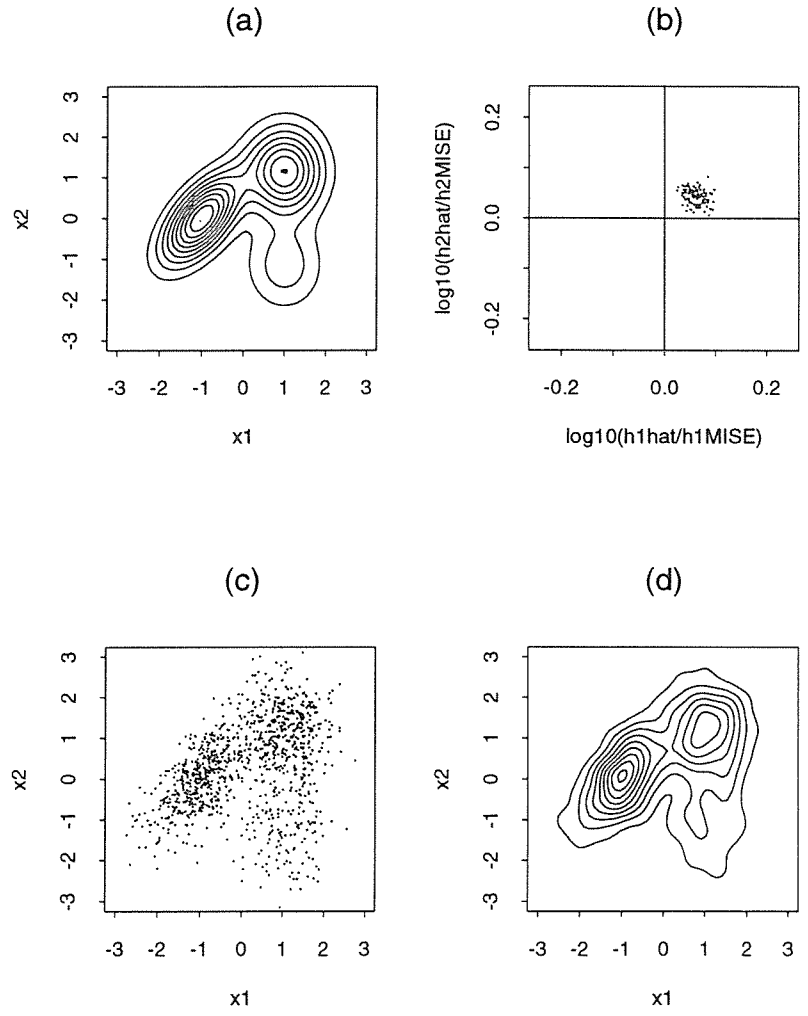

Figure 5. Simulation results for Density (K). A description of each plot is given in the text.

cially for density (F). In this case the selected h_2 bandwidths tend to be about 65% larger than $h_{2,\text{MISE}}$, reflecting the fact that the normal scale rule based on the sample covariance matrix is inappropriate for densities with widely separated modes. See Janssen, Marron, Veraverbeke and Sarle (1992) for a discussion of and solution to this problem in the univariate context.

We also applied the same selector to the plasma lipid data, a bivariate data set analysed by Scott, Gotto, Cole and Gorry (1978) and Silverman (1986 pp.81–83). These data consist of pairs of observed plasma lipid concentrations of cholesterol and triglyceride, for 320 diseased patients. A scatterplot of the data is shown in Figure 6a. In Figure 6b a contour plot of the density estimate based on the $\mathcal{D} - (\mathcal{S}, \mathcal{S}) - \mathcal{N}_S$ selection strategy is shown. The bimodality, which was used to divide the patients into two distinct groups by Scott *et al.* 1978, is also present in this automatically generated density estimate.

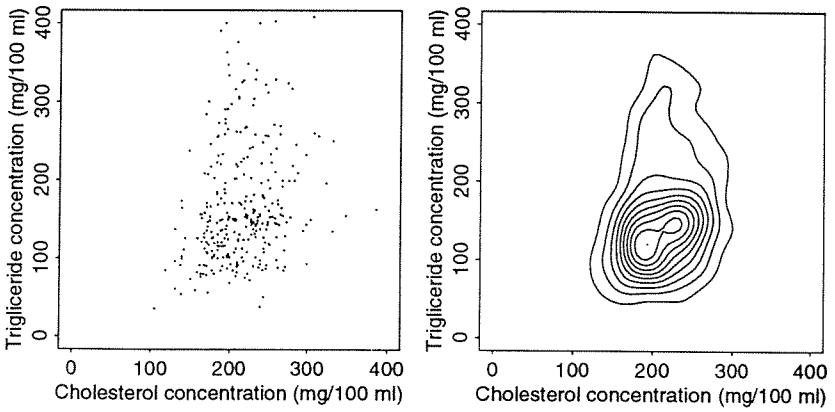

Figure 6. (a) *Scatterplot of the plasma lipid data ($n = 320$). (b) Gaussian kernel density estimate of the data using the $\mathcal{D} - (\mathcal{S}, \mathcal{S}) - \mathcal{N}_S$ strategy.*

Interestingly, Silverman (1986) uses this data set to demonstrate how a "density estimate will detect or highlight features that are not at all obvious in the scatterplot" while Terrell (1990) oversmooths and obtains a "plausible conservative" unimodal estimate.

7. Conclusion

The multivariate plug-in bandwidth selector developed here shows considerable promise. Both the theory given in Section 5 and the small simulation study given in Section 6 suggest that it is more stable and reliable than some previously proposed multivariate bandwidth selection procedures. Further assessment of its practical performance and comparison with cross-validatory selectors, including that of Sain *et al*, 1992, through a

large scale simulation study would be an important future project. Another subject worth investigating is the fine tuning of the plug-in strategies in an attempt to lessen their dependence on normal scale rules. This could involve ideas presented in the univariate context by Janssen *et al* (1992) and Scott (1992b).

Acknowledgments

We are grateful to Dr Keith Baggerly for his helpful discussions and to Professor David Scott for providing us with the plasma lipid data.

References

Chiu, S.-T. (1991). Bandwidth selection for kernel density estimation. *Ann. Statist.* **19**, 1883–1905.

Chiu, S.-T. (1992). An automatic bandwidth selector for kernel density estimation. *Biometrika*, **79**, 771–782.

Hall, P. and Johnstone, I. (1992). Empirical functionals and efficient smoothing parameter selection (with discussion). *J. Roy. Statist. Soc. Ser. B*, **54**, 475–530.

Hall, P. and Marron, J. S. (1987a). Extent to which least-squares cross-validation minimises integrated squared error in nonparametric density estimation. *Prob. Theory Rel. Fields*, **74**, 567 - 581.

Hall, P. and Marron, J. S. (1987b). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.*, **6**, 109–114.

Hall, P. and Marron, J. S. (1991). Lower bounds for bandwidth selection in density estimation. *Probab. Theory Rel. Fields*, **92**, 149–173.

Hall, P., Marron, J. S. and Park, B. U. (1992). Smoothed cross-validation. *Probab. Theory Rel. Fields*, **92**, 1–20.

Hall, P., Sheather, S. J., Jones, M. C. and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, **78**, 263–269.

Janssen, P., Marron, J. S., Veraverbeke, N. and Sarle, W. (1992). Scale measures for bandwidth selection. Unpublished manuscript.

Jones, M. C. (1992). Potential for automatic bandwidth choice in variations on kernel density estimation. *Statist. Probab. Lett.*, **13**, 351–356.

Jones, M. C., Marron, J. S. and Park, B. U. (1991) A simple root n bandwidth selector. *Ann. Statist.*, **19**, 1919–1932.

Jones, M.C., Marron, J. S., and Sheather, S. J. (1992). Progress in data-based bandwidth selection for kernel density estimation. *University of New South Wales, Australian Graduate School of Management, Working Paper Series 92-014.*

Jones, M. C. and 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.

Kendall, M. G. and Stuart, A. (1966). *The Advanced Theory of Statistics*, Vol I, Hafner Publishing Co., New York.

Marron, J. S. (1986). Will the art of smoothing ever become a science? *Contemp. Math.*, **59**, 169–178.

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

Sain, S. R., Baggerly, K. A. and Scott, D. W. (1992). Cross-validation of multivariate densities. *Department of Statistics, Rice University, Technical Report 92-10.*

Scott, D. W. (1992a). *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: Wiley.

Scott, D. W. (1992b). Constrained oversmoothing and upper bounds on smoothing parameters in regression and density estimation. *Department of Statistics, Rice University, Technical Report 92-8.*

Scott, D. W., Gotto, A. M., Cole, J. S. and Gorry, G. A. (1978). Plasma lipids as collateral risk factors in coronary heart disease – a study of 371 males with chest pain. *J. Chronic Diseases*, **31**, 337–345.

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

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

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, **12**, 1285 – 97.

Terrell, G. R. (1990). The maximal smoothing principle in density estimation. *J. Amer. Statist. Assoc.*, **85**, 470 – 477.

Wand, M. P. (1992). Error analysis for general multivariate kernel estimators. *J. Nonpar. Statist.*, **2**, 1-15.

Wand, M. P. (1993). Fast computation of multivariate kernel estimators. *University of New South Wales, Australian Graduate School of Management, Working Paper Series 93-007.*

Wand, M. P. and Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.*, **88**, 520–528.