Nonparametric Statistics, 2003, Vol. 15(1), pp. 17–30

Taylor & Francis
Taylor & Francis Group

PLUG-IN BANDWIDTH MATRICES FOR BIVARIATE KERNEL DENSITY ESTIMATION

TARN DUONG* and MARTIN L. HAZELTON

School of Mathematics and Statistics, University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia

(Received April 2002; In final form August 2002)

We consider bandwidth matrix selection for bivariate kernel density estimators. The majority of work in this area has been directed towards selection of diagonal bandwidth matrices, but full bandwidth matrices can give markedly better performance for some types of target density. Our methodological contribution has been to develop a new version of the plug-in selector for full bandwidth matrices. Our approach has the advantage, in comparison to existing full bandwidth matrix plug-in techniques, that it will always produce a finite bandwidth matrix. Furthermore, it requires computation of significantly fewer pilot bandwidths. Numerical studies indicate that the performance of our bandwidth selector is best when implemented with two pilot estimation stages and applied to sphered data. In this case our methodology performs at least as well as any competing method considered, while being simpler to implement than its competitors.

Keywords: Asymptotic; MISE; Nonparametric smoothing; Pilot estimation; Positive-definite

1 INTRODUCTION

Kernel density estimation has become a popular tool for visualizing the distribution of univariate data [see Ref. 5, for example, for an overview]. Univariate kernel density estimation has received considerable attention in the literature, partly because of its practical utility, and partly because it provides a simple testing ground for learning about nonparametric smoothing. Kernel density estimation for multivariate data has received significantly less attention. The lower level of interest in the multivariate context may be explained, to some extent, by the difficulties in viewing high dimensional density functions. Scott [4] described a variety of techniques for visualizing such multivariate functions, but while many of these visualization devices are ingenious, interpretation of the resulting types of plot requires significant experience. For this reason, if no other, the use of high-dimensional density estimation as a tool for exploratory data analysis appears relatively uncommon amongst practitioners.

Bivariate kernel density estimation sits at an important junction between the univariate and high-dimensional multivariate cases. From a practical standpoint, bivariate density estimates have a utility and accessibility that is akin to that of their univariate cousins, largely because

* Corresponding author.

ISSN 1048-5252 print; ISSN 1029-0311 online
DOI: 10.1080/1048525021000049711

they can be viewed using familiar perspective ('wire-frame') or contour plots. From a theoretical viewpoint, bivariate density estimation is an excellent setting for understanding aspects of multivariate kernel smoothing. There are important aspects of bivariate kernel estimation that have no univariate analogue (such as the orientation of the kernel functions) yet can be generalized to higher dimensional cases with relatively little effort.

For a bivariate sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ the kernel density estimate is defined by

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

where $\boldsymbol{x} = (x_1, x_2)^T$ and $\boldsymbol{X}_i = (X_{i1}, X_{i2})^T$, $i = 1, 2, \ldots, n$. Here $K(\boldsymbol{x})$ is the bivariate kernel (which we assume to be a probability density function); $\mathbf{H}$ is the bandwidth matrix which is symmetric and positive-definite; and $K_{\mathbf{H}}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\boldsymbol{x})$. The choice of $\mathbf{H}$ is crucially important in determining the performance of $\hat{f}$. Bivariate bandwidth selection is a difficult problem which may be simplified (at the expense of flexibility) by imposing constraints on $\mathbf{H}$. For example, $\mathbf{H}$ may be restricted to the class of diagonal (positive-definite) matrices, or to the class of (positive) multiples of the identity matrix. The merits of imposing restrictions on $\mathbf{H}$ have been investigated by Wand and Jones [6]. These authors conclude that choosing a diagonal bandwidth matrix will sometimes be adequate, but that in other cases there is much to be gained by selecting a full (*i.e.*, unconstrained) bandwidth matrix. While the use of a full bandwidth matrix requires an additional smoothing parameter (in comparison to diagonal $\mathbf{H}$), it permits arbitrary orientation of the kernel function. This orientation could be chosen in an automatic fashion by constraining $\mathbf{H}$ to be a (positive) multiple of the sample correlation matrix, but Wand and Jones demonstrated that this is inappropriate, in general.

A large body of published work now exists on bandwidth selection for univariate kernel density estimation. See Ref. [2], for example, for a review. Cross-validation, bootstrap and plug-in methods have proved popular in this context, and all of these technologies have been transferred into the bivariate (and more generally multivariate) setting. Sain *et al.* [3] considered cross-validation and bootstrap methods for bandwidth selection for multivariate density estimators. However, these authors restricted their attention to estimators constructed using product kernels (which is essentially equivalent to using a diagonal $\mathbf{H}$). Wand and Jones [7] looked at plug-in bandwidth selection. These authors showed that it is impossible to derive an explicit expression for the plug-in estimator of $\mathbf{H}$ for general multivariate kernel density estimators. Wand and Jones therefore concentrated most of their efforts on diagonal bandwidth matrices for bivariate density estimation, since explicit plug-in estimates are available in this context, and hence analysis is more straightforward that in the general case.

This paper is concerned with plug-in methods for selecting a full bandwidth matrix for bivariate kernel density estimation. This is a problem of some significance in light of the conclusions of Wand and Jones [6]. We operate within the general framework developed in Ref. [7], but our aim is to adapt the work of these authors to improve practical performance. Our principal methodological contribution concerns the pilot estimation of a matrix of functionals of the target density, crucial to the calculation of the plug-in estimates. Wand and Jones [7] suggested the method be calibrated so as to optimize estimation of this matrix of functionals on an element by element basis. We note, however, that this approach can result in an estimate that lacks the positive-definiteness of the target matrix. Even if the estimated matrix has this property it may be almost singular, which can lead to unstable bandwidth selection. We prefer to optimize the estimates of all elements of the matrix

estimate using a single, common tuning parameter. Following this approach allows us to ensure that the matrix estimate will be positive definite, and also lightens the computational burden of implementing the plug-in bandwidth matrix selector.

The remainder of the paper is structured as follows. In the next section we discuss optimization of bandwidth matrices with respect to the mean integrated squared error (MISE) of $\hat{f}$. Following [7] we give an asymptotic version of the MISE, and outline how this can be estimated using plug-in techniques. In Section 3 we turn our attention to pilot estimation in the plug-in method, and describe a new method for selecting the pilot smoothing parameters. We incorporate this novel methodology into practical algorithms for plug-in bandwidth matrices in Section 4. The practical performance of our methodology is compared with existing plug-in techniques (including those for diagonal $\mathbf{H}$) via numerical studies in Section 5. The results are encouraging. In Section 6 we draw together the findings in the paper, and suggest some avenues for further research.

2 OPTIMAL BANDWIDTH MATRICES

In order to measure the performance of $\hat{f}$ we shall (in common with the great majority of researchers in this field) use the mean integrated squared error (MISE) criterion,

$$\operatorname{MISE} \hat{f}(\cdot; \mathbf{H}) = \mathbb{E}\operatorname{ISE} \hat{f}(\cdot; \mathbf{H}) = \mathbb{E} \int_{\mathbb{R}^2} [\hat{f}(\boldsymbol{x}; \mathbf{H}) - f(\boldsymbol{x})]^2 \, \mathrm{d}\boldsymbol{x}.$$

Here f denotes the target density, from which $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are henceforth assumed to be a random sample. Our aim in bandwidth selection is to estimate

$$\mathbf{H}_{\operatorname{MISE}} = \operatorname*{arg\,min}_{\mathbf{H} \in \mathcal{H}} \operatorname{MISE} \hat{f}(\cdot; \mathbf{H}),$$

where $\mathcal{H}$ is the space of all symmetric, positive definite 2×2 matrices. It is well known that the optimal bandwidth $\mathbf{H}_{\operatorname{MISE}}$ does not have a closed form. In order to make progress it is usual to employ an asymptotic analysis. It can be shown (see [8, Chapter 4] for instance) that (under conditions to be specified)

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

where

$$\operatorname{AMISE} \hat{f}(\cdot\,; \mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2} R(K) + \frac{1}{4}\mu_2(K)^2 (\operatorname{vech}^T \mathbf{H}) \boldsymbol{\Psi}_4 (\operatorname{vech} \mathbf{H}) \tag{2}$$

where $R(K) = \int_{\mathbb{R}^2} K(\boldsymbol{x})^2 \, \mathrm{d}\boldsymbol{x}$, $\mu_2(K)\mathbf{I} = \int_{\mathbb{R}^2} \boldsymbol{x}\boldsymbol{x}^T K(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$ with $\mu_2(K) < \infty$ and vech is the vector half operator (see [8, chapter 4]). The $\boldsymbol{\Psi}_4$ matrix is the 3×3 matrix given by

$$\boldsymbol{\Psi}_4 = \int_{\mathbb{R}^2} \operatorname{vech}[2D^2 f(\boldsymbol{x}) - \operatorname{dg} D^2 f(\boldsymbol{x})] \operatorname{vech}^T [2D^2 f(\boldsymbol{x}) - \operatorname{dg} D^2 f(\boldsymbol{x})] \, \mathrm{d}\boldsymbol{x}$$

where $D^2 f(\boldsymbol{x})$ is the Hessian matrix of f and $\operatorname{dg} \mathbf{A}$ is matrix $\mathbf{A}$ with all of its non-diagonal elements set to zero. Sufficient conditions for the validity of the expansions defined by

Eqs. (1) and (2) are that all entries in $D^2 f(\boldsymbol{x})$ are square integrable and all entries of $\mathbf{H} \to 0$ and $n^{-1}|\mathbf{H}|^{-1/2} \to 0$, as $n \to \infty$.

If we introduce some more notation, we can explicitly state an expression for the matrix $\boldsymbol{\Psi}_4$ in terms of its individual elements. Let $\mathbf{r} = (r_1, r_2)$ where the r_1, r_2 are non-negative integers. Let $|\mathbf{r}| = r_1 + r_2$, then the $\mathbf{r}$th partial derivative of f can be written as

$$f^{(\mathbf{r})}(\boldsymbol{x}) = \frac{\partial^{|\mathbf{r}|}}{\partial_{x_1}^{r_1} \partial_{x_2}^{r_2}} f(\boldsymbol{x})$$

and the integrated density derivative functional is

$$\psi_{\mathbf{r}} = \int_{\mathbb{R}^2} f^{(\mathbf{r})}(\boldsymbol{x}) f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

Note that if $\boldsymbol{X}$ has density f then $\mathbb{E} f^{(\mathbf{r})}(\boldsymbol{X}) = \psi_{\mathbf{r}}$. Also,

$$\int_{\mathbb{R}^2} f^{(\mathbf{r})}(\boldsymbol{x}) f^{(\mathbf{s})}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = (-1)^{|\mathbf{r}|} \int_{\mathbb{R}^2} f^{(\mathbf{r}+\mathbf{s})}(\boldsymbol{x}) f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

This then implies that

$$\boldsymbol{\Psi}_4 = \begin{bmatrix} \psi_{40} & 2\psi_{31} & \psi_{22} \\ 2\psi_{31} & 4\psi_{22} & 2\psi_{13} \\ \psi_{22} & 2\psi_{13} & \psi_{04} \end{bmatrix}. \tag{3}$$

(Note that the subscript 4 on $\boldsymbol{\Psi}$ relates to the order of the derivatives involved.)

Equations (2) and (3) combine to give a tractable approximation, AMISE, to the MISE. Plug-in methods of selecting the bandwidth matrix make use of the tractability of AMISE by seeking to estimate

$$\mathbf{H}_{\mathrm{AMISE}} = \operatorname*{arg\,min}_{\mathbf{H} \in \mathcal{H}} \mathrm{AMISE}\, \hat{f}(\cdot; \mathbf{H})$$

rather than $\mathbf{H}_{\mathrm{MISE}}$. Of course, the AMISE is a functional of the unknown target density, through $\boldsymbol{\Psi}_4$. Hence we require pilot estimates of the $\psi_{\mathbf{r}}$ functionals that can be 'plugged-in' to provide an estimate $\hat{\boldsymbol{\Psi}}_4$. This in turn produces an estimate $\widehat{\mathrm{AMISE}}$ that can be numerically minimized to give the plug-in bandwidth matrix, $\hat{\mathbf{H}}$. We note that this process is facilitated if $\mathbf{H}$ is assumed to be diagonal, since then $\hat{\mathbf{H}}$ can be written down in closed form (see Ref. 7). Nonetheless, it is far from clear that the simplifications obtained through using a diagonal bandwidth matrix warrant the loss of flexibility that this restriction of $\mathbf{H}$ entails.

3 PILOT FUNCTIONAL ESTIMATION

In order to implement plug-in selection of $\mathbf{H}$ we require pilot estimates of the integrated density derivative functionals, $\psi_{\mathbf{r}}$. If we note that $\psi_{\mathbf{r}} = \mathbb{E} f^{(\mathbf{r})}(\boldsymbol{X})$ then a natural estimator of $\psi_{\mathbf{r}}$ (following [7]) is

$$\hat{\psi}_{\mathbf{r}}(\mathbf{G}) = n^{-1} \sum_{i=1}^{n} \hat{f}^{(\mathbf{r})}(\boldsymbol{X}_i; \mathbf{G}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{\mathbf{G}}^{(\mathbf{r})}(\boldsymbol{X}_i - \boldsymbol{X}_j). \tag{4}$$

where $\mathbf{G}$ is a pilot bandwidth matrix (usually different to $\mathbf{H}$). (This is known as the leave-in-diagonals estimator as it includes the non-stochastic $i = j$ terms; cf. Ref. 1). An important aspect of this pilot estimation is the choice of $\mathbf{G}$. The finite sample properties of $\psi_{\mathbf{r}}$ as a function of $\mathbf{G}$ are intractable, but we can again make progress through asymptotic expansions. It can be shown that the bias of $\hat{\psi}_{\mathbf{r}}$ can be expressed as

$$\operatorname{Bias} \hat{\psi}_{\mathbf{r}}(\mathbf{G}) = n^{-1} K_{\mathbf{G}}^{(\mathbf{r})}(\mathbf{0}) + \frac{1}{2} \mu_2(K) \int_{\mathbb{R}^2} \operatorname{tr}[\mathbf{G} D^2 f(\boldsymbol{x})] f^{(\mathbf{r})}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} + o(\operatorname{tr} \mathbf{G})$$

while the variance can be expanded as

$$\begin{aligned} \operatorname{Var} \hat{\psi}_{\mathbf{r}}(\mathbf{G}) = {} & 2n^{-2} \psi_{\mathbf{0}} \int_{\mathbb{R}^2} K_{\mathbf{G}}^{(\mathbf{r})}(\boldsymbol{x})^2 \, \mathrm{d}\boldsymbol{x} + o(n^{-2} t(\mathbf{G})) \\ & + 4n^{-1} \left[\int_{\mathbb{R}^2} f^{(\mathbf{r})}(\boldsymbol{x})^2 f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \psi_{\mathbf{r}}^2 \right] + o(n^{-1}). \end{aligned}$$

for some smooth function t. However, the explicit form for t is rather complicated for a general $\mathbf{G}$; to find $K_{\mathbf{G}}^{(\mathbf{r})}$ requires many applications of the chain rule and the resulting expression quickly becomes unwieldy. We therefore follow the lead of Wand and Jones, and consider pilot bandwidth matrices of the form $\mathbf{G} = g^2 \mathbf{I}$ (where $\mathbf{I}$ is the 2×2 identity matrix). While this form of $\mathbf{G}$ may appear very restrictive and, moreover, inappropriate for many data sets, the deficiencies of this approach are perhaps less extreme than might appear at first sight. In the first place, the data may be pre-scaled so as to improve the applicability of this form of $\mathbf{G}$. We return to this matter in the next section. Secondly, this choice of $\mathbf{G}$ does not affect convergence rates for the estimates $\hat{\psi}_{\mathbf{r}}$.

Setting $\mathbf{G} = g^2 \mathbf{I}$ the bias and variance expressions simplify as follows:

$$\operatorname{Bias} \hat{\psi}_{\mathbf{r}}(g) = n^{-1} g^{-|\mathbf{r}|-2} K^{(\mathbf{r})}(\mathbf{0}) + \frac{1}{2} g^2 \mu_2(K) \sum_{i=1}^{2} \psi_{\mathbf{r}+2\mathbf{e}_i} + o(g^2)$$

where $\mathbf{e}_i$ is the ith elementary vector (*i.e.* a vector of length 2 with 1 in the ith position and 0 elsewhere) and

$$\operatorname{Var} \hat{\psi}_{\mathbf{r}}(g) = 2n^{-2} g^{-2|\mathbf{r}|-2} \psi_{\mathbf{0}} R(K^{(\mathbf{r})}) + o(n^{-2} g^{-2|\mathbf{r}|-2} + n^{-1})$$

provided that $K^{(\mathbf{r})}$ is square integrable and $g \to 0$ and $n^{-2}g^{-2|\mathbf{r}|-2} \to 0$ as $n \to \infty$. Combining these results gives an asymptotic from for the mean square error (MSE) of $\hat{\psi}_{\mathbf{r}}$:

$$\begin{aligned}\operatorname{AMSE} \hat{\psi}_{\mathbf{r}}(g) &= 2n^{-2}g^{-2|\mathbf{r}|-2}\psi_{\mathbf{0}}R(K^{(\mathbf{r})}) \\ &\quad + \left[n^{-1}g^{-|\mathbf{r}|-2}K^{(\mathbf{r})}(\mathbf{0}) + \frac{1}{2}g^2\mu_2(K)\sum_{i=1}^{2}\psi_{\mathbf{r}+2\mathbf{e}_i}\right]^2. \end{aligned} \tag{5}$$

Wand and Jones [7] suggested that for each separate $\mathbf{r}$, one should select a data-driven estimate of the bandwidth $g \equiv g_{\mathbf{r}}$ which minimizes $\operatorname{AMSE}\hat{\psi}_{\mathbf{r}}$. However, this approach will produce a matrix estimate $\hat{\mathbf{\Psi}}_4$ which may not be positive-definite. Plugging an estimate which lacks this property in Eq. (2) will produce an estimated AMISE surface without a finite global minimum (since it will decrease monotonically in some direction). Alternatively, $\hat{\mathbf{\Psi}}_4$ may be positive definite but almost singular, which can lead to serious numerical instability when seeking the minimizer of the estimated AMISE. This behaviour in the estimated $\hat{\mathbf{\Psi}}_4$ is an example of an important aspect of estimation for high-dimensional structures, namely that MSE optimal estimation of each element of the structure may produce clearly sub-optimal estimates (in many senses) of the structure as a whole.

Our preferred approach is to employ a common value of g for estimation of all elements of $\mathbf{\Psi}_4$. If K is multivariate normal (as we shall assume henceforth) then this is bound to produce a positive-definite estimate $\hat{\mathbf{\Psi}}_4$. To see this, note first that $\mathbf{\Psi}_4$ is positive-definite for any (continuous) target density f. It is easy to show that $\hat{\mathbf{\Psi}}_4$ is the $\mathbf{\Psi}_4$ matrix corresponding to $f = \hat{f}(\cdot; 2^{-1}g^2\mathbf{I})$ under the aforementioned condition on K, and hence the matrix estimate has the required property. We note that our approach has the added advantage of parsimony, since we need select only a single g rather than separate pilot bandwidths for each possible $\mathbf{r}$. It remains to describe a methodology for selecting a common g. We propose to estimate the bandwidth that minimizes the sum of AMSE (SAMSE) for $\hat{\mathbf{\psi}}_{\mathbf{r}}$; that is

$$g_{4,\mathrm{SAMSE}} = \operatorname*{arg\,min}_{g>0} \operatorname{SAMSE} \hat{\mathbf{\Psi}}_4$$

where

$$\operatorname{SAMSE} \hat{\mathbf{\Psi}}_4 \equiv \operatorname{SAMSE}_4(g) = \sum_{\mathbf{r}:|\mathbf{r}|=4} \operatorname{AMSE} \hat{\psi}_{\mathbf{r}}(g). \tag{6}$$

It is clear from Eqs. (5) and (6) that $g_{4,\mathrm{SAMSE}}$ will depend on the functionals $\psi_{\mathbf{r}+2\mathbf{e}_i}$ for $|\mathbf{r}| = 4$. These functionals are elements of $\mathbf{\Psi}_6$, and hence pilot estimation of this matrix will be necessary in order to derive a data-driven version of $g_{4,\mathrm{SAMSE}}$. For this second stage of pilot estimation we could employ the bandwidth $g_{6,\mathrm{SAMSE}}$; *i.e.* the minimizer of $\operatorname{SAMSE}\hat{\mathbf{\Psi}}_6$. Generalizing, we will be interested in the SAMSE optimal pilot bandwidth $g_{j,\mathrm{SAMSE}}$ for $j = |\mathbf{r}| = 4, 6, 8, \ldots$. Fortunately $g_{j,\mathrm{SAMSE}}$ is available in closed form, as we now show.

We have, from Eq. (5),

$$\begin{aligned}\operatorname{SAMSE}_j(g) &= \sum_{\mathbf{r}:|\mathbf{r}|=j} \operatorname{AMSE} \hat{\psi}_{\mathbf{r}}(g) \\ &= 2n^{-2}g^{-2j-2}A_1 + n^{-2}g^{-2j-4}A_2 + n^{-1}g^{-j}A_3 + \frac{1}{4}g^4A_4\end{aligned}$$

where

$$A_1 = \sum_{\mathbf{r}:|\mathbf{r}|=j} R(K^{(\mathbf{r})}),$$
$$A_2 = \sum_{\mathbf{r}:|\mathbf{r}|=j} K^{(\mathbf{r})}(\mathbf{0})^2,$$
$$A_3 = \mu_2(K) \sum_{\mathbf{r}:|\mathbf{r}|=j} K^{(\mathbf{r})}(\mathbf{0}) \left(\sum_{i=1}^{2} \psi_{\mathbf{r}+2\mathbf{e}_i} \right),$$
$$A_4 = \mu_2(K)^2 \sum_{\mathbf{r}:|\mathbf{r}|=j} \left(\sum_{i=1}^{2} \psi_{\mathbf{r}+2\mathbf{e}_i} \right)^2.$$

Note that A_1, A_2 and A_4 are positive by construction. Furthermore, $A_3 < 0$ under our assumption that K is multivariate normal. To see this, note that when all elements of $\mathbf{r}$ are even then $K^{(\mathbf{r})}(\mathbf{0})$ and $\psi_{\mathbf{r}+2\mathbf{e}_i}$ are of opposite sign; and when at least one of these elements is odd then $K^{(\mathbf{r})}(\mathbf{0}) = 0$. Now, the SAMSE expression can be simplified as the first term is $O(n^{-2}g^{-2j-2})$ and the second term is $O(n^{-2}g^{-2j-4})$, which means the latter always dominates the former. If we remove the first term (which is the asymptotic variance) we are left with

$$\mathrm{SAMSE}_j(g) = n^{-2}g^{-2j-4}A_2 + n^{-1}g^{-j}A_3 + \frac{1}{4}g^4 A_4. \tag{7}$$

Then differentiating this with respect to g gives

$$\frac{\partial}{\partial g}\mathrm{SAMSE}_j(g) = -(2j+4)n^{-2}g^{-2j-5}A_2 - jn^{-1}g^{-j-1}A_3 + g^3 A_4.$$

Setting this to zero and dividing by $-g^3$, we obtain

$$(2j+4)n^{-2}g^{-2j-8}A_2 + jn^{-1}g^{-j-4}A_3 - A_4 = 0.$$

This is a quadratic in $n^{-1}g^{-j-4}$ which can be solved to give the jth order SAMSE-optimal pilot bandwidth as

$$g_{j,\mathrm{SAMSE}} = \left[\frac{(4j+8)A_2}{(-jA_3 + \sqrt{j^2A_3^2 + (8j+16)A_2A_4})n} \right]^{1/(j+4)}. \tag{8}$$

4 PRACTICAL PLUG-IN ALGORITHMS

In the previous section we saw that optimal SAMSE smoothing of $\hat{\psi}_{\mathbf{r}}$ functionals of order $j = |\mathbf{r}|$ requires pilot estimates of corresponding functionals of order $j+2$. To implement a practical plug-in methodology it is therefore necessary to forego the full SAMSE approach at some given maximum order $j_{\max}$. For functionals of this order we simply employ normal reference estimates,

$$\hat{\psi}_{\mathbf{r}}^{\mathrm{NR}} = (-1)^{|\mathbf{r}|}\phi_{2\mathbf{S}}^{(\mathbf{r})}(\mathbf{0})$$

where $\phi_{\mathbf{\Sigma}}(\boldsymbol{x})$ is the multivariate normal density with zero mean and covariance matrix $\mathbf{\Sigma}$ evaluated at $\boldsymbol{x}$; and $\mathbf{S}$ is the covariance matrix of the data. We are now in a position to give the basic structure of our plug-in algorithm. Note that this algorithm is indexed by the number of stages, m, at which $\psi_{\mathbf{r}}$ functionals are estimated by kernel methods (as opposed to normal reference).

4.1 Algorithm for *m*-Stage Plug-in Bandwidth Matrix Selection

1. Set $j_{\max} = 2m + 4$. Obtain normal reference estimates $\hat{\psi}_{\mathbf{r}}^{\mathrm{NR}}$ for $|\mathbf{r}| = j_{\max}$. Plug these estimates into the SAMSE-optimal bandwidth $g_{j_{\max}-2,\mathrm{SAMSE}}$.
2. For $j = j_{\max} - 2, j_{\max} - 4, \ldots, 6$:
   (a) Calculate kernel estimates of $\psi_{\mathbf{r}}$ functionals of order $j = |\mathbf{r}|$ using plug-in estimate of $g_{j,\mathrm{SAMSE}}$.
   (b) Substitute $\hat{\psi}_{\mathbf{r}}$ estimates into Eq. (8) to give plug-in estimate of $g_{j-2,\mathrm{SAMSE}}$.
3. Employ $g_{4,\mathrm{SAMSE}}$ to produce kernel estimate $\hat{\mathbf{\Psi}}_4$. Plug this estimate into Eq. (2) to give $\widehat{\mathrm{AMISE}}$.
4. Numerically minimize $\widehat{\mathrm{AMISE}}$ to obtain required plug-in bandwidth matrix $\hat{\mathbf{H}}_{\mathrm{AMISE}}$.

This algorithm uses pilot bandwidths of the form $\mathbf{G} = g^2\mathbf{I}$ which will be clearly inappropriate if the dispersion of the data differs markedly between the two coordinate directions. Therefore the data should be pre-transformed before the algorithm is employed. More specifically, we propose that the algorithm is applied to transformed data $\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \ldots, \boldsymbol{X}_n^*$, where the transformation is either *sphering*

$$\boldsymbol{X}^* = \mathbf{S}^{-1/2}\boldsymbol{X}$$

where $\mathbf{S}$ is the sample covariance matrix of the untransformed data; or *scaling*

$$\boldsymbol{X}^* = \mathbf{S}_D^{-1/2}\boldsymbol{X}$$

where $\mathbf{S}_D = \mathrm{diag}(s_1^2, s_2^2)$ and s_1^2, s_2^2 are the diagonal elements of $\mathbf{S}$ (*i.e.* marginal sample variances). The plug-in bandwidth matrix $\hat{\mathbf{H}}_{\mathrm{AMISE}}^*$ for the sphered or scaled data can be back transformed to the original scale by $\hat{\mathbf{H}}_{\mathrm{AMISE}} = \mathbf{S}^{1/2}\hat{\mathbf{H}}_{\mathrm{AMISE}}^*\mathbf{S}^{1/2}$ or $\hat{\mathbf{H}}_{\mathrm{AMISE}} = \mathbf{S}_D^{1/2}\hat{\mathbf{H}}_{\mathrm{AMISE}}^*\mathbf{S}_D^{1/2}$, as appropriate.

In practice we employ a quasi-Newton (variable metric) method of numerical minimization at stage 4 of the algorithm. In the simulation study reported in Section 5.1, we did not encounter any significant computational difficulties using this approach.

5 NUMERICAL STUDIES

This section is split in two parts. The first of these reports on a simulation study used to compare various plug-in bandwidth matrix selectors, while the second considers density estimation for a real data set.

5.1 A Simulation Study

Here we seek to compare the performance of our plug-in bandwidth matrix selector to existing plug-in methodologies. We consider six target densities (labelled A through F), each of

which is a normal mixture described in Figures 1 and 2. All but density F were included in the numerical studies in Ref. [6]; they exhibit a range of characteristics that we might wish to detect using kernel density estimation. From each target density we generated 400 data sets of size $n = 100$ and the results are detailed below. (We also generated 400 data sets of size $n = 1000$, but the results are similar to those for $n = 100$ and so have been largely excluded for the sake of brevity.) For each data set we constructed bivariate kernel density estimates using multivariate normal K and bandwidth matrix selected using the following methods:

- Wand and Jones' [7] 2-stage plug-in diagonal bandwidth matrix selector, which we label D2;

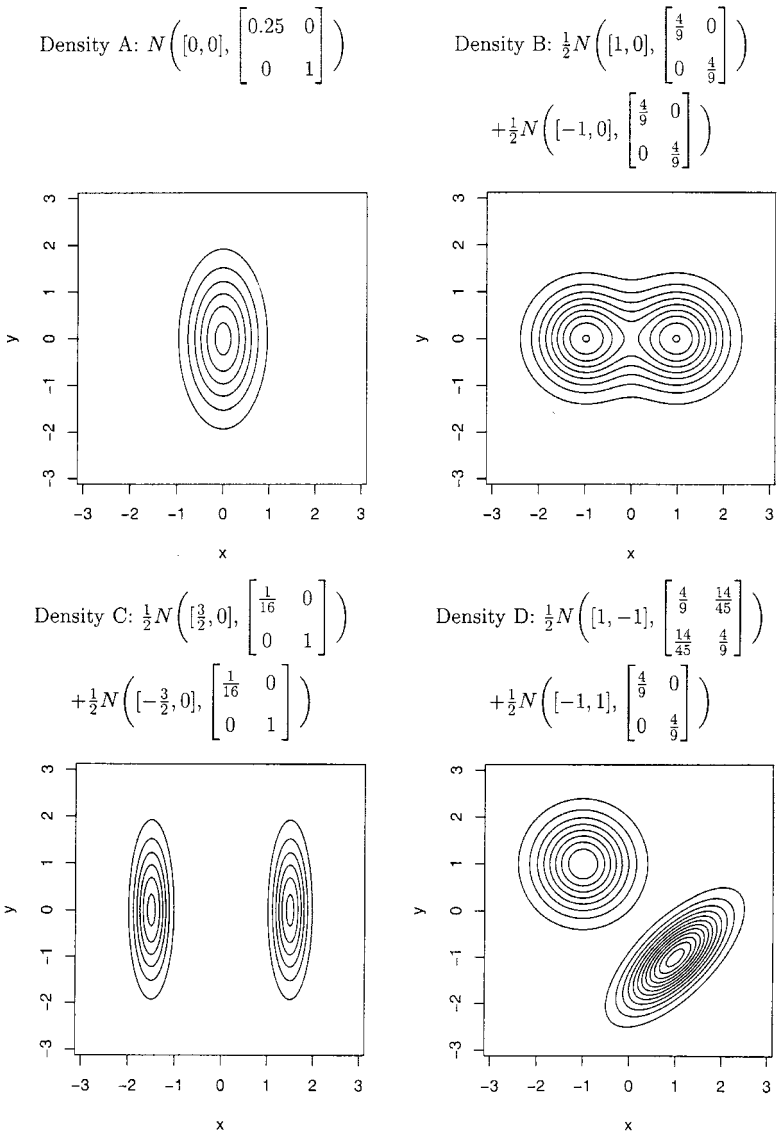

FIGURE 1 Test densities A, B, C and D.

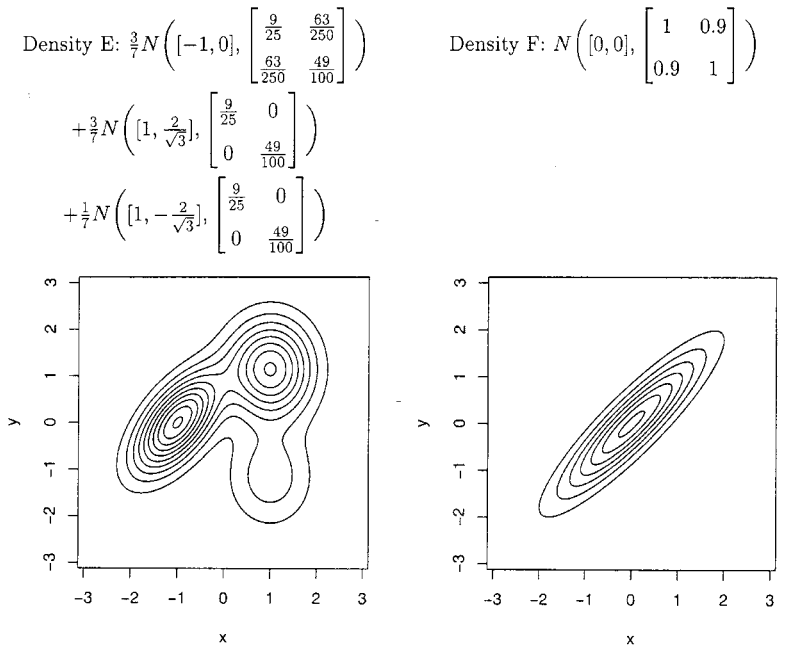

FIGURE 2 Test densities E and F.

- Wand and Jones' [7] 1-stage and 2-stage plug-in full bandwidth matrix selectors, labelled F1 and F2 respectively;
- Plug-in bandwidth matrix selectors using our 1-stage and 2-stage SAMSE based algorithm, labelled S1 and S2 respectively.

All but the diagonal bandwidth matrix selector were implemented using both pre-scaling and pre-sphering of the data. We add an asterisk superscript to the method label to indicate the latter type of transformation (*e.g.* F2*).

Before examining the results as a whole, it is important to note that methods F1 and F2 failed to produce plug-in bandwidths for some data sets. This occurred when the estimate $\hat{\mathbf{\Psi}}_4$ failed to be positive-definite. The failure rate (as a percentage) is classified by target density and sample size in Tables I and II. A number of aspects of these results deserve particular note. First, the failure rates of both F1 and F2 are sufficiently large (for certain target densities) that they cannot be ignored from the viewpoint of the practical user. Secondly, the failures occurred for the densities which are not oriented in parallel to the coordinate axes. Thirdly, the failure rates do not appear to diminish with the larger sample size. Wand and Jones' [7] full bandwidth matrix selector did not encounter such problems when applied

TABLE I Percentage Failure Rate for F1 Bandwidth Matrix Selector.

	Target density					
n	*A*	*B*	*C*	*D*	*E*	*F*
100	0.00	0.00	0.00	0.50	0.50	6.75
1000	0.00	0.00	0.00	2.75	0.00	5.25

TABLE II Percentage Failure Rate for F2 Bandwidth Matrix Selector.

	Target density					
n	*A*	*B*	*C*	*D*	*E*	*F*
100	0.00	0.00	0.00	1.75	0.25	4.75
1000	0.00	0.00	0.00	4.75	0.00	3.25

to sphered data. For every data set the matrix estimate $\hat{\mathbf{\Psi}}_4$ was positive definite for both F1* and F2*. While it remains theoretically possible for either of these methods to fail, this seems likely only when the structure of the target density is very intricate, for example, when f is composed of several components with long, thin elliptical contours at a variety of orientations to the coordinate axes.

The integrated squared error was computed for each (successfully) estimated density. (Note that this is available in closed form because each f is a normal mixture and K is normal; see Ref. 6). The efficacy of each methodology is compared using box plots of log(ISE) in Figures 3 and 4.

A striking aspect of the results is that relative performance depends strongly on target density shape. For densities *A* and *B* all methods considered produce very comparable results. In particular, D2 is no worse than the more complex full bandwidth matrix selectors in these cases. The picture is somewhat similar for target density *E*, although the two-stage full bandwidth matrix selectors have a small advantage over alternative methods. For target density *C*,

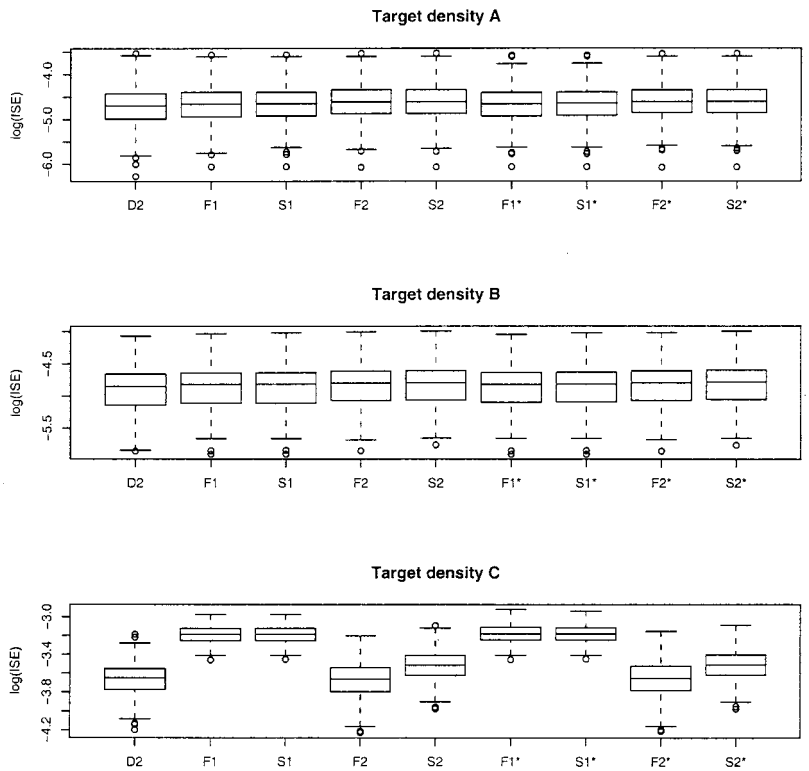

FIGURE 3 Box plots of log(ISE) for samples of size $n = 100$ from densities A, B and C.

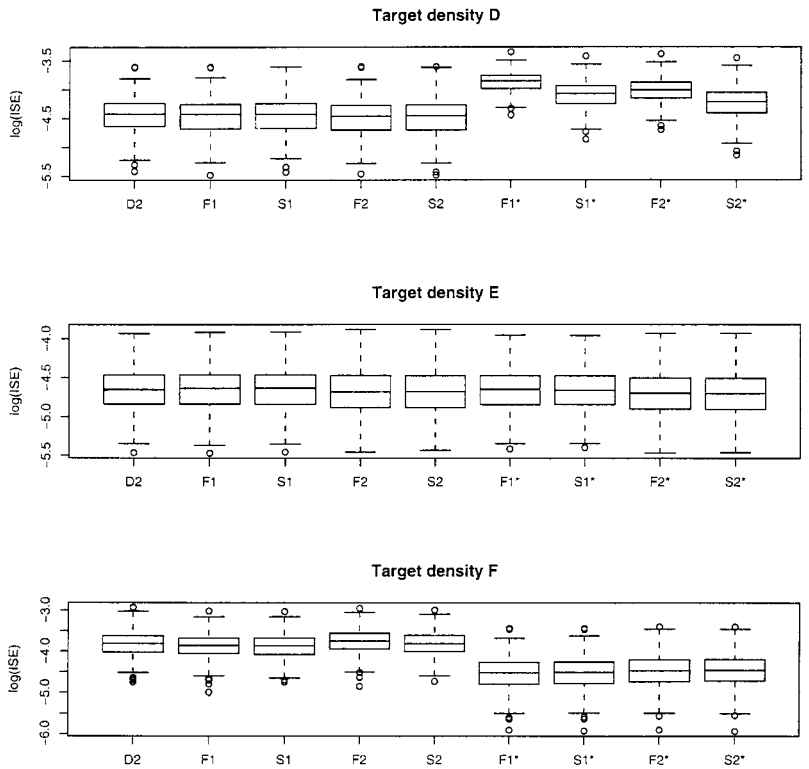

FIGURE 4 Box plots of log(ISE) for samples of size $n = 100$ from densities D, E and F.

one-stage full bandwidth matrix selectors perform rather poorly. The problem is that this density is so poorly represented by a normal reference distribution, and so the effects of this poor representation need to be mollified by an additional stage of pilot estimation. Turning to density *D*, this provides an excellent example of a situation in which sphering the data is detrimental (in comparison to simple scaling). This transformation corrupts important structure in *f*, largely because the orientation of the density as a whole is completely at odds with the orientation of the individual components of the density. (Note that the overall correlation is -0.58 while the individual mixture components have correlation zero and 0.7.) Of all the methods based on sphered data, only S2* comes close to competing with methods using scaled data. Finally, the pattern of results for density *F* is almost the exact reverse of that for density *D*. Here sphering the data proves a positive boon, as might be expected given the orientation of *f* to the coordinate axes. We note that D2 does very poorly with this target density. Naturally the performance of a diagonal matrix bandwidth selector could be hugely improved by sphering of the data in this case. However, we concur with Wand and Jones [6] that the implementation of diagonal bandwidth matrix selection with data sphering is not generally advisable. Indeed, such an approach performs very badly when *f* has a shape similar to target density *D*.

5.2 Density Estimation for Old Faithful Geyser Data

In this section we consider data from the 'Old Faithful' geyser in Yellowstone National Park, USA, as described in Ref. [5] (amongst many others). This data set consists of

$$\text{D2: } \hat{\mathbf{H}}_{\text{AMISE}} = \begin{bmatrix} 0.02582 & 0 \\ 0 & 5.686 \end{bmatrix} \qquad \text{S2}^* : \hat{\mathbf{H}}_{\text{AMISE}} = \begin{bmatrix} 0.0565 & 0.5604 \\ 0.5604 & 10.503 \end{bmatrix}$$

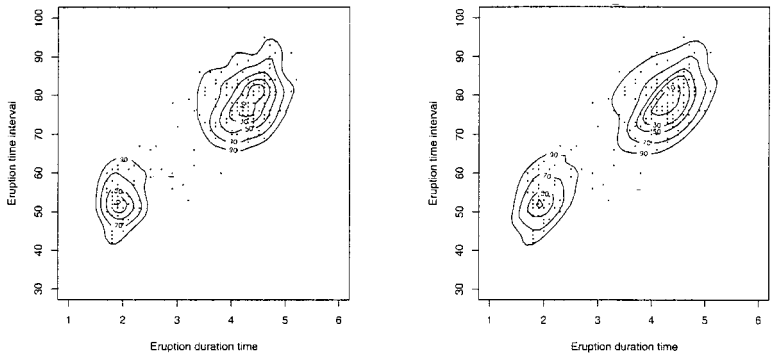

FIGURE 5 'Old Faithful' geyser density estimate using D2 and S2*.

222 observations on duration of eruptions of the geyser, each of which is paired with the waiting time for the following eruption (both time intervals measured in minutes). The data set is strongly bimodal with an overall orientation angled away from the coordinate axes, and so provides a relatively challenging setting for implementation of the bandwidth matrix selectors. The bandwidth matrix and density estimate for D2 are displayed in Figure 5. The corresponding results for the full bandwidth matrix selectors applied to scaled data are similar to D2 and have been omitted for brevity. The results for S2* are also in Figure 5. The other methods employing sphered data are similar to S2* and they have also been omitted. A major feature of these results is the difference between (i) methods using a diagonal bandwidth matrix, or a full bandwidth matrix applied to scaled data, and (ii) methods using sphered data. In particular, the former group of methods provide density estimates in which the lower left mode runs almost parallel to the waiting time axis. For the sphered data methods the orientation of this mode is at a marked angle to this axis. We also note that the elements of the bandwidth matrices are larger for the sphered data methods than the scaled data ones, producing smoother estimates.

6 CONCLUSIONS

In this paper we have considered bandwidth matrix selection for bivariate kernel density estimators. This is an important problem both from a practical and theoretical standpoint. The majority of work in this area has been directed towards selection of diagonal bandwidth matrices, but we reiterate Wand and Jones' [6] viewpoint that full bandwidth matrices can give markedly better performance for some types of target density. Our methodological contribution has been to develop a new type of plug-in selector for full bandwidth matrices. Our methodology has the advantage, in comparison to the full bandwidth matrix techniques outlined by Wand and Jones [7], that it will always produce a finite bandwidth matrix. Furthermore, our approach requires computation of significantly fewer pilot bandwidths.

The simulation study in Section 5.1 had two purposes. It provided information as to the optimal implementation of our technology (*i.e.* whether to use scaled or sphered data, and

whether to use one or two stages of pilot estimation) and also compared the performance of various competing methodologies. Our overall findings were that the two-stage methods are to be preferred to their one-stage counterparts. Amongst the two-stage methods, D2, F2 and S2 all performed badly when estimating target density *F*. Furthermore, F2 failed to produce finite bandwidth matrices on a significant number of occasions. This suggests that the optimal implementation of our bandwidth matrix selector involves two pilot stages, and sphering of the data; *i.e.* S2*. The principal competitor to S2* is F2*. The performance of these methods was almost indistinguishable on most target densities, while the advantage of F2* on density *C* is offset by the advantage of S2* on density *D*. However, it should be recalled that S2* is the simpler method to implement, requiring just two pilot bandwidths (excluding normal reference bandwidths) in comparison to the twelve required by F2*. Indeed, S2* is even more parsimonious than D2 in this regard, with the diagonal bandwidth matrix selector needing seven pilot bandwidths.

We finish by mentioning some possible extensions of our work, and some avenues for further research into bandwidth selection for bivariate (and multivariate) density estimation. It is a straightforward exercise to extend our SAMSE technology to higher dimensional density estimation. We note that the benefits of the SAMSE approach in terms of parsimony of pilot bandwidth selection will become increasingly great as the dimensionality of the data increases. Nonetheless, whether the relative lack of flexibility in the SAMSE method will tell against it in terms of performance at higher dimensions remains to be discovered. On a more general issue regarding multivariate bandwidth selection, we note that plug-in methods are the only univariate technique that has been transferred to higher dimensions for full bandwidth selection. An open research question is whether cross-validation and bootstrap methods (so far restricted to the case of diagonal bandwidth matrices) can provide an attractive practical alternative to the plug-in approach in this context.

Acknowledgements

The first author acknowledges the financial support of an Australian Postgraduate Award at the University of Western Australia.

References

[1] Jones, M. C. and Sheather, S. J. (1991). Using nonstochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics and Probability Letters*, **11**, 511–514.
[2] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, **91**, 401–407.
[3] Sain, S. R., Baggerly, K. A. and Scott, D. W. (1994). Cross-validation of multivariate densities. *Journal of the American Statistical Association*, **89**, 807–817.
[4] Scott, D. W. (1992). *Multivariate Density Estimation; Theory, Practice and Visualization.* Wiley-Interscience, New York.
[5] Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* Springer, New York.
[6] Wand, M. P. and Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, **88**, 520–528.
[7] —. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, **9**, 97–116.
[8] —. (1995). *Kernel Smoothing.* Chapman & Hall, London.