ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Computational Statistics & Data Analysis ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

A Bayesian approach to bandwidth selection for multivariate kernel density estimation

Xibin Zhang, Maxwell L. King*, Rob J. Hyndman

Department of Econometrics and Business Statistics, Monash University, Clayton, Victoria 3800, Australia

Received 30 November 2004; received in revised form 31 May 2005; accepted 16 June 2005

Abstract

Kernel density estimation for multivariate data is an important technique that has a wide range of applications. However, it has received significantly less attention than its univariate counterpart. The lower level of interest in multivariate kernel density estimation is mainly due to the increased difficulty in deriving an optimal data-driven bandwidth as the dimension of the data increases. We provide Markov chain Monte Carlo (MCMC) algorithms for estimating optimal bandwidth matrices for multivariate kernel density estimation. Our approach is based on treating the elements of the bandwidth matrix as parameters whose posterior density can be obtained through the likelihood cross-validation criterion. Numerical studies for bivariate data show that the MCMC algorithm generally performs better than the plug-in algorithm under the Kullback–Leibler information criterion, and is as good as the plug-in algorithm under the mean integrated squared error (MISE) criterion. Numerical studies for five-dimensional data show that our algorithm is superior to the normal reference rule. Our MCMC algorithm is the first data-driven bandwidth selector for multivariate kernel density estimation that is applicable to data of any dimension.

Keywords: Cross-validation; Kullback–Leibler information; Mean integrated squared errors; Sampling algorithms; Monte Carlo kernel likelihood

* Corresponding author. Tel.: +61 3 99052449; fax: +61 3 99058039.
E-mail address: max.king@buseco.monash.edu.au (M.L. King).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.06.019

1. Introduction

Multivariate kernel density estimation is an important technique in multivariate data analysis and has a wide range of applications (see, for example, Scott, 1992; Aït-Sahalia, 1996; Donald, 1997; Stanton, 1997; Aït-Sahalia and Lo, 1998; de Valpine, 2004). However, its widespread usefulness has been limited by the difficulty in computing an optimal data-driven bandwidth. We remedy this deficiency in this paper.

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)'$ denote a d-dimensional random vector with density $f(\boldsymbol{x})$ defined on $\boldsymbol{R}^d$, and let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ be an independent random sample drawn from $f(\boldsymbol{x})$. The general form of the kernel estimator of $f(\boldsymbol{x})$ is (Wand and Jones, 1995)

$$\hat{f}_H(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} K_H(\boldsymbol{x} - \boldsymbol{x}_i),$$

where $K_H(\boldsymbol{x}) = |H|^{-1/2} K\left(H^{-1/2}\boldsymbol{x}\right)$, $K(\cdot)$ is a multivariate kernel function, and H is a symmetric positive definite $d \times d$ matrix known as the bandwidth matrix.

The bandwidth matrix can be restricted to a class of positive definite diagonal matrices, and then the corresponding kernel function is known as a product kernel. However, there is much to be gained by choosing a full bandwidth matrix, where the corresponding kernel smoothing is equivalent to pre-rotating the data by an optimal amount and then using a diagonal bandwidth matrix (Wand and Jones, 1993). It has been widely recognized that the performance of a kernel density estimator is primarily determined by the choice of bandwidth, and only in a minor way by the choice of kernel function (see, for example, Izenman, 1991; Scott, 1992; Simonoff, 1996).

A large body of literature exists on bandwidth selection for univariate kernel density estimation (see, for example, Marron, 1987; Jones et al., 1996 for surveys). However, the literature on bandwidth selection for multivariate data is very limited. To choose an optimal bandwidth matrix, a criterion must be used; one such criterion is the integrated squared error (ISE) given by

$$\mathrm{ISE}(H) = \int_{R^d} \left[\hat{f}_H(\boldsymbol{x}) - f(\boldsymbol{x})\right]^2 \mathrm{d}\boldsymbol{x},$$

where $f(\boldsymbol{x})$ denotes the target density. The least-squares cross-validation method aims to derive a bandwidth that minimizes ISE (see, for example, Härdle, 1991; Sain et al., 1994). However, the difficulty in deriving a numerical optimal bandwidth grows as the dimension of data increases. Another criterion for choosing an optimal bandwidth is the mean integrated squared error (MISE) expressed as

$$\mathrm{MISE}(H) = E \int_{R^d} \left[\hat{f}_H(\boldsymbol{x}) - f(\boldsymbol{x})\right]^2 \mathrm{d}\boldsymbol{x}.$$

It is well known that the optimal bandwidth that minimizes MISE does not have a closed form. In order to make progress under this criterion, it is usual to employ an asymptotic approximation. When data are observed from the multivariate normal density and the diagonal bandwidth matrix, denoted by $H = \mathrm{diagonal}\,(h_1, h_2, \ldots, h_d)$, is employed, the optimal

bandwidth that minimizes MISE can be approximated by (Scott, 1992; Bowman and Azzalini, 1997)

$$h_i = \sigma_i \left\{ \frac{4}{(d+2)n} \right\}^{1/(d+4)},$$

for $i = 1, 2, \ldots, d$, where σ_i is the standard deviation of the ith variate and can be replaced by its sample estimator in practical implementations. We call this the "normal reference rule". This method is often used in practice, in the absence of any other practical bandwidth selection schemes, despite the fact that most interesting data are non-Gaussian, and that a full bandwidth matrix is preferable.

Sain et al. (1994) derived an estimate of the asymptotic MISE (AMISE) for bivariate densities and employed biased cross-validation to estimate the optimal bandwidth. However, their method cannot be directly extended to the general multivariate setting. Wand and Jones (1995) showed that under certain smoothness assumptions on the target density, the AMISE is expressed as

$$\begin{aligned} \mathrm{AMISE}(H) = {} & \frac{1}{n} |H|^{-1/2} \int_{R^2} K^2(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\ & + \frac{1}{4} \int_{R^2} \boldsymbol{x}\boldsymbol{x}' K^2(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \ \left(\mathrm{vech}' H\right) \Psi_4(\mathrm{vech}\, H), \end{aligned}$$

where 'vech' is the vector half operator and Ψ_4 is a matrix whose elements are functionals of the unknown target density $f(\boldsymbol{x})$. An estimate of the optimal bandwidth can be derived using the plug-in method, which aims to minimize $\mathrm{AMISE}(H)$ by plugging an estimate of Ψ_4 in the above equation. For bivariate data, Wand and Jones (1994) presented a plug-in algorithm, which requires auxiliary smoothing parameters. The technology for choosing these auxiliary smoothing parameters is not well developed. Duong and Hazelton (2003) argued that the full bandwidth matrix selectors suggested by Wand and Jones (1994) fail to produce plug-in bandwidths for some data sets. In response to this problem, Duong and Hazelton (2003) presented an alternative plug-in algorithm, which has the advantage that it always produces a finite bandwidth matrix and requires computation of fewer pilot bandwidths. However, these plug-in algorithms cannot be directly extended to the general multivariate setting.

The maximum likelihood cross-validation criterion (discussed in Section 2) leads to an optimal bandwidth that minimizes the Kullback–Leibler information. The likelihood cross-validation bandwidth selector requires a numerical optimization procedure, which becomes increasingly difficult to implement as the dimension of data increases (see, for example, Härdle, 1991). However, from a Bayesian perspective, we can treat nonzero components of H as parameters, whose posterior density can be obtained through the likelihood cross-validation criterion. A posterior estimate of H can be derived through the Markov chain Monte Carlo (MCMC) technique. One important advantage of the MCMC technique for estimating optimal bandwidths is that it is applicable to data of any dimension, not only to bivariate data. Moreover, the sampling algorithm involves no increased difficulty as the dimension of the data increases.

To our knowledge, the only previous paper employing a Bayesian approach to bandwidth selection for kernel density estimation is Brewer (2000). He derived adaptive bandwidths for

univariate kernel density estimation, treating the bandwidths as parameters and estimating them via MCMC simulations. Brewer (2000) showed that the proposed Bayesian approach is superior to methods of Abramson (1982) and Sain and Scott (1996).

Schuster and Gregory (1981) demonstrated that in some circumstances, likelihood cross-validation produces inconsistent estimates for univariate kernel density estimation. However, Brewer (2000) argued that the MCMC approach to adaptive bandwidth selection may avoid the inconsistency problem by choosing an appropriate prior and using a kernel with infinite support. The same argument applies to the case considered here.

In this paper, we present MCMC algorithms for estimating the optimal bandwidth matrix for multivariate kernel density estimation through the likelihood cross-validation criterion, and sampling algorithms are developed for both diagonal and full bandwidth matrices. The rest of this paper is organized as follows. Section 2 briefly discusses the likelihood cross-validation criterion and presents MCMC algorithms for both diagonal and full bandwidth matrices. In Section 3, we examine the performance of MCMC algorithms with data generated from known bivariate densities. We find that the MCMC algorithm generally performs better than either the plug-in algorithm or the normal reference rule in the bivariate setting. Section 4 applies the MCMC bandwidth selectors to data generated from known multivariate densities, and we find that the MCMC algorithm performs much better than the normal reference rule (there are no other bandwidth selection methods available in this case). Section 5 illustrates the use of the MCMC algorithm for bandwidth selection with an application to some earthquake data and to estimation of financial data based on Monte Carlo kernel likelihood. We provide conclusions in Section 6.

2. MCMC for optimal bandwidth selection

2.1. Likelihood cross-validation

Kullback–Leibler information is a measure of distance between two densities. Our interest is in choosing the approximate density $\hat{f}_H(\boldsymbol{x})$ to minimize its distance from the target density $f(\boldsymbol{x})$. In this case, Kullback–Leibler information is defined as

$$\begin{aligned} d_{\mathrm{KL}}\left(f, \hat{f}_H\right) &= \int_{R^d} \log\left[\frac{f(\boldsymbol{x})}{\hat{f}_H(\boldsymbol{x})}\right] f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &= \int_{R^d} \log\left[f(\boldsymbol{x})\right] f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_{R^d} \log\left[\hat{f}_H(\boldsymbol{x})\right] f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \end{aligned} \tag{1}$$

which is nonnegative. We want to find an optimal bandwidth that minimizes $d_{\mathrm{KL}}\left(f, \hat{f}_H\right)$, or, equivalently, maximizes

$$E\,\log\left[\hat{f}_H(\boldsymbol{x})\right] = \int_{R^d} \log\left[\hat{f}_H(\boldsymbol{x})\right] f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

which can be approximated by

$$\hat{E}\ \log\left[\hat{f}_H(\boldsymbol{x})\right] = \frac{1}{n}\sum_{i=1}^{n}\log\hat{f}_H\left(\boldsymbol{x}_i\right) = \frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{1}{n}\sum_{j=1}^{n}K_H\left(\boldsymbol{x_i}-\boldsymbol{x}_j\right)\right]. \tag{2}$$

If we directly maximize (2) with respect to H, the resulting bandwidth is a matrix of zeros. A way out of this dilemma is to estimate $f_H\left(\boldsymbol{x}_i\right)$ based on the subset $\left\{\boldsymbol{x}_j : j \neq i\right\}$, and to approximate $E\ \log\left[\hat{f}_H(\boldsymbol{x})\right]$ by (Härdle, 1991)

$$L\left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \mid H\right) = \frac{1}{n}\sum_{i=1}^{n}\log\hat{f}_{H,i}\left(\boldsymbol{x}_i\right), \tag{3}$$

where $\hat{f}_{H,i}$ is the leave-one-out estimator

$$\hat{f}_{H,i}\left(\boldsymbol{x}_i\right) = \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}|H|^{-1/2}K\left(H^{-1/2}\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right)\right).$$

The likelihood cross-validation criterion is to select H by maximizing $n^{-1}L(\cdot \mid H)$.

Solving this maximization problem requires a numerical procedure, which becomes increasingly difficult to implement as the dimension increases. However, when nonzero components of H are treated as parameters, the logarithmic likelihood of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is provided by (3), and the posterior density of the parameters is proportional to the product of the joint prior density of nonzero components of H and the likelihood. As the MCMC technique is very powerful in sampling a high-dimensional vector of parameters, it can be employed to obtain a posterior estimate for the bandwidth matrix.

It is worth noting that rather than the MISE criterion and Kullback–Leibler information criterion for optimal bandwidth selection, one can use the criterion of maximizing accuracy of the mode location. However, under this criterion, it is impossible to obtain the likelihood of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ given nonzero bandwidths as parameters. In contrast, Kullback–Leibler information provides the possibility of deriving the likelihood, which we use to construct the posterior density.

2.2. Sampling a diagonal bandwidth matrix

When H is diagonal, the kernel density estimator of $f(\boldsymbol{x})$ is

$$\hat{f}_{\boldsymbol{h}}(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h_1h_2\cdots h_d}K\left(\frac{x_1-x_{j,1}}{h_1}, \frac{x_2-x_{j,2}}{h_2}, \ldots, \frac{x_d-x_{j,d}}{h_d}\right),$$

where $\boldsymbol{h} = (h_1, h_2, \ldots, h_d)'$ is a vector of bandwidths with positive values. The leave-one-out estimator is

$$\hat{f}_{\boldsymbol{h},i}\left(\boldsymbol{x}_i\right) = \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{1}{h_1h_2\cdots h_d}K\left(\frac{x_{i,1}-x_{j,1}}{h_1}, \frac{x_{i,2}-x_{j,2}}{h_2}, \ldots, \frac{x_{i,d}-x_{j,d}}{h_d}\right),$$

for $i = 1, 2, \ldots, n$. We treat the bandwidth $\boldsymbol{h}$ as a vector of parameters, given which, the log likelihood function of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is

$$L\left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n | \boldsymbol{h}\right) = \sum_{i=1}^{n} \log\, \hat{f}_{\boldsymbol{h},i}\left(\boldsymbol{x}_i\right). \tag{4}$$

We assume that the prior density of each component of $\boldsymbol{h}$ is (up to a normalizing constant)

$$\pi\left(h_k | \lambda\right) \propto \frac{1}{1 + \lambda h_k^2}, \tag{5}$$

for $k = 1, 2, \ldots, d$, where λ is a hyperparameter controlling the shape of the prior density. According to Bayes theorem, the posterior of $\boldsymbol{h}$ is (up to a normalizing constant)

$$\pi\left(\boldsymbol{h} | \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\right) \propto \left[\prod_{k=1}^{d} \frac{1}{1 + \lambda h_k^2}\right] \times \prod_{i=1}^{n} \hat{f}_{\boldsymbol{h},i}\left(\boldsymbol{x}_i\right), \tag{6}$$

from which we can sample $\boldsymbol{h}$ using the Metropolis–Hastings algorithm. The ergodic average or the posterior mean of $\boldsymbol{h}$ acts as an estimator of optimal bandwidth.

The likelihood appearing in the posterior density given by (4) is flat when components of $\boldsymbol{h}$ are large. If we use uniform priors for the components of $\boldsymbol{h}$ and employ the random-walk Metropolis–Hastings algorithm to sample $\boldsymbol{h}$, the update of $\boldsymbol{h}$ has a negligible effect when components of $\boldsymbol{h}$ are already very large. In order to make the sampling algorithm work appropriately, sufficient prior information on components of $\boldsymbol{h}$ is required to put a low prior probability on the "problematic" region in the parameter space, where the likelihood function is flat. In this sense, the effect of the prior given by (5) seems to be a penalty on the likelihood.

In a different context, Bauwens and Lubrano (1998) used a similar prior for the degrees-of-freedom parameter of the t-distribution. They proved that with a diffuse prior on the degrees-of-freedom parameter on $(0, \infty)$, the resulting posterior density is not integrable, and a prior of the form of (5) provides integrability. In our case, we can show that a diffuse prior on each component of $\boldsymbol{h}$ results in a posterior that is integrable on $(0, \infty)$. Hence, the purpose of the prior given by (5) is not to provide integrability but to make the sampling algorithm work appropriately.

As well as the leave-one-out method, there are some other cross-validation methods, such as the fixed-fraction version cross-validation discussed by van der Laan et al. (2004). They showed that the fixed-fraction cross-validation method has good asymptotic properties for model selection. The fixed-fraction cross-validation is generally appropriate for the proposed Bayesian framework for choosing an optimal bandwidth.

2.3. Sampling a full bandwidth matrix

As the bandwidth matrix is symmetric positive definite, we can obtain its Cholesky decomposition $H = LL'$, where L is a lower triangular matrix. Let $B = L^{-1}$ which is also

lower triangular. Then the kernel estimator of $f(\boldsymbol{x})$ is

$$\hat{f}_B(\boldsymbol{x}) = \frac{1}{n}|B|\sum_{i=1}^{n} K\left(B\left(\boldsymbol{x}-\boldsymbol{x}_i\right)\right),$$

and the leave-one-out estimator of $f(\boldsymbol{x})$ is

$$\hat{f}_{B,i}(\boldsymbol{x}_i) = \frac{1}{n-1}|B|\sum_{\substack{j=1\\ j\neq i}}^{n} K\left(B\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right)\right).$$

We treat nonzero elements of the bandwidth matrix as parameters, whose posterior density can be obtained based on the likelihood function given in (3). We assume that the prior density of each nonzero component of B is (up to a normalizing constant)

$$\pi\left(b_{ij}\,\middle|\,\lambda\right) \propto \frac{1}{1+\lambda b_{ij}^2} \tag{7}$$

for $j \leqslant i$ and $i = 1, 2, \ldots, d$. Using Bayes theorem, we can obtain the posterior density of B (up to a normalizing constant)

$$\pi\left(B\,|\,\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\right) \propto \left[\prod_{i=1}^{d}\prod_{j=1}^{i}\frac{1}{1+\lambda b_{ij}^2}\right] \times \prod_{i=1}^{n}\hat{f}_{B,i}(\boldsymbol{x}_i), \tag{8}$$

from which we sample all elements of B using the Metropolis–Hastings algorithm. The ergodic average or the posterior mean of B acts as an estimator of optimal bandwidth.

2.4. Transformation of data

The plug-in algorithm for bandwidth selection developed by Duong and Hazelton (2003) uses a simple form for the pilot bandwidths, which is inappropriate when the dispersion of the data differs markedly between the two variates. Hence, Duong and Hazelton (2003) suggested that the data be pre-scaled before the plug-in algorithm is implemented.

Given a set of bivariate data denoted by $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, let S denote the sample variance–covariance matrix with diagonal components s_1^2 and s_2^2. Duong and Hazelton (2003) defined the sphering and scaling transformations, respectively, by

$$\boldsymbol{x}_i^* = S^{-1/2}\boldsymbol{x}_i \quad \text{and} \quad \boldsymbol{x}_i^* = S_d^{-1/2}\boldsymbol{x}_i, \tag{9}$$

for $i = 1, 2, \ldots, n$, where $S_d = \text{diagonal}\left(s_1^2, s_2^2\right)$. When the optimal bandwidth matrix, denoted by $\hat{H}^*$, for the transformed data is obtained, the optimal bandwidth matrix for the original data can be calculated through the reverse transformation, $\hat{H} = S^{1/2}\hat{H}^*\left(S^{1/2}\right)'$ or $\hat{H} = S_d^{1/2}\hat{H}^*S_d^{1/2}$.

To sample a bandwidth matrix, we shall use the random-walk Metropolis–Hastings algorithm, in which scaling (and possibly sphering) is of prime importance because the algorithm has to mix different scales of different variates (and to incorporate correlations

between variates). This kind of scaling (and sphering) is incorporated in the proposal density and is different from the scaling and sphering pre-transformations of the data defined in (9). If we make a scaling or sphering pre-transformation of the data, for which we derive an estimate of the optimal bandwidth, then we have to make a reverse transformation to derive an estimated bandwidth for the original data. However, the sampling algorithm can directly produce an estimated bandwidth for the original data, even though a certain kind of scaling and sphering might be involved.

If we choose a sphering transformation of data and use the diagonal bandwidth matrix, the resulting bandwidth estimator for the original data is a full matrix. When the variates are correlated and the diagonal bandwidth matrix is used, the bandwidth matrix estimator obtained through the sphering transformation of the original data might produce a better performance than that obtained directly from the original data, because the sphering transformation is equivalent to pre-rotating the data (see, for example, Wand and Jones, 1993).

3. Numerical studies with bivariate densities

This section examines the performance of the proposed MCMC methods for bandwidth selection via several sets of bivariate data, generated from known densities. As the true density is known in each case, the performance of the bandwidth can be measured by the accuracy of the corresponding kernel density estimator via Kullback–Leibler information.

Kullback–Leibler information defined in (1) is the mean of $\log\left(f(\boldsymbol{x})/\hat{f}_H(\boldsymbol{x})\right)$ under density $f(\boldsymbol{x})$, and so it measures the discrepancy of the estimated density from the true density. If a large number of random vectors, denoted by $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, can be drawn from $f(\boldsymbol{x})$, Kullback–Leibler information can be estimated by

$$\hat{d}_{\mathrm{KL}}\left(f, \hat{f}_H\right) = \frac{1}{N}\sum_{i=1}^{N} \log\left(f\ (\boldsymbol{x}_i)/\ \hat{f}_H\ (\boldsymbol{x}_i)\right). \tag{10}$$

3.1. True densities

We consider four target densities labelled A, B, C and D, respectively. Contour plots of these densities are shown in Fig. 1. Density A is a mixture of two bivariate normal densities, with high correlation and bimodality:

$$f_A\left(\boldsymbol{x}|\,\mu_1, \Sigma_1, \mu_2, \Sigma_2\right) = \tfrac{1}{2}\,\phi\left(\boldsymbol{x}|\,\mu_1, \Sigma_1\right) + \tfrac{1}{2}\,\phi\left(\boldsymbol{x}|\,\mu_2, \Sigma_2\right),$$

where $\phi(\boldsymbol{x}\mid\mu, \Sigma)$ denotes a multivariate normal density with mean μ and variance–covariance matrix Σ, and

$$\mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} -1.5 \\ -1.5 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}.$$

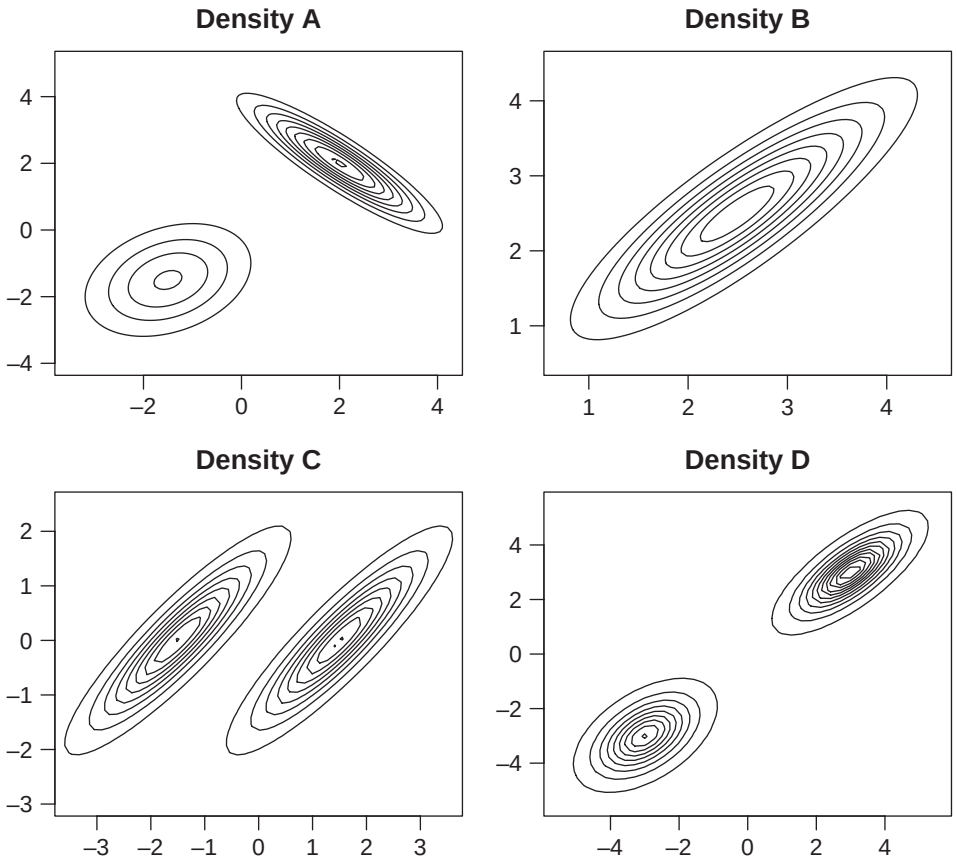

Fig. 1. Contour graphs of the proposed bivariate densities.

Density B is a bivariate skew-normal density with high correlation:

$$f_B(\boldsymbol{x} \mid \mu, \Sigma, \alpha) = 2\phi(\boldsymbol{x} \mid \mu, \Sigma)\, \Phi\left(\alpha' w^{-1/2}(\boldsymbol{x} - \mu)\right),$$

where $\Phi(\cdot)$ is the cumulative density function of a standard bivariate normal distribution and w is a diagonal matrix with diagonal elements the same as those of Σ. This distribution has been studied by Azzalini and Dalla Valle (1996), Azzalini and Capitanio (1999, 2003), Jones (2001) and Jones and Faddy (2003) among others. Here, α is a shape parameter capturing the skewness. When $\alpha = 0$, this density becomes the usual normal density. For the purpose of generating a set of data, we use the following parameters:

$$\mu = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}, \quad \alpha = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}.$$

Density C is a mixture of two bivariate Student t densities:

$$f_C\left(\boldsymbol{x}|\,\mu_1,\mu_2,\Sigma,\nu\right)=\tfrac{1}{2}\,t_d\left(\boldsymbol{x}|\,\mu_1,\Sigma,\nu\right)+\tfrac{1}{2}\,t_d\left(\boldsymbol{x}|\,\mu_2,\Sigma,\nu\right),$$

where

$$t_d(\boldsymbol{x}\,|\,\mu,\Sigma,\nu)=\frac{\Gamma((\nu+d)/2)}{(\nu\pi)^{d/2}\Gamma(\nu/2)|\Sigma|^{1/2}}\left[1+\frac{1}{\nu}\,(\boldsymbol{x}-\mu)'\Sigma^{-1}(\boldsymbol{x}-\mu)\right]^{-(d+\nu)/2},\quad(11)$$

has location parameter μ, dispersion matrix Σ and degrees-of-freedom ν, and with parameters set to

$$\mu_1=\begin{pmatrix}-1.5\\0\end{pmatrix},\quad \mu_2=\begin{pmatrix}1.5\\0\end{pmatrix},\quad \Sigma=\begin{pmatrix}1&0.9\\0.9&1\end{pmatrix}$$

and $\nu=5$. Density C exhibits heavy tail behaviour, high correlation and bimodality.

Density D is a mixture of two bivariate Student t densities, but has thicker tails than density C:

$$f_D\left(\boldsymbol{x}|\,\mu_1,\mu_2,\Sigma,\nu\right)=\tfrac{1}{2}\,t_d\left(\boldsymbol{x}|\,\mu_1,\Sigma_1,\nu\right)+\tfrac{1}{2}\,t_d\left(\boldsymbol{x}|\,\mu_2,\Sigma_2,\nu\right),$$

where $\nu=3$,

$$\mu_1=\begin{pmatrix}3\\3\end{pmatrix},\quad \Sigma_1=\begin{pmatrix}1&0.75\\0.75&1\end{pmatrix},\quad \mu_2=\begin{pmatrix}-3\\-3\end{pmatrix}\quad\text{and}\quad \Sigma_2=\begin{pmatrix}1&0.5\\0.5&1\end{pmatrix}.$$

3.2. Bandwidth matrix selectors

From each of the proposed bivariate densities, we generate data sets of size $n=200$, 500 and 1000, respectively. For each data set, we calculate the bivariate kernel density estimator using the standard bivariate Gaussian kernel function and bandwidth matrix selected through each of the following selectors:

M_1: MCMC algorithm for full bandwidth matrix without pre-transformation of data;
M_2: MCMC algorithm for full bandwidth matrix with scaling transformation of data;
M_3: MCMC algorithm for full bandwidth matrix with sphering transformation of data;
M_4: MCMC algorithm for diagonal bandwidth matrix without pre-transformation;
M_5: MCMC algorithm for diagonal bandwidth matrix with scaled data;
M_6: MCMC algorithm for diagonal bandwidth matrix with sphered data;
P_1: Plug-in selector of full bandwidth matrix with scaling transformation of data;
P_2: Plug-in selector of full bandwidth matrix with sphering transformation of data;
P_3: Plug-in selector of diagonal bandwidth matrix with scaling transformation of data;
P_4: Plug-in selector of diagonal bandwidth matrix with sphering transformation of data;
N_1: The normal reference rule approach for a diagonal bandwidth.

The plug-in bandwidth selector refers to the algorithm developed by Duong and Hazelton (2003). We have not included the plug-in algorithms of Wand and Jones (1993), because their algorithm for full bandwidth matrix selection sometimes fails to produce finite bandwidths for some data sets. When their algorithm works, its performance is similar to the plug-in

algorithm developed by Duong and Hazelton (2003). See, Duong and Hazelton (2003) for further discussion of these two plug-in algorithms.

3.3. MCMC outputs and sensitivity analysis

The hyperparameter of prior densities defined in (7) is initially set to λ=1 which represents a very flat prior. Given a data set generated from a bivariate density, we sample the diagonal and full bandwidth matrices from their corresponding posterior densities defined in (8) using the random-walk Metropolis–Hastings algorithm, in which the proposal density is the multivariate standard normal density. In order to prevent a false impression of convergence, the tuning parameter was chosen so that the acceptance rate was between 0.2 and 0.3.

The burn-in period is set at 5000 iterations, and the number of total recorded iterations is 25,000. The initial value of B is set to the identity matrix. After we obtain the sampled path of B for each data set, we calculate the ergodic average (or posterior mean) and the batch-mean standard error (see, for example, Roberts, 1996), where the number of batches is 50 and there are 500 draws in each batch. The ergodic average acts as an estimator of optimal bandwidth.

We ran our sampling algorithms on a parallel unix system, whose processor is a 64-bit EV6.8AL with 834 MHz and 2G RAM. The required CPU time is around 2 min for a sample size of 200, 15 min for a sample size of 500 and 55 min for a sample size of 1000. There is no obvious difference in computing time between the full bandwidth matrix sampler and the diagonal bandwidth matrix sampler (see Table 7).

We used the batch-mean standard error and the simulation inefficiency factor (SIF) to check the mixing performance of the sampling algorithm (see, for example, Roberts, 1996; Kim et al., 1998; Tse et al., 2004). We use $f_D(\cdot)$ as an example to illustrate the mixing performance of the sampling algorithm. Table 1 presents a summary of MCMC outputs obtained through M_1 and M_6. Both SIF and the batch-mean standard error show that all the simulated chains have mixed very well. We found a similar mixing performance for the other sampling algorithms, and for the other data sets.

We examined the robustness of the results to prior choices by trying values of $\lambda = 0.1$ and 5, as well as $\lambda = 1$. The mixing performance and posterior mean of each sampler was similar in all cases.

3.4. Accuracy of MCMC bandwidth selectors

In order to estimate the Kullback–Leibler information, we generated $N = 100,000$ bivariate random vectors from the true density and calculated the estimated Kullback–Leibler information defined by (10), which is employed to measure the distance between the bivariate kernel density estimator and the corresponding true density. Table 2 presents the estimated Kullback–Leibler information for each density and each bandwidth selector. The simulation study reveals the following evidence.

- For data sets generated from f_C and f_D, the MCMC bandwidth selector performs better than the corresponding plug-in bandwidth selector; for data sets generated from f_A, both selectors have a similar performance; for data sets generated from f_B, the MCMC

Table 1
MCMC results for data generated from $f_D(\cdot)$

Sample size	Bandwidths	Mean	Standard deviation	Batch-mean standard error	SIF	Acceptance rate
200	$1/b_{11}$	0.70	0.08	0.0017	10.32	0.224
	$1/b_{22}$	0.75	0.07	0.0015	11.77	
500	$1/b_{11}$	0.68	0.05	0.0011	11.72	0.207
	$1/b_{22}$	0.66	0.05	0.0009	8.73	
1000	$1/b_{11}$	0.69	0.03	0.0006	9.83	0.216
	$1/b_{22}$	0.61	0.03	0.0007	11.65	
200	b_{11}	1.18	0.15	0.0035	14.48	0.245
	b_{21}	−1.38	0.34	0.0164	57.58	
	b_{22}	1.69	0.21	0.0098	51.78	
500	b_{11}	1.10	0.08	0.0016	11.41	0.265
	b_{21}	−1.58	0.27	0.0137	65.54	
	b_{22}	1.91	0.19	0.1920	52.87	
1000	b_{11}	1.27	0.07	0.0015	11.68	0.267
	b_{21}	−0.79	0.11	0.0028	16.02	
	b_{22}	1.61	0.08	0.0016	9.45	

The first panel is obtained through the algorithm for a diagonal bandwidth matrix (M_6), while the second panel is obtained through the algorithm for a full bandwidth matrix (M_1).

bandwidth selector performs better than the plug-in bandwidth selector except when using a sphering transformation for a full bandwidth matrix.

- For each data set generated, the MCMC bandwidth selector performs better than the normal reference rule.
- The scaling transformation adds nothing to the performance of MCMC algorithms for sampling both diagonal and full bandwidth matrices.
- The sphering transformation of data is only helpful to the MCMC algorithm for sampling a diagonal bandwidth matrix when two variates are correlated, such as for densities B and D. For uncorrelated data, and for sampling a full bandwidth matrix, sphering can degrade performance. This is also supported by Wand and Jones (1993).
- The MCMC algorithm for a diagonal bandwidth matrix applied after sphering does not perform quite as well as the full bandwidth approach. However, the simplicity of using a diagonal bandwidth matrix makes this an attractive approach, especially with high-dimensional data.

It seems reasonable to interpret the choice between diagonal and full bandwidth matrices as a bias-variance tradeoff between diagonal and full covariance matrices, because there are more parameters in a full bandwidth matrix than in a diagonal matrix. For high-dimensional data, one might prefer more biased, lower variance estimates of a diagonal matrix over less biased but highly variable estimates of a full bandwidth matrix. Also sample size has a role to play in this choice—the larger the sample, the greater the confidence we can have in estimating the full set of parameters. When sphering is necessary, we found that the

Table 2
Estimated Kullback–Leibler information for bivariate densities

	Sample size	Kullback–Leibler information										
		M_1	M_2	M_3	M_4	M_5	M_6	P_1	P_2	P_3	P_4	N_1
$\hat{E}(\ln f_A) = -3.099$	200	0.131	0.129	0.158	0.154	0.154	0.228	0.129	0.213	0.153	0.192	0.375
	500	0.074	0.075	0.091	0.094	0.094	0.150	0.075	0.124	0.093	0.112	0.284
	1000	0.042	0.042	0.054	0.058	0.058	0.095	0.040	0.067	0.056	0.067	0.235
$\hat{E}(\ln f_B) = -1.822$	200	0.032	0.032	0.053	0.089	0.089	0.037	0.100	0.050	0.119	0.105	0.114
	500	0.021	0.021	0.037	0.048	0.047	0.022	0.047	0.023	0.055	0.089	0.085
	1000	0.018	0.018	0.040	0.040	0.040	0.021	0.038	0.021	0.043	0.065	0.071
$\hat{E}(\ln f_C) = -3.072$	200	0.299	0.296	0.247	0.394	0.392	0.361	0.357	0.345	0.391	0.325	0.410
	500	0.121	0.121	0.129	0.226	0.226	0.220	0.223	0.197	0.263	0.230	0.327
	1000	0.084	0.084	0.101	0.161	0.161	0.140	0.144	0.135	0.187	0.163	0.255
$\hat{E}(\ln f_D) = -3.850$	200	0.256	0.254	0.281	0.260	0.260	0.258	0.487	0.417	0.488	0.268	0.461
	500	0.219	0.221	0.249	0.240	0.240	0.217	0.333	0.298	0.345	0.240	0.385
	1000	0.149	0.149	0.150	0.178	0.178	0.149	0.260	0.222	0.274	0.173	0.299

Table 3
Numerical mean and standard deviation of ISEs for $f_D(\cdot)$

Sample size	Mean			Standard deviation			Difference between ISEs		
	MCMC (1)	PI (2)	NRR (3)	MCMC	PI	NRR	(1)–(2)	(1)–(3)	(2)–(3)
200	0.0077	0.0092	0.0176	0.00199	0.00136	0.00097	−0.00152 (0.00177)	−0.00998 (0.00151)	−0.00847 (0.00085)
500	0.0065	0.0060	0.0149	0.00179	0.00085	0.00061	0.00047 (0.00155)	−0.00842 (0.00147)	−0.00889 (0.00058)
1000	0.0049	0.0041	0.0128	0.00123	0.00057	0.00045	0.00081 (0.00107)	−0.00789 (0.00099)	−0.00870 (0.00032)

'PI' refers to the plug-in method, and 'NRR' the normal reference rule. Values in parentheses are the corresponding standard deviations.

performance of a full bandwidth matrix is better than that of a diagonal bandwidth matrix (as indicated by f_B).

We also employed the MISE criterion to examine the performance of optimal bandwidths obtained through the MCMC algorithm, the bivariate plug-in algorithm and the normal reference rule. We computed numerical MISEs for algorithms M_6, P_4 and N_1 through 50 data sets of sample sizes 200, 500 and 1000, each of which was generated from $f_D(\cdot)$. Results are given in the second column of Table 3, which shows that M_6 performs slightly better than P_4 for sample size 200, and slightly poorer than P_4 for sample sizes 500 and 1000.

When one bandwidth selector has a lower MISE than another method, it is useful to look at the standard deviation of the ISE. M_6 has less bias and larger variations than P_4 for sample size 200, while for sample sizes 500 and 1000, both bias and variation of M_6 are larger than those of P_4. In addition, both bandwidth selectors have less bias and larger variations than the normal reference rule.

We also computed the average difference between the ISEs of any two bandwidth selectors. The difference in ISE between M_6 and P_4 is insignificant, but the difference in ISE between M_6 and N_1, as well as that between P_4 and N_1, are significant. Both M_6 and P_4 perform significantly better than N_1. Hence, the empirical experience shows that M_6 and P_4 have a similar accuracy while M_6 is more variable that P_4, and that both M_6 and P_4 are significantly less biased and more variable than N_1. As the computation of numerical MISE is time-consuming, we have not computed MISE for the other bandwidth selectors, and for data sets generated from the other densities.

4. Numerical studies with multivariate densities

In this section, we examine the accuracy of the MCMC approach in the general multivariate setting. Our examples use $d = 5$.

4.1. True densities and bandwidth selectors

We consider five target densities labelled E, F, G, H and I, respectively. Density E is a multivariate normal density with location parameter μ and variance–covariance matrix defined as

$$\Sigma = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}, \tag{12}$$

where $\rho = 0.9$ and $\mu = (2, 2, 2, 2, 2)'$. This density is produced by a first-order autoregressive linear time series model.

Density F is a mixture of two multivariate normal densities,

$$f_F\left(\boldsymbol{x}|\,\mu_1, \mu_2, \Sigma\right) = \tfrac{1}{2}\,\phi\left(\boldsymbol{x}|\,\mu_1, \Sigma\right) + \tfrac{1}{2}\,\phi\left(\boldsymbol{x}|\,\mu_2, \Sigma\right),$$

where $\mu_1 = (2, 2, 2, 2, 2)'$, $\mu_2 = (-1.5, -1.5, -1.5, -1.5, -1.5)'$ and Σ is the 5×5 identity matrix.

Density G is a mixture of two multivariate Student t densities,

$$f_G\left(\boldsymbol{x}|\,\mu_1, \mu_2, \Sigma, \nu\right) = \tfrac{1}{2}\,t_d\left(\boldsymbol{x}|\,\mu_1, \Sigma, \nu\right) + \tfrac{1}{2}\,t_d\left(\boldsymbol{x}|\,\mu_2, \Sigma, \nu\right),$$

with $t_d(\cdot)$ defined in (11), $\mu_1 = (2, 2, 2, 2, 2)'$, $\mu_2 = (-1.5, -1.5, -1.5, -1.5, -1.5)'$, Σ is the identity matrix, and $\nu = 3$.

Density H is the multivariate skew normal density,

$$f_H(\boldsymbol{x}\,|\,\mu, \Sigma, \alpha) = 2\phi(\boldsymbol{x}\,|\,\mu, \Sigma)\,\Phi\left(\alpha' w^{-1/2}(\boldsymbol{x} - \mu)\right),$$

where $\Phi(\cdot)$ is the cumulative density function of a standard multivariate normal distribution and w is a diagonal matrix with diagonal elements the same as those of Σ. To generate a set of data, we define these parameters as $\mu = (2, 2, 2, 2, 2)'$, Σ as (12) with $\rho = 0.9$ and skewness parameter vector $\alpha = (-0.5, -0.5, -0.5, -0.5, -0.5)'$.

Density I is the multivariate skew t density,

$$f_I(\boldsymbol{x}\,|\,\mu, \Sigma, \nu, \alpha) = 2t_d(\boldsymbol{x}\,|\,\mu, \Sigma, \nu)T_d(\tilde{\boldsymbol{x}}\,|\,\nu + d),$$

where $t_d(\cdot)$ is the multivariate t density defined in (11), $T_d(\cdot\ |\ \nu + d)$ is the cumulative density function of a multivariate t distribution with mean $\mathbf{0}$, identity dispersion matrix and degrees-of-freedom $\nu + d$, and

$$\tilde{\boldsymbol{x}} = \alpha' w^{-1/2}(\boldsymbol{x} - \mu)\left(\frac{\nu + d}{(\boldsymbol{x} - \mu)'\Sigma^{-1}(\boldsymbol{x} - \mu) + \nu}\right)^{1/2},$$

with w the diagonal matrix with diagonal elements the same as those of Σ.

Table 4
MCMC results for data generated from $f_E(\cdot)$ with sample size 1500

	Bandwidths	Mean	Standard deviation	Batch-mean standard error	SIF	Acceptance rate
Diagonal matrix	$1/b_{11}$	0.56	0.03	0.0009	21.85	0.250
	$1/b_{22}$	0.58	0.03	0.0009	24.34	
	$1/b_{33}$	0.56	0.03	0.0009	29.25	
	$1/b_{44}$	0.58	0.03	0.0010	36.42	
	$1/b_{55}$	0.58	0.03	0.0009	34.14	
Full matrix	b_{11}	1.81	0.10	0.0042	41.83	0.272
	b_{21}	−0.15	0.15	0.0106	130.54	
	b_{22}	1.73	0.09	0.0033	36.26	
	b_{31}	0.11	0.18	0.0143	155.34	
	b_{32}	−0.15	0.13	0.0076	85.27	
	b_{33}	1.80	0.10	0.0031	25.31	
	b_{41}	−0.12	0.14	0.0084	93.56	
	b_{42}	−0.09	0.14	0.0099	133.07	
	b_{43}	−0.02	0.14	0.0083	93.30	
	b_{44}	1.74	0.10	0.0041	46.56	
	b_{51}	0.00	0.14	0.0084	88.95	
	b_{52}	0.07	0.14	0.0098	120.43	
	b_{53}	0.05	0.16	0.0114	134.69	
	b_{54}	0.18	0.13	0.0087	103.13	
	b_{55}	1.78	0.10	0.0042	47.31	

From each of the proposed multivariate densities, we generated data sets of sizes 500, 1000 and 1500. Then we applied the proposed MCMC algorithms to each data set to estimate the optimal bandwidth, where the multivariate standard Gaussian kernel is used. As the normal reference rule discussed in Scott (1992) and Bowman and Azzalini (1997) is the only viable alternative, we shall compare the performance of MCMC bandwidth selectors M_1 to M_6 with that of the alternative bandwidth selector N_1. The MCMC algorithm and parameter settings are the same as those in bivariate examples.

4.2. MCMC outputs and sensitivity analysis

Table 4 shows MCMC output obtained from $f_E(\cdot)$ with size 1500 to illustrate the mixing performance of the sampling algorithm. Both the batch-mean standard error and SIF show that all the sampled chains have mixed very well.

The numerical study shows that all algorithms for a diagonal bandwidth matrix have a similar mixing performance, and that all algorithms for a full bandwidth matrix have a similar mixing performance. However, the algorithm for a diagonal bandwidth matrix usually has a better mixing performance than that for a full bandwidth matrix. Similar results were found for the other data sets. Again, we found that the MCMC results are insensitive to changes in λ.

Table 5
Estimated Kullback–Leibler information for multivariate densities

	Sample size	Kullback–Leibler information						
		M_1	M_2	M_3	M_4	M_5	M_6	N_1
$\hat{E}(\ln f_E) = -7.9283$	500	0.178	0.177	0.539	0.441	0.441	0.186	1.262
	1000	0.127	0.126	0.505	0.304	0.304	0.162	1.235
	1500	0.118	0.117	0.470	0.276	0.276	0.141	1.545
$\hat{E}(\ln f_F) = -7.7934$	500	0.224	0.224	0.548	0.223	0.223	0.381	1.772
	1000	0.148	0.148	0.438	0.144	0.144	0.303	1.604
	1500	0.152	0.151	0.402	0.149	0.149	0.291	1.571
$\hat{E}(\ln f_G) = -9.2232$	500	0.774	0.771	1.147	0.746	0.746	0.915	2.222
	1000	0.687	0.685	1.149	0.677	0.677	0.846	1.862
	1500	0.696	0.696	1.029	0.679	0.680	0.845	1.992
$\hat{E}(\ln f_H) = -7.5123$	500	0.182	0.180	0.668	0.335	0.334	0.206	1.319
	1000	0.141	0.140	0.466	0.272	0.272	0.153	1.112
	1500	0.127	0.126	0.423	0.242	0.242	0.148	1.100
$\hat{E}(\ln f_I) = -7.3760$	500	0.288	0.282	0.725	0.479	0.479	0.247	1.342
	1000	0.142	0.141	0.662	0.331	0.331	0.166	1.204
	1500	0.109	0.109	0.537	0.270	0.270	0.147	1.318

4.3. Accuracy of MCMC bandwidth selectors

To estimate the Kullback–Leibler information, we generated $N = 100,000$ random vectors from the true density and calculated the estimated Kullback–Leibler information defined by (10). Table 5 presents these results for each density and each bandwidth selector.

The simulation study reveals the following evidence. First, all MCMC bandwidth selectors perform much better than the normal reference rule. Second, the scaling transformation adds nothing to the performance of MCMC algorithms for either the diagonal or full matrices. Third, the sphering transformation of data is only useful for the diagonal bandwidth matrix when variables are correlated (such as with densities E, H and I). When there is no correlation, or with the full bandwidth matrix, sphering degrades performance.

As we did in the bivariate case, we employed the MISE criterion to compare the performance of optimal bandwidths obtained through the MCMC algorithm and the normal reference rule. We computed numerical MISEs for algorithms M_6 and N_1 through 50 data sets of sample size 500, 1000 and 1500, each of which was generated from $f_H(\cdot)$. The ISE obtained through M_6 is less than that obtained through N_1 for every data set. A summary of numerical ISEs is given in Table 6, which shows that the average difference between ISEs of M_6 and N_1 is highly significant. As the numerical MISE is computationally intensive, we have not computed MISEs for the other bandwidth selectors, and for data sets generated from the other densities.

The CPU time required by the sampling algorithm (under the same conditions described in Section 3.3) for a diagonal bandwidth matrix is 19 min for a sample size of 500, 77 min for

Table 6
Numerical MISEs for the five-dimension density $f_H(\cdot)$

Sample size	MISE		Difference between ISEs	
	MCMC	NRR	MCMC & NRR	Standard deviation
500	0.000195	0.000499	−0.000304	0.000023
1000	0.000144	0.000421	−0.000278	0.000015
1500	0.000125	0.000391	−0.000265	0.000008

Table 7
CPU time for samplers of diagonal and full bandwidths (in minutes)

Sample size	Dimension = 2		Dimension = 5	
	Diagonal matrix	Full matrix	Diagonal matrix	Full matrix
200	2	2	–	–
500	14	15	19	26
1000	54	56	77	102
1500	–	–	177	238

a sample size of 1000 and 177 min for a sample size of 1500. The computing time required by the sampling algorithm for a full bandwidth matrix is 26 min for a sample size of 500, 102 min for a sample size of 1000 and 238 min for a sample size of 1500 (see Table 7).

5. Applications of MCMC bandwidth selectors

5.1. An application to earthquake data

We now apply the methodology to a trivariate data set discussed in Scott (1992). These data represent the epicenters of 510 earthquake tremors that occurred beneath the Mt St Helens volcano in the two months leading up to its eruption in March 1982. The three variables represent latitude, longitude and log-depth below the surface. Scott (1992, plate 8) gave several contours of a kernel density estimate of these data, where the bandwidths appear to have been chosen subjectively. We repeat this plot with the optimal bandwidth computed through our method.

We used the MCMC algorithms M_1 and M_5 to obtain optimal bandwidths, where the hyperparameter $\lambda = 1$, the burn-in period consists of 5000 iterations, and the recorded period contains 25,000 iterations. Table 8 tabulates a summary of results. Both the batch-mean standard error and SIF show that all sampled chains have mixed very well.

Using the estimated diagonal bandwidth matrix, we computed a kernel density estimator. (The estimate using the full bandwidth matrix was almost identical in this case.) The 98% highest density region (Hyndman, 1996) is plotted in Fig. 2. The surface was computed using the algorithm of Amenta et al. (1998). Note that the detached shells represent outliers in the

Table 8
MCMC results obtained from the earthquake data

	Bandwidths	Mean	Standard deviation	Batch-mean standard error	SIF	Acceptance rate
Diagonal matrix	$1/b_{11}$	0.003	0.0001	0.000003	9.07	0.254
	$1/b_{22}$	0.003	0.0001	0.000003	12.60	
	$1/b_{33}$	0.715	0.0383	0.000873	12.96	
Full matrix	b_{11}	311.65	0.07	0.002	15.80	0.246
	b_{21}	101.53	0.10	0.005	62.21	
	b_{22}	388.57	0.10	0.003	15.84	
	b_{31}	147.45	0.13	0.008	89.38	
	b_{32}	97.21	0.16	0.011	118.86	
	b_{33}	1.65	0.27	0.012	47.54	

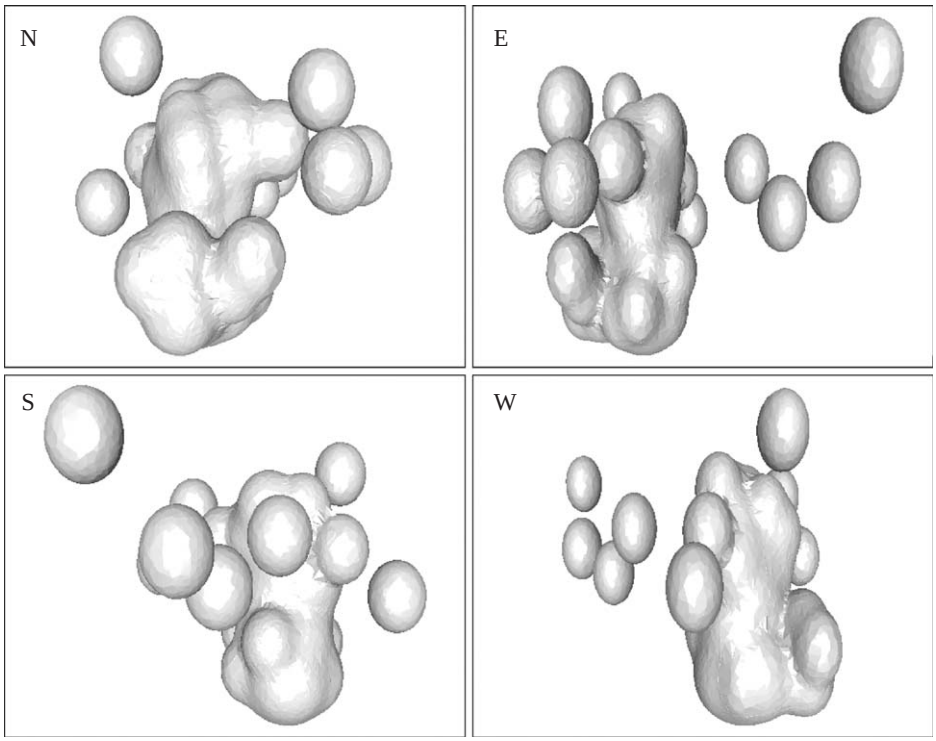

Fig. 2. The 98% highest density region for the earthquake data showing four views looking from north, east, south and west. Negative log-depth is on the vertical axis, and various combinations of latitude and longitude are on the horizontal axes.

data; the large central shell represents the bulk of the epicenters. The figure clearly shows clustering of the epicenters, revealing structure that was not discovered by Scott (1992) using a subjective bandwidth. It would be interesting to identify the clusters with geological

features, although this information is not available to us. As the plug-in bandwidth selectors are only applicable to bivariate data, we cannot obtain an optimal bandwidth through this method for comparison purposes.

5.2. Bandwidth selection for a Monte Carlo kernel likelihood

A difficulty for likelihood-based analysis such as maximum likelihood estimation and likelihood ratio testing with state-space models is that likelihood calculations require a high-dimensional integration of state variables. Let $\mathbf{y}$ denote a vector of observations and θ a parameter vector. Let $\theta^{(j)}$ denote the jth recorded draw of θ during MCMC iterations, for $j = 1, 2, \ldots, m$. de Valpine (2004) presented a Monte Carlo kernel likelihood (MCKL), which is an importance-sampled kernel estimator of the likelihood (up to a normalizing constant)

$$\hat{L}_H(\theta) = \frac{1}{m}\sum_{j=1}^{m} K_H\left(\theta - \theta^{(j)}\right)\frac{1}{p\left(\theta^{(j)}\right)}, \tag{13}$$

where $K_H(x) = |H|^{-1/2}K\left(H^{-1/2}x\right)$ with $K(\cdot)$ being a multivariate kernel function, H a symmetric positive definite matrix, and $p(\cdot)$ the prior of θ. The maximum likelihood estimate (MLE) of θ can be obtained by maximizing $\hat{L}_H(\theta)$ with respect to θ.

The MCKL involves selecting a bandwidth (chosen subjectively in de Valpine, 2004) for the posterior sample $\left\{\theta^{(j)} : j = 1, 2, \ldots, m\right\}$. de Valpine (2004) indicated that further work on automated bandwidth selection would facilitate the application of MCKL. To obtain an optimal bandwidth, we can use bandwidth selectors proposed in Section 2.

Consider the stochastic volatility (SV) model (Jacquier et al., 2004)

$$\begin{aligned} y_t &= \exp(\alpha_t/2)\,\lambda_t^{1/2}\varepsilon_t, \\ \alpha_{t+1} &= \mu + \phi(\alpha_t - \mu) + \sigma u_{t+1}, \end{aligned} \tag{14}$$

where $\varepsilon_t \sim IN(0, 1)$, $u_{t+1} \sim IN(0, 1)$, $cov(\varepsilon_t, u_{t+1}) = \rho$ and $\lambda_t \sim IG(\nu/2, \nu/2)$, which is equivalent to the fact that ν/λ_t follows a χ^2 distribution with ν degrees of freedom, and the marginal distribution of $v_t = \sqrt{\lambda_t}\varepsilon_t$ is Student t with ν degrees of freedom. The parameter vector is $\theta=(\phi, \mu, \rho, \sigma, \nu)'$, and the data set consists of 1134 continuously compounded daily returns of the Dow Jones industrial average index from the 1st January 2000 to 30th June 2004, excluding weekends and holidays. We employed the sampling algorithm provided by Zhang and King (2004) to obtain a posterior sample of θ, as well as the posterior average of θ, which is $(19.8826, 0.1938, -0.4615, -0.3372, 0.9694)'$.

To derive the optimal bandwidth for the posterior sample, we employed the algorithm for sampling a diagonal bandwidth matrix with scaling transformation of data discussed in Section 2. The estimated bandwidth is $\boldsymbol{h} = (2.5114, 0.0136, 0.0444, 0.0834, 0.0039)'$ and is employed in the MCKL. The MLE of θ is $(21.4499, 0.2822, -0.6008, -0.3894, 0.864)'$, which was obtained by numerically maximizing the MCKL. Even though the normalizing constant of the MCKL is unknown, likelihood-based analysis can be conducted using the particle filter algorithm, which aims to approximate the likelihood at the MLE of θ (see,

for example, Kitagawa, 1996; Kim et al., 1998; Zhang and King, 2004). Using the particle filter, we found that the value of the likelihood computed at the MLE of θ is -1567.95.

The application of the normal reference rule to the same posterior sample resulted in a bandwidth vector of $(2.5599, 0.0116, 0.0346, 0.0701, 0.0044)'$, which led to a MLE of θ of $(21.4499, 0.2548, -0.5977, -0.3895, 0.8611)'$. When the likelihood was evaluated at this estimate using the particle filter, we obtained a likelihood value of -1570.28. Hence, the bandwidth obtained through our MCMC sampler produced a maximum likelihood estimate with a larger maximized likelihood than that calculated via the normal reference rule.

The application of our bandwidth selector to the MCKL indicates the strength of a computational approach to bandwidth selection for multivariate kernel density estimation, because it is much easier to numerically optimize an objective function than it is to work out the theoretical optimum in this case.

6. Conclusion

This paper presents MCMC algorithms to estimate the optimal bandwidth for multivariate kernel density estimation via the likelihood cross-validation criterion. This represents the first data-driven bandwidth selection method for density estimation with more than two variables. Our numerical studies show that the sampling algorithms have a very good performance in achieving convergence of the simulated Markov chains, and are insensitive to prior choices.

Under the Kullback–Leibler information criterion, we have found that the MCMC algorithm generally performs better than the bivariate plug-in algorithm of Duong and Hazelton (2003) and the normal reference rule discussed in Scott (1992) and Bowman and Azzalini (1997). Under the MISE criterion, the MCMC algorithm works as well as Duong and Hazelton's (2003) plug-in algorithm, and both algorithms are superior to the normal reference rule. Under both criteria, our sampling algorithm is superior to the normal reference rule for higher-dimensional data. Apart from its performance, the other great advantage of our sampling algorithm is that it is applicable to data of any dimension, although the computing time required does increase as the dimension of data increases. In addition, our bandwidth selector provides a data-driven method for the problem of choosing an automated bandwidth for the MCKL—identified by de Valpine (2004) as a gap in literature. The effectiveness of our bandwidth selector in this case has been illustrated through an empirical example.

Acknowledgements

We wish to thank the Editor, Associate Editor and referees for their very insightful comments that have substantially improved the paper. We extend our sincere thanks to Faming Liang for sharing his coding skills and resources, David Scott for providing the earthquake data, Tarn Duong and Martin Hazelton for providing their R library to compute bivariate plug-in bandwidths, and the Victorian Partnership for Advanced Computing for computational support. We thank Martin Hazelton, Gael Martin, Mervyn Silvapulle and Dabao

Zhang for helpful comments. The second author acknowledges support from the Australian Research Council. Any remaining errors are, of course, ours only.

References

Abramson, I., 1982. On bandwidth variation in kernel estimates—a square root law. Ann. Statist. 10, 1217–1223.
Aït-Sahalia, Y., 1996. Testing continuous-time models of the spot interest rate. Rev. Financial Stud. 9, 385–426.
Aït-Sahalia, Y., Lo, A.W., 1998. Nonparametric estimation of state-price densities implicit in financial asset prices. J. Finance 53, 499–547.
Amenta, N., Bern, M., Kamvysselis, M., 1998. A new Voronoi-based surface reconstruction algorithm. Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, pp. 415–421.
Azzalini, A., Capitanio, A., 1999. Statistical applications of the multivariate skew normal distribution. J. Roy. Statist. Soc. Ser. B 61, 579–602.
Azzalini, A., Capitanio, A., 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t*-distribution. J. Roy. Statist. Soc. Ser. B 66, 367–389.
Azzalini, A., Dalla Valle, A., 1996. The multivariate skew normal distribution. Biometrika 83, 715–726.
Bauwens, L., Lubrano, M., 1998. Bayesian inference on GARCH models using the Gibbs sampler. Econometrics J. 1, C23–C26.
Bowman, A.W., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis. Oxford University Press, London.
Brewer, M.J., 2000. A Bayesian model for local smoothing in kernel density estimation. Statist. Computing 10, 299–309.
de Valpine, P., 2004. Monte Carlo state-space likelihood by weighted posterior kernel density estimation. J. Amer. Statist. Assoc. 99, 523–536.
Donald, S.G., 1997. Inference concerning the number of factors in a multivariate non-parametric relationship. Econometrica 65, 103–131.
Duong, T., Hazelton, M.L., 2003. Plug-in bandwidth selectors for bivariate kernel density estimation. J. Nonparametric Statist. 15, 17–30.
Härdle, W., 1991. Smoothing Techniques with Implementation in S. Springer, New York.
Hyndman, R.J., 1996. Computing and graphing highest density regions. Amer. Statist. 50, 120–126.
Izenman, A.J., 1991. Recent developments in nonparametric density estimation. J. Amer. Statist. Assoc. 86, 205–224.
Jacquier, E., Polson, N.G., Rossi, P.E., 2004. Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. J. Econometrics 122, 185–212.
Jones, M.C., 2001. A skew *t* distribution. In: Charalambides, C.A., Koutras, M.V., Balakrishnan, N. (Eds.), Probability and Statistical Models with Applications: A Volume in Honor of Theophilos Cacoullos. Chapman & Hall, London, pp. 269–278.
Jones, M.C., Faddy, M.J., 2003. A skew extension of the *t*-distribution, with applications. J. Roy. Statist. Soc. Ser. B 66, 159–174.
Jones, M.C., Marron, J.S., Sheather, S.J., 1996. A brief survey of bandwidth selection for density estimation. J. Amer. Statist. Assoc. 91, 401–407.
Kim, S., Shephard, N., Chib, S., 1998. Stochastic volatility: likelihood inference and comparison with ARCH models. Rev. Econom. Stud. 65, 361–393.
Kitagawa, G., 1996. Monte Carlo filter and smoother for Gaussian nonlinear state space models. J. Comput. Graphical Statist. 5, 1–25.
Marron, J.S., 1987. A comparison of cross-validation techniques in density estimation. Ann. Statist. 15, 152–162.
Roberts, G.O., 1996. Markov chain concepts related to sampling algorithms. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, London, pp. 45–57.
Sain, S.R., Baggerly, K.A., Scott, D.W., 1994. Cross-validation of multivariate densities. J. Amer. Statist. Assoc. 89, 807–817.
Sain, S.R., Scott, D.W., 1996. On locally adaptive density estimation. J. Amer. Statist. Assoc. 91, 1525–1534.

Schuster, E.F., Gregory, C.G., 1981. On the nonconsistency of maximum likelihood non-parametric density estimators. In: Eddy, W.F. (Ed.), Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface. Springer, New York, pp. 295–298.

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

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York.

Stanton, R., 1997. A nonparametric model of term structure dynamics and the market price of interest rate risk. J. Finance 52, 1973–2002.

Tse, Y.K., Zhang, X., Yu, J., 2004. Estimation of hyperbolic diffusion with Markov chain Monte Carlo simulation. Quantitative Finance 4, 158–169.

van der Laan, M.J., Dudoit, S., Keles, S., 2004. Asymptotic optimality of likelihood-based cross-validation. Statist. Appl. Genetics Molec. Biol. 4 (1), Article 4.

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

Wand, M.P., Jones, M.C., 1994. Multivariate plug-in bandwidth selection. Comput. Statist. 9, 97–116.

Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman & Hall, London.

Zhang, X., King, M.L., 2004. Box-Cox stochastic volatility models with heavy tails and correlated errors. Mimeo, Monash University.