Bandwidth Selection for Multivariate Kernel Density Estimation Using MCMC

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

Department of Econometrics and Business Statistics

Monash University, Clayton, Victoria, 3800, Australia

Correspondence to: Max.King@buseco.monash.edu.au

July 2004

Abstract: Kernel density estimation for multivariate data is an important technique that has a wide range of applications in econometrics and finance. 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 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 errors (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 kernel density estimation with more than two variables, and the sampling algorithm involves no increased difficulty as the dimension of data increases.

Key Words: Bandwidth matrices; Cross-validation; Kullback-Leibler information; Mean integrated squared errors; Sampling algorithms.

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). 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 1993):

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

where $K_H(\boldsymbol{x}) = |H|^{-1/2}K(H^{-1/2}\boldsymbol{x})$, $K(\cdot)$ 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. 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, Marron and Sheather 1996 for surveys). However, the literature on bandwidth selection for multivariate data is very limited. Sain, Baggerly and Scott (1994) discussed the performance of bootstrap and cross-validation methods for bandwidth selection in multivariate density estimation and found that the complexity of finding an optimal bandwidth grows prohibitively as the dimension of data increases. Wand and Jones (1994) presented a less variable cross-validation algorithm using the plug-in method, which requires auxiliary smoothing parameters. The technology for choosing these auxiliary smoothing parameters is not well developed. Duong and Hazelton (2003) showed 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 for bandwidth selection for bivariate data. This plug-in method has the advantage that it always produces a finite bandwidth matrix and requires computation of fewer pilot bandwidths. However, it cannot be directly extended to the general multivariate setting.

When data are observed from the multivariate normal density and the diagonal bandwidth matrix, denoted by $H = \text{diagonal}(h_1, h_2, \cdots, h_d)$, 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.

In theory, the cross-validation criterion can be employed for estimating the optimal bandwidth for multivariate data. However, it usually involves a numerical optimization, which becomes increasingly difficult as the dimension of data increases. The MCMC approach avoids this problem.

To our knowledge, the only previous paper employing an MCMC 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, the 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.

To estimate the optimal bandwidth matrix for multivariate data, we treat the bandwidth matrix H as a parameter matrix. The posterior density of H can be obtained through the likelihood function derived via the likelihood cross-validation criterion. MCMC algorithms can be developed to sample H from its posterior, and the ergodic average or the posterior mean acts as an estimator of the optimal bandwidth matrix. One important advantage of the MCMC approach to bandwidth matrix selection 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 data increases.

In this paper, we present MCMC algorithms for estimating 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. Section 5 illustrates the use of the MCMC algorithm for bandwidth selection with an application to some earthquake data. We provide conclusions in Section 6.

2 MCMC for optimal bandwidth selection

2.1 Likelihood cross-validation

The 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 true density $f(\boldsymbol{x})$. In this case, the Kullback-Leibler information is defined as

$$d_{KL}(f, \hat{f}_H) = \int \log\left(\frac{f(\boldsymbol{x})}{\hat{f}_H(\boldsymbol{x})}\right) f(\boldsymbol{x}) d\boldsymbol{x} = \int \log[f(\boldsymbol{x})] f(\boldsymbol{x}) d\boldsymbol{x} - \int \log[\hat{f}_H(\boldsymbol{x})] f(\boldsymbol{x}) d\boldsymbol{x}\,, \qquad (1)$$

which is nonnegative. We want to find an optimal bandwidth that minimizes $d_{KL}(f, \hat{f}_H)$, or, equivalently, maximizes $\mathrm{E}\log[\hat{f}_H(\boldsymbol{x})] = \int \log[\hat{f}_H(\boldsymbol{x})] f(\boldsymbol{x}) d\boldsymbol{x}$, in which the logarithmic pseudo-likelihood function can be approximated by (see, for example, Härdle 1991)

$$L(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \mid H) = \sum_{i=1}^{n} \log \hat{f}_{H,i}(\boldsymbol{x}_i), \qquad (2)$$

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

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

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

2.2 Sampling a diagonal bandwidth matrix

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

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

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

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

We treat the bandwidth $\boldsymbol{h}$ as a vector of parameters, given which, the likelihood function of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ is

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

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

$$\pi(h_i | \lambda) \propto \frac{1}{1 + \lambda\, h_i^2}, \tag{3}$$

for $i = 1, 2, \cdots, d$, where λ is a hyperparameter controlling the shape of the prior density. Uniform priors of bandwidths are generally unsuitable, because the update of each h_i has a negligible effect when h_i is already very large. However, the prior of h_i defined in (3) can prevent the update of h_i from getting too large. In a different context, Bauwens and Lubrano (1998) used a similar prior for the degree-of-freedom parameter of the t-distribution. The purpose of such priors is to put a low prior probability on the "problematic" region in the parameter space, where the likelihood function is flat. The joint prior of $\boldsymbol{h}$, denoted by $\pi(\boldsymbol{h}|\lambda)$, is the product of marginal priors defined in (3). According to Bayes theorem, the logarithmic posterior of $\boldsymbol{h}$ is (up to an additive constant)

$$\pi(\boldsymbol{h} | \lambda, \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n) \propto \log \pi(\boldsymbol{h}|\lambda) + \sum_{i=1}^{n} \log \hat{f}_{\boldsymbol{h},i}(\boldsymbol{x}_i), \tag{4}$$

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.

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

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

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

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

We treat non-zero elements of the bandwidth matrix as parameters, whose posterior density can be obtained based on the likelihood function given in (2). We assume that the prior density of each non-zero component of B is (up to a normalizing constant)

$$\pi(b_{ij} \mid \lambda) \propto \frac{1}{1+\lambda\, b_{ij}^2} \tag{5}$$

for $j \leq i$ and $i = 1, 2, \ldots, d$. The joint prior of all elements of B is the product of marginal priors defined in (5). Using Bayes theorem, we can obtain the logarithmic posterior of B (up to an additive constant)

$$\pi(B \mid \lambda, \boldsymbol{x}) \propto \sum_{i=1}^{d}\sum_{j=1}^{i} \log \pi(b_{ij} \mid \lambda) + \sum_{i=1}^{n} \log \hat{f}_{B,i}(\boldsymbol{x}_i), \tag{6}$$

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

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

for $i = 1, 2, \ldots, n$, where $S_d = \text{diagonal}(s_1^2, s_2^2)$. 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, $\hat{H} = S^{1/2}\hat{H}^*(S^{1/2})'$ or $\hat{H} = S_d^{1/2}\hat{H}^* S_d^{1/2}$.

In contrast, the MCMC algorithm does not require such pre-transformations of data. However, if we choose to make 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 original data might produce a better performance than that obtained directly from the original data, because the sphering transformation is equivalent to pre-rotating 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.

The Kullback-Leibler information defined in (1) is the mean of $\log(f(\boldsymbol{x})/\hat{f}_H(\boldsymbol{x}))$ 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 $f(\boldsymbol{x})$, the Kullback-Leibler information can be estimated by

$$\hat{d}_{KL}(f, \hat{f}_H) = \frac{1}{N}\sum_{i=1}^{N} \log(f(\boldsymbol{x}_i)/\hat{f}_H(\boldsymbol{x}_i)). \tag{7}$$

3.1 True densities

We consider five target densities labelled A, B, C, D and E. Contour plots of these densities are shown in Figure 1.

Density A is bivariate normal with high correlation between two variates:

$$f_A(\boldsymbol{x} \mid \mu, \Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)'\Sigma^{-1}(\boldsymbol{x}-\mu)\right),$$

with mean μ and variance-covariance matrix Σ given by

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}.$$

Density B is a mixture of two bivariate normal densities, with high correlation and bimodality:

$$f_B(\boldsymbol{x} \mid \mu_1, \Sigma_1, \mu_2, \Sigma_2) = \frac{1}{2} f_A(\boldsymbol{x} \mid \mu_1, \Sigma_1) + \frac{1}{2} f_A(\boldsymbol{x} \mid \mu_2, \Sigma_2),$$

where

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

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

$$f_C(\boldsymbol{x} \mid \mu, \Sigma, \alpha) = 2\phi(\boldsymbol{x} \mid \mu, \Sigma)\, \Phi(\alpha' W^{-1/2}(\boldsymbol{x}-\mu)),$$

where $\phi(\cdot \mid \mu, \Sigma)$ is the bivariate normal density with mean μ and variance-covariance matrix Σ, $\Phi(\cdot)$ 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 recently been studied by Azzalini and Dalla Valle (1996), Azzalini and Capitanio (1999, 2003), Jones (2001) and Jones and Faddy (2003) among many others. Here α is the shape parameter capturing the skewness of the distribution. When $\alpha = 0$, the density f_C becomes the usual normal density. For the purpose of generating a set of data, we use the following parameters,

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

Density D is a mixture of two bivariate Student t densities

$$f_D(\boldsymbol{x} \mid \mu_1, \mu_2, \Sigma, \nu) = \frac{1}{2} t_d(\boldsymbol{x} \mid \mu_1, \Sigma, \nu) + \frac{1}{2} t_d(\boldsymbol{x} \mid \mu_2, \Sigma, \nu),$$

where

$$t_d(\boldsymbol{x} \mid \mu, \Sigma, \nu) = \frac{\Gamma((\nu + d)/2)}{(\nu\pi)^{d/2}\Gamma(\nu/2)|\Sigma|^{1/2}} \left[1 + \frac{1}{\nu}(\boldsymbol{x} - \mu)'\Sigma^{-1}(\boldsymbol{x} - \mu)\right]^{-(d+\nu)/2}, \tag{8}$$

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

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

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

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

$$f_E(\boldsymbol{x} \mid \mu_1, \mu_2, \Sigma, \nu) = \frac{1}{2} t_d(\boldsymbol{x} \mid \mu_1, \Sigma_1, \nu) + \frac{1}{2} t_d(\boldsymbol{x} \mid \mu_2, \Sigma_2, \nu),$$

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 bivariate Gaussian kernel function and bandwidth matrix selected through each of the following selectors.

M1: MCMC algorithm for full bandwidth matrix without pre-transformation of data;

M2: MCMC algorithm for full bandwidth matrix with scaling transformation of data;

M3: MCMC algorithm for full bandwidth matrix with sphering transformation of data;

M4: MCMC algorithm for diagonal bandwidth matrix without pre-transformation of data;

M5: MCMC algorithm for diagonal bandwidth matrix with scaling transformation of data;

M6: MCMC algorithm for diagonal bandwidth matrix with sphering transformation of data;

P1: Plug-in selector of full bandwidth matrix with scaling transformation of data;

P2: Plug-in selector of full bandwidth matrix with sphering transformation of data;

P3: Plug-in selector of diagonal bandwidth matrix with scaling transformation of data;

P4: Plug-in selector of diagonal bandwidth matrix with sphering transformation of data;

A1: 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 a discussion on the advantages of their plug-in algorithm.

3.3 MCMC outputs and sensitivity analysis

The hyperparameter of prior densities defined in (5) is initially set to $\lambda = 1$ 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 (6) using the random-walk Metropolis-Hastings algorithm, in which the proposal density is the multivariate standard normal density, and the tuning parameter is chosen so that the acceptance rate is between 0.2 and 0.3.

The burn-in period is set at 5,000 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 use the batch-mean standard error and the simulation inefficiency factor (SIF) to check the mixing performance of the sampling algorithm (see, for example, Kim, Shephard and Chib 1998; Tse, Zhang and Yu 2004). We use $f_E(\cdot)$ 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. To demonstrate the mixing performance of these samplers graphically, we plot the sampled paths, their associated autocorrelation functions and histograms in Figure 2, which also reveals that these simulated chains have mixed very well. We found 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 $\lambda = 5$, 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 generate $N = 100{,}000$ bivariate random vectors from the true density and calculate the estimated Kullback-Leibler information defined by (7), 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_D and f_E, the MCMC bandwidth selector performs better than the corresponding plug-in bandwidth selector; for data sets generated from f_B, both selectors have a similar performance; for data sets generated from f_A and f_C, the MCMC 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 A, C and E. 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.

We also employ 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 compute numerical MISEs for algorithms M_6, P_4 and A_1 through 50 data sets of sample sizes 200, 500 and 1000, each of which was generated from $f_E(\cdot)$. Results are given in 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. We also compute the average difference between integrated squared errors (ISE) of any two bandwidth selectors. The difference of ISEs between M_6 and P_4 is insignificant, but the difference of ISEs between M_6 and A_1, as well as that between P_4 and A_1, are significant. Both M_6 and P_4 perform significantly better than A_1.

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 F, G, H, I and J, respectively.

Density F is a multivariate normal density with location parameter μ and variance-covariance matrix defined as

$$\Sigma = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{d-1} \\ \rho & 1 & \rho & \cdots & \rho^{d-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{d-3} \\ \cdots & & & \cdots & \\ \rho^{d-1} & \rho^{d-2} & \rho^{d-3} & \cdots & 1 \end{pmatrix}, \tag{9}$$

where $\rho = 0.9$ and $\mu = (2, 2, 2, 2, 2)'$.

Density G is a mixture of two multivariate normal densities,

$$f_G(\boldsymbol{x} \mid \mu_1, \mu_2, \Sigma) = \frac{1}{2} f_A(\boldsymbol{x} \mid \mu_1, \Sigma) + \frac{1}{2} f_A(\boldsymbol{x} \mid \mu_2, \Sigma),$$

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

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

$$f_H(\boldsymbol{x} \mid \mu_1, \mu_2, \Sigma, \nu) = \frac{1}{2} t_d(\boldsymbol{x} \mid \mu_1, \Sigma, \nu) + \frac{1}{2} t_d(\boldsymbol{x} \mid \mu_2, \Sigma, \nu),$$

with $t_d(\cdot)$ defined in (8), $\mu_1 = (2, 2, 2, 2, 2)'$, $\mu_2 = (-1.5, -1.5, -1.5, -1.5, -1.5)'$, Σ is the identity matrix, and $\nu = 3$.

Density I is the multivariate skew normal density,

$$f_I(\boldsymbol{x} \mid \mu, \Sigma, \alpha) = 2\phi(\boldsymbol{x} \mid \mu, \Sigma)\, \Phi(\alpha' W^{-1/2}(\boldsymbol{x} - \mu)),$$

where $\phi(\cdot \mid \mu, \Sigma)$ is the multivariate normal density with location parameter μ and variance-covariance matrix Σ, $\Phi(\cdot)$ 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 $\mu = (2, 2, 2, 2, 2)'$, variance-covariance matrix

Σ defined in (9) with $\rho = 0.9$, and skewness parameter $\alpha = (-0.5, -0.5, -0.5, -0.5, -0.5)'$.

Density J is the multivariate skew t density,

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

where $t_d(\cdot)$ is the multivariate t density defined in (8), $T_d(\cdot \mid \nu + d)$ 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.

From each of the proposed multivariate densities, we generate data sets of sizes 500, 1000 and 1500. Then we apply 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 A_1. The MCMC algorithm and parameter settings are the same as in bivariate examples.

4.2 MCMC outputs and sensitivity analysis

Table 4 shows MCMC output obtained from $f_F(\cdot)$ 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. Using the output obtained through M_1, we plot the sampled paths, their associated autocorrelation functions and histograms in Figure 3, which shows that the simulated chains via this algorithm have mixed 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 with the other data sets. Again, we found that MCMC results are insensitive to changes in λ.

4.3 Accuracy of MCMC bandwidth selectors

To estimate the Kullback-Leibler information, we generate N =100,000 random vectors from the true density and calculate the estimated Kullback-Leibler information defined by (7). 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 F, I and J). When there is no correlation, or with the full bandwidth matrix, sphering degrades performance.

As we did in the bivariate case, we employ 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 A_1 through 50 data sets of sample size 500, 1000 and 1500, each of which was generated from $f_I(\cdot)$. The ISE obtained through M_6 is less than that obtained through A_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 A_1 is highly significant.

5 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) gives several contours of a kernel density estimate of these data, where the bandwidths appear to have been chosen subjectively. We repeat this plot, but using our optimal bandwidth methodology.

We use the MCMC algorithms M_1 and M_5 to obtain optimal bandwidths, where the hyperparameter $\lambda = 1$, the burn-in period consists of 5,000 iterations, and the recorded period contains 25,000 iterations. Table 7 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 compute 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 Figure 4. The surface was computed using the algorithm of Amenta, Bern and Kamvysselis (1998). Note that the detached shells represent outliers in the 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.

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 the superior performance, the other great advantage of our sampling algorithm is that it is applicable to higher dimensional data with no increase in the complexity, although the computing time required does increase.

Acknowledgements

We extend our sincere thanks to Faming Liang for sharing his coding skills and resources, David Scott for providing the earthquake data, and Tarn Duong and Martin Hazelton for providing their R library to compute bivariate plug-in bandwidths. We thank Martin Hazelton and Gael Martin for helpful comments.

References

Abramson, I. (1982), "On Bandwidth Variation in Kernel Estimates – A Square Root Law," *The Annals of Statistics*, 10, 1217-1223.

Aït-Sahalia, Y. (1996), "Testing Continuous-Time Models of the Spot Interest Rate," *Review of Financial Studies*, 9, 385-426.

Aït-Sahalia, Y., and Lo, A.W. (1998), "Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices," *The Journal of Finance*, 53, 499-547.

Amenta, N., Bern, M., and Kamvysselis, M. (1998) "A New Voronoi-based Surface Reconstruction Algorithm", *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, 415–421.

Azzalini, A., and Capitanio, A. (1999), "Statistical Applications of the Multivariate Skew Normal Distribution," *Journal of the Royal Statistical Society*, Series B, 61, 579-602.

Azzalini, A., and Capitanio, A. (2003), "Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew *t*-distribution," *Journal of the Royal Statistical Society*, Series B, 66, 367-389.

Azzalini, A., and Dalla Valle, A. (1996), "The Multivariate Skew Normal Distribution," *Biometrika*, 83, 715-726.

Bauwens, L., and Lubrano, M. (1998), "Bayesian Inference on GARCH Models Using the Gibbs Sampler," *Econometrics Journal*, 1, C23-C26.

Bowman, A.W., and Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis*, London: Oxford University Press.

Brewer, M.J. (2000), "A Bayesian Model for Local Smoothing in Kernel Density Estimation," *Statistics and Computing*, 10, 299-309.

Donald, S.G. (1997), "Inference Concerning the Number of Factors in a Multivariate Nonparametric Relationship," *Econometrica*, 65, 103-131.

Duong, T., and Hazelton, M.L. (2003), "Plug-in Bandwidth Selectors for Bivariate Kernel Density Estimation," *Journal of Nonparametric Statistics*, 15, 17-30.

Härdle, W. (1991), *Smoothing Techniques with Implementation in S*, New York: Springer-Verlag.

Hyndman, R.J. (1996), "Computing and Graphing Highest Density Regions," *American Statistician*, 50, 120-126.

Izenman, A.J. (1991), "Recent Developments in Nonparametric Density Estimation," *Journal of the American Statistical Association*, 86, 205-224.

Jones, M.C. (2001), "A Skew *t* Distribution," in *Probability and Statistical Models with Applications: A Volume in Honor of Theophilos Cacoullos*," eds. C.A. Charalambides, M.V. Koutras, and N. Balakrishnan, London: Chapman & Hall, pp. 269-278.

Jones, M.C., and Faddy, M.J. (2003), "A Skew Extension of the t-distribution, with Applications," *Journal of the Royal Statistical Society*, Series B, 66, 159-174.

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.

Kim, S., Shephard, N., and Chib, S. (1998), "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models," *Review of Economic Studies*, 65, 361-393.

Marron, J.S. (1987), "A Comparison of Cross-Validation Techniques in Density Estimation," *Annals of Statistics*, 15, 152-162.

Roberts, G.O. (1996), "Markov Chain Concepts Related to Sampling Algorithms," in *Markov Chain Monte Carlo in Practice*, eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, London: Chapman & Hall, pp. 45-57.

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.

Sain, S.R., and Scott, D.W. (1996), "On Locally Adaptive Density Estimation," *Journal of the American Statistical Association*, 91, 1525-1534.

Schuster, E.F., and Gregory C.G. (1981), "On the Nonconsistency of Maximum Likelihood Nonparametric Density Estimators," in *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, eds. W.F. Eddy, New York: Springer-Verlag, pp.295-298.

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

Simonoff, J.S. (1996), *Smoothing Methods in Statistics*, New York: Springer-Verlag.

Stanton, R. (1997), "A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk," *The Journal of Finance*, 52, 1973-2002.

Tse, Y.K., Zhang, X., and Yu, J. (2004), "Estimation of Hyperbolic Diffusion with Markov Chain Monte Carlo Simulation," *Quantitative Finance*, 4, 158-169.

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.

Wand, M.P., and Jones, M.C. (1994), "Multivariate Plug-in Bandwidth Selection," *Computational Statistics*, 9, 97-116.

Wand, M.P., and Jones, M.C. (1995), *Kernel Smoothing*, London: Chapman & Hall.

Table 1: *MCMC results for data generated from $f_E(\cdot)$. 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).*

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	

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	A_1
	200	0.046	0.046	0.066	0.101	0.101	0.045	0.117	0.044	0.120	0.203	0.205
$\hat{E}(\ln f_A) =$	500	0.025	0.024	0.041	0.051	0.051	0.026	0.054	0.030	0.057	0.132	0.134
-2.003	1000	0.017	0.017	0.018	0.038	0.038	0.020	0.043	0.019	0.047	0.094	0.096
	200	0.131	0.129	0.158	0.154	0.154	0.228	0.129	0.213	0.153	0.192	0.375
$\hat{E}(\ln f_B) =$	500	0.074	0.075	0.091	0.094	0.094	0.150	0.075	0.124	0.093	0.112	0.284
-3.099	1000	0.042	0.042	0.054	0.058	0.058	0.095	0.040	0.067	0.056	0.067	0.235
	200	0.032	0.032	0.053	0.089	0.089	0.037	0.100	0.050	0.119	0.105	0.114
$\hat{E}(\ln f_C) =$	500	0.021	0.021	0.037	0.048	0.047	0.022	0.047	0.023	0.055	0.089	0.085
-1.822	1000	0.018	0.018	0.040	0.040	0.040	0.021	0.038	0.021	0.043	0.065	0.071
	200	0.299	0.296	0.247	0.394	0.392	0.361	0.357	0.345	0.391	0.325	0.410
$\hat{E}(\ln f_D) =$	500	0.121	0.121	0.129	0.226	0.226	0.220	0.223	0.197	0.263	0.230	0.327
-3.072	1000	0.084	0.084	0.101	0.161	0.161	0.140	0.144	0.135	0.187	0.163	0.255
	200	0.256	0.254	0.281	0.260	0.260	0.258	0.487	0.417	0.488	0.268	0.461
$\hat{E}(\ln f_E) =$	500	0.219	0.221	0.249	0.240	0.240	0.217	0.333	0.298	0.345	0.240	0.385
-3.850	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 MISEs for the bivariate density $f_E(\cdot)$. "PI" refers to the plug-in method, and "NRR" the normal reference rule. Values in parentheses are the corresponding standard deviations.*

sample	MISE			difference between ISEs		
size	MCMC	PI	NRR	MCMC & PI	MCMC & NRR	PI & NRR
200	0.0077	0.0092	0.0176	-0.00152	-0.00998	-0.00847
				(0.00177)	(0.00151)	(0.00085)
500	0.0065	0.0060	0.0149	0.00047	-0.00842	-0.00889
				(0.00155)	(0.00147)	(0.00058)
1000	0.0049	0.0041	0.0128	0.00081	-0.00789	-0.00870
				(0.00107)	(0.00099)	(0.00032)

Table 4: *MCMC results for data generated from $f_F(\cdot)$. The sample size is 1500.*

	bandwidths	mean	standard deviation	batch-mean standard error	SIF	acceptance rate
diagonal	$1/b_{11}$	0.56	0.03	0.0009	21.85	0.250
matrix	$1/b_{22}$	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	b_{11}	1.81	0.10	0.0042	41.83	0.272
matrix	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	

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	A_1
	500	0.178	0.177	0.539	0.441	0.441	0.186	1.262
$\hat{E}(\ln f_F) =$	1000	0.127	0.126	0.505	0.304	0.304	0.162	1.235
-7.9283	1500	0.118	0.117	0.470	0.276	0.276	0.141	1.545
	500	0.224	0.224	0.548	0.223	0.223	0.381	1.772
$\hat{E}(\ln f_G) =$	1000	0.148	0.148	0.438	0.144	0.144	0.303	1.604
-7.7934	1500	0.152	0.151	0.402	0.149	0.149	0.291	1.571
	500	0.774	0.771	1.147	0.746	0.746	0.915	2.222
$\hat{E}(\ln f_H) =$	1000	0.687	0.685	1.149	0.677	0.677	0.846	1.862
-9.2232	1500	0.696	0.696	1.029	0.679	0.680	0.845	1.992
	500	0.182	0.180	0.668	0.335	0.334	0.206	1.319
$\hat{E}(\ln f_I) =$	1000	0.141	0.140	0.466	0.272	0.272	0.153	1.112
-7.5123	1500	0.127	0.126	0.423	0.242	0.242	0.148	1.100
	500	0.288	0.282	0.725	0.479	0.479	0.247	1.342
$\hat{E}(\ln f_J) =$	1000	0.142	0.141	0.662	0.331	0.331	0.166	1.204
-7.3760	1500	0.109	0.109	0.537	0.270	0.270	0.147	1.318

Table 6: *Numerical MISEs for the 5-dimension density $f_E(\cdot)$.*

sample size	MISE MCMC	NRR	difference between ISEs 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: *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	

Figure 1: *Contour graphs of the proposed bivariate densities.*

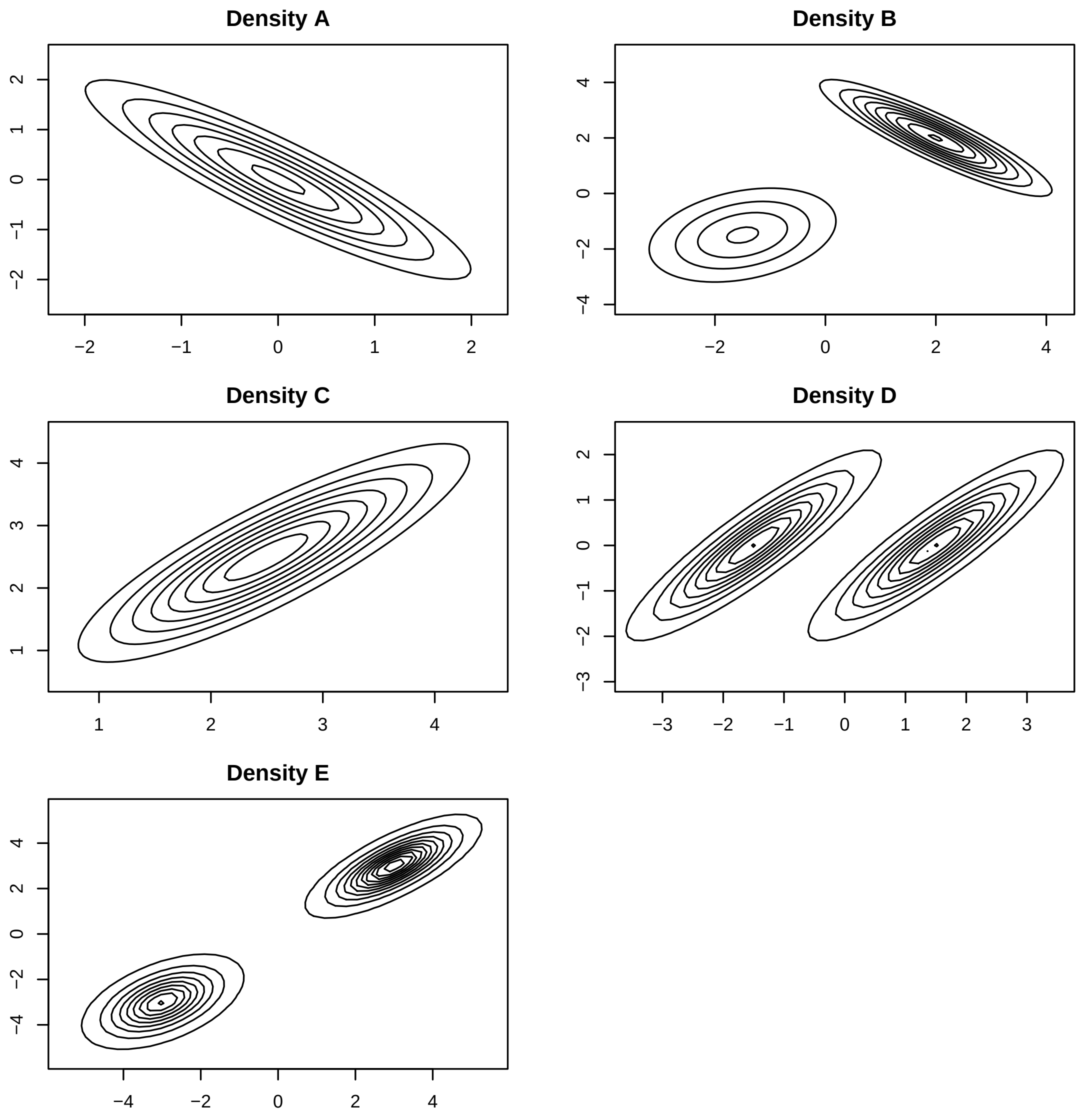

Figure 2: *MCMC results for the simulated bivariate data. Columns (from left to right) show the sampled paths, their associated autocorrelation functions and histograms. The first two rows represent nonzero elements of $H^{1/2}$, while the rest three rows represent elements of B.*

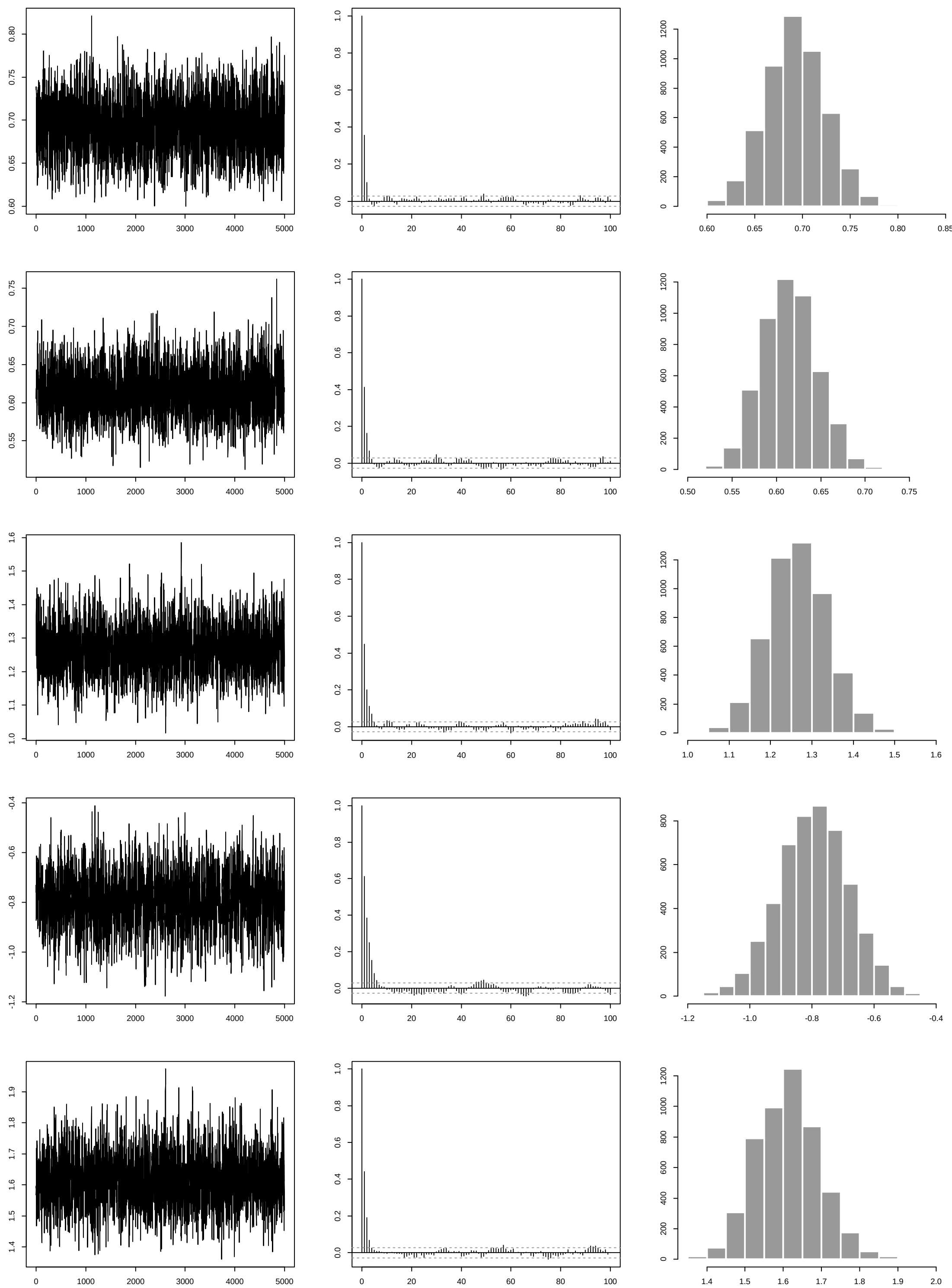

Figure 3: *MCMC results for the simulated multivariate data. Columns (from left to right) represent the sampled paths, their associated autocorrelation functions and histograms. Rows represent nonzero elements of* $H^{1/2}$.

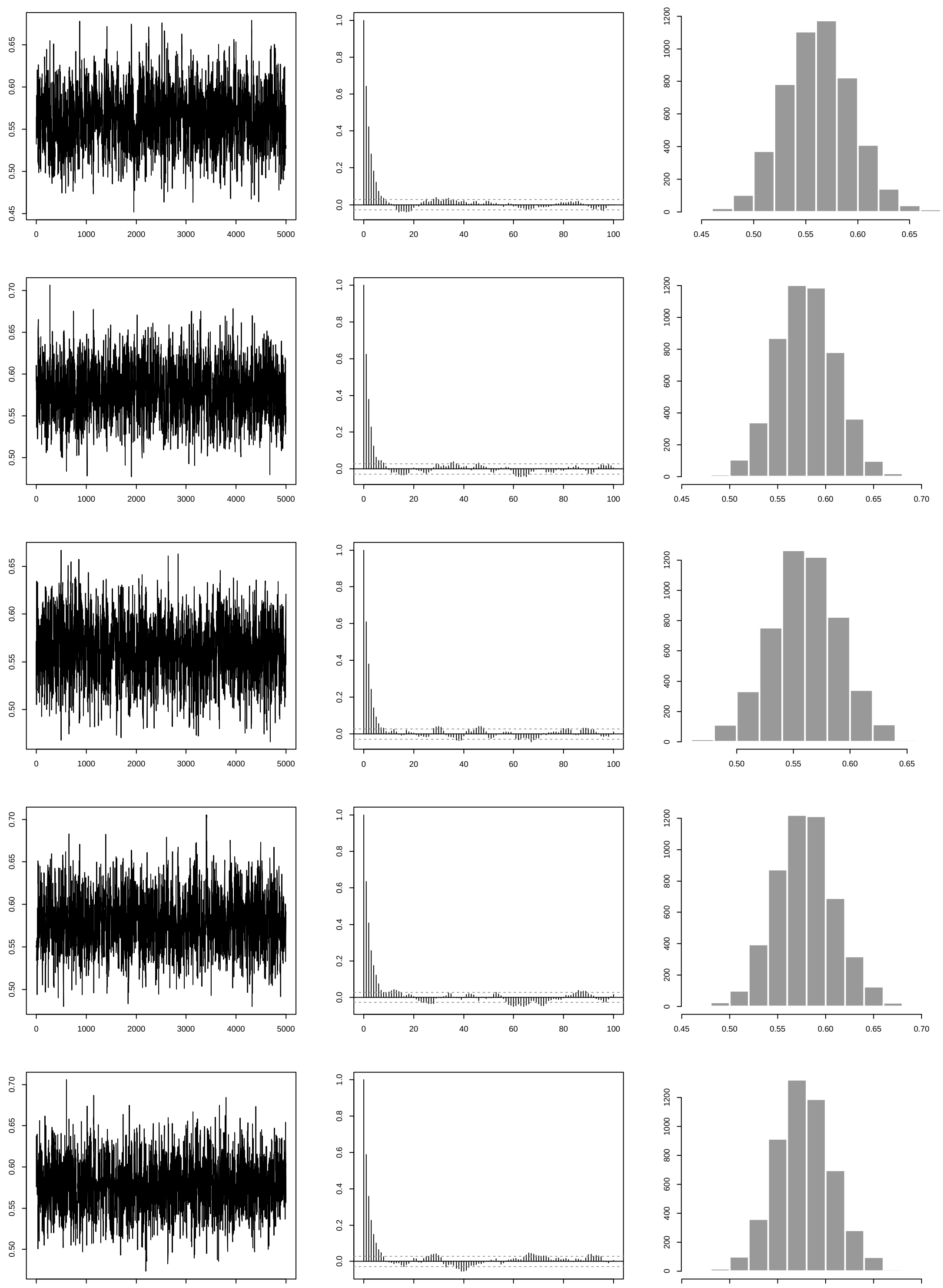

Figure 4: *The 98% highest density region for the earthquake data showing four views looking from north, east, south and west. Negative log-depth is on the vertical axis, and various combinations of latitude and longitude are on the horizontal axes.*