Flexible Multivariate Density Estimation With Marginal Adaptation

Paolo GIORDANI, Xiuyan MUN, Minh-Ngoc TRAN, and Robert KÖHN

This article is concerned with multivariate density estimation. We discuss deficiencies in two popular multivariate density estimators—mixture and copula estimators, and propose a new class of estimators that combines the advantages of both mixture and copula modeling, while being more robust to their weaknesses. Our method adapts any multivariate density estimator using information obtained by separately estimating the marginals. We propose two marginally adapted estimators based on a multivariate mixture of normals and a mixture of factor analyzers estimators. These estimators are implemented using computationally efficient split-and-elimination variational Bayes algorithms. It is shown through simulation and real-data examples that the marginally adapted estimators are capable of improving on their original estimators and compare favorably with other existing methods. Supplementary materials for this article are available online.

Key Words: Copula; Mixture of factor analyzers; Mixture of normals; Nonparametric; Variational Bayes.

1. INTRODUCTION

This article is concerned with estimation of multivariate densities. The goal of the article is to develop fast and flexible methods for constructing estimators of multivariate densities that are computationally reliable and fast, and that are statistically efficient, especially in high dimensions.

A common approach to multivariate density estimation is to use a parametric density such as a normal or a $t$. Such densities are relatively easy to estimate and there is extensive finite-sample inference available for them (e.g., see Anderson 2003). Estimation methods and inference for more general parametric densities such as symmetric and skew symmetric elliptic densities are also available (Genton 2004). An important advantage of parametric densities is that they can be applied to high-dimensional problems because of the relatively small number of parameters involved. However, the small number of parameters can also...
be a disadvantage if the data-generating process differs significantly from the parametric model. For example, if bivariate data are modeled as a normal distribution, then just one parameter (the correlation coefficient) is available to capture the dependence between its two marginals. Another drawback of this approach is that all the marginals are restricted to the same model (e.g., normal or \( t \)). Such a restriction may be unrealistic in real-data analysis.

To overcome the problems encountered using parametric methods, there is a large and growing literature on nonparametric density estimation. One popular approach is kernel density estimation. Sheather (2004) surveyed the univariate case and made it clear that the critical aspect of the method is the choice of smoothing or bandwidth parameter. We infer from Sheather’s article that it is challenging to successfully apply kernel density estimation in higher dimensions. A second approach is to use finite mixture models; for example, see Titterington, Smith, and Makov (1985) and McLachlan and Peel (2000). A related approach is to use Dirichlet process mixtures (e.g., see Neal 2000), which in many applications are equivalent to infinite mixtures of normals with constraints on the mixing probabilities. Mixture modeling has been shown successful in modeling complex and nonstandard data (Titterington, Smith, and Makov 1985; McLachlan and Peel 2000; Ghahramani and Beal 2000; McGrory and Titterington 2007). However, as will be demonstrated in Section 2, a potential problem with estimating densities by mixture models, especially in high dimensions, is that in finite samples, the implied model for each marginal may not be even close to the best model for that marginal.

Copula models such as the normal copula (NC) or \( t \) copula (tC) provide extra flexibility in multivariate density estimation by separately modeling the marginals and then linking them through a joint parametric model of dependence such as the multivariate normal or \( t \) distribution (Joe 1997; Nelsen 1999; Trivedi and Zimmer 2005). In general, however, using parametric copula models to capture dependence is likely to have the same disadvantages as those of conventional parametric models. Furthermore, as will be demonstrated in Section 2, another problem with copula modeling may occur when the transformed data through the marginal cumulative distribution function (cdf) transformation may become more complex than the original one.

Our article proposes a new class of multivariate density estimators that attempts to capture the advantages of copula estimators and mixture estimators, while being more robust to their weaknesses. Our approach takes any multivariate density estimator and modifies it to take into account the information obtained by separately estimating the marginal densities of the target. Necessary and sufficient conditions are given for such a marginally adapted multivariate density estimator to perform better in terms of the Kullback–Leibler (KL) divergence than the original estimator. In particular, our focus is on a marginally adapted multivariate mixture of normals (MN) (e.g., see McLachlan and Peel 2000) and a mixture of factor analyzers (MFA) (Ghahramani and Hinton 1997) estimators. These marginally adapted estimators can be expected to improve on the original MN and the original MFA estimators.

General alternative methods for modifying density estimators have been given by Hjort and Glad (1995) and Efron and Tibshirani (1996), who focused mainly on univariate density estimates. Our work has similar aims to those of Spiegelman and Park (2003), who modified kernel estimators to take into account parametric models for the marginals and discussed
problems with the Hjort and Glad (1995) approach. All of these methods are based on kernel density estimation, are quite different from our approach, and demonstrate the methodology on, at most, two-dimensional examples.

The MN and the MFA models are the basic building blocks of our estimators. They are not simple models to fit because the likelihood and posteriors can be multimodal and badly behaved. A commonly used method is using expectation–maximization (EM) algorithms for fitting and some model selection criterion, such as the Bayesian information criterion (BIC) or cross-validation (CV), for selecting the number of components/factors (McLachlan and Peel 2000; Ghahramani and Hinton 1997). Using CV may be too time-demanding and does not apply in time-series data analysis, while from our experience, BIC does not work well. Beal and Ghahramani (2006) also observed that BIC works poorly compared with a variational Bayes (VB) approach. However, they considered model selection in graphical rather than mixture modeling. We conjecture that there are at least two reasons for this. The first is that the Laplace approximation on which BIC is based becomes less effective as the dimension of the parameter space increases. The second is that the Laplace approximation often gives poor estimates in complex contexts, such as mixture models with multiple modes.

The second contribution of the article is to present and implement computationally efficient algorithms for fitting the MN and MFA models. We take a Bayesian approach and consider a VB approach for fitting. Using VB for fitting MN and MFA models has been previously considered by a number of authors; for example, see Corduneanu and Bishop (2001), McGrory and Titterington (2007), and Ghahramani and Beal (2000). Unlike EM algorithms, which only produce point estimates, VB can, to some extent, give an approximate distribution to the true posterior. Our VB algorithms, combined with the split-and-eliminate/merge idea (Ueda et al. 2000), can automatically determine the number of components and number of factors, besides overcoming the local maxima problem inherent in fitting mixture models. The MATLAB programs implementing these algorithms are included in the online supplementary materials.

The rest of this article is organized as follows. Section 2 presents the motivation and details of the marginal adaptation idea. Section 3 proposes two marginally adapted estimators based on MNs and MFAs. Sections 4 and 5 present simulation studies and real-data examples, respectively. The VB algorithms for fitting the mixture models are presented in Appendices A and B in the online supplementary materials.

2. MARGINALLY ADAPTED MULTIVARIATE DENSITIES

2.1 Motivation

When a multivariate density is estimated, its marginal densities are also implicitly estimated. A problem we observe is that the implied marginals may be poor estimates of the true marginals, especially in high dimensions. This happens when the model under consideration for the joint distribution is overparameterized in the sense that many valid constraints on the model parameters are not imposed. Implicit estimation of the marginals may then perform poorly when the degree of overparameterization is high. As an extreme example, consider a multivariate distribution with \( d \) independent marginals—each is a
Figure 1. Left: The Kullback–Leibler (KL) divergence between the true and the implicitly estimated marginals against dimension $d$. Middle and right: The difference between the KL divergence of the implicitly and directly estimated marginals and that of the true marginals when the true model is a normal copula and a mixture of three components, respectively. For each $d$, the KL values are averaged over 50 replications and the $d$ marginals.

univariate MN with $m$ components. Then, the joint distribution is a multivariate MN with $m^d$ components. In this case, a multivariate MN quickly becomes highly overparameterized as $d$ gets larger, because many constraints on the model parameters are not imposed. As a result, the implied marginals are poorly estimated.

Figure 1 summarizes the results of a simulation experiment to illustrate this issue further in a less extreme case. The data-generating process is a mixture of three normals, described in Example 2, Section 4. The left panel of the figure plots the KL divergence between the true and the implicitly estimated marginals against the dimension $d$, where the fitted densities are also MNs with three components. Here, the KL divergence values are averaged over the $d$ marginals and 50 replications. The plot shows that the implicitly estimated marginals get worse as the dimension increases.

The second observation we make is that implicit estimation of marginals may be in some cases practically less efficient than direct estimation, even when the true model is used to fit the joint distribution. The middle and right panels of Figure 1 plot the difference between the KL divergence of the implicitly and directly estimated marginals to that of the true marginals when the true model is an NC and a mixture of three normals (described in Examples 1 and 2, Section 4), respectively. The fitted models for the joint distribution and marginals are mixtures of three normals. We are simulating two cases: in the first case, the fitted model for the joint distribution (i.e., a mixture of three normals) is different from the true (an NC), while in the second case, the fitted model for the joint distribution is the same as the true (both are mixtures of three normals). A positive value of the difference means that the directly estimated marginals are closer to the true marginals than the implicitly estimated ones. The KL divergence values here are also averaged over the $d$ marginals and 50 replications. The plots show that the implicit estimation becomes less efficient as $d$ increases. We note that even when the fitted model for the joint distribution is the same as the true model, the implicit estimation of its marginals is still less efficient than the direct estimation when $d$ is sufficiently large. We conjecture that the large number of parameters
in the joint model that need to be estimated makes the estimation practically less efficient, while direct estimation of the marginals does not deteriorate with the dimension. However, note that this is not the case when \( d < 5 \) in our example. Sometimes, the implicit estimation can be more efficient. Typically, this happens if there is a high dependence between the marginals, \( d \) is sufficiently small, and we have sufficient data.

Copula models offer a way to overcome the problem of the deficiency in the implicit estimation of marginals by estimating the marginal and joint distributions separately (Joe 1997; Nelsen 1999). Suppose that the marginals are already given or separately estimated with (continuous) cumulative distribution functions (cdf) \( F_i \). If we wish to model the joint distribution of \( y \) by a copula implied by the joint distribution of another multivariate variable \( x \) with joint cdf \( H \) and \( i \)th marginal cdf \( H_i \), we first need to transform the original data via \( H_i^{-1} \circ F_i \), and then fit \( H \) to the transformed data. The hope is that the transformed data will be easier to model, which is typically the case in many cases. However, in some cases, such as with mixture data, the transformed data may become more complex than the original one.

To illustrate, consider data generated by a mixture of three well-separated bivariate normals

\[
p(y) = \frac{1}{3} \phi_2(y; 0, I) + \frac{1}{3} \phi_2(y; -51, I) + \frac{1}{3} \phi_2(y; 51, I),
\]

(1)

where \( 1 \) is a vector of ones. Suppose that we want to model the joint distribution by an NC. The first row of Figure 2 shows 1000 observations generated from this mixture model, together with the transformed data through an NC, that is, the \( H_i \) are standard normals. Clearly, the transformed data look much more difficult to model. The second row of Figure 2 plots 1000 observations generated from a scale mixture of two normals,

\[
p(y) = 0.6\phi_2(y; 0, I) + 0.4\phi_2(y; 0, 16I),
\]

(2)

together with the transformed data. Clearly, more than two components are needed to adequately capture the joint density in the transformed data. The simulation results in Section 4 confirm that for mixture data, all the copula estimators perform poorly.

In the following, we propose a new class of multivariate density estimators that attempts to capture the advantages of mixture estimators and copula estimators, while being more robust to their weaknesses.

### 2.2 Marginal Adaptation

Suppose that \( p(y) \) is the true \( d \)-dimensional multivariate density with marginals \( p_i(y_i), i = 1, \ldots, d \), that we wish to estimate and \( f(y) \) is an estimate of \( p(y) \) with implied marginal densities \( f_i(y_i), i = 1, \ldots, d \). For example, \( f(y) \) can be the density of an MN or an MFA. Assume that \( h_i(y_i) \) is an estimator of the \( i \)th marginal \( p_i(y_i) \), which may be estimated separately from \( f(y) \) and directly from the data. We assume that each \( h_i(y_i) \) is at least as good as the implied marginal \( f_i(y_i) \). Note that the implied marginal \( f_i(y_i) \) itself can be considered as a candidate in choosing a good estimator for the \( i \)th marginal. The selection can be based on some performance measurement, such as the log-predictive density score (LPDS) defined in Section 4.
The question addressed in this article is whether, and in what sense, it is possible to modify $f(y)$ by using the directly estimated marginals $h_i(y_i)$ to obtain a better estimator of $p(y)$.

To do so, we define

$$p_{f,h}(y) = k_{f,h}^{-1} f(y) \prod_{i=1}^{d} \frac{h_i(y_i)}{f_i(y_i)}$$

as a marginally adapted version of $f(y)$, where $k_{f,h}$ is a normalizing constant. To ensure that $p_{f,h}(y)$ is a density, we also assume for now that the ratios $h_i(y_i)/f_i(y_i)$ are bounded for $i = 1, \ldots, d$. This assumption is later relaxed.

We note at this stage that copula estimators are marginally adjusted by construction.

Lemma 1 gives necessary and sufficient conditions for $p_{f,h}(y)$ to be closer to the true density $p(y)$ than to $f(y)$ in the KL divergence, where the KL divergence between densities $p(y)$ and $f(y)$ is defined as

$$\text{KL}(p \parallel f) = \int p(y) \log \left( \frac{p(y)}{f(y)} \right) dy.$$
Lemma 1. Using the notation above, we have

\[
\text{KL}(p \| f) - \text{KL}(p \| pf,h) = -\log(k_{f,h}) + \sum_{i=1}^{d} (\text{KL}(p_i \| f_i) - \text{KL}(p_i \| h_i)).
\]  \(\text{(4)}\)

Proof.

\[
\text{KL}(p \| f) - \text{KL}(p \| pf,h) = \int p(y) \log \frac{pf,h(y)}{f(y)} dy
\]

\[
= -\log(k_{f,h}) + \sum_{i=1}^{d} \int p_i(y_i) \log \frac{h_i(y_i)}{f_i(y_i)} dy_i
\]

\[
= -\log(k_{f,h}) + \sum_{i=1}^{d} (\text{KL}(p_i \| f_i) - \text{KL}(p_i \| h_i)).
\]  \(\square\)

By Lemma 1, \(\text{KL}(p \| f) > \text{KL}(p \| pf,h)\), that is \(pf,h\) is closer to the true \(p\) than to \(f\) in the KL divergence if and only if

\[-\log(k_{f,h}) + \sum_{i=1}^{d} (\text{KL}(p_i \| f_i) - \text{KL}(p_i \| h_i)) > 0.\]  \(\text{(5)}\)

We call (5) the dominance condition, and the number on the left-hand side, the dominance condition value.

In general, the ratios \(h_i(y_i)/f_i(y_i)\) are not necessarily bounded. To ensure boundedness, we define

\[f_{i,\epsilon}(y_i) = (1 - \epsilon)f_i(y_i) + \epsilon h_i(y_i), \quad i = 1, \ldots, d,\]

with \(0 \leq \epsilon < 1\). Then, \(f_{i,\epsilon}(y_i)\) is a density, and \(h_i(y_i)/f_{i,\epsilon}(y_i) \leq 1/\epsilon\) for \(\epsilon > 0\). Let

\[p_{f,h,\epsilon}(y) = k_{f,h,\epsilon}^{-1} f(y) \prod_{i=1}^{d} \frac{h_i(y_i)}{f_{i,\epsilon}(y_i)},\]  \(\text{(6)}\)

where \(k_{f,h,\epsilon}\) is a normalizing constant that makes \(p_{h,\epsilon}(y)\) a density. Similarly to Lemma 1, we have the following lemma.

Lemma 2. The marginally adapted estimator \(p_{f,h,\epsilon}\) is better than the original estimator \(f\), that is,

\[\text{KL}(p \| f) > \text{KL}(p \| pf_{f,h,\epsilon}),\]

if and only if the following dominance condition holds

\[\text{DC} \equiv \text{DC}(f, h, \epsilon) := -\log(k_{f,h,\epsilon}) + \sum_{i=1}^{d} (\text{KL}(p_i \| f_{i,\epsilon}) - \text{KL}(p_i \| h_i)) > 0.\]  \(\text{(7)}\)
We note the following lemma, whose proof is straightforward.

**Lemma 3.** If $f$ is a consistent estimator of $p$ and the $h_i$ are consistent estimators of the $p_{i_1}, i = 1, \ldots, d$, then the corresponding marginally adapted estimator is a consistent estimator of $p$.

### 2.2.1 Estimating the Normalizing Constant

In general, there is no analytic expression for the normalizing constant $k_{f,h,\epsilon}$, but we can estimate it using importance sampling as

$$
\hat{k}_{f,h,\epsilon} = \frac{1}{M} \sum_{t=1}^{M} g(y_t), \quad \text{where} \quad g(y) = \frac{f(y)}{r(y)} \prod_{i=1}^{d} \frac{h_i(y_i)}{f_i(\epsilon)(y_i)}, \quad (8)
$$

$r(y)$ is the importance density, and $\{y_t, t = 1, \ldots, M\}$ is a sample from $r(y)$. The estimate (8) converges almost surely to the true value of $k_{f,h,\epsilon}$ as $M \to \infty$. Ideally, the importance density should be chosen so that it is straightforward to sample from and is a good approximation to the true density $p(y)$. If it is possible to simulate from $f(y)$, then $f(y)$ is a straightforward but not necessarily optimal choice for $r(y)$. In practice, the original estimator itself $f(y)$ or a tC estimator of $p(y)$ can be a good choice. Ultimately, the choice of $r(y)$ should be based on the variance of $\hat{k}_{f,h,\epsilon}$.

Assuming that $f(y)/r(y)$ is bounded, this makes $g(y)$ bounded by construction. In this case, $\hat{k}_{f,h,\epsilon}$ has the finite variance

$$
\text{var}(\hat{k}_{f,h,\epsilon}) = \frac{1}{M} \int (g(y) - k_{f,h,\epsilon})^2 r(y) dy,
$$

which can be estimated from the sample of $r(y)$ as

$$
\hat{\text{var}}(\hat{k}_{f,h,\epsilon}) = \frac{1}{M^2} \sum_{t=1}^{M} (g(y_t) - \hat{k}_{f,h,\epsilon})^2. \quad (9)
$$

Typically, the variance (9) increases when $\epsilon$ decreases. In practice, $\epsilon$ should be selected such that the variance (9) is reasonably small.

### 2.2.2 Sampling From the Marginally Adapted Estimator

Drawing from a marginally adapted distribution can be performed by independent Metropolis–Hastings sampling using a proposal density $r(y)$, which may be $f(y)$ or a tC approximation to $p(y)$.

### 3. MARGINALY ADAPTED MIXTURE ESTIMATORS

Our article considers two particular marginally adapted estimators: marginally adapted mixture of normals (MAMN) and marginally adapted mixture of factor analyzers (MAMFA).
3.1 Marginally Adapted Mixture of Normals (MAMN)

The MAMN specifies the original estimator \( f(y) \) as a multivariate MN density of the form

\[
    f(y) = \sum_{j=1}^{G} \pi_j \phi_d(y; \mu_j, V_j),
\]

(10)

where \( \pi_j \geq 0 \) are mixing probabilities, \( \pi_1 + \cdots + \pi_G = 1 \), and \( \phi_d(y; \mu, V) \) denotes the \( d \)-dimensional normal density with mean \( \mu \) and covariance matrix \( V \). Hereafter, an MN with \( G \) components will be denoted by MN(\( G \)). The marginal densities of \( f(y) \) are also MNs and therefore straightforward to compute. It is also straightforward to sample from \( f(y) \) for use in the importance sampling estimation in (8).

We fit the MN model (10) by a new split-and-elimination variational Bayes (SEVB) method, which is outlined in Appendix A. Our SEVB algorithm provides a very computationally efficient method for fitting MNs, in which the number of components is automatically determined. Fitting MNs with VB has been previously considered by a number of authors; for example, see Corduneanu and Bishop (2001) and McGrory and Titterington (2007). Our SEVB algorithm is in some cases more efficient than the existing ones in the sense that it can quickly remove inappropriate components and restore components when necessary. The details are documented in Appendix A.

3.2 Marginally Adapted Mixture of Factor Analyzers (MAMFA)

The MAMFA (Ghahramani and Hinton 1997) provides an effective way to parsimoniously model high-dimensional densities. MFA modeling has the advantages of flexibility from mixture modeling and dimensionality reduction from the factor representation. An MFA density has the form of (10), with the covariance matrices decomposed into \( V_j = \Lambda_j \Lambda_j' + \Psi_j \), where the \( \Lambda_j \) are loading matrices of order \( d \times k_j \) and the \( \Psi_j \) are diagonal matrices. Typically, the local dimensions \( k_j \) are much smaller than \( d \), which leads to a significant reduction in dimensionality of the unknown parameter space. However, fitting an MFA is challenging. We present in Appendix B an efficient SEVB algorithm, which is adapted from Ghahramani and Beal (2000), for fitting this model. Our algorithm automatically determines the number of components \( G \) as well as the local dimensions \( k_j \).

Note that once the MFA density \( f(y) \) is estimated, it is straightforward to compute the implied marginals \( f_i(y_i) \). We will denote by MFA(\( G; k_1, \ldots, k_G \)) an MFA model with \( G \) components and local dimensions \( k_1, \ldots, k_G \).

4. Simulation Studies

This section reports on some simulation experiments to compare the performance of our marginally adapted estimators with that of some other widely used multivariate density estimators, including an NC, a tC, an MN, and an MFA. We also report for the marginally adapted estimators estimates of the dominance condition value (7). We use univariate MNs for direct marginal estimation in our marginally adapted estimators and the copula estimators. We use the LPDS to measure the performance of the estimators. Let \( D_T \) be a
testing set; the LPDS of an estimator \( \hat{p} \) is defined as
\[
\text{LPDS}(\hat{p}) = \frac{1}{|D^T|} \sum_{y_i \in D^T} \log \hat{p}(y_i),
\]
(11)
where \(|D^T|\) is the number of observations in \( D^T \). Maximizing LPDS(\( \hat{p} \)) is equivalent to minimizing the KL divergence between the true density \( p(y) \) and the estimator \( \hat{p}(y) \). Therefore, the larger the LPDS, the better the estimator. In the simulation studies reported below, we consider several data-generating processes and the testing set \( D^T \) consists of 50,000 independent observations.

The number of samples \( M \) in the importance sampling estimation (8) of the normalizing constant is set to \( 10^5 \). We use the original estimator \( f(y) \) as the importance density in the following examples. Alternatively, a \( tC \) estimator can also be used. We use \( \epsilon = 0.05 \) and 0.5 in low- and higher-dimensional cases, respectively. This is to ensure that the importance sampling estimate is reliable. Indeed, the variances (9) in all the following examples have an order of \( 10^{-7} \) (details not shown), which shows that the estimation of the normalizing constant is reliable.

**Example 1 (NC data-generating process).** In the first simulation study of our proposed estimators, data are generated from an NC model. All marginals are the same, with each a univariate mixture of three normals with means \(-3, 0, 3\), variances \( 1, 2, 3 \), and mixing probabilities \( 0.5, 0.3, 0.2 \). The distribution of the underlying copula variable \( x \) is \( N_d(0, \Sigma) \), with \( \Sigma_{ij} = 0.5^{|i-j|} \).

We consider a simple low-dimensional case with \( d = 3 \) and a higher-dimensional case with \( d = 10 \). We try two sample sizes, \( n = 200 \) and \( n = 500 \), in each case. Table 1 summarizes the simulation results over 50 replications. Figure 3 shows the boxplots of the dominance condition values and the LPDS values for all four combinations of \( d \) and \( n \). The copula-based estimators are the best as the data are generated from a copula model. In all cases, the marginally adapted estimators significantly improve on their original estimators, especially in the higher-dimensional cases.

**Example 2 (MN data-generating process).** In this example, we generate data from a mixture of three normals with means \( \mu_1 = 0_{d \times 1}, \mu_2 = -31_{d \times 1}, \mu_3 = 31_{d \times 1} \), covariance matrices \( \Sigma_1 = I_d, \Sigma_2 = (0.5^{|i-j|})_{ij}, \Sigma_3 = ((-0.5)^{|i-j|})_{ij} \), and mixing probabilities

| \( d \) | \( n \) | DC | LPDS |
|---|---|---|---|
| 3 | 200 | 0.1112 | 0.1189 | \(-7.00\) | \(-7.00\) | \(-7.12\) | \(-7.16\) | \(-7.01\) | \(-7.04\) |
| 500 | 0.1225 | 0.1770 | \(-6.92\) | \(-6.92\) | \(-7.08\) | \(-7.15\) | \(-6.96\) | \(-6.97\) |
| 10 | 200 | 0.3839 | 0.3845 | \(-23.12\) | \(-23.12\) | \(-23.54\) | \(-23.53\) | \(-23.16\) | \(-23.15\) |
| 500 | 0.5661 | 0.5690 | \(-22.77\) | \(-22.79\) | \(-23.41\) | \(-23.42\) | \(-22.84\) | \(-22.85\) |

NOTE: In each case, the maximum LPDS is in bold.
Figure 3. The data-generating process is a normal copula: boxplots over 50 replications of the dominance condition values and the LPDS values for various estimators in four cases, with $d = 3$ (first row), $d = 10$ (second row) and $n = 200$ in the two left panels and $n = 500$ in the two right panels of each row.

The results are summarized in Table 2 and Figure 4. The copula-based estimators perform poorly, which confirms the discussion in Section 2.1. When $d$ is small ($d = 3$), as expected, the MN estimator is the best because the true model is an MN, and marginal adaptation does not improve on the MN. However, when $d$ is large ($d = 10$), the MN estimator is no longer the best even when the data-generating process is an MN. This confirms our discussion in Section 2.1 that implicitly estimated marginals may perform badly when the dimension $d$ is large. Marginal adaptation helps improve on the MN

Table 2. The data-generating process is a mixture of normals: the averaged values of the dominance condition values and the LPDS values for various estimators over 50 replications

|   | DC                  |          | LPDS                |
|---|---------------------|----------|---------------------|
|   | MAMN               | MAMFA    | NC                  | tC      | MN     | MFA    | MAMN   | MAMFA |
| 3 | 200                | −0.0065  | 0.1079              | −5.91   | −5.90  | −5.30  | −5.69  | −5.30  | −5.58  |
|   | 500                | −0.0047  | 0.1088              | −5.85   | −5.84  | −5.17  | −5.51  | −5.18  | −5.40  |
| 10| 200                | 0.0048   | 0.0374              | −17.38  | −17.29 | −16.21 | −16.13 | −16.21 | −16.09 |
|   | 500                | 0.1108   | 0.1229              | −17.25  | −17.17 | −15.20 | −15.66 | −15.09 | −15.54 |

NOTE: In each case, the maximum LPDS is in bold.
estimator in this case. The MAMFA always improves on the MFA, and it seems that the dimension reduction property of the MFA helps in higher-dimensional cases.

**Example 3 (A high-dimensional example).** We consider a high-dimensional example in which data are generated from the MN model as in Example 2 but with $d = 50$. Table 3 and Figure 5 summarize the results. As expected, the MFA and MAMFA estimators outperform the others in this example. Marginal adaptation helps improve on the MN and MFA estimators.

**Table 3.** High-dimensional example: the averaged values of the dominance condition values and the LPDS values for various estimators over 50 replications

| $n$  | MAMN | MAMFA | NC      | tC     | MN      | MFA    | MAMN   | MAMFA   |
|------|------|-------|---------|--------|---------|--------|--------|---------|
| 200  | 0.02 | 0.44  | −94.16  | −89.70 | −81.91  | −74.40 | −81.89 | −73.95  |
| 500  | 0.54 | 0.58  | −89.44  | −87.08 | −75.44  | −74.21 | −74.90 | −73.63  |

**NOTE:** In each case, the maximum LPDS is in bold.
5. REAL-DATA EXAMPLES

For the real-data examples considered in this section, we use cross-validated LPDS to measure performance. Suppose that the dataset $D$ is split into roughly $B$ equal parts $D_1, \ldots, D_B$, then the $B$-fold cross-validated LPDS is defined as

$$LPDS(\hat{p}) = \frac{1}{|D|} \sum_{j=1}^{B} \sum_{y_i \in D_j} \log \hat{p}(y_i|D \setminus D_j).$$

For direct marginal estimation, we consider four candidate estimators: a scale location $t$, a kernel density, a directly estimated MN, and the implicitly estimated MN itself. We select the best estimator among these four based on the cross-validated LPDS for use in our marginally adapted estimators and the copula estimators.

5.1 IRIS PLANTS

The Iris plants’ dataset consists of 150 observations of the length and width of the plants’ sepal and petals and consists of three varieties—*Iris setosa*, *Iris virginica*, and *Iris versicolor*—in equal proportions. This is a four-dimensional density estimation problem. The MN(2) and MFA(2;1,1) models are selected by our SEVB algorithm, with computation (CPU) time 0.3 and 1.2 sec, respectively, on an Intel Core i7 3.20-GHz desktop.

The best estimator for the first marginal is the implicitly estimated estimator, and the best estimators for the last three marginals are directly estimated MNs. After a few trials by checking the variance (9), we use $\epsilon = 0.1$ and $\epsilon = 0.5$ in the MAMN and MAMFA estimators, respectively. The normalizing constants of these two estimators are $1.0209 \times 10^{-6}$ and $1.0819 \times 10^{-6}$, respectively. The numbers in brackets are the order of the variance (9) in the importance sampling estimate (8).

The five-fold cross-validated LPDS values of the NC, tC, MN, MFA, MAMN, and MAMFA estimators are $-2.49, -2.50, -1.70, -2.74, -1.68$, and $-1.89$, respectively. The copula-based estimators perform poorly. We conjecture that this may be due to the
mixture structure of the data. The MAMN estimator improves on the MN estimator and outperforms the others. The MAMFA improves on the MFA estimator, although MFA does not work well in this low-dimensional example. The CPU time taken to do all the fitting and compute the LPDS values is 83 sec.

5.2 PERCENT BODY FAT DATA

Johnson (1996) introduced a dataset in which percent body fat and 13 simple body measurements (such as weight, height, abdomen circumference) were recorded for 252 men. After omitting observations 39 (because a weight value of 363.15 pounds is unusually large), 42 (because a height value of 29.5 inches is unreasonable), and 182 (because the percent body fat value is 0), we obtain a dataset of size 249. We are interested in estimating the multivariate distribution of these 14 variables.

The MN(1) and MFA(1;9) models are selected by our algorithm for the joint distribution, which suggests that this dataset has no mixture structure. The CPU times taken for fitting the MN and MFA estimators are 0.8 and 18 sec, respectively. The best models for the marginals vary between a kernel density and a directly estimated MN (details not shown). The normalizing constants of the MAMN and MAMFA estimators are $1.0576 \times 10^{-6}$ and $1.1128 \times 10^{-6}$, respectively. We use $\epsilon = 0.5$ in both the MAMN and the MAMFA estimators.

The LPDS values of the NC, tC, MN, MFA, MAMN, and MAMFA estimators are $-32.54$, $-32.16$, $-33.03$, $-33.22$, $-32.97$, and $-33.09$, respectively. The CPU time taken to do all the fitting and compute the LPDS values is 184 sec. The copula-based estimators work slightly better than the others. This result is consistent with the conclusion in Section 4 that when no more than one cluster is found in the data, copula-based estimators often work well.

5.3 PLASMODIUM GENE EXPRESSION LEVEL

Malaria is an infectious disease caused by the parasitic protozoan Plasmodium, which is a major concern in developing countries. The study of Plasmodium molecular biology is thus of great importance in developing an effective antimalaria treatment and vaccine strategy. Our example considers the relative expression level of 4221 genes taken at 46 time points over a 48-hr period in the life cycle of the parasite. The gene expression data have been further processed by Jasra et al. (2007) using $K$-means clustering and principal component analysis to reduce the number of observations from 4221 to 1000 and the number of variables from 46 to 6.

We fit our models to this dataset. The MN(4) and MFA(4;3,3,3,3) models are selected for the joint distribution. The CPU times taken for fitting these two estimators are 102 and 322 sec, respectively. The best estimators for the first three marginals are kernel densities, and for the last three, are a scale location $t$, a kernel density, and a directly estimated MN, respectively. The normalizing constants of the MAMN and MAMFA estimators are $1.0236 \times 10^{-6}$ and $1.0136 \times 10^{-6}$, respectively. The $\epsilon$ is set to 0.1 in both estimators. The five-fold cross-validated LPDS values of the NC, tC, MN, MFA, MAMN, and MAMFA
estimators are $-11.84, -11.74, -10.78, -11.65, -10.73,$ and $-11.49,$ respectively. The MAMN performs the best in this example. The CPU time taken is 2319 sec.

6. CONCLUSION

Our article proposes a methodology for adapting any multivariate density estimator by taking into account separate estimates of its marginal densities. These separate density estimates can be parametric or nonparametric. We consider, in particular, base estimators that are either mixtures of multivariate normal densities or mixtures of normal factor analyzers. We also propose a computationally efficient VB estimation methodology that automatically chooses the number of components for the MN estimators and the number of components and number of factors for the MFA estimators. Using simulated and empirical examples, we show that the new estimators often substantially improve on their base estimators and rarely perform much worse.

Based on our empirical results, we recommend fitting two estimators to each dataset, a marginally adapted mixture estimator and a copula-based estimator, and choosing between them based on an LPDS criterion. If only one estimator is to be computed in an applied problem, then we would recommend fitting a marginally adapted mixture estimator to the data if the data are multimodal, and a copula-based estimator, if they are not.

SUPPLEMENTARY MATERIALS

All the supplementary materials are contained in supp.zip.

Appendices: Appendices A and B (appendix.pdf) describe our VB fitting algorithms for the mixture of normals and mixture of factor analyzers models.

MATLAB programs: The MATLAB programs implementing the VB fitting algorithms are included in the subfolder. The subfolder also contains the datasets used in the article.

ACKNOWLEDGMENTS

The research of Xiuyan Mun, Minh-Ngoc Tran, and Robert Kohn was partially supported by the Australian Research Council grant DP0667069. The authors thank Professor Wenxin Jiang for the current form of Lemma 1 and Professor Jasra for the genome data.

[Received January 2012. Revised February 2012]

REFERENCES

Anderson, T. (2003), An Introduction to Multivariate Statistical Analysis (3rd ed.), New York: Wiley. [814]
Beal, M., and Ghahramani, Z. (2006), “Variational Bayesian Learning of Directed Graphical Models With Hidden Variables,” Bayesian Analysis, 1, 793–832. [816]
Corduneanu, A., and Bishop, C. (2001), “Variational Bayesian Model Selection for Mixture Distributions,” in Artificial Intelligence and Statistics 2001: Proceedings of the Eighth International Workshop (Vol. 14), eds. T. Jaakkola and T. Richardson, San Francisco, CA: Morgan Kaufmann, pp. 27–34. [816,822]
Efron, B., and Tibshirani, R. (1996), “Using Specially Designed Exponential Families for Density Estimation,” The Annals of Statistics, 24, 2431–2461. [815]
Genton, M. (ed.) (2004), *Skew-Elliptical Distributions and Their Applications*, New York: Chapman & Hall. [814]

Ghahramani, Z., and Beal, M. J. (2000), “Variational Inference for Bayesian Mixtures of Factor Analyzers,” in *Advances in Neural Information Processing Systems* (Vol. 12), eds. S. A. Solla, T. K. Leen, and K. R. Muller, Cambridge, MA: MIT Press, pp. 449–455. [815,816,822]

Ghahramani, Z., and Hinton, G. (1997), “The EM Algorithm for Mixtures of Factor Analyzers,” Technical Report, Department of Computer Science, University of Toronto. [815,816,822]

Hjort, N., and Glad, I. (1995), “Nonparametric Density Estimation With a Parametric Start,” *The Annals of Statistics*, 23, 882–904. [815]

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, London: Chapman & Hall. [815,818]

Johnson, R. W. (1996), “Fitting Percentage of Body Fat to Simple Body Measurements,” *Journal of Statistics Education* [on-line], 4 (1). Available at http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html. [827]

McGrory, C. A., and Titterington, D. M. (2007), “Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions,” *Computational Statistics and Data Analysis*, 51, 5352–5367. [815,816,822]

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley. [815,816]

Neal, R. M. (2000), “Markov Chain Sampling Methods for Dirichlet Process Mixture Models,” *Journal of Computational and Graphical Statistics*, 9, 249–265. [815]

Nelsen, R. (1999), *An Introduction to Copulas*, New York: Springer-Verlag. [815,818]

Sheather, S. (2004), “Density Estimation,” *Statistical Science*, 19, 588–597. [815]

Spiegelman, C., and Park, E. (2003), “Nonparametric Multivariate Density Estimates That Incorporate Marginal Parametric Density Information,” *The American Statistician*, 57, 183–188. [815]

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley. [815]

Trivedi, P., and Zimmer, D. (2005), “Copula Modelling: An Introduction for Practitioners,” *Foundations and Trends in Econometrics*, 1, 1–111. [815]

Ueda, N., Nakano, R., Ghahramani, Z., and Hinton, G. E. (2000), “SMEM Algorithm for Mixture Models,” *Neural Computation*, 12, 2109–2128. [816]