Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 6242–6251
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Clustering with mixtures of log-concave distributions☆

George T. Chang, Guenther Walther*

Department of Statistics, 390 Serra Mall, Stanford University, Stanford, CA 94305, USA

Received 10 August 2006; received in revised form 4 January 2007; accepted 6 January 2007
Available online 1 February 2007

Abstract

The EM algorithm is a popular tool for clustering observations via a parametric mixture model. Two disadvantages of this approach are that its success depends on the appropriateness of the assumed parametric model, and that each model requires a different implementation of the EM algorithm based on model-specific theoretical derivations. We show how this algorithm can be extended to work with the flexible, nonparametric class of log-concave component distributions. The advantages of the resulting algorithm are: first, it is not restricted to parametric models, so it no longer requires to specify such a model and its results are no longer sensitive to a misspecification thereof. Second, only one implementation of the algorithm is necessary. Furthermore, simulation studies based on the normal mixture model show that there seems to be no noticeable performance penalty of this more general nonparametric algorithm vis-a-vis the parametric EM algorithm in the special case where the assumed parametric model is indeed correct.
© 2007 Elsevier B.V. All rights reserved.

MSC: 62G07; 62G20; 62G35

Keywords: EM algorithm; Log-concave distribution; Clustering; Normal copula

1. Introduction

Clustering concerns the assignment of each of n observations $X_1, \ldots, X_n$ to one of k groups. One popular way to approach this task is via a finite mixture model, see e.g. McLachlan and Peel (2000): the data X_i are assumed i.i.d. with a density $f(x)$ that admits a representation

$$f(x) = \sum_{m=1}^{k} \pi_m f_m(x), \tag{1}$$

where the mixture proportions $\pi_1, \ldots, \pi_k$ are nonnegative and sum to unity, and the component distribution f_m models the conditional density of the data in the mth group. Typically one assumes a parametric formulation $f_m(x) = f(\theta_m, x)$ for the component distributions, such as a normal model, see e.g. Fraley and Raftery (2002). Then the fitting of the mixture model (1) as well as the assignment of the data to the k groups has an elegant solution in terms of the EM algorithm, see e.g. McLachlan and Krishnan (1997). The EM algorithm iteratively assigns the data based on the current

☆ Work supported by NSF Grant DMS-0505682 and NIH Grant 5R33HL068522.
* Corresponding author. Tel.: +1 650 723 3066; fax: +1 650 725 8977.
E-mail addresses: gtchang@stanford.edu (G.T. Chang), gwalther@stanford.edu (G. Walther).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.01.008

maximum likelihood estimates of the component distributions, and then updates those estimates $\hat{\pi}_m$, $\hat{\theta}_m$ based on these assignments.

One key advantage of using a mixture model for clustering is that it not only provides an assignment of the data to the k groups, but also a measure of uncertainty for the assignment of each observation via the posterior probabilities of component membership (see (2) in Section 3 below).

Problems arise when the parametric model is misspecified. Then the accuracy of the clustering may deteriorate, and the measure of uncertainty may be considerably off. In addition, each parametric model requires a different implementation of the EM algorithm based on attendant theoretical derivations. For these reasons, it would be helpful to have an EM-type clustering algorithm with nonparametric component distributions. Such a methodology would provide a universal software implementation with flexible component distributions. Indeed, nonparametric extensions of parametric models have proved quite successful in discriminant analysis, the supervised counterpart to the problem under consideration here, see e.g. Hastie and Tibshirani (1996) and Lin and Jeon (2003). In contrast, there seems to be little existing work on mixture models with nonparametric components for clustering, presumably because it is not obvious how to develop such methodology in the unsupervised case. Hunter et al. (2006) give methodology to estimate a location mixture of symmetric univariate components.

In this paper we will model each component as a log-concave density, i.e. as a density whose logarithm is a concave function. This model has the advantage that it includes most common parametric distributions (the prime example being the normal density, whose logarithm is a quadratic), and it is flexible enough to allow e.g. skewness. See e.g. Walther (2001, 2002) for a further discussion of this model. Moreover, it turns out that the MLE of a log-concave density exists uniquely, so there is a hope that one can mimic the EM-type clustering algorithm that works so successfully in the parametric context. During the preparation of this paper we became aware of the work of Eilers and Borgdorff (2006), who use penalized smoothing to move a nonparametric estimate of the component distribution 'towards' a log-concave form. This approach requires the choice of a tuning parameter. In contrast, we use the fact that the log-concave MLE exists uniquely and gives an algorithm for its computation. Thus, our approach is free of tuning parameters. In addition, we show how to generalize this approach to the multivariate situation where dependence is present in each component.

2. The univariate model and the MLE

Our model for the univariate case posits that each component f_m in (1) is a log-concave density, i.e. $\log f_m(x)$ is a concave function. An EM-type algorithm requires the computation of the MLE of each f_m. Theory and algorithms for this task have been developed in Walther (2002) and in Rufibach (2006). We briefly summarize the relevant results.

Given data $X_1, \ldots, X_n$ i.i.d. from f, the MLE $\hat{f}$ of f under the restriction that f be log-concave exists uniquely and has support $[X_{(1)}, X_{(n)}]$. $\log \hat{f}$ is a piecewise linear function whose knots are a subset of $\{X_1, \ldots, X_n\}$. The MLE can be computed e.g. using the Iterative Convex Minorant Algorithm described in Jongbloed (1998). The resulting algorithms for computing the log-concave MLE $\hat{f}$ as given in Walther (2002) and Rufibach (2006) provide as output $\hat{f}(X_i)$, $i = 1, \ldots, n$. This is all that is needed for an EM-type algorithm; of course one can easily compute the entire density $\hat{f}$ by linearly interpolating between $\log \hat{f}(X_{(i)})$ and $\log \hat{f}(X_{(i+1)})$ and then exponentiating. Further, as shown in Walther (2002) and Rufibach (2006), it is straightforward to incorporate weights for the data as is required for an EM-type algorithm.

3. Clustering with an EM-type algorithm

Our methodology is as follows: first we run the usual EM algorithm for a Gaussian mixture to convergence. We then use the outcome of this clustering as starting value for five more iterations of the EM algorithm, where in the M-step we now compute for each component the log-concave MLE instead of the Gaussian MLE.

The motivation for this approach is that the Gaussian mixture model should provide a clustering that is roughly correct, and that the subsequent log-concave MLE provides a correction by e.g. adjusting for skewness. Without a formal proof that this EM-type algorithm converges, we need to put a bound on the number of iterations. From our experience, the algorithm seems to converge quickly and five iterations are sufficient. Another advantage of running EM for a Gaussian mixture first is that the MLEs can be computed quickly in the M-step, whereas the computation of the log-concave MLEs is much more computer-intensive.

In more detail, the EM algorithm for a Gaussian mixture provides posterior probabilities that the ith observation belongs to the mth component

$$\tau_m(X_i) := \hat{\pi}_m \hat{f}_m(X_i) \Big/ \sum_{j=1}^{k} \hat{\pi}_j \hat{f}_j(X_i) \,, \tag{2}$$

where $\hat{f}_m$ is the Gaussian MLE for the mth component, $i = 1, \ldots, n$, $m = 1, \ldots, k$, see Chapter 2.7 in McLachlan and Krishnan (1997). In the second part of our algorithm, where we switch to the log-concave MLE, the M-step comprises the following computations: we use the $\tau_m(X_i)$ as weights for the X_i when we compute the log-concave MLE $\hat{f}_m$ for the mth component, and we set as usual $\hat{\pi}_m = \sum_{i=1}^{n} \tau_m(X_i)/n$, $m = 1, \ldots, k$. The E-step consists of the usual computation (2), where $\hat{f}_m$ is now the log-concave MLE from the preceding M-step, and the $\hat{\pi}_m$ are likewise obtained in the preceding M-step.

4. Comparison with parametric EM

We compared the result of our methodology with that obtained with the EM algorithm for the Gaussian mixture model.

In our first example we drew 500 observations from a gamma(2,1) distribution, then with probability 0.6 each observation was shifted to the right by 5. This mixture density is plotted in the left panel of Fig. 1 (dotted line). The dashed line gives the fitted model obtained by the Gaussian EM algorithm, and the solid line gives the fitted model obtained by the log-concave EM algorithm. One observes that the log-concave EM algorithm provides a noticeably better solution. While this mixture has homoscedastic components, neither of the two algorithms was restricted to that effect. Further details of the implementation are given in the Appendix.

We examined the quality of the resulting clustering as follows: for each simulated observation X_i we recorded from which component it came. After running the EM algorithm, we classified each observation according to the plug-in sample version of the Bayes (optimal) rule, i.e. via $\arg\max_m \tau_m(X_i)$. We then recorded the number of misclassified observations in each of 1000 sets of simulations. The left panel of Fig. 2 plots these 1000 numbers of misclassified observations for the log-concave EM algorithm vs. those obtained with the Gaussian EM algorithm. One sees that the log-concave EM algorithm provides a clear improvement.

As mentioned in the Introduction, a key advantage of a mixture model is that it provides a measure of uncertainty for the assignment of each observation via the posterior probabilities of component membership (2). Thus, as our second performance criterion, we investigate how well both algorithms estimate the corresponding population version:

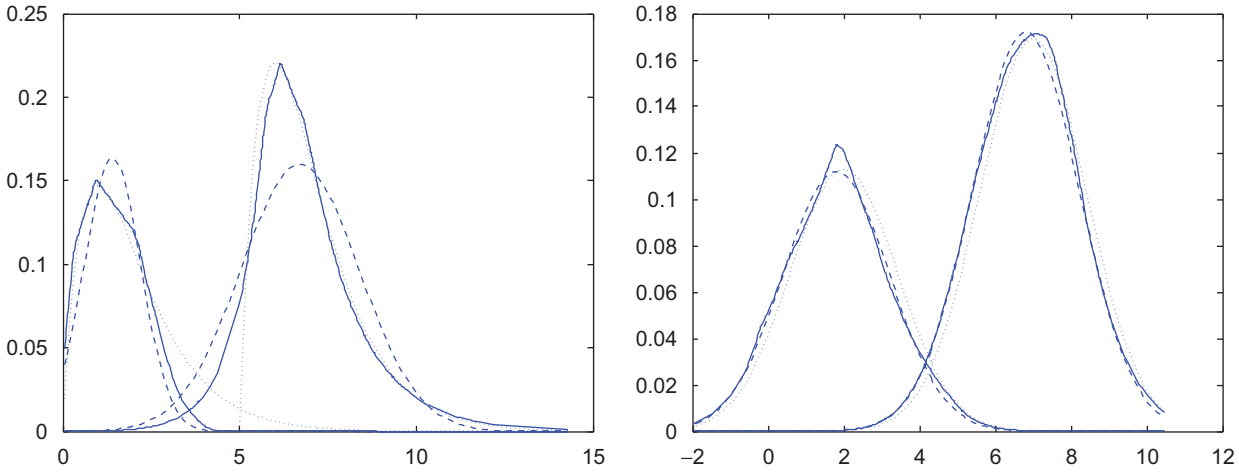

Fig. 1. Solutions of the Gaussian EM algorithm (dashed) and the log-concave EM algorithm (solid) for 500 observations from a location mixture (dotted) of gamma distributions (left) and normal distributions (right).

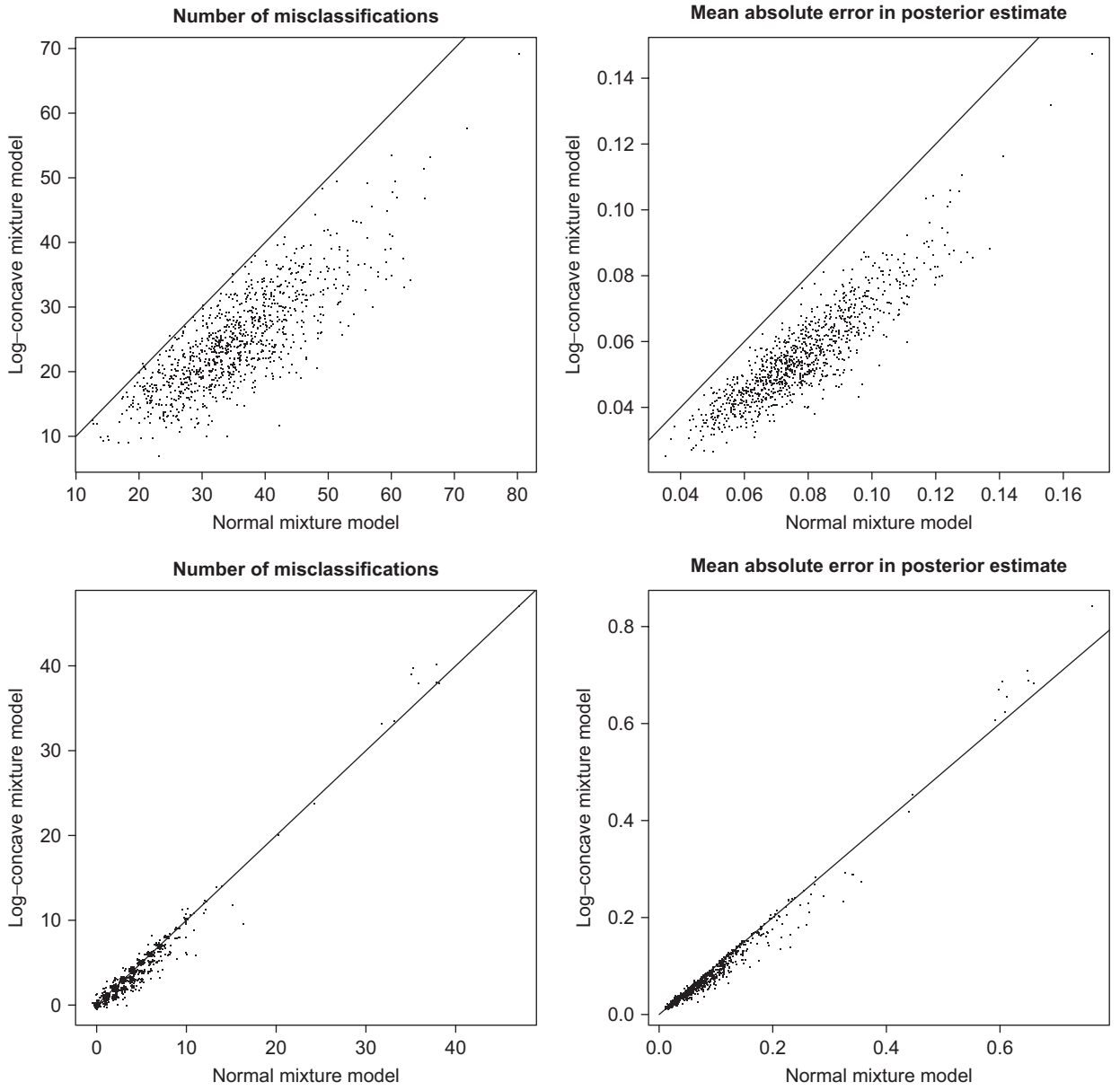

Fig. 2. Top left: number of observations misallocated by the log-concave EM algorithm vs. the Gaussian EM algorithm for 1000 sets of simulations from the gamma location mixture with sample size 500. Each plotted point represents the result of one simulation. Top right: the mean absolute errors of estimated membership probabilities for the 1000 sets of simulations. The solid line marks the identity. Jitter has been added in the left panels for better visualization. The bottom row shows the analogous results for sample size 50.

the right panel of Fig. 2 shows $1/n\sum_{i=1}^{n}|\tau_1(X_i) - t_1(X_i)|$, for both algorithms in each of 1000 simulations, where $t_m(X_i) := \pi_m f_m(X_i)/\sum_{j=1}^{2}\pi_j f_j(X_i)$. Again, the log-concave EM algorithm gives a clear improvement.

One expects that a nonparametric technique performs less favorably if the sample size is small. The bottom row of Fig. 2 repeats the above simulation study for a sample size of 50 to examine this effect. The log-concave EM algorithm still gives a better performance.

The improvements of the log-concave EM algorithm over the Gaussian EM algorithm are to be expected in this case, as the data come from a model that is not Gaussian. What price does one have to pay for the flexibility of the

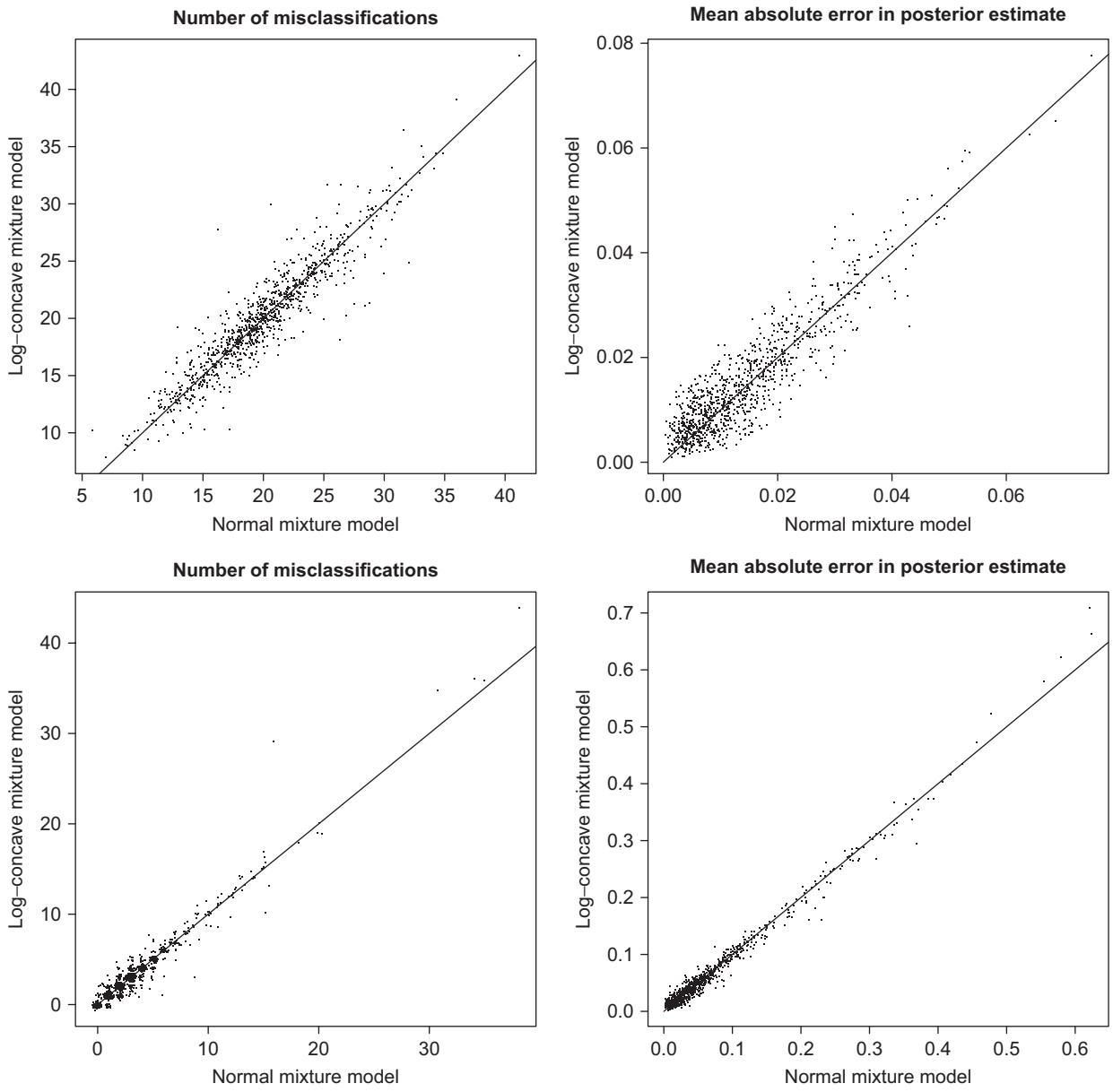

Fig. 3. Top left: number of observations misallocated by the log-concave EM algorithm vs. the Gaussian EM algorithm for 1000 sets of simulations from the normal mixture with sample size 500. Top right: the mean absolute errors of estimated membership probabilities for the 1000 sets of simulations. The solid line marks the identity. Jitter has been added in the left panels for better visualization. The bottom row shows the analogous results for sample size 50.

log-concave EM algorithm in the case where the data come in fact from a Gaussian location mixture, a set-up for which the Gaussian EM algorithm is tailor-made? We repeated the above simulation experiment, but now we drew the 500 observations from the model $0.4N(2, 2) + 0.6N(7, 2)$. The fitted models are given in the right panel of Fig. 1, and the accuracy of the clustering is given in Fig. 3, both for sample size of 500 and a sample size of 50. Those plots do not show any noticeable difference in performance between the two algorithms. We thus conclude that the log-concave EM-algorithm is a flexible, nonparametric tool that seems to give superior results in the large class of mixtures with log-concave components, without any noticeable penalty in the special case of a Gaussian mixture.

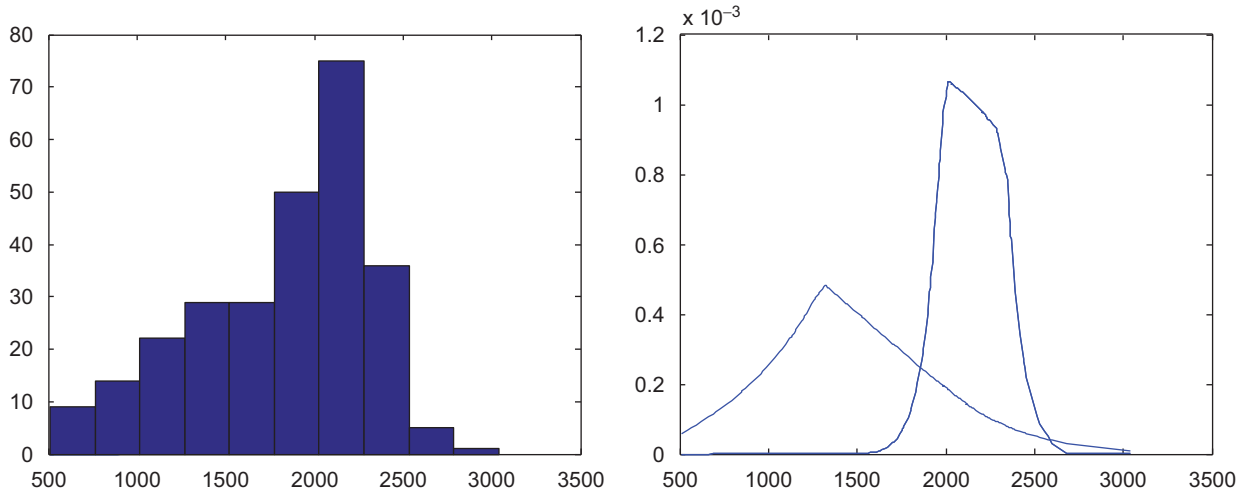

Fig. 4. Left: histogram of the flow cytometry data. Right: estimated components via the log-concave EM algorithm.

As a final application of the univariate log-concave EM algorithm, we analyzed the flow cytometry data described in Walther (2001). Fig. 4 shows the estimated components when we set $k = 2$ as suggested in Walther (2001). The shape of the estimated components indicates that a nonparametric analysis is indeed more appropriate.

5. A multivariate extension

In the multivariate set-up, we observe n i.i.d observations $(X_{i1}, \ldots, X_{id})$ in $\mathbf{R}^d$. Log-concave distributions are defined in a multivariate situation just as in the univariate case, but the computation of the MLE appears to be much more complicated. For this reason, we will work with the following simpler, yet flexible model: we only require that the univariate marginal distributions be log-concave. Then we model the dependence structure with a normal copula. That is, let $(N_1, \ldots, N_d)$ be multivariate normal with mean $\mathbf{0}$ and a correlation matrix $\mathbf{\Sigma}$ as covariance matrix. Let $F_1, \ldots, F_d$ be cdfs of arbitrary univariate log-concave distributions. Then the model for the distribution within a component (cluster) is $(X_{i1}, \ldots, X_{id}) \stackrel{d}{=} (F_1^{-1}\Phi(N_1), \ldots, F_d^{-1}\Phi(N_d))$, where Φ denotes the standard normal cdf. The marginal log-concave distributions F_j and the correlation matrix $\mathbf{\Sigma}$ are allowed to vary between components (clusters). The joint density in the mth component (cluster) is thus

$$f_m(x_1, \ldots, x_d) \;=\; \phi_{\mathbf{0},\mathbf{\Sigma}}(\Phi^{-1}F_1(x_1), \ldots, \Phi^{-1}F_d(x_d)) \prod_{j=1}^{d} \frac{f_j(x_j)}{\phi_{\mathbf{0},\mathbf{I}}(\Phi^{-1}F_j(x_j))}. \tag{3}$$

Fig. 5 shows the density contours of two distributions that belong to this model. The first distribution has a gamma(3,1) distribution as x-marginal, a standard normal distribution as y-marginal, and the identity matrix as $\mathbf{\Sigma}$. The second distribution has a beta(8,8) distribution as x-marginal, a beta(2,8) distribution as y-marginal, and a matrix with ones on the diagonal and 0.6 on the off-diagonal as $\mathbf{\Sigma}$. These examples show that the model under consideration allows for a flexible choice of cluster shape.

The resulting EM algorithm is quite similar to the univariate case. Again we use the results of a Gaussian EM algorithm as starting values, and then add five iterations of the log-concave EM algorithm. In the M-step we compute for each component (cluster) the d marginal log-concave MLEs just as in the univariate case. To obtain the MLE of $\mathbf{\Sigma}$ we use the following alternative interpretation of our model: $(\Phi^{-1}F_1(X_{i1}), \ldots, \Phi^{-1}F_d(X_{id}))$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. Thus we compute $Y_{ij} := \Phi^{-1}\hat{F}_j(X_{ij})$ for $i = 1, \ldots, n$, $j = 1, \ldots, d$, where $\hat{F}_j$ is the cdf of the log-concave MLE. Then we estimate $\mathbf{\Sigma}$ by the sample covariance matrix of the Y_{ij}, weighted by the $\tau_m(X_{ij})$, just as in the case of the Gaussian EM algorithm.

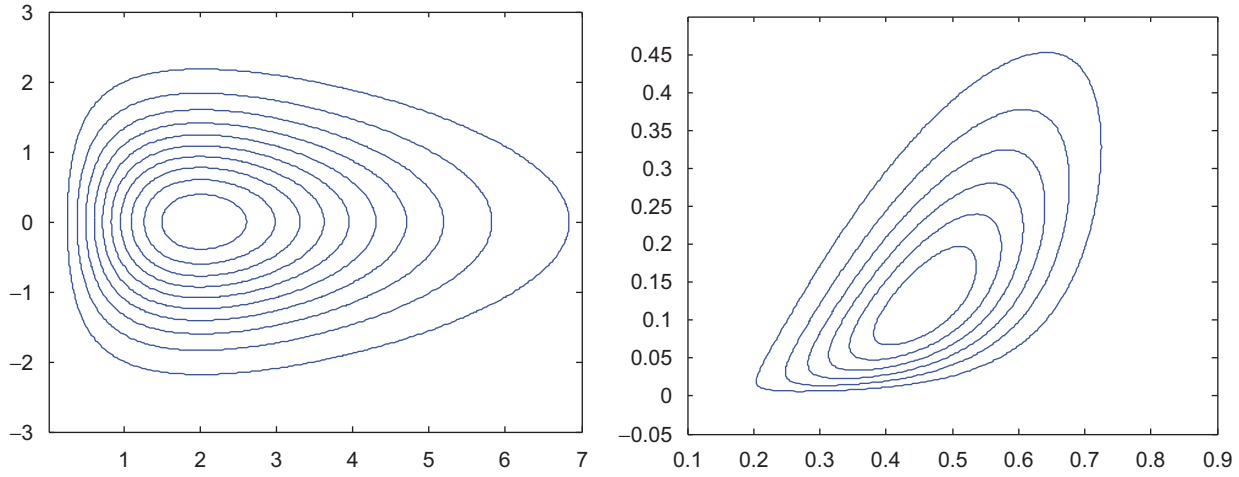

Fig. 5. Density contours of two distributions that belong to the multivariate log-concave model.

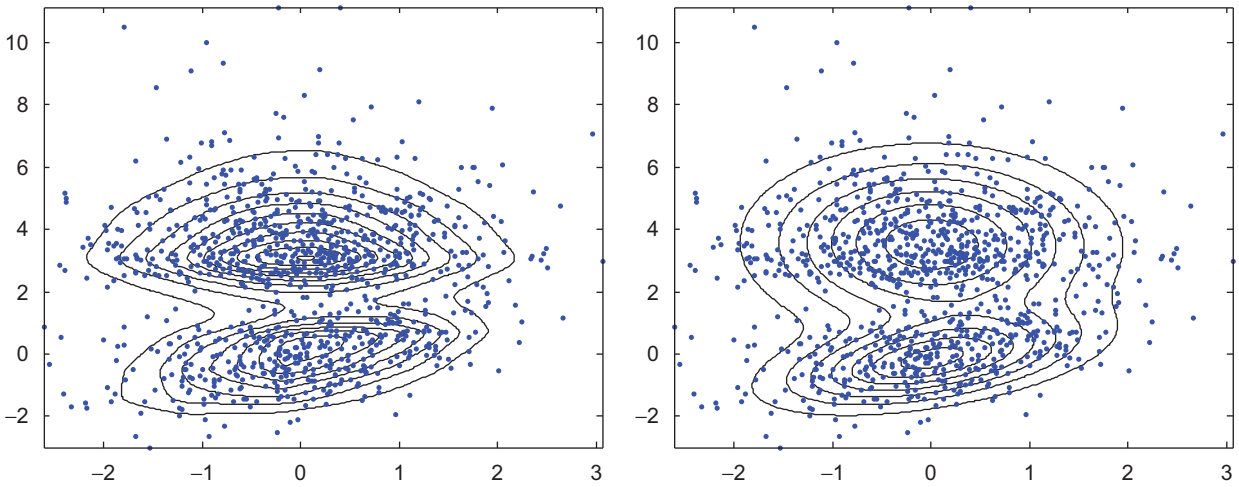

Fig. 6. Contours of the estimated model of the log-concave EM algorithm (left) and the Gaussian EM algorithm (right) based on the plotted observations. The underlying distribution has a skewed (shifted gamma) distribution in the y-direction of the top component.

We compared our multivariate log-concave EM algorithm to the Gaussian EM algorithm in two examples. In the first example we simulated 1000 observations with probability 0.4 from a bivariate normal with mean at the origin and covariance matrix [1 0.5; 0.5 1], and with probability 0.6 from a distribution whose x-coordinate is standard normal and whose y-coordinate is independently gamma(2,1) shifted by 2. The contours lines of the density estimates resulting from both algorithms are given in Fig. 6.

Fig. 7 compares the resulting clusterings over 1000 sets of observations, both for sample size 1000 and for sample size 100. The log-concave EM algorithm shows a clear improvement.

Fig. 8 repeats this simulation study for data drawn from the following normal mixture: with probability 0.4 the observation is drawn from a bivariate normal with mean at the origin and covariance matrix [2 1; 1 2], and with probability 0.6 from a bivariate normal with mean $(5, 5)'$ and the same covariance matrix. The Gaussian EM algorithm is tailor-made for this situation, but the simulations show that the more flexible log-concave EM algorithm seems to incur no significant performance penalty, even for the smaller sample size of 100. While we have not done any

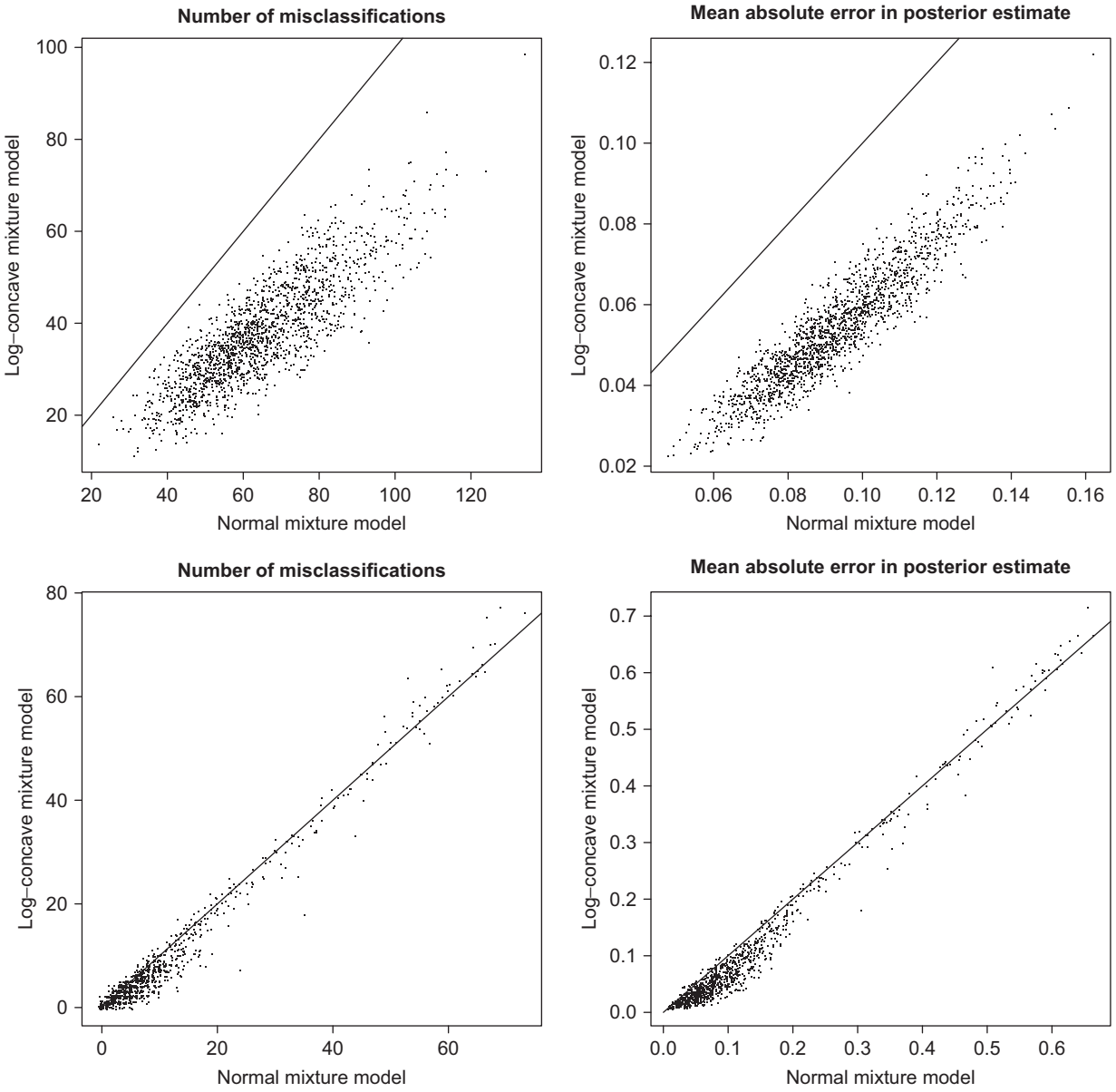

Fig. 7. Top left: number of observations misallocated by the log-concave EM algorithm vs. the Gaussian EM algorithm for 1000 sets of simulations from the multivariate mixture with a skewed (gamma) distribution in one component and sample size 1000. Top right: the mean absolute errors of estimated membership probabilities for the 1000 sets of simulations. The solid line marks the identity. Jitter has been added to the left panels for better visualization. The bottom row shows the analogous results for sample size 100.

simulations in higher dimensions, we note that by estimating only the marginal cluster distributions nonparametrically, we avoid the 'curse of dimensionality'.

6. Conclusion

We have shown how the parametric EM algorithm for clustering can be extended to allow for a flexible, nonparametric class of component distributions. The advantages of this algorithm are that it is not restricted to parametric models, that it no longer requires to specify such a model for the component distributions and hence that it is not sensitive to

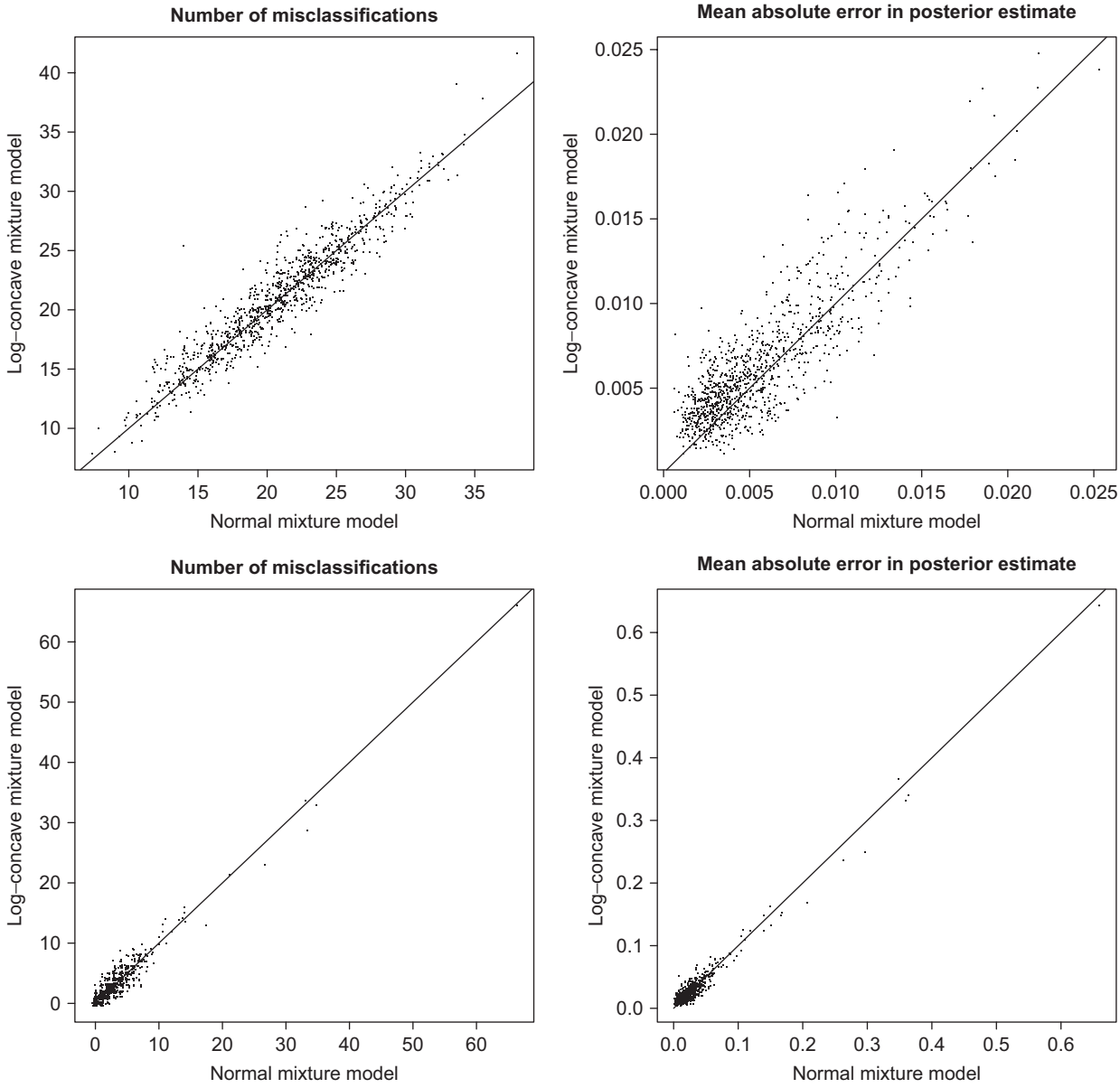

Fig. 8. Top left: number of observations misallocated by the log-concave EM algorithm vs. the Gaussian EM algorithm for 1000 sets of simulations from the multivariate normal mixture with sample size 1000. Top right: the mean absolute errors of estimated membership probabilities for the 1000 sets of simulations. The solid line marks the identity. Jitter has been added to the left panels for better visualization. The bottom row shows the analogous results for sample size 100.

a misspecification thereof, and that only one implementation of the algorithm is necessary. At the same time, there seems to be no noticeable performance penalty of this more general nonparametric algorithm vis-a-vis the parametric EM algorithm in the special case where the specified parametric model is indeed correct.

7. Problems for further research

We left open the question of identifiability of log-concave mixtures and the problem of selecting the number of components (clusters). Following our motivation that the log-concave MLE provides a correction to the fitted Gaussian

mixture model, a reasonable suggestion would be to employ one of the criteria for selecting the number of components in the Gaussian mixture model, see e.g. Chapter 6 in McLachlan and Peel (2000). A direct way to select the number of components in a nonparametric context is developed in Walther (2007).

Appendix A. Computational details

We used the following standard procedure with the Gaussian EM algorithm to avoid getting stuck in bad local maxima of the likelihood, see McLachlan and Krishnan (1997): we restarted the algorithm 20 times and selected the solution with the highest likelihood. At each restart, we initialized the component means with a random sample from a normal distribution with mean equal to the sample mean and variance equal to the sample variance. The component variances were initialized to equal the sample variance, and the mixing proportions were initialized to be equal. The EM algorithm was terminated once the relative change in log-likelihood dropped below 10^{-8}.

In the multivariate case of the log-concave EM algorithm, for each component m we iterated through the marginals $j = 1, \ldots, d$ and compute the log-concave MLE $\hat{f}_j$. (For simplicity we suppress the dependence on the component m.) As $\log \hat{f}_j$ is piecewise linear between the observations, it is straightforward to compute the corresponding cdf $\hat{F}_j$. We found that an approximation by linear interpolation of $\hat{f}_j$ allows simpler code without significant difference in the results. To avoid problems when computing $Y_{ij} := \Phi^{-1}\hat{F}_j(X_{ij})$, we rescaled $\hat{F}_j$ such that $\hat{F}_j(X_{(1)j}) = 1/(n+1)$ and $\hat{F}_j(X_{(n)j}) = n/(n+1)$.

References

Eilers, P.H.C., Borgdorff, M.W., 2006. Non-parametric log-concave mixtures. Manuscript.
Fraley, C.F., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. J. Amer. Statist. Assoc. 97, 611–631.
Hastie, T.J., Tibshirani, R.J., 1996. Discriminant analysis by Gaussian mixtures. J. Roy. Statist. Soc. Ser. B 58, 155–176.
Hunter, D.R., Wang, S., Hettmansperger, T.P., 2006. Inference for mixtures of symmetric distributions. Ann. Statist., in press.
Jongbloed, G., 1998. The iterative convex minorant algorithm for nonparametric estimation. J. Comput. Graph. Statist. 7, 310–321.
Lin, Y., Jeon, Y., 2003. Discriminant analysis through a semi-parametric model. Biometrika 90, 379–392.
McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley, New York.
McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.
Rufibach, K., 2006. Computing maximum likelihood estimators of a log-concave density function. J. Statist. Comput. Simul., in press.
Walther, G., 2001. Multiscale maximum likelihood analysis of a semiparametric model, with applications. Ann. Statist. 29, 1297–1319.
Walther, G., 2002. Detecting the presence of mixing with multiscale maximum likelihood. J. Amer. Statist. Assoc. 97, 508–513.
Walther, G., 2007. Oscillation analysis for the mixture complexity. Manuscript in preparation.