Global Convergence of Model Reference Adaptive Search for Gaussian Mixtures

Jeffrey W. Heath
Department of Mathematics
University of Maryland, College Park, MD 20742, jheath@math.umd.edu

Michael C. Fu
Robert H. Smith School of Business
University of Maryland, College Park, MD 20742, mfu@rhsmith.umd.edu

Wolfgang Jank
Robert H. Smith School of Business
University of Maryland, College Park, MD 20742, wjank@rhsmith.umd.edu

January 19, 2007

Abstract

While the Expectation-Maximization (EM) algorithm is a popular and convenient tool for mixture analysis, it only produces solutions that are locally optimal, and thus may not achieve the globally optimal solution. This paper introduces a new algorithm, based on a recently introduced global optimization approach called Model Reference Adaptive Search (MRAS), designed to produce globally optimal solutions of finite mixture models. We propose the MRAS mixture model algorithm for the estimation of Gaussian mixtures, which relies on the Cholesky decomposition to simulate random positive definite covariance matrices, and we provide a theoretical proof of global convergence to the optimal solution of the likelihood function. Numerical experiments illustrate the effectiveness of the proposed algorithm in finding global optima in settings where the classical EM fails to do so.

1 Introduction

A mixture model is a statistical model where the probability density function is a convex sum of multiple density functions. Mixture models provide a flexible and powerful mathematical approach to modeling many natural phenomena in a wide range of fields (McLachlan and Peel, 2000). One particularly convenient attribute of mixture models is that they provide a natural framework for clustering data, where the data are assumed to originate from a mixture of probability distributions, and the cluster memberships of the data points are unknown. Mixture models are highly popular and widely applied in many fields, including biology, genetics, economics, engineering, and marketing. Mixture models also form the basis of many modern supervised and unsupervised classification methods such as neural networks or mixtures of experts. In mixture analysis, the goal is to estimate the parameters of the underlying mixture distributions by maximizing the likelihood function of the mixture density with respect to the observed data.

One of the most popular methods for obtaining this goal is the Expectation-Maximization (EM) algorithm. The EM algorithm has gained popularity in mixture analysis, primarily because of its many convenient properties. One of these properties is that it guarantees an increase in the likelihood function in every iteration (Dempster et al., 1977). Moreover, because the algorithm operates on the log-scale, the EM updates are analytically simple and numerically stable for distributions that belong to the exponential family, such as Gaussian. However, the major drawback of EM is that it is a *local* optimization method only; that is, it converges to a local optimum of the likelihood function (Wu, 1983). This is a problem because with increasing data-complexity (e.g., higher dimensionality of the data and/or increasing number of clusters), the number of local optima in the mixture likelihood increases. Furthermore, the EM algorithm is a *deterministic* method; i.e., it converges to the same stationary point if initiated from the same starting value. So, depending on its starting values, there is a chance that the EM algorithm can get stuck in a sub-optimal solution, one that may be far from the global (and true) solution.

There have been relatively few attempts at systematically addressing the shortcomings of EM in the mixture model context. Perhaps the most common approach in practice is to simply re-run EM from multiple (e.g., randomly chosen) starting values, and then select the parameter value that provides the best solution obtained from all runs (see Biernacki et al., 2003). In addition to being computationally burdensome, especially when the parameter space is large, this approach is somewhat ad-hoc. More systematic approaches involve using *stochastic* versions of the EM algorithm such as the Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990). Alternative approaches rely on producing ergodic Markov chains that exhaustively explore every point in the parameter space (see e.g., Diebolt and Robert, 1990; Cao and West, 1996; Celeux and Govaert, 1992). Another approach that has been proposed recently is to use methodology from the global optimization literature. In that context, Jank (2006a) proposes a Genetic Algorithm version of the MCEM algorithm to overcome local solutions in the mixture likelihood (see also Jank, 2006b; Tu et al., 2006). Botev and Kroese (2004) propose the Cross-Entropy (CE) method for Gaussian mixtures with data of small dimension, and Kroese et al. (2006) use CE in vector quantization clustering. However, while many of these approaches promise a better performance in empirical studies, they stop short of guaranteeing global convergence. In this paper we propose a new and previously unexplored approach for mixture analysis, and we rigorously prove its convergence to the global optimum. Our approach is based on the ideas of Model Reference Adaptive Search (Hu et al., 2006). A related paper (Heath et al., 2007) focuses on computational issues, proposing and comparing several additional global optimization algorithms based on the principles of EM updating, for which proving theoretical convergence to the global optimum is still an open problem.

Model Reference Adaptive Search (MRAS) is a method that was first proposed in the field of operations research and is designed to attain globally optimal solutions to general multi-extremal continuous optimization problems (Hu et al., 2006). As in the CE method (Boer et al., 2005), MRAS produces estimates to optimization problems by iteratively generating candidate solutions in each iteration from a parametric sampling distribution. The candidates are all scored according to an objective function, and the highest scoring candidates are used to update the parameters of the sampling distribution by minimizing the Kullback-Leibler (KL) divergence between the sampling distribution and the current reference model. In this way, the properties of the *best* candidates in each iteration are retained. Due to the choice of the reference model sequence, the updating scheme of MRAS leads to a more general framework than the CE method, and allows for rigorous analysis of theoretical convergence (Hu et al., 2006).

Applying MRAS to the Gaussian mixture model problem directly is rather challenging, because the simulation of appropriate candidate solutions in every iteration is computationally intensive and the candidate solutions have to satisfy certain mixture model constraints (e.g., positive-definiteness of the mixture covariance matrices). Simulating these candidate solutions in a naive manner results in an extremely inefficient algorithm. In this paper, we propose a new and efficient implementation of MRAS for mixture models. In particular, we show that by representing the mixture covariance matrices by their corresponding Cholesky factorizations allows for unconstrained simulation of the covariance matrices in the MRAS mixture model algorithm.

Many global optimization algorithms that perform well empirically have no theoretical convergence proofs. In fact, many algorithms are ad-hoc or based on heuristics that do not allow for a rigorous mathematical investigation of their convergence properties. A particularly attractive feature of MRAS is that it leads to a rather general framework in which theoretical convergence of a particular instantiated algorithm can be proved rigorously, so we use this framework to prove convergence to the global optimum of Gaussian mixtures. To the best of our knowledge, this is the first mixture analysis algorithm that has provable global convergence. In addition to providing theoretical justification that the algorithm is not merely an ad-hoc heuristic, the convergence proof also gives insight into the performance of the algorithm.

The rest of the paper begins with the description of the mathematical framework of finite mixture models in Section 2. We proceed in Section 3 by explaining MRAS in general, and our implementation for Gaussian mixturs in particular. We also prove that our implementation converges to the global optimum. In Section 4 we carry out numerical experiments to investigate how the MRAS mixture model algorithm performs relative to the classical EM algorithm. We conclude and discuss future work in Section 5.

2 Finite Mixture Models

We begin by presenting the mathematical framework of finite mixture models. Assume there are n observed data points, $y = \{y_1, ..., y_n\}$, in some p-dimensional space. Assume that data is known to have been derived from g distinct probability distributions, weighted according to the vector $\pi = (\pi_1, ..., \pi_g)$, where the weights are positive and sum to one. Each component of the mixture has an associated probability density $f_j(\,\cdot\,;\psi_j)$, where ψ_j represents the parameters of the j^{th} mixture component. The mixture model parameters that need to be estimated are $\theta = (\pi_j;\psi_j)_{j=1}^{g}$; that is, both the weights and the probability distribution parameters for each of the g components. We write the mixture density of the data point y_i as:

$$\tilde{f}(y_i;\theta) = \sum_{j=1}^{g} \pi_j f_j(y_i;\psi_j).$$

The typical approach to estimating the parameters θ with respect to the observed data y is via maximization of the likelihood function:

$$L(y,\theta) = \prod_{i=1}^{n} \tilde{f}(y_i;\theta),$$

which is equivalent to maximization of the log-likelihood function:

$$\begin{aligned} \ell(y,\theta) = \log L(y,\theta) &= \sum_{i=1}^{n} \log \tilde{f}(y_i;\theta) \\ &= \sum_{i=1}^{n} \log \sum_{j=1}^{g} \pi_j f_j(y_i;\psi_j). \end{aligned}$$

Maximization of the log-likelihood function in the mixture model problem is non-trivial, primarily because the likelihood function L typically contains many local maxima, especially when the number of components g and/or the data-dimension p is large.

Consider the following example for illustration. We simulate 40 points from two univariate Gaussian distributions with means $\mu_1 = 0$ and $\mu_2 = 2$, variances $\sigma_1^2 = .001$ and $\sigma_2^2 = 1$, and each weight equal to .5. Notice that in this relatively simple example, we have 5 parameters to optimize (because the second weight is uniquely given by the first weight). Figure 1 shows the log-likelihood function plotted against only one parameter-component, μ_1. All other parameters are held constant at their true values. Notice the large number of local maxima to the right of the optimal value of $\mu_1 \approx 0$. Clearly, if we start the EM algorithm at, say, 3, it could get stuck far away from the global (and true) solution. This demonstrates that a very simple situation can already cause problems with respect to global and local optima.

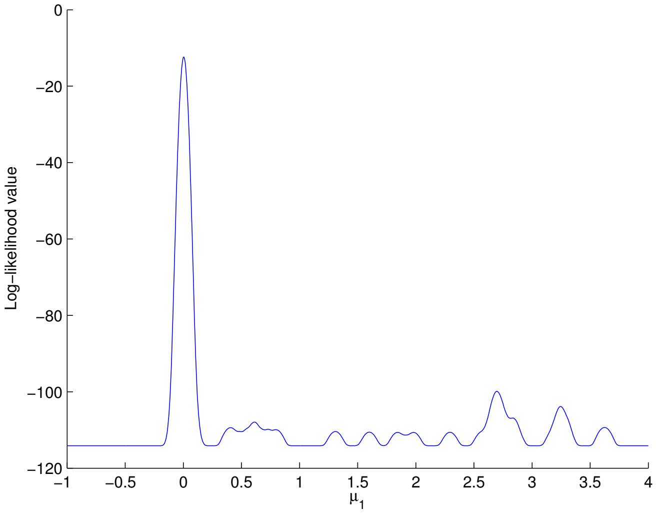

Figure 1: Plot of the log-likelihood function of the data set described above with parameters $\mu = (0, 2)$, $\sigma^2 = (.001, 1)$, and $\pi = (.5, .5)$, plotted against varying values of the mean component μ_1.

Henceforth, we will assume that the number of components g in the mixture are known. Methods for estimating g from the data are discussed in Fraley and Raftery (1998). In principle, one could combine these methods with the MRAS mixture model algorithm that we propose in this paper. The only adjustment that needs to be made is that the log-likelihood function as the optimization criterion be replaced by a suitable model-selection criterion such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) (McLachlan and Peel, 2000).

3 Global Optimization in Finite Mixture Models

Because the likelihood function of mixture models typically has a large number of local maxima, finding the global maximum can be a difficult task. Many optimization methods solely guarantee convergence to a local optimum, and are not necessarily concerned with systematically finding the global optimum. In the following we discuss a method specifically designed to find the global optimum. The method was first introduced in the field of operations research and is referred to as *Model Reference Adaptive Search*. We first explain how the method works in principle; then we adapt the method to the mixture model setting.

3.1 Model Reference Adaptive Search

Model Reference Adaptive Search (MRAS) is a global optimization tool that estimates the global optimum by generating candidate solutions from a parametric sampling distribution in each iteration. Hu et al. (2006) introduce MRAS as a method that produces solutions to the following global optimization problem:

$$x^* \in \operatorname*{argmax}_{x \in \chi} H(x), \ \chi \subseteq \Re^n.$$

MRAS achieves this goal by utilizing a sequence of intermediate reference distributions on the solution space to guide the parameter updates.

The basic methodology of MRAS can be described as follows. In the k^{th} iteration, we generate N_k candidate solutions, $X_1, X_2, ..., X_{N_k}$, according to a sampling distribution $\tilde{g}(\ \cdot\ ; \Upsilon^{(k)})$, where $\Upsilon^{(k)}$ represents the sampling parameters of the k^{th} iteration. After sampling the candidates, we score them according to the objective function, i.e., we compute the objective function value $H(X_i)$ for each candidate X_i. We then obtain an *elite* pool of candidates by selecting the top ρ-percentile scoring candidates. These elite candidates are used to update the parameters of the sampling distribution for the next iteration. An outline of MRAS is given in Figure 2.

In MRAS, the value of the percentile ρ changes over the course of the algorithm to ensure that the current iteration's candidates improve upon the candidates in the previous iteration. Let the lowest objective function score among the elite candidates in any iteration k be denoted as γ_k. We introduce a parameter ϵ, a very small positive number, to ensure that the increment in the $\{\gamma_k\}$ sequence is strictly bounded below. If $\gamma_k < \gamma_{k-1} + \epsilon$, increase ρ until $\gamma_k \geq \gamma_{k-1} + \epsilon$, effectively reducing the number of elite candidates. If, however, no such percentile ρ exists, then the number of candidates is increased in the next iteration by a factor of α

```
Result: Return highest-scoring estimate for $x^*$
Initialize $\Upsilon^{(0)}$.
$k \leftarrow 0$.
repeat
    Generate $N_k$ i.i.d. candidate vectors $X_1, ..., X_{N_k}$ from the sampling distribution $\tilde{g}(\ \cdot\ ; \Upsilon^{(k)})$.
    Compute the objective function values $H(X_1), H(X_2), ..., H(X_{N_k})$.
    Select the elite candidates by taking the top scoring $\rho$-percentile candidate vectors.
    Compute the updated MRAS sampling parameters $\Upsilon^{(k+1)}$ via (1) using the elite candidates.
    $k \leftarrow k + 1$.
until Stopping criterion is met.
```

Figure 2: MRAS Outline

(where $\alpha > 1$), such that $N_{k+1} = \alpha N_k$. The sampling parameters are then updated as follows:

$$\Upsilon^{(k+1)} := \underset{\Upsilon \in \Theta}{\operatorname{argmax}} \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{[S(H(X_i))]^k}{\tilde{g}(X_i; \Upsilon^{(k)})} I_{\{H(X_i) \geq \gamma_{k+1}\}} \ln g(X_i; \Upsilon), \tag{1}$$

where $S : \Re \to \Re^+$ is a strictly increasing function to account for cases where the objective function value $H(X)$ is negative for a given X, and $I_{\{\cdot\}}$ denotes the indicator function such that:

$$I_{\{A\}} := \begin{cases} 1, & \text{if event A holds,} \\ 0, & \text{otherwise.} \end{cases}$$

The sampling distribution $\tilde{g}$ in MRAS is generally chosen from the natural exponential family. We choose to sample candidate solutions from the Gaussian distribution for our implementation, such that $\Upsilon = (\xi, \Omega)$, where ξ and Ω are the mean and covariance matrix of the MRAS sampling distribution, respectively. The main idea of MRAS is that the sampling parameters will converge to a degenerate distribution centered on the optimal solution; i.e., the sequence of means $\xi^{(0)}, \xi^{(1)}, ...$ will converge to the optimal vector X^*, representing the optimal solution x^*, as the sequence of the sampling covariance matrices $\Omega^{(0)}, \Omega^{(1)}, ...$ converges to the zero matrix. Table 1 provides a list of the mixture model parameters and the MRAS parameters.

While MRAS is generally very versatile, applying it to the mixture model context is not straightforward. Part of the reason is that the mixture model requires simulation of candidate solutions that satisfy the mixture model constraints. In the following, we propose a solution via the Cholesky decomposition in order to assure an efficient implementation of MRAS.

Table 1: List of the model and MRAS parameters.

Mixture Model parameters	MRAS parameters
n = number of data points	$\Upsilon^{(k)}$ = sampling parameters in k^{th} iteration
$y_i = i^{th}$ data point	$\tilde{g}(\,\cdot\,;\Upsilon^{(k)})$ = sampling density
p = dimension of data	N_k = number of candidates in k^{th} iteration
g = number of mixture components	X_i = candidate vector
π_j = weight of j^{th} mixture component	ρ = elite candidate percentile
ψ_j = probability distribution parameters of j^{th} component	γ_k = lowest objective score of elite candidates in k^{th} iteration
$f_j(\,\cdot\,;\psi_j)$ = probability density of j^{th} component	λ = sampling weight
θ = model parameters to estimate	$S : \Re \rightarrow \Re^+$ = strictly increasing function
θ^* = model parameters that represent the global optimum	X^* = candidate vector representing the global optimum
$\ell(y, \theta)$ = log-likelihood function	ϵ = lower bound on the increase of each γ_k
μ_j = Gaussian mixture mean vector	$\xi^{(k)}$ = Gaussian sampling mean vector
Σ_j = Gaussian mixture covariance matrix	$\Omega^{(k)}$ = Gaussian sampling covariance matrix
	χ = constrained domain for candidate vectors
	$H(\,\cdot\,)$ = objective function
	Θ = constrained domain for sampling parameters

3.2 MRAS algorithm for Gaussian Mixture Models

As pointed out above, MRAS requires, in every iteration, the simulation of candidate solutions from within the parameter space. In the Gaussian mixture model, these candidate solutions must include the mixture weights $\pi = (\pi_1, ..., \pi_g)$ and the probability distribution parameters $\psi_j = (\mu_j, \Sigma_j)$ for $j = 1, ..., g$, where μ_j is the mean vector and Σ_j is the covariance matrix of the j^{th} component. Simulating covariance matrices is involved, since they need to be positive definite. Naive approaches (e.g., via simulating matrices randomly and consequently selecting only those that are positive definite) can be extremely inefficient (see e.g., Heath et al., 2007). In the following, we propose a new method to simulate positive definite covariance matrices for the MRAS mixture model algorithm. This method relies on the Cholesky decomposition. Recall the following theorem (see e.g., Thisted, 1988) regarding the Cholesky decomposition of a symmetric positive definite matrix:

Theorem 1. *A real, symmetric matrix A is symmetric positive definite (s.p.d.) if and only if it has a Cholesky decomposition such that $A = U^T U$, where U is a real-valued upper triangular matrix.*

Because covariance matrices are s.p.d., each covariance matrix has a corresponding Cholesky factorization U. Therefore, one way to stochastically generate covariance matrices in the MRAS mixture model is to generate the components of the U matrix from the Cholesky decomposition instead of the components of the covariance matrix Σ directly. Note that only the $\frac{p(p+1)}{2}$ upper right-hand components of U must be generated for each $p \times p$ covariance matrix (all other components are necessarily zero). Then the covariance matrix can be constructed from the simulated Cholesky factors, ensuring that the covariance matrix is s.p.d.

One potential problem with this method is that the Cholesky factorization for a symmetric positive definite matrix is not unique. For a Cholesky factorization U of Σ, we can multiply any subset of rows of U by -1 and obtain a different Cholesky factorization of the same Σ. Thus, there is not a unique global optimum in the MRAS mixture model algorithm. However, in their discussion of parameterizations of positive definite matrices, Pinheiro and Bates (1996) note that if the diagonal elements of the Cholesky factorization U are required to be positive, then the Cholesky factorization U is unique. Thus, by restricting the diagonal elements of U to be positive, we can circumvent the uniqueness problem of the Cholesky factorization mentioned above. We therefore choose to construct the covariance matrices in the MRAS mixture model algorithm by sampling the diagonal components of U from a truncated Gaussian distribution (accepting all positive values), and subsequently computing the covariance matrix $\Sigma = U^T U$.

MRAS can now be applied to the estimation of Gaussian mixtures in the following way. We first sample candidate solutions X_i that correspond to the set of mixture parameters $\theta = (\mu_j, \Sigma_j, \pi_j)_{j=1}^{g}$, where the covariance matrices are represented by their corresponding Cholesky factorizations mentioned above. We then score each candidate with the log-likelihood function, and use the best-scoring candidates to update the sampling distribution. The goal is to obtain the optimal solution X^* containing the mixture means, Cholesky factorizations, and weights of the *global* maximum likelihood estimate $\theta^* = (\mu_j^*, \Sigma_j^*, \pi_j^*)_{j=1}^{g}$. We provide the MRAS mixture model algorithm in Figure 2. Note that the MRAS parameter λ is a small constant which assigns a probability to sample from the initial sampling distribution $g(\,\cdot\,; \Upsilon^{(0)})$ in any subsequent iteration. Also, $g(\,\cdot\,; \xi, \Omega)$ is the multivariate Gaussian density, i.e.,

$$g(X; \xi, \Omega) = \frac{1}{\sqrt{(2\pi)^p |\Omega|}} \exp\left(-\frac{1}{2}(X-\xi)^T \Omega^{-1} (X-\xi)\right),$$

where $|\Omega|$ denotes the determinant of the matrix Ω.

The stopping criterion for the MRAS mixture model algorithm that we use is to stop when the increase of the best log-likelihood value over k iterations falls below a specified tolerance.

```
Data: Data points $y_1, y_2, ..., y_n$
Result: Return highest-scoring estimate for $X^*$.
Initialize $\xi^{(0)}$ and $\Omega^{(0)}$.
$k \leftarrow 0$.
repeat
    Generate $N_k$ i.i.d. candidate vectors $X_1, ..., X_{N_k}$ from the sampling distribution
    $\tilde{g}(\,\cdot\,; \xi^{(k)}, \Omega^{(k)}) := (1-\lambda)g(\,\cdot\,; \xi^{(k)}, \Omega^{(k)}) + \lambda g(\,\cdot\,; \xi^{(0)}, \Omega^{(0)})$.
    Compute the log-likelihoods values $\ell(y, X_1), \ell(y, X_2), ..., \ell(y, X_{N_k})$.
    Select the elite candidates by taking the top scoring $\rho_{k-1}$-percentile candidate vectors, and define
    $\tilde{\gamma}_k(\rho_k)$ as the $\rho_k$-percentile log-likelihood score obtained of all candidates in iteration $k$.
    if $k = 0$ or $\tilde{\gamma}_k(\rho_k) \geq \gamma_k + \frac{\epsilon}{2}$ then
        $\gamma_{k+1} \leftarrow \tilde{\gamma}_k(\rho_k), \rho_{k+1} \leftarrow \rho_k$, and $N_{k+1} \leftarrow N_k$.
    else
        find the largest $\tilde{\rho} \in (\rho_k, 100)$ such that $\tilde{\gamma}_k(\tilde{\rho}) \geq \gamma_k + \frac{\epsilon}{2}$.
        if such a $\tilde{\rho}$ exists then
            $\gamma_{k+1} \leftarrow \tilde{\gamma}_k(\tilde{\rho})$, $\rho_{k+1} \leftarrow \tilde{\rho}$, and $N_{k+1} \leftarrow N_k$,
        else
            $\gamma_{k+1} \leftarrow \gamma_k$, $\rho_{k+1} \leftarrow \rho_k$, and $N_{k+1} \leftarrow \alpha N_k$.
        end
    end
    Update the sampling parameters according to:

    $\xi^{(k+1)} \leftarrow \frac{\sum_{i=1}^{N_k} S(\ell(y, X_i))^k / \tilde{g}(X_i, \xi^{(k)}, \Omega^{(k)}) I_{\{\ell(y, X_i) \geq \gamma_{k+1}\}} X_i}{\sum_{i=1}^{N_k} S(\ell(y, X_i))^k / \tilde{g}(X_i, \xi^{(k)}, \Omega^{(k)}) I_{\{\ell(y, X_i) \geq \gamma_{k+1}\}}},$   (2)

    $\Omega^{(k+1)} \leftarrow \frac{\sum_{i=1}^{N_k} S(\ell(y, X_i))^k / \tilde{g}(X_i, \xi^{(k)}, \Omega^{(k)}) I_{\{\ell(y, X_i) \geq \gamma_{k+1}\}} (X_i - \xi^{(k+1)})(X_i - \xi^{(k+1)})^T}{\sum_{i=1}^{N_k} S(\ell(y, X_i))^k / \tilde{g}(X_i, \xi^{(k)}, \Omega^{(k)}) I_{\{\ell(y, X_i) \geq \gamma_{k+1}\}}}.$   (3)

    $k \leftarrow k + 1$.
until Stopping criterion is met.
```

Figure 3: MRAS Mixture Model Algorithm

3.3 Preventing Degenerate Solutions

Maximizing the log-likelihood function in the Gaussian mixture model can lead to unbounded solutions, if the parameter space is not properly constrained. In fact, we can make the log-likelihood value arbitrarily large by letting one of the component means be equal to a single data point, and then letting the generalized variance, or determinant of the covariance matrix, of that component be arbitrarily small. Such a solution is referred to as a degenerate, or spurious, solution. In order to prevent degenerate solutions in practice, it is necessary to constrain the parameter space in such a way as to avoid exceedingly small variance components in the univariate case, or exceedingly small generalized variances in the multivariate case.

One constraint that achieves this goal is to limit the relative size of the generalized variances of the mixture components (McLachlan and Peel, 2000) and it is given by:

$$\min_{i,j} \frac{|\Sigma_i|}{|\Sigma_j|} \geq c > 0,$$

where $|\Sigma|$ denotes the determinant of the matrix Σ. To avoid degenerate solutions, we will use the following constraint instead:

$$\min_{j} |\Sigma_j| \geq c > 0. \tag{4}$$

In each of these constraints, determining the appropriate value of c is difficult when no prior information on the problem structure is known. If the MRAS mixture model algorithm generates a covariance matrix that violates the constraint given by (4), we discard the candidate and re-generate a new one.

3.4 Global Convergence of MRAS

In this section we discuss the global convergence properties of MRAS mixture model algorithm in the finite mixture model problem. We must first revert to the general MRAS framework, where Hu et al. (2006) provide a convergence proof of MRAS to the globally optimal solution when using a sampling distribution $g(\,\cdot\,;\Upsilon)$ that belongs to the exponential family. Before we discuss the theorem, we provide the definition of the exponential family of distributions, as well as some required assumptions for the theorem.

Definition 1. *A parameterized family of p.d.f.'s* $\{g(\,\cdot\,;\Upsilon), \Upsilon \in \Theta \subseteq \Re^m\}$ *on* χ *is said to belong to the exponential family if there exists functions* $h : \Re^n \to \Re, \Gamma : \Re^n \to \Re^m$, *and* $K : \Re^m \to \Re$ *such that*

$$g(x;\Upsilon) = \exp\{\Upsilon^T \Gamma(x) - K(\Upsilon)\}h(x), \ \forall \Upsilon \in \Theta,$$

where $K(\Upsilon) = \ln \int_{x\in\chi} \exp\{\Upsilon^T\Gamma(x)\}h(x)dx$.

The following assumptions are referenced in the statement of Theorem 2.

Assumptions:

A1. For any given constant $\xi < H(x^*)$, the set $\{x : H(x) \geq \xi\} \cap \chi$ has a strictly positive Lebesgue measure.

A2. For any given constant $\delta > 0$, $\sup_{x\in A_\delta} H(x) < H(x^*)$, where $A_\delta \colon = \{x : \|x - x^*\| \geq \delta\} \cap \chi$ and the supremum over the empty set is defined to be $-\infty$.

A3. There exists a compact set Π_ϵ such that $\{x : H(x) \geq H(x^*) - \epsilon\} \cap \chi \subseteq \Pi_\epsilon$. Moreover, $g(x; \Upsilon^{(0)})$ is bounded away from zero on Π_ϵ, i.e., $g_* = \inf_{x \in \Pi_\epsilon} g(x; \Upsilon^{(0)}) > 0$.

A4. The parameter vector $\Upsilon^{(k)}$ computed in (1) is an interior point of Θ for all k.

In Theorem 2, Hu et al. (2006) show global convergence of MRAS to the optimal solution x^* when using the multivariate Gaussian sampling distribution. As the number of iterations tends to infinity, the sampling distribution tends toward a degenerate distribution centered on the optimal solution x^*.

Theorem 2. *If multivariate Gaussian p.d.f.'s are used in MRAS, i.e.,*

$$g(X; \xi^{(k)}, \Omega^{(k)}) = \frac{1}{\sqrt{(2\pi)^n |\Omega^{(k)}|}} \exp\left(-\frac{1}{2}(X - \xi^{(k)})^T (\Omega^{(k)})^{-1} (X - \xi^{(k)})\right),$$

$\epsilon > 0, \alpha > (\beta S^*)^2$*, and Assumptions A1, A2, A3, and A4 are satisfied, then*

$$\lim_{k \to \infty} \xi^{(k)} = x^*, \textit{and} \lim_{k \to \infty} \Omega^{(k)} = 0_{n \times n} \ \textit{w.p.}\ 1$$

3.5 Proving Global Convergence of the MRAS Mixture Model Algorithm

In order to show that Corollary 2 applies to the MRAS mixture model algorithm algorithm, we must show that Assumptions A1, A2, A3, and A4 hold true in the maximization of the likelihood function of the mixture density. So, for our purposes, the objective function $H(x)$ discussed in the general MRAS framework is the log-likelihood of the mixture density:

$$\ell(y, \theta) = \sum_{i=1}^{n} \log \sum_{j=1}^{g} \pi_j f_j(y_i; \mu_j, \Sigma_j).$$

In the MRAS mixture model algorithm, we are estimating the vector X^* representing the optimal means, weights, and Cholesky factorizations of the covariance matrices, i.e., the vector X^* representing the optimal solution $\theta^* = (\mu_i^*, \Sigma_i^*, \pi_i^*)_{i=1}^g$. Therefore, we are trying to solve the optimization problem:

$$X^* \in \operatorname*{argmax}_{X \in \chi} \ell(y, X).$$

Before we prove the global convergence of the MRAS mixture model algorithm, we first provide the following useful lemmas. Lemma 1 shows that a continuous function that is bounded above and possesses a unique optimal maximizer on a constrained space $\chi \in \Re^n$ satisfies Assumption A1.

Lemma 1. *For a continuous function $H(x)$, $x \in \chi \in \Re^n$, where H is bounded above and there exists a unique optimal maximizer x^* s.t. $H(x) < H(x^*)$, $\forall x \neq x^*$, then $\forall \xi < H(x^*)$, the set $\{x : H(x) \geq \xi\}$ has strictly positive Lebesgue measure, and thereby Assumption A1 is satisfied.*

Proof: Choose $\xi < H(x^*)$ and let $\epsilon = H(x^*) - \xi > 0$. By continuity of H, $\exists \delta > 0$ s.t. $\forall x \in \{x : \|x - x^*\| < \delta\}$, then $|H(x) - H(x^*)| < \epsilon$. By rewriting the left- and right-hand sides of the inequality, we see that $H(x^*) - H(x) < H(x^*) - \xi$, i.e., $\xi < H(x)$, $\forall x \in \{x : \|x - x^*\| < \delta\}$. Since the set $\{x : \|x - x^*\| \leq \frac{\delta}{2}\} \subseteq \{x : H(x) \geq \xi\}$, then $m\left(\{x : H(x) \geq \xi\}\right) \geq m\left(\{x : \|x - x^*\| \leq \frac{\delta}{2}\}\right) > 0$. □

Lemma 2 gives an inequality relating the determinants of two positive definite $n \times n$ matrices with the determinant of their convex combination (see e.g., Horn and Johnson, 1990).

Lemma 2. *For positive definite $n \times n$ matrices A and B,*

$$\det\left(\alpha A + (1-\alpha)B\right) \geq (\det A)^{\alpha}(\det B)^{1-\alpha}, \ \textit{where } \alpha \in (0,1).$$

In Lemma 3 we extend the statement of Lemma 2 to a convex combination of an arbitrary number of positive definite $n \times n$ matrices. In the proof, we make use of two properties of positive definite matrices: for positive definite matrices A, B, and scalar $\alpha > 0$, then αA and $A + B$ are both positive definite as well (Johnson, 1970).

Lemma 3. *For positive definite $n \times n$ matrices A_j, $j = 1, ..., k$,*

$$\det\left(\sum_{j=1}^{k} \alpha_j A_j\right) \geq \prod_{j=1}^{k} (\det A_j)^{\alpha_j},$$

for any set of $\{\alpha_j\}_{j=1}^{k}$ s.t. $\alpha_j > 0$ and $\sum_{j=1}^{k} \alpha_j = 1$.

Proof: We prove this lemma by induction.

i. Base case: $k = 2$, shown by Lemma 2.

ii. Assuming the lemma holds for k, we show it holds for $k+1$, i.e., for any set $\{\tilde{\alpha}_j\}_{j=1}^{k+1}$ s.t. $\tilde{\alpha}_j > 0$ and $\sum_{j=1}^{k+1} \tilde{\alpha}_j = 1$, then $\det\left(\sum_{j=1}^{k+1} \tilde{\alpha}_j A_j\right) \geq \prod_{j=1}^{k+1} (\det A_j)^{\tilde{\alpha}_j}$,.

Define $\alpha_j = \frac{\tilde{\alpha}_j}{1 - \tilde{\alpha}_{k+1}}$, for $j = 1, ..., k$. Thus, $\sum_{j=1}^{k} \alpha_j = 1$ and the induction assumption can be applied to $\{A_1, ..., A_k\}$ for this set of $\{\alpha_j\}_{j=1}^{k}$. Then,

$$\begin{aligned}
\det\left(\sum_{j=1}^{k+1}\tilde{\alpha}_j A_j\right) &= \det\left(\sum_{j=1}^{k}\tilde{\alpha}_j A_j + \tilde{\alpha}_{k+1}A_{k+1}\right)\\
&= \det\left((1-\tilde{\alpha}_{k+1})\sum_{j=1}^{k}\alpha_j A_j + \tilde{\alpha}_{k+1}A_{k+1}\right)\\
&\geq \left[\det\left(\sum_{j=1}^{k}\alpha_j A_j\right)\right]^{1-\tilde{\alpha}_{k+1}} (\det A_{k+1})^{\tilde{\alpha}_{k+1}} \text{ (by Lemma 2)}\\
&\geq \left[\prod_{j=1}^{k}(\det A_j)^{\alpha_j}\right]^{1-\tilde{\alpha}_{k+1}} (\det A_{k+1})^{\alpha_{\widetilde{k+1}}} \text{ (by induction assumption)}\\
&= \left[\prod_{j=1}^{k}(\det A_j)^{\tilde{\alpha}_j}\right](\det A_{k+1})^{\tilde{\alpha}_{k+1}}\\
&= \prod_{j=1}^{k+1}(\det A_j)^{\tilde{\alpha}_j}.
\end{aligned}$$

Therefore, we have shown by induction that the statement of the lemma is true. □

Constraining the parameter space is necessary for the proof of the MRAS mixture model algorithm convergence theorem. As mentioned in Section 3.3, we must place additional constraints on the parameter space in order to prevent degenerate clusters and an unbounded log-likelihood value. Specifically, these constraints are $|U_j^T U_j| \geq c > 0,\ j = 1, ..., g$, i.e., bounding the generalized variances of the covariance matrices below. We simplify this constraint by relying on a convenient property of determinants of positive definite matrices: for positive definite A, B, $\det AB = \det A \det B$. So, for the Cholesky decomposition $\Sigma = U^T U$, $|\Sigma| = |U^T||U| = |U|^2$. Equivalently, we write $|U| \geq \sqrt{c}$. Since U is an upper-triangular matrix, $|U|$ is equal to the product of its diagonal elements. So, the constraint $|U^T U| \geq c$ can be written as $\prod_{i=1}^{p} U_{ii} \geq \sqrt{c}$.

One condition that is necessary for satisfying Assumption A2 is that the optimal candidate solution X^* be unique. By restricting the diagonal components of the Cholesky factorization U to be positive, its correponding covariance matrix Σ is unique. However, for a given optimal solution, any permutation of the cluster labels will result in an equivalent log-likelihood value to the problem, resulting in $g!$ optimal solutions and therefore a non-indentifiable formulation. To avoid this problem, we add the following constraint to the problem:

$$\mu_1(1) \leq \mu_2(1) \leq \mu_3(1) \leq ... \leq \mu_g(1),$$

where $\mu_i(1)$ represents the 1^{st} mean component of the i^{th} cluster. Although the inequalities in this constraint are not strict, the probability of multiple mean components of continuous random data being equal is zero. Therefore, this constraint mandates a unique ordering of the mixture components of θ^* w.p. 1 for continuous random data, resulting in a unique optimal candidate solution X^* to the MRAS Gaussian mixture model algorithm.

To allow us to prove convergence, we choose to bound the candidate means within a compact space based on the observed data set. In particular, we define y_{min} as the minimum value over all components of the data points $y_1, ..., y_n$. That is, $y_{min}(i) = \min_{j=1,...,n} y_j(i)$. Similarly, we define y_{max} as the maximum value over all components of the data points, i.e., $y_{max}(i) = \max_{j=1,...,n} y_j(i)$. We note that bounding the candidate mean components by y_{min} and y_{max} is not an unreasonable constraint; clearly, the means of the optimal clusters will not lie outside the range of the data.

We also place constraints on the components of the Cholesky factorizations when we generate the candidate vectors. We first calculate the sample variance of the data set, $Var(\{y_1, y_2, ..., y_n\})$, and then choose the maximum across all p components, i.e., $V_{max} = \max_{i=1,...,p} Var(\{y_1, y_2, ..., y_n\})$. And so, V_{max} represents an upper bound for the variance component of any cluster. Constraining the diagonal components of the Cholesky factorizations within the bounds $[0, V_{max}]$ and the off-diagonal non-zero components within $[-V_{max}, V_{max}]$ suffices, as the global optima will undeniably satisfy these constraints.

Therefore, the solution space with all of the necessary constraints is given by the following:

$$\chi = \begin{cases} \mu_j \in [y_{min}, y_{max}], & j = 1, ..., g \\ \quad \text{s.t. } \mu_1(1) \leq \mu_2(1) \leq ... \leq \mu_g(1) & \\ U_{j(ii)} \in [0, V_{max}], & j = 1, ..., g; \ i = 1, ..., p \\ \quad \text{s.t. } \prod_{i=1}^{p} U_{j(ii)} \geq \sqrt{c} > 0, & j = 1, ..., g \\ U_{j(ik)} \in [-V_{max}, V_{max}], & j = 1, ..., g; \ i = 1, ..., p-1; \ k = i+1, ..., p \\ \pi_j \in [0, 1], & j = 1, ..., g \\ \quad \text{s.t. } \sum_{j=1}^{g} \pi_j = 1 & \end{cases} \tag{5}$$

The number of parameters that we are estimating, namely the means, weights, and the upper-triangular entries of the Cholesky factorization (all other components are necessarily zero) for each cluster, is $d := g\frac{(p+1)(p+2)}{2}$, so we can consider the space χ to be d-dimensional. The MRAS sampling parameters $\Upsilon = (\xi, \Omega)$ belong to the space Θ, where $\Theta = \{\xi \in \chi,\ \Omega \text{ is s.p.d.}\}$.

Lemma 4. *The subspace $\chi \subseteq \Re^d$ is compact.*

Proof: For any vector $X \in \chi$, we note that all components of X are bounded as described in (5). We now show that the space χ is closed. The constraints that bound the space clearly constitute a closed subspace in $\Re^d$. The remaining constraints, namely $\mu_1(1) \leq \mu_2(1) \leq ... \leq \mu_g(1)$, $\prod_{i=1}^{p} U_{j(ii)} \geq \sqrt{c}$ for $j = 1, ..., g$, and $\sum_{j=1}^{g} \pi_j = 1$, each represent a closed subspace of $\Re^d$, because all inequalities on the constraints are not strict. Therefore, χ is a finite intersection of closed sets, and is thus closed. Because χ is both closed and bounded, then χ is compact. □

Lemma 5. *For a continuous function $H(x)$, $x \in \chi \in \Re^n$, where H is bounded above and there exists a unique optimal solution x^* s.t. $H(x) < H(x^*)$, $\forall x \neq x^*$ and χ is a compact space, then $\forall \delta > 0$, $\sup_{x \in A_\delta} H(x) < H(x^*)$, where $A_\delta := \{x : \|x - x^*\| \geq \delta\} \cap \chi$, and thereby Assumption A2 is satisfied.*

Proof: We prove this lemma directly:
We can rewrite $A_\delta = \chi \setminus \{x : \|x - x^*\| < \delta\}$, which is the complement of the open ball of radius δ around x^* intersected with χ. Therefore, since χ is a compact space, A_δ is a compact space as well.
Since $H(x)$ is a continuous function, it achieves its supremum on the compact space A_δ, i.e., $\exists \tilde{x} \in A_\delta$ s.t. $\sup_{x \in A_\delta} H(x) = H(\tilde{x})$.
And, because $H(x) < H(x^*)$, $\forall x \neq x^*$, we have:

$$\sup_{x \in A_\delta} H(x) = H(\tilde{x}) < H(x^*). \ \square$$

Now we give Theorem 3, where we show that Corollary 2 applies to MRAS mixture model algorithm in the global optimization of Gaussian finite mixture models.

Theorem 3. *For the maximization of the likelihood function of a mixture density of g Gaussian clusters, if the MRAS parameters are chosen s.t. $\epsilon > 0, \alpha > (\beta S^*)^2$, where $S^* := S(\ell(y, \theta^*))$, and we are optimizing over the compact space χ denoted by (5), then:*

$$\lim_{k \to \infty} \xi^{(k)} = X^*, and \lim_{k \to \infty} \Omega^{(k)} = 0_{d \times d} \ w.p.\ 1,$$

where X^* is the unique optimal vector representing the MLE $\theta^* = (\mu_j^*, \Sigma_j^*, \pi_j^*)_{j=1}^g$.

Proof: This proof consists of showing that Assumptions A1, A2, A3, and A4 apply to MRAS mixture model algorithm in the maximization of the log-likelihood of the Gaussian mixture density.

i. Because $\ell(y,\theta)$ is continuous on χ w.r.t. θ, then by Lemma 1, for any $\xi < \ell(y,\theta^*)$, the set $\{y : \ell(y,\theta) \geq \xi\} \cap \chi$ has a strictly positive Lebesgue measure. Thus, Assumption A1 is satisfied.

ii. By Lemma 5, since $\ell(y,\theta)$ is continuous on χ w.r.t. θ, then $\forall \delta > 0$, $\sup_{\theta \in A_\delta} \ell(y,\theta) < \ell(y,\theta^*)$, where $A_\delta := \{\theta : \|\theta - \theta^*\| \geq \delta\} \cap \chi$. And so, Assumption A2 is satisfied.

iii. By restricting the search space to a compact region, then the set $\{\theta : \ell(y,\theta) \geq \ell(y,\theta^*) - \epsilon\} \cap \chi$ is a subset of a compact set, namely χ itself. Moreover, using a multivariate Gaussian sampling distribution ensures that sampling any point in the entire solution space on the first iteration occurs with non-zero probability. Thus, A3 is shown.

iv. In order to show that the formulation satisfies A4, we first revisit the updating scheme of MRAS when the sampling distribution is multivariate Gaussian as given by Equations (2) and (3). It is evident that the mean of the sampling distribution, $\xi^{(t)}$, is simply a convex combination of the elite candidates. Since each candidate $X_i \in \chi$, then a convex combination of them will satisfy all of the constraints as well. One can verify this by noting that the space χ is convex; this is clearly evident for all of the constraints in the formulation, except for the degenerate cluster constraint, $|U_j| \geq \sqrt{c} > 0, \ j = 1, ..., g$, which we now address.

   We need to show that a convex combination of the top t candidates also satisfies this constraint, namely $\left|\sum_{j=1}^t \alpha_j U_j\right| \geq \sqrt{c}$. We note that as a direct application of Lemma 3, $\left|\sum_{j=1}^t \alpha_j U_j\right| \geq \prod_{j=1}^t |U_j|^{\alpha_j} \geq \min_j |U_j| \geq \sqrt{c}$. This shows that a convex combination of Cholesky factorizations satisfying the degenerate constraint will also satisfy the degenerate constraint.

   Also, because the candidates X_i are sampled from the probability distribution $\tilde{g}(\,\cdot\,;\xi^{(k)},\Omega^{(k)})$, then w.p. 1 each candidate lies in the interior of χ. Therefore, the updated mean vector $\xi^{(k+1)}$ will also lie in the interior of the space. Also, the updated $\Omega^{(k+1)}$ is clearly s.p.d. by construction, and thus A4 is satisfied. □

4 Numerical Experiment

We now demonstrate the performance of the MRAS mixture model algorithm. We compare it to the EM algorithm, since EM is the most common approach for estimating mixture models. We conduct the numerical experiment in Matlab and execute the algorithms on a 2.80 GHz Intel CPU with 1 GB RAM. This experiment provides empirical evidence that the MRAS mixture model algorithm has the potential to produce better solutions than EM. A more comprehensive computational study on the performance of the MRAS mixture model algorithm (and other mixture model algorithms) is contained in Heath et al. (2007).

4.1 Initial Parameters

We initialize the EM algorithm using uniform starting values over the solution space. That is, we initialize the means uniformly over the range of the data, the variances uniformly between 0 and the sample variance of the data, and the weights uniformly between 0 and 1. Then we normalize the weights so that they sum to one. The stopping criterion for the EM algorithm is set to $|\zeta_k - \zeta_{k-1}| \leq 10^{-5}$, where ζ_k is the log-likelihood value obtained in iteration k.

One of the benefits of MRAS is that its performance is relatively insensitive to its starting values (Hu et al., 2006). We initialize the mean of the sampling distribution $\xi^{(0)}$ of the MRAS mixture model algorithm as follows: we set the means equal to the mean of the data, we set the covariance matrices equal to diagonal matrices with the sample variances of the data along the diagonals, and we set each weight component equal to $\frac{1}{g}$. Also, we initialize the covariance matrix of the sampling distribution $\Omega^{(0)}$ as a diagonal matrix with diagonal entries large enough to ensure the exploration of the entire solution space; to that end, we set the i^{th} diagonal component of $\Omega^{(0)}$, $\Omega_i^{(0)}$, to a value so that the range of that parameter is encompassed in the interval $\left(\xi_i^{(0)} - 2\sqrt{\Omega_i^{(0)}}, \xi_i^{(0)} + 2\sqrt{\Omega_i^{(0)}}\right)$. Therefore, the entire range is within two sampling standard deviations of the initial mean.

We choose the additional parameter values for the MRAS algorithms based on Hu et al. (2006): $\lambda = .01$, $\epsilon = 10^{-5}$, $\rho_0 = 80$, $N_0 = 200$, and $S(\ell(y, X)) = \exp\left(-\frac{\ell(y,X)}{1000}\right)$. Additionally, we use the following stopping criterion: $|\zeta_k - \zeta_{k-10}| \leq .1$, where ζ_k is the best log-likelihood value attained in the first k iterations. However, we also run the MRAS mixture model algorithm a minimum of 50 iterations to ensure that the algorithms are given enough time to steer away from the initial solution and begin converging to the optimal solution. In other words, the stopping criterion is enforced only after 50 iterations, stopping the methods when no further improvement in the best log-likelihood is attained in the last 10 iterations. To prevent degenerate clusters, we use the constant $c = .01$ for the constraint given by (4). Also, we restrict the maximum value of N_k in any iteration of MRAS to be 1000 to limit the computational expense of any single iteration.

4.2 Estimating a Mixture of Survey Responses

The data set consists of 152 responses in a survey of MBA students. In that survey, students were asked about their perceived desirability of 10 different cars. The cars ranged from minivans and hatchbacks, to sport utility vehicles and sports cars. Students rated the desirability of each car on a scale of 1-10. Thus, the resulting data set comprises of 152 10-dimensional vectors. The goal is to find segments in the market of all MBA students with respect to car preferences.

We illustrate our methods on this data set in the following way. Assuming $g = 3$ mixture components, we first standardize the data to make it more amenable to the Gaussian assumption. We run each algorithm 10 times. We compute the best ("max") and worst ("min") solution found. We also compute the average and standard error of the 10 solutions. We compute the percent improvement of MRAS's best solution over the best one found by EM. We also report the average number of iterations and the average time to convergence. Table 2 shows the results.

Table 2: Simulation results on the survey data set.

Algorithm	Max	Min	Mean	Std. Error	% improvement over EM	Avg iters	Avg time
EM	-1622.1	-1942.2	-1797.6	28.20	-	13.6	0.14
MRAS	-1435.5	-1886.1	-1620.4	32.87	11.50%	268.4	297.91

From the results in Table 2, we notice that significant improvements in the likelihood values can be gained with the MRAS mixture model algorithm as compared to EM. In particular, we notice that the best solution found by MRAS (-1435.5) is about 12% better than EM's best solution (-1622.1). Furthermore, the worst solution found by MRAS (-1886.1) is still better than the worst solution of EM (-1942.2). On average, MRAS performs better than EM, with a slightly larger standard error. On the computational side, a typical run of MRAS requires far more iterations until convergence, which is another indication that the method explores the parameter space more exhaustively. However, since no lunch is free, the increased computational effort also results in a much longer runtime for MRAS, on average 3 orders of magnitude longer than a single run of EM for this experiment.

In order to gauge what can be gained from the global optimum, consider the graph in Figure 4, which depicts the best set of clusters obtained by each of the two methods, with the data projected onto the

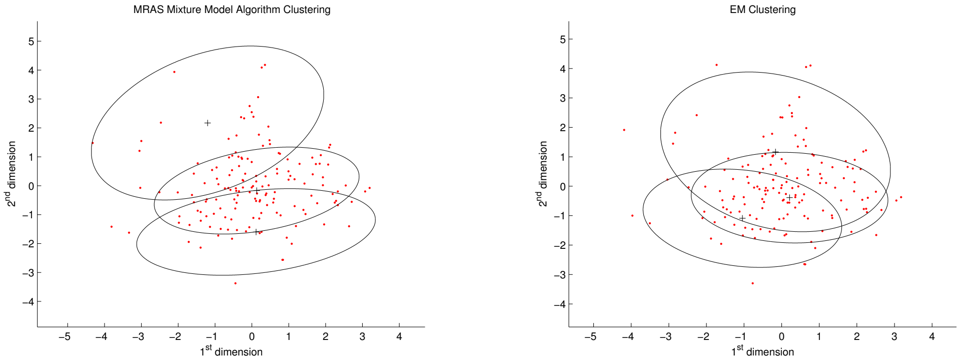

Figure 4: The best runs of MRAS mixture model algorithm and EM on the survey data set, with the data projected onto the first two principal components.

first two principal components. We can see that the best solution obtained by the MRAS mixture model algorithm (left panel) separates the data much better into 3 distinct clusters. In particular, in the left panel the cluster means are much better separated, as are the cluster shapes (i.e., the corresponding covariance matrices). The clusters in the left panel span the data set without the amount of overlap seen in the right panel. All-in-all, the cluster assignment corresponding to the best solution obtained by the MRAS mixture model algorithm appears to be supported much better by the observed data than that of the best solution obtained by EM.

5 Conclusion

In this paper we introduce the MRAS mixture model algorithm, designed to produce globally optimal solutions for Gaussian mixtures. The algorithm utilizes the Cholesky decomposition for the construction of the random covariance matrices. We present a proof of global convergence of the MRAS mixture model algorithm to the optimal solution for Gaussian mixtures. Our numerical experiment indicates that the proposed algorithm can find solutions missed by the classical EM, even when implemented using multiple starting points to compensate for its local convergence properties.

Perhaps the biggest current limitation of the MRAS mixture model algorithm is the computational time to convergence. Because it requires the generation of multiple candidate solutions in each iteration, the MRAS mixture model algorithm is a more computationally-intensive algorithm than EM. This is similar in nature to the Monte Carlo EM algorithm (Booth and Hobert, 1999; Levine and Casella, 2001; Levine and Fan, 2003; Jank, 2004; Caffo et al., 2003), which spends most of its computational effort on simulating from a suitable distribution and is thus much slower than its deterministic counterpart. That being said, our current implementation of the MRAS mixture model algorithm is not optimized for speed, and continuous advances in computing power and processor speed will make the computational disadvantages less practically important. At the end of the day, the decision that researchers faces is whether one wants fast but possibly highly inaccurate answers, or alternatively whether waiting a little longer is worth obtaining better solutions. The MRAS mixture model algorithm is one systematic way for finding those solutions.

References

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models. *Computational Statistics & Data Analysis*, 41(3-4):561–575.

Boer, P. D., Kroese, D., Mannor, S., and Rubinstein, R. (2005). A Tutorial on the Cross-Entropy Method. *Annals of Operations Research*, 134:19–67.

Booth, J. and Hobert, J. (1999). Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm. *Journal of the Royal Statistical Society B*, 61:265–285.

Botev, Z. and Kroese, D. (2004). Global Likelihood Optimization via the Cross-Entropy Method with an Application to Mixture Models. In *Proceedings of the 2004 Winter Simulation Conference*.

Caffo, B., Jank, W., and Jones, G. (2003). Ascent-Based Monte Carlo EM. *Journal of the Royal Statistical Society B*, 67:235–252.

Cao, G. and West, M. (1996). Practical Bayesian Inference Using Mixtures of Mixtures. *Biometrics*, 52:1334–1341.

Celeux, G. and Govaert, G. (1992). A Classification EM Algorithm for Clustering and Two Stochastic Versions. *Computational Statistics & Data Analysis*, 14(3):315–332.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, 39:1–22.

Diebolt, J. and Robert, C. (1990). Bayesian Estimation of Finite Mixture Distributions: Part II, Sampling Implementation. Technical Report III, Laboratoire de Statistique Théorique et Appliquée, Université Paris VI.

Fraley, C. and Raftery, A. (1998). How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. *The Computer Journal*, 41:578–588.

Heath, J., Fu, M., and Jank, W. (2007). New Global Optimization Algorithms for Model-Based Clustering. Working Paper, Smith School of Business, University of Maryland.

Horn, R. and Johnson, C. (1990). *Matrix Analysis.* Cambridge Univ. Press, New York.

Hu, J., Fu, M., and Marcus, S. (2006). A Model Reference Adaptive Search Algorithm for Global Optimization. Forthcoming in *Operations Research*.

Jank, W. (2004). Quasi-Monte Carlo Sampling to Improve the Efficiency of Monte Carlo EM. *Computational Statistics & Data Analysis*, 48:685–701.

Jank, W. (2006a). Ascent EM for Fast and Global Model-Based Clustering: An Application to Curve-Clustering of Online Auctions. *Computational Statistics & Data Analysis*, 51(2):747–761.

Jank, W. (2006b). The EM Algorithm, Its Stochastic Implementation and Global Optimization: Some Challenges and Opportunities for OR. In Alt, F., Fu, M., and Golden, B., editors, *Perspectives in Operations Research: Papers in Honor of Saul Gass' 80th Birthday*, pages 367–392. Springer, New York.

Johnson, C. (1970). Positive Definite Matrices. *The American Mathematical Monthly*, 77(3):259–264.

Kroese, D., Rubinstein, R., and Taimre, T. (2006). Application of the Cross-Entropy Method to Clustering and Vector Quantization. Forthcoming in *Journal of Global Optimization*.

Levine, R. and Casella, G. (2001). Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statistics*, 10:422–439.

Levine, R. and Fan, J. (2003). An Automated (Markov Chain) Monte Carlo EM Algorithm. *Journal of Statistical Computation and Simulation*, 74:349–359.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley, New York.

Pinheiro, J. and Bates, D. (1996). Unconstrained Parameterizations for Variance-Covariance Matrices. *Statistics and Computing*, 6:289–296.

Thisted, R. (1988). *Elements of Statistical Computing.* Chapman & Hall, London.

Tu, Y., Ball, M., and Jank, W. (2006). Estimating Flight Departure Delay Distributions - A Statistical Approach with Long-Term Trend and Short-Term Pattern. Forthcoming in the *Journal of the American Statistical Association*.

Wei, G. and Tanner, M. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85:699–704.

Wu, C. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103.