

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

December 12, 2006

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 the global optimization algorithm Model Reference Adaptive Search (MRAS), designed to produce globally-optimal solutions in the estimation of finite mixture models. We propose the MRAS mixture model algorithm for the estimation of Gaussian mixtures, which relies on the Cholesky decomposition to construct the random positive definite covariance matrices. In addition, we provide a theoretical proof of global convergence of the MRAS mixture model algorithm to the optimal solution for the maximization of the likelihood function of Gaussian mixtures. We conduct numerical experiments to evaluate the effectiveness of the proposed algorithms in comparison to classical EM.

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 the neural network or the mixture 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, especially 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. 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 repeatedly 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 only 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 (randomly chosen) starting values, and then select the parameter value that provides the best solution obtained from all runs (see Biernacki et al., 2003). While this approach can be very burdensome, especially when the parameter space is large, it also does not offer any systematic solution to the problem. More systematic approaches involve using *stochastic* versions of the EM algorithm such as the Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990). In that context, Jank (2006a) proposes a Genetic Algorithm version of the EM algorithm to overcome local solutions in the mixture likelihood (see also Jank, 2006b; Tu et al., 2006). Alternative approaches rely on producing ergodic Markov chains that, at convergence, have visited every point in the parameter space (see e.g. Diebolt and Robert, 1990; Cao and West, 1996; Celeux and Govaert, 1992). Yet an alternative approach which has been proposed only recently is to use methodology from the global optimization literature. 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.

Model Reference Adaptive Search (MRAS) is a method that was first proposed in the field of operations research and designed to attain globally optimal solutions to general multi-extremal continuous optimization problems (Hu et al., 2006). Similar to the Cross-Entropy method (De Boer et al., 2005) in nature, 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 (the “reference model”). The parameters of the new reference model are updated by minimizing the Kullback-Leibler (KL) divergence with respect to the current reference model and the top candidate solutions. In this way, the properties of the *best* candidates in each iteration are retained. Additionally, the updating scheme of MRAS leads to a more general framework in which theoretical convergence of a particular instantiated algorithm can be rigorously proved (Hu et al., 2006). In this paper we propose several modifications to the formulation of estimating Gaussian mixtures in order to apply MRAS to the mixture model setting.

Applying MRAS directly to Gaussian mixtures cannot be done directly; therefore, we modify the formulation of the estimation of Gaussian mixtures for our proposed MRAS mixture model algorithm. In particular, it is necessary to generate random positive definite covariance matrices for each candidate solution. We show that satisfying this constraint by directly simulating the covariance matrix components is inefficient; therefore, we update the Cholesky factorizations of the covariance matrices instead. Representing the covariance matrices by their corresponding Cholesky factorizations in the candidate vectors allows for unconstrained simulation of the covariance matrices in the MRAS mixture model algorithm.

A particularly attractive attribute of the MRAS mixture model algorithm is that it *provably converges* 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. The proof gives theoretical justification that the algorithm is not merely an ad-hoc heuristic, but is truly an algorithm for producing globally-optimal solutions to Gaussian mixtures.

The rest of the paper begins with the description of the mathematical framework of Gaussian mixtures in Section 2. We proceed in Section 3 by explaining MRAS, leading to our discussion of the proposed MRAS mixture model algorithm and its global convergence proof. In Section 4 we carry out numerical experiments

to investigate how the MRAS mixture model algorithm performs in the estimation of Gaussian mixtures with respect to the classical EM algorithm. We conclude and discuss future work in Section 5.

2 Gaussian Mixture Models

We begin by presenting the mathematical framework of Gaussian mixture models. Assume there are n observed data points, $x = \{x_1, \dots, x_n\}$, in some p -dimensional space. Assume that data is known to have been derived from g distinct probability distributions, weighted according to the vector $\pi = (\pi_1, \dots, \pi_g)$, where the weights sum to one. Each component of the mixture has an associated Gaussian probability density $f(\cdot; \psi_j)$, where the parameters $\psi_j = (\mu_j; \Sigma_j)$ consists of the mean vector μ_j and the covariance matrix Σ_j . The parameters of the model that need to be estimated are $\theta = (\pi; \psi)$; that is, both the weights and the Gaussian distribution parameters for each of the g components. We write the mixture density of the data point x_i as:

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

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

$$L(x, \theta) = \prod_{i=1}^n f(x_i; \theta),$$

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

$$\begin{aligned} \log L(x, \theta) &= \sum_{i=1}^n \log f(x_i; \theta) \\ &= \sum_{i=1}^n \log \sum_{j=1}^g \pi_j f(x_i; \psi_j). \end{aligned}$$

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.

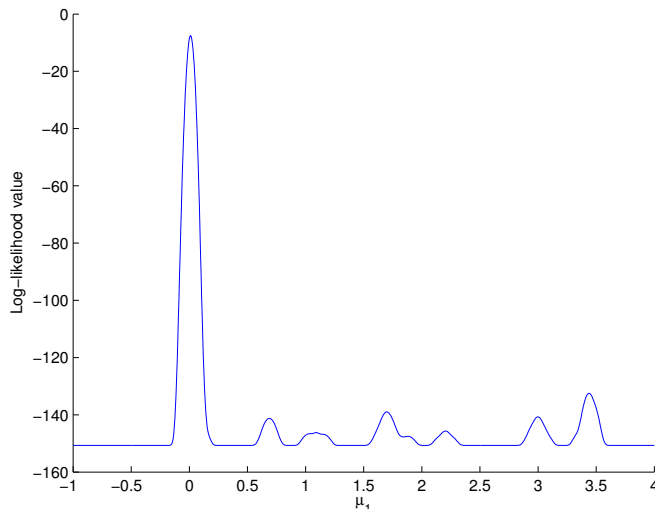


Figure 1: Plot of the log-likelihood function of the data set described above with parameters $\mu = (0, 2)$, $\sigma^2 = (.001, 1)$, and $\pi = (.5, .5)$, plotted against varying values of the mean component μ_1 .

Notice that in the following 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 Akaike information criterion (AIC) or Bayesian information criterion (BIC) (McLachlan and Peel, 2000).

3 Global Optimization in Gaussian Mixtures

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. We discuss one such method from the field of operations research that is designed specifically for global optimization, namely Model Reference Adaptive Search. Further, we expand on how MRAS can be applied 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:

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

The method that MRAS uses to achieve 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 by the following. We generate N_k candidate solutions, X_1, X_2, \dots, X_{N_k} , in each iteration k according to the sampling distribution and compute the value of the objective function $H(X_i)$ for each candidate X_i . The elite candidates in each iteration are selected by taking the top p -percentile candidates. The value of p 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 t be denoted as γ_t . If $\gamma_t < \gamma_{t-1}$, we increase p until $\gamma_t \geq \gamma_{t-1}$, effectively reducing the number of elite candidates. If, however, no such percentile p exists, then the number of candidates N is increased by a factor of α (where $\alpha > 1$), such that $N_{k+1} = \alpha N_k$. 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 $\mu_M^{(0)}, \mu_M^{(1)}, \dots$ will converge to the optimal vector X^* , representing the optimal solution x^* , as the sequence of the sampling covariance matrices $\Sigma_M^{(0)}, \Sigma_M^{(1)}, \dots$ converges to the zero matrix.

The sampling distribution we use is multivariate Gaussian, with corresponding sampling parameters of μ_M and Σ_M , representing the sampling mean and covariance matrix, respectively. Additionally, Hu et al. (2006) provide a proof for global convergence of MRAS in continuous optimization. In the following, we describe a version of MRAS suitable for mixture models.

3.2 MRAS algorithm for mixture models

Selecting the parameter vector (the candidate vectors) to optimize with MRAS in the estimation of Gaussian mixtures is non-trivial. Although a straightforward representation of the means and weight components is

possible, finding a suitable representation for the covariance matrices is less clear. Simulating random positive definite covariance matrices directly can be very inefficient (see e.g., Heath et al., 2007). We propose a method to simulate positive definite covariance matrices in the MRAS mixture model algorithm that relies on 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 spd 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 spd, each covariance matrix has a corresponding Cholesky factorization U . Therefore, one possible 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 itself. 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 spd.

One potential problem with this method is that the Cholesky factor for a symmetric positive definite matrix is not unique. For a Cholesky factor U of Σ , we can multiply any subset of rows of U by -1 and obtain a different Cholesky factor of the same Σ . Thus there is not a unique optimal X^* 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 factor U are required to be positive, then the Cholesky factor 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 choose to construct the covariance matrices in the MRAS mixture model algorithm by sampling the U_{ii} components from a truncated Gaussian distribution (accepting all values in the interval $(0, \infty)$).

MRAS can be applied to the estimation of Gaussian mixtures by sampling candidate solutions X_i that represent the parameters $(\mu_j, \Sigma_j, \pi_j)_{j=1}^g$, where the covariance matrices are represented by their corresponding Cholesky factorizations. We then score each candidate according to the log-likelihood function, and use the best-scoring candidates to update the sampling distribution. The goal is to obtain the optimal solution X^* representing the maximum likelihood estimate (μ^*, Σ^*, π^*) . Below we provide an outline of the MRAS mixture model algorithm. Note that $f(\cdot; \mu, \Sigma)$ is the multivariate Gaussian density, i.e.,

$$f(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right).$$

Other parameters are λ , a small constant, the multiplicative constant α , and $S(\cdot)$, a strictly increasing function on $\mathfrak{R} \rightarrow \mathfrak{R}^+$ (Hu et al., 2006).

MRAS Mixture Model Algorithm

1. Initialize $\mu_M^{(0)}$ and $\Sigma_M^{(0)}$. Set $t = 1$.
2. **repeat**
3. Generate N_{t-1} i.i.d. candidate vectors $X_1, \dots, X_{N_{t-1}}$ from the sampling distribution $f(\cdot; \mu_M^{(t-1)}, \Sigma_M^{(t-1)}) := (1 - \lambda)f(\cdot; \mu_M^{(t-1)}, \Sigma_M^{(t-1)}) + \lambda f(\cdot; \mu_M^{(0)}, \Sigma_M^{(0)})$.
4. Compute the log-likelihoods of the data with respect to each N_{t-1} candidate solution X_i .
5. Select the elite candidates by taking the top scoring p_{t-1} -percentile candidate vectors, and define $\tilde{\gamma}_t(p_{t-1})$ as the p_{t-1} -percentile log-likelihood score obtained of all candidates in iteration t .
6. If $t = 1$ or $\tilde{\gamma}_t(p_{t-1}) \geq \gamma_{t-1} + \frac{\epsilon}{2}$,
7. Set $\gamma_t = \tilde{\gamma}_t(p_{t-1})$, $p_t = p_{t-1}$ and $N_t = N_{t-1}$.
8. Else find the largest $\tilde{p} \in (p_{t-1}, 100)$ such that $\tilde{\gamma}_t(\tilde{p}) \geq \gamma_{t-1} + \frac{\epsilon}{2}$.
9. If such a \tilde{p} exists, then set $\gamma_t = \tilde{\gamma}_t(\tilde{p})$, $p_t = \tilde{p}$, and $N_t = N_{t-1}$.
10. Else set $\gamma_t = \gamma_{t-1}$, $p_t = p_{t-1}$, and $N_t = \alpha N_{t-1}$.
11. Update the sampling parameters according to:

$$\mu_M^{(t)} = \frac{\sum_{i=1}^{N_{t-1}} S(L(X_i))^{t-1} / f(X_i, \mu_M^{(t-1)}, \Sigma_M^{(t-1)}) I_{\{L(X_i) \geq \gamma_t\}} X_i}{\sum_{i=1}^{N_{t-1}} S(L(X_i))^{t-1} / f(X_i, \mu_M^{(t-1)}, \Sigma_M^{(t-1)}) I_{\{L(X_i) \geq \gamma_t\}}},$$

$$\Sigma_M^{(t)} = \frac{\sum_{i=1}^{N_{t-1}} S(L(X_i))^{t-1} / f(X_i, \mu_M^{(t-1)}, \Sigma_M^{(t-1)}) I_{\{L(X_i) \geq \gamma_t\}} (X_i - \mu_M^{(t)})(X_i - \mu_M^{(t)})^T}{\sum_{i=1}^{N_{t-1}} S(L(X_i))^{t-1} / f(X_i, \mu_M^{(t-1)}, \Sigma_M^{(t-1)}) I_{\{L(X_i) \geq \gamma_t\}}}.$$
12. Set $t = t + 1$.
13. **until** stopping criterion is met.
14. Return highest-scoring candidate solution obtained.

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.

3.3 Preventing Degenerate Clusters

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 the cluster mean be equal to a single data point, and then let the generalized variance, or determinant of the covariance matrix, of that cluster be arbitrarily small. In order to prevent this from

happening 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 generalized variances in the multivariate case.

Hathaway (1985) discusses the use of a constraint that limits the relative size of the variance components in the univariate Gaussian case, given by:

$$\min_{i,j}(\sigma_i/\sigma_j) \geq c > 0. \quad (1)$$

This constraint’s multivariate counterpart provides a consistent formulation of the multivariate Gaussian mixture model (McLachlan and Peel, 2000) and is given by the following:

$$\min_{i,j} \frac{|\Sigma_i|}{|\Sigma_j|} \geq c > 0. \quad (2)$$

To avoid degenerate constraints, we will use the following constraint:

$$\min_j |\Sigma_j| \geq c > 0. \quad (3)$$

However, in each of these constraints, determining the appropriate value of c is difficult when no prior information on the problem structure is known. Unless otherwise specified, we use a value of $c = .01$. If one of our algorithms generates a covariance matrix that violates this constraint, we discard it and re-generate a new one. In addition, if more than half of the candidates of a given cluster violate the constraint on a given iteration, we consider that cluster a “lost cause” and re-initialize that cluster’s parameters to the initial values used in iteration 1.

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 $f(\cdot, \theta)$ that belongs to the natural exponential family. Before we discuss the theorem, we provide the definition of the natural exponential family of distributions, as well as some required assumptions for the theorem.

Definition 1. *A parameterized family of p.d.f.’s $\{f(\cdot, \theta), \theta \in \Theta \subseteq \mathbb{R}^m\}$ on χ is said to belong to the natural*

exponential family if there exists functions $h(\cdot) : \mathfrak{R}^n \rightarrow \mathfrak{R}$, $\Gamma(\cdot) : \mathfrak{R}^n \rightarrow \mathfrak{R}^m$, and $K(\cdot) : \mathfrak{R}^m \rightarrow \mathfrak{R}$ such that

$$f(x, \theta) = \exp\{\theta^T \Gamma(x) - K(\theta)\} h(x), \quad \forall \theta \in \Theta,$$

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

The following assumptions are referenced in the statements of Theorem 2, Corollary 1, or both.

Assumptions:

- A1. For any given constant $\xi < H(x^*)$, the set $\{x : H(x) \geq \xi\} \cap \mathcal{X}$ has a strictly positive Lebesgue measure.
- A2. For any given constant $\delta > 0$, $\sup_{x \in A_\delta} H(x) < H(x^*)$, where $A_\delta := \{x : \|x - x^*\| \geq \delta\} \cap \mathcal{X}$, where the supremum over the empty set is defined to be $-\infty$.
- A3. There exists a compact set Π_ϵ such that $\{x : H(x) \geq H(x^*) - \epsilon\} \cap \mathcal{X} \subseteq \Pi_\epsilon$. Moreover, $f(x, \theta_0)$ is bounded away from zero on Π_ϵ , i.e., $f_* = \inf_{x \in \Pi_\epsilon} f(x, \theta_0) > 0$.
- A4. The parameter vector $\tilde{\theta}_{k+1}$ computed in MRAS is an interior point of Θ for all k .
- A5. $\sup_{\theta \in \Theta} \|\exp\{\theta^T \Gamma(x)\} \Gamma(x) h(x)\|$ is integrable/summable with respect to x , where θ , $\Gamma(\cdot)$, and $h(\cdot)$ are defined as in Definition 1.

In Theorem 2, Hu et al. (2006) show that for a sampling distribution $f(\cdot, \theta)$ belonging to the natural exponential family as defined in Definition 1, MRAS will converge to the unique optimal solution x^* .

Theorem 2. *Let $\epsilon > 0$, and define the ϵ -optimal set $\mathcal{O}_\epsilon := \{x : H(x) \geq H(x^*) - \epsilon\} \cap \mathcal{X}$. If assumptions A1, A3, A4, and A5 are satisfied, then w.p. 1 there exists a random variable $\kappa < \infty$ such that*

1. $\gamma_k > H(x^*) - \epsilon, \forall k \geq \kappa$,
2. $E_{\theta_{k+1}}[\Gamma(X)] \in \text{CONV}\{\Gamma(\mathcal{O}_\epsilon)\}, \forall k \geq \kappa$ w.p. 1, where $\text{CONV}\{\Gamma(\mathcal{O}_\epsilon)\}$ indicates the convex hull of the set $\Gamma(\mathcal{O}_\epsilon)$.

Furthermore, let β be a positive constant satisfying the condition that the set $\{x : S(H(X)) \geq \frac{1}{\beta}\}$ has a strictly positive Lebesgue measure. If assumptions A1, A2, A3, A4, and A5 are satisfied and $\alpha > (\beta S^*)^2$, where $S^* := S(H(x^*))$, then

3. $\lim_{k \rightarrow \infty} E_{\theta_k}[\Gamma(X)] = \Gamma(x^*)$ w.p. 1.

The following corollary directly follows Theorem 2, and shows global convergence of MRAS to the optimal solution x^* when using the multivariate normal sampling distribution. As the number of iterations tends to infinity, the sampling distribution tends toward a degenerate distribution centered on the optimal solution x^* .

Corollary 1. *If multivariate normal p.d.f.'s are used in MRAS, i.e.,*

$$f(x, \theta_k) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right),$$

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

$$\lim_{k \rightarrow \infty} \tilde{\mu}_k = x^*, \text{ and } \lim_{k \rightarrow \infty} \tilde{\Sigma}_k = 0_{n \times n} \text{ w.p. } 1$$

3.5 Proving Global Convergence of the MRAS mixture model algorithm

In order to show that Corollary 1 applies to the MRAS mixture model 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:

$$\ell(x, \theta) = \log L(x, \theta) = \sum_{i=1}^n \log \sum_{j=1}^g \pi_j f(x_i; \mu_j, \Sigma_j).$$

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

$$\theta^* \in \underset{\theta \in \chi}{\operatorname{argmax}} \ell(x, \theta).$$

Before we prove the global convergence of MRAS mixture model algorithm, we first provide the following useful lemmas. Lemma 1 shows that for a continuous function f that is bounded above and possesses a unique optimal maximizer on a space χ , then f satisfies Assumption A1 on χ .

Lemma 1. *For a continuous function $f(x)$, $x \in \chi \in \mathbb{R}^n$, where f is bounded above and there exists a unique optimal solution x^* s.t. $f(x) < f(x^*)$, $\forall x \neq x^*$, then $\forall \xi < f(x^*)$, the set $\{x : f(x) \geq \xi\}$ has strictly positive*

Lebesgue measure, and thereby f satisfies Assumption A1.

Proof: Choose $\xi < f(x^*)$ and let $\epsilon = f(x^*) - \xi$. By continuity of f , $\exists \delta > 0$ s.t. $\forall x \in \{x : \|x - x^*\| < \delta\}$, then $|f(x) - f(x^*)| < \epsilon$. By rewriting the left- and right-hand sides of the inequality, we see that $f(x^*) - f(x) < f(x^*) - \xi$, i.e., $\xi < f(x)$. Therefore, the set $\{x : \|x - x^*\| \leq \frac{\delta}{2}\} \subseteq \{x : f(x) \geq \xi\}$. And so, because the set $\{x : \|x - x^*\| \leq \frac{\delta}{2}\}$ has strictly positive Lebesgue measure, then the set $\{x : f(x) \geq \xi\}$ must as well. \square

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, B ,

$$\det(\alpha A + (1 - \alpha)B) \geq (\det A)^\alpha (\det B)^{1-\alpha},$$

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 $c > 0$, then cA and $A + B$ are both positive definite as well.

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

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

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. Assume true for k , i.e., $\det\left(\sum_{j=1}^k \alpha_j A_j\right) \geq \prod_{j=1}^k (\det A_j)^{\alpha_j}$, for any set $\{\alpha_j\}_{j=1}^k$ s.t. $\alpha_j > 0$ and $\sum_{j=1}^k \alpha_j = 1$.
- iii. Show true for $k + 1$, i.e., show $\det\left(\sum_{j=1}^{k+1} \tilde{\alpha}_j A_j\right) \geq \prod_{j=1}^{k+1} (\det A_j)^{\tilde{\alpha}_j}$, for any set $\{\tilde{\alpha}_j\}_{j=1}^{k+1}$ s.t. $\tilde{\alpha}_j > 0$ and $\sum_{j=1}^{k+1} \tilde{\alpha}_j = 1$.

Define $\alpha_j = \frac{\tilde{\alpha}_j}{1 - \tilde{\alpha}_{k+1}}$, for $j = 1, \dots, k$. Thus, $\sum_{j=1}^k \alpha_j = 1$ and then,

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

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

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 $|U_j^T U_j| \geq c > 0$, $j = 1, \dots, g$, 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 requirement that Assumption A2 requires is that the optimal solution θ^* be unique. By restricting the diagonal components of the Cholesky factorization U to be positive, its corresponding 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-

identifiable formulation. To avoid this problem, we add the following constraint to the problem:

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

where $\mu_i(1)$ represents the 1st mean component of the i^{th} cluster. Although the inequalities in this constraint are not strict, the probability of multiple mean components being exactly equal is zero, when considering continuous random data. Therefore, this constraint mandates a unique ordering of the clusters of θ^* , resulting in a unique optimal solution to the clustering problem.

Furthermore, we choose to bound the candidate means and Cholesky factorization components within a compact space based on the data points in order to ease the showing of Assumption A3 in the proof of Theorem 3. In particular, bound the means by defining x_{min} as the minimum value over all components of the data points X_1, \dots, X_n . That is, $x_{min(i)} = \min_{j=1, \dots, n} X_{j(i)}$. Similarly, we define x_{max} as the maximum value over all components of the data points, i.e., $x_{max(i)} = \max_{j=1, \dots, n} X_{j(i)}$. We note that bounding the candidate mean components by x_{min} and x_{max} is not an unreasonable constraint; clearly, the means of the optimal clusters will not lie outside of the data points themselves.

We place constraints on the components of the Cholesky factorizations by first calculating the sample variance of the data set, $Var(\{X_1, X_2, \dots, X_n\})$. We then choose the maximum across all p components, i.e., $V_{max} = \max_{i=1, \dots, p} Var(\{X_1, X_2, \dots, X_n\})$. And so, V_{max} represents an upper bound for the variance component of any cluster. Constraining the diagonal components within the bounds $[0, V_{max}]$ and the off-diagonal non-zero components within $[-V_{max}, V_{max}]$ suffices, as a cluster can have a variance component at least as large as V_{max} .

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

$$\chi = \left\{ \begin{array}{ll} \mu_j \in \mathbb{R}^p, & j = 1, \dots, g \\ \text{s.t. } \mu_j \in [x_{min}, x_{max}], & j = 1, \dots, g \\ \text{and } \mu_1(1) \leq \mu_2(1) \leq \dots \leq \mu_g(1) & \\ U_j \in \mathbb{R}^p \times \mathbb{R}^p, & j = 1, \dots, g \\ \text{s.t. } U_{j(ik)} \begin{cases} \in [0, V_{max}], i = k \\ \in [-V_{max}, V_{max}], i < k \\ = 0, i > k \end{cases} & j = 1, \dots, g; i, k = 1, \dots, p \\ \text{and } \prod_{i=1}^p U_{j(ii)} \geq \sqrt{c} > 0, & j = 1, \dots, g \\ \pi_j \in \mathbb{R}, & j = 1, \dots, g \\ \text{s.t. } \pi_j \geq 0, & j = 1, \dots, g \\ \text{and } \sum_{j=1}^g \pi_j = 1 & \end{array} \right. \quad (4)$$

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 $c := g \frac{(p+1)(p+2)}{2}$. So, we can consider the space χ to be c -dimensional. The MRAS sampling parameters $\theta_M = (\mu_M, \Sigma_M)$ belong to the space Θ , where $\Theta = (\mu_M \in \chi, \Sigma_M \text{ is spd})$.

Lemma 4. *The subspace $\chi \subseteq \mathbb{R}^c$ is compact.*

Proof: For the vector $X \in \chi$, clearly all components of X are bounded: the components of π are bounded by 0 and 1, the components of μ are bounded by x_{min} and x_{max} , and the components of U are bounded according to 4. Therefore, if we can show that the space χ is closed, then we will obtain compactness as well. \square

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

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 $f(x)$ is a continuous function, it achieves its supremum on the compact space A_δ , i.e., $\exists \tilde{x} \in A_\delta$ s.t.

$$\sup_{x \in A_\delta} f(x) = f(\tilde{x}).$$

And, because $f(x) < f(x^*)$, $\forall x \neq x^*$, we have:

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

Now we give Theorem 3, where we show that Corollary 1 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 $S^* := S(\ell(x, \theta^*))$, and we are optimizing over the constrained space χ denoted by 4, then the sampling parameters of MRAS mixture model algorithm will converge as follows:*

$$\lim_{k \rightarrow \infty} \tilde{\mu}_M^{(k)} = \theta^*, \text{ and } \lim_{k \rightarrow \infty} \tilde{\Sigma}_M^{(k)} = 0_{c \times c} \text{ w.p. } 1,$$

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

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 $\ell(x, \theta)$ is continuous on χ w.r.t. θ , then by Lemma 1, for any $\xi < \ell(x, \theta^*)$, the set $\{x : \ell(x, \theta) \geq \xi\} \cap \chi$ has a strictly positive Lebesgue measure. Thus, Assumption A1 is satisfied.
- ii. By Lemma 5, since $\ell(x, \theta)$ is continuous on χ w.r.t. θ , then $\forall \delta > 0$, $\sup_{\theta \in A_\delta} \ell(x, \theta) < \ell(x, \theta^*)$, 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 $\{\theta : \ell(x, \theta) \geq \ell(x, \theta^*) - \epsilon\} \cap \chi$ is a subset of a compact set, namely, χ itself. Moreover, using Gaussian as the 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:

$$\mu_M^{(t)} = \frac{\sum_{i=1}^{N_{t-1}} S(L(X_i))^{t-1} / f(X_i, \mu_M^{(t-1)}, \Sigma_M^{(t-1)}) I_{\{L(X_i) \geq \gamma_t\}} X_i}{\sum_{i=1}^{N_{t-1}} S(L(X_i))^{t-1} / f(X_i, \mu_M^{(t-1)}, \Sigma_M^{(t-1)}) I_{\{L(X_i) \geq \gamma_t\}}},$$

$$\Sigma_M^{(t)} = \frac{\sum_{i=1}^{N_{t-1}} S(L(X_i))^{t-1} / f(X_i, \mu_M^{(t-1)}, \Sigma_M^{(t-1)}) I_{\{L(X_i) \geq \gamma_t\}} (X_i - \mu_M^{(t)})(X_i - \mu_M^{(t)})^T}{\sum_{i=1}^{N_{t-1}} S(L(X_i))^{t-1} / f(X_i, \mu_M^{(t-1)}, \Sigma_M^{(t-1)}) I_{\{L(X_i) \geq \gamma_t\}}}$$

It is evident that the mean of the sampling distribution, $\mu_M^{(t)}$, is simply a convex combination of the elite candidates. Since each candidate $X_i \in \chi$, i.e., 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, $|U_j| \geq \sqrt{c} > 0$, $j = 1, \dots, g$, which we now address.

We need to show that a convex combination of the top k candidates also satisfies this constraint, namely $|\sum_{j=1}^k \alpha_j U_j| \geq \sqrt{c}$. We note that as a direct application of Lemma 3, $|\sum_{j=1}^k \alpha_j U_j| \geq \prod_{j=1}^k |U_j|^{\alpha_j} \geq \min_j |U_j| \geq \sqrt{c}$. 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 $f(\cdot; \mu_M^{(t-1)}, \Sigma_M^{(t-1)})$, then with probability one each candidate lies in the interior of χ . Therefore, the updated mean vector $\mu_M^{(t)}$ will also lie in the interior of the space. Also, the updated $\Sigma_M^{(t)}$ is clearly spd by construction, and thus A4 is satisfied. \square

4 Numerical Experiments

In the following numerical experiment, we demonstrate the performance of the MRAS mixture model algorithm in comparison with the original EM algorithm. The experiment was performed in Matlab and run on a 2.80 GHz Intel with 1 GB RAM. This experiment is given simply to provide empirical evidence that the MRAS mixture model algorithm has the potential to produce better solutions than EM in practice. A more comprehensive computational study on the performance of the MRAS mixture model algorithm and other mixture model algorithms can be seen in (Heath et al., 2007).

4.1 Initial Parameters

For the EM algorithm we use 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 $|\gamma_k - \gamma_{k-1}| \leq 10^{-5}$, where γ_k is the log-likelihood

value obtained in iteration k .

One of the benefits of MRAS is that its performance is virtually independent of its starting values for many practical purposes (Hu et al., 2006). We initialize the parameters $\mu_M^{(0)}$ 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 parameters $\Sigma_M^{(0)}$ 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 $\Sigma_M^{(0)}$, $\Sigma_{M(i)}^{(0)}$, to a value so that the range of that parameter is encompassed in the interval $\left(\mu_{M(i)}^{(0)} - 2\sqrt{\Sigma_{M(i)}^{(0)}}, \mu_{M(i)}^{(0)} + 2\sqrt{\Sigma_{M(i)}^{(0)}}\right)$. 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): we set $\lambda = .01$, $\epsilon = 10^{-5}$, $p_0 = 80$, $N_0 = 200$, and $S(L(X)) = \exp -\frac{L(X)}{1000}$. Additionally, we use the following stopping criterion: $|\gamma_k - \gamma_{k-10}| \leq .1$, where γ_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. 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 Clustering 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 SUVs 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 of clustering 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 $g = 3$ clusters, we first standardize the data to make it more amenable to the Gaussian assumption. We run each method 10 times. Table 1 shows the results.

From the results in Table 1, we notice that significant improvements in the likelihood values can be gained with the MRAS mixture model model algorithm as compared to EM. The improvement over EM

Table 1: 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	0	13.6	0.14
MRAS	-1435.5	-1886.1	-1620.4	32.87	11.50%	268.4	297.91

ranges between 10% and 20%. However, computational time is sacrificed to obtain the solutions produced by the MRAS mixture model algorithm, as a single run of EM is approximately 3 orders of magnitude faster than a single run of MRAS in this experiment.

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 performs well empirically. In fact, we show that the MRAS mixture model algorithm has the potential for significant gains of up to 10% over solutions obtained with classical EM, even when that latter is implemented using multiple starting points to compensate for its local convergence properties.

Perhaps the biggest limitation of the MRAS mixture model algorithm is the computational time to convergence. Because it requires the generation of multiple candidates in each iteration, the MRAS mixture model algorithm is by nature a more computationally-intensive algorithm than EM. That being said, our implementation of the MRAS mixture model algorithm has not been optimized, so the computational times could be improved. However, as typical with all global optimization algorithms, one must consider the speed vs. accuracy trade-off. In other words, the quickest answer is not necessarily the best answer.

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*, 413-4:561-575.

- Booth, J. G. and Hobert, J. P. (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. P. (2004). Global Likelihood Optimization via the Cross-Entropy Method with an Application to Mixture Models. *Proceedings of the 2004 Winter Simulation Conference*.
- Caffo, B. S., Jank, W. S., and Jones, G. L. (2005). Ascent-Based Monte Carlo EM. *Journal of the Royal Statistical Society, Series 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 version. *Computational Statistics & Data Analysis*, 143:315–332.
- De Boer, P., Kroese, D. P., Mannor, S., and Rubinstein R. Y. (2005). A Tutorial on the Cross-Entropy Method. *Annals of Operations Research*, 134:19–67.
- Diebolt, J. and Robert, C. P. (1990). Bayesian estimation of finite mixture distributions: Part II, Sampling implementation. *Technical Report III*. Paris: Laboratoire de Statistique Théorique et Appliquée, Université Paris VI.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, 39:1–22.
- Fraley, C. and A.E. Raftery. (1998). How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. *Technical Report No. 329*, Dept. Statistics, Univ. Washington, Seattle, WA.
- Glover, F. (1990). Tabu Search: A Tutorial. *Interfaces*, 20:74–94.
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Reading, MA: Addison-Wesley.
- Hathaway, R. J. (1985). A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. *The Annals of Statistics*, 132:795–800.
- Heath, J. Fu, M. and Jank, W. (2007). New Global Optimization Algorithms for Model-Based Clustering. Working paper.

- Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*, New York: Cambridge Univ. Press.
- Hu, J., Fu, M. C., and Marcus, S. I. (2006). A Model Reference Adaptive Search Algorithm for Global Optimization. Forthcoming in *Operations Research*.
- Jamshidian, M. and Jennrich, R. I. (1993). Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association*, 88:221–228.
- Jamshidian, M. and Jennrich, R. I. (1997). Acceleration of the EM algorithm by using Quasi-Newton methods. *Journal of the Royal Statistical Society B*, 59:569–587.
- Jank, W. (2004). Quasi-Monte Carlo Sampling to Improve the Efficiency of Monte Carlo EM. *Computational Statistics and Data Analysis*, 48:685–701.
- Jank, W. (2006). Ascent EM for Fast and Global Model-Based Clustering: An Application to Curve-Clustering of Online Auctions. Forthcoming in *Computational Statistics and Data Analysis*.
- Jank, W. (2006). The EM Algorithm, Its Stochastic Implementation and Global Optimization: Some Challenges and Opportunities for OR. Forthcoming in Alt, Fu, and Golden (Eds.) *Perspectives in Operations Research: Papers in Honor of Saul Gass' 80th Birthday*, New York: Springer.
- Johnson, C. R. (1970). Positive Definite Matrices. *The American Mathematical Monthly*, **77**:259-264.
- Kroese, D. P., Porotsky, S., and Rubinstein, R. Y. (2004). The Cross-Entropy Method for Continuous Multi-Extremal Optimization.
- Kroese, D. P., Rubinstein, R. Y., 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. (2004). An automated (Markov Chain) Monte Carlo EM algorithm. *Journal of Statistical Computation and Simulation*, 74:349–359.
- Liu, C. and Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81:633–648.

- Louis, T. A. (1982). Finding the Observed Information Matrix when using the EM algorithm. *Journal of the Royal Statistical Society B*, 44:226–233.
- McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*, New York: Wiley.
- Meng, X.-L. (1994). On the rate of convergence of the ECM algorithm. *The Annals of Statistics*, 22:326–339.
- Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267–278.
- Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parameterizations for variance-covariance matrices. *Statistics and Computing*, 6:289–296.
- Reeves, C. R. (1997). Genetic algorithms for the operations researcher. *INFORMS Journal on Computing*, **9** 3:231–250.
- Rubinstein, R. Y. (1997). Optimization of Computer Simulation Models with Rare Events. *European Journal of Operations Research*, 99:89–112.
- Rubinstein, R. Y. (2005). The Simulated Entropy Method for Combinatorial and Continuous Optimization. *Methodology and Computing in Applied Probability*, 2:127–190.
- Rubinstein, R. Y., and Kroese, D. P. (2005). *The cross-entropy method: A unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*, New York: Springer-Verlag.
- Thisted, R. A. (1988). *Elements of Statistical Computing*, London: Chapman & Hall.
- Tu Y., Ball M., and Jank W. (2006). Estimating Flight Departure Delay Distributions - A Statistical Approach with Long-Term Trend and Short-Term Pattern. *Robert H. Smith School Research Paper No. RHS 06-034*, available at SSRN: <http://ssrn.com/abstract=923628>.
- Wei, G. C. G. and Tanner, M. A. (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. F. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, **11** 1:95–103.