

On Convergence Properties of the EM Algorithm for Gaussian Mixtures

Lei Xu

*Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology, Cambridge, MA 02139 USA and
Department of Computer Science, The Chinese University of Hong Kong, Hong Kong*

Michael I. Jordan

*Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology, Cambridge, MA 02139 USA*

We build up the mathematical connection between the “Expectation-Maximization” (EM) algorithm and gradient-based approaches for maximum likelihood learning of finite gaussian mixtures. We show that the EM step in parameter space is obtained from the gradient via a projection matrix P , and we provide an explicit expression for the matrix. We then analyze the convergence of EM in terms of special properties of P and provide new results analyzing the effect that P has on the likelihood surface. Based on these mathematical results, we present a comparative discussion of the advantages and disadvantages of EM and other algorithms for the learning of gaussian mixture models.

1 Introduction

The “Expectation-Maximization” (EM) algorithm is a general technique for maximum likelihood (ML) or maximum a posteriori (MAP) estimation. The recent emphasis in the neural network literature on probabilistic models has led to increased interest in EM as a possible alternative to gradient-based methods for optimization. EM has been used for variations on the traditional theme of gaussian mixture modeling (Ghahramani and Jordan 1994; Nowlan 1991; Xu and Jordan 1993a,b; Tresp *et al.* 1994; Xu *et al.* 1994) and has also been used for novel chain-structured and tree-structured architectures (Bengio and Frasconi 1995; Jordan and Jacobs 1994). The empirical results reported in these papers suggest that EM has considerable promise as an optimization method for such architectures. Moreover, new theoretical results have been obtained that link EM to other topics in learning theory (Amari 1994; Jordan and Xu 1995; Neal and Hinton 1993; Xu and Jordan 1993c; Yuille *et al.* 1994).

Despite these developments, there are grounds for caution about the promise of the EM algorithm. One reason for caution comes from con-

sideration of theoretical convergence rates, which show that EM is a first-order algorithm.¹ More precisely, there are two key results available in the statistical literature on the convergence of EM. First, it has been established that under mild conditions EM is guaranteed to converge toward a local maximum of the log likelihood l (Boyles 1983; Dempster *et al.* 1977; Redner and Walker 1984; Wu 1983). (Indeed the convergence is monotonic: $l(\Theta^{(k+1)}) \geq l(\Theta^{(k)})$, where $\Theta^{(k)}$ is the value of the parameter vector Θ at iteration k .) Second, considering EM as a mapping $\Theta^{(k+1)} = M(\Theta^{(k)})$ with fixed point $\Theta^* = M(\Theta^*)$, we have $\Theta^{(k+1)} - \Theta^* \approx [\partial M(\Theta^*) / \partial \Theta^*](\Theta^{(k)} - \Theta^*)$ when $\Theta^{(k+1)}$ is near Θ^* , and thus

$$\|\Theta^{(k+1)} - \Theta^*\| \leq \left\| \frac{\partial M(\Theta^*)}{\partial \Theta^*} \right\| \cdot \|\Theta^{(k)} - \Theta^*\|$$

with

$$\left\| \frac{\partial M(\Theta^*)}{\partial \Theta^*} \right\| \neq 0$$

almost surely. That is, EM is a first-order algorithm.

The first-order convergence of EM has been cited in the statistical literature as a major drawback. Redner and Walker (1984), in a widely cited article, argued that superlinear (quasi-Newton, method of scoring) and second-order (Newton) methods should generally be preferred to EM. They reported empirical results demonstrating the slow convergence of EM on a gaussian mixture model problem for which the mixture components were not well separated. These results did not include tests of competing algorithms, however. Moreover, even though the convergence toward the “optimal” parameter values was slow in these experiments, the convergence in likelihood was rapid. Indeed, Redner and Walker acknowledge that their results show that . . . “even when the component populations in a mixture are poorly separated, the EM algorithm can be expected to produce in a very small number of iterations parameter values such that the mixture density determined by them reflects the sample data very well.” In the context of the current literature on learning, in which the predictive aspect of data modeling is emphasized at the expense of the traditional Fisherian statistician’s concern over the “true” values of parameters, such rapid convergence in likelihood is a major desideratum of a learning algorithm and undercuts the critique of EM as a “slow” algorithm.

¹For an iterative algorithm that converges to a solution Θ^* , if there is a real number γ_0 and a constant integer k_0 , such that for all $k > k_0$, we have

$$\|\Theta^{(k+1)} - \Theta^*\| \leq q \|\Theta^{(k)} - \Theta^*\|^{\gamma_0}$$

with q being a positive constant independent of k , then we say that the algorithm has a convergence rate of order γ_0 . Particularly, an algorithm has first-order or linear convergence if $\gamma_0 = 1$, superlinear convergence if $1 < \gamma_0 < 2$, and second-order or quadratic convergence if $\gamma_0 = 2$.

In the current paper, we provide a comparative analysis of EM and other optimization methods. We emphasize the comparison between EM and other first-order methods (gradient ascent, conjugate gradient methods), because these have tended to be the methods of choice in the neural network literature. However, we also compare EM to superlinear and second-order methods. We argue that EM has a number of advantages, including its naturalness at handling the probabilistic constraints of mixture problems and its guarantees of convergence. We also provide new results suggesting that under appropriate conditions EM may in fact approximate a superlinear method; this would explain some of the promising empirical results that have been obtained (Jordan and Jacobs 1994), and would further temper the critique of EM offered by Redner and Walker. The analysis in the current paper focuses on unsupervised learning; for related results in the supervised learning domain see Jordan and Xu (1995).

The remainder of the paper is organized as follows. We first briefly review the EM algorithm for gaussian mixtures. The second section establishes a connection between EM and the gradient of the log likelihood. We then present a comparative discussion of the advantages and disadvantages of various optimization algorithms in the gaussian mixture setting. We then present empirical results suggesting that EM regularizes the condition number of the effective Hessian. The fourth section presents a theoretical analysis of this empirical finding. The final section presents our conclusions.

2 The EM Algorithm for Gaussian Mixtures

We study the following probabilistic model:

$$P(x | \Theta) = \sum_{j=1}^K \alpha_j P(x | m_j, \Sigma_j) \quad (2.1)$$

and

$$P(x | m_j, \Sigma_j) = \frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} e^{-1/2(x-m_j)^T \Sigma_j^{-1}(x-m_j)}$$

where $\alpha_j \geq 0$ and $\sum_{j=1}^K \alpha_j = 1$, d is the dimension of x . The parameter vector Θ consists of the mixing proportions α_j , the mean vectors m_j , and the covariance matrices Σ_j .

Given K and given N independent, identically distributed samples $\{x^{(t)}\}_1^N$, we obtain the following log likelihood:²

$$l(\Theta) = \log \prod_{t=1}^N P(x^{(t)} | \Theta) = \sum_{t=1}^N \log P(x^{(t)} | \Theta) \quad (2.2)$$

²Although we focus on maximum likelihood (ML) estimation in this paper, it is straightforward to apply our results to maximum a posteriori (MAP) estimation by multiplying the likelihood by a prior.

which can be optimized via the following iterative algorithm (see, e.g., Dempster *et al.* 1977):

$$\begin{aligned}
 \alpha_j^{(k+1)} &= \frac{\sum_{t=1}^N h_j^{(k)}(t)}{N} \\
 m_j^{(k+1)} &= \frac{\sum_{t=1}^N h_j^{(k)}(t) x^{(t)}}{\sum_{t=1}^N h_j^{(k)}(t)} \\
 \Sigma_j^{(k+1)} &= \frac{\sum_{t=1}^N h_j^{(k)}(t) [x^{(t)} - m_j^{(k)}][x^{(t)} - m_j^{(k)}]^T}{\sum_{t=1}^N h_j^{(k)}(t)}
 \end{aligned} \tag{2.3}$$

where the posterior probabilities $h_j^{(k)}$ are defined as follows:

$$h_j^{(k)}(t) = \frac{\alpha_j^{(k)} P(x^{(t)} \mid m_j^{(k)}, \Sigma_j^{(k)})}{\sum_{i=1}^K \alpha_i^{(k)} P(x^{(t)} \mid m_i^{(k)}, \Sigma_i^{(k)})}$$

3 Connection between EM and Gradient Ascent

In the following theorem we establish a relationship between the gradient of the log likelihood and the step in parameter space taken by the EM algorithm. In particular we show that the EM step can be obtained by premultiplying the gradient by a positive definite matrix. We provide an explicit expression for the matrix.

Theorem 1. *At each iteration of the EM algorithm equation 2.3, we have*

$$\mathcal{A}^{(k+1)} - \mathcal{A}^{(k)} = P_{\mathcal{A}}^{(k)} \frac{\partial l}{\partial \mathcal{A}} \Big|_{\mathcal{A}=\mathcal{A}^{(k)}} \tag{3.1}$$

$$m_j^{(k+1)} - m_j^{(k)} = P_{m_j}^{(k)} \frac{\partial l}{\partial m_j} \Big|_{m_j=m_j^{(k)}} \tag{3.2}$$

$$\text{vec}[\Sigma_j^{(k+1)}] - \text{vec}[\Sigma_j^{(k)}] = P_{\Sigma_j}^{(k)} \frac{\partial l}{\partial \text{vec}[\Sigma_j]} \Big|_{\Sigma_j=\Sigma_j^{(k)}} \tag{3.3}$$

where

$$P_{\mathcal{A}}^{(k)} = \frac{1}{N} \{\text{diag}[\alpha_1^{(k)}, \dots, \alpha_K^{(k)}] - \mathcal{A}^{(k)} (\mathcal{A}^{(k)})^T\} \tag{3.4}$$

$$P_{m_j}^{(k)} = \frac{\Sigma_j^{(k)}}{\sum_{t=1}^N h_j^{(k)}(t)} \tag{3.5}$$

$$P_{\Sigma_j}^{(k)} = \frac{2}{\sum_{t=1}^N h_j^{(k)}(t)} \Sigma_j^{(k)} \otimes \Sigma_j^{(k)} \tag{3.6}$$

where \mathcal{A} denotes the vector of mixing proportions $[\alpha_1, \dots, \alpha_K]^T$, j indexes the mixture components ($j = 1, \dots, K$), k denotes the iteration number, “ $\text{vec}[B]$ ” denotes the vector obtained by stacking the column vectors of the matrix B , and “ \otimes ” denotes the Kronecker product. Moreover, given the constraints $\sum_{j=1}^K \alpha_j^{(k)} = 1$ and $\alpha_j^{(k)} \geq 0$, $P_{\mathcal{A}}^{(k)}$ is a positive definite matrix and the matrices $P_{m_j}^{(k)}$ and $P_{\Sigma_j}^{(k)}$ are positive definite with probability one for N sufficiently large.

The proof of this theorem can be found in the Appendix.

Using the notation $\Theta = [m_1^T, \dots, m_K^T, \text{vec}[\Sigma_1]^T, \dots, \text{vec}[\Sigma_K]^T, \mathcal{A}^T]^T$, and $P(\Theta) = \text{diag}[P_{m_1}, \dots, P_{m_K}, P_{\Sigma_1}, \dots, P_{\Sigma_K}, P_{\mathcal{A}}]$, we can combine the three updates in Theorem 1 into a single equation:

$$\Theta^{(k+1)} = \Theta^{(k)} + P(\Theta^{(k)}) \left. \frac{\partial l}{\partial \Theta} \right|_{\Theta = \Theta^{(k)}} \quad (3.7)$$

Under the conditions of Theorem 1, $P(\Theta^{(k)})$ is a positive definite matrix with probability one. Recalling that for a positive definite matrix B , we have $(\partial l / \partial \Theta)^T B (\partial l / \partial \Theta) > 0$, we have the following corollary:

Corollary 1. For each iteration of the EM algorithm given by equation 2.3, the search direction $\Theta^{(k+1)} - \Theta^{(k)}$ has a positive projection on the gradient of l .

That is, the EM algorithm can be viewed as a variable metric gradient ascent algorithm for which the projection matrix $P(\Theta^{(k)})$ changes at each iteration as a function of the current parameter value $\Theta^{(k)}$.

Our results extend earlier results due to Baum and Sell (1968), who studied recursive equations of the following form:

$$\begin{aligned} x^{(k+1)} &= T(x^{(k)}), \\ T(x^{(k)}) &= [T(x^{(k)})_1, \dots, T(x^{(k)})_K], \\ T(x^{(k)})_i &= \frac{x_i^{(k)} \partial J / \partial x_i^{(k)}}{\sum_{i=1}^K x_i^{(k)} \partial J / \partial x_i^{(k)}} \end{aligned}$$

where $x_i^{(k)} \geq 0$, $\sum_{i=1}^K x_i^{(k)} = 1$, where J is a polynomial in $x_i^{(k)}$ having positive coefficients. They showed that the search direction of this recursive formula, i.e., $T(x^{(k)}) - x^{(k)}$, has a positive projection on the gradient of J with respect to the $x^{(k)}$ (see also Levinson *et al.* 1983). It can be shown that Baum and Sell’s recursive formula implies the EM update formula for \mathcal{A} in a gaussian mixture. Thus, the first statement in Theorem 1 is a special case of Baum and Sell’s earlier work. However, Baum and Sell’s theorem is an existence theorem and does not provide an explicit expression for the matrix $P_{\mathcal{A}}$ that transforms the gradient direction into the EM direction. Our theorem provides such an explicit form for $P_{\mathcal{A}}$. Moreover, we generalize Baum and Sell’s results to handle the updates for m_j and Σ_j , and we provide explicit expressions for the positive definite transformation matrices P_{m_j} and P_{Σ_j} as well.

It is also worthwhile to compare the EM algorithm to other gradient-based optimization methods. *Newton's method* is obtained by premultiplying the gradient by the inverse of the Hessian of the log likelihood:

$$\Theta^{(k+1)} = \Theta^{(k)} + H(\Theta^{(k)})^{-1} \frac{\partial l}{\partial \Theta^{(k)}}. \quad (3.8)$$

Newton's method is the method of choice when it can be applied, but the algorithm is often difficult to use in practice. In particular, the algorithm can diverge when the Hessian becomes nearly singular; moreover, the computational costs of computing the inverse Hessian at each step can be considerable. An alternative is to approximate the inverse by a recursively updated matrix $B^{(k+1)} = B^{(k)} + \eta \Delta B^{(k)}$. Such a modification is called a *quasi-Newton method*. Conventional quasi-Newton methods are unconstrained optimization methods, however, and must be modified to be used in the mixture setting (where there are probabilistic constraints on the parameters). In addition, quasi-Newton methods generally require that a one-dimensional search be performed at each iteration to guarantee convergence. The EM algorithm can be viewed as a special form of quasi-Newton method in which the projection matrix $P(\Theta^{(k)})$ in equation 3.7 plays the role of $B^{(k)}$. As we discuss in the remainder of the paper, this particular matrix has a number of favorable properties that make EM particularly attractive for optimization in the mixture setting.

4 Constrained Optimization and General Convergence

An important property of the matrix P is that the EM step in parameter space automatically satisfies the probabilistic constraints of the mixture model in equation 2.1. The domain of Θ contains two regions that embody the probabilistic constraints: $\mathcal{D}_1 = \{\Theta : \sum_{j=1}^K \alpha_j^{(k)} = 1\}$ and $\mathcal{D}_2 = \{\Theta : \alpha_j^{(k)} \geq 0, \Sigma_j \text{ is positive definite}\}$. For the EM algorithm the update for the mixing proportions α_j can be rewritten as follows:

$$\alpha_j^{(k+1)} = \frac{1}{N} \sum_{i=1}^N \frac{\alpha_j^{(k)} P(x^{(i)} | m_j^{(k)}, \Sigma_j^{(k)})}{\sum_{i=1}^K \alpha_i^{(k)} P(x^{(i)} | m_i^{(k)}, \Sigma_i^{(k)})}$$

It is obvious that the iteration stays within \mathcal{D}_1 . Similarly, the update for Σ_j can be rewritten as:

$$\Sigma_j^{(k+1)} = \frac{1}{\sum_{i=1}^N h_j^{(k)}(t)} \sum_{i=1}^N \frac{\alpha_j^{(k)} P(x^{(i)} | m_j^{(k)}, \Sigma_j^{(k)})}{\sum_{i=1}^K \alpha_i^{(k)} P(x^{(i)} | m_i^{(k)}, \Sigma_i^{(k)})} [x^{(i)} - m_j^{(k)}][x^{(i)} - m_j^{(k)}]^T$$

which stays within \mathcal{D}_2 for N sufficiently large.

Whereas EM automatically satisfies the probabilistic constraints of a mixture model, other optimization techniques generally require modification to satisfy the constraints. One approach is to modify each iterative

step to keep the parameters within the constrained domain. A number of such techniques have been developed, including feasible direction methods, active sets, gradient projection, reduced-gradient, and linearly constrained quasi-Newton. These constrained methods all incur extra computational costs to check and maintain the constraints and, moreover, the theoretical convergence rates for such constrained algorithms need not be the same as that for the corresponding unconstrained algorithms. A second approach is to transform the constrained optimization problem into an unconstrained problem before using the unconstrained method. This can be accomplished via penalty and barrier functions, Lagrangian terms, or reparameterization. Once again, the extra algorithmic machinery renders simple comparisons based on unconstrained convergence rates problematic. Moreover, it is not easy to meet the constraints on the covariance matrices in the mixture using such techniques.

A second appealing property of $P(\Theta^{(k)})$ is that each iteration of EM is guaranteed to increase the likelihood (i.e., $l(\Theta^{(k+1)}) \geq l(\Theta^{(k)})$). This monotonic convergence of the likelihood is achieved without step-size parameters or line searches. Other gradient-based optimization techniques, including gradient descent, quasi-Newton, and Newton's method, do not provide such a simple theoretical guarantee, even assuming that the constrained problem has been transformed into an unconstrained one. For gradient ascent, the step size η must be chosen to ensure that $\|\Theta^{(k+1)} - \Theta^{(k-1)}\| / \|(\Theta^{(k)} - \Theta^{(k-1)})\| \leq \|I + \eta H(\Theta^{(k-1)})\| < 1$. This requires a one-dimensional line search or an optimization of η at each iteration, which requires extra computation, which can slow down the convergence. An alternative is to fix η to a very small value, which generally makes $\|I + \eta H(\Theta^{(k-1)})\|$ close to one and results in slow convergence. For Newton's method, the iterative process is usually required to be near a solution, otherwise the Hessian may be indefinite and the iteration may not converge. Levenberg–Marquardt methods handle the indefinite Hessian matrix problem; however, a one-dimensional optimization or other form of search is required for a suitable scalar to be added to the diagonal elements of Hessian. Fisher scoring methods can also handle the indefinite Hessian matrix problem, but for nonquadratic nonlinear optimization Fisher scoring requires a stepsize η that obeys $\|I + \eta BH(\Theta^{(k-1)})\| < 1$, where B is the Fisher information matrix. Thus, problems similar to those of gradient ascent arise here as well. Finally, for the quasi-Newton methods or conjugate gradient methods, a one-dimensional line search is required at each iteration. In summary, all of these gradient-based methods incur extra computational costs at each iteration, rendering simple comparisons based on local convergence rates unreliable.

For large-scale problems, algorithms that change the parameters immediately after each data point (“on-line algorithms”) are often significantly faster in practice than batch algorithms. The popularity of gradient descent algorithms for neural networks is in part to the ease of obtaining on-line variants of gradient descent. It is worth noting that on-line

variants of the EM algorithm can be derived (Neal and Hinton 1993; Titterton 1984), and this is a further factor that weighs in favor of EM as compared to conjugate gradient and Newton methods.

5 Convergence Rate Comparisons

In this section, we provide a comparative theoretical discussion of the local convergence rates of constrained gradient ascent and EM.

For gradient ascent a local convergence result can be obtained by Taylor expanding the log likelihood around the maximum likelihood estimate Θ^* . For sufficiently large k we have

$$\|\Theta^{(k+1)} - \Theta^*\| \leq \|I + \eta H(\Theta^*)\| \|\Theta^{(k)} - \Theta^*\| \quad (5.1)$$

and

$$\|I + \eta H(\Theta^*)\| \leq \lambda_M[I + \eta H(\Theta^*)] = r \quad (5.2)$$

where H is the Hessian of l , η is the step size, and $r = \max\{|1 - \eta\lambda_M[-H(\Theta^*)]|, |1 - \eta\lambda_m[-H(\Theta^*)]|\}$, where $\lambda_M[A]$ and $\lambda_m[A]$ denote the largest and smallest eigenvalues of A , respectively.

Smaller values of r correspond to faster convergence rates. To guarantee convergence, we require $r < 1$ or $0 < \eta < 2/\lambda_M[-H(\Theta^*)]$. The minimum possible value of r is obtained when $\eta = 1/\lambda_M[H(\Theta^*)]$ with

$$\begin{aligned} r_{\min} &= 1 - \lambda_m[H(\Theta^*)]/\lambda_M[H(\Theta^*)] \\ &\equiv 1 - \kappa^{-1}[H(\Theta^*)] \end{aligned}$$

where $\kappa[H] = \lambda_M[H]/\lambda_m[H]$ is the *condition number* of H . Larger values of the condition number correspond to slower convergence. When $\kappa[H] = 1$ we have $r_{\min} = 0$, which corresponds to a superlinear rate of convergence. Indeed, Newton's method can be viewed as a method for obtaining a more desirable condition number—the inverse Hessian H^{-1} balances the Hessian H such that the resulting condition number is one. Effectively, Newton can be regarded as gradient ascent on a new function with an effective Hessian that is the identity matrix: $H_{\text{eff}} = H^{-1}H = I$. In practice, however, $\kappa[H]$ is usually quite large. The larger $\kappa[H]$ is, the more difficult it is to compute H^{-1} accurately. Hence it is difficult to balance the Hessian as desired. In addition, as we mentioned in the previous section, the Hessian varies from point to point in the parameter space, and at each iteration we need to recompute the inverse Hessian. Quasi-Newton methods approximate $H(\Theta^{(k)})^{-1}$ by a positive matrix $B^{(k)}$ that is easy to compute.

The discussion thus far has treated unconstrained optimization. To compare gradient ascent with the EM algorithm on the constrained mix-

ture estimation problem, we consider a gradient projection method:

$$\Theta^{(k+1)} = \Theta^{(k)} + \eta \Pi_k \frac{\partial l}{\partial \Theta^{(k)}} \quad (5.3)$$

where Π_k is the projection matrix that projects the gradient $\partial l / \partial \Theta^{(k)}$ into \mathcal{D}_1 . This gradient projection iteration will remain in \mathcal{D}_1 as long as the initial parameter vector is in \mathcal{D}_1 . To keep the iteration within \mathcal{D}_2 , we choose an initial $\Theta^{(0)} \in \mathcal{D}_2$ and keep η sufficiently small at each iteration.

Suppose that $E = [e_1, \dots, e_m]$ is a set of independent unit basis vectors that spans the space \mathcal{D}_1 . In this basis, $\Theta^{(k)}$ and $\Pi_k(\partial l / \partial \Theta^{(k)})$ become $\Theta_c^{(k)} = E^T \Theta^{(k)}$ and $\partial l / \partial \Theta_c^{(k)} = E^T (\partial l / \partial \Theta^{(k)})$, respectively, with $\|\Theta_c^{(k)} - \Theta_c^*\| = \|\Theta^{(k)} - \Theta^*\|$. In this representation the projective gradient algorithm equation 5.3 becomes simple gradient ascent: $\Theta_c^{(k+1)} = \Theta_c^{(k)} + \eta (\partial l / \partial \Theta_c^{(k)})$. Moreover, equation 5.1 becomes $\|\Theta^{(k+1)} - \Theta^*\| \leq \|E^T [I + \eta H(\Theta^*)]\| \|\Theta^{(k)} - \Theta^*\|$. As a result, the convergence rate is bounded by

$$\begin{aligned} r_c &= \|E^T [I + \eta H(\Theta^*)]\| \\ &\leq \sqrt{\lambda_M[E^T [I + \eta H(\Theta^*)][I + \eta H(\Theta^*)]^T E]} \\ &= \sqrt{\lambda_M[E^T [I + 2\eta H(\Theta^*) + \eta^2 H^2(\Theta^*)]E]} \end{aligned}$$

Since $H(\Theta^*)$ is negative definite, we obtain

$$r_c \leq \sqrt{1 + \eta^2 \lambda_M^2[-H_c] - 2\eta \lambda_m[-H_c]} \quad (5.4)$$

In this equation $H_c = E^T H(\Theta^*) E$ is the Hessian of l restricted to \mathcal{D}_1 .

We see from this derivation that the convergence speed depends on $\kappa[H_c] = \lambda_M[-H_c] / \lambda_m[-H_c]$. When $\kappa[H_c] = 1$, we have

$$\sqrt{1 + \eta^2 \lambda_M^2(-H_c) - 2\eta \lambda_m[-H_c]} = 1 - \eta \lambda_m[-H_c]$$

which in principle can be made to equal zero if η is selected appropriately. In this case, a superlinear rate is obtained. Generally, however, $\kappa[H_c] \neq 1$, with smaller values of $\kappa[H_c]$ corresponding to faster convergence.

We now turn to an analysis of the EM algorithm. As we have seen EM keeps the parameter vector within \mathcal{D}_1 automatically. Thus, in the new basis the connection between EM and gradient ascent (cf. equation 3.7) becomes

$$\Theta_c^{(k+1)} = \Theta_c^{(k)} + E^T P(\Theta^{(k)}) \frac{\partial l}{\partial \Theta}$$

and we have

$$\|\Theta^{(k+1)} - \Theta^*\| \leq \|E^T [I + PH(\Theta^*)]\| \|\Theta^{(k)} - \Theta^*\|$$

with

$$r_c = \|E^T [I + PH(\Theta^*)]\| \leq \sqrt{\lambda_M[E^T [I + PH(\Theta^*)][I + PH(\Theta^*)]^T E]}$$

The latter equation can be further manipulated to yield

$$r_c \leq \sqrt{1 + \lambda_M^2[E^T PHE] - 2\lambda_m[-E^T PHE]} \quad (5.5)$$

Thus we see that the convergence speed of EM depends on

$$\kappa[E^T PHE] = \lambda_M[E^T PHE]/\lambda_m[E^T PHE]$$

When

$$\kappa[E^T PHE] = 1, \quad \lambda_M[E^T PHE] = 1$$

we have

$$\sqrt{1 + \lambda_M^2[E^T PHE] - 2\lambda_m[-E^T PHE]} = (1 - \lambda_M[-E^T PHE]) = 0$$

In this case, a superlinear rate is obtained. We discuss the possibility of obtaining superlinear convergence with EM in more detail below.

These results show that the convergence of gradient ascent and EM both depend on the shape of the log likelihood as measured by the condition number. When $\kappa[H]$ is near one, the configuration is quite regular, and the update direction points directly to the solution yielding fast convergence. When $\kappa[H]$ is very large, the l surface has an elongated shape, and the search along the update direction is a zigzag path, making convergence very slow. The key idea of Newton and quasi-Newton methods is to reshape the surface. The nearer it is to a ball shape (Newton's method achieves this shape in the ideal case), the better the convergence. Quasi-Newton methods aim to achieve an effective Hessian whose condition number is as close as possible to one. Interestingly, the results that we now present suggest that the projection matrix P for the EM algorithm also serves to effectively reshape the likelihood yielding an effective condition number that tends to one. We first present empirical results that support this suggestion and then present a theoretical analysis.

We sampled 1000 points from a simple finite mixture model given by

$$p(x) = \alpha_1 p_1(x) + \alpha_2 p_2(x)$$

where

$$p_i(x) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{1}{2} \frac{(x - m_i)^2}{\sigma_i^2}\right\}$$

The parameter values were as follows: $\alpha_1 = 0.7170$, $\alpha_2 = 0.2830$, $m_1 = -2$, $m_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$. We ran both the EM algorithm and gradient ascent on the data. The initialization for each experiment is set randomly, but is the same for both the EM algorithm and the gradient algorithm. At each step of the simulation, we calculated the condition number of the Hessian ($\kappa[H(\Theta^{(k)})]$), the condition number determining the rate of convergence of the gradient algorithm ($\kappa[E^T H(\Theta^{(k)})E]$), and the condition number determining the rate of convergence of EM ($\kappa[E^T P(\Theta^{(k)})H(\Theta^{(k)})E]$). We also

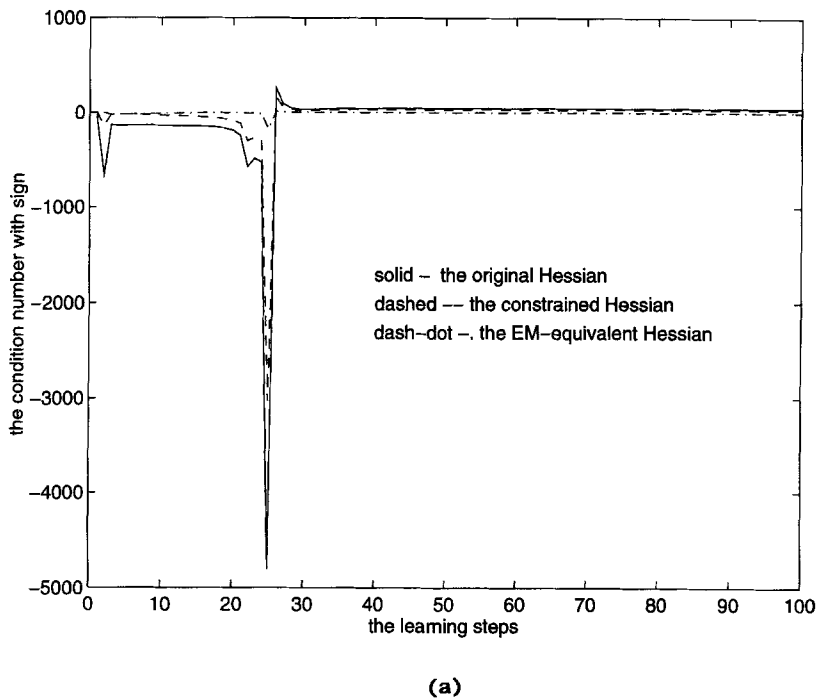


Figure 1a: Experimental results for the estimation of the parameters of a two-component gaussian mixture. (a) The condition numbers as a function of the iteration number.

calculated the largest eigenvalues of the matrices $H(\Theta^{(k)})$, $E^T H(\Theta^{(k)}) E$, and $E^T P(\Theta^{(k)}) H(\Theta^{(k)}) E$. The results are shown in Figure 1. As can be seen in Figure 1a, the condition numbers change rapidly in the vicinity of the 25th iteration. This is because the corresponding Hessian matrix is indefinite before the iteration enters the neighborhood of a solution. Afterward, the Hessians quickly become definite and the condition numbers converge.³ As shown in Figure 1b, the condition numbers converge toward the values $\kappa[H(\Theta^{(k)})] = 47.5$, $\kappa[E^T H(\Theta^{(k)}) E] = 33.5$, and $\kappa[E^T P(\Theta^{(k)}) H(\Theta^{(k)}) E] = 3.6$. That is, the matrix P has greatly reduced the condition number, by factors of 93 and 13.2. This significantly improves the shape of l and speeds up the convergence.

³Interestingly, the EM algorithm converges soon afterward as well, showing that for this problem EM spends little time in the region of parameter space in which a local analysis is valid.

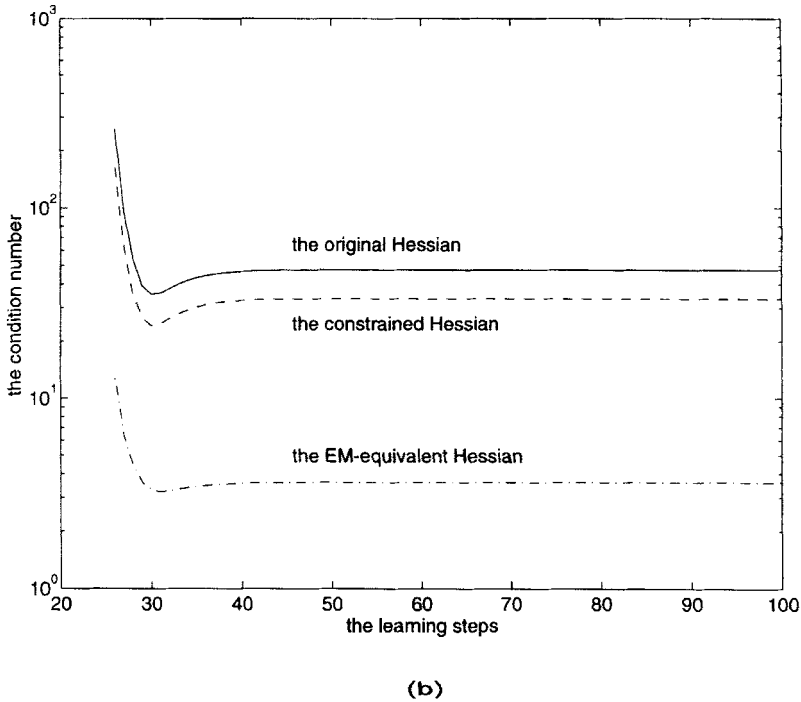


Figure 1b: (b) A zoomed version of (a) after discarding the first 25 iterations. The terminology “original, constrained, and EM-equivalent Hessians” refers to the matrices H , $E^T H E$, and $E^T P H E$, respectively.

We ran a second experiment in which the means of the component gaussians were $m_1 = -1$ and $m_2 = 1$. The results are similar to those shown in Figure 1. Since the distance between two distributions is reduced into half, the shape of l becomes more irregular (Fig. 2). The condition number $\kappa[H(\Theta^{(k)})]$ increases to 352, $\kappa[E^T H(\Theta^{(k)}) E]$ increases to 216, and $\kappa[E^T P(\Theta^{(k)}) H(\Theta^{(k)}) E]$ increases to 61. We see once again a significant improvement in the case of EM, by factors of 3.5 and 5.8.

Figure 3 shows that the matrix P has also reduced the largest eigenvalues of the Hessian from between 2000 to 3000 to around 1. This demonstrates clearly the stable convergence that is obtained via EM, without a line search or the need for external selection of a learning stepsize.

In the remainder of the paper we provide some theoretical analyses that attempt to shed some light on these empirical results. To illustrate the issues involved, consider a degenerate mixture problem in which

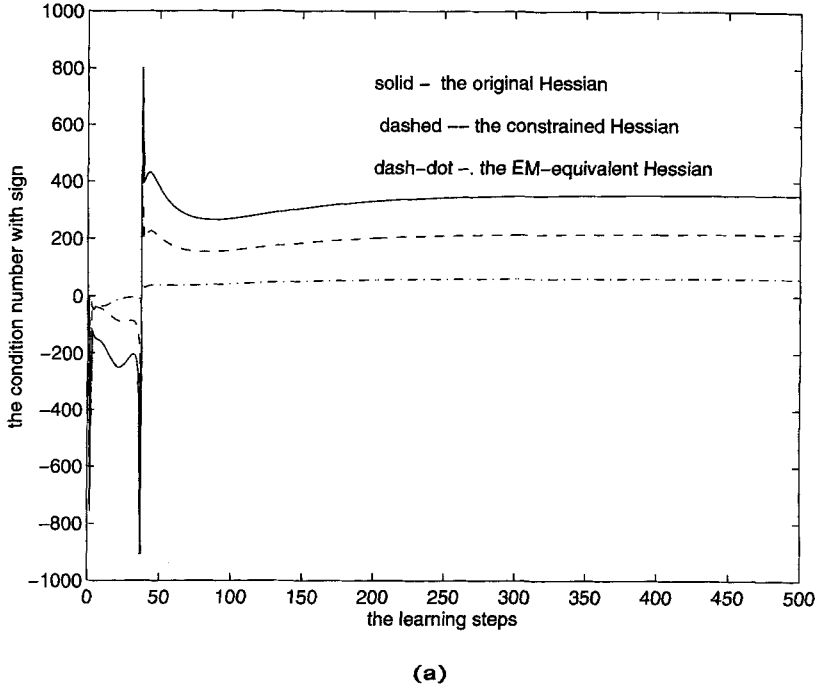


Figure 2: Experimental results for the estimation of the parameters of a two-component gaussian mixture (cf. Fig. 1). The separation of the gaussians is half the separation in Figure 1.

the mixture has a single component. (In this case $\alpha_1 = 1$.) Let us furthermore assume that the covariance matrix is fixed (i.e., only the mean vector m is to be estimated). The Hessian with respect to the mean m is $H = -N\Sigma^{-1}$ and the EM projection matrix P is Σ/N . For gradient ascent, we have $\kappa[E^T H E] = \kappa[\Sigma^{-1}]$, which is larger than one whenever $\Sigma \neq cI$. EM, on the other hand, achieves a condition number of one exactly ($\kappa[E^T P H E] = \kappa[P H] = \kappa[I] = 1$ and $\lambda_M[E^T P H E] = 1$). Thus, EM and Newton's method are the same for this simple quadratic problem. For general nonquadratic optimization problems, Newton retains the quadratic assumption, yielding fast convergence but possible divergence. EM is a more conservative algorithm that retains the convergence guarantee but also maintains quasi-Newton behavior. We now analyze this behavior in more detail. We consider the special case of estimating the means in a gaussian mixture when the gaussians are well separated.

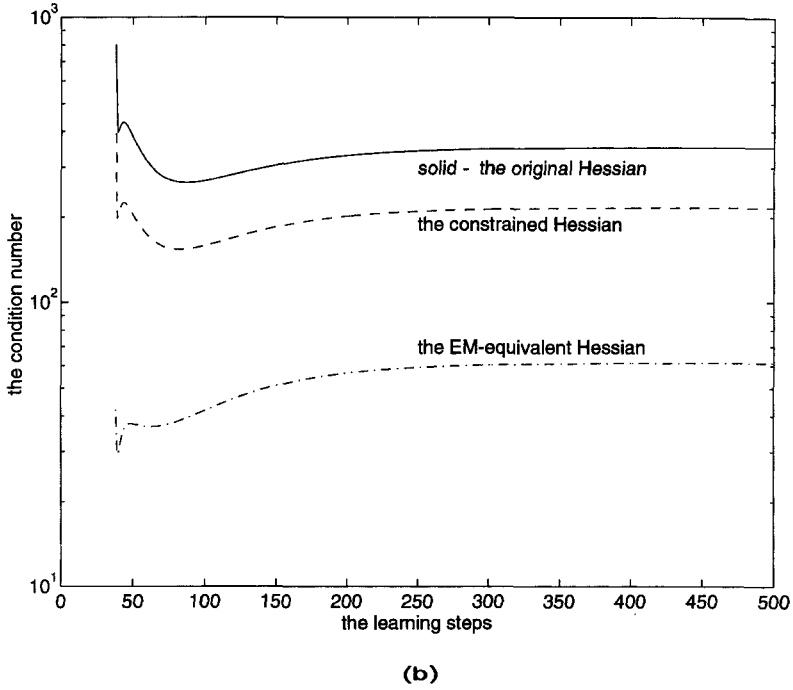


Figure 2: Continued.

Theorem 2. Consider the EM algorithm in equation 2.3, where the parameters α_j and Σ_j are assumed to be known. Assume that the K gaussian distributions are well separated, such that for sufficiently large k the posterior probabilities $h_j^{(k)}(t)$ are approximately zero or one. For such k , the condition number associated with EM is approximately one, which is smaller than the condition number associated with gradient ascent. That is

$$\kappa[E^T P(\Theta^{(k)}) H(\Theta^{(k)}) E] \approx 1 \quad (5.6)$$

$$\kappa[E^T P(\Theta^{(k)}) H(\Theta^{(k)}) E] < \kappa[E^T H(\Theta^{(k)}) E] \quad (5.7)$$

Furthermore, we have also

$$\lambda_M[E^T P(\Theta^{(k)}) H(\Theta^{(k)}) E] \approx 1 \quad (5.8)$$

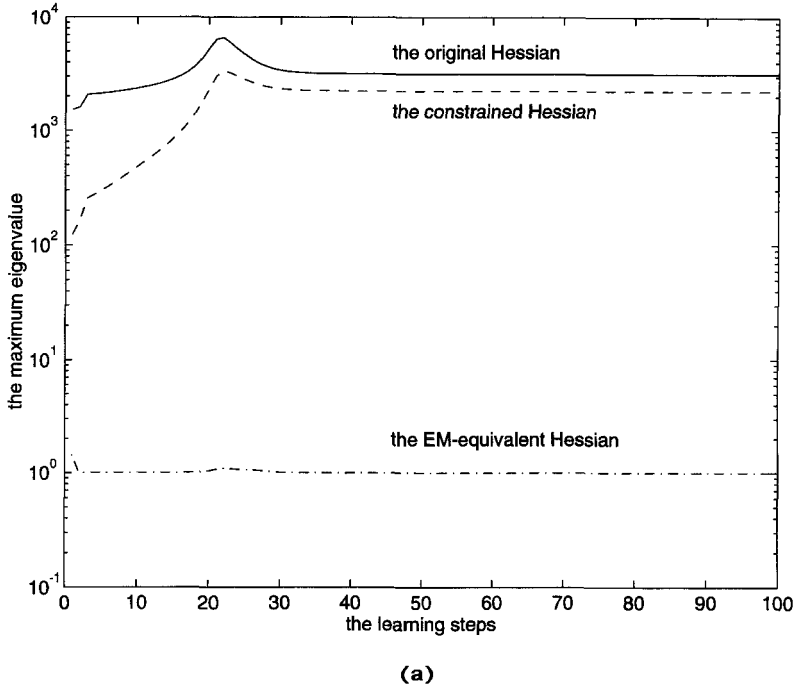


Figure 3: The largest eigenvalues of the matrices H , $E^T H E$, and $E^T P H E$ plotted as a function of the number of iterations. The plot in (a) is for the experiment in Figure 1; (b) is for the experiment reported in Figure 2.

Proof. The Hessian is

$$H = \begin{bmatrix} H_{11} & H_{12} & \cdots & H_{1K} \\ H_{21} & H_{22} & \cdots & H_{2K} \\ \vdots & \vdots & & \vdots \\ H_{K1} & H_{K2} & \cdots & H_{KK} \end{bmatrix} \quad (5.9)$$

where

$$\begin{aligned} H_{ij} &\equiv \frac{\partial^2 l}{\partial m_i \partial m_j^T} \\ &= -(\Sigma_j^{(k)})^{-1} \sum_{t=1}^N \delta_{ij} h_j^{(k)}(t) \\ &\quad + (\Sigma_j^{(k)})^{-1} \left[\sum_{t=1}^N \gamma_{ij}(x^{(t)}) (x^{(t)} - m_j)(x^{(t)} - m_i)^T \right] (\Sigma_i^{(k)})^{-1} \end{aligned} \quad (5.10)$$

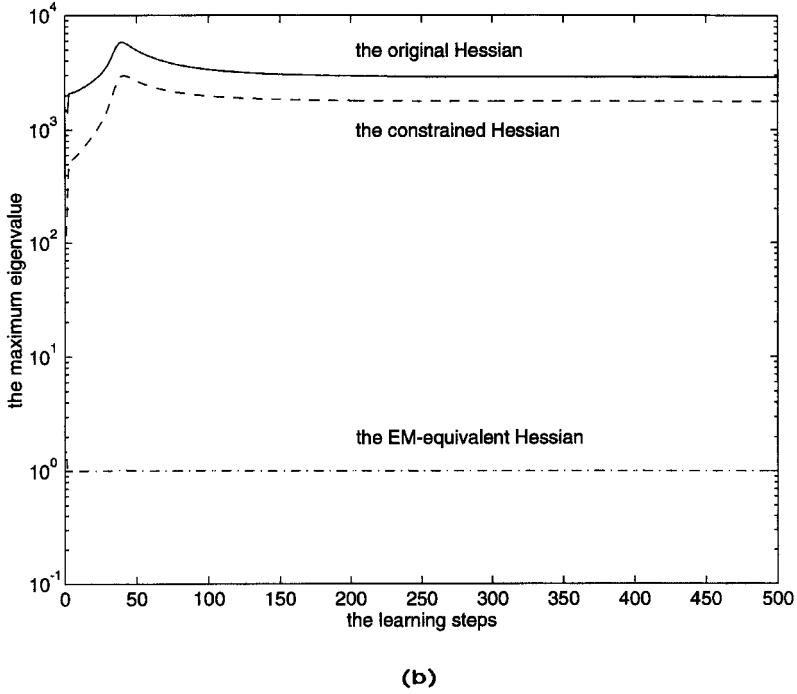


Figure 3: Continued. The plot in (a) is for the experiment in Figure 1; (b) is for the experiment reported in Figure 2.

with $\gamma_{ij}(x^{(t)}) = [\delta_{ij} - h_i^{(k)}(t)]h_j^{(k)}(t)$. The projection matrix P is

$$P^{(k)} = \text{diag}[P_{11}^{(k)}, \dots, P_{KK}^{(k)}]$$

where

$$P_{jj}^{(k)} = \Sigma_j^{(k)} / \sum_{t=1}^N h_j^{(k)}(t)$$

Given that $h_j^{(k)}(t)[1 - h_j^{(k)}(t)]$ is negligible for sufficiently large k [since $h_j^{(k)}(t)$ are approximately zero or one], the second term in equation 5.10 can be neglected, yielding $H_{ii} \approx -(\Sigma_j^{(k)})^{-1} \sum_{t=1}^N h_j^{(k)}(t)$ and $H = \text{diag}[H_{11}, \dots, H_{KK}]$. This implies that $PH \approx -I$ and $E^T PHE \approx -I$, thus $\kappa[E^T PHE] \approx 1$ and $\lambda_M[E^T PHE] \approx 1$, whereas usually $\kappa[E^T HE] > 1$. \square

This theorem, although restrictive in its assumptions, gives some indication as to why the projection matrix in the EM algorithm appears to

condition the Hessian, yielding improved convergence. In fact, we conjecture that equations 5.7 and 5.8 can be extended to apply more widely, in particular to the case of the full EM update in which the mixing proportions and covariances are estimated, and also, within limits, to cases in which the means are not well separated. To obtain an initial indication as to possible conditions that can be usefully imposed on the separation of the mixture components, we have studied the case in which the second term in equation 5.10 is neglected only for H_{ii} and is retained for the H_{ij} components, where $j \neq i$. Consider, for example, the case of a univariate mixture having two mixture components. For fixed mixing proportions and fixed covariances, the Hessian matrix (equation 5.9) becomes

$$H = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$$

and the projection matrix (equation 5.10) becomes

$$P = \begin{bmatrix} -h_{11}^{-1} & 0 \\ 0 & -h_{22}^{-1} \end{bmatrix}$$

where

$$h_{ii} = -\frac{1}{\sigma_i^{2(k)}} \sum_{t=1}^N h_i^{(k)}(t), \quad i = 1, 2$$

and

$$h_{ij} = \frac{1}{\sigma_i^{2(k)} \sigma_j^{2(k)}} \sum_{t=1}^N [1 - h_i^{(k)}(t)] h_j^{(k)}(t) (x^{(t)} - m_j)^T (x^{(t)} - m_i), \quad i \neq j = 1, 2$$

If H is negative definite (i.e., $h_{11}h_{22} - h_{12}h_{21} < 0$), then we can show that the conclusions of equation 5.7 remain true, even for gaussians that are not necessarily well separated. The proof is achieved via the following lemma:

Lemma 1. Consider the positive definite matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

For the diagonal matrix $B = \text{diag}[\sigma_{11}^{-1}, \sigma_{22}^{-1}]$, we have $\kappa[B\Sigma] < \kappa[\Sigma]$.

Proof. The eigenvalues of Σ are the roots of $(\sigma_{11} - \lambda)(\sigma_{22} - \lambda) - \sigma_{21}\sigma_{12} = 0$, which gives

$$\begin{aligned} \lambda_M &= \frac{\sigma_{11} + \sigma_{22} + \gamma}{2} \\ \lambda_m &= \frac{\sigma_{11} + \sigma_{22} - \gamma}{2} \\ \gamma &= \sqrt{(\sigma_{11} + \sigma_{22})^2 - 4(\sigma_{11}\sigma_{22} - \sigma_{21}\sigma_{12})} \end{aligned}$$

and

$$\kappa[\Sigma] = \frac{\sigma_{11} + \sigma_{22} + \gamma}{\sigma_{11} + \sigma_{22} - \gamma}$$

The condition number $\kappa[\Sigma]$ can be written as $\kappa[\Sigma] = (1+s)/(1-s) \equiv f(s)$, where s is defined as follows:

$$s = \sqrt{1 - \frac{4(\sigma_{11}\sigma_{22} - \sigma_{21}\sigma_{12})}{(\sigma_{11} + \sigma_{22})^2}}$$

Furthermore, the eigenvalues of $B\Sigma$ are the roots of $(1-\lambda)(1-\lambda) - (\sigma_{21}\sigma_{12})/(\sigma_{11}\sigma_{22}) = 0$, which gives $\lambda_M = 1 + \sqrt{(\sigma_{21}\sigma_{12})/(\sigma_{11}\sigma_{22})}$ and $\lambda_m = 1 - \sqrt{(\sigma_{21}\sigma_{12})/(\sigma_{11}\sigma_{22})}$. Thus, defining $r = \sqrt{(\sigma_{21}\sigma_{12})/(\sigma_{11}\sigma_{22})}$, we have $\kappa[B\Sigma] = (1+r)/(1-r) = f(r)$.

We now examine the quotient s/r :

$$\frac{s}{r} = \frac{1}{r} \sqrt{1 - \frac{4(1-r^2)}{(\sigma_{11} + \sigma_{22})^2/(\sigma_{11}\sigma_{22})}}$$

Given that $(\sigma_{11} + \sigma_{22})^2/(\sigma_{11}\sigma_{22}) \geq 4$, we have $s/r > 1/r\sqrt{1 - (1-r^2)} = 1$. That is, $s > r$. Since $f(x) = (1+x)/(1-x)$ is a monotonically increasing function for $x > 1$, we have $f(s) > f(r)$. Therefore, $\kappa[B\Sigma] < \kappa[\Sigma]$. \square

We think that it should be possible to generalize this lemma beyond the univariate, two-component case, thereby weakening the conditions on separability in Theorem 2 in a more general setting.

6 Conclusions

In this paper we have provided a comparative analysis of algorithms for the learning of gaussian mixtures. We have focused on the EM algorithm and have forged a link between EM and gradient methods via the projection matrix P . We have also analyzed the convergence of EM in terms of properties of the matrix P and the effect that P has on the likelihood surface.

EM has a number of properties that make it a particularly attractive algorithm for mixture models. It enjoys automatic satisfaction of probabilistic constraints, monotonic convergence without the need to set a learning rate, and low computational overhead. Although EM has the reputation of being a slow algorithm, we feel that in the mixture setting the slowness of EM has been overstated. Although EM can indeed converge slowly for problems in which the mixture components are not well separated, the Hessian is poorly conditioned for such problems and thus other gradient-based algorithms (including Newton's method) are also likely to perform poorly. Moreover, if one's concern is convergence in likelihood, then EM generally performs well even for these ill-conditioned problems. Indeed the algorithm provides a certain amount

of safety in such cases, despite the poor conditioning. It is also important to emphasize that the case of poorly separated mixture components can be viewed as a problem in model selection (too many mixture components are being included in the model), and should be handled by regularization techniques.

The fact that EM is a first-order algorithm certainly implies that EM is no panacea, but does not imply that EM has no advantages over gradient ascent or superlinear methods. First, it is important to appreciate that convergence rate results are generally obtained for unconstrained optimization, and are not necessarily indicative of performance on constrained optimization problems. Also, as we have demonstrated, there are conditions under which the condition number of the effective Hessian of the EM algorithm tends toward one, showing that EM can approximate a superlinear method. Finally, in cases of a poorly conditioned Hessian, superlinear convergence is not necessarily a virtue. In such cases many optimization schemes, including EM, essentially revert to gradient ascent.

We feel that EM will continue to play an important role in the development of learning systems that emphasize the predictive aspect of data modeling. EM has indeed played a critical role in the development of hidden Markov models (HMMs), an important example of predictive data modeling.⁴ EM generally converges rapidly in this setting. Similarly, in the case of hierarchical mixtures of experts the empirical results on convergence in likelihood have been quite promising (Jordan and Jacobs 1994; Waterhouse and Robinson 1994). Finally, EM can play an important conceptual role as an organizing principle in the design of learning algorithms. Its role in this case is to focus attention on the “missing variables” in the problem. This clarifies the structure of the algorithm and invites comparisons with statistical physics, where missing variables often provide a powerful analytic tool (Yuille *et al.* 1994).

Appendix: Proof of Theorem 1

1. We begin by considering the EM update for the mixing proportions α_i . From equations 2.1 and 2.2, we have

$$\left. \frac{\partial l}{\partial \alpha} \right|_{\mathcal{A}=\mathcal{A}^{(k)}} = \sum_{l=1}^N \frac{[P(x^{(l)}, \theta_1^{(k)}), \dots, P(x^{(l)}, \theta_K^{(k)})]^T}{\sum_{i=1}^K \alpha_i^{(k)} P(x^{(l)}, \theta_i^{(k)})}$$

⁴In most applications of HMMs, the “parameter estimation” process is employed solely to yield models with high likelihood; the parameters are not generally endowed with a particular meaning.

Premultiplying by $P_{\mathcal{A}}^{(k)}$, we obtain

$$\begin{aligned}
 & P_{\mathcal{A}}^{(k)} \left. \frac{\partial l}{\partial \mathcal{A}} \right|_{\mathcal{A}=\mathcal{A}^{(k)}} \\
 &= \frac{1}{N} \sum_{t=1}^N \frac{\{[\alpha_1^{(k)} P(x^{(t)}, \theta_1^{(k)}), \dots, \alpha_K^{(k)} P(x^{(t)}, \theta_K^{(k)})]^T - \mathcal{A}^{(k)} \sum_{i=1}^K \alpha_i^{(k)} P(x^{(t)}, \theta_i^{(k)})\}}{\sum_{i=1}^K \alpha_i^{(k)} P(x^{(t)}, \theta_i^{(k)})} \\
 &= \frac{1}{N} \sum_{t=1}^N [h_1^{(k)}(t), \dots, h_K^{(k)}(t)]^T - \mathcal{A}^{(k)}
 \end{aligned}$$

The update formula for \mathcal{A} in equation 2.3 can be rewritten as

$$\mathcal{A}^{(k+1)} = \mathcal{A}^{(k)} + \frac{1}{N} \sum_{t=1}^N [h_1^{(k)}(t), \dots, h_K^{(k)}(t)]^T - \mathcal{A}^{(k)}$$

Combining the last two equations establishes the update rule for \mathcal{A} (equation 2.4). Furthermore, for an arbitrary vector u , we have $Nu^T P_{\mathcal{A}}^{(k)} u = u^T \text{diag}[\alpha_1^{(k)}, \dots, \alpha_K^{(k)}] u - (u^T \mathcal{A}^{(k)})^2$. By Jensen's inequality we have

$$\begin{aligned}
 u^T \text{diag}[\alpha_1^{(k)}, \dots, \alpha_K^{(k)}] u &= \sum_{j=1}^K \alpha_j^{(k)} u_j^2 \\
 &> \left(\sum_{j=1}^K \alpha_j^{(k)} u_j \right)^2 \\
 &= (u^T \mathcal{A}^{(k)})^2
 \end{aligned}$$

Thus, $u^T P_{\mathcal{A}}^{(k)} u > 0$ and $P_{\mathcal{A}}^{(k)}$ is positive definite given the constraints $\sum_{j=1}^K \alpha_j^{(k)} = 1$ and $\alpha_j^{(k)} \geq 0$ for all j .

2. We now consider the EM update for the means m_i . It follows from equations 2.1 and 2.2 that

$$\left. \frac{\partial l}{\partial m_j} \right|_{m_j=m_j^{(k)}} = \sum_{t=1}^N h_j^{(k)}(t) (\Sigma_j^{(k)})^{-1} [x^{(t)} - m_j^{(k)}]$$

Premultiplying by $P_{m_j}^{(k)}$ yields

$$\begin{aligned}
 P_{m_j}^{(k)} \left. \frac{\partial l}{\partial m_j} \right|_{m_j=m_j^{(k)}} &= \frac{1}{\sum_{t=1}^N h_j^{(k)}(t)} \sum_{t=1}^N h_j^{(k)}(t) x^{(t)} - m_j^{(k)} \\
 &= m_j^{(k+1)} - m_j^{(k)}
 \end{aligned}$$

From equation 2.3, we have $\sum_{t=1}^N h_j^{(k)}(t) > 0$; moreover, $\Sigma_j^{(k)}$ is positive definite with probability one assuming that N is large enough such that

the matrix is of full rank. Thus, it follows from equation 3.5 that $P_{m_j}^{(k)}$ is positive definite with probability one.

3. Finally, we prove the third part of the theorem. It follows from equations 2.1 and 2.2 that

$$\frac{\partial l}{\partial \Sigma_j} \Big|_{\Sigma_j = \Sigma_j^{(k)}} = -\frac{1}{2} \sum_{t=1}^N h_j^{(k)}(t) (\Sigma_j^{(k)})^{-1} \{ \Sigma_j^{(k)} - [x^{(t)} - m_j^{(k)}][x^{(t)} - m_j^{(k)}]^T \} (\Sigma_j^{(k)})^{-1}$$

With this in mind, we rewrite the EM update formula for $\Sigma_j^{(k)}$ as

$$\begin{aligned} \Sigma_j^{(k+1)} &= \Sigma_j^{(k)} + \frac{1}{\sum_{t=1}^N h_j^{(k)}(t)} \sum_{t=1}^N h_j^{(k)}(t) [x^{(t)} - m_j^{(k)}][x^{(t)} - m_j^{(k)}]^T - \Sigma_j^{(k)} \\ &= \Sigma_j^{(k)} + \frac{2\Sigma_j^{(k)}}{\sum_{t=1}^N h_j^{(k)}(t)} V_{\Sigma_j} \Sigma_j^{(k)} \end{aligned}$$

where

$$\begin{aligned} V_{\Sigma_j} &= -\frac{1}{2} \sum_{t=1}^N h_j^{(k)}(t) (\Sigma_j^{(k)})^{-1} \{ \Sigma_j^{(k)} - [x^{(t)} - m_j^{(k)}][x^{(t)} - m_j^{(k)}]^T \} (\Sigma_j^{(k)})^{-1} \\ &= \frac{\partial l}{\partial \Sigma_j} \Big|_{\Sigma_j = \Sigma_j^{(k)}} \end{aligned}$$

That is, we have

$$\Sigma_j^{(k+1)} = \Sigma_j^{(k)} + \frac{2\Sigma_j^{(k)}}{\sum_{t=1}^N h_j^{(k)}(t)} \frac{\partial l}{\partial \Sigma_j} \Big|_{\Sigma_j = \Sigma_j^{(k)}} \Sigma_j^{(k)}$$

Utilizing the identity $\text{vec}[ABC] = (C^T \otimes A)\text{vec}[B]$, we obtain

$$\text{vec}[\Sigma_j^{(k+1)}] = \text{vec}[\Sigma_j^{(k)}] + \frac{2}{\sum_{t=1}^N h_j^{(k)}(t)} (\Sigma_j^{(k)} \otimes \Sigma_j^{(k)}) \frac{\partial l}{\partial \Sigma_j} \Big|_{\Sigma_j = \Sigma_j^{(k)}}$$

Thus $P_{\Sigma_j}^{(k)} = \frac{2}{\sum_{t=1}^N h_j^{(k)}(t)} (\Sigma_j^{(k)} \otimes \Sigma_j^{(k)})$. Moreover, for an arbitrary matrix U , we have

$$\begin{aligned} \text{vec}[U]^T (\Sigma_j^{(k)} \otimes \Sigma_j^{(k)}) \text{vec}[U] &= \text{tr}(\Sigma_j^{(k)} U \Sigma_j^{(k)} U^T) \\ &= \text{tr}[(\Sigma_j^{(k)} U)^T (\Sigma_j^{(k)} U)] \\ &= \text{vec}[\Sigma_j^{(k)} U]^T \text{vec}[\Sigma_j^{(k)} U] \geq 0 \end{aligned}$$

where equality holds only when $\Sigma_j^{(k)} U = 0$ for all U . Equality is impossible, however, since $\Sigma_j^{(k)}$ is positive definite with probability one when N is sufficiently large. Thus it follows from equation 3.6 and $\sum_{t=1}^N h_j^{(k)}(t) > 0$ that $P_{\Sigma_j}^{(k)}$ is positive definite with probability one. \square

Acknowledgments

This project was supported in part by a Ho Sin-Hang Education Endowment Foundation Grant and by the HK RGC Earmarked Grant CUHK250/94E, by a grant from the McDonnell-Pew Foundation, by a grant from ATR Human Information Processing Research Laboratories, by a grant from Siemens Corporation, by Grant IRI-9013991 from the National Science Foundation, and by Grant N00014-90-J-1942 from the Office of Naval Research. Michael I. Jordan is an NSF Presidential Young Investigator.

References

- Amari, S. 1995. Information geometry of the EM and em algorithms for neural networks. *Neural Networks* **8**, (5) (in press).
- Baum, L. E., and Sell, G. R. 1968. Growth transformation for functions on manifolds. *Pac. J. Math.* **27**, 211–227.
- Bengio, Y., and Frasconi, P., 1995. An input-output HMM architecture. In *Advances in Neural Information Processing Systems 7*, G. Tesauro, D. S. Touretzky, and J. Alspector, eds. MIT Press, Cambridge MA.
- Boyles, R. A. 1983. On the convergence of the EM algorithm. *J. Royal Stat. Soc.* **B45**(1), 47–50.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc.* **B39**, 1–38.
- Ghahramani, Z. and Jordan, M. I. 1994. Function approximation via density estimation using the EM approach. In *Advances in Neural Information Processing Systems 6*, J. D. Cowan, G. Tesauro, and J. Alspector, eds., pp. 120–127. Morgan Kaufmann, San Mateo, CA.
- Jordan, M. I., and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Comp.* **6**, 181–214.
- Jordan, M. I., and Xu, L. 1995. Convergence results for the EM approach to mixtures-of-experts architectures. *Neural Networks* (in press).
- Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. 1983. An introduction to the application of the theory of probabilistic functions of Markov process to automatic speech recognition. *Bell Sys. Techni. J.* **62**, 1035–1072.
- Neal, R. N., and Hinton, G. E. 1993. *A New View of the EM Algorithm that Justifies Incremental and Other Variants*. University of Toronto, Department of Computer Science preprint.
- Nowlan, S. J. 1991. *Soft Competitive Adaptation: Neural Network Learning Algorithms Based on Fitting Statistical Mixtures*. Tech. Rep. CMU-CS-91-126, CMU, Pittsburgh, PA.
- Redner, R. A., and Walker, H. F. 1984. Mixture densities, maximum likelihood, and the EM algorithm, *SIAM Rev.* **26**, 195–239.
- Titterton, D. M. 1984. Recursive parameter estimation using incomplete data. *J. of Royal Stat. Soc.* **B46**, 257–267.
- Tresp, V., Ahmad, S., and Neuneier, R. 1994. Training neural networks with deficient data. In *Advances in Neural Information Processing Systems 6*, J. D.

- Cowan, G. Tesauro, and J. Alspector, eds. Morgan Kaufmann, San Mateo, CA.
- Waterhouse, S. R., and Robinson, A. J. 1994. Classification using hierarchical mixtures of experts. *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 177–186.
- Wu, C. F. J. 1983. On the convergence properties of the EM algorithm. *Ann. Stat.* **11**, 95–103.
- Xu, L., and Jordan, M. I. 1993a. Unsupervised learning by EM algorithm based on finite mixture of Gaussians. *Proc. WCNN'93*, Portland, OR, **II**, 431–434.
- Xu, L., and Jordan, M. I. 1993b. EM learning on a generalized finite mixture model for combining multiple classifiers. *Proc. WCNN'93*, Portland, OR, **IV**, 227–230.
- Xu, L., and Jordan, M. I. 1993c. *Theoretical and Experimental Studies of the EM Algorithm for Unsupervised Learning Based on Finite Gaussian Mixtures*. MIT Computational Cognitive Science, Tech. Rep. 9302, Dept. of Brain and Cognitive Science, MIT, Cambridge, MA.
- Xu, L., Jordan, M. I., and Hinton, G. E. 1994. A modified gating network for the mixtures of experts architecture. *Proc. WCNN'94*, San Diego, **2**, 405–410.
- Yuille, A. L., Stolorz, P., and Utans, J. 1994. Statistical physics, mixtures of distributions and the EM algorithm. *Neural Comp.* **6**, 334–340.

This article has been cited by:

1. S. Venkatesh, S. Jayalalith, S. Gopal. 2014. Mixture Density Estimation Clustering Based Probabilistic Neural Network Variants for Multiple Source Partial Discharge Signature Analysis. *Journal of Applied Sciences* **14**, 1496-1505. [[CrossRef](#)]
2. Theresa Springer, Karsten Urban. 2014. Comparison of the EM algorithm and alternatives. *Numerical Algorithms* **67**, 335-364. [[CrossRef](#)]
3. Marco Aste, Massimo Boninsegna, Antonino Freno, Edmondo Trentin. 2014. Techniques for dealing with incomplete data: a tutorial and survey. *Pattern Analysis and Applications* . [[CrossRef](#)]
4. Sylvain Calinon, Danilo Bruno, Milad S. Malekzadeh, Thrishantha Nanayakkara, Darwin G. Caldwell. 2014. Human-robot skills transfer interfaces for a flexible surgical robot. *Computer Methods and Programs in Biomedicine* **116**, 81-96. [[CrossRef](#)]
5. Sang Hyoung Lee, Il Hong Suh, Sylvain Calinon, Rolf Johansson. 2014. Autonomous framework for segmenting robot trajectories of manipulation task. *Autonomous Robots* . [[CrossRef](#)]
6. Álvaro Gómez-Losada, Antonio Lozano-García, Rafael Pino-Mejías, Juan Contreras-González. 2014. Finite mixture models to characterize and refine air quality monitoring networks. *Science of The Total Environment* **485-486**, 292-299. [[CrossRef](#)]
7. Zhi-Yong Liu, Hong Qiao, Li-Hao Jia, Lei Xu. 2014. A graph matching algorithm based on concavely regularized convex relaxation. *Neurocomputing* **134**, 140-148. [[CrossRef](#)]
8. Shuo-Wen Jin, Qiusheng Gu, Song Huang, Yong Shi, Long-Long Feng. 2014. COLOR-MAGNITUDE DISTRIBUTION OF FACE-ON NEARBY GALAXIES IN SLOAN DIGITAL SKY SURVEY DR7. *The Astrophysical Journal* **787**, 63. [[CrossRef](#)]
9. Kim-Han Thung, Chong-Yaw Wee, Pew-Thian Yap, Dinggang Shen. 2014. Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion. *NeuroImage* **91**, 386-400. [[CrossRef](#)]
10. Iain L. MacDonald. 2014. Numerical Maximisation of Likelihood: A Neglected Alternative to EM?. *International Statistical Review* n/a-n/a. [[CrossRef](#)]
11. Alireza Fallahi, Hassan Khotanlou, Mohammad Pooyan, Hassan Hashemi, Mohammad Ali Oghabian. 2014. SEGMENTATION OF UTERINE USING NEIGHBORHOOD INFORMATION AFFECTED POSSIBILISTIC FCM AND GAUSSIAN MIXTURE MODEL IN UTERINE FIBROID PATIENTS MRI. *Biomedical Engineering: Applications, Basis and Communications* **26**, 1450010. [[CrossRef](#)]
12. Alexander Krull, André Steinborn, Vaishnavi Ananthanarayanan, Damien Ramunno-Johnson, Uwe Petersohn, Iva M. Tolić-Nørrelykke. 2014. A divide and

conquer strategy for the maximum likelihood localization of low intensity objects. *Optics Express* **22**, 210. [[CrossRef](#)]

13. Marcos López de Prado, Matthew D. Foreman. 2013. A mixture of Gaussians approach to mathematical portfolio oversight: the EF3M algorithm. *Quantitative Finance* 1-18. [[CrossRef](#)]
14. Kevin R. Haas, Haw Yang, Jih-Wei Chu. 2013. Expectation-Maximization of the Potential of Mean Force and Diffusion Coefficient in Langevin Dynamics from Single Molecule FRET Data Photon by Photon. *The Journal of Physical Chemistry B* **117**, 15591-15605. [[CrossRef](#)]
15. Jouni Pohjalainen, Okko Räsänen, Serdar Kadioglu. 2013. Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Computer Speech & Language* . [[CrossRef](#)]
16. Abhishek Srivastav, Ashutosh Tewari, Bing Dong. 2013. Baseline building energy modeling and localized uncertainty quantification using Gaussian mixture models. *Energy and Buildings* **65**, 438-447. [[CrossRef](#)]
17. Rita Simões, Christoph Mönninghoff, Martha Dlugaj, Christian Weimar, Isabel Wanke, Anne-Marie van Cappellen van Walsum, Cornelis Slump. 2013. Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images. *Magnetic Resonance Imaging* **31**, 1182-1189. [[CrossRef](#)]
18. Junjun Xu, Haiyong Luo, Fang Zhao, Rui Tao, Yiming Lin, Hui Li. 2013. The WiMap. *International Journal of Advanced Pervasive and Ubiquitous Computing* 3:10.4018/japuc.20110101, 29-38. [[CrossRef](#)]
19. Huaifei Hu, Haihua Liu, Zhiyong Gao, Lu Huang. 2013. Hybrid segmentation of left ventricle in cardiac MRI using gaussian-mixture model and region restricted dynamic programming. *Magnetic Resonance Imaging* **31**, 575-584. [[CrossRef](#)]
20. References 629-711. [[CrossRef](#)]
21. Manuel Samuelides Response Surface Methodology and Reduced Order Models 17-64. [[CrossRef](#)]
22. J. Andrew Howe, Hamparsum Bozdogan. 2013. Robust mixture model cluster analysis using adaptive kernels. *Journal of Applied Statistics* **40**, 320-336. [[CrossRef](#)]
23. Ulvi Baspinar, Huseyin Selcuk Varol, Volkan Yusuf Senyurek. 2013. Performance Comparison of Artificial Neural Network and Gaussian Mixture Model in Classifying Hand Motions by Using sEMG Signals. *Biocybernetics and Biomedical Engineering* **33**, 33-45. [[CrossRef](#)]
24. Jian Sun. 2012. A Fast MEANSHIFT Algorithm-Based Target Tracking System. *Sensors* **12**, 8218-8235. [[CrossRef](#)]
25. Dawei Li, Lihong Xu, Erik Goodman. 2012. On-line EM Variants for Multivariate Normal Mixture Model in Background Learning and Moving Foreground Detection. *Journal of Mathematical Imaging and Vision* . [[CrossRef](#)]

26. Miin-Shen Yang, Chien-Yo Lai, Chih-Ying Lin. 2012. A robust EM clustering algorithm for Gaussian mixture models. *Pattern Recognition* **45**, 3950-3961. [[CrossRef](#)]
27. Erik Cuevas, Felipe Sención, Daniel Zaldivar, Marco Pérez-Cisneros, Humberto Sossa. 2012. A multi-threshold segmentation approach based on Artificial Bee Colony optimization. *Applied Intelligence* **37**, 321-336. [[CrossRef](#)]
28. Sung-Mahn Ahn, Suhn-Beom Kwon. 2012. Choosing the Tuning Constant by Laplace Approximation. *Communications of the Korean statistical society* **19**, 597-605. [[CrossRef](#)]
29. Chul-Hee Lee, Sung-Mahn Ahn. 2012. Parallel Implementations of the Self-Organizing Network for Normal Mixtures. *Communications of the Korean statistical society* **19**, 459-469. [[CrossRef](#)]
30. A. Ramezani, B. Moshiri, B. Abdulhai, A.R. Kian. 2012. Distributed maximum likelihood estimation for flow and speed density prediction in distributed traffic detectors with Gaussian mixture model assumption. *IET Intelligent Transport Systems* **6**, 215. [[CrossRef](#)]
31. Ziheng Wang, Ryan M. Hope, Zuoguan Wang, Qiang Ji, Wayne D. Gray. 2012. Cross-subject workload classification with a hierarchical Bayes model. *NeuroImage* **59**, 64-69. [[CrossRef](#)]
32. Zheng You, Jian Sun, Fei Xing, Gao-Fei Zhang. 2011. A Novel Multi-Aperture Based Sun Sensor Based on a Fast Multi-Point MEANSHIFT (FMMS) Algorithm. *Sensors* **11**, 2857-2874. [[CrossRef](#)]
33. Yu-Ren Lai, Kuo-Liang Chung, Guei-Yin Lin, Chyou-Hwa Chen. 2011. Gaussian mixture modeling of histograms for contrast enhancement. *Expert Systems with Applications* . [[CrossRef](#)]
34. Sung-Mahn Ahn, Myeong-Kyun Kim. 2011. A Self-Organizing Network for Normal Mixtures. *Communications of the Korean statistical society* **18**, 837-849. [[CrossRef](#)]
35. S. Venkatesh, S. Gopal. 2011. Robust Heteroscedastic Probabilistic Neural Network for multiple source partial discharge pattern recognition – Significance of outliers on classification capability. *Expert Systems with Applications* **38**, 11501-11514. [[CrossRef](#)]
36. Jae-Sik Jeong. 2011. A Finite Mixture Model for Gene Expression and Methylation Profiles in a Bayesian Framework. *Korean Journal of Applied Statistics* **24**, 609-622. [[CrossRef](#)]
37. Yan Yang, Jinwen Ma. 2011. Asymptotic Convergence Properties of the EM Algorithm for Mixture of Experts. *Neural Computation* **23**:8, 2140-2168. [[Abstract](#)] [[Full Text](#)] [[PDF](#)] [[PDF Plus](#)] [[Supplementary Content](#)]
38. Jian Yu, Miin-Shen Yang, E. Stanley Lee. 2011. Sample-weighted clustering methods. *Computers & Mathematics with Applications* . [[CrossRef](#)]

39. A. Mailing, B. Cernuschi-Frías. 2011. A Method for Mixed States Texture Segmentation with Simultaneous Parameter Estimation. *Pattern Recognition Letters* . [[CrossRef](#)]
40. Lei Shi, Shikui Tu, Lei Xu. 2011. Learning Gaussian mixture with automatic model selection: A comparative study on three Bayesian related approaches. *Frontiers of Electrical and Electronic Engineering in China* **6**, 215-244. [[CrossRef](#)]
41. A. Lanatá, G. Valenza, C. Mancuso, E.P. Scilingo. 2011. Robust multiple cardiac arrhythmia detection through bispectrum analysis. *Expert Systems with Applications* **38**, 6798-6804. [[CrossRef](#)]
42. Xu Lei. 2011. Codimensional matrix pairing perspective of BYY harmony learning: hierarchy of bilinear systems, joint decomposition of data-covariance, and applications of network biology. *Frontiers of Electrical and Electronic Engineering in China* **6**, 86-119. [[CrossRef](#)]
43. Rikiya Takahashi. 2011. Sequential minimal optimization in convex clustering repetitions. *Statistical Analysis and Data Mining* n/a-n/a. [[CrossRef](#)]
44. Xiao-liang Tang, Min Han. 2010. Semi-supervised Bayesian ARTMAP. *Applied Intelligence* **33**, 302-317. [[CrossRef](#)]
45. Lei Xu. 2010. Bayesian Ying-Yang system, best harmony learning, and five action circling. *Frontiers of Electrical and Electronic Engineering in China* **5**, 281-328. [[CrossRef](#)]
46. Lei Xu. 2010. Machine learning problems from optimization perspective. *Journal of Global Optimization* **47**, 369-401. [[CrossRef](#)]
47. Behrooz Safarinejadian, Mohammad B. Menhaj, Mehdi Karrari. 2010. A distributed EM algorithm to estimate the parameters of a finite mixture of components. *Knowledge and Information Systems* **23**, 267-292. [[CrossRef](#)]
48. D. P. Vetrov, D. A. Kropotov, A. A. Osokin. 2010. Automatic determination of the number of components in the EM algorithm of restoration of a mixture of normal distributions. *Computational Mathematics and Mathematical Physics* **50**, 733-746. [[CrossRef](#)]
49. Erik Cuevas, Daniel Zaldivar, Marco Pérez-Cisneros. 2010. Seeking multi-thresholds for image segmentation with Learning Automata. *Machine Vision and Applications* . [[CrossRef](#)]
50. Abdenaceur Boudlal, Benayad Nsiri, Driss Aboutajdine. 2010. Modeling of Video Sequences by Gaussian Mixture: Application in Motion Estimation by Block Matching Method. *EURASIP Journal on Advances in Signal Processing* **2010**, 1-9. [[CrossRef](#)]
51. Otilia Boldea. 2009. Maximum Likelihood Estimation of the Multivariate Normal Mixture Model. *Journal of the American Statistical Association* **104**, 1539-1549. [[CrossRef](#)]

52. Roy Kwang Yang Chang, Chu Kiong Loo, M. V. C. Rao. 2009. Enhanced probabilistic neural network with data imputation capabilities for machine-fault classification. *Neural Computing and Applications* **18**, 791-800. [[CrossRef](#)]
53. Guobao Wang, Larry Schultz, Jinyi Qi. 2009. Statistical Image Reconstruction for Muon Tomography Using a Gaussian Scale Mixture Model. *IEEE Transactions on Nuclear Science* **56**, 2480-2486. [[CrossRef](#)]
54. Hyeyoung Park, Tomoko Ozeki. 2009. Singularity and Slow Convergence of the EM algorithm for Gaussian Mixtures. *Neural Processing Letters* **29**, 45-59. [[CrossRef](#)]
55. Siddhartha Ghosh, Dirk Froebrich, Alex Freitas. 2008. Robust autonomous detection of the defective pixels in detectors using a probabilistic technique. *Applied Optics* **47**, 6904. [[CrossRef](#)]
56. O. Michailovich, A. Tannenbaum. 2008. Segmentation of Tracking Sequences Using Dynamically Updated Adaptive Learning. *IEEE Transactions on Image Processing* **17**, 2403-2412. [[CrossRef](#)]
57. C NGUYEN, K CIOS. 2008. GAKREM: A novel hybrid clustering algorithm. *Information Sciences* **178**, 4205-4227. [[CrossRef](#)]
58. C.K. Reddy, Hsiao-Dong Chiang, B. Rajaratnam. 2008. TRUST-TECH-Based Expectation Maximization for Learning Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**, 1146-1157. [[CrossRef](#)]
59. Dongbing Gu. 2008. Distributed EM Algorithm for Gaussian Mixtures in Sensor Networks. *IEEE Transactions on Neural Networks* **19**, 1154-1166. [[CrossRef](#)]
60. Michael J. Boedigheimer, John Ferbas. 2008. Mixture modeling approach to flow cytometry data. *Cytometry Part A* **73A**:10.1002/cyto.a.v73a:5, 421-429. [[CrossRef](#)]
61. Xing Yuan, Zhenghui Xie, Miaoling Liang. 2008. Spatiotemporal prediction of shallow water table depths in continental China. *Water Resources Research* **44**, n/a-n/a. [[CrossRef](#)]
62. John J. Kasianowicz, Sarah E. Henrickson, Jeffery C. Lerman, Martin Misakian, Rekha G. Panchal, Tam Nguyen, Rick Gussio, Kelly M. Halverson, Sina Bavari, Devanand K. Shenoy, Vincent M. Stanford The Detection and Characterization of Ions, DNA, and Proteins Using Nanometer-Scale Pores . [[CrossRef](#)]
63. Michael Lynch, Ovidiu Ghita, Paul F. Whelan. 2008. Segmentation of the Left Ventricle of the Heart in 3-D+t MRI Data Using an Optimized Nonrigid Temporal Model. *IEEE Transactions on Medical Imaging* **27**, 195-203. [[CrossRef](#)]
64. A. Haghbin, P. Azmi. 2008. Precoding in downlink multi-carrier code division multiple access systems using expectation maximisation algorithm. *IET Communications* **2**, 1279. [[CrossRef](#)]
65. Estevam R. Hruschka, Eduardo R. Hruschka, Nelson F. F. Ebecken. 2007. Bayesian networks for imputation in classification problems. *Journal of Intelligent Information Systems* **29**, 231-252. [[CrossRef](#)]

66. Oleg Michailovich, Yogesh Rath, Allen Tannenbaum. 2007. Image Segmentation Using Active Contours Driven by the Bhattacharyya Gradient Flow. *IEEE Transactions on Image Processing* **16**, 2787-2801. [[CrossRef](#)]
67. L. XU. 2007. A unified perspective and new results on RHT computing, mixture based learning, and multi-learner based problem solving#. *Pattern Recognition* **40**, 2129-2153. [[CrossRef](#)]
68. J. W. F. Robertson, C. G. Rodrigues, V. M. Stanford, K. A. Robinson, O. V. Krasilnikov, J. J. Kasianowicz. 2007. Single-molecule mass spectrometry in solution using a solitary nanopore. *Proceedings of the National Academy of Sciences* **104**, 8207-8211. [[CrossRef](#)]
69. Miguel A. Carreira-Perpinan. 2007. Gaussian Mean-Shift Is an EM Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**, 767-776. [[CrossRef](#)]
70. Chunhua Shen, Michael J. Brooks, Anton van den Hengel. 2007. Fast Global Kernel Density Mode Seeking: Applications to Localization and Tracking. *IEEE Transactions on Image Processing* **16**, 1457-1469. [[CrossRef](#)]
71. Z. LIU, J. ALMHANA, V. CHOULAKIAN, R. MCGORMAN. 2006. Online EM algorithm for mixture with application to internet traffic modeling. *Computational Statistics & Data Analysis* **50**, 1052-1071. [[CrossRef](#)]
72. L.C. Khor. 2006. Robust adaptive blind signal estimation algorithm for underdetermined mixture. *IEE Proceedings - Circuits, Devices and Systems* **153**, 320. [[CrossRef](#)]
73. J. MA, S. FU. 2005. On the correct convergence of the EM algorithm for Gaussian mixtures#. *Pattern Recognition* **38**, 2602-2611. [[CrossRef](#)]
74. Grzegorz A. Rempala, Richard A. Derrig. 2005. Modeling Hidden Exposures in Claim Severity Via the Em Algorithm. *North American Actuarial Journal* **9**, 108-128. [[CrossRef](#)]
75. Carlos Ordonez, Edward Omiecinski. 2005. Accelerating EM clustering to find high-quality solutions. *Knowledge and Information Systems* **7**, 135-157. [[CrossRef](#)]
76. M. YANG. 2005. Estimation of parameters in latent class models using fuzzy clustering algorithms. *European Journal of Operational Research* **160**, 515-531. [[CrossRef](#)]
77. J. Fan, H. Luo, A.K. Elmagarmid. 2004. Concept-Oriented Indexing of Video Databases: Toward Semantic Sensitive Retrieval and Browsing. *IEEE Transactions on Image Processing* **13**, 974-992. [[CrossRef](#)]
78. Balaji Padmanabhan, Alexander Tuzhilin. 2003. On the Use of Optimization for Data Mining: Theoretical Interactions and eCRM Opportunities. *Management Science* **49**, 1327-1343. [[CrossRef](#)]
79. Z. Zhang. 2003. EM algorithms for Gaussian mixtures with split-and-merge operation. *Pattern Recognition* **36**, 1973-1983. [[CrossRef](#)]

80. Meng-Fu Shih, A.O. Hero. 2003. Unicast-based inference of network link delay distributions with finite mixture models. *IEEE Transactions on Signal Processing* **51**, 2219-2228. [[CrossRef](#)]
81. R.D. Nowak. 2003. Distributed EM algorithms for density estimation and clustering in sensor networks. *IEEE Transactions on Signal Processing* **51**, 2245-2253. [[CrossRef](#)]
82. Sin-Horng Chen, Wen-Hsing Lai, Yih-Ru Wang. 2003. A new duration modeling approach for mandarin speech. *IEEE Transactions on Speech and Audio Processing* **11**, 308-320. [[CrossRef](#)]
83. Y. Matsuyama. 2003. The α -EM algorithm: surrogate likelihood maximization using α -logarithmic information measures. *IEEE Transactions on Information Theory* **49**, 692-706. [[CrossRef](#)]
84. M.A.F. Figueiredo, A.K. Jain. 2002. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 381-396. [[CrossRef](#)]
85. J PERES, R OLIVEIRA, S FEYODEAZEVEDO. 2001. Knowledge based modular networks for process modelling and control. *Computers & Chemical Engineering* **25**, 783-791. [[CrossRef](#)]
86. Zheng Rong Yang, M. Zwolinski. 2001. Mutual information theory for adaptive mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**, 396-403. [[CrossRef](#)]
87. C Harris. 2001. State estimation and multi-sensor data fusion using data-based neurofuzzy local linearisation process models. *Information Fusion* **2**, 17-29. [[CrossRef](#)]
88. H. Yin, N.M. Allinson. 2001. Self-organizing mixture networks for probability density estimation. *IEEE Transactions on Neural Networks* **12**, 405-411. [[CrossRef](#)]
89. Lei Xu. 2001. Best Harmony, Unified RPCL and Automated Model Selection for Unsupervised and Supervised Learning on Gaussian Mixtures, Three-Layer Nets and ME-RBF-SVM Models. *International Journal of Neural Systems* **11**, 43-69. [[CrossRef](#)]
90. H. Yin, N.M. Allinson. 2001. Bayesian self-organising map for Gaussian mixtures. *IEE Proceedings - Vision, Image, and Signal Processing* **148**, 234. [[CrossRef](#)]
91. C.J. Harris, X. Hong. 2001. Neurofuzzy mixture of experts network parallel learning and model construction algorithms. *IEE Proceedings - Control Theory and Applications* **148**, 456. [[CrossRef](#)]
92. Qiang Gan, C.J. Harris. 2001. A hybrid learning scheme combining EM and MASMOD algorithms for fuzzy local linearization modeling. *IEEE Transactions on Neural Networks* **12**, 43-53. [[CrossRef](#)]
93. Jinwen Ma, Lei Xu, Michael I. Jordan. 2000. Asymptotic Convergence Rate of the EM Algorithm for Gaussian Mixtures. *Neural Computation* **12**:12, 2881-2907. [[Abstract](#)] [[PDF](#)] [[PDF Plus](#)]

94. Dirk Husmeier. 2000. The Bayesian Evidence Scheme for Regularizing Probability-Density Estimating Neural Networks. *Neural Computation* **12**:11, 2685-2717. [[Abstract](#)] [[PDF](#)] [[PDF Plus](#)]
95. Ashish Singhal, Dale E. Seborg. 2000. Dynamic data rectification using the expectation maximization algorithm. *AIChE Journal* **46**:10.1002/aic.v46:8, 1556-1565. [[CrossRef](#)]
96. P. Hedelin, J. Skoglund. 2000. Vector quantization based on Gaussian mixture models. *IEEE Transactions on Speech and Audio Processing* **8**, 385-401. [[CrossRef](#)]
97. Man-Wai Mak, Sun-Yuan Kung. 2000. Estimation of elliptical basis function parameters by the EM algorithm with application to speaker verification. *IEEE Transactions on Neural Networks* **11**, 961-969. [[CrossRef](#)]
98. J. Peres, R. Oliveira, S. Feyeo de Azevedo Knowledge based modular networks for process modelling and control 247-252. [[CrossRef](#)]
99. M. Zwolinski, Z.R. Yang, T.J. Kazmierski. 2000. Using robust adaptive mixing for statistical fault macromodelling. *IEE Proceedings - Circuits, Devices and Systems* **147**, 265. [[CrossRef](#)]
100. K Chen. 1999. Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks* **12**, 1229-1252. [[CrossRef](#)]
101. N.M. Allinson, H. Yin Self-Organising Maps for Pattern Recognition 111-120. [[CrossRef](#)]
102. Z Yang. 1998. Robust maximum likelihood training of heteroscedastic probabilistic neural networks. *Neural Networks* **11**, 739-747. [[CrossRef](#)]
103. E Alpaydin. 1998. Soft vector quantization and the EM algorithm. *Neural Networks* **11**, 467-477. [[CrossRef](#)]
104. Athanasios Kehagias, Vassilios Petridis. 1997. Time-Series Segmentation Using Predictive Modular Neural Networks. *Neural Computation* **9**:8, 1691-1709. [[Abstract](#)] [[PDF](#)] [[PDF Plus](#)]
105. A.V. Rao, D. Miller, K. Rose, A. Gersho. 1997. Mixture of experts regression modeling by deterministic annealing. *IEEE Transactions on Signal Processing* **45**, 2811-2820. [[CrossRef](#)]
106. James R. Williamson. 1997. A Constructive, Incremental-Learning Network for Mixture Modeling and Classification. *Neural Computation* **9**:7, 1517-1543. [[Abstract](#)] [[PDF](#)] [[PDF Plus](#)]
107. Yiu-Ming Cheung, Wai-Man Leung, Lei Xu. 1997. Adaptive Rival Penalized Competitive Learning and Combined Linear Predictor Model for Financial Forecast and Investment. *International Journal of Neural Systems* **08**, 517-534. [[CrossRef](#)]
108. Lei Xu, Shun-ichi Amari Combining Classifiers and Learning Mixture-of-Experts 243-252. [[CrossRef](#)]

109. Junjun Xu, Haiyong Luo, Fang Zhao, Rui Tao, Yiming Lin, Hui LiThe WiMap 31-41. [[CrossRef](#)]
110. Pasquale De Meo, Antonino Nocera, Domenico UrsinoA Component-Based Framework for the Integration and Exploration of XML Sources 343-377. [[CrossRef](#)]