

Learning Mixture Models With the Regularized Latent Maximum Entropy Principle

Shaojun Wang, Dale Schuurmans, Fuchun Peng, and Yunxin Zhao, *Senior Member, IEEE*

Abstract—This paper presents a new approach to estimating mixture models based on a recent inference principle we have proposed: the latent maximum entropy principle (LME). LME is different from Jaynes' maximum entropy principle, standard maximum likelihood, and maximum *a posteriori* probability estimation. We demonstrate the LME principle by deriving new algorithms for mixture model estimation, and show how robust new variants of the expectation maximization (EM) algorithm can be developed. We show that a regularized version of LME (RLME), is effective at estimating mixture models. It generally yields better results than plain LME, which in turn is often better than maximum likelihood and maximum a posterior estimation, particularly when inferring latent variable models from small amounts of data.

Index Terms—Expectation maximization (EM), iterative scaling, latent variables, maximum entropy, mixture models, regularization.

I. INTRODUCTION

MIXTURE models are among the most enduring, well-established modeling techniques in statistical machine learning. In a typical application, sample data is thought of as originating from various possible sources, where the data from each particular source is modeled by a familiar form. Given labeled and unlabeled data from a weighted combination of these sources, the goal is to estimate the generating mixture distribution, that is, the nature of each source and the ratio with which each source is present.

The most popular computational method for estimating parametric mixture models is the expectation-maximization (EM) algorithm, first formalized by [10]. EM is an iterative parameter-optimization technique that is guaranteed to converge to a local maxima in likelihood or posterior probability. It is widely applicable to latent variable models, has proven useful for applications in estimation, regression and classification, and also has well investigated theoretical foundations [10], [20], [26]. However, a number of key issues remain unresolved. For example, since the likelihood function or posterior probability for mixture models typically has multiple local maxima, there is a question of which local maximizer to choose as the final estimate. Fisher's classical maximum likelihood estimation (MLE) principle states that the desired estimate corresponds to the global maximizer of the likelihood function or posterior probability, in

situations where the likelihood function is bounded over the parameter space. Unfortunately, in many cases, such as mixtures of Gaussians with unequal covariances, the likelihood function is unbounded. In such situations, the choice of local maxima is not obvious, and the final selection requires careful consideration in practice. Another open issue is generalization. That is, in practice, it is often observed that estimating mixture models by MLE leads to overfitting (poor generalization) particularly when faced with limited training data. The maximum a posterior (MAP) estimation principle is developed to alleviate the overfitting problem, however in situation of unbounded case, there is no prior existed to overcome this problem.

To address these issues, we have recently proposed a new statistical machine learning framework for density estimation and pattern classification, which we refer to as the latent maximum entropy (LME) principle [25]. Although classical statistics is heavily based on parametric models, such models can sometimes be restrictive and can lead to departures from reality. As data becomes more abundant in the form of modern applications such as data mining, more flexible nonparametric models often become more appropriate [13]. However, when only a "small" amount of data is available, such as in statistical language modeling, such restrictive models are sometimes the best we can do without overfitting. The alternative principle we propose, LME, is a nonparametric approach based on matching a set of features in the data (i.e., sufficient statistics, weak learners, or basis functions). The technique becomes parametric when we necessarily have to approximate the principle. LME is an extension to Jaynes' maximum entropy (ME) principle that explicitly incorporates latent variables in the formulation, and thereby extends the original principle to cases where data components are missing. The resulting principle is different from both maximum likelihood estimation and standard maximum entropy, but often yields better estimates in the presence of hidden variables and limited training data. In this paper, we show a further extension of LME, the regularized LME principle, and demonstrate its advantage over LME as well as MAP for estimating mixture models.

II. MOTIVATION

We first repeat a standard example used to motivate the LME extension to Jaynes' standard ME principle. Assume we observe a random variable Y that reports people's heights in a population. Given sample data $\tilde{Y} = (y_1, \dots, y_T)$, one might believe that simple statistics such as the sample mean and sample mean square of Y are well represented in the data. If so, then Jaynes' ME principle [14] suggests that one should infer a distribution for Y that has maximum entropy, subject to the con-

Manuscript received March 15, 2003; revised October 23, 2003.

S. Wang and D. Schuurmans are with the Department of Computing Science, University of Alberta, Alberta T6G 2E8, Canada.

F. Peng is with the Department of Computer Science, University of Massachusetts, Amherst, MA 01003 USA.

Y. Zhao is with the Department of Computer Engineering and Computer Science, University of Missouri, Columbia, MO 65201 USA.

Digital Object Identifier 10.1109/TNN.2004.828755

straints that the mean and mean square values of Y match the sample values; that is, that $EY = m_1$ and $EY^2 = m_2$, where $m_1 = (1/T) \sum_{t=1}^T y_t$ and $m_2 = (1/T) \sum_{t=1}^T y_t^2$, respectively. In this case, it is known that the maximum entropy solution is a Gaussian density with mean m_1 and variance $m_2 - m_1^2$, $p(y) = N(y; m_1, m_2 - m_1^2)$; a consequence of the well-known fact that a Gaussian random variable has the largest differential entropy of any random variable for a specified mean and variance [6].

However, assume further that after observing the data we find that there are actually two peaks in the histogram. Obviously the standard ME solution would not be the most appropriate model for such bimodal data, because it will continue to postulate a unimodal distribution. However, the existence of the two peaks might be due to the fact that there are two subpopulations in the data, male and female, each of which have different height distributions. In this case, each height measurement Y has an accompanying (hidden) gender label C that indicates which subpopulation the measurement is taken from. One way to incorporate this information is to *explicitly* add the missing label data. That is, we could let $X = (Y, C)$, where Y denotes a person's height and C is the gender label, and then obtain *labeled* measurements $(y_1, c_1, \dots, y_T, c_T)$. The problem then is to find a joint model $p(x) = p(y, c)$ that maximizes entropy while matching the expectations over $\delta_k(c)$, $y\delta_k(c)$, and $y^2\delta_k(c)$, for $k = 1, 2$. In this fully observed data case, *where we witness the gender label C* , the ME principle poses a separable optimization problem that has a unique solution: $p(x) = p(y, c)$ is a mixture of two Gaussian distributions specified by $p(c) = \theta_c = (N_c/T)$ and $p(y|c) = N(y; \mu_c, \sigma_c^2)$, where $\mu_c = (1/N_c) \sum_{t=1}^T y_t \delta_c(c_t)$ and $\sigma_c^2 = (1/N_c) \sum_{t=1}^T (y_t - \mu_c)^2 \delta_c(c_t)$ for $c = 1, 2$.

Unfortunately, obtaining fully labeled data is tedious or impossible in most realistic situations. In cases where variables are unobserved, Jaynes' ME principle, which is maximally non-committal with respect to missing information, becomes insufficient. For example, if the gender label is unobserved, one would still be reduced to inferring a unimodal Gaussian as above. To cope with missing but nonarbitrary hidden structure, we must extend the ME principle to account for the underlying causal structure in the data.

III. THE LME PRINCIPLE

To briefly recap, but also generalize the LME principle introduced in [25], let $X \in \mathcal{X}$ be a random variable denoting the complete data, $Y \in \mathcal{Y}$ be the observed incomplete data, and $Z \in \mathcal{Z}$ be the missing data. That is, $X = (Y, Z)$. If we let $p(x)$ and $p(y)$ denote the densities of X and Y , respectively, and let $p(z|y)$ denote the conditional density of Z given Y , then $p(y) = \int_{z \in \mathcal{Z}} p(x) \mu(dz)$ where $p(x) = p(y)p(z|y)$.

LME Principle: Given features f_1, \dots, f_N , specifying the properties we would like to match in the data, select a joint probability model p^* from the space of all distributions \mathcal{P} over \mathcal{X} to maximize the joint entropy

$$\max_p H(p) = - \int_{x \in \mathcal{X}} p(x) \log p(x) \mu(dx) \quad (1)$$

subject to the constraints

$$\int_{x \in \mathcal{X}} f_i(x) p(x) \mu(dx) = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x) p(z|y) \mu(dz) \quad i = 1 \dots N, \quad Y \text{ and } Z \text{ not independent} \quad (2)$$

where $x = (y, z)$.

Here $\tilde{p}(y)$ is the empirical distribution over the observed data, and \mathcal{Y} denotes the set of observed Y values. Intuitively, the constraints specify that we require the expectations of $f_i(X)$ in the complete model to match their empirical expectations on the complete data Y , taking into account the structure of the dependence of the unobserved component Z on Y .

In many cases we will also find it useful to consider an interesting generalization of the LME principle that seeks joint distributions p that minimize the relative entropy between p and a reference (default) distribution q .

Generalized LME Principle: Given a default distribution q , select a joint probability model p^* from the space of all distributions \mathcal{P} over \mathcal{X} to minimize the relative entropy

$$\min_p D(p(x)||q(x)) = D(p(z)||q(z)) + D(p(y|z)||q(y|z)) \quad (3)$$

subject to the constraints

$$\int_{x \in \mathcal{X}} f_i(x) p(x) \mu(dx) = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x) p(z|y) \mu(dz) \quad i = 1 \dots N, \quad Y \text{ and } Z \text{ not independent.} \quad (4)$$

Notice that (1) and (2) are special cases of (3) and (4) when we set the default distribution q to be uniform.

Before we apply the LME principle to mixture models, we first consider a small improvement that will prove useful. In many statistical modeling situations, the constraints used in the maximum entropy principle are subject to errors due to the empirical data, especially in a very sparse domain. One way to gain robustness to these errors is to relax the constraints but add a penalty to the entropy of the joint model [7].

Regularized Generalized LME Principle (RLME): Given a default distribution q on \mathcal{X} , select a joint probability model p^* from the space of all distributions \mathcal{P} over \mathcal{X} to minimize the regularized relative entropy

$$\min_{p,a} D(p(x)||q(x)) + U(a) = D(p(z)||q(z)) + D(p(y|z)||q(y|z)) + U(a) \quad (5)$$

subject to the constraints

$$\int_{x \in \mathcal{X}} f_i(x) p(x) \mu(dx) = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x) p(z|y) \mu(dz) - a_i \quad i = 1 \dots N, \quad Y \text{ and } Z \text{ not independent.} \quad (6)$$

Here $a = (a_1, \dots, a_N)$, a_i is the error for each constraint, and $U: \mathfrak{R}^N \rightarrow \mathfrak{R}$ is a smoothing convex function [7] which has its minimum at 0. The regularization term U can be used to penalize deviations in more reliably observed constraints to

a greater degree than deviations in less reliably observed constraints.

Again, (3) and (4) are special cases of (5) and (6) when we set the cost function U to be constant. So in the following we will refer (5) and (6) as the RLME principle.

Unfortunately, there is no simple solution for p^* in (5) and (6). However, a good approximation can be obtained by restricting the model to have an exponential form

$$p_\lambda(x) = \Phi_\lambda^{-1} q(x) \exp\left(\sum_{i=1}^N \lambda_i f_i(x)\right)$$

where $\Phi_\lambda = \int_{x \in \mathcal{X}} q(x) \exp(\sum_{i=1}^N \lambda_i f_i(x)) \mu(dx)$ is a normalizing constant that ensures $\int_{x \in \mathcal{X}} p_\lambda(x) \mu(dx) = 1$. This restriction provides a free parameter λ_i for each feature function f_i . By adopting such a ‘‘log-linear’’ restriction, it turns out that we can formulate a practical iterative algorithm for finding feasible solutions (below) to approximately satisfying the RLME principle. Our algorithmic strategy then is to generate many feasible candidates (by restarting the iterative procedure at different initial points), evaluate their regularized entropy and select the best model. The hardest part of this process is generating feasible solutions.

IV. A TRAINING ALGORITHM FOR LOG-LINEAR MODELS

The key observation to finding feasible solutions is to note that they are intimately related to finding locally *maximum a posteriori* (MAP) solutions¹. Given a penalty function U over errors a , an associated *prior* U^* on λ can be obtained by setting U^* to the convex (Fenchel) conjugate [5] of U . For example, given a quadratic penalty $U(a) = \sum_{i=1}^N (1/2)\sigma_i^2 a_i^2$, the convex conjugate $U^*(\lambda) = \sum_{i=1}^N (\lambda_i^2/2\sigma_i^2)$ can be determined by setting $a_i^* = (\lambda_i/\sigma_i^2)$; which specifies a Gaussian prior on λ . Then, given a prior U^* , note that the standard MAP estimate maximizes the penalized log-likelihood $R(\lambda) = \sum_y \tilde{p}(y) \log p_\lambda(y) - U^*(\lambda)$. Our key result is that locally maximizing $R(\lambda)$ is equivalent to satisfying the feasibility constraints (6) of the RLME principle.

Theorem 1: Under the log-linear assumption, locally maximizing a posterior probability of log-linear models on incomplete data is equivalent to satisfying the feasibility constraints of the RLME principle. That is, the only distinction between MAP and RLME in log-linear models is that, among local maxima (feasible solutions), RLME selects the model with the maximum regularized entropy, whereas MAP selects the model with the maximum posterior probability.

Proof: Define $L(\lambda) = \sum_y \tilde{p}(y) \log p_\lambda(y)$, then $R(\lambda) = L(\lambda) - U^*(\lambda)$. So we have

$$\frac{\partial R(\lambda)}{\partial \lambda_i} = \frac{\partial L(\lambda)}{\partial \lambda_i} - \frac{\partial U^*(\lambda)}{\partial \lambda_i}$$

Similar as in [25], we can show that

$$\begin{aligned} \frac{\partial L(\lambda)}{\partial \lambda_i} &= - \int_{x \in \mathcal{X}} f_i(x) p_\lambda(x) \mu(dx) \\ &\quad + \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x) p_\lambda(z|y) \mu(dz). \end{aligned}$$

By the definition of the convex conjugate [5], we have $U^*(\lambda) = \sup_a \{\langle \lambda, a \rangle - U(a)\}$. Thus, for all $\lambda \in \Omega$

$$\begin{aligned} \frac{\partial U^*(\lambda)}{\partial \lambda_i} &= \frac{\partial \{\sup_a \{\langle \lambda, a \rangle - U(a)\}\}}{\partial \lambda_i} \\ &= \frac{\partial \{\langle \lambda, a^* \rangle - U(a^*)\}}{\partial \lambda_i} \\ &= a_i^*. \end{aligned}$$

By setting $\partial R(\lambda)/\partial \lambda_i = 0$, for $i = 1, \dots, N$, we obtain the original constraints (6). Therefore, the feasible solutions of (6) satisfy the conditions for the stationary points of the posterior probability function. This establishes the first part of the theorem. The remainder of the proof follows the same argument as for LME [25].

This connection allows us to exploit an EM algorithm [10] to find *feasible* solutions to the RLME principle. It is important to emphasize, however, that EM will only find alternative feasible solutions, while the RLME and penalized MAP principles will differ markedly in the feasible solutions they prefer. We illustrate this distinction below.

To formulate an EM algorithm for learning log-linear models, first decompose the penalized log-likelihood function $R(\lambda)$ into

$$R(\lambda) = \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \log p_\lambda(y) - U^*(\lambda) = Q(\lambda, \lambda') + H(\lambda, \lambda') \quad (7)$$

where $Q(\lambda, \lambda') = \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} p_{\lambda'}(z|y) \log p_\lambda(x) \mu(dz) - U^*(\lambda)$ and $H(\lambda, \lambda') = - \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} p_{\lambda'}(z|y) \log p_\lambda(z|y) \mu(dz)$. This is very similar to the standard decomposition used for deriving EM. For log-linear models, in particular, we have

$$\begin{aligned} Q(\lambda, \lambda^{(j)}) - G(\lambda^{(j)}) &= -\log(\Phi_\lambda) \\ &\quad + \sum_{i=1}^N \lambda_i \left(\sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x) p_{\lambda^{(j)}}(z|y) \mu(dz) \right) - U^*(\lambda) \quad (8) \end{aligned}$$

where $G(\lambda') = \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} p_{\lambda'}(z|y) \log q(x) \mu(dz)$.

Interestingly, it turns out that maximizing $Q(\lambda, \lambda^{(j)}) - G(\lambda^{(j)})$ as a function of λ for fixed $\lambda^{(j)}$ (the **M step**) is equivalent to solving another constrained optimization problem corresponding to a generalized maximum entropy principle, but a much simpler one than before.

Lemma 1: Maximizing $Q(\lambda, \lambda^{(j)}) - G(\lambda^{(j)})$ as a function of λ for fixed $\lambda^{(j)}$ is equivalent to solving

$$\begin{aligned} \min_{p, a} D(p(x)||q(x)) + U(a) \\ = \left(\int_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \mu(dx) + U(a) \right) \quad (9) \end{aligned}$$

¹In [25], the results are stated when the default model and cost function U are both chosen to be constant, but the results still hold in the present, more general situation.

$$\begin{aligned}
& \text{subject to } \int_{x \in \mathcal{X}} f_i(x)p(x)\mu(dx) \\
& = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x)p_{\lambda^{(j)}}(z|y)\mu(dz) - a_i, \\
& \quad i = 1 \dots N.
\end{aligned} \tag{10}$$

Proof: Define the Lagrangian $\Lambda(p, a, \lambda, \lambda^{(j)})$ by

$$\begin{aligned}
\Lambda(p, a, \lambda, \lambda^{(j)}) & = D(p||q) + U(a) + \sum_{i=1}^N \lambda_i \\
& \times \left(- \int_{x \in \mathcal{X}} p(x)f_i(x)\mu(dx) + \sum_{y \in \mathcal{Y}} \tilde{p}(y) \right. \\
& \quad \left. \times \int_{z \in \mathcal{Z}} p_{\lambda^{(j)}}(z|y)f_i(x)\mu(dz) - a_i \right). \tag{11}
\end{aligned}$$

Holding $\lambda^{(j)}$ fixed, compute the unconstrained maximum of the Lagrangian over $p \in \mathcal{P}$, to get

$$\begin{aligned}
p_\lambda & = \arg \min_{p \in \mathcal{P}} \Lambda(p, a, \lambda, \lambda^{(j)}) \\
& = q(x)\Phi_\lambda^{-1} \exp \left(\sum_{i=1}^N \lambda_i f_i(x) \right).
\end{aligned}$$

(This result is obtained by taking the derivative of (11) with respect to $p(x)$ and setting it to zero.) Now by plugging p_λ into $\Lambda(p_\lambda, a, \lambda, \lambda^{(j)})$, we obtain the dual function

$$\begin{aligned}
\Upsilon(\lambda, \lambda^{(j)}) & = \inf_{p, a} \Lambda(p, a, \lambda, \lambda^{(j)}) \\
& = \inf_a \Lambda(p_\lambda, a, \lambda, \lambda^{(j)}) \\
& = \inf_a \left(-\log(\Phi_\lambda) + \sum_{i=1}^N \lambda_i \sum_{y \in \mathcal{Y}} \tilde{p}(y) \right. \\
& \quad \left. \times \int_{z \in \mathcal{Z}} f_i(x)p_{\lambda^{(j)}}(z|y)\mu(dz) + U(a) - \langle \lambda, a \rangle \right) \\
& = \sup_a \left(-\log(\Phi_\lambda) + \sum_{i=1}^N \lambda_i \sum_{y \in \mathcal{Y}} \tilde{p}(y) \right. \\
& \quad \left. \times \int_{z \in \mathcal{Z}} f_i(x)p_{\lambda^{(j)}}(z|y)\mu(dz) - (\langle \lambda, a \rangle - U(a)) \right) \\
& = -\log(\Phi_\lambda) + \sum_{i=1}^N \lambda_i \sum_{y \in \mathcal{Y}} \tilde{p}(y) \\
& \quad \times \int_{z \in \mathcal{Z}} f_i(x)p_{\lambda^{(j)}}(z|y)\mu(dz) - U^*(\lambda)
\end{aligned}$$

which is exactly the $Q(\lambda^*, \lambda^{(j)}) - G(\lambda^{(j)})$ as given in (8). If we denote the optimal value of (9) subject to (10) as

$D(p_{\lambda^*}(\lambda^{(j)})||q) + U(a^*)$, then under the conditions where strong duality holds [4], [19] we have

$$\begin{aligned}
\max_\lambda Q(\lambda, \lambda^{(j)}) - G(\lambda^{(j)}) & = \max_\lambda \Upsilon(\lambda, \lambda^{(j)}) \\
& = \max_\lambda \min_a \Lambda(p_\lambda, a, \lambda, \lambda^{(j)}) \\
& = \max_\lambda \min_a \min_{p \in \mathcal{P}} \Lambda(p, a, \lambda, \lambda^{(j)}) \\
& = D(p_{\lambda^*}(\lambda^{(j)})||q) + U(a^*). \tag{12}
\end{aligned}$$

It is critical to realize that the new constrained optimization problem in Lemma 1 is much easier than maximizing (1) subject to (2) for log-linear models, because the right-hand side of the constraints (10) no longer depends on λ but rather on the fixed constants from the previous iteration $\lambda^{(j)}$. This means that maximizing (9) subject to (10) with respect to λ is now a convex optimization problem with linear constraints. The generalized iterative scaling algorithm (GIS) [8] or improved iterative scaling algorithm (IIS) [9] can be used to maximize $Q(\lambda, \lambda^{(j)}) - G(\lambda^{(j)})$ very efficiently.

From these observations, we can recover feasible log-linear models by using an algorithm that combines EM with nested iterative scaling to calculate the **M step**.

Assuming we use the Gaussian prior of λ , then the explicit iterative procedures we obtain will be

R-EM-IS Algorithm:

E step: Given $\lambda^{(j)}$, for each feature f_i , $i = 1, \dots, N$, calculate its current expectation $\eta_i^{(j)}$ with respect to $\lambda^{(j)}$ by:

$$\eta_i^{(j)} = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x)p_{\lambda^{(j)}}(z|y)\mu(dz)$$

M step: Perform S iterations of full parallel update of parameter values $\lambda_1, \dots, \lambda_N$ either by GIS or IIS as follows. Each update is given by

$$\lambda_i^{(j+\frac{s}{S})} = \lambda_i^{(j+\frac{s-1}{S})} + \gamma_i \tag{13}$$

where $\gamma_i^{(j+s/S)}$ satisfies

$$\begin{aligned}
& \int_{x \in \mathcal{X}} f_i(x)e^{\gamma_i f(x)} p_{\lambda^{(j+\frac{s-1}{S})}}(x)\mu(dx) + \\
& \frac{\lambda_i^{(j+\frac{s-1}{S})} + \gamma_i^{(j+\frac{s}{S})}}{\sigma_i^2} = \eta_i^{(j)} \tag{14}
\end{aligned}$$

where $f(x) = \sum_{k=1}^N f_k(x)$ and $s = 1, \dots, S$.

■

Provided that the E and **M steps** can both be computed, R-EM-IS can be shown to converge to a local maximum in

likelihood for log-linear models, and hence is guaranteed to yield feasible solutions to the RLME principle.

Theorem 2: The R-EM-IS algorithm monotonically increases the penalized likelihood function $R(\lambda)$, and all limit points of any R-EM-IS sequence $\{\lambda^{(j+s/S)}, j \geq 0, s = 1 \dots S\}$, belong to the set

$$\Theta = \left\{ \lambda \in \mathbb{R}^N : \frac{\partial R(\lambda)}{\partial \lambda} = 0 \right\}. \quad (15)$$

Therefore, R-EM-IS asymptotically yields feasible solutions to the RLME principle for log-linear models.

Proof: The proof basically follows the same line as in [25].

Thus, R-EM-IS provides an effective means to find feasible solutions to the RLME principle. (We note that Lauritzen [16] has suggested a similar algorithm, but did not supply a convergence proof. More recently, Riezler [22] has also proposed an algorithm equivalent to setting $S = 1$ in EM-IS. However, we have found $S > 1$ to be more effective in many cases.)

We can now exploit the R-EM-IS algorithm to develop a practical approximation to the RLME principle.

R-ME-EM-IS Algorithm:

Initialization: Randomly choose initial guesses for λ .

R-EM-IS: Run R-EM-IS to convergence, to obtain feasible λ^* .

Entropy calculation: Calculate the regularized relative entropy of p_{λ^*} with respect to $q(x)$.

Model selection: Repeat the above steps several times to produce a set of distinct feasible candidates. Choose the feasible candidate that achieves the lowest relative entropy. ■

This leads to a new estimation technique that we will compare to standard MAP below. One apparent complication, first, is that we need to calculate the entropies of the candidate models produced by R-EM-IS. However, it turns out that we do not need to calculate entropies explicitly because one can recover the entropy of *feasible* log-linear models simply as a byproduct of running R-EM-IS to convergence.

Corollary 1: If λ^* is feasible, then $Q(\lambda^*, \lambda^*) - G(\lambda^*) = D(p_{\lambda^*} \| q) + U(a^*)$, and $R(\lambda^*) = D(p_{\lambda^*} \| q) + U(a^*) + G(\lambda^*) + H(\lambda^*, \lambda^*)$.

Proof: By (7), we know that $R(\lambda) = Q(\lambda, \lambda) + H(\lambda, \lambda)$ for all $\lambda \in \Theta$. Let $\lambda^{(j+1)} = \arg \max_{\lambda} Q(\lambda, \lambda^{(j)})$. Then from (12) we obtain $Q(\lambda^{(j+1)}, \lambda^{(j)}) - G(\lambda^{(j)}) = \max_{\lambda} Q(\lambda, \lambda^{(j)}) - G(\lambda^{(j)}) = D(p_{\lambda^*}(\lambda^{(j)}) \| q) + U^*(\lambda^*)$. Now, using the same argument as in the proof of Theorem 2, we can show that all limit points of the sequence $\{\lambda^{(j+1)}, j \geq 0\}$ belong to the set Θ , and therefore $Q(\lambda, \lambda) - G(\lambda) = D(p_{\lambda} \| q) + U(a^*)$ for all $\lambda \in \Theta$. Thus, we have $R(\lambda) = D(p_{\lambda} \| q) + U(a^*) + G(\lambda) + H(\lambda, \lambda)$ for all $\lambda \in \Theta$.

Therefore, at a feasible solution λ^* , we have already calculated the regularized relative entropy, $Q(\lambda^*, \lambda^*)$, in the **M step** of R-EM-IS.

To draw a clear distinction between RLME and MAP, assume that the term $G(\lambda^*) + H(\lambda^*, \lambda^*)$ from Corollary 1 is constant across different feasible solutions. Then MAP, which maximizes $R(\lambda^*)$, will choose the model that has maximum posterior probability, whereas RLME, which minimizes $D(p_{\lambda^*}(X) \| q(X)) + U(a^*)$, will choose a model that has minimum regularized entropy with respect to default model q . (Of course, $G(\lambda^*) + H(\lambda^*, \lambda^*)$ will not be constant in practice and the comparison between MAP and RLME is not so straightforward, but this example does highlight their difference.) The fact that RLME and MAP are different raises the question of which method is the most effective when inferring a model from sample data. To address this question we turn to a comparison.

V. RLME FOR LEARNING MIXTURE MODELS

In the traditional approach to mixture models [20], the distribution of data is assumed to have a parametric form with unknown parameters. In our approach, we do not make assumptions about the form of the source but rather specify a set of features we would like to match in the data. Here we show that by choosing certain sets of features, we can recover familiar mixture models by LME principle. Then we present the corresponding regularized RLME version which leads to MAP estimation.

A. Gaussian Mixtures

Let $X = (Y, C)$, where Y is an observable M dimensional random vector and $C \in \{1, \dots, K\}$ denotes a hidden class index. Consider the features: $f_0^k(x) = \delta_k(c)$, $f_\ell^k(x) = y_\ell \delta_k(c)$, $f_{\ell, m}^k(x) = y_\ell y_m \delta_k(c)$, for $\ell, m = 1, \dots, M$, $k = 1, \dots, K$, where $\delta_k(c)$ denotes the indicator function of the event $c = k$. Then, given the observed data $\tilde{Y} = (y^1, \dots, y^T)$, the initial LME principle can be formulated as

$$\begin{aligned} \min_{p(x)} D(p(x) \| q(x)) &= D(p(c) \| q(c)) \\ &\quad + D(p(y|c) \| q(y|c)) \\ \text{subject to } \int_{x \in \mathcal{X}} \delta_k(c) p(x) \mu(dx) &= \sum_{y \in \tilde{Y}} \tilde{p}(y) \sum_{c=1}^K \delta_k(c) p(c|y) \\ \int_{x \in \mathcal{X}} y_\ell \delta_k(c) p(x) \mu(dx) &= \sum_{y \in \tilde{Y}} \tilde{p}(y) \sum_{c=1}^K y_\ell \delta_k(c) p(c|y) \\ \int_{x \in \mathcal{X}} y_\ell y_m \delta_k(c) p(x) \mu(dx) &= \sum_{y \in \tilde{Y}} \tilde{p}(y) \sum_{c=1}^K y_\ell y_m \\ &\quad \times \delta_k(c) p(c|y) \end{aligned}$$

Y and C not independent $\ell, m = 1, \dots, M$
 $k = 1, \dots, K. \quad (16)$

To find a feasible log-linear solution, we apply EM-IS as follows. First, start with an initial guess for the parameters, where we use the canonical parameterization $\lambda = (\lambda_0^k, \lambda_\ell^k, \lambda_{\ell, m}^k)$, $\ell, m = 1, \dots, M$ and $k = 1, \dots, K$, for the features. To execute

the **E step**, we then calculate the right-hand side feature expectations

$$\begin{aligned}\eta_0^{k,(j)} &= \frac{1}{T} \sum_{t=1}^T \sum_{c=1}^K \delta_k(c) \rho_t^{k,(j)} \\ \eta_\ell^{k,(j)} &= \frac{1}{T} \sum_{t=1}^T \sum_{c=1}^K y_\ell^t \delta_k(c) \rho_t^{k,(j)} \\ \eta_{\ell,m}^{k,(j)} &= \frac{1}{T} \sum_{t=1}^T \sum_{c=1}^K y_\ell^t y_m^t \delta_k(c) \rho_t^{k,(j)}\end{aligned}$$

where $\rho_t^{k,(j)} = p_{\lambda^{(j)}}(C = k|y^t) = p_{\lambda^{(j)}}(y^t|C = k)p_{\lambda^{(j)}}(C = k) / \sum_{c=1}^K p_{\lambda^{(j)}}(y^t|c)p_{\lambda^{(j)}}(c)$. To execute the **M step** we then formulate the simpler minimization problem with linear constraints, as in (9) and (10)

$$\begin{aligned}\min_{p(x)} D(p(x)||q(x)) &= D(p(c)||q(c)) \\ &\quad + D(p(y|c)||q(y|c)) \\ \text{subject to} \int_{x \in \mathcal{X}} \delta_k(c) p(x) \mu(dx) &= \eta_0^{k,(j)} \\ \int_{x \in \mathcal{X}} y_\ell \delta_k(c) p(x) \mu(dx) &= \eta_\ell^{k,(j)} \\ \int_{x \in \mathcal{X}} y_\ell y_m \delta_k(c) p(x) \mu(dx) &= \eta_{\ell,m}^{k,(j)}\end{aligned}\quad (17)$$

for $\ell, m = 1, \dots, M; k = 1, \dots, K$, where $x = (y, c)$. This problem can be solved analytically. In particular, when we choose the default model to be *improper uniform distribution* (that is, a distribution with infinite mass [18]), for (17) we can directly obtain the unique log-linear solution $p(x) = p(y, c)$, where $p(c) = (1/T) \sum_{t=1}^T \rho_t^{c,(j)}$ and $p(y|c) = N(y; \mu_c, \Sigma_c)$ with $\mu_c = \sum_{t=1}^T y^t \rho_t^{c,(j)} / \sum_{t=1}^T \rho_t^{c,(j)}$ and $\Sigma_c = \sum_{t=1}^T (y^t - \mu_c)(y^t - \mu_c)^\top \rho_t^{c,(j)} / \sum_{t=1}^T \rho_t^{c,(j)}$ for $c = 1, \dots, K$. We then set $p_{\lambda^{(j+1)}} = p$ and repeat.

Therefore, EM-IS produces a model that has the form of a Gaussian mixture. So in this case, LME is more general than Jaynes' ME principle, because it can postulate a multimodal distribution over the observed component Y , whereas standard ME is reduced to producing a unimodal Gaussian here.² Interestingly, the update formula we obtain for $p_{\lambda^{(j)}} \rightarrow p_{\lambda^{(j+1)}}$ is equivalent to the standard EM update for estimating Gaussian mixture distributions. In fact, we find that in many natural situations EM-IS recovers standard EM updates as a special case (although there are other situations where EM-IS yields new it-

²Radford Neal has observed that dropping the dependence constraint between Y and C allows the unimodal ME Gaussian solution with a uniform mixing distribution to be a feasible global solution in this specific case. However, this model is ruled out by the dependence requirement.

erative update procedures that converge faster than standard parameter estimation formulas). Nevertheless, the final estimation principle we propose, which must select from among feasible solutions, is different from standard MLE.

To demonstrate the difference of regularized RLME with MAP estimate, we use conjugate prior for the Gaussian mixture model. As in [12], we take the Dirichlet density to model the prior knowledge about the mixture weights

$$p(w_1, \dots, w_K | \nu_1, \dots, \nu_K) \propto \prod_{k=1}^K w_k^{\nu_k - 1}. \quad (18)$$

Then for the mean and covariance of each Gaussian component, we use the joint conjugate prior density, a normal Wishart density of the form

$$\begin{aligned}p(\mu, \Sigma | \tau, m, \alpha, V) &\propto |\Sigma|^{\frac{(\alpha-n)}{2}} \exp\left(-\frac{\tau}{2}(\mu - m)^T \Sigma (\mu - m)\right) \\ &\quad \times \exp\left(-\frac{1}{2} \text{tr}(V \Sigma)\right)\end{aligned}\quad (19)$$

where (τ, m, α, V) are the prior density parameters such that $\alpha > n - 1, \tau > 0, \mu$ is a n -dimensional vector and V is $n \times n$ positive-definite matrix. Thus, the joint prior density is the product of the prior density defined in (18) and (19).

The joint prior density for the *natural parameterization* can be derived correspondingly [3] and its log form is the convex conjugate cost function U^* . The corresponding penalty function U can be derived by using the property of *Fenchel biconjugation* [5], that is the conjugate of the conjugate of a convex function is the original convex function, $U = U^{**}$. For example, if we chose $U^*(\lambda) = \|\lambda\|_1 = \sum_{i=1}^N |\lambda_i|$, the Laplacian prior on λ , then $U(a) = \begin{cases} 0 & \|a\|_\infty = \max_{i=1}^N |a_i| \leq 1 \\ \infty & \text{otherwise} \end{cases}$ which corresponds to absolute inequality constraints. However knowing the explicit form of U is not necessary, since when we calculate the regularized entropy, we use the value of auxiliary function Q for each feasible log-linear solution (Corollary 1).

The EM re-estimation formulas can be derived as follows:

$$w_k = \frac{(\nu_k - 1) + \sum_{t=1}^T \rho_t^k}{\sum_{k=1}^K \left((\nu_k - 1) + \sum_{t=1}^T \rho_t^k \right)} \quad (20)$$

$$\mu_k = \frac{\tau \mu_k + \sum_{t=1}^T \rho_t^k y_t}{\tau_k + \sum_{t=1}^T \rho_t^k} \quad (21)$$

(see (22) at the bottom of the page). Once we obtain the estimates of w_k, μ_k, Σ_k , for $k = 1, \dots, K$, we can then transform them into the natural parameterization and calculate the regularized entropy and penalized likelihood. We then choose the highest regularized entropy estimate as the final regularized RLME estimate and highest penalized likelihood estimate as the final MAP estimate.

$$\Sigma_k = \frac{\mu_k + \sum_{t=1}^T \rho_t^k (y_t - \mu_k)(y_t - \mu_k)' + \tau_t (m_k - \mu_k)(m_k - \mu_k)'}{(\alpha_k - n) + \sum_{t=1}^T \rho_t^k} \quad (22)$$

To compare the relative benefits of estimating Gaussian mixture models using RLME versus MAP, we conducted experiments on synthetic and real data.

Experiments on Synthetic Data: As a first case study, we considered a simple three component mixture model where the mixing component C is unobserved but a two dimensional vector $Y \in \mathbb{R}^2$ is observed. Thus, the features we match in the data are of the same form as in Section V. Given sample data $\tilde{\mathcal{Y}} = (y_1, \dots, y_T)$ the idea is to infer a log-linear model $p(x) = p(y, c)$ such that $c \in \{1, 2, 3\}$.

We are interested in determining which method yields better estimates of various underlying models p^* used to generate the data. We measure the quality of an estimate p_λ by calculating the *cross entropy* from the correct marginal distribution $p^*(y)$ to the estimated marginal distribution $p_\lambda(y)$ on the *observed* data component Y

$$D(p^*(y) \| p_\lambda(y)) = \int_{y \in \mathcal{Y}} p^*(y) \log \frac{p^*(y)}{p_\lambda(y)} \mu(dy).$$

The goal is to minimize the cross entropy between the marginal distribution of the estimated model p_λ and the correct marginal p^* . A cross entropy of zero is obtained only when $p_\lambda(y)$ matches $p^*(y)$.

We consider a variety of experiments with different models and different sample sizes to test the robustness of both RLME and MAP to sparse training data, high-variance data, and deviations from log-linearity in the underlying model. In particular, we used the following experimental design.

- 1) fix a generative model $p^*(x) = p^*(y, c)$;
- 2) generate a sample of observed data $\tilde{\mathcal{Y}} = (y_1, \dots, y_T)$ according to $p^*(y)$;
- 3) run R-EM-IS to generate multiple feasible solutions by restarting from 300 random initial vectors λ . We generated initial vectors λ by generating mixture weights θ_c from a uniform prior, and independently generating each component of the mean vectors μ_c and covariance matrices σ_c^2 by choosing numbers uniformly from $\{-4, -2, 0, 2, 4\}$ and $\{0.5, 2.5\}$;
- 4) calculate regularized entropy and posterior probability for each candidate;
- 5) select the maximum regularized entropy candidate p_{RLME} as the RLME estimate, and the maximum posterior probability candidate p_{MAP} in the interior of the parameter space as the MAP estimate;
- 6) calculate the cross entropy from $p^*(y)$ to the marginals $p_{\text{RLME}}(y)$ and $p_{\text{MAP}}(y)$, respectively;
- 7) repeat Steps 2 to 6 500 times and compute the average of the respective cross entropies. That is, average the cross entropy over 500 repeated trials for each sample size and each method, in each experiment;
- 8) repeat Steps 2 to 7 for different sample sizes T ; and
- 9) repeat Steps 1 to 8 for different models $p^*(x)$.

Scenario 1: In the first experiment, we generated the data according to a three component Gaussian mixture model that has the form expected by the estimators. Specifically, we used a uniform mixture distribution $\theta_c = (1/3)$ for $c = 1, 2, 3$, where

the component Gaussians were specified by the mean vectors $[0 \ -3]^\top$, $[0 \ 0]^\top$, $[0 \ 3]^\top$ and covariance matrices $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, respectively. Fig. 1 is the scatter plot of this scenario.

Figs. 2 and 3 first show that the average posterior probabilities and average regularized entropies of the models produced by RLME and MAP, respectively, behave as expected. MAP clearly achieves higher posterior probability than RLME, however RLME clearly produces models that have significantly higher regularized entropy than MAP. The interesting outcome is that the two estimation strategies obtain significantly different cross entropies. Fig. 4 reports the average cross entropy obtained by MAP and RLME as a function of sample size, and shows the somewhat surprising result that RLME achieves substantially lower cross entropy than MAP. RLME's advantage is especially pronounced at small sample sizes, and persists even to sample sizes as large as 10 000 (Fig. 4).

Although one might have expected an advantage for RLME because of a "regularization" effect, this does not completely explain RLME's superior performance at large sample sizes. (We return to a more thorough discussion of RLME's regularization properties in Section VI.

This first experiment considered a favorable scenario where the underlying generative model has the same form as the distributional assumptions made by the estimators. We next consider situations where these assumptions are violated.

Scenario 2: In our second experiment, we used a generative model that was a mixture of *five* Gaussian distributions over \mathbb{R}^2 . Specifically, we generated data by sampling from a uniform distribution over mixture components $\theta_c = (1/5)$ for $c = 1, \dots, 5$, and then generated the observed data $Y \in \mathbb{R}^2$ by sampling from the corresponding Gaussian distribution, where these distributions had means $[2 \ 0]^\top$, $[0 \ 0]^\top$, $[0 \ 2]^\top$, $[-2 \ 0]^\top$, $[0 \ -2]^\top$ and covariances $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, respectively. Fig. 5 is the scatter plot of this scenario.

The RLME and MAP estimators still only inferred three component mixtures in this case and, hence, were each making an incorrect assumption about the underlying model.

Fig. 6 shows that RLME still obtained a significantly lower cross entropy than MAP at small sample sizes, but lost its advantage at larger sample sizes. At a crossover point of $T = 1000$ data points, MAP began to produce slightly better estimates than RLME, but only marginally so. Overall, RLME still appears to be a safer estimator for this problem, but it is not uniformly dominant.

Scenario 3: Our third experiment attempted to test how robust the estimators were to high-variance data generated by a heavy tailed distribution. This experiment yielded our most dramatic results. We generated data according to a three component mixture (which was correctly assumed by the estimators) but then used a Laplacian distribution instead of a Gaussian distribution to generate the Y observations. This model generated data that was much more variable than data generated by a Gaussian mixture and challenged the estimators significantly. The specific parameters we used in this experiment were $\theta_c = (1/3)$

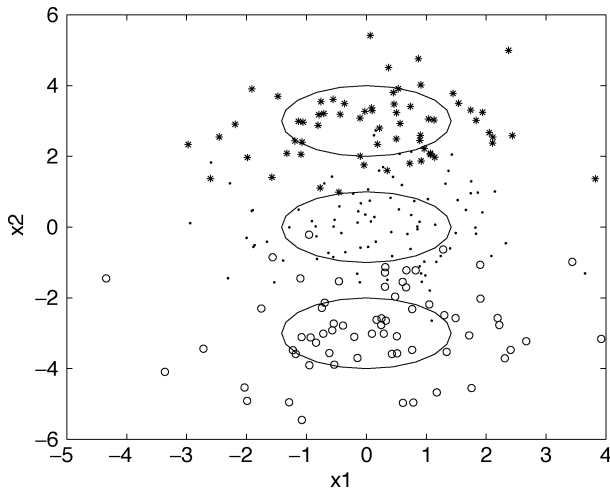


Fig. 1. Scatter plot of 200 training data in scenario 1.

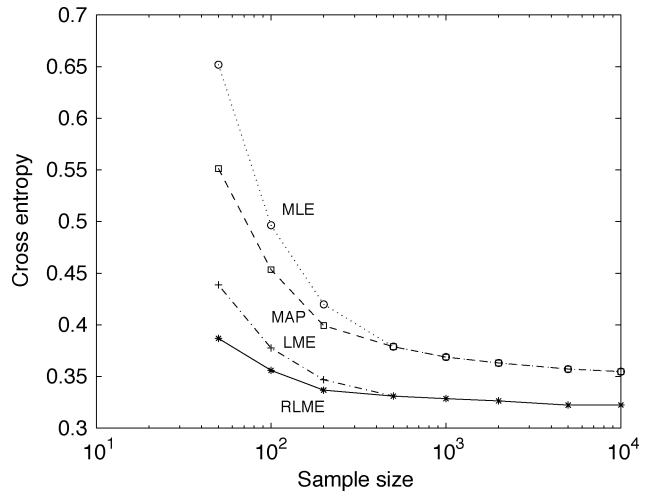


Fig. 4. Average cross entropy between the true distribution and the MAP estimates versus the RLME estimates in experiment 1.

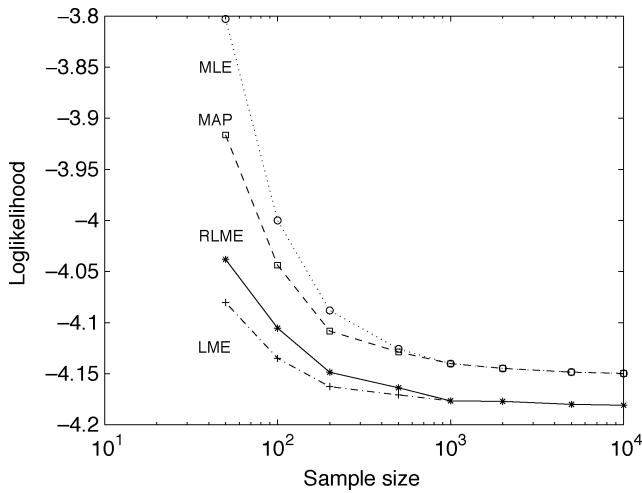


Fig. 2. Average posterior probability of the MAP estimates versus the RLME estimates in experiment 1.

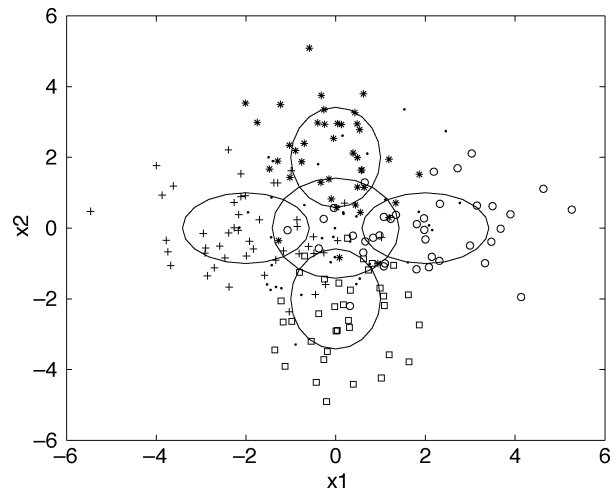


Fig. 5. Scatter plot of 200 training data in scenario 2.

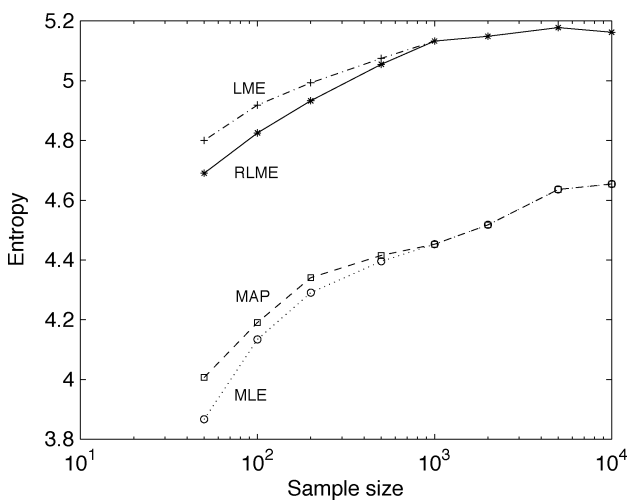


Fig. 3. Average regularized entropy of the MAP estimates versus the RLME estimates in experiment 1.

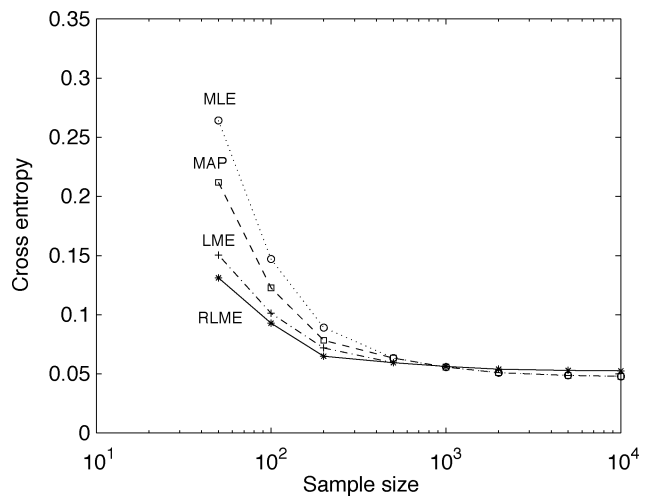


Fig. 6. Average cross entropy between the true distribution and the MAP estimates versus the RLME estimates in experiment 2.

for $c = 1, 2, 3$, and means $[2 \ 0]^T$, $[0 \ 0]^T$, $[0 \ 2]^T$, and “covari-

ances” $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ for the Laplacians. Fig. 7 is the scatter plot of this scenario.

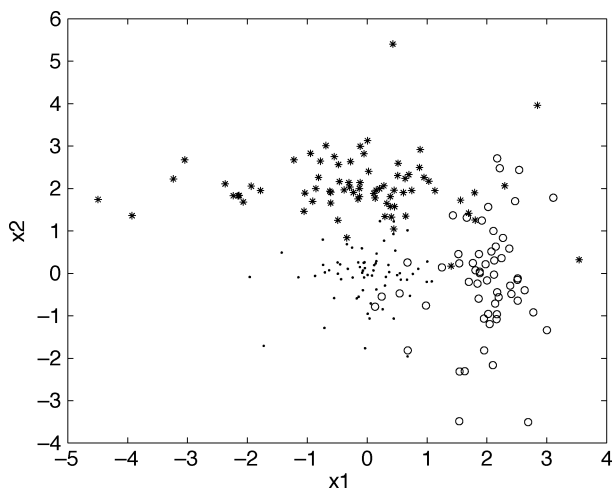


Fig. 7. Scatter plot of 200 training data in scenario 3.

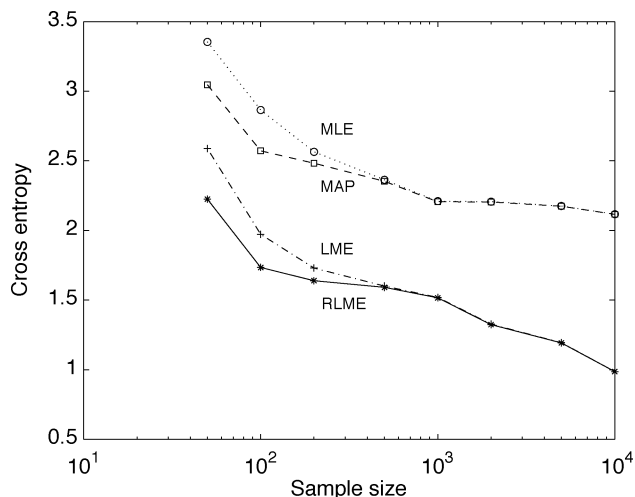


Fig. 8. Average cross entropy between the true distribution and MAP estimates versus the regularized RLME estimates in Gaussian mixture experiment 3.

Fig. 8 shows that RLME produces significantly better estimates than MAP in this case, and even improved its advantage at larger sample sizes. Clearly, MAP is not a stable estimator when subjected to heavy tailed data when this is not expected. RLME proves to be far more robust in such circumstances and clearly dominates MAP.

Fig. 8 shows that RLME still produces significantly better estimates than MAP in this case. Comparing with Fig. 8, we notice that when the data is small, the regularization term causes the estimates to be closer to the true distribution, however when the sample size gets large, this effect diminishes.

Scenario 4: However, there are other situations where MAP appears to be a slightly better estimator than RLME when sufficient data is available. Fig. 10 shows the results of subjecting the estimators to data generated from a three component Gaussian mixture, $\theta_c = (1/3)$, $c = 1, 2, 3$, with means $[2\ 0]^T$, $[0\ 0]^T$, $[0\ 2]^T$ and covariances $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, respectively. Fig. 9 is the scatter plot of this scenario.

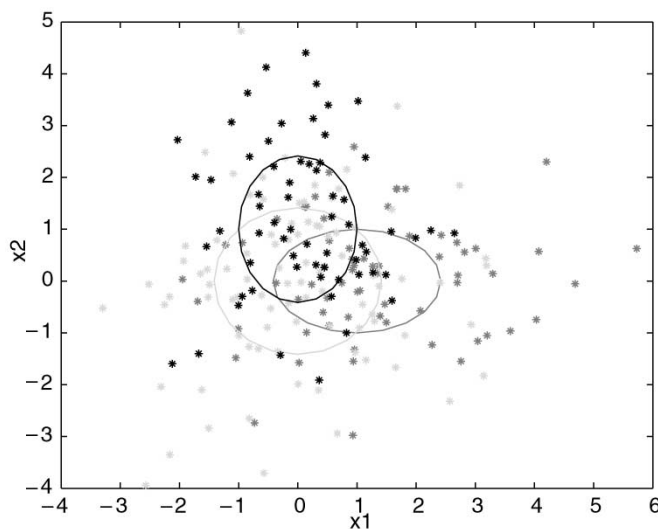


Fig. 9. Scatter plot of 200 training data in scenario 4.

In this case, RLME still retains a sizable advantage at small sample sizes, but after a sample size of $T = 500$, MAP begins to demonstrate a persistent advantage (Fig. 10).

Overall, these results suggest that maximum a posterior probability estimation (MAP) is effective at large sample sizes, as long as the presumed model is close to the underlying data source. If there is a mismatch between the assumption and reality however, or if there is limited training data, then RLME appears to offer a significantly safer and more effective alternative. Of course, these results are far from definitive, and further experimental and theoretical analysis is required to give completely authoritative answers.

Experiment on Iris Data: To further confirm our observation, we consider a classification problem on the well known set of *Iris* data as originally collected by Anderson and first analyzed by [11]. The data consists of measurements of the length and width of both sepals and petals of 50 plants for each of three types of *Iris* species *setosa*, *versicolor*, and *virginica*. In our experiments, we intentionally ignore the types of species, and use the data for unsupervised learning and clustering of multivariate Gaussian mixture models. That is, we train the model

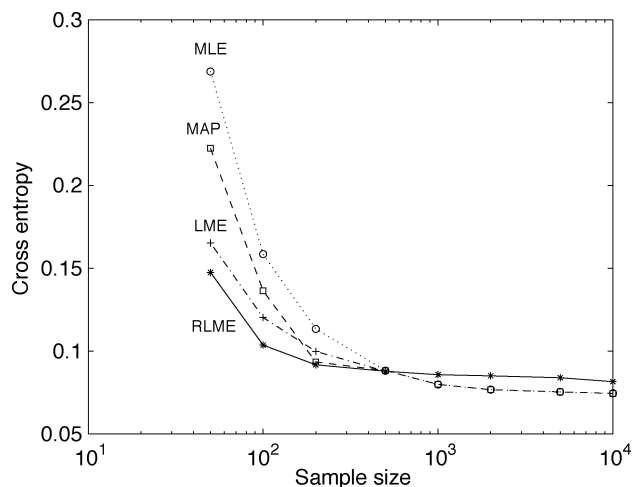


Fig. 10. Average cross entropy between the true distribution and the MAP estimates versus the RLME estimates in experiment 4.

by the LME/RLME principle, and we use the Bayes' decision

rule which selects the class c having the highest posterior probability $p(c|x)$ for clustering. Among 150 samples, we uniformly choose 100 samples as training data, and the rest 50 samples as test data. Again, we start from 300 initial points, where each initial point is chosen as the following: first, we calculate the sample mean and covariance matrix of the training data, then perturb the sample mean using the sample variance as the initial mean, and take sample covariance as the covariance for each class. To measure the performance of the estimates, we use the empirical test set likelihood and clustering error rate. We repeat this procedure 100 times. Table I shows the averaged results. Two observations can be drawn from these results: 1) the test data is more likely under the RLME estimates than MAP and also that the clustering error rate is cut in half; a similar relationship holds true for the LME versus MLE comparison; 2) as expected, the RLME estimates give better results in likelihood and clustering error rate on test data than MAP estimates; this is also true for the LME and MLE comparison.

B. Dirichlet Mixtures

Of course, the RLME principle is much more general than merely being applicable to estimating Gaussian mixture models. It can easily be applied to any form of parametric mixture model (and many other models beyond these, see Section VI). Here we present an alternative application of RLME to estimating a mixture of Dirichlet sources.

Assume the observed data has the form of an M dimensional probability vector $y = (y_1, \dots, y_M)$ such that $0 \leq y_\ell \leq 1$ for $\ell = 1, \dots, M$ and $\sum_{\ell=1}^M y_\ell = 1$. That is, the observed variable is a random vector $Y = (Y_1, \dots, Y_M) \in [0, 1]^M$, which happens to be normalized. There is also an underlying class variable $C \in \{1, \dots, K\}$ that is unobserved. Let $X = (Y, C)$. Given an observed sequence of TM -dimensional probability vectors $\tilde{Y} = (y^1, \dots, y^T)$, where $y^t = (y_1^t, \dots, y_M^t)$ for $t = 1, \dots, T$, we attempt to infer a latent maximum entropy model that matches expectations on the features $f_0^k(x) = \delta_k(c)$ and $f_\ell^k(x) = (-\log y_\ell)\delta_k(c)$ for $\ell = 1, \dots, M$ and $k = 1, \dots, K$, where $x = (y, c)$. By setting the penalty function U to be constant, we start with the initial LME formulation as follows:

$$\begin{aligned} & \min_{p(x)} D(p(x)||q(x)) \\ & = D(p(c)||q(c)) + D(p(y|c)||q(y|c)) \\ \text{subject to } & \int_{x \in \mathcal{X}} \delta_k(c)p(x)\mu(dx) \\ & = \sum_{y \in \tilde{Y}} \tilde{p}(y) \sum_c \delta_k(c)p(c|y)\mu(dx) \\ & \int_{x \in \mathcal{X}} (-\log y_\ell)\delta_k(c)p(x)\mu(dx) \\ & = \sum_{y \in \tilde{Y}} \tilde{p}(y) \sum_c (-\log y_\ell)\delta_k(c)p(c|y)\mu(dx) \\ & \ell = 1, \dots, M, k = 1, \dots, K \\ & \text{and } Y, C \text{ not independent.} \end{aligned} \quad (23)$$

TABLE I
COMPARISON OF LME, MLE, RLME, AND MAP ON THE *IRIS* TEST DATA SET

	LOG-LIKELIHOOD	ERROR RATE
LME	5.5889	0.1220
MLE	5.3770	0.2446
RLME	5.6173	0.0935
MAP	5.4285	0.1691

Here $\tilde{p}(y) = (1/T)$ and $\delta_k(c)$ denotes the indicator function of the event $c = k$. Due to the nonlinear mapping caused by $p(c|y)$ there is no closed form solution to (23). However, as for Gaussian mixtures, we can apply EM-IS to obtain feasible log-linear models for this problem. To perform the **E step**, one can calculate the feature expectations

$$\begin{aligned} \eta_0^{k,(j)} &= \frac{1}{T} \sum_{t=1}^T \sum_{c=1}^K \delta_k(c) \rho_t^{k,(j)} \\ \eta_\ell^{k,(j)} &= \frac{1}{T} \sum_{t=1}^T \sum_{c=1}^K (-\log y_\ell^t) \delta_k(c) \rho_t^{k,(j)} \end{aligned} \quad (24)$$

for $\ell = 1, \dots, M$, $k = 1, \dots, K$, where $\rho_t^{k,(j)} = p_{\lambda^{(j)}}(C = k|y^t) = p_{\lambda^{(j)}}(y^t|C = k)p_{\lambda^{(j)}}(C = k) / \sum_{c=1}^K p_{\lambda^{(j)}}(y^t|c)p_{\lambda^{(j)}}(c)$. Note that these expectations can be calculated efficiently, as in Section V.

To perform the **M step** we then formulate the simpler generalized maximum entropy problem with linear constraints, as in (9) and (10)

$$\begin{aligned} & \min_{p(x)} D(p(x)||q(x)) = D(p(c)||q(c)) + D(p(y|c)||q(y|c)) \\ \text{subject to } & \int_{x \in \mathcal{X}} \delta_k(c)p(x)\mu(dx) = \eta_0^{k,(j)} \\ & \int_{x \in \mathcal{X}} (-\log y_\ell)\delta_k(c)p(x)\mu(dx) = \eta_\ell^{k,(j)} \end{aligned} \quad (25)$$

for $\ell = 1, \dots, M$ and $k = 1, \dots, K$. When we choose the default model to be uniform over the latent class variable and the observation on intervals $[0, 1]^M$, for this problem we can obtain a log-linear solution of the form $p(x) = p(y, c)$ where $p(c) = (1/T) \sum_{t=1}^T \rho_t^k$ and the class conditional model $p(y|c)$ is a Dirichlet distribution with parameters $\alpha_\ell^c = 1 - \lambda_\ell^c$; that is $p(y|c) = \Gamma(\sum_{\ell=1}^M \alpha_\ell^c) (\prod_{\ell=1}^M \Gamma(\alpha_\ell^c))^{-1} \prod_{\ell=1}^M y_\ell^{\alpha_\ell^c - 1}$. However, we still need to solve for the parameters α_ℓ^c . By plugging in the form of the Dirichlet distribution, the feature expectations (25) will have an explicit formula, and the constraints on the parameters α_ℓ^c can then be written

$$-\Psi(\alpha_\ell^{c,(j)}) + \Psi\left(\sum_{m=1}^M \alpha_m^{c,(j)}\right) = \eta_\ell^{k,(j)}$$

for $\ell = 1, \dots, M$ and $k = 1, \dots, K$, where Ψ is the digamma function. The solution can be obtained by iterating the fixed-point equations

$$\Psi\left(\alpha_\ell^{c, (j+\frac{s}{S})}\right) = \Psi\left(\sum_{m=1}^M \alpha_m^{c, (j+\frac{s-1}{S})}\right) - \eta_\ell^{k, (j)}$$

for $\ell = 1, \dots, M$ and $k = 1, \dots, K$. This iteration corresponds to a well known technique for locally monotonic maximizing the likelihood of a Dirichlet mixture [21]. Thus, EM-IS recovers a classical likelihood maximization algorithm as a special case. However, as before, this only yields feasible solutions, from which we have to select a final estimate.

Now consider RLME for (23) and we use the quadratic penalty $U(a) = \sum_{i=1}^N (1/2)\sigma_i^2 a_i^2$. By plugging in the form of the Dirichlet distribution, the feature expectations (14) will have an explicit formula, and the constraints on the parameters γ_ℓ^c can then be written

$$-\Psi\left(\alpha_\ell^{c, (j)} - \gamma_\ell^{c, (j+\frac{s}{S})}\right) + \Psi\left(\sum_{m=1}^M \alpha_m^{c, (j)} - \gamma_\ell^{c, (j+\frac{s}{S})}\right) + \frac{1 - \alpha_\ell^{c, (j)} + \gamma_\ell^{c, (j+\frac{s}{S})}}{\sigma_i^2} = \eta_\ell^{k, (j)}$$

for $\ell = 1, \dots, M$ and $k = 1, \dots, K$, and the solution can be obtained by iterating the fixed-point equations as above or Newton-Raphson procedure.

Dirichlet Mixture Experiment: To compare model selection based on the RLME versus MAP principles for this problem, we conducted an experiment on a mixture of Dirichlet sources. In this experiment, we generate the data according to a three component Dirichlet mixture, with mixing weights $\theta_c = 1/6, 1/2, 1/3$ and component Dirichlet distributions specified by the α parameters $[1 \ 2]^\top$, $[3 \ 1]^\top$, and $[5 \ 2]^\top$, respectively. The inferred model is three component Dirichlet mixture. The initial mixture weights were generated from a uniform prior, and each α was generated by choosing numbers uniformly from $\{0.1, 0.5, 1, 2.5, 5\}$. Fig. 11 shows the cross entropy results of RLME and MAP averaged over 10 repeated trials for each fixed training sample size. The outcome in this case shows a significant advantage for RLME.

C. Poisson Mixtures

Our last example considers a discrete distribution, the Poisson mixture model, which has received considerable attention recently for the analysis of data in the form of counts [20].

Assume that the observed data y takes on values in $0, 1, 2, \dots, \infty$, and also that there is an underlying class variable $C \in \{1, \dots, K\}$ which is unobserved. Let $X = (Y, C)$. Given an observed sequence of TM -dimensional probability vectors $\tilde{Y} = (y^1, \dots, y^T)$, where $y^t = (y_1^t, \dots, y_M^t)$ for $t = 1, \dots, T$, we attempt to infer a latent maximum entropy model that matches expectations on the features $f_0^k(x) = \delta_k(c)$ and $f^k(x) = y\delta_k(c)$ for $\ell = 1, \dots, M$ and $k = 1, \dots, K$, where $x = (y, c)$. In this case, we only consider the *generalized* LME principle by setting the penalty function U to be constant,

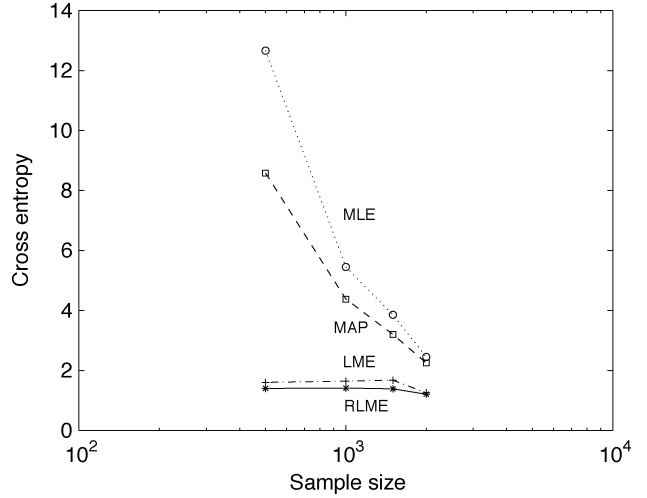


Fig. 11. Average cross entropy between the true distribution and MAP versus RLME estimates in the Dirichlet mixture experiment.

and formulate the problem as finding a joint distribution according to the principle

$$\min_{p(x)} D(p(x)||q(x)) = D(p(c)||q(c)) + D(p(y|c)||q(y|c))$$

$$\text{subject to } \int_{x \in \mathcal{X}} \delta_k(c)p(x)\mu(dx) = \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \times \sum_c \delta_k(c)p(c|y)\mu(dx)$$

$$\int_{x \in \mathcal{X}} y\delta_k(c)p(x)\mu(dx) = \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \times \sum_c y\delta_k(c)p(c|y)\mu(dx)$$

$$k = 1, \dots, K \quad \text{and} \quad Y, C \text{ not independent}$$

where $q(C, Y)$ is a default joint model with $q(c)$ uniform and $q(Y|C)$ defined to be a Poisson distribution with a default parameter λ_{def}^c for each c^3 ; and $\tilde{p}(y) = (1/T)$ and $\delta_k(c)$ denotes the indicator function of the event $c = k$. Due to the nonlinear mapping caused by $p(c|y)$ there is no closed form solution to (26). However, again as for Gaussian or Dirichlet mixtures, we can apply EM-IS to obtain feasible log-linear models for this problem. To perform the *E step*, one can calculate the feature expectations

$$\eta_0^{k, (j)} = \frac{1}{T} \sum_{t=1}^T \sum_{c=1}^K \delta_k(c) \rho_t^{k, (j)}$$

$$\eta^{k, (j)} = \frac{1}{T} \sum_{t=1}^T \sum_{c=1}^K y \delta_k(c) \rho_t^{k, (j)} \quad (26)$$

$k = 1, \dots, K$, where $\rho_t^{k, (j)} = p_{\lambda^{(j)}}(C = k|y^t) = p_{\lambda^{(j)}}(y^t|C = k)p_{\lambda^{(j)}}(C = k) / \sum_{c=1}^K p_{\lambda^{(j)}}(y^t|c)p_{\lambda^{(j)}}(c)$.

³We can use an *improper* uniform distribution over $0, 1, 2, \dots, \infty$, or any Poisson distribution as a default conditional model.

Note that these expectations can be calculated efficiently, as in Section V.

To perform the *M step* we then formulate the simpler generalized maximum entropy problem with linear constraints, as in (9) and (10)

$$\begin{aligned} \min_{p(x)} D(p(x)||q(x)) &= D(p(C)||q(C)) \\ &\quad + D(p(Y|C)||q(Y|C)) \\ \text{subject to } \int_{x \in \mathcal{X}} \delta_k(c)p(x)\mu(dx) &= \eta_0^{k,(j)} \\ \int_{x \in \mathcal{X}} y\delta_k(cp(x))\mu(dx) &= \eta^{k,(j)} \end{aligned}$$

for $k = 1, \dots, K$. For this problem we can obtain a log-linear solution of the form $p(x) = p(y, c)$ where $p(c) = (1/T) \sum_{t=1}^T \rho_t^k$ and the class conditional model $p(y|c)$ is a Poisson distribution with parameters $\lambda^c = \log \eta^{k,(j)}$; that is $p(y|c) = \exp(-\eta^{k,(j)})((\eta^{k,(j)})^y / y!)$. Interestingly, this conditional distribution is independent of the default conditional distribution given by λ_{def}^c , so the λ_{def}^c parameters turn out to be irrelevant.

Poisson Mixture Experiment: To compare model selection based on the LME versus MLE principles for this problem, we conducted an experiment on a mixture of Poisson sources. In this experiment, we generate the data according to a three component Poisson mixture, with mixing weights $\theta_c = .2, .3, .5$ and component Poissons specified by the λ parameters 10, 2, and 5, respectively. The inferred model is three component Poisson mixture. The initial mixture weights were generated from a uniform prior, and each λ was generated by choosing numbers uniformly from $\{0, 0.05, 0.1, 5, 10\}$. Fig. 12 shows the cross entropy results of LME and MLE averaged over 25 repeated trials for each fixed training sample size. The outcome in this case shows a significant advantage for LME.

VI. CONCLUSION

A few comments are in order. It appears that LME adds more than just a fixed regularization effect to MLE. In fact, as we have demonstrated in this paper, one can add a regularization term to the LME principle (to obtain RLME) in the same way one can add a regularization term to the MLE principle (to obtain MAP). LME behaves more like an *adaptive* rather than fixed regularizer [23], because we see no real under-fitting from LME/RLME on large data samples, even though LME chooses far “smoother” models than MAP at smaller sample sizes. In fact, LME/RLME can demonstrate a stronger regularization effect than any standard penalization method: In the well known case where EM-IS converges to a degenerate solution (i.e., such that the determinant of the covariance matrix of Gaussian goes to zero, or the parameters of Dirichlet or Poisson go to zero) no finite penalty can counteract the resulting unbounded likelihood. However, the LME/RLME principles can automatically filter out degenerate models, because such models have a differential entropy of $-\infty$ and any nondegenerate model will be preferred. Eliminating degenerate models by the LME principle solves one of the main practical problems with mixture estimation.

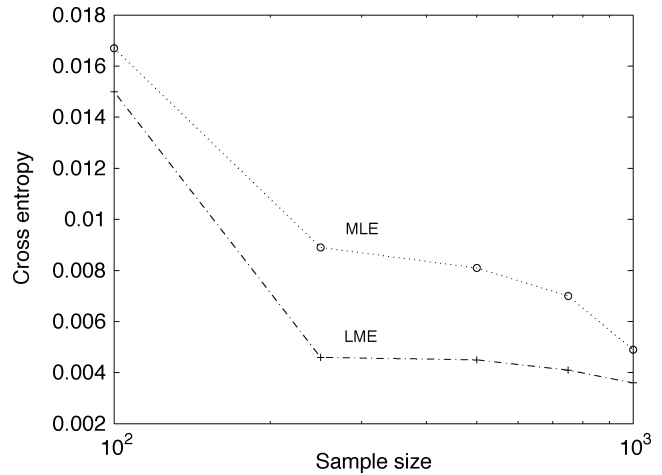


Fig. 12. Average cross entropy between the true distribution and MLE versus LME estimates in the poisson mixture experiment.

Another observation is that all of our experiments show that MAP and RLME reduce the cross entropy error when sample size is increased. However, we have not yet proved that the RLME principle is statistically consistent, that is, that it is guaranteed to converge to zero cross entropy in the limit of large samples when the underlying model has a log-linear form in the same features considered by the estimator. We are actually interested in a stronger form of consistency that requires the estimator to converge to the best *representable* log-linear model (i.e., the one with minimum cross entropy error) for any underlying distribution, even if the minimum achievable cross entropy is nonzero. Previous work [2] has studied the approximation error by sequences of exponential families in a rather simple one dimensional complete data case. Determining the statistical consistency of LME, in either sense, remains an important topic for future research.

In this paper, the number of mixture components is pre-defined. The authors are currently investigating information theoretic techniques to automatically determine the optimal number of mixture components. Also, the R-ME-EM-IS procedure, which uses random restarts to produce different feasible solutions, is computationally expensive. It is worthwhile to develop an analogous deterministic annealing algorithm [24] for finding feasible regularized maximum entropy log-linear models for RLME.

The purpose of using a generalized Kullback–Leibler (KL) divergence instead of Shannon’s entropy $H(p)$ that is we want to obtain a distribution which deviates from an a priori distribution q as little as possible, subject to the constraints. It turns out that some models, like Gaussian and Dirichlet mixtures, are sensitive to this choice of default distribution, whereas other models are not, such as the Poisson mixture.

In other work [25], we observe that the LME principle can be applied to other statistical models beyond mixtures, such as hidden Markov models [15] and Boltzmann machines [1]. We have begun to investigate these models, and in each case, have identified new parameter optimization methods based on EM-IS, and new statistical estimation principles based on ME-EM-IS.

A final remark is that the log-linear models given by the R-EM-IS algorithm for RLME is applicable to *undirected* graphical models with canonical parameters. It is well known that a decomposable graphical model can be represented by either an undirected graphical model or a directed graphical model [17]. If a decomposable graphical model is represented by a directed graphical model, then the parameters are those we are familiar with instead of the natural parameters, and they can be estimated by the common EM algorithm. This leads to a natural question: if we run the common EM algorithm and get local maxima, how can one select the model that has regularized maximum entropy? In fact, it can be shown that for each local maxima, the entropy value under undirected graphical model representation is the same as the negative value of auxiliary function. This is because there is an underlying “invariance of parameterization” property to represent a probabilistic model.

REFERENCES

- [1] D. Ackley, G. Hinton, and T. Sejnowski, “A learning algorithm for Boltzmann machines,” *Cogn. Sci.*, vol. 9, pp. 147–169, 1985.
- [2] A. Barron and C. Sheu, “Approximation of density functions by sequences of exponential families,” *Ann. Statist.*, vol. 19, pp. 1347–1369, 1991.
- [3] J. Bernardo and A. Smith, *Bayesian Theory*. New York: Wiley, 2000.
- [4] D. Bertsekas, *Nonlinear Programming*. Nashua, NH: Athena Scientific, 1999.
- [5] J. Borwein and A. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*. New York: Springer-Verlag, 2000.
- [6] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [7] I. Csiszar, “Maxent, mathematics, and information theory,” in *Maximum Entropy and Bayesian Methods*, K. Hanson and R. Silver, Eds. Norwell, MA: Kluwer, 1996, pp. 35–50.
- [8] J. Darroch and D. Ratchliff, “Generalized iterative scaling for log-linear models,” *Ann. Math. Statist.*, vol. 43, no. 5, pp. 1470–1480, 1972.
- [9] S. Della Pietra, V. Della Pietra, and J. Lafferty, “Inducing features of random fields,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 4, pp. 380–393, Apr. 1997.
- [10] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood estimation from incomplete data via the EM algorithm,” *J. Roy. Statist. Soc. Series B*, vol. 39, pp. 1–38, 1977.
- [11] R. Fisher, “The use of multiple measurements in taxonomic problems,” *Ann. Eugenics*, vol. 7, no. II, pp. 179–188, 1936.
- [12] J. Gauvain and C. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.
- [13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning—Data Mining, Inference and Prediction*. New York: Springer-Verlag, 2001.
- [14] E. Jaynes, *Papers on Probability, Statistics, and Statistical Physics*, R. Rosenkrantz and R. Rosenkrantz, Eds. Amsterdam, The Netherlands: Reidel, 1983.
- [15] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: probabilistic models for segmenting and labeling sequence data,” in *Proc. Int. Conf. Machine Learning*, Williamstown, MA, 2001, pp. 282–289.
- [16] S. Lauritzen, “The EM-algorithm for graphical association models with missing data,” *Computat. Statist. Data Anal.*, vol. 1, pp. 191–201, 1995.
- [17] ———, *Graphical Models*. Oxford, U.K.: Clarendon Press, 1996.
- [18] E. Lehmann and G. Casella, *Theory of Point Estimation*. New York: Springer-Verlag, 1998.
- [19] D. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.
- [20] G. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.
- [21] T. Minka, Estimating a Dirichlet Distribution, 2000.
- [22] S. Riezler, “Probabilistic Constraint Logic Programming,” Ph.D. dissertation, Univ. Stuttgart, Germany, 1999.
- [23] A. Tikhonov, *Ill-Posed Problems in Natural Sciences*. Philadelphia, PA: Coronet Books, 1992.
- [24] N. Ueda and R. Nakano, “Deterministic annealing EM algorithm,” *Neural Netw.*, vol. 11, pp. 272–282, 1998.
- [25] S. Wang, D. Schuurmans, and Y. Zhao, The Latent Maximum Entropy Principle, 2003.
- [26] C. Wu, “On the convergence properties of the EM algorithm,” *Ann. Statist.*, vol. 11, pp. 95–103, 1983.



Shaojun Wang received the B.E. and M.E. degrees in electric power and energy systems in electrical engineering from Tsinghua University, Beijing, China, in 1988 and 1992, respectively, the M.S. degree in mathematics and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign, in 1998 and 2000, respectively.

He worked as a Postdoctoral Fellow in the Center for Automated Learning and Discovery, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, School of Computer Science, University of Waterloo, Waterloo, ON, Canada, and Department of Statistics, University of Toronto, Toronto, ON, Canada, from 2000 to 2003. His primary research interest is statistical machine learning. He is especially interested in developing methods to solve machine learning problems that arise in human-machine interaction such as text and natural language processing, information retrieval and data mining, speech and biological signal processing. He is also interested in the underlying statistical learning theory.



Dale Schuurmans received the Ph.D. degree in computer science from the University of Toronto, Toronto, ON, Canada and the M.Sc. and B.Sc. degrees in computing science and mathematics from the University of Alberta, Canada.

He is currently an Associate Professor of computing science and the Canada Research Chair in machine learning at the University of Alberta. He has previously been an Associate and Assistant Professor of computer science at the University of Waterloo, a Postdoctoral Fellow at the University of Pennsylvania, a Researcher at the NEC Research Institute, and a Research Associate at the National Research Council Canada. His research interests include machine learning, optimization, and search.



Fuchun Peng received the Ph.D. degree in computer science from the University of Waterloo, Waterloo, ON, Canada, in 2003.

He is a currently a Senior Postdoctoral Research Associate at the Computer Science Department, University of Massachusetts, Amherst. His research interests focus on natural language processing, information extraction, machine learning, and data mining.



Yunxin Zhao (S'86–M'88–SM'94) received the Ph.D. degree from University of Washington, Seattle, in 1988.

She was Senior Research Staff and Project Leader of Speech Technology Laboratory, Panasonic Technologies, Inc., from 1988 to 1994. She was an Assistant Professor for the Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, from 1994 to 1998. She is currently a Professor in the Department of Computer Engineering and Computer Science, University of

Missouri, Columbia. Her research interests are in spoken language processing, automatic speech recognition, multimedia interface, multimodal human–computer interaction, statistical pattern recognition, blind systems identification and estimation, speech and signal processing, and biomedical applications.

Dr. Zhao was Associate Editor of IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING and member of IEEE Speech Technical Committee. She received the 1995 NSF Career Award and is listed in *American Men and Women of Science*, 1998.