Learning Mixture Models with the Latent Maximum Entropy Principle

Shaojun Wang[†‡] SJWANG@CS.UWATERLOO.CA
Dale Schuurmans[‡] DALE@CS.UWATERLOO.CA
Fuchun Peng[‡] F3PENG@CS.UWATERLOO.CA
Yunxin Zhao[*] ZHAOY@MISSOURI.EDU
[†]Department of Statistics, University of Toronto, Toronto, Ontario M5S 3G3, Canada
[‡]School of Computer Science, Unversity of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[*]Department of CSCE, University of Missouri at Columbia, Columbia, MO 65211, USA

Abstract

We present a new approach to estimating mixture models based on a new inference principle we have proposed: the latent maximum entropy principle (LME). LME is different both from Jaynes' maximum entropy principle and from standard maximum likelihood estimation. We demonstrate the LME principle by deriving new algorithms for mixture model estimation, and show how robust new variants of the EM algorithm can be developed. Our experiments show that estimation based on LME generally yields better results than maximum likelihood estimation, particularly when inferring latent variable models from small amounts of data.

1. 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 (Dempster et al. 1977). EM is an iterative parameter optimization technique that is guaranteed to converge to a local maxima in likelihood. 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 (Dempster et al. 1977; McLachlan and Peel 2000; Wu 1983). However, a number of key issues remain unresolved. For example, since the likelihood function 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, 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 over-fitting (poor generalization) particularly when faced with limited training data.

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 (Wang et al. 2003). 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 demonstrate the use of LME for estimating mixture models.

2. Motivation

The easiest way to motivate LME is with an example. Assume we observe a random variable Y that reports people's heights in a population. Given sample data $\tilde{Y} = (y_1, ..., y_T)$, 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 (Jaynes 1983) suggests that one should infer a distribution for Y that has maximum entropy, subject to the constraints 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 $m_1 = \frac{1}{T}\sum_{t=1}^{T} y_t$ and $m_2 = \frac{1}{T}\sum_{t=1}^{T} y_t^2$ respectively. In

Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

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 (Cover and Thomas 1991).

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 bi-modal data, because it will continue to postulate a uni-modal distribution. However, the existence of the two peaks might be due to the fact that there are two sub-populations 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 sub-population 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 $(y_1, c_1, ..., y_T, c_T)$. 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 $p(c) = \theta_c = \frac{N_c}{T}$ and $p(y|c) = N(y; \mu_c, \sigma_c^2)$, where $\mu_c = \frac{1}{N_c}\sum_{t=1}^{T} y_t\,\delta_c(c_t)$ and $\sigma_c^2 = \frac{1}{N_c}\sum_{t=1}^{T}(y_t - \mu_c)^2\,\delta_c(c_t)$ 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 noncommittal with respect to missing information, becomes insufficient. For example, if the gender label is unobserved, one would still be reduced to inferring a uni-modal Gaussian as above. To cope with missing but non-arbitrary hidden structure, we must extend the ME principle to account for the underlying causal structure in the data.

3. The latent maximum entropy principle

To formulate the LME principle, 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 $f_1, ...f_N$, 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

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

subject to the constraints

$$\int_{x\in\mathcal{X}} f_i(x)\,p(x)\;\mu(dx) = \sum_{y\in\tilde{\mathcal{Y}}}\tilde{p}(y)\int_{z\in\mathcal{Z}} f_i(x)\,p(z|y)\;\mu(dz)$$

$$i = 1...N, \quad Y \text{ and } Z \text{ not independent} \qquad (2)$$

where $x = (y, z)$. Here $\tilde{p}(y)$ is the empirical distribution over the observed data, and $\tilde{\mathcal{Y}}$ 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 incomplete data Y, taking into account the structure of the dependence of the unobserved component Z on Y.

Unfortunately, there is no simple solution for p^* in (1,2). However, a good approximation can be obtained by restricting the model to have an exponential form

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

where $\Phi_\lambda = \int_{x\in\mathcal{X}}\exp\left(\sum_{i=1}^{N}\lambda_i f_i(x)\right)\mu(dx)$ 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 algorithm for approximately satisfying the LME principle.

4. A training algorithm for log-linear models

To derive a practical training algorithm for log-linear models, we exploit the following intimate connection between LME and maximum likelihood estimation (MLE).

Theorem 1 *Under the log-linear assumption, maximizing the likelihood of log-linear models on incomplete data is equivalent to satisfying the feasibility constraints of the LME principle. That is, the only distinction between MLE and LME in log-linear models is that, among local maxima (feasible solutions), LME selects the model with the maximum entropy, whereas MLE selects the model with the maximum likelihood (Wang et al. 2003).*

This connection allows us to exploit an EM algorithm (Dempster et al. 1977) to find *feasible* solutions to the LME principle. It is important to emphasize, however, that EM will only find alternative feasible solutions, while the LME and MLE 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 log-likelihood function $L(\lambda)$ into

$$L(\lambda) \quad = \quad \sum_{y\in\tilde{\mathcal{Y}}}\tilde{p}(y)\log p_\lambda(y) \quad = \quad Q(\lambda, \lambda') \; + \; H(\lambda, \lambda')$$

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

$$Q(\lambda, \lambda^{(j)}) = -\log(\Phi_\lambda) + \sum_{i=1}^{N} \lambda_i \left(\sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x)\, p_{\lambda^{(j)}}(z|y)\, \mu(dz) \right) \quad (3)$$

Interestingly, it turns out that maximizing $Q(\lambda, \lambda^{(j)})$ as a function of λ for fixed $\lambda^{(j)}$ (the M step) is equivalent to solving another constrained optimization problem corresponding to a maximum entropy principle; but a much simpler one than before (Wang et al. 2003).

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

$$\max_p \quad H(p) = -\int_{x \in \mathcal{X}} p(x) \log p(x) \mu(dx) \quad (4)$$

$$\text{subject to} \quad \int_{x \in \mathcal{X}} f_i(x)\, p(x)\, \mu(dx) = \quad (5)$$

$$\sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x)\, p_{\lambda^{(j)}}(z|y)\, \mu(dz), \quad i = 1...N$$

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 (5) no longer depends on λ but rather on the fixed constants from the previous iteration $\lambda^{(j)}$. This means that maximizing (4) subject to (5) with respect to λ is now a convex optimization problem with linear constraints. The generalized iterative scaling algorithm (GIS) (Darroch et al. 1972) or improved iterative scaling algorithm (IIS) (Della Pietra et al. 1997) can be used to maximize $Q(\lambda, \lambda^{(j)})$ 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.

EM-IS algorithm:
E step: Given $\lambda^{(j)}$, for each feature f_i, $i = 1, ..., N$, calculate its current expectation $\eta_i^{(j)}$ with respect to $\lambda^{(j)}$ by: $\eta_i^{(j)} = \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x)\, p_{\lambda^{(j)}}(z|y)\, \mu(dz)$.
M step: Perform S iterations of full parallel update of parameter values $\lambda_1, ..., \lambda_N$ either by GIS or IIS as follows. Each update is given by $\lambda_i^{(j+s/S)} = \lambda_i^{(j+(s-1)/S)} + \gamma_i$, such that γ_i satisfies $\int_{x \in \mathcal{X}} f_i(x) e^{\gamma_i f(x)} p_{\lambda^{(j+(s-1)/S)}}(x) \mu(dx) = \eta_i^{(j)}$, where $f(x) = \sum_{k=1}^{N} f_k(x)$ and $s = 1, ..., S$. ∎

Provided that the E and M steps can both be computed, 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 LME principle.

Theorem 2 *The EM-IS algorithm monotonically increases the likelihood function $L(\lambda)$, and all limit points of any EM-IS sequence $\{\lambda^{(j+s/S)}, j \geq 0, s = 1..S\}$, belong to the set $\Theta = \{\lambda \in \Re^N : \partial L(\lambda)/\partial\lambda = 0\}$. Therefore, EM-IS asymptotically yields feasible solutions to the LME principle for log-linear models (Wang et al. 2003).*

Thus, EM-IS provides an effective means to find feasible solutions to the LME principle. (We note that Lauritzen (1995) has suggested a similar algorithm, but did not supply a convergence proof. More recently, Riezler (1999) 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 EM-IS algorithm to develop a practical approximation to the LME principle.

ME-EM-IS algorithm:
Initialization: Randomly choose initial guesses for λ.
EM-IS: Run EM-IS to convergence, to obtain feasible λ^*.
Entropy calculation: Calculate the entropy of p_{λ^*}.
Model selection: Repeat the above steps several times to produce a set of distinct feasible candidates. Choose the feasible candidate that achieves the highest entropy. ∎

This leads to a new estimation technique that we will compare to standard MLE below. One apparent complication, first, is that we need to calculate the entropies of the candidate models produced by 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 EM-IS to convergence.

Corollary 1 *If λ^* is feasible, then $Q(\lambda^*, \lambda^*) = -H(p_{\lambda^*})$ and $L(\lambda^*) = -H(p_{\lambda^*}) + H(\lambda^*, \lambda^*)$.*

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

To draw a clear distinction between LME and MLE, assume that the term $H(\lambda^*, \lambda^*)$ from Corollary 1 is constant across different feasible solutions. Then MLE, which maximizes $L(\lambda^*)$, will choose the model that has lowest entropy, whereas LME, which maximizes $H(p_{\lambda^*})$, will chose a model that has least likelihood. (Of course, $H(\lambda^*, \lambda^*)$ will not be constant in practice and the comparison between MLE and LME is not so straightforward, but this example does highlight their difference.) The fact that LME and MLE 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.

5. LME for learning Gaussian mixtures

In the traditional approach to mixture models (McLachlan et al. 2000), 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.

Let $X = (Y, C)$, where Y is an observable M dimensional random vector and $C \in \{1, ..., K\}$ 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 $\ell, m = 1, ..., M$, $k = 1, ..., K$, where $\delta_k(c)$ denotes the indicator function of the event $c = k$. Then, given the observed data $\tilde{\mathcal{Y}} = (y^1, ..., y^T)$, the LME principle can be formulated as

$$\begin{aligned}
\max_{p(x)} \; H(X) \quad &= \quad H(C) + H(Y|C) \quad \text{subject to} \\
\int_{x\in\mathcal{X}} \delta_k(c) p(x) \, \mu(dx) \quad &= \quad \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_{c=1}^{K} \delta_k(c) p(c|y) \qquad (6) \\
\int_{x\in\mathcal{X}} y_\ell \, \delta_k(c) p(x) \, \mu(dx) \quad &= \quad \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_{c=1}^{K} y_\ell \, \delta_k(c) p(c|y) \\
\int_{x\in\mathcal{X}} y_\ell \, y_m \, \delta_k(c) p(x) \, \mu(dx) \quad &= \quad \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_{c=1}^{K} y_\ell \, y_m \, \delta_k(c) p(c|y) \\
Y \text{ and } C \text{ not independent} \qquad & \qquad \ell, m = 1, ..., M; \, k = 1, ..., K
\end{aligned}$$

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 $\lambda = (\lambda_0^k, \lambda_\ell^k, \lambda_{\ell,m}^k)$, $\ell, m = 1, ..., M$ and $k = 1, ..., K$, for the features. To execute the E step, we then calculate the right hand side feature expectations

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

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 maximization problem with linear constraints, as in (4,5)

$$\begin{aligned}
\max_{p(x)} \; H(X) = H(C) + H(Y|C) \quad & \text{subject to} \\
\int_{x\in\mathcal{X}} \delta_k(c) \;\; p(x) \, \mu(dx) \quad &= \quad \eta_0^{k,(j)} \\
\int_{x\in\mathcal{X}} y_\ell \, \delta_k(c) \;\; p(x) \, \mu(dx) \quad &= \quad \eta_\ell^{k,(j)} \qquad (7) \\
\int_{x\in\mathcal{X}} y_\ell \, y_m \, \delta_k(c) \;\; p(x) \, \mu(dx) \quad &= \quad \eta_{\ell,m}^{k,(j)}
\end{aligned}$$

for $\ell, m = 1, ..., M$; $k = 1, ..., K$, where $x = (y, c)$. This problem can be solved analytically. In particular, for (7) we can directly obtain the unique log-linear solution $p(x) = p(y, c)$, where $p(c) = \frac{1}{T} \sum_{t=1}^{T} \rho_t^{c,(j)}$ 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 $c = 1, ..., K$. 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 multi-modal distribution over the observed component Y, whereas standard ME is reduced to producing a uni-modal Gaussian here.[1] Interestingly, the update formula we obtain for $p_{\lambda^{(j)}} \to p_{\lambda^{(j+1)}}$ 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 iterative 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.

6. Gaussian mixture experiments

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

Synthetic experiments 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 $Y \in \Re^2$ is observed. Thus, the features we match in the data are of the same form as in Section 5. Given sample data $\tilde{\mathcal{Y}} = (y_1, ..., y_T)$ 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

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

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

[1] Radford Neal has observed that dropping the dependence constraint between Y and C allows the uni-modal 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.

LME and MLE 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 $\tilde{\mathcal{Y}} = (y_1, ..., y_T)$ according to $p^*(y)$.
3. Run 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 $\{.5, 2.5\}$.
4. Calculate entropy and likelihood for each candidate.
5. Select the maximum entropy candidate p_{LME} as the LME estimate, and the maximum likelihood candidate p_{MLE} in the interior of the parameter space as the MLE estimate.
6. Calculate the cross entropy from $p^*(y)$ to the marginals $p_{LME}(y)$ and $p_{MLE}(y)$ 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.
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 $\theta_c = \frac{1}{3}$ 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 $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ respectively.

Figures 1 and 2 first show that the average log-likelihoods and average entropies of the models produced by LME and MLE respectively behave as expected. MLE clearly achieves higher log-likelihood than LME, however LME clearly produces models that have significantly higher entropy than MLE. The interesting outcome is that the two estimation strategies obtain significantly different cross entropies. Figure 3 reports the average cross entropy obtained by MLE and LME as a function of sample size, and shows the somewhat surprising result that LME achieves substantially lower cross entropy than MLE. LME's advantage is especially pronounced at small sample sizes, and persists even to sample sizes as large as 10,000 (Figure 3).

Although one might have expected an advantage for LME because of a "regularization" effect, this does not completely explain LME's superior performance at large sample sizes. (We return to a more thorough discussion of LME's regularization properties in Section 8.)

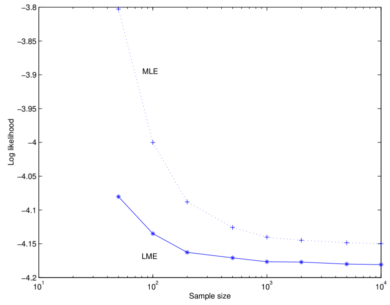

Figure 1. Average log-likelihood of the MLE estimates versus the LME estimates in Experiment 1.

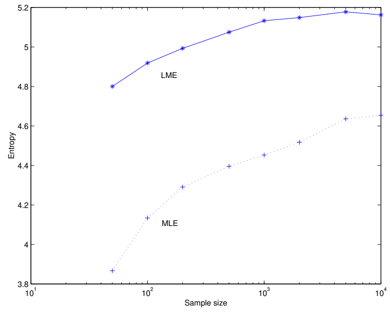

Figure 2. Average entropy of the MLE estimates versus the LME estimates in Experiment 1.

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 $\Re^2$. Specifically, we generated data by sampling from a uniform distribution over mixture components $\theta_c = \frac{1}{5}$ for $c = 1, ..., 5$, and then generated the observed data $Y \in \Re^2$ 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. The LME

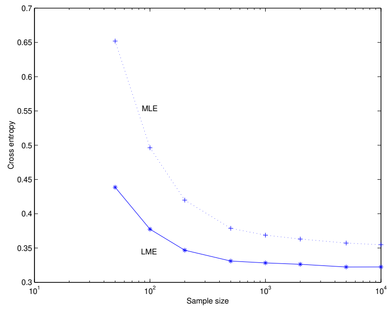

Figure 3. Average cross entropy between the true distribution and the MLE estimates versus the LME estimates in Experiment 1.

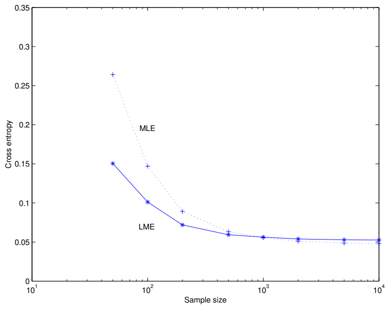

Figure 4. Average cross entropy between the true distribution and the MLE estimates versus the LME estimates in Experiment 2.

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

Figure 4 shows that LME still obtained a significantly lower cross entropy than MLE at small sample sizes, but lost its advantage at larger sample sizes. At a crossover point of $T = 1000$ data points, MLE began to produce slightly better estimates than LME, but only marginally so. Overall, LME 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 $\theta_c = \frac{1}{3}$ for $c = 1, 2, 3$, and means $[2\ 0]^\top$, $[0\ 0]^\top$, $[0\ 2]^\top$ and "covariances" $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ for the Laplacians.

Figure 5 shows that LME produces significantly better estimates than MLE in this case, and even improved its advantage at larger sample sizes. Clearly, MLE is not a stable estimator when subjected to heavy tailed data when this is not expected. LME proves to be far more robust in such circumstances and clearly dominates MLE.

Scenario 4 However, there are other situations where MLE appears to be a slightly better estimator than LME when sufficient data is available. Figure 6 shows the results of subjecting the estimators to data generated from a three component Gaussian mixture, $\theta_c = \frac{1}{3}$, $c = 1, 2, 3$, with means $[2\ 0]^\top$, $[0\ 0]^\top$, $[0\ 2]^\top$ and covariances $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ respectively. In this case, LME still re-

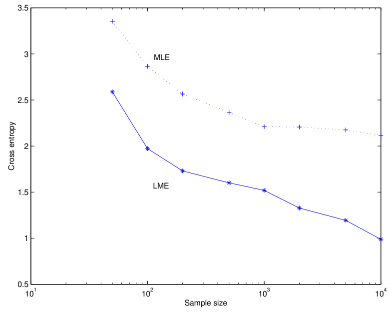

Figure 5. Average cross entropy between the true distribution and the MLE estimates versus the LME estimates in Experiment 3.

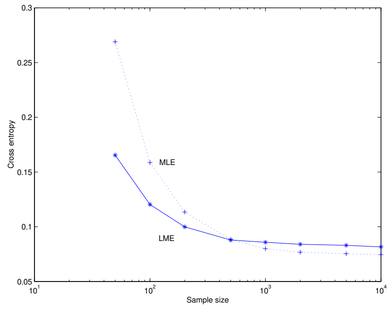

Figure 6. Average cross entropy between the true distribution and the MLE estimates versus the LME estimates in Experiment 4.

tains a sizeable advantage at small sample sizes, but after a sample size of $T = 500$, MLE begins to demonstrate a persistent advantage.

Overall, these results suggest that maximum likelihood estimation (MLE) 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 LME 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 Fisher (1936). 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. 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 sam-

Table 1. Comparison of LME and MLE on *Iris* data set.

	LOG-LIKELIHOOD	ERROR RATE
LME	5.58886	0.1220
MLE	5.37704	0.2446

ple 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 1 shows the averaged results. We see that the test data is more likely under the LME estimates, and also that the clustering error rate is cut in half.

7. LME for learning Dirichlet mixtures

Of course, the LME 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—cf. Section 8). Here we present an alternative application of LME to estimating a mixture of Dirichlet sources.

Assume the observed data has the form of an M dimensional probability vector $y = (y_1, ..., y_M)$ such that $0 \le y_\ell \le 1$ for $\ell = 1, ..., M$ and $\sum_{\ell=1}^M y_\ell = 1$. That is, the observed variable is a random vector $Y = (Y_1, ..., Y_M) \in [0,1]^M$, which happens to be normalized. There is also an underlying class variable $C \in \{1, ..., K\}$ that is unobserved. Let $X = (Y, C)$. Given an observed sequence of T M-dimensional probability vectors $\tilde{\mathcal{Y}} = (y^1, ..., y^T)$, where $y^t = (y_1^t, ..., y_M^t)$ for $t = 1, ..., T$, 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 $\ell = 1, ..., M$ and $k = 1, ..., K$, where $x = (y, c)$. In this case, the LME principle is

$$\begin{aligned}
&\max_{p(x)} \ H(X) = H(C) + H(Y|C) \quad \text{subject to} \\
&\int_{x \in \mathcal{X}} \delta_k(c)\, p(x)\, \mu(dx) = \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \sum_c \delta_k(c)\, p(c|y)\, \mu(dx) \\
&\int_{x \in \mathcal{X}} (-\log y_\ell)\, \delta_k(c)\, p(x)\, \mu(dx) \\
&\qquad = \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \sum_c (-\log y_\ell)\, \delta_k(c)\, p(c|y)\, \mu(dx) \\
&\ell = 1, ..., M, \ k = 1, ..., K \quad \text{and } Y, C \text{ not independent}
\end{aligned} \tag{8}$$

Here $\tilde{p}(y) = \frac{1}{T}$ 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 (8). 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

$$\eta_0^{k,(j)} = \frac{1}{T} \sum_{t=1}^T \sum_{c=1}^K \delta_k(c)\ \rho_t^{k,(j)} \tag{9}$$

$$\eta_\ell^{k,(j)} = \frac{1}{T} \sum_{t=1}^T \sum_{c=1}^K (-\log y_\ell^t)\, \delta_k(c)\ \rho_t^{k,(j)}$$

for $\ell = 1, ..., M$, $k = 1, ..., K$, 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 5.

To perform the M step we then formulate the simpler maximum entropy problem with linear constraints, as in (4,5)

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

for $\ell = 1, ..., M$ and $k = 1, ..., K$. For this problem we can obtain a log-linear solution of the form $p(x) = p(y, c)$ where $p(c) = \frac{1}{T} \sum_{t=1}^T \rho_k^t$ and the class conditional model $p(y|c)$ is a Dirichlet distribution with parameters $\alpha_\ell^c = 1 - \lambda_l^c$; that is $p(y|c) = \Gamma\left(\sum_{\ell=1}^M \alpha_\ell^c\right) \left(\prod_{\ell=1}^M \Gamma(\alpha_\ell^c)\right)^{-1} \prod_{\ell=1}^M y_\ell^{\alpha_\ell^c - 1}$. However, we still need to solve for the parameters α_ℓ^c. By plugging in the form of the Dirichlet distribution, the feature expectations (9) will have an explicit formula, and the constraints on the parameters α_ℓ^c can then be written

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

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

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

for $\ell = 1, ..., M$ and $k = 1, ..., K$. This iteration corresponds to a well known technique for locally monotonic maximizing the likelihood of a Dirichlet mixture (Minka, 2000). 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.

Dirichlet mixture experiment To compare model selection based on the LME versus MLE 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 $\theta_c = \frac{1}{6}, \frac{1}{2}, \frac{1}{3}$ and component Dirichlets specified by the α parameters $[1\ 2]^\top$, $[3\ 1]^\top$ and $[5\ 2]^\top$ respectively. The initial mixture weights were generated from a uniform prior, and each α was generated by choosing numbers uniformly from $\{0.1, 0.5, 1, 2.5, 5\}$. Figure 7 shows the cross

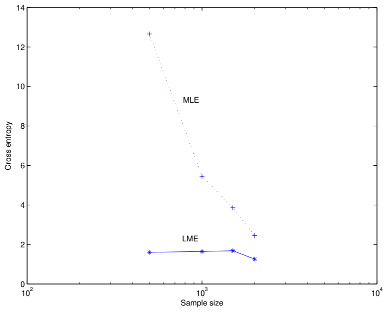

Figure 7. Average cross entropy between true distribution and MLE versus LME estimates in the Dirichlet mixture experiment.

entropy results of LME and MLE averaged over 10 repeated trials for each fixed training sample size. The outcome in this case shows a significant advantage for LME.

8. 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 demonstrate in (Wang et al. 2003), one can add a regularization term to the LME principle in the same way one can add a regularization term to the MLE principle. LME behaves more like an *adaptive* rather than fixed regularizer, because we see no real under-fitting from LME on large data samples, even though LME chooses far "smoother" models than MLE at smaller sample sizes. In fact, LME 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 goes to zero) no finite penalty can counteract the resulting unbounded likelihood. However, the LME principle can automatically filter out degenerate models, because such models have a differential entropy of $-\infty$ and any non-degenerate model will be preferred. Eliminating degenerate models by the LME principle solves one of the main practical problems with Gaussian mixture estimation.

Another observation is that all of our experiments show that MLE and LME reduce cross entropy error when the sample size is increased. However, we have not yet proved that the LME 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. Determining the statistical consistency of LME, in either sense, remains an important topic for future research.

In this paper, by randomly choosing different starting points, we take the feasible log-linear model with maximum entropy value as the LME estimate. This procedure is computationally expensive. Thus it is worthwhile to develop an analogous deterministic annealing ME-EM-IS algorithm to automatically find feasible maximum entropy log-linear model for LME (Ueda and Nakano 1998).

Finally, we point out that the LME principle can be applied to other statistical models beyond mixtures, such as hidden Markov models (Lafferty et al. 2001) and Boltzmann machines (Ackley et al. 1985). 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.

References

Ackley, D., Hinton, G., Sejnowski, T. (1985). A learning algorithm for Boltzmann machines. *Cogn. Sci.*, *9*, 147-169.

Cover, T., Thomas, J. (1991). *Elements of Information Theory*. John Wiley & Sons, Inc.

Darroch, J., Ratchliff, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, *43-5*, 1470-1480.

Della Pietra, S., Della Pietra, V., Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19-4*, 380-393.

Dempster, A., Laird, N., Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *J. Royal Statis. Soc. B*, *39*, 1-38.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7, II*, 179-188.

Jaynes, E. (1983). *Papers on Probability, Statistics, and Statistical Physics*. R. Rosenkrantz (ed). D. Reidel.

Lafferty, J., McCallum, A., Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proceedings ICML-2001*.

Lauritzen, S. (1995). The EM-algorithm for graphical association models with missing data. *Comput. Statist. Data Analysis*, *1*, 191-201.

McLachlan, G., Peel, D. (2000). *Finite mixture models*. John Wiley & Sons, Inc.

Minka, T. (2000). Estimating a Dirichlet distribution. Manuscript.

Riezler, S. (1999). *Probabilistic constraint logic programming*. Ph.D. Thesis, University of Stuttgart.

Ueda, N., Nakano, R. (1998). Deterministic annealing EM algorithm. *Neural Networks*, *11*, 272-282.

Wang, S., Schuurmans, D., Zhao, Y. (2003). The latent maximum entropy principle. manuscript

Wu, C. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, *11*, 95-103.