$\mathcal{U}$-Likelihood and $\mathcal{U}$-Updating Algorithms: Statistical Inference in Latent Variable Models

Jaemo Sung[1], Sung-Yang Bang[1], Seungjin Choi[1], and Zoubin Ghahramani[2]

[1] Department of Computer Science,
POSTECH, Republic of Korea
{emtidi, sybang, seungjin}@postech.ac.kr
[2] Gatsby Computational Neuroscience Unit,
University College London, 17 Queen Square,
London WC1N 3AR, England
zoubin@gatsby.ucl.ac.uk

Abstract. In this paper we consider latent variable models and introduce a new *$\mathcal{U}$-likelihood* concept for estimating the distribution over hidden variables. One can derive an estimate of parameters from this distribution. Our approach differs from the Bayesian and Maximum Likelihood (ML) approaches. It gives an alternative to Bayesian inference when we don't want to define a prior over parameters and gives an alternative to the ML method when we want a better estimate of the distribution over hidden variables. As a practical implementation, we present a *$\mathcal{U}$-updating algorithm* based on the mean field theory to approximate the distribution over hidden variables from the $\mathcal{U}$-likelihood. This algorithm captures some of the correlations among hidden variables by estimating reaction terms. Those reaction terms are found to penalize the likelihood. We show that the $\mathcal{U}$-updating algorithm becomes the EM algorithm as a special case in the large sample limit. The useful behavior of our method is confirmed for the case of mixture of Gaussians by comparing to the EM algorithm.

1 Introduction

Latent variable models are important tools for probabilistic methods and have wide applications in machine learning, computer vision, pattern recognition, and speech processing, to name a few. The Bayesian and the Maximum Likelihood (ML) approaches have been extensively studied for learning such models in the past decades.

In Bayesian Inference [1], we define a prior over parameters $P(\boldsymbol{\theta})$ and from this all inference is automatically performed. In particular, using this prior we can compute the *marginal probability* of data set $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ and hidden variable set $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$:

$$P(Y, X) = \int P(Y, X|\boldsymbol{\theta}) P(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{1}$$

J. Gama et al. (Eds.): ECML 2005, LNAI 3720, pp. 377–388, 2005.

We can further marginalize out hidden variables to get the marginal probability of just data set Y:

$$P(Y) = \sum_X P(Y, X), \tag{2}$$

assuming all hidden variables are discrete. The Bayesian approach basically provides a way of solving the overfitting problem by eliminating model parameters by integrating over them.

In certain settings it may be undesirable to define a prior over parameters. For example, many statisticians don't like the subjective nature of Bayesian inference even though all modelling contains an element of subjectivity. Moreover, in some cases it is very difficult to define one's prior belief about the model parameters. This motivates the use of ML method for parameter estimation.

Starting from the likelihood of the parameters, which is the probability of data set Y given the parameters, in the ML approach one can find the parameters which maximize the likelihood function given by

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_X P(Y, X|\boldsymbol{\theta}). \tag{3}$$

We wish to find the parameters that maximize the likelihood: $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$. From this estimate of parameters, we can find the distribution over hidden variables $P(X|Y, \boldsymbol{\theta}^*)$, regarding $\boldsymbol{\theta}^*$ as true parameters. The fundamental problem of the ML approach is overfitting since it considers only a single estimate of ML parameters, whereas the Bayesian approach solves this problem by integrating over parameters.

If we don't want a Bayesian approach, we can still eliminate the parameters, not by marginalizing over them as in (1), but by maximizing over them. This is the key of our work. By doing so, we obtain a method somewhat analogous to the Bayesian approach without specifying the parameter prior.

In this paper, we introduce a new concept, *U-likelihood*, to infer the distribution over hidden variables, which differs from the Bayesian and the ML approaches. We obtain the U-likelihood, which is an analogous quantity to the marginal probability of data set in (2), by marginalizing the maximum of the complete-data likelihood over hidden variables. We show that the U-likelihood can be bounded using variational method [2] and this gives the joint distribution over hidden variables $Q(X)$ (Sec. 3). Like the Bayesian and the ML approaches, the exact $Q(X)$ is intractable to compute for large data sets. As a practical implementation, we introduce a *$\mathcal{U}$-updating algorithm*, which iteratively solves mean field equations for hidden variables under $Q(X)$. We show that the $\mathcal{U}$-updating algorithm estimates the reaction of all the other hidden variables and it penalizes the likelihood term to alleviate the overfitting problem. The EM algorithm appears as a special case of $\mathcal{U}$-updating algorithm in the large sample limit. (Sec. 4). We demonstrate the useful behavior of our $\mathcal{U}$-updating algorithm, compared to the EM algorithm, through the example of mixtures of Gaussians on synthetic and real data sets (Sec. 5).

2 General Framework

Throughout this paper, we assume that data set $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ of n data points is always given. Let $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ denote hidden variable set. Allowing $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$, to be multidimensional, we assume that a complete data point $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ is IID from a sampling distribution which is parameterized by parameter vector $\boldsymbol{\theta}$ such as $P(\boldsymbol{y}_i, \boldsymbol{x}_i|\boldsymbol{\theta})$. For simplicity, we here focus on the discrete type hidden variable $\boldsymbol{x}_i$, but the understanding of the case of continuous hidden variables is straightforward by exchanging sums into integrals.

For Bayesian inference, we can form a lower bound of $\log P(Y)$ in (2) for any $Q(X)$ using Jensen's inequality:

$$\log P(Y) \geq \sum_X Q(X) \log \frac{P(Y,X)}{Q(X)} \equiv \mathcal{F}_{\mathcal{B}}(Q(X)), \tag{4}$$

and then we can find $Q(X)$ by maximizing $\mathcal{F}_{\mathcal{B}}$. The maximization of $\mathcal{F}_{\mathcal{B}}$ is equivalent to the minimization of the Kullback-Leibler divergence between $Q(X)$ and $P(X|Y) = P(Y,X)/P(Y)$. Therefore, at maxima of $\mathcal{F}_{\mathcal{B}}$, $Q(X)$ gives the exact $P(X|Y)$. However, for most models of interest this is intractable to compute. For example, for a mixture model with m components, the sum $\sum_X$ of $P(Y)$ in (2) contains m^n terms. As practical implementations, MCMC [3] methods, the Expectation-Propagation (EP) [4] and the variational Bayes (VB) [5] methods were introduced but we will not tackle them in this paper.

For the ML approach, we can form the lower bound of log likelihood in a similar way to (4):

$$\log \mathcal{L}(\boldsymbol{\theta}) \geq \sum_X Q(X) \log \frac{P(Y,X|\boldsymbol{\theta})}{Q(X)} \equiv \mathcal{F}_{\mathcal{L}}(Q(X), \boldsymbol{\theta}). \tag{5}$$

The maximization of $\mathcal{F}_{\mathcal{L}}$ is equivalent to the maximization of $\mathcal{L}$ since if $(Q^*, \boldsymbol{\theta}^*)$ occurs at maxima of $\mathcal{F}_{\mathcal{L}}$, then $\boldsymbol{\theta}^*$ occurs at maxima of $\mathcal{L}(\boldsymbol{\theta})$ and $Q^*(X)$ becomes $P(X|Y, \boldsymbol{\theta}^*)$. Since the global maximization of $\mathcal{F}_{\mathcal{L}}$ is intractable in most cases like in the Bayesian approach, the well-known EM algorithm [6] independently maximizes $\mathcal{F}_{\mathcal{L}}$ w.r.t. Q or $\boldsymbol{\theta}$ by fixing the other as a practical implementation. Refer [6, 7] for more details on the EM algorithm.

3 $\mathcal{U}$-Likelihood

If we don't want a Bayesian approach, we can still eliminate the parameters, not by marginalizing over them as in (1), but by maximizing over them. We start by defining the $\mathcal{U}$-function which is the maximum of the complete-data likelihood:

$$\mathcal{U}(Y,X) \equiv \max_{\boldsymbol{\theta}} P(Y,X|\boldsymbol{\theta}) = P(Y,X|\widehat{\boldsymbol{\theta}}(Y,X)) > 0, \tag{6}$$

where $\widehat{\boldsymbol{\theta}}\,(Y,X)$ denotes the ML parameter estimator, a function of the complete-data set, defined by

$$\widehat{\boldsymbol{\theta}}\,(Y,X) \equiv \arg\max_{\boldsymbol{\theta}} P(Y,X|\boldsymbol{\theta}). \tag{7}$$

The $\mathcal{U}$-function is analogous to the marginal probability in (1) except that it maximizes over parameters rather than integrating over parameters. Another way to think about it is that instead of a parameter prior, we substitute in (1) a delta function of the ML parameter estimate on the complete-data set, e.g. $P(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\,(Y, X))$. This is certainly not coherent from the point of view of Bayesian inference since the prior cannot depend on the data, but we will see some of the interesting properties of this approach.

We can take the $\mathcal{U}$-function and marginalize out the hidden variable set X:

$$\mathcal{U}(Y) \equiv \sum_X \mathcal{U}(Y, X). \tag{8}$$

We call this quantity *$\mathcal{U}$-likelihood* and it is analogous to $P(Y)$ in (2). Another view of it is that it forms an upper bound of the likelihood function: $\mathcal{U} \geq \mathcal{L}^* \geq \mathcal{L}(\boldsymbol{\theta})$, where $\mathcal{L}^* = \max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ denotes the maximum likelihood value. Analogous to (4), we can lower bound it for any distribution $Q(X)$ over hidden variables:

$$\log \mathcal{U}(Y) \geq \sum_X Q(X) \log \frac{\mathcal{U}(Y, X)}{Q(X)} \equiv \mathcal{F}_{\mathcal{U}}(Q(X)). \tag{9}$$

We can use the optimal $Q(X)$ maximizing $\mathcal{F}_{\mathcal{U}}$ as a joint conditional distribution over hidden variables given data set: $P(X|Y)$. The next theorem shows the form of $Q(X)$ at the maxima of $\mathcal{F}_{\mathcal{U}}$.

Theorem 1. *The optimal joint distribution $Q(X)$ maximizing the lower bound $\mathcal{F}_{\mathcal{U}}(Q(X))$ is of the form*

$$Q(X) = \frac{\mathcal{U}(Y, X)}{\mathcal{U}(Y)}. \tag{10}$$

Proof: Let $Q'(X) = \frac{\mathcal{U}(Y,X)}{\mathcal{U}(Y)}$. Then, the Kullback-Leibler divergence between $Q(X)$ and $Q'(X)$ is given by $KL\,[Q \| Q'] = \log \mathcal{U}(Y) - \mathcal{F}_{\mathcal{U}}(Q)$. It follows from Gibbs inequality that $KL\,[Q \| Q'] = 0$ when $Q(X) = Q'(X)$, implying that $\mathcal{F}_{\mathcal{U}}(Q)$ is maximized at $Q(X) = Q'(X)$. ■

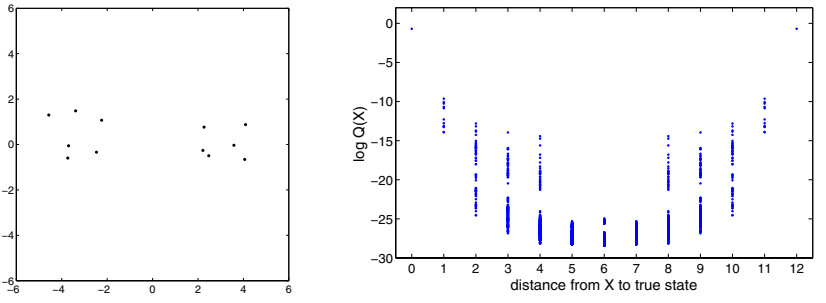
We illustrate the joint distribution $Q(X)$ in (10) when data set Y consists of 12 data points generated from the mixture of two Gaussians. The true X^* is a binary vector with 12 components. Figure 1 plots $\log Q(X)$ as a function of Manhattan distance from the true X^*. This demonstrates that $Q(X)$ tends to give higher probability to hidden states that are similar to the true states.

We have seen the relationship of $\mathcal{U}$-likelihood to Bayesian inference. We can also see a simple relationship to maximum likelihood methods:

$$\text{Maximum likelihood} \quad : \quad \mathcal{L}^* = \max_{\boldsymbol{\theta}} \sum_X P(Y, X|\boldsymbol{\theta}) \;, \tag{11}$$

$$\mathcal{U}\text{-likelihood} \quad : \quad \mathcal{U} = \sum_X \max_{\boldsymbol{\theta}} P(Y, X|\boldsymbol{\theta}) \;, \tag{12}$$

where we here dropped the data dependency. The former gives a single value of the model parameters $\boldsymbol{\theta}^*$, from which a distribution over hidden variables can



(a) (b)

Fig. 1. Demonstration of $Q(X)$ in Theorem 1 given data set Y of 12 data points generated from the mixture of two Gaussians, i.e. $\mathcal{N}([-3,0],\mathbf{I})$ and $\mathcal{N}([3,0],\mathbf{I})$. (a) data set Y. (b) $\log Q(X)$ as a function of Manhattan distance from the true state, $\sum_{i=1}^{12} |\boldsymbol{x}_i - \boldsymbol{x}_i^*|$. Each dot indicates a state of 2^{12} possible configurations of X. The symmetrical phenomenon stems from the identifiability of the mixture of Gaussians.

be derived: $P(X|Y,\boldsymbol{\theta}^*)$. The latter no longer gives a single value of parameters. However, it may give a better estimate of the distribution over hidden variables, $Q(X)$, which captures some of correlations among hidden variables. From this distribution $Q(X)$ one can derive an estimate of parameters.

We outline some of the possible advantages of $\mathcal{U}$-likelihood approach over the Bayesian and the ML approaches:

1. High dimensional integrals like (1) required for Bayesian inference can be intractable. For many models, the optimum of $\boldsymbol{\theta}$ given the complete-data set (Y,X) is a simple function of the sufficient statistics. So no explicit optimization is necessary to compute (6).
2. Many researchers may not wish to define a prior over parameters. The $\mathcal{U}$-likelihood method provides an alternative.
3. Optimizing over parameters in the ML method is often fraught with local optima. By optimizing out parameters, the $\mathcal{U}$-likelihood method is sometimes found to have better convergence properties than the ML method. That is, it can find good solutions without falling into local optima as often. We show this empirically.

The distribution $Q(X)$ in (10) may be intractable to compute, excepting for small n, since it requires all possible configurations of X. As a practical implementation, we will use a mean field approximation and present the $\mathcal{U}$-updating algorithm as an alternative to the EM algorithm in the next section.

4 $\mathcal{U}$-Updating Algorithm

We start by considering a case where the sampling distribution of the complete-data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ is in the exponential family with the following form

$$P(\boldsymbol{y}_i, \boldsymbol{x}_i|\boldsymbol{\theta}) = f(\boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i))g(\boldsymbol{\theta}) \exp\Big\{\boldsymbol{\phi}(\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i)\Big\}, \tag{13}$$

where $\boldsymbol{\phi}(\boldsymbol{\theta})$ is a vector of natural parameters and $\boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i)$ is a vector of sufficient statistics. The normalizing constant is denoted by $g(\boldsymbol{\theta})$. Probability distributions of the exponential family have been widely used in latent variable models such as mixture of Gaussians, factor analysis, hidden Markov models, state-space models, and so on. For the case of the exponential family, the complete-data likelihood depends on the complete-data set only through sufficient statistics:

$$P(Y, X|\boldsymbol{\theta}) = \Bigg[\prod_{i=1}^{n} f(\boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i))\Bigg] g(\boldsymbol{\theta})^n \exp\Big\{\boldsymbol{\phi}(\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{s}(Y, X)\Big\}, \tag{14}$$

where $\boldsymbol{s}(Y, X) = \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i)$. Moreover, a closed-form solution of $\widehat{\boldsymbol{\theta}}$ and $\mathcal{U}$-function always exists as a function of sufficient statistics: $\widehat{\boldsymbol{\theta}}(Y, X) = \widehat{\boldsymbol{\theta}}(\boldsymbol{s}(Y, X))$ and $\mathcal{U}(Y, X) = \mathcal{U}(\boldsymbol{s}(Y, X))$.

The mean field theory [8], originally from statistical physics, has been widely used in the machine learning community to approximate joint distributions in graphical models when exact inference is intractable because of highly-coupled interactions among variables. Consider the marginal distribution over $\boldsymbol{x}_i$:

$$Q(\boldsymbol{x}_i) = \sum_{X_{\setminus i}} Q(X) = \frac{1}{\mathcal{U}(Y)} \sum_{X_{\setminus i}} \mathcal{U}(\boldsymbol{s}(Y, X)), \tag{15}$$

where $X_{\setminus i}$ denotes a subset of hidden variables where $\boldsymbol{x}_i$ is excluded: $X_{\setminus i} = X \setminus \boldsymbol{x}_i$. In general, the exact calculation of $Q(\boldsymbol{x}_i)$ is intractable since it requires all possible realizations of $X_{\setminus i}$. Assuming weak dependencies among hidden variables, the mean field theory suggests that the influence of the other hidden variables $\boldsymbol{s}_j(\boldsymbol{y}_j, \boldsymbol{x}_j)$ in the marginal distribution $Q(\boldsymbol{x}_i)$ can be approximated by the expected values $\langle \boldsymbol{s}_j(\boldsymbol{y}_j, \boldsymbol{x}_j) \rangle$. This leads to the mean field distributions $Q_i(\boldsymbol{x}_i)$:

$$Q_i(\boldsymbol{x}_i) \equiv \frac{1}{\mathcal{U}_i} \mathcal{U}(\bar{\boldsymbol{s}}_i(\boldsymbol{x}_i)) \approx Q(\boldsymbol{x}_i), \tag{16}$$

where $\bar{\boldsymbol{s}}_i(\boldsymbol{x}_i) = \boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) + \sum_{j=1, j\neq i}^{n} \langle \boldsymbol{s}_j(\boldsymbol{y}_j, \boldsymbol{x}_j) \rangle$ and $\mathcal{U}_i = \sum_{\boldsymbol{x}_i} \mathcal{U}(\bar{\boldsymbol{s}}_i(\boldsymbol{x}_i))$ is the normalizing constant. The joint distribution $Q(X)$ is approximated by the factored form with all mean field distributions $Q_i(\boldsymbol{x}_i)$: $Q(X) \approx \prod_{i=1}^{n} Q_i(\boldsymbol{x}_i)$. Moreover, the expected sufficient statistics $\langle \boldsymbol{s}(Y, X) \rangle$ can be obtained by solving self-consistent equations called mean field equations, which are stationary conditions:

$$\langle \boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) \rangle = \sum_{\boldsymbol{x}_i} \boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i)\, Q_i(\boldsymbol{x}_i)\ . \tag{17}$$

Table 1. $\mathcal{U}$-updating algorithm and EM algorithm

$\mathcal{U}$-updating algorithm	EM algorithm
Initialize $\langle \boldsymbol{s} \rangle^{(0)} = \sum_{i=1}^{n} \langle \boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) \rangle^{(0)}$.	
Set $\langle \boldsymbol{s} \rangle^{(1)} = \langle \boldsymbol{s} \rangle^{(0)}$.	
Repeat $t = 1, 2, \ldots$ until convergence .	Initialize $\boldsymbol{\theta}^{(0)}$.
Repeat $i = 1, \ldots, n$.	Repeat $t = 1, 2, \ldots$ until convergence.
Update $Q_i^{(t)}(\boldsymbol{x}_i) \propto \mathcal{U}(\bar{\boldsymbol{s}}_i^{(t)}(\boldsymbol{x}_i))$,	E-Step :
$\bar{\boldsymbol{s}}_i^{(t)}(\boldsymbol{x}_i) =$	Update $Q_i^{(t)}(\boldsymbol{x}_i) = P(\boldsymbol{x}_i \vert \boldsymbol{y}_i, \boldsymbol{\theta}^{(t)})$
$\langle \boldsymbol{s} \rangle^{(t)} + \boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) - \langle \boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) \rangle^{(t-1)}$.	for all $i = 1, \ldots, n$.
Refine	M-Step :
$\langle \boldsymbol{s} \rangle^{(t)} \leftarrow$	Estimate $\boldsymbol{\theta}^{(t+1)} = \widehat{\boldsymbol{\theta}}(\langle \boldsymbol{s} \rangle^{(t)})$
$\langle \boldsymbol{s} \rangle^{(t)} + \langle \boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) \rangle^{(t)} - \langle \boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) \rangle^{(t-1)}$	with $\langle \boldsymbol{s} \rangle^{(t)} = \sum_{i=1}^{n} \langle \boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) \rangle^{(t)}$
with $\langle \boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) \rangle^{(t)}$ under new $Q_i^{(t)}(\boldsymbol{x}_i)$.	under new $\{Q_i^{(t)}(\boldsymbol{x}_i)\}$.
End (Repeat)	End (Repeat)
Set $\langle \boldsymbol{s} \rangle^{(t+1)} = \langle \boldsymbol{s} \rangle^{(t)}$.	
End (Repeat)	

Therefore, the distribution $Q_i(\boldsymbol{x}_i)$ in (16) can be computed by iterative procedure solving the mean field equations in (17). This iterative procedure is referred to as the *$\mathcal{U}$-updating algorithm* and gives an alternative to the EM algorithm.

The Table 1 summarizes the $\mathcal{U}$-updating algorithm in comparison with the EM algorithm. In order to estimate ML parameter $\boldsymbol{\theta}^{(t+1)}$ in M-Step, the EM algorithm requires distributions $P(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{\theta}^{(t)})$ in E-Step built on the ML parameter $\boldsymbol{\theta}^{(t)}$ which may be overfitted to the data set at the previous iteration.Therefore, the overfitting effects may accumulate throughout iterations in the EM algorithm. However, the $\mathcal{U}$-updating algorithm alleviates this overfitting-accumulation problem by estimating the reaction of all the other hidden variables, which penalizes the likelihood. Therefore, it can give better distribution $Q_i(\boldsymbol{x}_i)$ than the EM algorithm. We can simply use all $Q_i(\boldsymbol{x}_i)$ resulted from the $\mathcal{U}$-updating algorithm to estimate parameters like the M-Step of the EM algorithm.

In order to see how the $\mathcal{U}$-updating algorithm penalizes the likelihood, decompose the $\mathcal{U}$-function:

$$\mathcal{U}(\bar{\boldsymbol{s}}_i(\boldsymbol{x}_i)) = \alpha_i(\boldsymbol{x}_i)\, \beta_i(\boldsymbol{x}_i), \tag{18}$$

where $\alpha_i(\boldsymbol{x}_i) = P(\boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) \,|\, \widehat{\boldsymbol{\theta}}(\bar{\boldsymbol{s}}_i(\boldsymbol{x}_i)))$ and $\beta_i(\boldsymbol{x}_i) = \prod_{j=1, \neq i}^{n} \rho(\langle \boldsymbol{s}_j(\boldsymbol{y}_j, \boldsymbol{x}_j) \rangle \,|\, \widehat{\boldsymbol{\theta}}(\bar{\boldsymbol{s}}_i(\boldsymbol{x}_i)))$, given by

$$\rho(\langle \boldsymbol{s}_j(\boldsymbol{y}_j, \boldsymbol{x}_j) \rangle \,|\, \widehat{\boldsymbol{\theta}}(\bar{\boldsymbol{s}}_i(\boldsymbol{x}_i))) = f(\langle \boldsymbol{s}_j \rangle) g(\widehat{\boldsymbol{\theta}}(\bar{\boldsymbol{s}}_i(\boldsymbol{x}_i))) \exp\left\{ \boldsymbol{\phi}(\widehat{\boldsymbol{\theta}}(\bar{\boldsymbol{s}}_i(\boldsymbol{x}_i)))^{\mathrm{T}} \langle \boldsymbol{s}_j \rangle \right\}.$$

The term $\alpha_i(\boldsymbol{x}_i)$ is the likelihood on the complete data point i. The term $\beta_i(\boldsymbol{x}_i)$ can be interpreted as a reaction of all the other hidden variables via the expected values $\langle \boldsymbol{s}_j(\boldsymbol{y}_j, \boldsymbol{x}_j) \rangle$. When computing $Q_i(\boldsymbol{x}_i)$, the $\mathcal{U}$-updating algorithm therefore penalizes the likelihood $\alpha_i(\boldsymbol{x}_i)$ by estimating the reaction $\beta_i(\boldsymbol{x}_i)$ of the other hidden variables, which captures some correlations among hidden variables.

The $\mathcal{U}$-updating algorithm generalizes the EM algorithm since if we ignore the reaction term $\beta_i(\boldsymbol{x}_i)$ in (18), it will be same to the EM algorithm. The following theorem states the behavior of $\mathcal{U}$-updating algorithm in the large sample limit.

Theorem 2. *For the case of the exponential family, the $\mathcal{U}$-updating algorithm is equivalent to the EM algorithm in the limit of large samples.*

Proof: In the large sample limit, the sufficient statistic $\boldsymbol{s}(Y, X)$ will be insensitive to one hidden variable: $\boldsymbol{s}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) + \sum_{j=1,\neq i}^{n} \langle \boldsymbol{s}_j(\boldsymbol{y}_j, \boldsymbol{x}_j) \rangle \approx \langle \boldsymbol{s}(Y, X) \rangle$ as $n \to \infty$. Therefore, in the large sample limit, the reaction term $\beta_i(\boldsymbol{x}_i)$ becomes a constant and $Q_i(\boldsymbol{x}_i)$ of the $\mathcal{U}$-updating algorithm becomes the distribution resulted from the E-step of the EM algorithm:

$$Q_i(\boldsymbol{x}_i) = \frac{P(\boldsymbol{y}_i, \boldsymbol{x}_i \,|\, \widehat{\boldsymbol{\theta}}(\langle \boldsymbol{s} \rangle))}{\sum_{\boldsymbol{x}_i'} P(\boldsymbol{y}_i, \boldsymbol{x}_i' \,|\, \widehat{\boldsymbol{\theta}}(\langle \boldsymbol{s} \rangle))} = P(\boldsymbol{x}_i \,|\, \boldsymbol{y}_i, \widehat{\boldsymbol{\theta}}(\langle \boldsymbol{s} \rangle)) \,. \tag{19}$$

From the fact that $\widehat{\boldsymbol{\theta}}(\langle \boldsymbol{s} \rangle)$ gives the ML parameter in the M-step of the EM algorithm, the $\mathcal{U}$-updating algorithm is equivalent to the EM algorithm in the large sample limit. ■

5 Numerical Experiments

5.1 Mixture of Gaussians

For the p-dimensional observational vector $\boldsymbol{y}_i \in \mathrm{R}^p$, the mixture model [9, 10] of m components with parameter $\boldsymbol{\theta}$ is generally defined as $P(\boldsymbol{y}_i|\boldsymbol{\theta}) = \sum_{k=1}^{m} P(\boldsymbol{y}_i|x_i = k, \boldsymbol{\theta}) P(x_i = k|\boldsymbol{\theta})$, where $x_i \in \{k = 1, \ldots, m\}$ denotes the hidden variable indicating which mixture component is in charge of generating $\boldsymbol{y}_i$. The components are labelled by k. Although our method can be applied to an arbitrary mixture model, for simplicity, we consider the case of Gaussian components. In this case, the mixture model parameterized by $\boldsymbol{\theta} = (\{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}, \{w_k\})$ is given by $P(\boldsymbol{y}_i|\boldsymbol{\theta}) = \sum_{k=1}^{m} \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\, w_k$, where $w_k = P(x_i = k|\boldsymbol{\theta})$ is the mixing proportion satisfying $\sum_{k=1}^{m} w_k = 1$ and $\mathcal{N}(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = P(\boldsymbol{y}_i|x_i = k, \boldsymbol{\theta})$ denotes the kth Gaussian component distribution with the mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. The sampling distribution of the mixture of Gaussians is given by

$$P(\boldsymbol{y}_i, x_i|\boldsymbol{\theta}) = \prod_{k=1}^{m} \Big[\mathcal{N}(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\, w_k \Big]^{\delta_k(x_i)} , \tag{20}$$

where $\delta_k(x_i)$ denotes the Kronecker delta function given by $\delta_k(x_i) = 1$ for $x_i = k$ and $\delta_k(x_i) = 0$ for $x_i \neq k$.

Let (Y, X) denote the complete data set of n IID observations, where $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ and $X = \{x_1, \ldots, x_n\}$. Since the sampling distribution $P(\boldsymbol{y}_i, x_i|\boldsymbol{\theta})$ is in the exponential family, the complete data likelihood $P(Y, X|\boldsymbol{\theta})$ and the ML parameter estimator $\widehat{\boldsymbol{\theta}}(Y, X)$ become the function of the sufficient statistics.

Table 2. $\mathcal{U}$-updating Algorithm : Mixture of Gaussians

Initialize $\langle\gamma_k\rangle^{(0)} = \sum_{i=1}^{n}\langle\delta_k(x_i)\rangle^{(0)}$,
$\langle\boldsymbol{\xi}_k\rangle^{(0)} = \sum_{i=1}^{n}\langle\delta_k(x_i)\rangle^{(0)}\,\boldsymbol{y}_i$,
$\langle\boldsymbol{\lambda}_k\rangle^{(0)} = \sum_{i=1}^{n}\langle\delta_k(x_i)\rangle^{(0)}\,\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}$,
where $\langle\delta_k(x_i)\rangle^{(0)} = Q_i^{(0)}(x_i = k)$.
Set $\langle\gamma_k\rangle^{(1)} = \langle\gamma_k\rangle^{(0)}$, $\langle\boldsymbol{\xi}_k\rangle^{(1)} = \langle\boldsymbol{\xi}_k\rangle^{(0)}$ and $\langle\boldsymbol{\lambda}_k\rangle^{(1)} = \langle\boldsymbol{\lambda}_k\rangle^{(0)}$.
Repeat $t = 1, 2, 3, \ldots$ until convergence.
Repeat $i = 1, \ldots, n$.
1) Update $Q_i^t(x_i) \propto \prod_{k=1}^{m}\left(\bar{\gamma}_k^{(t)}(x_i)^{1+\frac{p}{2}}\,|\bar{\boldsymbol{C}}_k^{(t)}(x_i)|^{-\frac{1}{2}}\right)^{\bar{\gamma}_k^{(t)}(x_i)}$,
where $\bar{\gamma}_k^{(t)}(x_i) = \langle\gamma_k\rangle^{(t)} + [\,\delta_k(x_i) - \langle\delta_k(x_i)\rangle^{(t-1)}\,]$,
$\bar{\boldsymbol{\xi}}_k^{(t)}(x_i) = \langle\boldsymbol{\xi}_k\rangle^{(t)} + [\,\delta_k(x_i) - \langle\delta_k(x_i)\rangle^{(t-1)}\,]\,\boldsymbol{y}_i$,
$\bar{\boldsymbol{\lambda}}_k^{(t)}(x_i) = \langle\boldsymbol{\lambda}_k\rangle^{(t)} + [\,\delta_k(x_i) - \langle\delta_k(x_i)\rangle^{(t-1)}\,]\,\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}$,
$\bar{\boldsymbol{C}}_k^{(t)}(x_i) = \bar{\boldsymbol{\lambda}}_k^{(t)}(x_i) - \bar{\gamma}_k^{(t)}(x_i)^{-1}\,\bar{\boldsymbol{\xi}}_k^{(t)}(x_i)\bar{\boldsymbol{\xi}}_k^{(t)}(x_i)^{\mathrm{T}}$.
2) Refine sufficient statistics
$\langle\gamma_k\rangle^{(t)} \leftarrow \langle\gamma_k\rangle^{(t)} + [\,\langle\delta_k(x_i)\rangle^{(t)} - \langle\delta_k(x_i)\rangle^{(t-1)}\,]$,
$\langle\boldsymbol{\xi}_k\rangle^{(t)} \leftarrow \langle\boldsymbol{\xi}_k\rangle^{(t)} + [\,\langle\delta_k(x_i)\rangle^{(t)} - \langle\delta_k(x_i)\rangle^{(t-1)}\,]\,\boldsymbol{y}_i$,
$\langle\boldsymbol{\lambda}_k\rangle^{(t)} \leftarrow \langle\boldsymbol{\lambda}_k\rangle^{(t)} + [\,\langle\delta_k(x_i)\rangle^{(t)} - \langle\delta_k(x_i)\rangle^{(t-1)}\,]\,\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}$
with $\langle\delta_k(x_i)\rangle^{(t)} = Q_i^{(t)}(x_i)$.
End (Repeat)
Set $\langle\gamma_k\rangle^{(t+1)} = \langle\gamma_k\rangle^{(t)}$, $\langle\boldsymbol{\xi}_k\rangle^{(t+1)} = \langle\boldsymbol{\xi}_k\rangle^{(t)}$ and $\langle\boldsymbol{\lambda}_k\rangle^{(t+1)} = \langle\boldsymbol{\lambda}_k\rangle^{(t)}$.
End (Repeat)

Therefore, the $\mathcal{U}$-function is also a function of the sufficient statistics $\boldsymbol{s}(Y, X) = (\{\gamma_k, \boldsymbol{\xi}_k, \boldsymbol{\lambda}_k\})$:

$$\mathcal{U}\big(\{\gamma_k, \boldsymbol{\xi}_k, \boldsymbol{\lambda}_k\}\big) = c\prod_{k=1}^{m}\left(\gamma_k{}^{1+\frac{p}{2}}\,|\boldsymbol{C}_k|^{-\frac{1}{2}}\right)^{\gamma_k}, \tag{21}$$

where $\boldsymbol{C}_k = \boldsymbol{\lambda}_k - \gamma_k^{-1}\boldsymbol{\xi}_k\boldsymbol{\xi}_k^{\mathrm{T}}$ and

$$\gamma_k = \sum_{i=1}^{n}\delta_k(x_i), \quad \boldsymbol{\xi}_k = \sum_{i=1}^{n}\delta_k(x_i)\,\boldsymbol{y}_i, \quad \boldsymbol{\lambda}_k = \sum_{i=1}^{n}\delta_k(x_i)\,\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}, \tag{22}$$

and c is a constant. Using $\langle\delta_k(x_i)\rangle = Q_i(x_i = k)$, we present the $\mathcal{U}$-updating algorithm for the mixture of Gaussians in Table 2. We can simply obtain the estimate of the parameters by $\boldsymbol{\theta}^* = \widehat{\boldsymbol{\theta}}\big(\{\langle\gamma_k\rangle, \langle\boldsymbol{\xi}_k\rangle, \langle\boldsymbol{\lambda}_k\rangle\}\big)$ under all $Q_i(\boldsymbol{x}_i)$ resulted from the $\mathcal{U}$-updating algorithm like as the M-Step of the EM algorithm, where the ML parameter estimator is given by

$$\widehat{\boldsymbol{\theta}}\big(\{\gamma_k, \boldsymbol{\xi}_k, \boldsymbol{\lambda}_k\}\big) = \left(\left\{\widehat{w}_k = \frac{\gamma_k}{n}\,,\ \widehat{\boldsymbol{\mu}}_k = \frac{\boldsymbol{\xi}_k}{\gamma_k}\,,\ \widehat{\boldsymbol{\Sigma}}_k = \frac{\boldsymbol{C}_k}{\gamma_k}\right\}\right). \tag{23}$$

5.2 Numerical Results

In order to demonstrate $\mathcal{U}$-updating algorithm in comparison with the EM algorithm, we first used the data set of 800 data points generated from the mixture

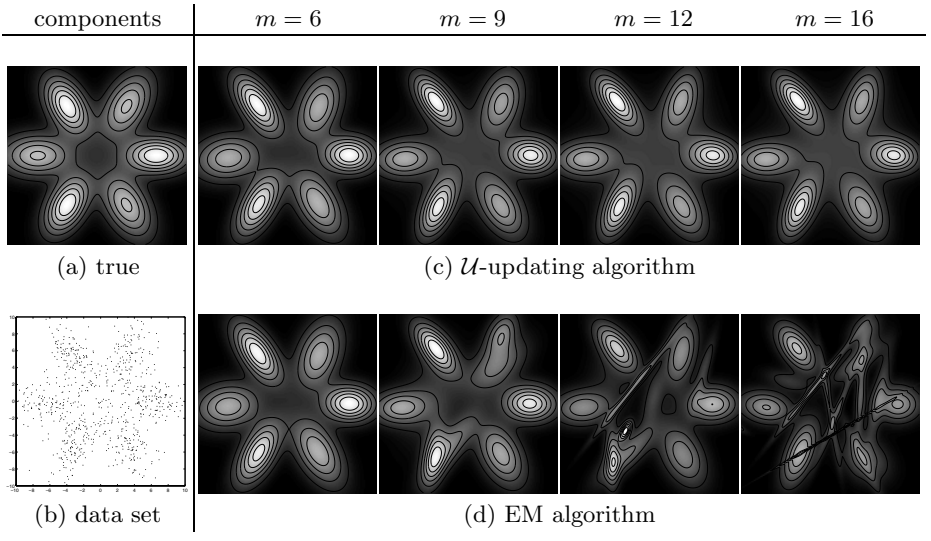


Fig. 2. Results on a mixture of 6 well-clustered Gaussian components: (a) true Gaussian-mixture distribution, where the more bright, the higher probability is there. (b) 800 data points generated from the true distribution. (c) and (d) learned distributions by $\mathcal{U}$-updating and EM algorithms when the models have the components $m = 6, 9, 12, 16$.

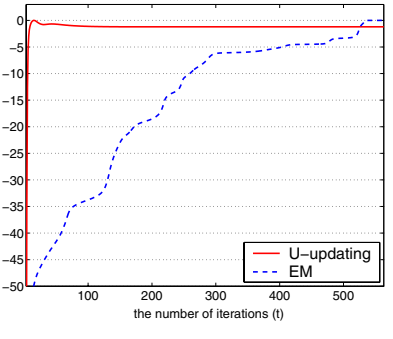


Fig. 3. Intermediate log likelihood values, subtracted from the maximum value, of the $\mathcal{U}$-updating and the EM algorithms in the case of $m = 16$ on data set shown in Figure 2

of 6 well-clustered Gaussian components having equal mixing proportion w_k but having different volume. Both algorithms started by the same initial guess from k-means algorithm. Figure 2 shows that the $\mathcal{U}$-updating algorithm alleviates the overfitting in comparison with the EM algorithm. Although models were more complicated than the true model ($m = 6$), the $\mathcal{U}$-updating algorithm demonstrated that all of the learned distributions ($m = 6, 9, 12, 16$) were very similar

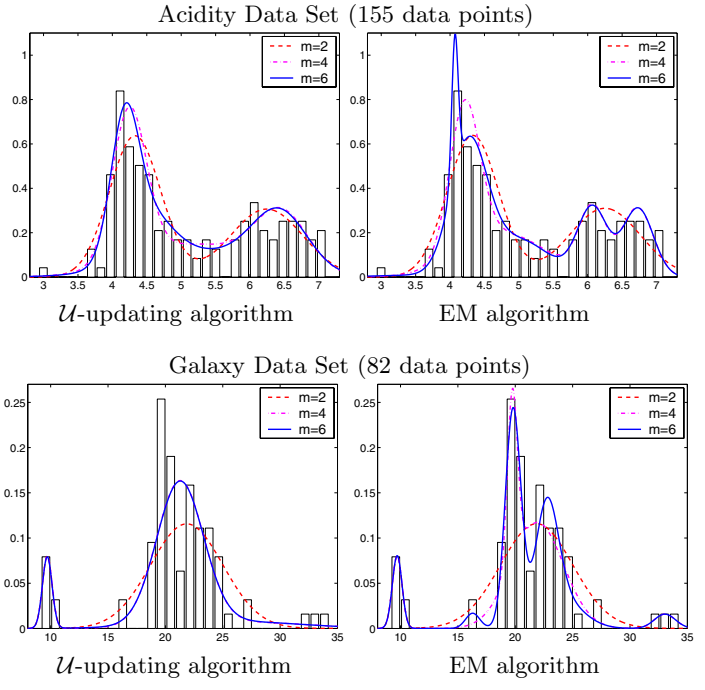


Fig. 4. Learned distributions by $\mathcal{U}$-updating and EM algorithms when the models have the components $m = 2, 4, 6$

to the true distribution. However, for the EM algorithm, the more complicated the model we considered, the more overfitted the distribution that resulted.

As a practical issue, overfitting leads to slow convergence. Figure 3 shows the convergence curves in term of log likelihood subtracted from the maximum value in the case of $m = 16$. By penalizing the likelihood term α_i by the reaction term β_i, the $\mathcal{U}$-updating algorithm achieved much faster convergence, approximately more than three times, than the EM algorithm. The $\mathcal{U}$-updating algorithm met the convergence threshold, that was $\sqrt{\sum_{k=1}^{m} |w_k^{(t)} - w_k^{(t-1)}|^2} < 10^{-4}$, after 153 iterations, whereas the EM algorithm met the same threshold after 563 iterations.

Next, we used real data sets, acidity and galaxy data sets shown in [10]. Figure 4 shows the learned distributions when the models have 2, 4, and 6 components and the Table 3 shows that the optimized mixing proportions $\hat{w}_k$ when the model has 6 components.

6 Conclusions

In this paper, we introduced the $\mathcal{U}$-likelihood approach for learning latent variable models, which differs from the Bayesian and the ML approaches. We presented some advantages of our approach over them in section 3. Our $\mathcal{U}$-likelihood method gives an alternative to Bayesian inference and the ML method when we

Table 3. Optimized mixing proportions ($\hat{w}_k$) of learned model having 6 components

Component	1	2	3	4	5	6
Acidity Data Set						
$\mathcal{U}$-updating algorithm	0.295	0.259	0.187	0.086	0.086	0.086
EM algorithm	0.386	0.188	0.171	0.169	0.073	0.013
Galaxy Data Set						
$\mathcal{U}$-updating algorithm	0.267	0.267	0.267	0.083	0.058	0.058
EM algorithm	0.403	0.277	0.171	0.085	0.037	0.026

don't want to use these. As a practical implementation, we presented the $\mathcal{U}$-updating algorithm to compute the distribution over hidden variables, which was found to penalize the likelihood by estimating the reaction of the other hidden variables and to alleviate the overfitting-accmulation problem of the EM algorithm.

We leave some of issues for the future work: 1) How can we more accurately approximate $Q(X)$ in (10) than the $\mathcal{U}$-updating algorithm. 2) How can we perform the model selection in the framework of the $\mathcal{U}$-likelihood. 3) Comparison with the Bayesian approach, e.g. EP and VB.

References

1. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, and A. Gelman. *Bayesian Data Analysis*. Chapman & Hall/CRC, 1995.
2. M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 72(2):183–233, 1999.
3. R. M. Neal. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993.
4. T. Minka. Expectation Propagation for approximate Bayesian inference. In *Proc. Uncertainty in Artificial Intelligence*, 2001.
5. Z. Ghahramani and M. J. Beal. Propagation algorithms for variational bayesian learning. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001.
6. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.
7. R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1988.
8. M. Opper and D. Saad, editors. *Advanced Mean Field Methods : Theory and Practice*. MIT Press, 2001.
9. G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley-Interscience, 2000.
10. S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B*, 59:731–792, 1997.