EM algorithm and its applications for mixture models

by

Zsolt Szabó

Submitted to

Central European University

Department of Mathematics and its Applications

In partial fulfilment of the requirements for the degree of

Master of Science

Supervisor: Marianna Bolla

Budapest, Hungary

2012

Table of Contents

Introduction

In statistics it is important to estimate the parameters of a data set if the family of the background distribution is known. The easiest way to do so is the well known Maximum Likelihood method. However, sometimes some part of the data is not known or the likelihood function cannot be maximized explicitly. In this case one needs to use iterative algorithms. One of the methods is the Expectation-Maximization (EM) algorithm.

The EM algorithm, first published by Dempster, Rubin and Laird [3], is an iterative algorithm scheme that is applicable for several purposes in statistics and also in other sciences, like computational neuroscience. This is not a proper algorithm in terms of the computer science, since it does not describe the exact steps of the algorithm. Rather it is a scheme, though, we will refer to it as an algorithm. The implementations vary widely.

The main advantage of the algorithm is the fact that it converges in a wide variety of cases. The disadvantages are that it may converge to a local maximum and the rate of convergence could be slow. In practice, in case of not properly chosen initial parameters, it converges to a wrong solution.

Briefly, the setting of the algorithm is the following. Let $\mathcal{Y}$ be the space of the observed data and $\mathcal{X}$ be that of the directly not observed data. We assume a parametric distribution $f(\cdot|\theta)$ on the complete space where θ is a parameter. Choose an arbitrary $\theta^{(0)}$ in the parameter space as an initialization. Then the iterative steps of the algorithm are as follows.

E-step Given $\theta^{(m)}$ calculate $Q(\theta|\theta^{(m)}) = E(\log f(\mathbf{X}|\mathbf{Y}, \theta)|\theta^{(m)})$

M-step Let $\theta^{(m+1)}$ be such that maximizes $Q(\theta|\theta^{(m)})$.

In the thesis we describe the EM algorithm. First, we examine it in the

exponential families of distributions as a special case. Then we present a generalization of the algorithm (GEM) where E-step cannot be done easily, based on [3].

An application of the EM algorithm that we present is the mixture models. In mixture models a set of data points is given, the points are coming from different distributions randomly. Our task is to find the parameters of the distributions of the components if the parametric family of distributions and the number of different components are known.

We apply the EM algorithm for mixtures of normally distributed random variables. We would like to estimate the parameters of each normal and the mixing multinomial distribution. We implement the algorithm in Octave and run simulations on data produced by random number generator. We examine how the results depend on the initialization.

A new application of the EM algorithm is clustering of generalized random graphs. The model is the following. Consider n vertices of a graph. We put each of them independently to one of the k clusters with multinomial distribution. Then consider each pair of vertices. If one of the vertices is from the ith cluster and the other is from the jth cluster then there is an edge between them with probability p_{ij}, independently $(i, j = 1, \ldots, k)$. However, we do not know the cluster memberships. We only know the adjacency matrix of a given graph as an incomplete data specification. The algorithm is applied to get the classification of the vertices.

The structure of the thesis is the following. In Chapter 2 we present the theory of the EM algorithm. Then in Chapter 3 we show directions to implement the algorithm for mixture models, and implement the algorithm for Gaussian mixtures. Then in Chapter 4 we present the model of generalized random graphs and implement the EM algorithm for clustering them.

Chapter 2

The EM algorithm and its theoretical background

In this chapter after introducing the basic statistical setting, we give the theoretical steps of the EM in case of exponential families in Section 2.1. Then in Section 2.2 we describe the algorithm more generally and give some properties of it regarding the convergence.

The Expectation Maximization algorithm is an algorithm scheme that is used to find the maximum likelihood estimates of parameters of a data given a model that the data comes from. It is used in those cases that something is missing according to a model or for some reasons the usual maximum likelihood estimations cannot be performed.

First of all, we have two sample spaces. Let $\mathcal{Y}$ be the space of observed data. Then we choose a model for the specific problem. The model specifies the complete data space $\mathcal{X}$. We define a mapping from $\mathcal{X}$ to $\mathcal{Y}$, in most of the cases this maps several points from $\mathcal{X}$ to the same $\mathbf{y} \in \mathcal{Y}$. Let us denote this mapping by $\mathbf{x} \to \mathbf{y}(\mathbf{x})$. Conversely, $\mathcal{X}(\mathbf{y}) = \{\mathbf{x} : \mathcal{Y}(\mathbf{x}) = \mathbf{y}\}$.

The chosen model should contain a parametric family of distributions $f(\cdot|\theta)$ for the complete data $\mathbf{X}$ and the parameter space Θ. The supposed distribution

of $\mathbf{Y}$ is given by the integral

$$g(\mathbf{y}|\theta) = \int_{\mathcal{X}(\mathbf{y})} f(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Our task is to find the parameter θ which fits the data the best. So we would like to maximize the loglikelihood function,

$$L(\theta) = \log g(\mathbf{y}|\theta). \tag{2.1}$$

To do this, first, we initialize the algorithm by some $\theta^{(0)} \in \Theta$. Then by the iteration the estimate approaches the true parameter value θ^*. In Section 2.1 we describe the iteration steps of the algorithm for the special case of distributions of exponential family.

2.1 The EM algorithm for exponential families

First we introduce the EM algorithm in case of distributions belonging to the exponential family. The reason for doing this is the simplicity of the description and the ease of computation in this case. We define the exponential families of distributions as in [6].

Definition 1 (Exponential Family)**.** *Let $f(\cdot|\theta)$ be a parametric density function, with parameter $\theta = (\theta_1, \theta_2, \dots \theta_k)^T$. We say that $f(\cdot|\theta)$ is of exponential family, if it can be written as*

$$f(\mathbf{x}|\theta) = \frac{b(\mathbf{x})e^{\mathbf{t}^T(\mathbf{x})\mathbf{h}(\theta)}}{a(\theta)},$$

where $\mathbf{t}(\mathbf{x})$ is a q-dimensional (coloumn) function of $\mathbf{x}$, h is a q-dimensional function of θ, $a(\theta)$ is the normalizing factor

$$a(\theta) = \int_{\mathcal{X}} b(\mathbf{x})e^{\mathbf{t}^T(\mathbf{x})\mathbf{h}(\theta)}.$$

If $\mathbf{h}$ is the identity function, we say that θ is the natural parameter of the distribution f.

From now on, we consider only natural parameters without loss of generality.

Now we can describe the EM algorithm for exponential families. First, choose an arbitrary $\theta^{(0)} \in \Theta$ as initialization of the algorithm. Then the iterative steps of the algorithm are as follows.

Let us suppose that after m steps of the algorithm an estimation for the parameter θ is given as $\theta^{(m)}$. Then the $(m+1)$th iteration of the algorithm is the following.

E-step: If $\theta^{(m)}$ is given, let $\mathbf{t}^{(m+1)} = E(\mathbf{t}|\mathbf{Y}\theta^{(m)})$.

M-step: Solve the equation $E(\mathbf{t}(\mathbf{X})|\theta) = \mathbf{t}^{(m+1)}$ for θ, let $\theta^{(m+1)}$ be the solution. Here, E stands for the conditional expectation, M stands for maximization of the likelihood function. Note that in case of exponential distributions the M-step is equivalent to performing a maximum likelihood estimate on the parameter θ given the statistic $t^{(m)}$. To see that, we use the notation of Definition 1. Differentiating by components,

$$\frac{\partial \ln(f(\mathbf{x}|\theta))}{\partial \theta} = \mathbf{t}(\mathbf{x}) - \frac{\partial \ln(a(\theta))}{\partial \theta} = \mathbf{0},$$

where the second term is:

$$\begin{aligned} \frac{\partial \ln(a(\theta))}{\partial \theta} &= \frac{1}{a(\theta)} \frac{\partial a(\theta)}{\partial \theta} = \frac{1}{a(\theta)} \int_{\mathcal{X}} \frac{\partial}{\partial \theta} b(\mathbf{x}) e^{\mathbf{t}^T(\mathbf{x})\theta} \mathrm{d}(\mathbf{x}) = \\ &= \frac{1}{a(\theta)} \int_{\mathcal{X}} \mathbf{t}(\mathbf{x}) b(\mathbf{x}) e^{\mathbf{t}^T(\mathbf{x})\theta} \mathrm{d}(\mathbf{x}) = E(\mathbf{t}(\mathbf{x})|\theta). \end{aligned} \tag{2.2}$$

So the solution for the equation

$$E(\mathbf{t}(\mathbf{X})|\theta) = \mathbf{t}^{(m+1)} \tag{2.3}$$

is the ML estimate for θ. Note that in the M-step we maximize $\ln f(\mathbf{x}|\theta) = \ln b(\mathbf{x}) + \mathbf{t}^T(\mathbf{x})\theta - \ln a(\theta)$, which is equivalent to maximize $\mathbf{t}^T(\mathbf{x})\theta - \ln a(\theta)$. Now we would like to show that by repeated steps of the iteration, θ tends to the place of the maximum of (2.1).

We introduce some notations. The conditional density of $\mathbf{X}$ given $\mathbf{Y}$ and θ is

$$k(\mathbf{x}|\mathbf{y}, \theta) = \frac{f(\mathbf{x}|\theta)}{g(\mathbf{y}|\theta)},$$

so $L(\theta)$ can be written as

$$L(\theta) = \ln f(\mathbf{x}|\theta) - \ln k(\mathbf{x}|\mathbf{y},\theta). \tag{2.4}$$

$k(\mathbf{x}|\mathbf{y},\theta)$ can be computed by the definition of g:

$$\begin{aligned} k(\mathbf{x}|\mathbf{y},\theta) &= \frac{f(\mathbf{x}|\theta)}{\int_{\mathcal{X}(\mathbf{y})} f(\mathbf{x}|\theta)\mathrm{d}\mathbf{x}} = \frac{\frac{b(\mathbf{x})\exp(\mathbf{t}^T(\mathbf{x})\theta)}{a(\theta)}}{\frac{\int_{\mathcal{X}(\mathbf{y})} b(\mathbf{x})\exp(\mathbf{t}^T(\mathbf{x})\theta)\mathrm{d}\mathbf{x}}{a(\theta)}} = \\ &= \frac{b(\mathbf{x})\exp(\mathbf{t}^T(\mathbf{x})\theta)}{\int_{\mathcal{X}(\mathbf{y})} b(\mathbf{x})\exp(\mathbf{t}^T(\mathbf{x})\theta)\mathrm{d}\mathbf{x}} \end{aligned}$$

Therefore the conditional density is also of exponential form. We introduce the notation $a(\theta|\mathbf{y}) = \int_{\mathcal{X}(\mathbf{y})} b(\mathbf{x})\exp(\mathbf{t}^T(\mathbf{x})\theta)\mathrm{d}\mathbf{x}$.

By differentiating $\ln a(\theta|\mathbf{y})$ similarly to (2.2) we get

$$\begin{aligned} \frac{\partial \ln(a(\theta|\mathbf{y}))}{\partial\theta} &= \frac{1}{a(\theta|\mathbf{y})}\frac{\partial a(\theta|\mathbf{y})}{\partial\theta} = \frac{1}{a(\theta|\mathbf{y})}\int_{\mathcal{X}} \frac{\partial}{\partial\theta} b(\mathbf{x}) e^{\mathbf{t}^T(\mathbf{x})\theta}\mathrm{d}(\mathbf{x}) = \\ &= \frac{1}{a(\theta|\mathbf{y})}\int_{\mathcal{X}(\mathbf{y})} \mathbf{t}(\mathbf{x}) b(\mathbf{x}) e^{\mathbf{t}^T(\mathbf{x})\theta}\mathrm{d}(\mathbf{x}) = E(\mathbf{t}(\mathbf{x})|\mathbf{y},\theta). \end{aligned} \tag{2.5}$$

Using (2.4) we get

$$L(\theta) = -\ln(a(\theta)) + \mathbf{b}(\mathbf{x})\theta - (\mathbf{b}(\mathbf{x})\theta - \ln a(\theta|y)) = -\ln a(\theta) + \ln a(\theta|\mathbf{y}). \tag{2.6}$$

Differentiating (2.6) componentwise and using (2.2) and (2.5)

$$\frac{\partial L(\theta)}{\partial\theta} = -E(\mathbf{t}(\mathbf{x})|\theta) + E(\mathbf{t}(\mathbf{x})|\mathbf{y},\theta).$$

So if L has a global maximum at θ' in the interior of Θ, then

$$0 = -E(\mathbf{t}(\mathbf{x})|\theta') + E(\mathbf{t}(\mathbf{x})|\mathbf{y},\theta'). \tag{2.7}$$

So if $\theta^{(m)}$ converges to a maximum point θ^* then θ^* is a fixed point of the algorithm. Therefore, by the E-step, $\mathbf{t}^* = E(\mathbf{t}(\mathbf{x})|\mathbf{Y},\theta^*)$. On the other hand, by

(2.3) $\mathbf{t}^* = E(\mathbf{t}(\mathbf{x})|\theta^*)$ in the M-step. By this two, we have

$$E(\mathbf{t}(\mathbf{x})|\mathbf{y}, \theta^*) = E(\mathbf{t}(\mathbf{x})|\theta^*),$$

which is exactly the solution of the equation (2.7).

2.2 EM in general

In this section we introduce the EM in a general setting of which EM is a special case.

We describe the EM algorithm for an arbitrary distribution and also relax the maximization step. Now we have a parametric distribution function $f(\cdot|\theta)$ as above. We define a bivariate function Q of θ and θ':

$$Q(\theta'|\theta) = E(\log(f(\mathbf{x}|\theta'))|\mathbf{y}, \theta)$$

which is assumed to exist for all pairs $(\theta', \theta) \in \Theta \times \Theta$, and the log function denotes the logarithm of base 2.

The iteration $\theta^{(m)} \to \theta^{(m+1)}$ described in [3] is the following.

E-step Compute the function $Q(\theta|\theta^{(m)})$.

M-step Choose $\theta^{(m+1)} \in \Theta$ to maximize $Q(\theta|\theta^{(m)})$.

Remarks. In case of exponential families

$$Q(\theta|\theta^{(m)}) = E(\log(f(x|\theta))|Y, \theta^{(m)}) =$$
$$= E(\log(b(x))|y, \theta^{(m)}) + E(\mathbf{t}^T(x)\theta|y, \theta^{(m)}) - E(\log(a(\theta))|y, \theta^{(m)}) \qquad (2.8)$$

In order to maximize (2.8) in θ we can omit the first term. Because of the conditional expectation and θ being independent of y and $\theta^{(m)}$, the second term can be written as $E(t^T(x)\theta|y, \theta^{(m)}) = t^{T(m)}\theta$. In the third term $a(\theta)$ does not depend on the conditions therefore we can simply write $E(\log(a(\theta))|y, \theta^{(m)}) = \log a(\theta)$. Therefore, $t^{T(m)}\theta - \log a(\theta)$ is to be maximized as in M-step of EM for exponential families.

In practice, it is not convenient to maximize $Q(\theta|\theta^{(m)})$ in every step for every

θ, sometimes it is not even possible. So Dempster et al. in [3] defined an iteration which can be used in an easier way.

Definition 2. *We assume everything as in the description of EM algorithm. Let us have a mapping $M : \Theta \to \Theta$, $\theta \mapsto M(\theta)$. The mapping M defines a Generalized EM (GEM) algorithm if*

$$Q(M(\theta)|\theta) \geq Q(\theta|\theta)$$

for every $\theta \in \Theta$.

The GEM algorithm increases the value of Q instead of maximizing it. The convergence property does not change.

Definition 3 (Kullback–Leibler divergence)**.** *Let $(\mathcal{X}, \mathcal{F}, P)$ be a probability space and let pdf's f and g be defined on it. The Kullback–Leibler divergence (or relative entropy) D_{KL} of f and g is*

$$D_{KL}(f||g) = \int_{\mathcal{X}} \log \frac{f(x)}{g(x)} f(x) \mathrm{d}x.$$

First, we introduce the notation

$$H(\theta'|\theta) = E(\log k(\mathbf{x}|\mathbf{y}, \theta')|\mathbf{y}, \theta),$$

which is similar to the K-L divergence.

The following easy lemma is important to prove the convergence of the GEM.

Lemma 1. *The Kullback–Leibler divergence is nonnegative and equality occurs if and only if $f = g$ almost everywhere:*

$$D_{KL}(f||g) \geq 0. \tag{2.9}$$

Proof. By basic analysis we know that

$$\log x \leq x - 1,$$

with equality if and only if $x = 1$. We substitute x by $\frac{f(x)}{g(x)}$ we get

$$-(f(x)\log\frac{f(x)}{g(x)}) = f(x)\log\frac{g(x)}{f(x)} \leq f(x)(\frac{g(x)}{f(x)} - 1) = g(x) - f(x). \quad (2.10)$$

Multiplying both side by -1 and integrating over $\mathcal{X}$ we get the inequality.

$$\int_{\mathcal{X}} f(x)\log\frac{f(x)}{g(x)} \geq \int_{\mathcal{X}} (f(x) - g(x))\mathrm{d}x = 0.$$

In the equality, the only if part is easy as $f(x) = g(x)$ almost everywhere then in the inequality 2.10, equality occurs for almost every $x \in \mathcal{X}$. For the other direction, we prove it by contradiction. If $f(x) \neq g(x)$ almost everywhere, then $f(x)\log\frac{f(x)}{g(x)} > 0$ on a set whose measure is larger than 0. Therefore the integral of that nonnegative function is strictly larger than 0. □

Corollary 1.

$$H(\theta'|\theta) \leq H(\theta|\theta) \quad (2.11)$$

Proof. By Lemma 1 we can see that

$$H(\theta|\theta) - H(\theta'|\theta) = E(\log k(\mathbf{x}|\mathbf{y},\theta)|\mathbf{y},\theta) - E(\log k(\mathbf{x}|\mathbf{y},\theta')|\mathbf{y},\theta)$$
$$E(\log k(\mathbf{x}|\mathbf{y},\theta) - \log k(\mathbf{x}|\mathbf{y},\theta')|\mathbf{y},\theta) = E(\log\frac{k(\mathbf{x}|\mathbf{y},\theta)}{k(\mathbf{x}|\mathbf{y},\theta')}|\mathbf{y},\theta).$$

The last expected value is the K–L divergence between $k(\mathbf{x}|\mathbf{y},\theta)$ and $k(\mathbf{x}|\mathbf{y},\theta')$ thus is positive. □

Another important property of $H(\theta'|\theta)$ is the following:

$$Q(\theta'|\theta) - H(\theta'|\theta) = L(\theta'). \quad (2.12)$$

To show (2.12) we expand both Q and H.

$$Q(\theta'|\theta) - H(\theta'|\theta) = E(\log f(\mathbf{x}|\theta')|\mathbf{y},\theta) - E(\log k(\mathbf{x}|\mathbf{y},\theta')|\mathbf{y},\theta)$$
$$= E(\log f(\mathbf{x}|\theta')|\mathbf{y},\theta) - E(\log f(\mathbf{x}|\theta') - \log g(\mathbf{y}|\theta')|\mathbf{y},\theta) = E(\log g(\mathbf{y}|\theta')|\mathbf{y},\theta),$$

since $g(\mathbf{y}|\theta')$ is measurable with respect to $\mathbf{y}$, we can simply write $E(g(\mathbf{y}|\theta'))$, which is $L(\theta')$ by definition.

The main advantage of GEM to other iterative algorithms that the convergence is proved for a wide range of settings. In particular, GEM increases the likelihood and the maximum point is a fixed point of the mapping. This is discussed by the following theorem by Dempster et al.

Theorem 2. *For every GEM algorithm*

$$L(M(\theta)) \geq L(\theta) \qquad \forall \theta \in \Theta,$$

with equality if and only if the following two equations hold,

$$Q(M(\theta)|\theta) = Q(\theta|\theta)$$

and

$$k(\mathbf{x}|\mathbf{y}, M(\theta)) = k(\mathbf{x}|\mathbf{y}, \theta)$$

almost everywhere.

Proof. We would like to show that $L(M(\theta)) - L(\theta) \geq 0$. To do so we expand it as in (2.4),

$$\begin{aligned} L(M(\theta)) - L(\theta) &= Q(M(\theta)|\theta) - H(M(\theta)|\theta) - (Q(\theta|\theta) - H(\theta|\theta)) = \\ &= [Q(M(\theta)|\theta) - Q(\theta|\theta)] + [H(\theta|\theta)) - H(M(\theta)|\theta)] \end{aligned}$$

The first term is nonnegative by the definition of GEM. The second term is larger than 0, by Lemma 1 with equality if and only if $k(\mathbf{x}|\mathbf{y}, M(\theta)) = k(\mathbf{x}|\mathbf{y}, \theta)$. □

The following statements are obvious consequences of Theorem 2

Corollary 2. *Suppose for some $\theta^* \in \Theta$, $L(\theta^*) \geq L(\theta)$ for all $\theta \in \Theta$, then for every GEM,*

1. $L(M(\theta^*)) = L(\theta^*)$,
2. $Q(M(\theta^*)|\theta^*) = Q(\theta^*|\theta^*)$,

3. $k(\mathbf{x}|\mathbf{y}, M(\theta^*)) = k(\mathbf{x}|\mathbf{y}, \theta^*)$ *almost everywhere.*

Proof. Statement 1 is trivial as θ^* maximizes L. Statement 2 and 3 are immediate consequences of the Theorem 2, because these are the two conditions for attaining equality.

□

Moreover, if there is a unique global maximum, the following is also true.

Corollary 3. *If there is a global maximum* θ^* *of* $L(\theta)$*, then for every GEM algorithm*

$$M(\theta^*) = \theta^*.$$

Proof. An iteration step cannot decrease the value of L, which is bounded as it has a global maximum at θ^* □

We present further properties of the EM algorithm. Wu in [7] proved the convergence properties of the GEM algorithm. That article contains some further general assumptions, namely, $\Theta \subset \mathbb{R}$, L is continuous in Θ and differentiable in the interior of Θ, and the set $\Theta_{\theta_0} = \{\theta \in \Theta : L(\theta) \geq L(\theta_0)\}$ is compact for any θ_0 such that $L(\theta_0) > -\infty$.

Lemma 3. *Having the three assumption, any sequence* $\{L(\theta^{(m)})\}_{m\geq 0}$ *defined by a GEM is bounded from above.*

Proof. By the compactness, and Theorem 2, $\{\theta^{(m)})\}_{m\geq 0}$ is in a compact subset of Θ. Therefore, we can regard L as a continuous function on a compact subset of $\mathbb{R}$. Thus, it has a maximum. □

By Lemma 3 and Theorem 2, for every sequence $\theta^{(m)}$ given by a GEM, $L(\theta^{(m)})$ converges.

Wu gives a definition of point-to-set maps.

Definition 4. *Let* X *be a set, a mapping from* X *to the subsets of* X *is called a* point-to-set map*. It is said to be* closed *at a point* x *if* $x_k \to x$ *and* $y_k \to y$*,* $y_k \in A(x_k)$ *then* $y \in A(x)$*.*

In case of point-to-set maps continuity implies closedness.

So we can consider that in a GEM, the iterative step M is a point-to-set map which maps to the subset of *Theta* for which L is larger. And we can choose an arbitrary point in order to get a sequence of a GEM. Wu's Theorem 1 is based on Zangwill's Global convergence theorem.

Theorem 4. *Let the sequence* $\{x_k\}$ *be generated by* $x_{k+1} \in M(x_k)$*, where* M *is a point-to-set map on* X*. Let a solution set* $\Gamma \subset X$ *be given and suppose that:*

1. *all points* x_k *are contained in a compact set* $S \subset X$*;*
2. M *is closed over the complement of* Γ*:*
3. *there is a continuous function* α *on* X *such that if* $x \notin \Gamma$*, then* $\alpha(y) > \alpha(x)$ *for all* $y \in \Gamma$ *and if* $x \in \Gamma$*, then* $\alpha(y) > \alpha(x)$ *for all* $y \in M(x)$

Then all the limit points of $\{x_k\}$ *are in the solution set* Γ *and* $a(x_k)$ *converges monotonically to* $\alpha(x)$ *for some* $x \in \Gamma$*.*

Wu considers the following two subsets. Let M be a point-to-set map in a GEM iteration and α is the loglikelihood L here. The solution sets are

$$\mathcal{M} = \{x : x \text{ is a local maximum in the interior of } \Theta\}$$

and

$$\mathcal{I} = \{x : x \text{ is a stationary point in the interior of } \Theta\}.$$

Theorem 5. *Let* $\theta^{(m)}$ *be a sequence generated by a GEM,* $\theta^{(m+1)} \in M(\theta^{(m)})$ *and suppose that (i)* M *is a closed point-to-set map over the complement of* $\mathcal{I}$ *(or* $\mathcal{M}$*), (ii)* $L(\theta^{(m+1)}) > L(\theta^{(m)})$ *for all* $\theta^{(m)} \notin \mathcal{I}$ *(or* $\mathcal{M}$*). Then all the limit points of* $\{\theta^{(m)}\}$ *are stationary points (or local maxima) of* L*, and* $\{L(\theta^{(m)})\}$ *converges monotonically to* $L^* = L(\theta^*)$ *for some* $\theta^* \in \mathcal{I}$ *(or* $\mathcal{M}$*).*

So under some mild conditions on the L function, the EM algorithm converges to a local maximum.

There is an information theoretic result given by Csiszár and Shields in [2]. This only applies when $\mathcal{X} = A^n$ and $\mathcal{Y} = B^n$, where $|A|$ and$|B|$ are finite. They introduce a special case of KL divergence for these finite distributions, the I-divergence, denoted by $D(P||Q)$. The description of steps are as

follows. First, choose a distribution $Q^{(0)}$ on the complete data space A. In the $(m+1)$th iteration in the E-step, by the conditional expectation of the complete data, we can calculate an empirical distribution $\hat{P}^{(m+1)}$ on the complete data. Then by the M-step we perform a usual ML estimate on the data. Since the ML estimate minimizes the KL divergence between the empirical distribution, and the possible distributions in the parametric family, in the M-step $D(P^{(m+1)}||Q^{(m)}) \geq D(P^{(m+1)}||Q^{(m+1)})$. Moreover, they prove in the E-step, that $D(P^{(m+1)}||Q^{(m)}) \geq D(P^{(m)}||Q^{(m)})$ So we get a non-increasing sequence of divergences, hence a convergent sequence. Let us denote the empirical distribution of the incomplete data by P^T and the distribution of the incomplete data by $Q^{(m)T}$ if the distribution of the complete data $Q^{(m)}$ is given. If the possible distributions $\mathcal{Q}$ is compact and convex then $Q^{(m)}$ converges to a minimizer distribution of $D(P^T||Q^T)$, that is, again by the divergence minimization property of the ML estimator, the ML estimator of the incomplete data.

So in the finite case, the conditions for converging to the true ML estimator of the incomplete data are easy to check.

Chapter 3

Mixtures

The EM algorithm is a good tool for missing data analysis. Here not necessarily the data is missing but there may be latent variables. A typical missing data setting is when there are a number of hidden variables. Several models use hidden variables in a wide variety of fields. One specific class of such models are finite mixtures. Mixtures are studied by a number of articles for example by Fraley and Raferty [4] and by Celeux and Goavert [1].

3.1 Finite mixtures

Mixture models in general can be described as follows. We have a set of observations, the data $\mathbf{Y} = (\mathbf{Y}_1, \dots \mathbf{Y}_n)$. It is known that each data point comes from a distribution F_i $(i = 1, \dots, k)$, with pdf f_i, where f_i's are from a parametric family of distributions.

The number of categories k is assumed to be known and usually much smaller than the number of data points. We also assume a parametric family of distributions for f_i. Our task is to infer the parameters of f_i and say something about the distribution of a specific data point.

In the most simple model the mixing variable is from a multinomial distribution. So in the model, for each data point Y_i belongs a hidden variable Δ_i. For the simplicity of notation, Δ_i is a random vector, with zero elements ex-

cept for one entry, which is 1. We denote the jth entry of Δ_i by Δ_{ij}. This is the indicator that the sample point Y_i comes from the distribution f_i. Let $\theta = (\phi_1, \dots \phi_k, \pi_1, \dots, \pi_k)$, where ϕ_i's correspond to the distributions f_i, and π_i are mixing parameters.

The likelihood of the model is given by

$$g(\mathbf{y}|\theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j f_j(y_i),$$

while the likelihood of the complete data is given by

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} \Delta_{ij} \pi_j f_j(y_i).$$

In the complete likelihood, the sum is only for convenience of the notation, it contains only one nonzero summand. So by taking the logarithm we get

$$\log f(\mathbf{x}|\theta) = \sum_{i=1}^{n} \sum_{j=1}^{k} \Delta_{ij} \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{k} \Delta_{ij} \log f_j(y_i). \tag{3.1}$$

So the EM algorithm for mixtures in the $(m+1)$th step is the following.

Initialization: choose an arbitrary $\theta^{(0)} \in \Theta$.

E-step: Calculate the function $Q(\theta|\theta^{(m)}) = E(\log f(\mathbf{X}|\theta)|\mathbf{Y}, \theta^{(m)})$

M-step: Maximize $Q(\theta|\theta^{(m)})$ in θ, and let it be $\theta^{(m+1)}$.

Remark: The choice of the initial value of θ can lead to false results during the iteration. So the initialization of the parameters should be somewhat realistic. For example, in case of normal variables we have to choose different mean to each initial distribution as we want to distinguish between them. Looking at the histogram of the data can give us a hint.

By (3.1) the E-step is to calculate

$$E(\sum_{i=1}^{n} \sum_{j=1}^{k} \Delta_{ij} \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{k} \Delta_{ij} \ln f_j(Y_i|\theta)|\mathbf{Y}, \theta^{(m)}).$$

It is enough to calculate $E(\Delta_{ij}|\mathbf{Y}, \theta^{(m)})$ since the function in the expectation only

depends on the random variables Δ_i. By Bayes therorem

$$E(\Delta_{i,j}|Y_i, \theta^{(m)}) = P(\Delta_{i,j} = 1|Y_i = y_i, \theta^{(m)}) \tag{3.2}$$

$$= \frac{f_j(Y_i)P(\Delta_{i,j} = 1|\theta^{(m)})}{\sum_{l=1}^{k} f_l(Y_i)P(\Delta_{i,l} = 1|\theta^{(m)})} = \frac{f_j(Y_i)\pi_j^{(m)}}{\sum_{l=1}^{k} f_l(Y_i|\theta)\pi_l^{(m)}}. \tag{3.3}$$

Note that

$$\sum_{j=1}^{k} \pi_{ij}^{(m)} = 1 \tag{3.4}$$

for all $i = 1 \dots n$. Then the function Q is the following

$$Q(\theta|\theta^{(m)}) = \sum_{i=1}^{n}\sum_{j=1}^{k} \pi_{ij}^{(m+1)} \log \pi_j + \sum_{i=1}^{n}\sum_{j=1}^{k} \pi_{ij}^{(m+1)} \log f_j(Y_i|\phi_j). \tag{3.5}$$

We should maximize it with respect to π_j and ϕ_j. Since π_j appears only in the first sum, ϕ_j in the second sum, we can maximize them separately. In this generality we can only maximize the first sum. The second sum has to be maximized on a case-by-case basis.

So the first sum is

$$\sum_{i=1}^{n}\sum_{j=1}^{k} \pi_{ij}^{(m+1)} \log \pi_j$$

with constraint $\sum_{j=1}^{k} \pi_j = 1$, which can be maximized by Lagrange multiplier. Introducing the Lagrange function

$$L(\pi_1, \dots, \pi_k, \lambda) = \sum_{i=1}^{n}\sum_{j=1}^{k} \pi_{ij}^{(m+1)} \log \pi_j + \lambda(\sum_{j=1}^{k} \pi_j - 1),$$

the partial derivatives are

$$\frac{\partial L(\pi_1, \dots, \pi_k, \lambda)}{\partial \pi_l} = \frac{\sum_{i=1}^{n} pi_{il}^{(m)}}{\pi_l} - \lambda$$

and

$$\frac{\partial L(\pi_1, \dots, \pi_k, \lambda)}{\partial \lambda} = \sum_{j=1}^{k} \pi_j - 1$$

The first equation to be zero we get

$$\lambda = \frac{\sum_{i=1}^{n} \pi_{il}^{(m)}}{\pi_l}$$

Therefore the left hand side must be the same for all l. So the solution is

$$\pi_l = \frac{\sum_{i=1}^{n} \pi_{il}^{(m)}}{\sum_{i=1}^{n} \sum_{j=1}^{k} \pi_{ij}^{(m)}} = \frac{\sum_{i=1}^{n} \pi_{il}^{(m)}}{n}, \tag{3.6}$$

with last step by (3.4).

3.2 Normal mixtures

Two-component mixture of normals are studied by Hastie, Tibshirani, Friedman in [5], where they give the algorithm for the mixture of two normals without calculations. In case of normal mixtures we have a data of mixture of k normally distributed random variables. The distributions are $\mathcal{N}(\mu_j, \sigma_j^2)$.

The likelihood function of the observed data in case of univariate normals is

$$L_\theta(\mathbf{Y}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(Y_i - \mu_j)^2}{2\sigma_j^2}}.$$

The parameter set in this case is $\theta = (\mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2, \pi_1, \ldots, \pi_k)$. The incomplete data is $\mathbf{Y}$, the data points we get. The complete data is the $(\Delta, \mathbf{Y})$ pair, where Δ contains the row vectors Δ_i, where Δ_i's are the hidden variables.

The initialization of the algorithm is to choose a $\theta^{(0)}$. For this case, we need to choose a mixing parameter set $\pi^{(0)}$ and for each group a parameter set $\mu_i^{(0)}$ and a $\sigma_i^{2(0)}$ for the given normal distribution.

Now we describe the iterative steps of the algorithm in details. The E-step is as it is described in (3.5).

By the arguments of Section 3.1 we only have to find the responsibilities as conditional expectations, calculating Q is redundant for the implementation of

the algorithm. Thus in the E-step we calculate

$$\pi_{i,j}^{(m+1)} = E(\Delta_{i,j}|Y_i = y_i, \theta) = \frac{f_j(y_i)\pi_j^{(m)}}{\sum_{l=1}^{k} f_l(y_i)\pi_l^{(m)}}. \tag{3.7}$$

In the M-step we maximize $Q(\theta|\theta^{(m)})$ with respect to θ. The maximization with respect to the mixing variables are given by (3.6). Our task is find the maximizer parameters of

$$\sum_{i=1}^{n}\sum_{j=1}^{k} \pi_{ij}^{(m+1)} \log\left(\frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(y_i-\mu_j)^2}{2\sigma_j^2}}\right).$$

We calculate this as in case of the usual ML estimate of a simple Gaussian data.

$$\frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{k} \pi_{ij}^{(m+1)}\left(-\log\left(\sqrt{2\pi\sigma_j^2}\right) - \frac{(y_i-\mu_j)^2}{2\sigma_j^2}\right)}{\partial \mu_l} = \sum_{i=1}^{n} \pi_{il}^{(m+1)} \frac{2(y_i-\pi_l)}{2\sigma_j^2} = 0$$

So

$$\mu_l^{(m+1)} = \frac{\sum_{i=1}^{n} \pi_{ij}^{(m+1)} y_i}{\sum_{i=1}^{n} \pi_{ij}^{(m+1)}}, \quad l = 1, \ldots, k. \tag{3.8}$$

The maximum with respect to σ_l^2:

$$\begin{aligned}
\frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{k} \pi_{ij}^{(m+1)}\left(-\log\left(\sqrt{2\pi\sigma_j^2}\right) - \frac{(y_i-\mu_j)^2}{2\sigma_j^2}\right)}{\partial \sigma_l^2} &= \\
= \sum_{i=1}^{n} \pi_{ij}^{(m+1)}\left(-\frac{1}{2\sigma_l^2} + \frac{(y_i-\mu_j)^2}{2(\sigma_l^2)^2}\right) &= \\
\sum_{i=1}^{n} \pi_{ij}^{(m+1)}\left(-1 + \frac{(y_i-\mu_l)^2}{\sigma_l^2}\right) &= 0
\end{aligned}$$

So the solution for σ_l^2 is

$$\frac{\sum_{i=1}^{n} \pi_{ij}^{(m+1)}(y_i-\mu_l)^2}{\sum_{i=1}^{n} \pi_{ij}^{(m+1)}} = \sigma_l^2.$$

Since the maximum point with respect to μ_l was independent of σ_j^2 we get

$$(\sigma_l^2)^{(m+1)} = \frac{\sum_{i=1}^{n} \pi_{ij}^{(m+1)} (y_i - \mu_l^{(m+1)})}{\sum_{i=1}^{n} \pi_{ij}^{(m+1)}}, \quad l = 1, \ldots, k. \tag{3.9}$$

To sum up, (3.7) gives the E-step, (3.8) and (3.9) gives the M-step of an implementation of the EM algorithm for univariate normal mixtures. For multivariate normal mixtures, the E-step is almost the same, with substituting to the multivariate density function. The M-step is similar, we have to calculate the weighted average and the weighted variance of the data.

3.3 Simulation results

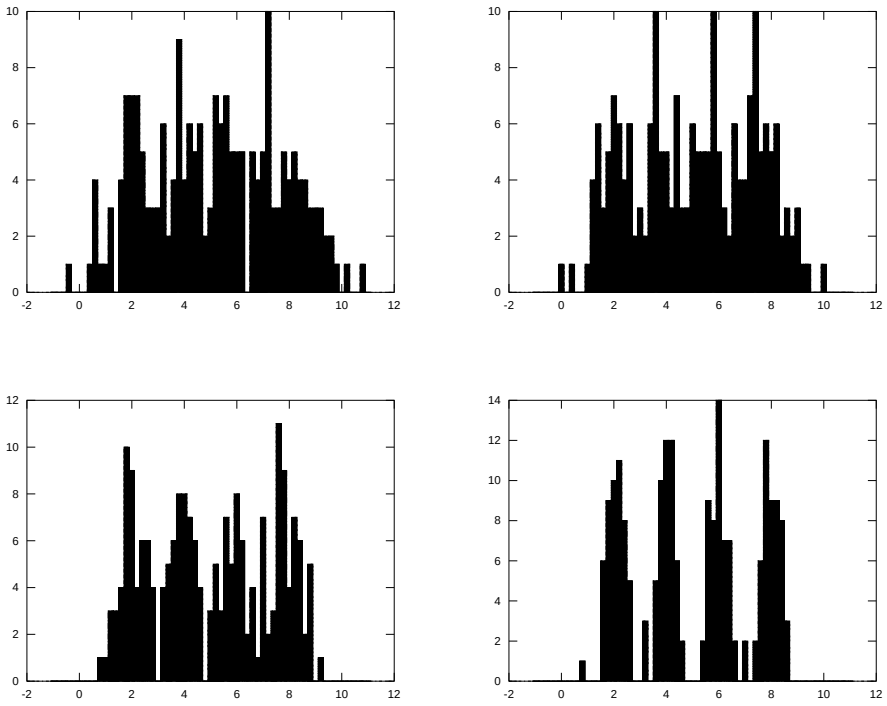
Here are some simulation results with the EM algorithm for normal mixtures. The simulation was processed by GNU Octave version 3.2.4. Four separate experiments are given. In the first part of the program random data is generated with the following parameters. In all the four cases the random data is a mixture of four normals, there are 50 data point from each distribution. The expected values in the four cases are 2, 4, 6 and 8, respectively. In each component variance is equal. In the first case the variance is 1, in the second case the variance is 0.5, in the third case the variance is 0.3, and in the last case the variance is 0.1.

In Figure 3.1 we can see the histograms of the data in the four cases.

In the upper left figure we can see the histogram for variance 1. It is not even clear that the data is a mixture of four normals. The data with variance 0.5 is on the upper right histogram. There are some peaks in it,therefore, it is somewhat obvious that it is a mixture. On the lower left histogram we can see the data with variance 0.3. It is clear that it comes form four different distributions. On the lower right histogram we can see that in case of variance 0.1, the data points can be distinguished easily by inspection. The reason for examining the four different cases is testing the performance of the algorithm in four different setting, starting with a data with the four categories being almost invisible, to a data with clearly visible categories.

The statistician can set up a initial value of the mean of distribution by looking

Figure 3.1: Histogram of the observed data

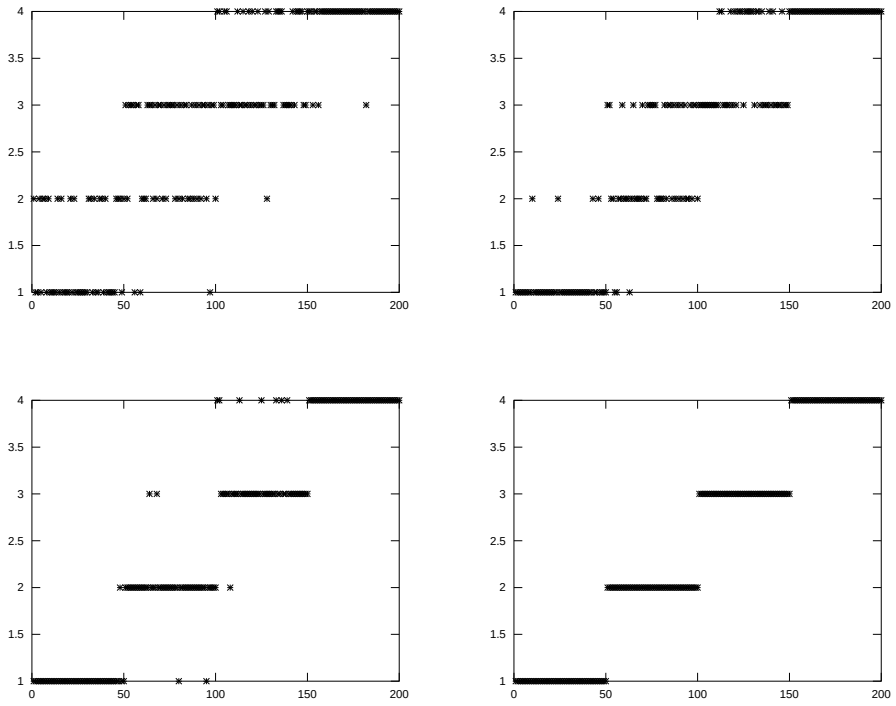


at the data range. The choice of the initial means must be realistic.

After 100 iteration of the algorithm, in Figure 3.3 we can see which distribution each data point is the most likely in. On the upper left figure the membership is not really clear. In fact, we cannot expect so, because in case of variance 1 and mean distance 2, the normal data points overlap, and it is easy to see that in case of mixture of normals with given parameters, sections and half lines gives the categories for each data points which category is the most probable. In the upper right figure, there are less misclassifications of the algorithm, In the lower left figure there are only a few wrong decisions, and on the lower right there is no misclassification at all.

We can see this results on the confusion matrix. The confusion matrix is defined by elements. If C is the confusion matrix, C_{ij} is the number of data points that originally comes from the ith distribution and after the iteration it gets to the jth distribution.

Figure 3.2: Clusters after the iteration



31	19	0	0
3	18	29	0
0	1	28	21
0	0	3	47

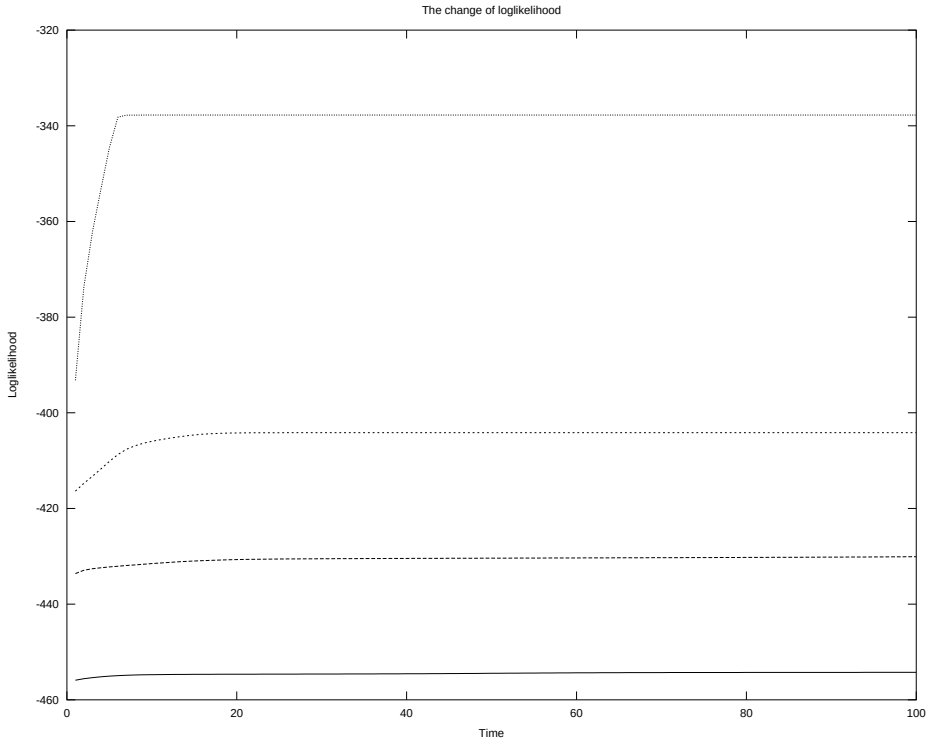
46	4	0	0
3	27	20	0
0	0	31	19
0	0	0	50

49	1	0	0
2	46	2	0
0	1	42	7
0	0	0	50

50	0	0	0
0	50	0	0
0	0	50	0
0	0	0	50

In figure 3.3 we can see the how the value of loglikelihood increases during the iteration.

Figure 3.3: The loglikelihood during the iteration



Chapter 4

Generalized random graphs

In this chapter we are introducing the concept of generalized random graphs and applying the the EM algorithm to divide them into homogeneous clusters. The implementation of the EM algorithm is similar to the one of normal mixtures.

Definition 5. *Let $G = (V, E)$ be a graph on n vertices. Let $k < n$ be fixed. There is a hidden partition $V_1, \dots V_k$ of V and each vertex belongs to V_j randomly and independently with probability π_j $(j = 1 \dots, k)$. Knowing the cluster memberships, vertices form V_i and V_j are connected randomly and independently with probability p_{ij} $(p_{ij} = p_{ji})$. A random graph given by this model is called* generalized random graph(GRG).

Note that connections are possible within the clusters. The word cluster is used to distinguish homogeneous subsets of V, not maximal connected subgraphs.

In case of $k = 1$, it is the well-known $G(n, p)$ Erdős–Rényi random graph. Examining the generalized random graphs is important in some real-life problems, such as modeling social networks. Finding homogeneous groups is the main direction of research in GRG. We would like to apply the EM algorithm to obtain the clusters of the generalized random graphs, and get the connection probabilities.

In our case the adjacency matrix and the number of clusters are given, the cluster membership is unknown. So the observed data is the adjacency matrix **A** of the graph. The unobserved data in this model is the cluster memberships.

We introduce a hidden vector variable Δ_i for each vertex $v_i \in V$ as in Chapter 3. The parameter is $\theta = (\pi_1, \dots \pi_k, p_{rs}, \quad 1 \le r, s \le k)$.

The likelihood of the complete data is

$$f(\mathbf{X}, \mathbf{\Delta}|\theta) = \left(\prod_{i=1}^{n} \sum_{j=1}^{k} \Delta_{ij} \pi_j \right) \cdot \left(\prod_{1 \le i < j \le n} \sum_{r,s=1}^{k} \Delta_{ir} \Delta_{js} p_{rs}^{a_{ij}} (1 - p_{rs})^{1 - a_{ij}} \right)$$

The incomplete data likelihood is the sum of likelihood of the complete data on all the possible Δ's:

$$\sum_{\Delta} \left[\left(\prod_{i=1}^{n} \sum_{j=1}^{k} \Delta_{ij} \pi_j \right) \cdot \left(\prod_{1 \le i < j \le n} \sum_{r,s=1}^{k} \Delta_{ir} \Delta_{js} p_{rs}^{a_{ij}} (1 - p_{rs})^{1 - a_{ij}} \right) \right]$$

As above we would like to maximize the likelihood of the incomplete data with respect to π_i and p_{rs}. In order to do that we use the iteration given by the EM algorithm. The initialization step is as usual, choose an arbitrary $\theta^{(0)}$.

In the E-step we would like to calculate the function $Q(\theta|\theta^{(m)})$.

$$Q(\theta|\theta^{(m)}) = E(\log f(\mathbf{X}|\theta)|\mathbf{A}, \theta^{(m)}) = E\Bigg(\sum_{i=1}^{n} \sum_{r=1}^{k} \Delta_{ir} \log \pi_r + \\ + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{r=1}^{k} \sum_{s=1}^{k} \Delta_{ir} \Delta_{js} \log(p_{rs}^{a_{ij}} (1 - p_{rs})^{1 - a_{ij}}) | \mathbf{A}, \theta^{(m)} \Bigg)$$

Taking the expected value inside the sums and evaluating the logarithm we get

$$\sum_{i=1}^{n} \sum_{r=1}^{k} E(\Delta_{ir} \log \pi_r | \mathbf{A}, \theta^{(m)}) + \\ \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{r=1}^{k} \sum_{s=1}^{k} E(\Delta_{ir} \Delta_{js} (a_{ij} \log p_{rs} + (1 - a_{ij}) \log(1 - p_{rs})) | \mathbf{A}, \theta^{(m)})$$

Since we take the conditional expectation of constant multiples of Δ_{ir} and $\Delta_{ir}\Delta_{js}$, moreover, Δ_{ir} and Δ_{js} are independent when $i \neq j$, we can take the conditional expectation of Δ_{ir} and substitute it to the function $\log f(\mathbf{x}|\theta)$.

We define the responsibilities $\pi_{ir}^{(m+1)} = E(\Delta_{ir}|\mathbf{A}, \theta^{(m)})$. So the conditional

expectation is

$$\pi_{ir}^{(m+1)} = E(\Delta_{ir}|\mathbf{A}, \theta^{(m)}) = P(\Delta_{ir} = 1|\mathbf{A}, \theta^{(m)}). \tag{4.1}$$

Since Δ_{ir} is independent of Δ_j if $i \neq j$, and a_{jl} only depends on Δ_j and Δ_l, in the condition $\mathbf{A}$ those entries, neither of the endpoints of which is i , are redundant. Therefore, we can only take pairs of vertices $\{(i,j) : j \neq i\}$ as a condition. Let us denote this by S_i. So by the Bayes theorem

$$P(\Delta_{ir} = 1|S_i, \theta^{(m)}) = \frac{P(S_i|\Delta_{ir} = 1, \theta^{(m)})P(\Delta_{ir} = 1|\theta^{(m)})}{\sum_{s=1}^{k} P(S_i|\Delta_{is} = 1, \theta^{(m)})P(\Delta_{is} = 1|\theta^{(m)})}, \tag{4.2}$$

where $P(\Delta_{ir} = 1|\theta^{(m)}) = \pi_r^{(m)}$ and

$$\begin{aligned} P(S_i|\Delta_{ir} = 1, \theta^{(m)}) &= \sum_{\substack{\Delta \\ \Delta_{ir}=1}} \prod_{j\neq i} \sum_{t=1}^{k} \Delta_{jt}\pi_t \prod_{j\neq i} \sum_{t=1}^{k} \Delta_{jt}(p_{rt}^{(m)})^{a_{ij}}(1 - p_{rt}^{(m)})^{1-a_{ij}} \\ &= \prod_{j\neq i} \sum_{t=1}^{k} \pi_t (p_{rt}^{(m)})^{a_{ij}}(1 - p_{rt}^{(m)})^{1-a_{ij}}. \end{aligned}$$

Hence the essential part of the E-step is given by

$$\pi_{ir}^{(m+1)} = \frac{\prod_{j\neq i} \sum_{s=1}^{k} \pi_s (p_{rs}^{(m)})^{a_{ij}}(1 - p_{rs}^{(m)})^{1-a_{ij}}\pi_r}{\sum_{s=1}^{k} \prod_{j\neq i} \sum_{t=1}^{k} \pi_t (p_{st}^{(m)})^{a_{ij}}(1 - p_{st}^{(m)})^{1-a_{ij}}\pi_s}, \quad i = 1 \dots n,\ r = 1, \dots, k.$$

The M-step is to maximize

$$\sum_{i=1}^{n} \sum_{r=1}^{k} \pi_{ir}^{(m+1)} \log \pi_r +$$

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{r=1}^{k} \sum_{s=1}^{k} \pi_{ir}^{(m+1)} \pi_{js}^{(m+1)} (a_{ij} \log p_{rs} + (1 - a_{ij}) \log(1 - p_{rs}))$$

with respect to π_r and p_{rs}. Similarly to the case of normal mixtures, the two subset of parameters are independent, and the function to maximize is a sum that separately contains the two kinds of parameters, therefore we could maximize them separately.

The first one is

$$\sum_{i=1}^{n}\sum_{r=1}^{k}\pi_{ir}^{(m+1)}\log\pi_r \tag{4.3}$$

with the constraint $\sum_{s=1}^{k}\pi_s = 1$. We can see that (4.3) is the same as the first sum in case of normal mixtures. Therefore, we can write

$$\pi_r^{(m+1)} = \frac{\sum_{i=1}^{n}\pi_{ir}^{n+1}}{n} \quad r = 1,\dots k.$$

To get $p_{uv}^{(m+1)}$ we have to maximize

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{r=1}^{k}\sum_{s=1}^{k}\pi_{ir}^{(m+1)}\pi_{js}^{(m+1)}(a_{ij}\log(p_{rs}) + (1-a_{ij})\log(1-p_{rs})) \tag{4.4}$$

The partial derivative of (4.4) with respect to p_{uv} at the maximum is 0:

$$\begin{aligned}
\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\pi_{ir}^{(m+1)}\pi_{js}^{(m+1)}(a_{ij}\frac{1}{p_{uv}} - (1-a_{ij})\frac{1}{1-p_{uv}}) &= 0\\
\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\pi_{ir}^{(m+1)}\pi_{js}^{(m+1)}(a_{ij}(1-p_{uv}) - (1-a_{ij})p_{uv}) &= 0\\
\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\pi_{ir}^{(m+1)}\pi_{js}^{(m+1)}a_{ij}}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\pi_{ir}^{(m+1)}\pi_{js}^{(m+1)}} &= p_{rs} \qquad (4.5)
\end{aligned}$$

So (4.5) gives us $p_{uv}^{(m+1)}$.

The implementation of the EM algorithm for GRG is somewhat similar to the one of mixtures. Introducing the responsibilities saves some computational time, as in the E-step it is enough to calculate them. And using the responsibilities, Q can be maximized formally, without calculating Q itself.

Conclusion

Our thesis have presented the EM algorithm and shown the most important properties of it regarding the convergence. Then we have described by a typical model, the finite mixtures, how to implement the algorithm for a hidden variable model. The simulation results have shown that it is not always convenient to use the EM algorithm, however, if the analysis of the data indicates that it is a mixture, the simulations yield good results. We have given the implementation of the algorithm for generalized random graphs.

Bibliography

[1] Celeux, G., Govaert, G. *Gaussian Parsimonious Clustering Models*, Pattern recognition, Vol. 28, No. 5. pp. 781-793, 1995.

[2] Csiszár, I., Shields, P. C. *Information Theory and Statistics: A Tutorial* Foundations and Trends^{TM} in Communications and Information Theory Volume 1 Issue 4, 2004

[3] Dempster, A. B., Laird, N. M., Rubin, D. B. *Maximum Likelihood from Incomplete Data via the EM Algorithm*

[4] Fraley, C., Raferty, A. E. *Model-based clustering discriminant analysis, and density estimation* Journal of the American Statistical Association; Jun 2002;97, 458

[5] Hastie, T., Tibshirani, T. R., Friedman, J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* Second Edition, Springer, 2003

[6] Rao, C. R. *Linear Statistical Inference and Its Applications*, Second edition, John Wiley & Sons, 1973

[7] Wu, C. F. J. *On convergence properties of the EM algorithm* The Annals of Statistics, Vol. 11, No. 1, 95-103, 1983