TR-95-13

Mixture model for image understanding and the EM algorithm

Shotaro Akaho *

akaho@etl.go.jp

Received 7 April 1995

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba,
Ibaraki, 305, JAPAN

Abstract We present a mixture model that can be applied to the recognition of multiple objects in an image plane. The model consists of any shape of submodules. Each submodule is a probability density function of data points with scale and shift parameters, and the modules are combined with weight probabilities. We present the EM (Expectation-Maximization) algorithm to estimate those parameters. We also modify the algorithm in the case that data points are restricted in an attention window.

*Mathematical Informatics Section. This version is for electric distribution(ftp). The last update was on May 8, 1995.

1 Introduction

When we must solve a certain problem, it is effective to divide the problem into subproblems and then to integrate the results of the subproblems. Such a divide-and-conquer approach has successfully applied to a large number of information processing problems.

On the other hand, statistical modeling is an appropriate method to deal with a large number of data from real world environment, because those data are deteriorated by noise and some data might be missing.

We consider a mixture model that is a statistical model consisting of submodules. Each module is a parametric probability distribution and those modules are integrated by taking a weighted sum. In order to make our discussion concrete, we consider the application to an image recognition problem. Suppose there are scattered points in an image plane and clusters of those points form objects. Our purpose is to fit a mixture model to those objects, where each submodule corresponds to a model of some object. We assume that objects in a real image are scaled or shifted from the original object model.

Recently, the EM (Expectation and Maximization) algorithm, which is a technique to find a locally maximum likelihood estimation from incomplete data, has attracted much attention because it can be applied the parameter estimation of a lot of models of neural networks or related models, for example, Boltzmann machines with hidden units[3] and stochastic multi-layer perceptrons[2]. The EM algorithm has been successfully applied to hidden Markov model for speech recognition, which is known as Baum-Welch algorithm[4].

The EM algorithm can be also applied to a mixture model. Some kind of modules such as normal distribution can be easily trained by the EM algorithm. M.I. Jordan proposed hierarchical mixtures of experts model (HME) that is an extension of the mixture model[6, 7]. The framework of HME is supervised-learning whose statistical model is a conditional probability from inputs to outputs, and each module is designed to be a generalized linear model (GLIM)[8]. There is a fast algorithm to estimate parameters of GLIM, and the algorithm can be combined with the EM algorithm.

In this paper, we consider unsupervised learning, whose statistical model is an unconditional probability, and each module can be any kind of an object model. We adjust the scale and shift parameters to fit the module to an object. Each module is approximated by a normal mixture model and the parameters of the module are estimated in advance.

Another important technique to recognize multiple objects efficiently is attention. When there are a lot of objects in a image, it is not easy to fit a complicated model. However, if we restrict a region to estimate parameters, the data out of the region should be treated as missing values. We can also apply the EM algorithm to this problem.

In section 2, we describe the mixture model and introduce some examples of submodules. Mixture module with scale and shift parameters, which is the most important module we are concerned with, is described in 2.3. In section 3, we explain a general form of the EM algorithm. Fundamental equations of estimation by the EM algorithm are given by (21) and (22). In section 4, we present the EM algorithm for mixture models.

Recurrence formulae in each EM step are given by (32), (37) and (41) in 4.2.2. In section 5, we extend the algorithm to the case that the parameter is estimated only in a restricted region. The extended recurrence formulae are given by (49) and (50) and (54) in 5.5.

2 Mixture model

2.1 Mixture model

Suppose there are n probability density function $g_i(x \mid \boldsymbol{\theta}_i)$ $(i = 1, \ldots, n)$, where $\boldsymbol{\theta}_i$ denotes a parameter. For the sake of simplicity, random variable x is assumed to be one dimensional vector, but it is easy to generalize to higher dimension as shown in section 6. Each g_i is called "**module**". Mixture model is a probability density function defined as the weighted sum of those modules,

$$p(x \mid \boldsymbol{\xi}, \boldsymbol{\theta}^*) = \sum_{i=1}^{n} \frac{\xi_i}{\sum_k \xi_k} g_i(x \mid \boldsymbol{\theta}_i), \qquad \xi_i \in \mathcal{R}, \quad \xi_i > 0, \tag{1}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ and $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$.

The goal of the maximum likelihood estimation (MLE) is to estimate $\boldsymbol{\xi}$ and $\boldsymbol{\theta}^*$ that maximizes the likelihood from given samples of x,

$$\sum_x \log p(x \mid \boldsymbol{\xi}, \boldsymbol{\theta}^*), \tag{2}$$

where the summation is taken over all samples.

Since it is difficult to estimate $\boldsymbol{\xi}$ and $\boldsymbol{\theta}^*$ directly, let us introduce a hidden random variable $z \in \{1, 2, \ldots, n\}$, and observed x is generated from the module $g_z(x \mid \boldsymbol{\theta}_z)$. The joint distribution of (x, z) is given by

$$p(x, z \mid \boldsymbol{\xi}, \boldsymbol{\theta}^*) = \frac{\xi_z}{\sum_k \xi_k} g_z(x \mid \boldsymbol{\theta}_z). \tag{3}$$

If we regard observed sample x as the incomplete data of (x, z), we can apply the EM algorithm for the estimation.

It is also remarkable that we can combine any different type of modules. The difficulty of the estimation only depends on the form of each module g_i as long as the parameters of the module is independent of each other.

In the following subsection, we introduce several kinds of modules. Especially, mixture module described in subsection 2.3 can approximate a wide class of distributions. Parameters of all modules are successfully estimated by the EM algorithm in the sense that the estimation is explicitly obtained in each EM step. The explicit form of the estimation is described in section 4. In the following sections, we omit subscripts of $g_i(x \mid \boldsymbol{\theta}_i)$ as $g(x \mid \boldsymbol{\theta})$ unless confusing.

2.2 Simple modules

Uniform distribution The most simplest case is that $g(x \mid \boldsymbol{\theta})$ is a uniform distribution. In the application of object recognition, we can describe scattered noise by the uniform distribution module. Since the uniform distribution has a compact support, the domain of x must be bounded in order to keep the distribution regular. However, even if the domain is not bounded, we can construct the estimation algorithm formally.

Normal distribution Normal distribution is defined as

$$g(x \mid \boldsymbol{\theta}) = \mathrm{g}(x \mid \mu, \sigma^2) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \tag{4}$$

where $\boldsymbol{\theta} = (\mu, \sigma^2)$. The normal mixture model, which is the mixture model consisting of normal distribution modules, is widely applied to image recognition tasks because some kinds of objects in the image can be roughly approximated by normal distributions. Moreover, the estimation of parameters is easy since normal distribution belongs to an exponential family.

However, when it is desired to fit more complicated modules or completely different type of modules from normal distribution, it is difficult to find out appropriate modules so that the estimation of parameters is easy.

2.3 Mixture module with scale and shift parameters

In this subsection, we consider a class of density functions

$$\{ag(ax+b) \mid 0 < a, \quad a, b \in \mathcal{R}\}, \tag{5}$$

where $g(x)$ is an arbitrary smooth density function, and a denotes a scale parameter and b denotes a shift parameter.

Of course, it is difficult to estimate a and b in general. Therefore we approximate $g(x)$ in advance by normal mixture model as follows.

$$\hat{g}(x \mid \boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{j=1}^{m} \frac{\zeta_j}{\sum_k \zeta_k} \mathrm{g}(x \mid \mu_j, \sigma_j^2), \tag{6}$$

where $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_m)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_m^2)$

According to the result of the density approximation theory, $g(x)$ can be approximated by $\hat{g}(x \mid \boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ as precisely as possible with sufficient large m, when $g(x)$ satisfies some regularity conditions. The estimation of $\boldsymbol{\zeta}$,$\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ is described in 4.1.

We adopt

$$\{a\hat{g}(ax+b \mid \boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \mid a, b \in \mathcal{R}, \quad a > 0\} \tag{7}$$

as a module of mixture models instead of g. However, even if $\hat{g}$ is used, the estimation of a and b is not trivial, because a and b are common parameters in normal modules included in $a\hat{g}(ax+b \mid \boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\sigma})$. The estimation is described in section 4.2.

3 The EM algorithm

In this section, we describe a general form of the EM algorithm, which is formulated by Dempster et al[5].

3.1 Missing values and the EM algorithm

Observed data $\boldsymbol{y}$ is considered to be an incomplete data of a complete data $\boldsymbol{x}$, where the many-to-one map from $\boldsymbol{x}$ to $\boldsymbol{y}$ is assumed to be known.

The EM algorithm is applied to find the local maximum of likelihood $\sum_{\boldsymbol{y}}[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})]$, especially it is effective when the likelihood of the incomplete data is much more difficult to maximize than the likelihood of the complete data.

The algorithm starts from an initial solution $\boldsymbol{\theta}^{(0)}$ and it develops solution $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(t)}$ iteratively. Each iteration step consists of the expectation step (**E-step**) and the maximization step (**M-step**). In each iteration, the likelihood increases monotonously.

The EM algorithm

1. Let $\boldsymbol{\theta}^{(0)}$ be an initial solution.

2. Repeat the following two steps for $t = 0, 1, 2, \ldots,$

   (a) E-step :
   Calculate the expectation value of log-likelihood of complete data conditioned by observed samples and the current solution $\boldsymbol{\theta}^{(t)}$,

   $$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \sum_{\boldsymbol{y}} \left[\mathrm{E} \left[\log p(\boldsymbol{x} \mid \boldsymbol{\theta}) \mid \boldsymbol{y}, \boldsymbol{\theta}^{(t)} \right] \right], \tag{8}$$

   where the summation is taken over all samples.

   (b) M-step :
   Let $\boldsymbol{\theta}^{(t+1)}$ be such $\boldsymbol{\theta}$ that maximizes $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$.

3.2 Curved exponential family

An **exponential family** is a class of parametric probability density functions, and it is written in the following form,

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \exp\{\sum_i \theta_i F_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) + C(\boldsymbol{x})\}, \tag{9}$$

where θ_i is called **the natural parameter**. Let us consider a parameter defined by

$$\eta_i = \mathrm{E}[F_i(\boldsymbol{x}) \mid \boldsymbol{\theta}], \tag{10}$$

which is called **the expectation parameter**. An exponential family (9) defines the manifold whose local coordinate system is $\{\theta_i\}$. $\{\eta_i\}$ is also a coordinate system. The two coordinate systems are dual to each other[1], and they are related by the Legendre transformation,

$$\theta_i = \frac{\partial}{\partial \eta_i}\phi(\boldsymbol{\eta}), \quad \eta_i = \frac{\partial}{\partial \theta_i}\psi(\boldsymbol{\theta}), \quad \phi(\eta) = \sum_i \theta_i \eta_i - \psi(\boldsymbol{\theta}) \tag{11}$$

Example 1 We show that the normal mixture model (6) with hidden variables belongs to an exponential family. The density function with hidden variable $w \in \{1, 2, \ldots, m\}$ is given by

$$\begin{aligned}
p(x, w) &= \frac{\zeta_w}{\sum_k \zeta_k} \mathrm{g}(x \mid \mu_w, \sigma_w^2) && (12)\\
&= \exp\left\{ \log \zeta_w - \frac{(x-\mu_w)^2}{2\sigma_w^2} - \log\sqrt{2\pi\sigma_w^2} - \log \sum_k \zeta_k \right\} && (13)\\
&= \exp\left[\sum_j \delta(w-j) \left\{ \log \frac{\zeta_j}{\sigma_j} - \frac{(x-\mu_j)^2}{2\sigma_j^2} \right\} - \log(\sqrt{2\pi} \sum_k \zeta_k) \right] && (14)\\
&= \exp\left[\sum_j \left\{ (\log \frac{\zeta_j}{\sigma_j} - \frac{\mu_j^2}{2\sigma_j^2})\delta(w-j) + \frac{\mu_j}{\sigma_j^2} x \delta(w-j) \right.\right.\\
&\qquad \left.\left. -\frac{1}{2\sigma_j^2} x^2 \delta(w-j) \right\} - \log(\sqrt{2\pi} \sum_k \zeta_k) \right]. && (15)
\end{aligned}$$

Therefore, by putting

$$\theta_j = \log \frac{\zeta_j}{\sigma_j} - \frac{\mu_j^2}{2\sigma_j^2}, \qquad \theta_{m+j} = \frac{\mu_j}{\sigma_j^2}, \qquad \theta_{2m+j} = -\frac{1}{2\sigma_j^2}, \tag{16}$$

$$F_j(x, w) = \delta(w-j), \quad F_{m+j}(x, w) = x\delta(w-j), \quad F_{2m+j}(x, w) = x^2\delta(w-j), \tag{17}$$

where $j = 1, \ldots, m$ and δ denotes Kronecker's delta, we can show that $p(x, w)$ belongs to an exponential family. The expectation parameters are obtained by straight forward calculations,

$$\eta_j = \zeta_j, \qquad \eta_{m+j} = \zeta_j \mu_j, \qquad \eta_{2m+j} = \zeta_j(\mu_j^2 + \sigma_j^2). \tag{18}$$

An **curved exponential family** is defined as a submanifold of an exponential family. Suppose the local coordinate system is $\{u_i\}$, the probability distribution is written by

$$p(\boldsymbol{x} \mid \boldsymbol{u}) = \exp\{\sum_i \theta_i(\boldsymbol{u}) F_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta}(\boldsymbol{u})) + C(\boldsymbol{x})\}, \tag{19}$$

where $\dim(\boldsymbol{u}) < \dim(\boldsymbol{\theta})$.

A lot of typical distributions belong to an exponential family or a curved exponential family. Moreover a (curved) exponential family with hidden variables includes more interesting distributions such as the normal mixture models, the stochastic perceptrons and

the hierarchical mixtures of experts. In that case, the EM algorithm is reduced a simple form as shown below, which is one reason that the EM algorithm is widely applied.

Applying the E-step (8) to the curved exponential family (19), it follows

$$\begin{aligned} Q(\boldsymbol{u} \mid \boldsymbol{u}^{(t)}) &= \sum_{\boldsymbol{y}} \mathrm{E}[\log p(\boldsymbol{x} \mid \boldsymbol{u}) \mid \boldsymbol{y}, \boldsymbol{u}^{(t)}] \\ &= \sum_{\boldsymbol{y}} \left\{ \sum_i \theta_i(\boldsymbol{u}) \, \mathrm{E}[F_i(\boldsymbol{x}) \mid \boldsymbol{y}, \boldsymbol{u}^{(t)}] - \psi(\boldsymbol{\theta}(\boldsymbol{u})) + \mathrm{E}[C(\boldsymbol{x}) \mid \boldsymbol{y}, \boldsymbol{u}^{(t)}] \right\}. \end{aligned} \tag{20}$$

Next, we maximize $Q(\boldsymbol{u} \mid \boldsymbol{u}^{(t)})$ with respect to $\boldsymbol{u}$ in the M-step. Differentiating by u_j, it follows

$$\sum_{\boldsymbol{y}} \left\{ \sum_i \frac{\partial \theta_i(\boldsymbol{u})}{\partial u_j} \left\{ \mathrm{E}[F_i(\boldsymbol{x}) \mid \boldsymbol{y}, \boldsymbol{u}^{(t)}] - \eta_i(\boldsymbol{u}) \right\} \right\} = 0, \tag{21}$$

where we used the relation (11). Equation (21) is the fundamental equation of the EM algorithm for a curved exponential family. Especially, if the class of distribution belongs to an exponential family, namely if $\boldsymbol{u}$ is identical to $\boldsymbol{\theta}$, explicit solution is given by

$$\eta_i = \frac{1}{N} \sum_{\boldsymbol{y}} \mathrm{E}[F_i(\boldsymbol{x}) \mid \boldsymbol{y}, \boldsymbol{\theta}^{(t)}], \tag{22}$$

where N is the number of samples y.

4 The EM algorithm for the mixture model

4.1 Normal mixture model

As already shown in example 1, the normal mixture model belongs to the exponential family with hidden variables. Thus we can use (22) for the estimation. The conditional density function $q(w \mid x, \boldsymbol{\lambda})$ is given by

$$q(w \mid x, \boldsymbol{\lambda}) = \frac{p(x, w \mid \boldsymbol{\lambda})}{p(x \mid \boldsymbol{\lambda})} = \frac{\zeta_w \mathrm{g}(x \mid \mu_w, \sigma_w^2)}{\sum_k \zeta_k \mathrm{g}(x \mid \mu_k, \sigma_k^2)}. \tag{23}$$

where $\boldsymbol{\lambda} = (\boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_m)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_m^2)$.

Therefore from (16), (17), (18) and (22), we obtain the recurrence formula of the EM algorithm,

$$\zeta_j^{(t+1)} = \frac{1}{N} \sum_x q_j(x, \boldsymbol{\lambda}^{(t)}), \tag{24}$$

$$\zeta_j^{(t+1)} \mu_j^{(t+1)} = \frac{1}{N} \sum_x x q_j(x, \boldsymbol{\lambda}^{(t)}), \tag{25}$$

$$\zeta_j^{(t+1)} ((\mu_j^{(t+1)})^2 + (\sigma_j^{(t+1)})^2) = \frac{1}{N} \sum_x x^2 q_j(x, \boldsymbol{\lambda}^{(t)}), \tag{26}$$

where $q_j(x, \boldsymbol{\lambda}^{(t)}) = q(j \mid x, \boldsymbol{\lambda}^{(t)})$, $j = 1, \ldots, m$ and N is the number of observed samples.

4.2 Scale and shift parameter estimation

4.2.1 Mixture model with mixture modules

In this section, we present the EM algorithm for the mixture model (1) with mixture modules described in section 2.3. Each module g_i is a normal mixture module (7) with scale and shift parameters.

Suppose that observed sample x is generated from the w-th normal submodule of the z-th module. $w \in \{1, \ldots, m\}$ and $z \in \{1, \ldots, n\}$ are hidden variables. Parameters to estimate are the weight ξ_i, the scale parameter a_i and the shift parameter b_i, where $i = 1, \ldots, n$. Other parameters are fixed, The distribution of (x, z, w) is written as

$$p(x, z, w \mid \boldsymbol{\lambda}) = p(x, z, w \mid \boldsymbol{\xi}, \boldsymbol{a}, \boldsymbol{b}) = \frac{\xi_z}{\sum_j \xi_j} \frac{\zeta_{z,w}}{\sum_k \zeta_{z,k}} a_z \mathrm{g}(a_z x + b_z \mid \mu_{z,w}, \sigma^2_{z,w}), \tag{27}$$

where $\boldsymbol{\lambda} = (\boldsymbol{\xi}, \boldsymbol{a}, \boldsymbol{b})$, $\boldsymbol{a} = (a_1, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, \ldots, b_n)$.

Each module g_i may have a different number of submodules. Since $p(x, z, w \mid \boldsymbol{\xi}, \boldsymbol{a}, \boldsymbol{b})$ belongs to a curved exponential family, it is not trivial to optimize parameters. If we optimize $\boldsymbol{a}$ and $\boldsymbol{b}$ separately in each E-step, we can obtain an explicit recurrence formulae as shown below.

4.2.2 Recurrence formulae for the weight and the scale parameters

In this section, we optimize $\boldsymbol{a}$ with fixing $\boldsymbol{b}$. The recurrence formulae are given in (32) and (37).

$$\begin{aligned} p(x, z, w \mid \boldsymbol{\xi}, \boldsymbol{a}) \;=\; & \exp\Bigg[\sum_i \delta(z - i) \log \frac{\xi_i}{a_i} + \sum_{i,j} \delta(z - i)\delta(w - j) x \frac{a_i(b_i - \mu_{i,j})}{\sigma^2_{i,j}} \\ & - \sum_{i,j} \delta(z - i)\delta(w - j) x^2 \frac{a_i^2}{2\sigma^2_{i,j}} - \psi(\boldsymbol{\xi}, \boldsymbol{a}) + C(x, z, w) \Bigg], \end{aligned} \tag{28}$$

where $\psi(\boldsymbol{\xi}, \boldsymbol{a})$ does not depend on (x, z, w) and $C(x, z, w)$ does not depend on $\boldsymbol{\xi}$ and $\boldsymbol{a}$. It is the form of a curved exponential family with natural parameters

$$\theta_{i,0} = \log \frac{\xi_i}{a_i}, \quad \theta_{i,j} = \frac{a_i(b_i - \mu_{i,j})}{\sigma^2_{i,j}}, \quad \theta_{i,m+j} = -\frac{a_i^2}{2\sigma^2_{i,j}}, \tag{29}$$

where $i = 1, \ldots, n$, $j = 1, \ldots, m$ and m can depend on i. Expectation parameters are given by

$$\eta_{i,0} = \xi_i, \quad \eta_{i,j} = \xi_i \zeta_{i,j} \frac{\mu_{i,j} - b_i}{a_i} \quad \eta_{i,m+j} = \xi_i \zeta_{i,j} \left(\frac{(\mu_{i,j} - b_i)^2 + \sigma^2_{i,j}}{a_i^2} \right). \tag{30}$$

We apply the EM algorithm (21) to the above model. The conditional probability $q(z, w \mid x, \boldsymbol{\lambda})$ of the hidden variables, where $\boldsymbol{\lambda} = (\boldsymbol{\xi}, \boldsymbol{a})$, is given by

$$q(z, w \mid x, \boldsymbol{\lambda}) = \frac{p(x, z, w \mid \boldsymbol{\lambda})}{p(x \mid \boldsymbol{\lambda})} = \frac{\xi_z \frac{\zeta_{z,w}}{\sum_l \zeta_{z,l}} a_z \mathrm{g}(a_z x + b_z \mid \mu_{z,w}, \sigma^2_{z,w})}{\sum_{k,l} \xi_k \frac{\zeta_{k,l}}{\sum_{l'} \zeta_{k,l'}} a_k \mathrm{g}(a_k x + b_k) \mid \mu_{k,l}, \sigma^2_{k,l})}. \tag{31}$$

As for $\boldsymbol{\xi}$, we obtain

$$\xi_i^{(t+1)} = \frac{1}{N} \sum_j \sum_x q_{i,j}(x, \boldsymbol{\lambda}^{(t)}). \tag{32}$$

where $q_{i,j}(x, \boldsymbol{\lambda}) = q(i, j \mid x, \boldsymbol{\lambda})$.

From straight forward calculations for $\boldsymbol{a}$, we get the equation,

$$X_i a_i^2 + Y_i a_i - Z_i = 0, \tag{33}$$

where

$$X_i = \sum_j \sum_x \frac{x^2 q_{i,j}(x, \boldsymbol{\lambda}^{(t)})}{\sigma_{i,j}^2} \tag{34}$$

$$Y_i = \sum_j \sum_x \frac{(b_i - \mu_{i,j}) x q_{i,j}(x, \boldsymbol{\lambda}^{(t)})}{\sigma_{i,j}^2} \tag{35}$$

$$Z_i = N \xi_i^{(t+1)} \sum_j \zeta_{i,j}. \tag{36}$$

Since $a_i > 0$, we obtain the recurrence formulae for $\boldsymbol{a}$,

$$a_i^{(t+1)} = \frac{\sqrt{Y_i^2 + 4 X_i Z_i} - Y_i}{2 X_i}. \tag{37}$$

4.2.3 Recurrence formula for the shift parameter

In this section, we optimize $\boldsymbol{b}$ with fixing $\boldsymbol{a}$. The recurrence formula is given in (41).

In a similar way to the case of optimizing $\boldsymbol{a}$,

$$\begin{aligned} p(x, z, w \mid \boldsymbol{\lambda}) &= \exp\Bigg[\sum_i \delta(z - i) \log \xi_i - \sum_{i,j} \delta(z - i)\delta(w - j) \frac{b_i^2 - 2 b_i \mu_{i,j}}{2 \sigma_{i,j}^2} \\ &\quad - \sum_{i,j} \delta(z - i)\delta(w - j) x \frac{b_i a_i}{\sigma_{i,j}^2} - \psi'(\boldsymbol{\xi}, \boldsymbol{b}) + C'(x, z, w) \Bigg], \end{aligned} \tag{38}$$

where $\psi'(\boldsymbol{\xi}, \boldsymbol{b})$ does not depend on (x, z, w) and $C'(x, z, w)$ does not depend on $\boldsymbol{\xi}$ and $\boldsymbol{b}$. It is the form of a curved exponential family with natural parameters

$$\theta'_{i,0} = \log \xi_i, \quad \theta'_{i,j} = -\frac{b_i^2 - 2 b_i \mu_{i,j}}{2 \sigma_{i,j}^2}, \quad \theta'_{i,m+j} = \frac{b_i a_i}{\sigma_{i,j}^2}, \tag{39}$$

where $i = 1, \ldots, n$, $j = 1, \ldots, m$ and m can depend on i. Expectation parameters are given by

$$\eta'_{i,0} = \xi_i, \quad \eta'_{i,j} = \xi_i \zeta_{i,j}, \quad \eta'_{i,m+j} = \xi_i \zeta_{i,j} \frac{\mu_{i,j} - b_i}{a_i}. \tag{40}$$

Applying (21) to the above model, the recurrence formula for $\boldsymbol{\xi}$ is the same as (32). And we obtain the recurrence formulae for $\boldsymbol{b}$,

$$b_i^{(t+1)} = \frac{V_i}{U_i}, \tag{41}$$

where

$$U_i = \sum_j \sum_x \frac{q_{i,j}(x, \boldsymbol{\lambda}^{(t)})}{\sigma_{i,j}^2}, \tag{42}$$

$$V_i = \sum_j \sum_x \frac{(\mu_{i,j} - a_i x)}{\sigma_{i,j}^2} q_{i,j}(x, \boldsymbol{\lambda}^{(t)}). \tag{43}$$

5 Estimation in an attention window

5.1 Attention and missing values

Attention is an efficient method for the recognition of objects and the learning of environments. However, when we focus our attention to some region, the data out of the region are censored. We can apply the EM algorithm in such a case. In this section, we show a slight modification of the EM algorithm for that problem.

We can use the uniform distribution module by considering the attention region because the uniform distribution must have a compact support as shown in section 2.2.

5.2 Formulation

Let C be the region of attention. Suppose random variable x is observed only when $x \in C$, and the number of data $x \notin C$ is unknown. We take the mixture model (1) as a model of distribution.

Since the EM algorithm can be applied when the number of data is known, we consider the following algorithm where the number of missing values are estimated in each EM step.

1. Let $(\boldsymbol{\xi}^{(0)}, \boldsymbol{\theta}^{*(0)})$ be an initial parameter.
2. Repeat the following two steps for $t = 0, 1, 2, \ldots$,
   (a) Estimate $M^{(t)}$, the number of missing values, with fixing $(\boldsymbol{\xi}, \boldsymbol{\theta}^*) = (\boldsymbol{\xi}^{(t)}, \boldsymbol{\theta}^{*(t)})$.
   (b) Apply an EM step based on N observed samples and $M^{(t)}$ missing data. Let $(\boldsymbol{\xi}^{(t+1)}, \boldsymbol{\theta}^{*(t+1)})$ be the solution.

5.3 Estimation of the number of missing observations

When we fix parameters $(\boldsymbol{\xi}, \boldsymbol{\theta}^*)$, observations can be regarded as Bernoulli trials, where each observation drops on C in probability $P_C^{(t)} = \int_C p(x \mid \boldsymbol{\xi}^{(t)}, \boldsymbol{\theta}^{*(t)})\, \mathrm{d}x$. The number of samples on $\bar{C}$ is estimated by

$$M^{(t)} = \frac{1 - P_C^{(t)}}{P_C^{(t)}} N. \tag{44}$$

5.4 The EM algorithm with missing observations

The EM algorithm described in section 3 is the case that all N incomplete samples are given. In this section, we present the EM algorithm in the case that N incomplete samples in C are given and M incomplete samples in $\bar{C}$ are missing.

$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ in (8) is modified as

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \sum_{\boldsymbol{y}} \left[\mathrm{E}\left[\log p(\boldsymbol{x} \mid \boldsymbol{\theta}) \mid \boldsymbol{y}, \boldsymbol{\theta}^{(t)} \right] \right] + M \, \mathrm{E}\left[\log p(\boldsymbol{x} \mid \boldsymbol{\theta}) \mid \boldsymbol{\theta}^{(t)}, \bar{C} \right], \tag{45}$$

where the expectation of the second term is taken over $y \in \bar{C}$. The fundamental equation for the curved exponential family in (21) is modified as

$$\sum_i \frac{\partial \theta_i(\boldsymbol{u})}{\partial u_j} \left\{ \sum_{\boldsymbol{y}} \mathrm{E}[F_i(\boldsymbol{x}) \mid \boldsymbol{y}, \boldsymbol{u}^{(t)}] + M \, \mathrm{E}[F_i(\boldsymbol{x}) \mid \boldsymbol{u}^{(t)}, \bar{C}] - (N+M)\eta_i(u) \right\} = 0, \tag{46}$$

and the equation for the exponential family is given by

$$\eta_i = \frac{1}{N+M} \left\{ \sum_{\boldsymbol{y}} \mathrm{E}[F_i(\boldsymbol{x}) \mid \boldsymbol{y}, \boldsymbol{\theta}^{(t)}] + M \, \mathrm{E}[F_i(\boldsymbol{x}) \mid \boldsymbol{\theta}^{(t)}, \bar{C}] \right\}, \tag{47}$$

5.5 Mixture model with mixture modules

In this section, we give the explicit form of the above algorithm for the mixture model with mixture modules.

Let $m_{i,j}^{\langle k \rangle}(\boldsymbol{\lambda}, \bar{C})$ be the k-th moment of $x \in \bar{C}$ defined by

$$m_{i,j}^{\langle k \rangle}(\boldsymbol{\lambda}, \bar{C}) = \frac{\int_{\bar{C}} x^k p(x, i, j \mid \boldsymbol{\xi}, \boldsymbol{a}, \boldsymbol{b}) \, \mathrm{d}x}{\int_{\bar{C}} p(x \mid \boldsymbol{\xi}, \boldsymbol{a}, \boldsymbol{b}) \, \mathrm{d}x}, \tag{48}$$

where $\boldsymbol{\lambda} = (\boldsymbol{\xi}, \boldsymbol{a}, \boldsymbol{b})$.

From (32),

$$\xi_i^{(t+1)} = \frac{1}{N + M^{(t)}} \sum_j \left\{ \sum_x q_{i,j}(x, \boldsymbol{\lambda}^{(t)}) + M^{(t)} m_{i,j}^{\langle 0 \rangle}(\boldsymbol{\lambda}^{(t)}, \bar{C}) \right\}, \tag{49}$$

from (37),

$$a_i^{(t+1)} = \frac{\sqrt{Y_i^2 + 4X_i Z_i} - Y_i}{2X_i}, \tag{50}$$

where

$$X_i = \sum_j \frac{\sum_x \{x^2 q_{i,j}(x, \boldsymbol{\lambda}^{(t)})\} + M^{(t)} m_{i,j}^{\langle 2 \rangle}(\boldsymbol{\lambda}^{(t)}, \bar{C})}{\sigma_{i,j}^2} \tag{51}$$

$$Y_i = \sum_j (b_i - \mu_{i,j}) \frac{\sum_x \{x q_{i,j}(x, \boldsymbol{\lambda}^{(t)})\} + M^{(t)} m_{i,j}^{\langle 1 \rangle} \boldsymbol{\lambda}^{(t)}, \bar{C})}{\sigma_{i,j}^2} \tag{52}$$

$$Z_i = (N + M^{(t)}) \xi_i^{(t+1)} \sum_j \zeta_{i,j}. \tag{53}$$

And from (41),

$$b_i^{(t+1)} = \frac{V_i}{U_i}, \tag{54}$$

where

$$U_i = \sum_j \frac{\sum_x \{q_{i,j}(x, \boldsymbol{\lambda}^{(t)})\} + M^{(t)} m_{i,j}^{\langle 0 \rangle}(\boldsymbol{\lambda}^{(t)}, \bar{C})}{\sigma_{i,j}^2}, \tag{55}$$

$$V_i = \sum_j \frac{\sum_x \{(\mu_{i,j} - a_i x) q_{i,j}(x, \boldsymbol{\lambda}^{(t)})\} + M^{(t)} \{\mu_{i,j} m_{i,j}^{\langle 0 \rangle}(\boldsymbol{\lambda}^{(t)}, \bar{C}) - a_i m_{i,j}^{\langle 1 \rangle}(\boldsymbol{\lambda}^{(t)}, \bar{C})\}}{\sigma_{i,j}^2}. \tag{56}$$

The k-th moments (48) for $k = 0, 1, 2$ can be calculated from

$$\mathrm{G}_k(x \mid \mu, \sigma^2) = \int_{-\infty}^{x} x^k \mathrm{g}(x \mid \mu, \sigma^2)\, \mathrm{d}x. \tag{57}$$

Usually, the cumulative density function G_0 can be provided in a lot of numerical libraries.

$$\mathrm{G}_0(x \mid \mu, \sigma^2) = \int_{-\infty}^{x} \mathrm{g}(x \mid \mu, \sigma^2)\, \mathrm{d}x. \tag{58}$$

By using G_0, higher order moments are given by

$$\mathrm{G}_1(x \mid \mu, \sigma^2) = \mu \mathrm{G}_0(x \mid \mu, \sigma^2) - \sigma^2 \mathrm{g}(x \mid \mu, \sigma^2), \tag{59}$$

$$\mathrm{G}_2(x \mid \mu, \sigma^2) = (\mu^2 + \sigma^2) \mathrm{G}_0(x \mid \mu, \sigma^2) - (\mu + x) \sigma^2 \mathrm{g}(x \mid \mu, \sigma^2). \tag{60}$$

5.6 The EM algorithm for a uniform distribution module

Let us consider an additional uniform distributed module to the above model. If the distribution is defined on C, there are no missing values out of C. Let ξ_0 be the weight value corresponding to the uniform module.

The recurrence formulae for other parameters of ξ_0 are unchanged from (49), (50) and (54), if $q(z, w \mid x, \boldsymbol{\lambda})$ of (31) is redefined by

$$q(z, w \mid x, \boldsymbol{\lambda}) = \frac{\xi_z \zeta_{z,w} \mathrm{g}(a_k x + b_k \mid \mu_{z,w}, \sigma_{z,w}^2)}{\xi_0 / S + \sum_{k,l} \xi_j \zeta_{k,l} \mathrm{g}(a_k x + b_k) \mid \mu_{k,l}, \sigma_{k,l}^2)}. \tag{61}$$

for $z \neq 0$, where S is the volume of C. For $z = 0$, let

$$q(0 \mid x, \boldsymbol{\lambda}) = \frac{\xi_0 / S}{\xi_0 / S + \sum_{k,l} \xi_j \zeta_{k,l} \mathrm{g}(a_k x + b_k) \mid \mu_{k,l}, \sigma_{k,l}^2)}, \tag{62}$$

then the recurrence formula for ξ_0 is given by

$$\xi_0^{(t+1)} = \frac{1}{N + M^{(t)}} \sum_x q(0 \mid x, \boldsymbol{\lambda}^{(t)}). \tag{63}$$

6 Extension to higher dimension

Although we have written in one dimensional form in the preceding sections, higher dimension is more interesting case (e.g. two dimension for image understanding). We can generalize easily under the following assumptions.

- The correlation matrix of each normal distribution is a diagonal matrix.
- If we use an attention window, the window C is a rectangle that is parallel to each axis in order to calculate the moments explicitly by $G_0(x \mid \mu, \sigma^2)$ in (58).
- Scale parameter a is not generalized to include rotation.

7 Conclusion

We have presented the EM algorithm for a mixture model with mixture modules that can approximate any kind of the model of object. We also considered the case that data are restricted in a region.

Let us conclude this paper by stating some problems and future works.

It is desired that the model is extended to the case of supervised learning. However, the model distribution often becomes a highly curved exponential family and we cannot obtain the explicit solution in each EM step.

About attention, we must consider how to decide the attention region. That problem is related to the field of active vision or active learning, and it would be decided mainly by the restriction of computational resources and the amount of expected information.

It is also a future task to apply our algorithm to real world images.

Acknowledgement

The author would like to thank Dr. Suwa, Director of Information Science Division of Electrotechnical Laboratory, for affording an opportunity of this study. He also expresses his thanks to all members of Mathematical Informatics Section and Real World Computing Project Team of Electrotechnical Laboratory for their helpful discussions. A part of this work is supported by Real World Computing Program.

References

[1] Amari, S.: *Differential Geometrical Methods in Statistics*. Lecture Notes in Statistics. Springer-Verlag, 1985.

[2] Amari, S.: Information geometry of the EM and em algorithms for neural networks. Technical Report METR 94–04, University of Tokyo, 1994. (to appear in Neural Networks).

[3] Amari, S., Kurata, K., and Nagaoka, H.: Information geometry of Boltzmann machines. *IEEE Trans. Neural Networks*, Vol. 3, No. 2,, 1992.

[4] Baum, L., Petrie, T., Soules, G., and Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, Vol. 41, No. 1, pp. 164–171, 1970.

[5] Dempster, A., Laird, N., and Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.* B, Vol. 39, pp. 1–38, 1977.

[6] Jordan, M. I. and Jacobs, R. A.: Hierarchical mixtures of experts and the EM algorithm. In *Proc. of IJCNN'93*, pp. 1339–1344, Nagoya, 1993.

[7] Jordan, M. I. and Xu, L.: Convergence results for the EM approach to mixtures of experts architectures. MIT A.I. Memo No. 1458, 1993.

[8] McCullagh, P. and Nelder, J.: *Generalized Linear Models.* Chapman and Hall, 1983.