Contents

Chapter 1

The Expectation-Maximization (EM) Algorithm

Notation

- $1 \leq n \leq N$ - time index and range
- $0 \leq m < M$ - state index and range
- $\Re^L$ - observation space
- $Y_n \in \Re^L$ for $n = 1, \cdots, N$ - observed data and associated space
- $X_n \in \Omega$ with $|\Omega| = M$ - unobserved data and associated space
- $\theta^{(k)}$ - result of k^{th} EM iteration
- N_m, b_m, R_m - sufficient statistics for each class
- $\bar{N}_m$, $\bar{b}_m$, $\bar{R}_m$ - expected sufficient statistics for each class
- L - dimension of observation vector
- θ, $\tilde{\theta}$ - new and old parameters respectively
- π_m, μ_m, σ_m^2 - parameters for m^{th} mixture component
- $\theta = [\pi_0, \mu_0, \sigma_0^2, \cdots, \pi_{M-1}, \mu_{M-1}, \sigma_{M-1}^2]$ - full parameter vector

Hopefully, the previous chapters have convinced you that the methods of model-based image processing are quite promising. However, one crucial

question that arises is how should the model be chosen? Clearly, a poor choice of models will produce bad results, so this is a critical question.

One solution to this problem is to select a broad family of models that are controlled by a scalar or vector parameter. So for example, the prior model for an image may be controlled by a parameter which determines properties such as variation or edginess. These model parameters can then be estimated along with the unknown image. If the family of models is sufficiently broad, and the parameter estimation is sufficiently robust, then this approach should provide a general solution to the model selection problem.

However, there is still a catch with this approach: The undistorted image is typically not available to measure. In fact, if it were available, then we would not need to solve the the image restoration or reconstruction problem in the first place. This means that the parameters that specify a particular prior model must be estimated from indirect measurements of the corrupted, distorted, or transformed image.

In fact, the problem of estimating parameters from indirect or incomplete observations is a recurring theme in model-based image processing. Fortunately, there are some powerful tools that can be used to address this problem. Among these tools, the **expectation-maximization (EM)** algorithm is perhaps the most widely used. At its core, the EM algorithm provides a two-step process for parameter estimation. In the first step, one hypothesizes the knowledge of the missing data and uses it to calculate an ML estimate of the parameter (M-step); and in the second step, the hypothesized missing data is corrected using the previously estimated parameter values (E-step).

This cyclic process is quite intuitive, but will it work? In fact, naive implementation of this approach leads very bad estimates. However, the EM algorithm provides a formal framework for this iterative heuristic, which insures well defined properties of convergence and good behavior. The following section provides the formal justification behind the EM algorithm, along with an informal perspective of its value in imaging applications.

1.1 Introduction and Motivation for EM Algorithm

Imagine the following problem. You have measured the height of each plant in a garden. There are N plants, and you know that some have been regularly

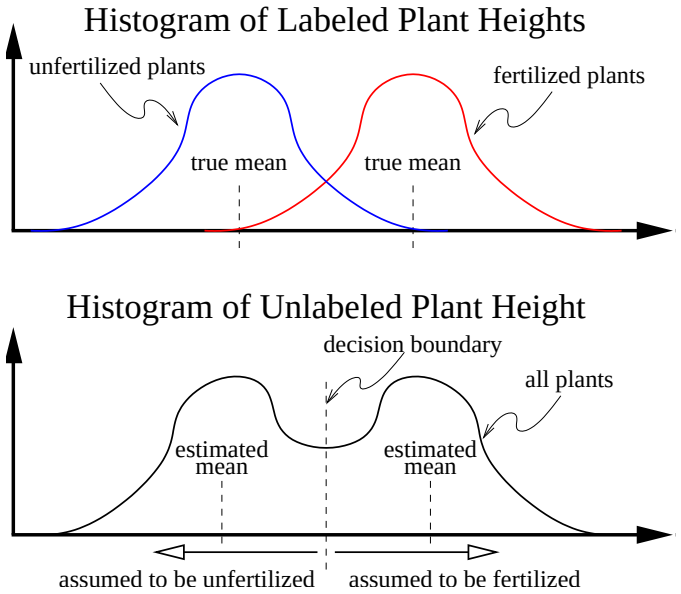

Figure 1.1: Example of the distribution we might expect in plant height for the two populations. Notice that the distributions for fertilized and unfertilized plants is different, with the fertilized plants having a larger mean height. However, if the labels of the plants are unknown, then the distribution appears to have two modes, as shown in the second plot. Notice that naive estimates of the mean, are systematically offset from the true values.

fertilized, and the remainder have not been fertilized at all. Unfortunately, the fertilizing records were lost. Your measurements, Y_n, of the plant height for the n^{th} plant could have been modeled as Gaussian with mean μ and variance, σ^2 if they had all been treated equally; but since they have not, the fertilized plants will, on average, be taller.

We can model this unknown treatment of the plants by a random variable X_n which is 1 if the plant has been fertilized, and 0 if it has not. If X_n where known, then the distribution of plant heights for each category could be assumed to be Gaussian, but with different means, and perhaps the same variance σ^2. This implies that the conditional distribution of Y_n given X_n will be Gaussian,

$$p(y_n|x_n) = \begin{cases} \frac{1}{z}\exp\left\{-\frac{1}{2\sigma^2}(y_n-\mu_1)^2\right\} & \text{for } x_n = 1 \\ \frac{1}{z}\exp\left\{-\frac{1}{2\sigma^2}(y_n-\mu_0)^2\right\} & \text{for } x_n = 0 \ , \end{cases}$$

with μ_0 and μ_1 denoting the means of the two different classes.

It might also be reasonable to assume the the class labels of the plants,

X_n, are i.i.d. Bernoulli random variables with

$$p(x_n) = \begin{cases} \pi_1 & \text{for } x_n = 1 \\ \pi_0 = 1 - \pi_1 & \text{for } x_n = 0 \end{cases},$$

where π_i parameterizes the Bernoulli distribution.

This type of stochastic process is sometimes referred to as a **doubly stochastic process** due to its hierarchical structure, with the distribution of Y_n being dependent on the value of the random variable X_n. The parameters of this doubly stochastic process (X_n, Y_n) are then given by $\theta = [\pi_0, \mu_0, \pi_1, \mu_1]$ where $\pi_0 + \pi_1 = 1$.

The question arises of how to estimate the parameter θ of this distribution? This is an important practical problem because we may want to measure the effect of fertilization on plant growth, so we would like the know how much μ_0 and μ_1 differ. Figure 1.1 illustrates the situation. Notice that the two populations create two modes in the distribution of plant height. In order to estimate the mean of each mode, it seems that we would need to know X_n, the label of each plant. However, casual inspection of the distribution of Fig. 1.1 suggests that one might be able to estimate the unknown means, μ_0 and μ_1, by looking at the combined distribution of the two populations.

One possibility for estimating μ_0 and μ_1 is to first estimate the labels X_n. This can be done by applying a threshold at the valley between the two modes, and classifying the value.

$$\hat{X}_n = \begin{cases} 0 & Y_n < \text{threshold} \\ 1 & Y_n \geq \text{threshold} \end{cases} \tag{1.1}$$

The result of this classification can be more compactly represented by the two sets $S_0 = \{n : \hat{X}_n = 0\}$ and $S_1 = \{n : \hat{X}_n = 1\}$. Using this notation, we can the estimate the two unknown means as

$$\begin{aligned} \hat{\mu}_0 &= \frac{1}{|S_0|} \sum_{n \in S_0} Y_n \\ \hat{\mu}_1 &= \frac{1}{|S_1|} \sum_{n \in S_1} Y_n \end{aligned},$$

where $|S_0|$ and $|S_1|$ denote the number of plants that have been classified as unfertilized and fertilized respectively.

While this is an intuitively appealing approach, it has a very serious flaw. Since we have separated the two groups by their height, it is inevitable that we will measure a larger value for μ_1 than we should (and also a smaller value of μ_0 than we should). In fact, even when the two means are quite different the resulting estimates of $\hat{\mu}_0$ and $\hat{\mu}_1$ will be systematically shifted no matter how many plants, N, are measured. This is much worse than simply being biased. These estimates are inconsistent because the estimated values do not converge to the true parameters as $N \to \infty$.

Another approach to solving this estimation problem is to attempt to directly estimate the value of the parameter vector $\theta = [\pi_0, \mu_0, \pi_1, \mu_1]$ from the data Y using the ML estimate. Formally, this can be stated as

$$\hat{\theta} = \arg \max_{\theta \in \Omega} \log p(y|\theta) \ .$$

This seems to be a much more promising approach since it is known that the ML estimate is not only consistent, but it is asymptotically efficient, which means that asymptotically it achieves the accuracy of the Cramer-Rao bound[4]. However, there is still a problem. The distribution $p(y|\theta)$ is not explicitly available for this problem. In fact, it requires the evaluation of a sum that makes direct ML estimation difficult.

$$p(y|\theta) = \sum_x p(y|x,\theta)p(x|\theta)$$

So we will see that it is no longer possible to calculate simple closed form expressions for the ML estimate of θ.

The purpose of the expectation-maximization (EM) algorithm is to provide a systematic methodology for estimating parameters, such as μ_k, when there is missing data. For this problem, we refer to the unknown labels, X, as the **missing data**; the combination of (X, Y) as the **complete data**; and Y along as the **incomplete data**. The EM algorithm provides a methodology for determining an ML estimate of the parameter vector, θ, from the incomplete data Y. We will see that the method is quite clever and perhaps surprising, but it results in some very intuitive algorithms when it is properly applied. In fact, the EM algorithm is more than a simple algorithm. It is a way of thinking about incomplete observations of data, and the resulting optimization algorithms required for ML parameter estimation.

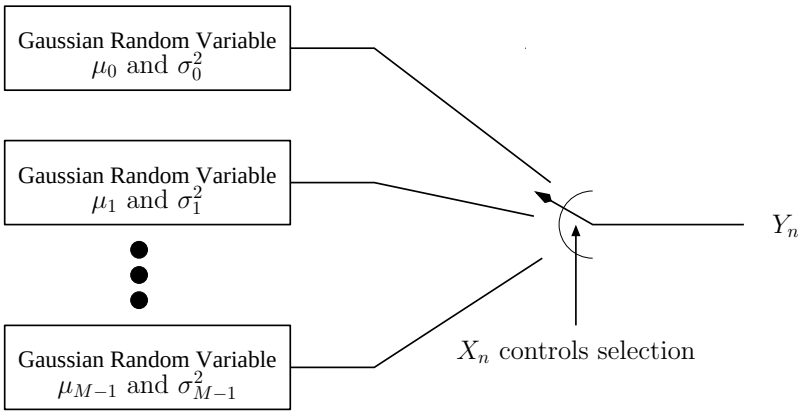

Figure 1.2: This figure graphically illustrates the creation of a random variable Y_n with a Gaussian mixture distribution. The output is selected among M possible Gaussian distributions using a switch. The switch is controlled using an independent random variable, X_n, that represents the class label.

1.2 Gaussian Mixture Distributions

Imagine an generalization of our example of the previous section in which the random variables $\{Y_n\}_{n=1}^N$ are assumed to be conditionally Gaussian and independent given the class labels $\{X_n\}_{n=1}^N$. In order to be a bit more general, we will assume that each label takes on M possible values. So that $X_n \in \{0, \cdots, M-1\}$.

Figure 1.2 graphically illustrates the behavior of the model. Each observation can come form one of M possible Gaussian distributions, and the section of the specific distribution is made using a switch controlled by the class label X_n. Using this model, the conditional distribution of Y_n given X_n is given by

$$p(y_n|x_n) = \frac{1}{\sqrt{2\pi\sigma_{x_n}^2}} \exp\left\{-\frac{1}{2\sigma_{x_n}^2}(y_n - \mu_{x_n})^2\right\} \ ,$$

where μ_{x_n} and $\sigma_{x_n}^2$ are the mean and variance of the sample when x_n is the label. From this, we can calculate the marginal distribution of y_n given by

$$\begin{aligned} p(y_n) &= \sum_{m=0}^{M-1} p(y_n|m)\pi_m \\ &= \sum_{m=0}^{M-1} \frac{\pi_m}{\sqrt{2\pi\sigma_m^2}} \exp\left\{-\frac{1}{2\sigma_m^2}(y_n - \mu_m)^2\right\} \ , \end{aligned} \tag{1.2}$$

where $\pi_m = P\{X_n = m\}$. If we further assume the the samples, Y_n, are i.i.d.

then the distribution of the entire sequence, $\{Y_n\}_{n=0}^{N}$, can be written as

$$p(y) = \prod_{n=1}^{N} \sum_{m=0}^{M-1} \frac{\pi_m}{\sqrt{2\pi\sigma_m^2}} \exp\left\{ -\frac{1}{2\sigma_m^2}(y_n - \mu_m)^2 \right\} \ . \tag{1.3}$$

A distribution with this form of (1.2) is known as a **Gaussian mixture distribution** with parameter vector $\theta = [\pi_0, \mu_0, \sigma_0^2, \cdots, \pi_{M-1}, \mu_{M-1}, \sigma_{M-1}^2]$. Gaussian mixtures are very useful because, for sufficiently large M, they can be used to approximate almost any distribution. This is because, intuitively, any density function can be approximated by a weighted sum of Gaussian density functions.

So the next question is, how to we estimate the parameter θ from the observations of Y_n? For now, we will make the simplifying assumption that we know both the values Y_n and their associated labels X_n. In order to calculate the ML estimate, we first compute N_m, the number of labels taking on class m.

$$N_m = \sum_{n=1}^{N} \delta(X_n - m)$$

From this, we can compute the ML estimate π_m as the fraction of labels taking on the class m.

$$\hat{\pi}_m = \frac{N_m}{N} \ .$$

The ML estimate of the mean for each class is then computed by only adding the values of Y_n with the class label $X_n = m$.

$$\hat{\mu}_m = \frac{1}{N_m} \sum_{n=1}^{N} Y_n \delta(X_n - m)$$

Finally, the ML estimate of the variance for each class is again computed by only summing over terms with the class label $X_n = m$.

$$\hat{\sigma}_m^2 = \frac{1}{N_m} \sum_{n=1}^{N} (Y_n - \hat{\mu}_m)^2 \, \delta(X_n - m)$$

Of course, for our problem the class labels X_n are not known, so the question remains of how we are going to calculate the ML estimate of θ from only the incomplete data Y.

The EM algorithm will provide an approach to computing the ML parameter estimation when we only have incomplete data. At its essence, the EM

algorithm works by replacing the unknown labels, X_n, with their expected value. However, in order to fully understand this, we will first need to develop the underlying theory of the method.

1.3 EM Algorithm Inequalities and Recursions

In this section, we derive the mathematical relationships that form the underpinnings of the EM algorithm. While these relationships may seem abstract, they are relatively simple to derive, and are very powerful, so the reader is strongly encourage to take the time to understand them. An early, succinct, and clear presentation of these results can be found in [2, 1].

The EM algorithm is based on two non-obvious mathematical insights. The first insight is that one can separate the log likelihood into the sum of two functions. In particular, the log likelihood can be expressed as

$$\log p(y|\theta) = Q(\theta, \tilde{\theta}) + H(\theta, \tilde{\theta}) \tag{1.4}$$

where $\tilde{\theta}$ is any value of the parameter, and the functions Q and H are defined as

$$\begin{aligned} Q(\theta, \tilde{\theta}) &\stackrel{\triangle}{=} \mathrm{E}\big[\log p(y, X|\theta)|Y = y, \tilde{\theta}\big] \\ H(\theta, \tilde{\theta}) &\stackrel{\triangle}{=} -\mathrm{E}\big[\log p(X|y, \theta)|Y = y, \tilde{\theta}\big] \ . \end{aligned}$$

where $p(y, x|\theta)$ is assumed to be strictly positive density function. We can prove this result using the following sequence of equalities.

$$\begin{aligned} \log p(y|\theta) &= \log \left\{ \frac{p(y, x|\theta)}{p(x|y, \theta)} \right\} \\ &= \log \left\{ \frac{p(y, X|\theta)}{p(X|y, \theta)} \right\} \\ &= \mathrm{E}\left[\log \left\{ \frac{p(y, X|\theta)}{p(X|y, \theta)} \right\} \middle| Y = y, \tilde{\theta} \right] \\ &= \mathrm{E}\big[\log p(y, X|\theta)|Y = y, \tilde{\theta}\big] - \mathrm{E}\big[\log p(X|y, \theta)|Y = y, \tilde{\theta}\big] \\ &= Q(\theta, \tilde{\theta}) + H(\theta, \tilde{\theta}) \end{aligned}$$

The first equality uses the fact that $p(y|\theta) = \frac{p(y,x|\theta)}{p(x|y,\theta)}$ for all values of the variable x. Since this ratio does not depend on the value of x, we can substitute

the random variable X, and the equality still holds. The remaining step simply use properties of the log and expectation.

The second insight is that that the function $H(\theta, \tilde{\theta})$ takes on its minimum value when $\theta = \tilde{\theta}$. More precisely, for all θ, $\tilde{\theta} \in \Omega$, we have that

$$H(\theta, \tilde{\theta}) \geq H(\tilde{\theta}, \tilde{\theta}) \ . \tag{1.5}$$

To see that this is true, we have the following set of inequalities

$$\begin{aligned}
0 &= \log\left\{\int p(x|y,\theta)dx\right\} \\
&= \log\left\{\int \frac{p(x|y,\theta)}{p(x|y,\tilde{\theta})} p(x|y,\tilde{\theta})dx\right\} \\
&\geq \int \log\left\{\frac{p(x|y,\theta)}{p(x|y,\tilde{\theta})}\right\} p(x|y,\tilde{\theta})dx \\
&= \int \log p(x|y,\theta)p(x|y,\tilde{\theta})dx - \int \log p(x|y,\tilde{\theta})p(x|y,\tilde{\theta})dx \\
&= -H(\theta,\tilde{\theta}) + H(\tilde{\theta},\tilde{\theta})
\end{aligned}$$

These two results of equations (1.4) and (1.5) now yield the fundamental result of the EM algorithm. If we adjust the parameter from an initial value of $\tilde{\theta}$ to a new value of θ, then the change in the log likelihood is lower bounded by the change in the Q function. This is formally stated as

$$\log p(y|\theta) - \log p(y|\tilde{\theta}) \geq Q(\theta,\tilde{\theta}) - Q(\tilde{\theta},\tilde{\theta}) \ . \tag{1.6}$$

The proof of this key result is quite simple.

$$\begin{aligned}
\log p(y|\theta) - \log p(y|\tilde{\theta}) &= Q(\theta,\tilde{\theta}) + H(\theta,\tilde{\theta}) - \left[Q(\tilde{\theta},\tilde{\theta}) + H(\tilde{\theta},\tilde{\theta})\right] \\
&= Q(\theta,\tilde{\theta}) - Q(\tilde{\theta},\tilde{\theta}) + H(\theta,\tilde{\theta}) - H(\tilde{\theta},\tilde{\theta}) \\
&\geq Q(\theta,\tilde{\theta}) - Q(\tilde{\theta},\tilde{\theta})
\end{aligned}$$

The inequality of (1.6) implies that by increasing the Q function, the log likelihood is also guaranteed to increase.

$$Q(\theta,\tilde{\theta}) > Q(\tilde{\theta},\tilde{\theta}) \Rightarrow p(y|\theta) > p(y|\tilde{\theta}) \tag{1.7}$$

From this result, the central concept of the EM algorithm becomes clear. Each iteration starts with an initial parameter $\tilde{\theta}$. Our objective is then to find a new parameter θ so that $Q(\theta,\tilde{\theta}) > Q(\tilde{\theta},\tilde{\theta})$. From equation (1.7), we

then know that this new parameter is then guaranteed to produce a larger value of the likelihood.

With this understanding in mind, we can now state the two-step recursion that defines the EM algorithm.

$$\text{E-step:} \quad Q(\theta, \theta^{(k)}) = \mathrm{E}\big[\log p(y, X|\theta)|Y = y, \theta^{(k)}\big] \tag{1.8}$$

$$\text{M-step:} \quad \theta^{(k+1)} = \arg\max_{\theta \in \Omega} Q(\theta, \theta^{(k)}) \tag{1.9}$$

Since each new value of the parameter $\theta^{(k)}$ is selected to maximize the Q function, we know that this iteration will produce a monotone increasing sequence of log likelihood values, $\log p(y|\theta^{(k+1)}) \geq \log p(y|\theta^{(k)})$.

1.4 EM for Gaussian Mixture Distributions

Now that we have the basic tools of the EM algorithm, we can use them to estimate the parameters of the Gaussian mixture distribution introduced in Section 1.2. In order to calculate the Q function for this problem, we will need a more suitable expression for the joint distribution of Y_n and X_n. It is easily verified that the following equality holds

$$\begin{aligned} \log p(y_n, x_n|\theta) &= \log\left\{p(y_n|\mu_{x_n}, \sigma_{x_n})\pi_{x_n}\right\} \\ &= \sum_{m=0}^{M-1} \delta(x_n - m)\left\{\log p(y_n|\mu_m, \sigma_m) + \log \pi_m\right\} \ , \end{aligned}$$

where $p(y_n|\mu_m, \sigma_m)$ denotes a Gaussian density with mean μ_i and standard deviation σ_m. Notice, that the delta function in the sum is only 1 when $x_n = m$, otherwise it is zero. Since the values of Y_n and X_n are assumed independent for different n, we know that

$$\begin{aligned} \log p(y, x|\theta) &= \sum_{n=1}^{N} \log p(y_n, x_n|\theta) \\ &= \sum_{n=1}^{N} \sum_{m=0}^{M-1} \delta(x_n - m)\left\{\log p(y_n|\mu_m, \sigma_m) + \log \pi_m\right\} \ . \end{aligned}$$

Using this expression, we can now calculate the Q function.

$$Q(\theta, \tilde{\theta}) = \mathrm{E}\big[\log p(y, X|\theta)|Y = y, \tilde{\theta}\big]$$

$$
\begin{aligned}
&= \mathrm{E}\left[\sum_{n=1}^{N}\sum_{m=0}^{M-1}\delta(X_n-m)\left\{\log p(y_n|\mu_m,\sigma_m)+\log\pi_m\right\}\middle| Y=y,\tilde{\theta}\right] \\
&= \sum_{n=1}^{N}\sum_{m=0}^{M-1} E\left[\delta(X_n-m)|Y=y,\tilde{\theta}\right]\left\{\log p(y_n|\mu_m,\sigma_m)+\log\pi_m\right\} \\
&= \sum_{n=1}^{N}\sum_{m=0}^{M-1} P\big\{X_n=m|Y=y,\tilde{\theta}\big\}\left\{\log p(y_n|\mu_m,\sigma_m)+\log\pi_m\right\}
\end{aligned}
$$

Now plugging in the explicit expression for the Gaussian distribution, we get the final expression for the Q function.

$$Q(\theta,\tilde{\theta}) = \sum_{n=1}^{N}\sum_{m=0}^{M-1}\left\{\frac{-1}{2\sigma_m^2}(y_n-\mu_m)^2 - \frac{1}{2}\log\left(2\pi\sigma_m^2\right)+\log\pi_m\right\} P\big\{X_n=m|Y=y,\tilde{\theta}\big\}$$

The posterior conditional probability, $P\big\{X_n=m|Y=y,\tilde{\theta}\big\}$, can be easily computed using Bayes rule. If we define the new notation that $f(m|y_n,\tilde{\theta}) \stackrel{\triangle}{=} P\big\{X_n=m|Y=y,\tilde{\theta}\big\}$, then this function is given by

$$f(m|y_n,\tilde{\theta}) = \frac{\dfrac{1}{\sqrt{2\pi}\tilde{\sigma}_m}\exp\left\{-\dfrac{1}{2\tilde{\sigma}_m^2}(y_n-\tilde{\mu}_m)^2\right\}\tilde{\pi}_m}{\displaystyle\sum_{j=0}^{M-1}\frac{1}{\sqrt{2\pi}\tilde{\sigma}_j}\exp\left\{-\frac{1}{2\tilde{\sigma}_j^2}(y_n-\tilde{\mu}_j)^2\right\}\tilde{\pi}_j}\ .$$

While this expression for $f(m|y_n,\tilde{\theta})$ may appear complex, it is quite simple to compute numerically.

In order to calculate the associated M-step, we must maximize $Q(\theta,\tilde{\theta})$ with respect to the parameter θ. This maximization procedure is much the same as is required for ML estimation of parameters. So for example, μ_k and σ_k are calculated in much the same way as the ML estimate of mean and variance are calculated for a Gaussian random variable. This approach results in the final EM update equations for this clustering problem.

$$\hat{N}_m^{(k+1)} = \sum_{n=1}^{N} f(m|y_n,\theta^{(k)}) \tag{1.10}$$

$$\hat{\pi}_m^{(k+1)} = \frac{\hat{N}_m^{(k+1)}}{N} \tag{1.11}$$

$$\hat{\mu}_m^{(k+1)} = \frac{1}{\hat{N}_i^{(k+1)}}\sum_{n=1}^{N} y_n f(m|y_n,\theta^{(k)}) \tag{1.12}$$

$$\left[\hat{\sigma}_m^2\right]^{(k+1)} = \frac{1}{\hat{N}_m^{(k+1)}}\sum_{n=1}^{N}\left(y_n-\hat{\mu}_m^{(k+1)}\right)^2 f(m|y_n,\theta^{(k)}) \tag{1.13}$$

1.5 Algorithmic Implementation of EM Clustering

In order to better understand the EM algorithm, it is useful to take a closer look at its algorithmic implementation. To do this, we will use the algorithmic notation of pseudo code programming where θ is the variable that contains the current value of the parameter. Also, let $P_{n,m}$ be the program variable that contains the current value of the posterior probability $P\{X_n = m | Y = y, \theta\}$.

Using these variables, the M-step is computed by evaluating the following expression for all values of n and m.

$$\text{E-step:} \quad P_{n,m} \leftarrow \frac{\frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left\{-\frac{1}{2\sigma_m^2}(y_n - \mu_m)^2\right\} \pi_m}{\sum_{j=0}^{M-1} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{1}{2\sigma_j^2}(y_n - \mu_j)^2\right\} \pi_j}$$

Notice that $\leftarrow$ indicates that the value is assigned to the program variable $P_{n,m}$. Once this is computed, we can update the model parameters by evaluating the following expressions for all values of m.

$$\text{M-step:} \quad N_m \leftarrow \sum_{n=1}^{N} P_{n,m} \tag{1.14}$$

$$\pi_m \leftarrow \frac{N_m}{N} \tag{1.15}$$

$$\mu_m \leftarrow \frac{1}{N_m} \sum_{n=1}^{N} y_n P_{n,m} \tag{1.16}$$

$$\sigma_m^2 \leftarrow \frac{1}{N_m} \sum_{n=1}^{N} (y_n - \mu_m)^2 P_{n,m} \tag{1.17}$$

In this form, the meaning of the EM algorithm starts to become more clear as shown in Fig. 1.3. With each update, we use the current estimate of the parameter, θ, to compute the matrix $P_{n,m}$, the probability that X_n has label m. These posterior probabilities are then used to assign each data point, Y_n, to the associated M clusters. Since the assignment is soft, each data point typically has partial membership in each cluster.

Once the data points are assign to the clusters, then the parameters of each cluster, $[\pi_m, \mu_m, \sigma_m^2]$, can be updated based on the weighted contributions of each sample. This results in an updated value for the parameter θ. However,

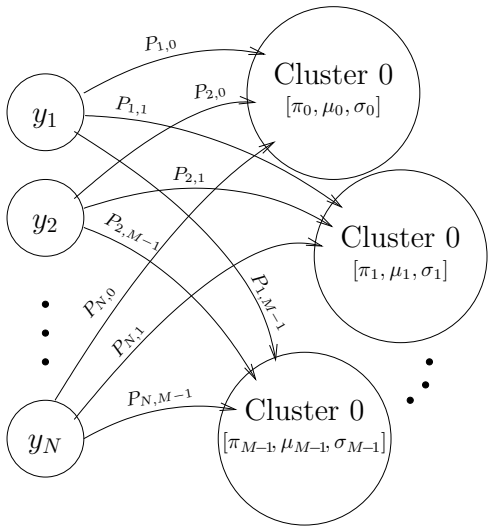

Figure 1.3: This figure illustrates the structure of the EM algorithm updates. In the E-step, the matrix $P_{n,m}$ is computed which contains the posterior probability that $X_n = m$. The entrees $P_{n,m}$ represent the soft assignment of each sample, Y_n, to a cluster m. Armed with this inform, the parameters of each cluster, (π_m, μ_m, σ_m), may be updated in the M-step.

with this new parameter, we must compute a new value for the matrix $P_{n,m}$, and the iterative process repeats.

In fact, this is a generalization of the heuristic method describe in the beginning of the chapter, except the hard classification heuristic of (1.1) is replaced with the software classification of the posterior probability. So the E-step can be viewed as a form of soft classification of the data to the available categories, while the M-step can be viewed as ML parameter estimation given those soft classifications.

1.6 EM Convergence and Substitute Functions

The previous section gives some intuition into the form of the EM iterations, but it does not fully explain why or how the EM algorithm converges. In order to better understand the convergence of the EM algorithm, it is useful to slightly modify the forms of the Q and H functions introduced in Section 1.3. So let us define the following modified functions.

$$\tilde{Q}(\theta, \tilde{\theta}) = Q(\theta, \tilde{\theta}) + H(\tilde{\theta}, \tilde{\theta})$$
$$\tilde{H}(\theta, \tilde{\theta}) = H(\theta, \tilde{\theta}) - H(\tilde{\theta}, \tilde{\theta})$$

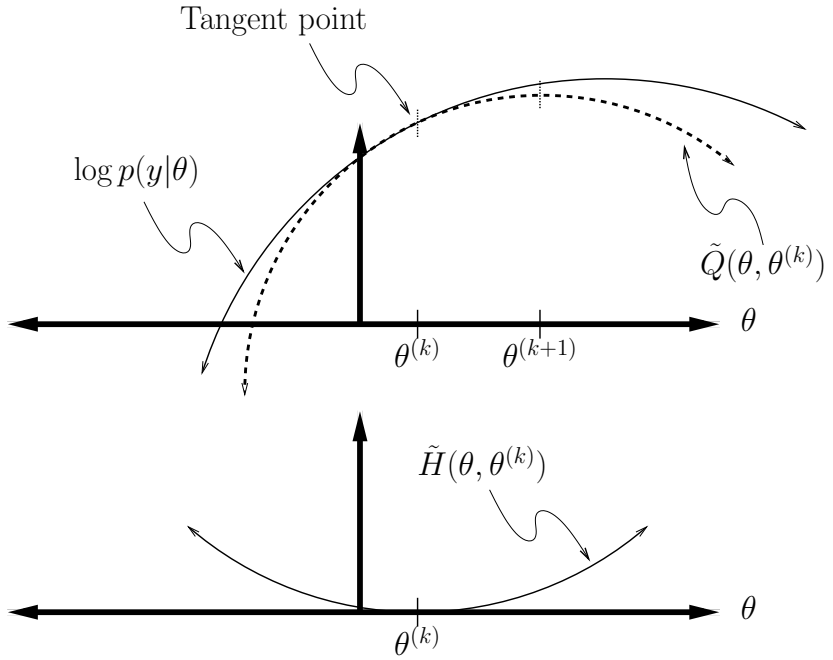

Figure 1.4: This illustrates why the EM algorithm increases the log likelihood with each step, and why it converges to a local maximum of the likelihood.

Since the EM updates only require optimization with respect to θ, these modified functions yield the same update for the M-step.

$$\theta^{(k+1)} = \arg\max_{\theta\in\Omega} Q(\theta, \theta^{(k)}) = \arg\max_{\theta\in\Omega} \tilde{Q}(\theta, \theta^{(k)})$$

As before, the log likelihood function is equal to the sum of these new functions

$$\log p(y|\theta) = \tilde{Q}(\theta, \tilde{\theta}) + \tilde{H}(\theta, \tilde{\theta}) \ .$$

However, we can now show that H has two important properties.

$$\tilde{H}(\theta, \tilde{\theta})\big|_{\theta=\tilde{\theta}} = 0 \tag{1.18}$$

$$\nabla\tilde{H}(\theta, \tilde{\theta})\big|_{\theta=\tilde{\theta}} = 0 \ , \tag{1.19}$$

where ∇ denotes the gradient with respect to the first argument of the function. The first property of (1.18) results from the modified definition of $\tilde{H}$. In order to see why the second property of (1.19) holds, notice that by (1.5), we know that the $\tilde{H}(\theta, \tilde{\theta})$ achieves a global minimum at $\theta = \tilde{\theta}$. Since $\tilde{H}$ is continuously differentiable, this implies that $\tilde{H}$ must have a gradient of zero at $\theta = \tilde{\theta}$. Figure 1.4 illustrates the situation. Notice that since $\tilde{H}$ has a minimum at $\theta = \tilde{\theta}$, its derivative must also be zero at this point.

From this, we get the next important result.

$$\begin{aligned} \nabla \log p(y|\tilde{\theta}) &= \nabla \tilde{Q}(\theta, \tilde{\theta})\big|_{\theta=\tilde{\theta}} + \nabla \tilde{H}(\theta, \tilde{\theta})\big|_{\theta=\tilde{\theta}} \\ &= \nabla \tilde{Q}(\theta, \tilde{\theta})\big|_{\theta=\tilde{\theta}} \end{aligned}$$

So at the point $\theta = \tilde{\theta}$, the gradient of the Q function is equal to the gradient of the log likelihood. Putting together all these properties of Q we have that

$$\log p(y|\tilde{\theta}) = \tilde{Q}(\tilde{\theta}, \tilde{\theta}) \tag{1.20}$$

$$\nabla \log p(y|\tilde{\theta}) = \nabla \tilde{Q}(\theta, \tilde{\theta})\big|_{\theta=\tilde{\theta}} \tag{1.21}$$

$$\log p(y|\theta) \geq \tilde{Q}(\theta, \tilde{\theta}) \ , \tag{1.22}$$

where the first property results from the modified definition of $\tilde{H}$, and the last property results from the positivity of $\tilde{H}$. In fact any function, $\tilde{Q}$, with these three properties is known as a **substitute function** for the log likelihood because these three properties allow one to substitute $\tilde{Q}$ for the true likelihood function while still guaranteeing monotone increase of the log likelihood.

Figure 1.4 illustrates the intuition behind the use of substitute functions. Notice that since the function $\tilde{Q}$ lower bounds $\log p(y|x)$, it is clear that increasing the value of $\tilde{Q}$ must increase the value of the log likelihood. Also notice that if the we find a value of *theta* so that $\nabla \tilde{Q}(\theta, \theta) = 0$, then we know that $\nabla \log p(y|\theta) = 0$, which in term implies a local maximum of the log likelihood.

However, this begs the next question, which is does the EM algorithm converge to the ML estimate of θ? It is difficult to give a simple and clear answer to this question because in optimization (as in life) many things can go wrong. For example, the iterations can become trapped in a local minimum of the likelihood. Even if the likelihood has no local minimum, there are strange technical problems that can occur, so it is difficult to make a simple statement about the convergence to the ML estimate that is absolutely true under all conditions.

But with that said, as a practical matter, the EM algorithm generally does converge to a local maximum of the log likelihood in typical estimation problems. For interested readers, the two references [5, 3] provide a more detailed introduction to the theory of convergence for the EM algorithm. However, in order to give some intuition regarding the convergence, we present some simplified results to illustrate the methods of proof.

Define the function $L(\theta) = \log p(y|\theta)$, then we know that for each iteration of the EM algorithm

$$L(\theta^{(k+1)}) \geq L(\theta^{(k)}) \ .$$

If the ML estimate exists, then for all k, we also know that $L(\theta^{(k)}) \leq L(\theta_{ML}) < \infty$. Therefore, the sequence $L(\theta^{(k)})$ must be a monotone increasing sequence that is bounded above, so it must therefore reach a limit which we denote by $L^* = \lim_{k\to\infty} L(\theta^{(k)})$. However, even though the limit exists, proving that that value L^* is the maximum of the likelihood estimate, or that the $\lim_{k\to\infty} \theta^{(k)}$ exists is a much more difficult matter.

The following simplified theorem gives some insight into the convergence of the EM algorithm. The basic result is that under normal conditions, the EM algorithm converges to a parameter value for which the gradient of the likelihood is zero. Typically, this means that the EM algorithm converges to a local or global maximum of the likelihood function.

Theorem 1 *Let $\theta^{(k)} \in \Omega$ be a sequence of parameter values generated by the EM updates. Assume that a) Ω is a open set, b) the functions Q and H are continuously differentiable on Ω^2 and the log likelihood is a strictly positive function, c) $\lim_{k\to\infty} \theta^{(k)} = \theta^*$ exists for $\theta^* \in \Omega$. Then*

$$\nabla \log p(y|\theta^*) = 0 \ .$$

Proof: First, we know that by the nature of the M-step, the updated parameter value, $\theta^{(k+1)}$, is a global, and therefore, local maximum of the Q function. This implies that

$$\begin{aligned} 0 &= \nabla Q(\theta, \theta^{(k)})\big|_{\theta=\theta^{(k+1)}} \\ &= \lim_{k\to\infty} \nabla Q(\theta^{(k+1)}, \theta^{(k)}) \\ &= \nabla Q\Big(\lim_{k\to\infty} \theta^{(k+1)}, \lim_{k\to\infty} \theta^{(k)}\Big) \\ &= \nabla Q(\theta^*, \theta^*) \\ &= \nabla \log p(y|\theta^*) \ . \end{aligned}$$

The assumptions of this proof are a bit artificial, but the result servers to illustrate the basic concepts of convergence with a very simple proof. In particular, assumption a) is used to insure that the solution does not fall on the boundary of a closed set because, if this happened, the gradient would

no longer need to be zero at a global maximum. Of course, it is possible to handle the case of solutions on a boundary, and in fact, this often occurs; however, it requires the use of the Karush-Kuhn-Tucker (KKT) conditions, which substantially complicates the analysis.

1.7 Simplified Methods for Deriving EM Updates

While the calculation of the Q function is typically complex, it is clear that there is a general pattern to the result. For example, in the clustering example of Section 1.2 the final EM update equations appear much like the conventions ML estimates, except that the means are weighted by the probability that a sample is from the particular class. This pattern turns out to have a underlying explanation which can be used to make derivation of the EM updates much simpler. In fact, it turns out that for all exponential distributions, the form of the EM update is quite simple.

In the following two sections, we derive and explain a easy method for determining the EM update equations for any exponential distribution. Armed with this technique, you will be able to write down the EM algorithm for most common distributions, without any need to calculate the Q function.

1.7.1 Exponential Distributions and Sufficient Statistics

In order to derive the simplified form of the EM update, we first must introduce the concept of an exponential distribution and its natural sufficient statistics. We start with two basic definitions.

> *Definition:* A **statistic** is any function, $T(Y)$, of the random vector Y.
>
> *Definition:* Let $p(y|\theta)$ be an family of density functions that is parameterized by $\theta \in \Omega$. A statistic $T(Y)$ is said to be **sufficient** for the parameter θ if there exist functions $g(\cdot,\cdot)$ and $h(\cdot)$ such that
>
> $$p(y|\theta) = h(y)\, g(T(y), \theta) \tag{1.23}$$
>
> for all $y \in I\!R^N$ and $\theta \in \Omega$.

Inituitively, a sufficient statistic distills all the information from the data, Y, necessary to estimate the parameter θ. For example, the ML estimator of

θ must be a function of the sufficient statistic $T(Y)$. To see this, notice that

$$\begin{aligned}\hat{\theta}_{ML} &= \arg\max_{\theta\in\Omega} \log p(y|\theta) \\ &= \arg\max_{\theta\in\Omega} \{\log h(y) + \log g(T(y),\theta)\} \\ &= \arg\max_{\theta\in\Omega} \log g(T(y),\theta) \\ &= f(T(y)) \ ,\end{aligned}$$

for some function $f(\cdot)$.

Many commonly used distributions such as Gaussian, exponential, Poisson, Bernoulli, and binomial have a structure which makes them particularly useful. These distributions are known as exponential families and have the following special property.

> *Definition:* A family of density functions $p(y|\theta)$ for $y \in R^N$ and $\theta \in \Omega$ is said to be a **k-parameter exponential family** if there exist functions $g(\theta) \in I\!R^k$, $s(y)$, $d(\theta)$ and statistic $T(y) \in I\!R^k$ such that
>
> $$p(y|\theta) = \exp\{< g(\theta), T(y) > + d(\theta) + s(y)\} \tag{1.24}$$
>
> for all $y \in I\!R^N$ and $\theta \in \Omega$ where $< \cdot, \cdot >$ denotes an inner product. We refer to $T(y)$ as the **natural sufficient statistic** or **natural statistic** for the exponential distribution.

Exponential distributions are extremely valuable because the log of its density forms an inner product that is easily manipulated when computing ML parameter estimates. Below are some examples of exponential distributions and their natural sufficient statistics.

Example 1.1: Let $\{Y_n\}_{n=1}^N$ be i.i.d. random vectors of dimension L with multivariate Gaussian distribution $N(\mu, R)$ and parameter $\theta = [\mu, R]$. Then $p(y|\theta)$ is an exponential distribution with natural sufficient statistics

$$\begin{aligned} b &= \sum_{n=1}^{N} y_n \\ S &= \sum_{n=1}^{N} y_n y_n^t \ . \end{aligned}$$

To see why this is true, we can use the result of equation (**??**) to write the density function for Y as

$$p(y|\theta) = \frac{1}{(2\pi)^{NL/2}}|R|^{-N/2}\exp\left\{-\frac{1}{2}\left(\mathrm{tr}\left\{SR^{-1}\right\} - 2b^tR^{-1}\mu + N\mu^tR^{-1}\mu\right)\right\} \ .$$

which is the form of (1.24) required by a natural sufficient statistic of the distribution. From these sufficient statistics, we can also compute the ML estimate, $\hat{\theta} = [\hat{\mu}, \hat{R}]$, as

$$\begin{aligned}\hat{\mu} &= b/N \\ \hat{R} &= (S/n) - \hat{\mu}\hat{\mu}^t \ .\end{aligned}$$

Of course, i.i.d. Gaussian random variables are just a special case of this example when the vector dimension is $L = 1$.

Example 1.2: Let $\{X_n\}_{n=0}^{N-1}$ be i.i.d. random variables which take on the discrete values in the set $\{0, \cdots, M-1\}$. The distribution of X_n is parameterized by $\theta = [\pi_0, \cdots, \pi_{M-1}]$ where $\pi_i = P\{X_n = i\}$. Then $p(x|\theta)$ is an exponential distribution with natural sufficient statistics

$$N_m = \sum_{n=1}^{N}\delta(X_n - m) \ .$$

To see why this is true, we can write the density function as

$$p(x|\theta) = \prod_{m=0}^{M-1}\pi_m^{N_m}$$

because N_m counts the number of times X_n takes on the value m. This can then be rewritten as

$$p(x|\theta) = \exp\left\{\sum_{m=0}^{M-1}N_m\log\pi_m\right\} \ ,$$

which is the form of (1.24) required by a natural sufficient statistic of the distribution. Again, from these sufficient statistics we can compute the ML estimate, $\hat{\theta} = [\hat{\pi}_0, \cdots, \hat{\pi}_{M-1}]$, as

$$\hat{\pi}_m = \frac{N_m}{N} \ .$$

Example 1.3: Here we consider two random processes, Y_n and X_n, with a structure similar to those used in the clustering problem of Sections 1.4 and 1.5, except we generalize the problem slightly by assuming that Y_n is a multivariate Gaussian random vector with conditional mean and covariance given by μ_m and R_m when $X_n = m$. Then $p(y, x|\theta)$ is an exponential distribution with natural sufficient statistics given by

$$N_m = \sum_{n=1}^{N} \delta(x_n - m) \tag{1.25}$$

$$b_m = \sum_{n=1}^{N} y_n \delta(x_n - m) \tag{1.26}$$

$$S_m = \sum_{n=1}^{N} y_n y_n^t \delta(x_n - m) \ . \tag{1.27}$$

The ML parameter estimate, $\hat{\theta}$, can then be computed from these sufficient statistics as

$$\hat{\pi}_m = \frac{N_m}{N} \tag{1.28}$$

$$\hat{\mu}_m = \frac{b_m}{N_m} \tag{1.29}$$

$$\hat{R}_m = \frac{S_m}{N_m} - \frac{b_m b_m^t}{N_m^2} \ . \tag{1.30}$$

The proof of these facts is left as an exercise.

Of course the problem with the ML estimates of (1.28), (1.29), and (1.30) is that we we may not know the labels X_n. This is the incomplete data problem that the EM algorithm addresses. In the next section, we will see how the EM algorithm can be easily applied for any such exponential distribution.

1.7.2 EM Update for Exponential Distributions

Let Y is the observed or incomplete data, and let X be the unobserved data. Then the EM updates turn out to have a particularly simple form, when (X, Y) have a distribution from an exponential family with parameter θ.

Since (X, Y) is assumed to be from an exponential family, then we know that

$$p(y, x|\theta) = \exp\{< g(\theta), T(y, x) > +d(\theta) + s(y, x)\}$$

for some natural sufficient statistic $T(y, x)$. Assuming the ML estimate of θ exists, then it is given by

$$\begin{aligned} \theta_{ML} &= \arg\max_{\theta\in\Omega} \{< g(\theta), T(y, x) > +d(\theta)\} \\ &= f(T(y, x)) \end{aligned} \tag{1.31}$$

where $f(\cdot)$ is some function of $T(y, x)$.

Recalling the form of the Q function, we have

$$Q(\theta, \tilde{\theta}) = E\left[\log p(y, X|\theta)|Y = y, \tilde{\theta}\right]$$

where Y is the observed data and X is the unknown data. Using the assumed structure of the exponential distribution, we have that

$$\begin{aligned} Q(\theta, \tilde{\theta}) &= E\left[\log p(y, X|\theta)|Y = y, \tilde{\theta}\right] \\ &= E\left[< g(\theta), T(y, X) > +d(\theta) + s(y, X)|Y = y, \tilde{\theta}\right] \\ &= < g(\theta), \bar{T}(y) > +d(\theta) + constant \end{aligned}$$

where

$$\bar{T}(y) = E\left[T(y, X)|Y = y, \tilde{\theta}\right]$$

is the conditional expectation of the sufficient statistic $T(y, X)$, and "*constant*" is a constant which does not depend on θ. Since our objective is to maximize Q with respect to θ, this constant can be dropped. A single update of the EM algorithm is then given by the recursion

$$\begin{aligned} \tilde{\theta}' &= \arg\max_{\theta\in\Omega} Q(\theta, \tilde{\theta}) \\ &= \arg\max_{\theta\in\Omega} \left\{< g(\theta), \bar{T}(y) > +d(\theta)\right\} \\ &= f(\bar{T}(y)) \end{aligned} \tag{1.32}$$

Intuitively, we see that the EM update of (1.32) has the same form as the computation of the ML estimate in equation (1.31), but with the expected value of the statistic, $\bar{T}$, replacing the actual statistic, T. To see how useful this result can be, we apply it to extend the clustering example of Sections 1.4 and 1.5.

Example 1.4: Let Y_n and X_n be as in the previous Example 1.3 in this chapter. So $\{Y_n\}_{n=0}^{N-1}$ are conditionally i.i.d. Gaussian random variables given the class labels $\{X_n\}_{n=1}^{N}$ with conditional mean and covariance given by μ_m and R_m when $X_n = m$; and X_n are assumed i.i.d. with $\pi_m = P\{X_n = m\}$ for $0 \leq m < M$.

Then we know from Example 1.3 that $p(y, x|\theta)$ is an exponential distribution with parameter

$$\theta = (\pi_0, \mu_0, R_0, \cdots, \pi_{M-1}, \mu_{M-1}, R_{M-1})$$

and natural sufficient statistics given by equations (1.25), (1.26), and (1.27).

In order to derive the EM update, we only need to replace the sufficient statistics in the ML estimate by their expected values. So to compute the EM update of the parameter θ, we first must compute the conditional expectation of the sufficient statistics. We can do this for the statistic N_m as follows.

$$\begin{aligned}
\bar{N}_m &= \mathrm{E}\big[N_m | Y = y, \theta^{(k)}\big] \\
&= \mathrm{E}\left[\sum_{n=1}^{N} \delta(x_n - m) \,\middle|\, Y = y, \theta^{(k)}\right] \\
&= \sum_{n=1}^{N} \mathrm{E}\big[\delta(x_n - m) \,|\, Y = y, \theta^{(k)}\big] \\
&= \sum_{n=1}^{N} P\{X_n = m | Y = y, \theta^{(k)}\}
\end{aligned}$$

Using a similar approach for all three sufficient statistics yields E-step of

$$\begin{aligned}
\bar{N}_m &= \sum_{n=1}^{N} P\{X_n = m | Y = y, \theta^{(k)}\} \\
\bar{b}_m &= \sum_{n=1}^{N} y_n P\{X_n = m | Y = y, \theta^{(k)}\} \\
\bar{S}_m &= \sum_{n=1}^{N} y_n y_n^t P\{X_n = m | Y = y, \theta^{(k)}\} \ .
\end{aligned}$$

Then in order to calculate the M-step, we simply uses these expected statistics in place of the conventional statistics in the equations for the ML estimator.

$$\pi_m^{(k+1)} = \frac{\bar{N}_m}{N}$$

$$\begin{aligned} \mu_m^{(k+1)} &= \frac{\bar{b}_m}{\bar{N}_m} \\ R_m^{(k+1)} &= \frac{\bar{S}_m}{\bar{N}_m} - \frac{\bar{b}_m \bar{b}_m^t}{\bar{N}_m^2} \ . \end{aligned}$$

With each repetition of this process, we increase the likelihood of the observations. Assuming that the likelihood has a minimum[1] and that one is not trapped in a local minimum, then repeated application of the EM iterations will converge to the ML estimate of θ.

[1]For the case of a Gaussian mixture with no lower bound on σ_k^2, the likelihood is not bounded. However, in practice a local minimum of the likelihood usually provides a good estimate of the parameters.

Chapter 1 Problems

1. Use the Q function in Section 1.2 to calculate the solutions to the M-step shown in equations (1.10), (1.12), and (1.13).

2. Derive the expression for the multivariate Gaussian density of $p(y|\theta)$ given in Example 1.1 with the following form.

$$p(y|\theta) = \frac{1}{(2\pi)^{NL/2}}|R|^{-N/2} \exp\left\{\frac{-1}{2}\left(\text{tr}\left\{SR^{-1}\right\} - 2b^tR^{-1}\mu + N\mu^tR^{-1}\mu\right)\right\} .$$

3. Show that the result stated in Example 1.3 is correct.

4. Let X and Y be two random vectors of dimensions M and L respectively that are jointly distributed with a Gaussian mixture. More specifically, the column vector $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$ has mixture density given by

$$p(z) = \sum_{i=0}^{M-1} \pi_i f(z|\mu_i, B_i)$$

where $f(z|\mu, B)$ is general notation for an N dimensional multivariate Gaussian density with mean $\mu \in \Re^N$ and inverse covariance $B \in \Re^{N\times N}$ given by

$$f(z|\mu, B) = \frac{1}{(2\pi)^{N/2}}|B|^{1/2} \exp\left\{-\frac{1}{2}(z-\mu)^tB(z-\mu)\right\} .$$

a) Show that the distribution of X is a Gaussian mixture.

b) Show that the conditional distribution of X given Y is a Gaussian mixture with the form

$$p(x|y) = \sum_{i=0}^{M-1} \pi_i(y) f(z|\mu_i(y), B_i) ,$$

and find expressions for its parameters $\mu_i(y)$, $\pi_i(y)$, and B_i.

c) Use the result of b) above to find an expression for the function $g(y)$ so that

$$\text{E}[X|Y] = g(Y) .$$

d) Is $g(Y)$ a linear function of Y? Why or why not?

5. Let X_n be N i.i.d. random variables with $P\{X_n = i\} = \pi_i$ for $i = 0, \cdots, M-1$. Also, assume that Y_n are conditionally independent given X_n with $p(y_n|x_n) \sim N(\mu_{x_n}, \gamma_{x_n})$. Use the results of this chapter to derive an EM algorithm for estimating the parameters $\{\pi_i, \mu_i, \gamma_i\}_{i=0}^{M-1}$.

6. Let X_n, W_n, and Y_n each be i.i.d. discrete time random processes with

$$Y_n = X_n + W_n$$

where $X_n \sim N(\mu, R)$ and $N \sim N(0, I)$.

a) Show that given N realizations, $X_1, X_2, \cdots, X_N$, the ML estimates of R and μ are

$$\mu = \frac{1}{N}\sum_{n=1}^{N} X_n$$

and

$$R = \frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)(X_n - \mu)^t \ .$$

b) Derive the EM algorithm for computing the ML estimates of R and μ from $\{Y_n\}_{n=1}^{N}$.

7. Let X_n be a sequence of i.i.d. random variables for $n = 1, \cdots, N$, and let Y_n be a sequence of random variables that are conditionally independent and Poisson with mean μ_i for $i = 0, \cdots, M-1$ given the corresponding values of X_n. More specifically, the distributions of X and Y are given by

$$\begin{aligned} P\{X_n = i\} &= \pi_i \\ p(y_n|x_n) &= \frac{1}{\mu_{x_n}} e^{-y_n/\mu_{x_n}} u(y_n) \end{aligned}$$

where $i \in \{0, \cdots, M-1\}$ and $u(\cdot)$ is the unit step function. Furthermore, let $\theta = [\pi_0, \mu_0, \cdots, \pi_{M-1}, \mu_{M-1}]$ be the parameter vector of the distribution.

a) Derive a closed form expression for the ML estimate of θ given both X and Y (i.e. the complete data).

b) Find the natural sufficient statistics for the exponential distribution of (X, Y).

c) Use the results of this chapter to derive the EM algorithm for computing the ML estimate of θ from Y.

Bibliography

[1] L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statistics*, 41(1):164–171, 1970.

[2] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statistics*, 37:1554–1563, 1966.

[3] E. Redner and H. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2), April 1984.

[4] S. Silvey. *Statistical Inference*. Chapman and Hall, London, 1975.

[5] C. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103, 1983.