

Estimation Maximization Algorithm with Applications

Benjamin Draves

December 12, 2017

Abstract

In several statistical models, missing or latent variables play a vital role in the structure of the stochastic system in question. Traditional estimation techniques implicitly assume that data generated from the model in question can be collected directly. In the presence of latent variables, however, this is usually not the case. In this scenario, classical methods fail to provide meaningful parameter estimates as the full joint distributed cannot be estimated directly. Bayesian analysis attempts to remedy this issue by placing a prior distribution over these latent variables but as the variables are hidden, there is no way to assess these prior assumptions. The *Estimation Maximization (EM) Algorithm* serves as a solution to this problem. By iteratively estimating the values of the latent variables then updating the model parameter estimate, the EM algorithm provides MLE estimates for statistical models where latent variables are critical to the global behavior of the model. This report provides a thorough introduction the EM estimation technique as we investigate the theory supporting the algorithm as well as provide several applications of the method to classical statistical problems and models.

1 Introduction

In several statistical models and applications, latent variables play a critical role in the global behavior of a stochastic system. In mixed effect modeling, statisticians address subject level effects by introducing random effect variables into the mean function of interest. While we never directly collect data on these entities, by explicitly building structure into our estimation procedure, we expand our frame of inference to allow for subject level variation. Oftentimes, practitioners either assume the latent variables follow some distribution or place a prior distribution over these variables and follow a Bayesian approach to inference. This allows for simple derivations that lead to inference of the fixed effects in the model. However, when practitioners are unaware of a correct distributional assumption for these latent variables, this question becomes much more difficult to answer. In these cases, statisticians look to leverage known correlation structure between observable variables in the model and the latent variable in question. In doing so, they utilize information from both the *incomplete data* sample as well as this correlation structure to estimate parameters that characterize the entire statistical model.

This line of logic is formalized in the Estimation Maximization (EM) algorithm. First introduced in 1977, the EM algorithm has been used in a wide variety of problems. In essence, the algorithm connects a large, usually very complex likelihood optimization problem with several smaller likelihood estimation problems that converge to the same optimal likelihood estimate. While the name implies this method is mostly algorithmic, the underlying theory, which we review here, is quite rich and gives rise to several desirable properties of the method.

This report is organized follow. In Section 2 we introduce the method and prove a result that implies convergence of the algorithm. In Section 3 we draw parallels between this method and standard Maximum Likelihood methods as well as a Bayesian estimation technique, Variational Bayes. In Section 4 we apply EM to mixture modeling and the Hidden Markov Model and in Section 5 we close with remarks on the method.

2 Development and Derivations

2.1 Introduction of the Algorithm

In several stochastic systems, statisticians are tasked with the *latent variable problem*. In this problem, practitioners attempt to model random variables that are not explicitly observable. To overcome this problem, inferential statements are made about the latent variables by examining random variables and the data they generated that are linked to the variables of interest. This indirect line of inference is utilized in several statistical models such as the Hidden Markov Model, mixture modeling, and mixed effect modeling in which all methods rely on the structure connecting the latent variable space with the observable data. This connection is the core foundation of the EM algorithm which we now develop.

Formally, the EM algorithm creates a sequence of estimates $\{\theta^{(r)}\}_{r=0}^{\infty}$ that converge to the MLE estimate, θ_{MLE} . For complete data situations (i.e. no latent variables or missing data), the algorithm is entirely derivative due to the arsenal of statistical techniques that already exist for maximum likelihood estimation. In the incomplete data situations (i.e. presence of latent variables or missing data), however, these techniques breakdown. One simple solution to this problem is to disregard the missing data or structure provided by the latent variables and find the MLE on the incomplete dataset. This naive approach however is not desirable as it disregards the true structure of the underlying problem. In this case, EM algorithm proves very useful as it connects the desired incomplete data likelihood with the complete data likelihood in which we are quite comfortable handling. Specifically, for the incomplete data $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ and the missing data $\mathbf{X} = (X_1, X_2, \dots, X_m)$, then by the law of total probability we have the relationship

$$g(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{x}|\theta) dx$$

where $g(\mathbf{y}|\theta)$ is the incomplete data density and $f(\mathbf{x}, \mathbf{y}|\theta)$ is the complete data density both parametrized by θ . From here we can define the respective likelihoods as $L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta)$ and $L(\theta|\mathbf{y}, \mathbf{x}) = f(\mathbf{x}, \mathbf{y}|\theta)$.

The power of the EM algorithm is that it allows us to maximize the incomplete data likelihood $L(\theta|\mathbf{y})$ through our knowledge of $L(\theta|\mathbf{x}, \mathbf{y})$. First, let $k(x)$ be some distribution over the missing data \mathbf{x} . Then by Jensen's Inequality we see that

$$\begin{aligned} \log L(\theta|\mathbf{y}) &= \log \int L(\theta|\mathbf{x}, \mathbf{y}) dx \\ &= \log \int k(x) \frac{L(\theta|\mathbf{x}, \mathbf{y})}{k(x)} dx \\ &\geq \int k(x) \log \frac{L(\theta|\mathbf{x}, \mathbf{y})}{k(x)} dx \\ &= \int k(x) \log L(\theta|\mathbf{x}, \mathbf{y}) dx - \int k(x) \log k(x) dx \\ &= \mathbb{E}_{k(\mathbf{x})}[\log L(\theta|\mathbf{x}, \mathbf{y})] - \mathbb{E}_{k(\mathbf{x})}[\log k(\mathbf{x})] \end{aligned}$$

From this we see that we can bound the incomplete data likelihood by the complete data likelihood minus some term¹ to account for the variability in our estimates of \mathbf{X} . But recall that Jensen's Inequality gives equality when the argument is constant with respect to $\mathbb{E}_{k(\mathbf{x})}(\cdot)$. Therefore we see that we can make this inequality an equality for $k(\mathbf{x}) \propto L(\theta|\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y}|\theta)$. We now construct a distribution for the missing data \mathbf{x} that has this property.

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

¹Technically, this term is the *entropy* of the missing data distribution $k(\mathbf{x})$

which is the conditional distribution of the missing data on the observed data. By construction the quantity $\frac{L(\theta|\mathbf{x},\mathbf{y})}{h(\mathbf{x}|\mathbf{y},\theta)} = g(\mathbf{y}|\mathbf{x},\theta)$ is constant with respect to \mathbf{x} . Therefore, we see that Jensen's Inequality gives us equality. Using this notation we have that

$$\log L(\theta|\mathbf{y}) = \mathbb{E} \left[\log L(\theta|\mathbf{y}, \mathbf{X}) | \mathbf{y}, \tilde{\theta} \right] - \mathbb{E} \left[\log h(\mathbf{X}|\theta, \mathbf{y}) | \mathbf{y}, \tilde{\theta} \right] \quad (1)$$

We can interpret this representation of the likelihood as a decomposition. That is the likelihood of the incomplete data can be written as the complete data likelihood with some reduction due to the stochasticity of the hidden variables in \mathbf{X} . Therefore, we see that maximum likelihood estimation is equivalent to maximizing this difference in expectation. As we will see in the next section, it is actually sufficient to maximize $\mathbb{E} \left[\log L(\theta|\mathbf{y}, \mathbf{X}) | \mathbf{y}, \tilde{\theta} \right]$ and disregard the second term in equation (1) to ensure that $L(\theta^{(r+1)}|\mathbf{y}) \geq L(\theta^{(r)}|\mathbf{y})$. With this piece of information, we can write the algorithm as follows.

1. Initialize $\theta^{(0)}$ to some randomly selected value.
2. Until convergence
 - (a) (**E Step**) Estimate $Q(\theta|\theta^{(r)}) := \mathbb{E} \left[\log L(\theta|\mathbf{y}, \mathbf{X}) | \mathbf{y}, \theta^{(r)} \right]$
 - (b) (**M Step**) Set $\theta^{(r+1)} = \arg \max_{\theta \in \Theta} Q(\theta|\theta^{(r)})$

Having defined the machinery driving the EM algorithm, we now turn to the monotonicity of the likelihood function evaluated at $\theta^{(r)}$ which implies convergence of our $\{\theta^{(r)}\}_{r=0}^{\infty}$.

2.2 Monotonicity of the EM Estimates

When analyzing the EM algorithm a natural first question is what is the behavior of the likelihood when evaluated at the constructed sequence $\{\theta^{(r)}\}_{r=0}^{\infty}$? As we will see, in the limit we have $\lim_{r \rightarrow \infty} \theta^{(r)} = \hat{\theta}_{MLE}$ but understanding the behavior of $L(\theta^{(r)}|\mathbf{y})$ will also prove useful. Our next result shows exactly how this sequence behaves which helps imply the convergence of the EM algorithm.

Lemma 1. *For densities $f(\cdot)$ and $g(\cdot)$ such that $f(x) > 0$ and $g(x) > 0$, we have*

$$\int g(x) \log f(x) dx \leq \int g(x) \log g(x) dx$$

Proof. Recall by Jensen's inequality, that for a concave function $h(\cdot)$ we have

$$\mathbb{E}[h(x)] \leq h[\mathbb{E}(x)]$$

Now, since \log is a concave function, we have

$$\begin{aligned} \int \log \left(\frac{f(x)}{g(x)} \right) g(x) dx &= \mathbb{E}_{g(X)} \left[\log \frac{f(x)}{g(x)} \right] \\ &\leq \log \left[\mathbb{E} \left(\frac{f(x)}{g(x)} \right) \right] \\ &= \log \left(\int \frac{f(x)}{g(x)} g(x) dx \right) \\ &= \log \left(\int f(x) dx \right) \\ &= 0 \end{aligned}$$

Using this with the property of logs, we arrive at our desired result.

$$\begin{aligned}\int \log \left(\frac{f(x)}{g(x)} \right) g(x) dx &\leq 0 \\ \int g(x) \log f(x) dx - \int g(x) \log g(x) dx &\leq 0 \\ \int g(x) \log f(x) dx &\leq \int g(x) \log g(x) dx\end{aligned}$$

□

Having this lemma, we are ready to give the monotonicity result.

Theorem 2. *The sequence $\{\theta^{(r)}\}_{r=0}^{\infty}$ given by the EM algorithm satisfies*

$$L(\theta^{(r+1)}|\mathbf{y}) \geq L(\theta^{(r)}|\mathbf{y})$$

Proof. It suffices to show $\log L(\theta^{(r+1)}|\mathbf{y}) \geq \log L(\theta^{(r)}|\mathbf{y})$. Recall by (1), with $\tilde{\theta} = \theta^{(r)}$ we can represent the log-likelihood as

$$\log L(\theta^{(r)}|\mathbf{y}) = \mathbb{E} \left[\log L(\theta^{(r)}|\mathbf{y}, \mathbf{X})|\mathbf{y}, \theta^{(r)} \right] - \mathbb{E} \left[\log h(\mathbf{X}|\theta^{(r)}, \mathbf{y})|\mathbf{y}, \theta^{(r)} \right]$$

Recall that we defined $\theta^{(r+1)} := \arg \max_{\theta} \mathbb{E} [\log L(\theta|\mathbf{y}, \mathbf{X})|\mathbf{y}, \theta^{(r)}]$. Hence we immediately see that for any $\theta \in \Theta$

$$\mathbb{E} \left[\log L(\theta^{(r+1)}|\mathbf{y}, \mathbf{X})|\mathbf{y}, \tilde{\theta} \right] \geq \mathbb{E} \left[\log L(\theta|\mathbf{y}, \mathbf{X})|\mathbf{y}, \tilde{\theta} \right]$$

Now considering right term in (1), we note by the Lemma

$$\begin{aligned}\mathbb{E} \left[\log h(\mathbf{X}|\mathbf{y}, \theta)|\tilde{\theta}, \mathbf{y} \right] &= \int \log h(\mathbf{X}|\mathbf{y}, \theta) h(\mathbf{X}|\tilde{\theta}, \mathbf{y}) dx \\ &\leq \int \log h(\mathbf{X}|\mathbf{y}, \tilde{\theta}) h(\mathbf{X}|\tilde{\theta}, \mathbf{y}) dx \\ &= \mathbb{E} \left[\log h(\mathbf{X}|\mathbf{y}, \tilde{\theta})|\tilde{\theta}, \mathbf{y} \right]\end{aligned}$$

Let $\tilde{\theta} = \theta^{(r)}$. Then, letting $\theta^{(r)} = \theta$ for the first inequality and letting $\theta = \theta^{(r+1)}$ in the second inequality, we arrive at our solution.

$$\begin{aligned}L(\theta^{(r)}|\mathbf{y}) &= \mathbb{E} \left[\log L(\theta^{(r)}|\mathbf{y}, \mathbf{X})|\mathbf{y}, \theta^{(r)} \right] - \mathbb{E} \left[\log h(\mathbf{X}|\theta^{(r)}, \mathbf{y})|\mathbf{y}, \theta^{(r)} \right] \\ &\leq \mathbb{E} \left[\log L(\theta^{(r+1)}|\mathbf{y}, \mathbf{X})|\mathbf{y}, \theta^{(r)} \right] - \mathbb{E} \left[\log h(\mathbf{X}|\theta^{(r)}, \mathbf{y})|\mathbf{y}, \theta^{(r)} \right] \\ &\leq \mathbb{E} \left[\log L(\theta^{(r+1)}|\mathbf{y}, \mathbf{X})|\mathbf{y}, \theta^{(r)} \right] - \mathbb{E} \left[\log h(\mathbf{X}|\theta^{(r+1)}, \mathbf{y})|\mathbf{y}, \theta^{(r)} \right] \\ &= L(\theta^{(r+1)}|\mathbf{y})\end{aligned}$$

□

Therefore, by this theorem, we see that as $\theta^{(r)}$ increases, so does $L(\theta^{(r)}|\mathbf{y})$. In standard cases², when we let $r \rightarrow \infty$, $L(\theta^{(r)}|\mathbf{y})$ attains its maximum which implies $\theta^{(r)} \rightarrow \theta_{MLE}$.

²There are some cases where the Likelihood function exhibits a nonstandard geometry (e.g. saddle points) for which the EM algorithm instead convergence on a local minimum. For simplicity, we disregard these issues here.

3 Comparison to Other Methods

3.1 Connection to Kullback–Leibler Divergence

The EM algorithm iteratively uses the connection between the complete data likelihood and the incomplete data likelihood to find optimal estimates of θ . Upon investigation of the way in which these likelihoods are related, we see there is a rich, information theoretic interpretation of the EM algorithm. In the derivation of (1) we used properties of Jensen’s inequality to arrive at a sharp bound on the incomplete data likelihood. However, if we consider the difference of the quantities given by Jensen’s, we recognize a familiar form. Consider the following difference.

$$\begin{aligned} \log L(\theta|\mathbf{y}) - \int k(x) \log \frac{L(\theta|\mathbf{y}, \mathbf{x})}{k(x)} dx &= \int k(x) \log L(\theta|\mathbf{y}) dx - \int k(x) \log \frac{L(\theta|\mathbf{y}, \mathbf{x})}{k(x)} dx \\ &= \int k(x) \log L(\theta|\mathbf{y}) dx - \int k(x) \log \frac{L(\theta|\mathbf{y}) h(\mathbf{x}|\mathbf{y}, \theta)}{k(x)} dx \\ &= \int k(x) \log L(\theta|\mathbf{y}) dx - \int k(x) \log L(\theta|\mathbf{y}) dx - \int k(x) \log \frac{h(\mathbf{x})}{k(x)} dx \\ &= KL(k||h) \end{aligned}$$

From this derivation we see that from Jensen’s inequality, the information we lose is exactly the KL divergence of k from h , the conditional distribution of the missing data. This motivates the decomposition

$$L(\theta|\mathbf{y}) = \int k(x) \log \frac{L(\theta|\mathbf{x}, \mathbf{y})}{k(x)} dx + KL(k||h) \quad (2)$$

Now notice that $KL(k||h) = 0$ if and only if $k \stackrel{a.s.}{=} h$. Therefore, when we set $k(x) = h(x|\theta, \mathbf{y})$ we are actually *minimizing the KL divergence*. Recall that Maximum Likelihood estimation can also be formulated as minimizing the KL divergence. Hence we can view the EM algorithm as a method that replaces a single, difficult, maximum likelihood problem with a sequence of easier to maximum likelihood problems. Moreover, we see that each $\theta^{(r)}$ is a maximum likelihood estimate. Therefore, each estimate in the sequence inherits the properties of maximum likelihood estimates such as invariance. While EM was specifically constructed to solve a single maximum likelihood problem, the sequence of estimates it constructs even has desirable properties.

3.2 Variational Bayes

In several ways, Variational Bayes is the Bayesian equivalent to the Estimation Maximization algorithm. In a Bayesian setting, we do not fix the unknown parameters θ . Instead, the parameters to the model are random variables and the data are fixed constants that inform our inference about the hidden variables. Therefore the Bayesian approach to the hidden variable problem includes both the parameters θ and the hidden variables \mathbf{X} as parameters to the model. In this framework we look to maximize the *marginal likelihood* of the data given by $\log f(\mathbf{y}) = \log \int f(\mathbf{y}, \mathbf{z}) d\mathbf{z}$ where $\mathbf{Z} = (\theta, \mathbf{X})$. As in the frequentist setting, this problem can be quite difficult. Instead Variational Bayes turns to maximizing simpler functions that serve as a lower bound on the marginal likelihood function. One can show using Jensen’s inequality

$$\log f(\mathbf{y}) \geq \int k(\mathbf{z}) \log \frac{f(\mathbf{z}, \mathbf{y})}{k(\mathbf{z})} d\mathbf{z}$$

Moreover, again we see that considering the difference between these two quantities gives rise the decomposition

$$\log f(\mathbf{y}) = \int k(\mathbf{z}) \log \frac{f(\mathbf{z}, \mathbf{y})}{k(\mathbf{z})} d\mathbf{z} + KL(k||h) \quad (3)$$

where $h(\cdot)$ is the true posterior distribution of \mathbf{Z} . Hence this maximization procedure is equivalent to minimizing the KL distance of the approximating distribution k and the true posterior distribution h . Recall that in the EM algorithm, the maximization step gave a single point estimate, $\theta^{(r)}$, which was equal to the missing data log likelihood at that point. The only difference in Variational Bayes is that we seek a *distribution*, k , that maximizes the lower bound or equivalently minimizes $KL(k||h)$. Therefore, if we let, $\tilde{k}(\mathbf{Z}) = \delta(\mathbf{Z} - \mathbf{Z}^*)$ be the point mass over the point \mathbf{Z}^* where the lower bound is equal to the marginal distribution, $\log f(\mathbf{y})$, our estimation procedure defines a sequence of estimates that are monotonically increasing to the mean of the true posterior distribution. But recall that this was exactly our approach in the EM algorithm. Therefore, we see that Variational Bayes reduces to the EM algorithm when we restrict the class of prior distributions of the latent variables to Dirac Delta distributions over the points that maximize the marginal likelihood.

4 Applications

4.1 Mixture Modeling

In several statistical applications, practitioners encounter data that is aggregated over multiple populations. In classification problems, statisticians attempt to partition their data into clusters that represent the latent population structure of the data. In this setting we consider the labels for each data point as *missing data* and use the EM algorithm to recover the latent subpopulation structure. Here we will consider the case where there are two subpopulations.

In order to formalize this problem, let $f(x)$ and $g(x)$ be densities and let p be an unknown probability. Then given a sample $\mathbf{X} = (X_1, X_2, \dots, X_n)$ from the mixture

$$\mathbf{X} \sim pf(x) + (1 - p)g(x)$$

an inferential task of interest is recovering the value of p from the sample \mathbf{X} . We anticipate that each X_i contains information about either $f(x)$ or $g(x)$ and p . But if we fail to attribute each X_i to its corresponding density, we will be inferring properties of p with an erroneous assumption from our sample. If we somehow knew the corresponding density for each X_i our estimation procedure would be straight forward. We will use EM to address this missing data problem.

Let $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)$ be the associated latent variable which determines which distribution X_i follows. Formally, our model can be written as

$$\begin{aligned} Z_i &\stackrel{iid}{\sim} \text{Bern}(p) \\ X_i | Z_i = 1 &\sim f(x) \\ X_i | Z_i = 0 &\sim g(x) \end{aligned}$$

Before we address the implementation of the EM algorithm, we will derive some distributions which will be helpful later on. Namely, we derive the full data distribution and the corresponding latent data distribution.

$$\begin{aligned} h(\mathbf{X}, \mathbf{Z} | p) &= h(\mathbf{X} | \mathbf{Z}, p) h(\mathbf{Z} | p) = \prod_{i=1}^n f(x_i)^{z_i} g(x_i)^{1-z_i} \prod_{i=1}^n p^{z_i} (1-p)^{1-z_i} = \prod_{i=1}^n [pf(x_i)]^{z_i} [(1-p)g(x_i)]^{1-z_i} \\ h(\mathbf{Z} | \mathbf{X}, p) &= \frac{h(\mathbf{X}, \mathbf{Z} | p)}{h(\mathbf{X} | p)} = \frac{\prod_{i=1}^n [pf(x_i)]^{z_i} [(1-p)g(x_i)]^{1-z_i}}{\prod_{i=1}^n h(X_i | p)} = \prod_{i=1}^n \frac{[pf(x_i)]^{z_i} [(1-p)g(x_i)]^{1-z_i}}{pf(x_i) + (1-p)g(x_i)} \end{aligned}$$

Notice from this derivation we also see that $Z_i | \mathbf{x}, p$ is a binomial random variable with parameter $pf(x)/[pf(x) + (1-p)g(x)]$. Now we calculate the expected complete-data log likelihood (the **E** step) which will define the

sequence of estimates of p given by the EM algorithm.

$$\begin{aligned}
\mathbb{E}[\log L(p|\mathbf{Z}, \mathbf{X})|\mathbf{X}, \tilde{p}] &= \mathbb{E} \left(\log \prod_{i=1}^n [pf(x_i)]^{z_i} [(1-p)g(x_i)]^{1-z_i} \middle| \tilde{p}, \mathbf{X} \right) \\
&= \mathbb{E} \left[\sum_{i=1}^n \log ([pf(x_i)]^{z_i} [(1-p)g(x_i)]^{1-z_i}) \middle| \tilde{p}, \mathbf{X} \right] \\
&= \sum_{i=1}^n \mathbb{E} \left[\log ([pf(x_i)]^{z_i} [(1-p)g(x_i)]^{1-z_i}) \middle| \tilde{p}, \mathbf{X} \right] \\
&= \sum_{i=1}^n \sum_{z_i=0}^1 P(z_i|\tilde{p}, \mathbf{X}) \log [pf(x_i)]^{z_i} [(1-p)g(x_i)]^{1-z_i} \\
&= \sum_{i=1}^n \sum_{z_i=0}^1 P(z_i|\tilde{p}, \mathbf{X}) [z_i \log(p) + z_i \log(f(x_i)) + (1-z_i) \log(1-p) + (1-z_i) \log(g(x_i))]
\end{aligned}$$

Now, recall we look to maximize this quantity with respect to the parameter p , so any term that does not contain p we can remove from consideration. Therefore maximizing the above is equivalent to maximizing the following

$$\mathbb{E}[\log L(p|\mathbf{X}, \mathbf{Z})|\tilde{p}, \mathbf{X}] \propto \log(1-p) \sum_{i=1}^n P(z_i = 0|\mathbf{X}, \tilde{p}) + \log(p) \sum_{i=1}^n P(z_i = 1|\mathbf{X}, \tilde{p}) \quad (4)$$

Now maximizing (4) (the **M** step) with respect to p , we see

$$\begin{aligned}
\frac{\partial}{\partial p} \log(1-p) \sum_{i=1}^n P(z_i = 0|\mathbf{X}, \tilde{p}) + \log(p) \sum_{i=1}^n P(z_i = 1|\mathbf{X}, \tilde{p}) &= 0 \\
-\frac{1}{1-p} \sum_{i=1}^n P(z_i = 0|\mathbf{X}, \tilde{p}) + \frac{1}{p} \sum_{i=1}^n P(z_i = 1|\mathbf{X}, \tilde{p}) &= 0 \\
(1-p) \sum_{i=1}^n P(z_i = 1|\mathbf{X}, \tilde{p}) &= p \sum_{i=1}^n P(z_i = 0|\mathbf{X}, \tilde{p}) \\
\sum_{i=1}^n P(z_i = 1|\mathbf{X}, \tilde{p}) &= p \left(\sum_{i=1}^n P(z_i = 0|\mathbf{X}, \tilde{p}) + \sum_{i=1}^n P(z_i = 1|\mathbf{X}, \tilde{p}) \right) \\
\sum_{i=1}^n P(z_i = 1|\mathbf{X}, \tilde{p}) &= p \left(\sum_{i=1}^n P(z_i = 0|\mathbf{X}, \tilde{p}) + \sum_{i=1}^n P(z_i = 1|\mathbf{X}, \tilde{p}) \right) \\
\hat{p} &= \frac{1}{n} \sum_{i=1}^n P(z_i = 1|\mathbf{X}, \tilde{p}) \\
\hat{p} &= \frac{1}{n} \sum_{i=1}^n \frac{\tilde{p}f(x_i)}{\tilde{p}f(x_i) + (1-\tilde{p})g(x_i)}
\end{aligned}$$

Now, letting $\tilde{p} = p^{(r)}$, we see that the sequence of estimates given by the EM algorithm can be written as

$$p^{(r+1)} = \frac{1}{n} \sum_{i=1}^n \frac{p^{(r)}f(x_i)}{p^{(r)}f(x_i) + (1-p^{(r)})g(x_i)}$$

Therefore, we see that current EM estimate $p^{(r+1)}$ of the parameter p is a weighted average of $p^{(r)}$ based on the density evaluated at the sample data. This shows that EM simultaneously relies on previous estimates

as well as the data to provide new estimates of p . Moreover, we note that by specifying distributional forms for f and g (e.g. Gaussian), this process also allows for estimation of distributional parameters. We would simply differentiate (4) with respect to the parameter in question (e.g. μ, σ^2) and find the r th EM estimate for these values. In this way, the EM technique can construct multiple sequences of parameter estimates simultaneously.

4.2 Hidden Markov Model

As we saw in the previous section, the EM algorithm is capable of estimating parameters from quite complex models. Another model for which the EM algorithm proves effective is the Hidden Markov Model (HMM). Suppose we have a sequence of random variables $\{X_t\}_{t=0}^N$ on D states that corresponds to a latent discrete time Markov Chain $\{Z_t\}_{t=0}^N$ over K states. Provided we have the following two properties

1. $P(Z_i|Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_N, X_1, \dots, X_N) = P(Z_i|Z_{i-1})$
2. $P(X_i|Z_1, \dots, Z_N, X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_N) = P(X_i|Z_i)$

then this model specifies a *Hidden Markov Model*. With several applications in bioinformatics and reinforcement learning, the HMM allows for practitioners to model simple stochastic processes that are hidden deep within a complex system. Using these simple assumptions we can find the joint distribution as follows,

$$f(x_1, \dots, x_N, z_1, \dots, z_N) = \underbrace{\left[p(z_1) \prod_{n=2}^N p(z_n|z_{n-1}) \right]}_{\text{Transition Probabilities}} \underbrace{\prod_{n=1}^N p(x_n|z_n)}_{\text{Emission Probabilities}}$$

We note by conditioning on the latent states z_n , we can factor the likelihood into two factors. The first factor represents the transition probabilities in the hidden Markov Chain while the second term we define as the *emission probabilities*. These conditional values represent the probability of observing a value x_i given the hidden state z_i . In a sense, the z_i emit the x_i values that represent the structure of the underlying Markov Chain. Suppose we have an initial probability vector π and a fixed transition matrix \mathbf{A} . Then we can write the transition probabilities as follows

$$p(z_n|z_{n-1}, \mathbf{A}) = \prod_{k=1}^K \prod_{j=1}^K A_{jk}^{z_{n-1}, k, z_n j} \quad p(z_1|\pi) = \prod_{k=1}^K \pi^{z_1 k}$$

where $z_{rs} = 1$ if and only of the n th variable is at state k . Moreover, we can write the emission probabilities as

$$p(x_n|z_n, \phi) = \prod_{k=1}^K p(x_n|\phi_k)^{z_{nk}}$$

where ϕ_k is the emission probabilities based on current latent state z_k . In general, the $K \times D$ matrix ϕ represents the conditional probabilities of $x_t|z_t = k$. With this in mind, letting $\mathbf{z} = (z_1, z_2, \dots, z_N)$, $\mathbf{x} = (x_1, x_2, \dots, x_N)$, and $\theta = (\pi, \mathbf{A}, \phi)$ we can write our conditional joint density as

$$p(\mathbf{x}, \mathbf{z}|\theta) = p(z_1|\pi) \prod_{n=2}^N p(z_n|z_{n-1}, \mathbf{A}) \prod_{n=1}^N p(x_n|z_n, \phi)$$

Having parameterized this problem through θ , a first inferential task would be to estimate θ . Seeing we only observe the emission variables X_t , we attempt to maximize the likelihood $L(\theta|\mathbf{x})$ but due to the dependence structure between \mathbf{x} and the unobservable \mathbf{z} , this problem is quiet difficult. Instead we turn to the EM algorithm and instead focus on maximizing the complete data likelihood $L(\theta|\mathbf{Z}, \mathbf{x})$ where \mathbf{Z} is estimated

iteratively. By doing so, we can attain estimates of the transition probabilities in the latent Markov Chain, the initial distribution, as well as the probabilistic structure connecting X_t and Z_t - quantities that characterize the Markov Chain, Z_t .

This example highlights the true power of the EM algorithm. While we never observe the latent stochastic process, this estimation technique allows us to attain maximum likelihood estimates of the transition matrix \mathbf{A} , the initial distribution π , and the emission probabilities ϕ . That is, we can attain estimates about a very complex probabilistic object through observing a sequence of independent random variables X_t that are only conditionally dependent on the Z_t .

5 Conclusion

The Estimation Maximization Algorithm provides a framework to find maximum likelihood estimates in statistical models that contain latent variables or missing data. In this framework classical estimation techniques fail to provide meaningful estimates as the full joint distribution cannot be estimated directly. As we showed here, the EM algorithm iteratively constructs lower bounds on the incomplete data likelihood and maximizes this lower bound upon each iteration. We proved that this procedure corresponds to maximization of the likelihood function in standard cases. We showed how this maximization method is connected to the minimization of the KL Divergence as well as its connection to the Variational Bayes estimation technique. Lastly, we showed the power of the EM algorithm in modeling subpopulation structure in mixture distributions and its ability to estimate characteristics of a HMM through emission variables.

6 References