Estimation Maximization Algorithm with Applications

Benjamin Draves

December 5, 2017

Abstract

1 Introduction

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 a 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 in seen 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 Estimation Maximization (EM) algorithm which we will now develop.

Formally, the EM algorithm creates a sequence of estimates $\{\theta^{(r)}\}_{r=0}^{\infty}$ that are guaranteed to 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. latent variables or missing), however, these techniques breakdown. One simple solution to this problem is to disregard the missing data 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 $h(\mathbf{x}|\theta, \mathbf{y})$ be the conditional density of the missing data $\mathbf{x}$

given the observed data $\mathbf{y}$. Then we see that

$$\begin{aligned}
h(\mathbf{x}|\theta, \mathbf{y}) &= \frac{f(\mathbf{x}, \mathbf{y}|\theta)}{g(\mathbf{y}|\theta)} \\
h(\mathbf{x}|\theta, \mathbf{y}) &= \frac{L(\theta|\mathbf{x}, \mathbf{y})}{L(\theta|\mathbf{y})} \\
\log\left[h(\mathbf{x}|\theta, \mathbf{y})\right] &= \log\left[L(\theta|\mathbf{x}, \mathbf{y})\right] - \log\left[L(\theta|\mathbf{y})\right] \\
\log\left[L(\theta|\mathbf{y})\right] &= \log\left[L(\theta|\mathbf{x}, \mathbf{y})\right] - \log\left[h(\mathbf{x}|\theta, \mathbf{y})\right]
\end{aligned}$$

From this, we see that maximization of $L(\theta|\mathbf{y})$ is equivalent to maximization of $\log\left[L(\theta|\mathbf{x}, \mathbf{y})\right] - \log\left[h(\mathbf{x}|\theta, \mathbf{y})\right]$ *which relies on unobserved data* $\mathbf{x}$. While this quantity is unknown with respect to the the missing data $\mathbf{x}$, one reasonable next step is to replace this value with its *expectation* under the conditional distribution $h(\cdot|\tilde{\theta}, \mathbf{y})$ for some value of θ, $\tilde{\theta}$. With this goal in mind, we can write

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

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 $\mathbf{X}$. Using these forms we see that

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

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 $\mathbb{E}\left[\log L(\theta|\mathbf{y}, \mathbf{X}) \middle| \mathbf{y}, \tilde{\theta}\right]$ 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 $Q(\theta|\theta^{(r)}) := \mathbb{E}\left[\log L(\theta|\mathbf{y}, \mathbf{X}) \middle| \mathbf{y}, \theta^{(r)}\right]$
   (b) (**M** Step) Set $\theta^{(r+1)} = \underset{\theta \in \Theta}{\arg\max}\, Q(\theta|\theta^{(r)})$

Having defined the machinery driving the EM algorithm, we now turn to the monotonicity of the likelihood function evaluated at $\theta^{(r)}$ and the convergence guarantees that follow.

2.2 Monotonicity & Convergence

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 $\lim_{r\to\infty} \theta^{(r)} = \hat{\theta}_{MLE}$

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

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

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

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

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

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

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.

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

□

3 Comparison to Other Methods

3.1 Variational Bayes and Kullback-Leibler Divergence

3.2 Entropy Interpretation

4 Applications

4.1 Exponential Family

4.2 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 the probability be an unknown probability. Then given a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 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

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

Before we address the implementation of the EM algorithm, we will derive some distributions which will be helpful later on. First we consider the complete data likelihood.

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

Using this we can derive the latent data distribution as follows.

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

Now we calculate the expected complete-data log likelihood (the **E** step) which we will show the sequence of estimates of p given by the EM algorithm.

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

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

$$\mathbb{E}[\log L(p|\mathbf{X}, \mathbf{Z})|\tilde{p}, \mathbf{X}] \propto \log(1-p) \sum_{i=1}^{n} P(z_i = 0|\mathbf{X}, \tilde{p}) z_i + \log(p) \sum_{i=1}^{n} P(z_i = 1|\mathbf{X}, \tilde{p}) \tag{2}$$

5 Data Example

6 Conclusion