Abstract

We survey the EM algorithm and its Monte Carlo-based extensions.

1 The EM Algorithm

The EM algorithm is a method for analyzing incomplete data which was formalized by Dempster et al. (1977). See McLachlan and Krishnan (2008) for an excellent book-length overview of the EM algorithm. We begin by discussing a probabilistic framework within which the EM algorithm is often applied. We then present the EM algorithm in detail. Finally, we discuss some limitations of this method. Throughout, we illustrate our presentation with a toy problem based on linear regression with a single, unobserved, covariate.

The EM algorithm consists of iterating two steps. First is the expectation, or "E", step, in which an objective function is constructed from the complete data likelihood. Second is the maximization, or "M", step, in which the previously computed objective function is maximized. These two steps are then alternated until some convergence criterion is met. Whatever value of θ the algorithm converges to is used as our parameter estimate. We now go into more detail on each of the two steps.

The E-step of the EM algorithm is where we construct the objective function which will be used to update our parameter estimate. This objective function is the conditional expectation of the complete data likelihood, given the observed data. If our complete data can be partitioned into an observed component, Y, and a missing component, X, then our objective function at iteration $k+1$ is given by

$$Q(\theta|\theta_k) = \mathbb{E}_{\theta_k}[\ell_c(\theta; y, X)|Y = y] \tag{1}$$

Where ℓ_c is the log-likelihood of the complete data model. Note that the conditional expectation uses our parameter estimate from the previous iteration.

The M-step of the EM algorithm consists of maximizing the objective function constructed in the previous E-step. That is, we define $\theta_{k+1} = \underset{\theta}{\operatorname{argmax}} Q(\theta|\theta_k)$. Typically, this optimization must be performed numerically via, e.g., gradient ascent or Newton's method. See Nocedal and Wright (2006) for details and other optimization algorithms.

We can combine the E- and M-steps of the EM algorithm into a single "update function". We write $M(\theta_k) = \underset{\theta}{\operatorname{argmax}} Q(\theta|\theta_k)$. The EM algorithm can thus be viewed as the iterative application of this update function, M.

1.1 Properties

Section intro...

1.1.1 Ascent Property and Generalized EM

An important feature of the EM algorithm is its so-called "ascent property". This property says that an iteration of the EM algorithm (have I explicitly defined "EM iteration"? Do I need to?) never decreases the observed data likelihood. This is somewhat surprising, since EM updates are computed without ever evaluating the observed data likelihood.

Proposition 1.1 (Ascent Property of EM)**.** *Let $\theta \in \Theta$, and $\theta' = M(\theta)$ be the EM update from θ. Then $\ell(\theta') \geq \ell(\theta)$.*

Proof. We begin by noting that the following decomposition holds for any value of x:

$$\ell(\theta; y) = \ell_c(\theta; y, x) - \ell_m(\theta; y, x) \tag{2}$$

Subtracting the values of both sides at θ from their values at θ' and taking conditional expectations, we get

$$\ell(\theta'; y) - \ell(\theta; y) = Q(\theta'|\theta) - Q(\theta|\theta) + \mathbb{E}_\theta[\ell_m(\theta; y, x) - \ell_m(\theta'; y, x)] \tag{3}$$

$$= Q(\theta'|\theta) - Q(\theta|\theta) + \mathrm{KL}(\theta||\theta') \tag{4}$$

Where the last term in line 4 is the Kullback-Leibler (KL) divergence from the missing data distribution with $\theta = \theta$ to the same distribution with $\theta = \theta'$. Note that KL divergences are always non-negative, so we get

$$\ell(\theta'; y) - \ell(\theta; y) \geq Q(\theta'|\theta) - Q(\theta|\theta) \tag{5}$$

Finally, since θ' maximizes $Q(\cdot|\theta)$, we have $\ell(\theta'; y) - \ell(\theta; y) \geq 0$. □

In our proof of the ascent property, we only required that $Q(\theta'|\theta) \geq Q(\theta|\theta)$, not that θ' maximize $Q(\cdot|\theta)$. This observation leads to the definition of the "Generalized EM Algorithm", which replaces the M-step with setting θ_{k+1} to any point in Θ such that $Q(\theta_{k+1}|\theta_k) \geq Q(\theta_k|\theta_k)$.

1.1.2 Recovering Observed Data Likelihood Quantities

Under regularity conditions, it is possible to compute both the score vector and the observed information matrix of the observed data likelihood using complete data quantities. These regularity conditions consist of being able to interchange the order of differentiation and integration for various functions (awk? I wasn't feeling well when I wrote this.).

Proposition 1.2. *Let f and ℓ be the density and log-likelihood of the observed data distribution. Let f_c, ℓ_c and f_m, ℓ_m be the corresponding quantities for the complete data and missing data distributions respectively.*

We start with the observed data score vector. Let f and f_c be densities for the observed and complete data distributions respectively.

1.2 Example: Linear Regression with an Unobserved Covariate

Consider the scenario where a measured quantity is known to depend linearly on another unobserved, but nevertheless well understood, quantity. For example, something, something, census data. We first present a model for such a scenario, then show how to directly analyze the observed data. Throughout the rest of this document, we will return to this example to illustrate how to perform an analysis when increasing portions of the calculations cannot be performed analytically (awk).

Let $X \sim \mathrm{M}(\mu, \tau^2)$, where $\mu \in \mathbb{R}$ and $\tau > 0$. Let $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ for some $\sigma > 0$, and $Y = X\beta + \varepsilon$ where $\beta \in \mathbb{R}$. We observe an iid sample of Ys, but not their corresponding Xs. We do however, treat μ and τ as known. Our goal is to estimate β and σ from this incomplete data.

Appendix A Likelihood for Linear Regression with Unobserved Covariates

In this appendix, we present details for the analysis of our linear regression example with unobserved covariates. See Section 1.2 for formulation of the model and definition of notation.

A.1 Observed Data Likelihood, Score and Information

The complete data distribution for our model can be written as follows.

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mu\beta \\ \mu \end{pmatrix}, \begin{bmatrix} \sigma^2 + \tau^2\beta^2 & \tau^2\beta \\ \tau^2\beta & \tau^2 \end{bmatrix} \right) \tag{6}$$

Since our observed data, Y, is a marginal of the complete data, we can read off the distribution of Y from Expression (6). That is, $Y \sim \text{N}(\mu\beta, \sigma^2 + \tau^2\beta^2)$.

Based on a sample of observed data, $y_1, \ldots, y_n$, our log-likelihood is as follows. For conciseness, let $\theta = (\beta, \sigma)$ be the vector of unknown parameters, and $\eta^2 = \sigma^2 + \tau^2\beta^2$ be the marginal variance of Y. In previous work, I used η for $\mathbb{V}Y$. Make sure I adjust any discussion and **CODE(!!!)** accordingly.

$$\ell(\theta; y) = -\frac{n}{2}\log(2\pi) - n\log(\eta) - \sum_{i=1}^{n} \frac{(y_i - \mu\beta)^2}{2\eta^2} \tag{7}$$

$$\equiv -n\log(\eta) - \sum_{i=1}^{n} \frac{(y_i - \mu\beta)^2}{2\eta^2} \tag{8}$$

Where $\equiv$ denotes equality up to additive constants which do not depend on θ.

The score vector is given by

$$S(\theta; y) = \frac{1}{\eta^4} \begin{pmatrix} -n\beta^3\tau^4 - \beta^2\mu\tau^2 \sum y_i + \beta[\tau^2 \sum y_i^2 - n\sigma^2(\tau^2 + \sigma^2)] + \mu\sigma^2 \sum y_i \\ \sigma[n\beta^2(\mu^2 - \tau^2) - 2\beta\mu \sum y_i + \sum y_i^2 - n\sigma^2] \end{pmatrix} \tag{9}$$

The observed information matrix is given by

$$I(\theta; y) = \frac{1}{\eta^6} \begin{bmatrix} I^{(1,1)} & I^{(1,2)} \\ I^{(1,2)} & I^{(2,2)} \end{bmatrix} \tag{10}$$

where

$$
\begin{aligned}
I^{(1,1)} &= -n\beta^4\tau^6 - 2\beta^3\mu\tau^4 \sum y_i + 3\beta^2\tau^2(3\tau^2 \sum y_i^2 - n\sigma^2\mu^2) \qquad (11)\\
&\quad + 6\beta\sigma^2\tau^2\mu \sum y_i + \sigma^2[n\sigma^2(\tau^2+\mu^2) - \tau^2 \sum y_i]\\
I^{(1,2)} &= 2n\beta^3\sigma\tau^2(\mu^2-\tau^2) - 6\beta^2\sigma\mu\tau^2 \sum y_i + 2\beta\sigma[2\tau^2 \sum y_i^2 - n\sigma^2(\mu^2+\tau^2)] \qquad (12)\\
&\quad + 2\mu\sigma^3 \sum y_i\\
I^{(2,2)} &= n\beta^4\tau^2(\tau^2-\mu^2) + 2\beta^3\mu\tau^2 \sum y_i + \beta^2(3n\mu^2\sigma^2 - \tau^2 \sum y_i^2) \qquad (13)\\
&\quad - 6\beta\mu\tau^2 \sum y_i + \sigma^2(3 \sum y_i^2 - n\sigma^2)
\end{aligned}
$$

As an aside, I did explore the above model with multiple covariates. Unfortunately, marginalizing out X consists of replacing each observed covariate vector with its mean, μ. This results in linearly dependent observations, so the model is overparameterized. I could probably incorporate an intercept term without introducing the overparameterization problem, but I don't think it's worth the effort. I'm not going to be able to sell anyone on the applicability of my model, and adding a third parameter won't really increase the pedagogical value.

Check for "Citation Needed" before publishing.

References

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1977.

Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions.* Wiley, 2nd edition, 2008.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization.* Springer, 2nd edition, 2006.

Index