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

$$\begin{aligned} \ell(\theta'; y) - \ell(\theta; y) &= Q(\theta'|\theta) - Q(\theta|\theta) + \mathbb{E}_\theta[\ell_m(\theta; y, x) - \ell_m(\theta'; y, x)] && (3) \\ &= Q(\theta'|\theta) - Q(\theta|\theta) + \mathrm{KL}(\theta||\theta') && (4) \end{aligned}$$

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.). Define $\mathcal{I}_c$ and $\mathcal{I}_m$.

Proposition 1.2. *The following identities hold (regularity conditions!):*

(i) $S(\theta; y) = \mathbb{E}_\theta[S_c(\theta; y, X)|Y = y]$

(ii) $I(\theta) = \mathcal{I}_c(\theta) - \mathcal{I}_m(\theta)$

Proof. We start with expression (i). Let Ω be the complete data sample space. Let $\mathcal{Y}$ and $\mathcal{X}$ be the observed and missing data sample spaces respectively. For every $y \in \mathcal{Y}$, let $\mathcal{X}(y) = \{x \in \mathcal{X} : (y, x) \in \Omega\}$. Note that $f(y;\theta) = \int_{\mathcal{X}(y)} f_c(y, x; \theta)dx$.

$$\begin{aligned}
\mathbb{E}_\theta[S_c(\theta; y, X)|Y = y] &= \int_{\mathcal{X}(y)} \nabla \ell_c(\theta; y) f_m(y, x; \theta)dx \\
&= \int_{\mathcal{X}(y)} \frac{f_m(y, x; \theta)}{f_c(y, x; \theta)} \nabla f_c(\theta; y) \\
&= \int_{\mathcal{X}(y)} \frac{1}{f(y;\theta)} \nabla f_c(\theta; y) \\
&= \frac{1}{f(y;\theta)} \int_{\mathcal{X}(y)} \nabla f_c(\theta; y) \\
&= \frac{1}{f(y;\theta)} \nabla \int_{\mathcal{X}(y)} f_c(\theta; y) \\
&= \frac{1}{f(y;\theta)} \nabla f(y;\theta) \\
&= S(\theta; y)
\end{aligned}$$

Proceeding now to (ii), we decompose the observed data log-likelihood as

$$\ell(\theta; y) = \ell_c(\theta; y, x) - \ell_m(\theta; y, x)$$

Differentiating twice and taking conditional expectations of both sides yields the required result. □

An alternative to Proposition 1.2 part (ii) which involves only conditional expectations of complete data quantities is given in the following proposition.

Proposition 1.3. *Let $\hat{\theta}$ be a stationary point of the observed data log-likelihood. Assuming regularity conditions, we can write the observed information of the observed data distribution at $\hat{\theta}$ as*

$$I(\hat{\theta}) = \mathcal{I}_c(\hat{\theta}) - \mathbb{E}_{\hat{\theta}}[S_c(\hat{\theta}) S_c(\hat{\theta})|Y = y] \tag{6}$$

Proof. We follow the derivation of Louis (1982). For brevity, we write $f(\theta)$ and $f_c(\theta)$ for

$f(y;\theta)$ and $f(y,x;\theta)$ respectively. Consider the following two Hessians:

$$\nabla^2 \ell(\theta) = \nabla \left[\int_{\mathcal{X}(y)} \frac{\nabla f_c(\theta) dx}{f(\theta)} \right] \tag{7}$$

$$= \int_{\mathcal{X}(y)} \frac{\nabla^2 f_c(\theta)}{f(\theta)} dx - \frac{1}{f(\theta)^2} \left(\int_{\mathcal{X}(y)} \nabla f_c(\theta) dx \right) \left(\int_{\mathcal{X}(y)} \nabla f_c(\theta) dx \right)^T \tag{8}$$

$$= \mathbb{E}_\theta \left[\frac{\nabla^2 f_c(\theta)}{f_c(\theta)} \middle| Y = y \right] - \mathbb{E}_\theta \left[\frac{\nabla f_c(\theta)}{f_c(\theta)} \middle| Y = y \right] \mathbb{E}_\theta \left[\frac{\nabla f_c(\theta)}{f_c(\theta)} \middle| Y = y \right]^T \tag{9}$$

$$= \mathbb{E}_\theta \left[\frac{\nabla^2 f_c(\theta)}{f_c(\theta)} \middle| Y = y \right] - S(\theta; y) S(\theta; y)^T \tag{10}$$

$$\nabla^2 \ell_c(\theta) = \nabla \left(\frac{\nabla f_c(\theta)}{f_c(\theta)} \right) \tag{11}$$

$$= \frac{\nabla^2 f_c(\theta)}{f_c(\theta)} - S_c(\theta) S_c(\theta)^T \tag{12}$$

Combining lines 10 and 12, we get

$$\nabla^2 \ell(\theta) = \mathbb{E}_\theta[\nabla^2 \ell_c(\theta) | Y = y] + \mathbb{E}_\theta[S_c(\theta) S_c(\theta)^T | Y = y] - S(\theta; y) S(\theta; y)^T \tag{13}$$

Finally, evaluating this last line at $\theta = \hat{\theta}$ makes the rightmost term vanish, thereby yielding the required expression. □

Proposition 1.3 is known as Louis' standard error formula. Other decompositions for the observed information matrix of the observed data likelihood do exist; see, e.g., Oakes (1999); McLachlan and Krishnan (2008). However, the one due to Louis will be most useful to us later.

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.

2 The Monte Carlo EM Algorithm

The Monte Carlo EM, or MCEM, algorithm was first proposed by Wei and Tanner (1990), and replaces the conditional expectation in the E-step of the EM algorithm with a Monte Carlo average. That is, at each iteration we generate observations from the conditional distribution of the missing data given the observed data, and average the complete data likelihood over this Monte Carlo sample. This does alleviate the challenge of computing a potentially intractable conditional expectation, but does introduce a number of other difficulties. In the rest of this section, we outline some of these difficulties and methods which have been proposed to address them. We focus particularly on the practical aspects of implementing the MCEM algorithm. For an excellent survey of the theoretical considerations for the MCEM algorithm, see Neath (2013).

2.1 Quantifying Monte Carlo Uncertainty

In their seminal work, Wei and Tanner (1990) highlight two important challenges with implementing this method: choosing the Monte Carlo sample size at each iteration, and deciding when to terminate the algorithm. It turns out that solutions to these two problems are often connected via their link to the uncertainty in our Monte Carlo conditional expectations. The recommendations given by Wei and Tanner on how to solve these problems are mostly informal, but much of the later work on the MCEM algorithm centers around developing more precise solutions (too strong of a statement?).

Chan and Ledolter (1995) present an alternative method for choosing the Monte Carlo sample size based on starting with a pilot study and using information near the optimal parameter estimate[1]to choose a Monte Carlo size for the rest of the analysis. In contrast to Wei and Tanner (1990), Chan and Ledolter use a fixed Monte Carlo size for their analysis.

The method of Booth and Hobert (1999) centers on treating each iteration of the MCEM algorithm as an M-estimation problem targeting the deterministic EM update. This framework is quite natural, as an iteration of the MCEM algorithm consists of maximizing a Monte Carlo approximation to the EM objective function. Provided that certain regularity conditions are satisfied (see, e.g., van der Vaart, 1998), we can estimate the asymptotic standard error of the MCEM update as an estimate of the EM update with the same starting value, with the sampling variability induced by Monte Carlo simulation. Booth and Hobert then recommend constructing an asymptotic confidence set for the EM update, and increasing the Monte Carlo size if this confidence set contains the previous iteration's parameter estimate[2].

[1]Chan and Ledolter (1995) present an identity which expresses the observed data log-likelihood ratio as the conditional expectation of the corresponding complete data log-likelihood ratio. Replacing the conditional expectation with a Monte Carlo average gives a natural estimate of the observed data log-likelihood ratio. This approach closely resembles the Monte Carlo Maximum Likelihood method of Geyer (1994). See Section ???.

[2]For myself: As far as I can tell, in their paper, Booth and Hobert don't suggest that you check whether

Appendix A Likelihood for Linear Regression with an Unobserved Covariate

In this appendix, we present details for the analysis of our linear regression example with a single, unobserved, covariate. 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.

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mu\beta \\ \mu \end{pmatrix}, \begin{bmatrix} \sigma^2 + \tau^2\beta^2 & \tau^2\beta \\ \tau^2\beta & \tau^2 \end{bmatrix} \right) \tag{14}$$

Since our observed data, Y, is a marginal of the complete data, we can read off the distribution of Y from Expression (14). 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. Write $\theta = (\beta, \sigma)$ for the vector of unknown parameters, and $\eta^2 = \sigma^2 + \tau^2\beta^2$ for the marginal variance of Y. In previous work, I used η for $\mathbb{V}Y$. Make sure I adjust any discussion and **CODE(!!!)** accordingly.

$$\ell(\theta; y) = -\frac{n}{2}\log(2\pi) - n\log(\eta) - \sum_{i=1}^{n} \frac{(y_i - \mu\beta)^2}{2\eta^2} \tag{15}$$

$$\equiv -n\log(\eta) - \sum_{i=1}^{n} \frac{(y_i - \mu\beta)^2}{2\eta^2} \tag{16}$$

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

The score vector is given by

$$S(\theta; y) = \frac{1}{\eta^4} \begin{pmatrix} -n\beta^3\tau^4 - \beta^2\mu\tau^2 \sum y_i + \beta[\tau^2 \sum y_i^2 - n\sigma^2(\tau^2 + \sigma^2)] + \mu\sigma^2 \sum y_i \\ \sigma[n\beta^2(\mu^2 - \tau^2) - 2\beta\mu \sum y_i + \sum y_i^2 - n\sigma^2] \end{pmatrix} \tag{17}$$

The observed information matrix is given by

$$I(\theta; y) = \frac{1}{\eta^6} \begin{bmatrix} I^{(1,1)} & I^{(1,2)} \\ I^{(1,2)} & I^{(2,2)} \end{bmatrix} \tag{18}$$

---

augmenting the Monte Carlo size produces a confidence set which does not contain the previous iteration's estimate. I think they're just suggesting you increase M and proceed with the next iteration. This differs from Caffo et al. (2005) who require that we keep checking and augmenting M until the condition is satisfied. I don't think this needs to go in the paper, but it's of some interest if you want to implement the method of Booth and Hobert

where

$$I^{(1,1)} = -n\beta^4\tau^6 - 2\beta^3\mu\tau^4\sum y_i + 3\beta^2\tau^2(3\tau^2\sum y_i^2 - n\sigma^2\mu^2) \tag{19}$$
$$+ 6\beta\sigma^2\tau^2\mu\sum y_i + \sigma^2[n\sigma^2(\tau^2+\mu^2) - \tau^2\sum y_i]$$
$$I^{(1,2)} = 2n\beta^3\sigma\tau^2(\mu^2-\tau^2) - 6\beta^2\sigma\mu\tau^2\sum y_i + 2\beta\sigma[2\tau^2\sum y_i^2 - n\sigma^2(\mu^2+\tau^2)] \tag{20}$$
$$+ 2\mu\sigma^3\sum y_i$$
$$I^{(2,2)} = n\beta^4\tau^2(\tau^2-\mu^2) + 2\beta^3\mu\tau^2\sum y_i + \beta^2(3n\mu^2\sigma^2 - \tau^2\sum y_i^2) \tag{21}$$
$$- 6\beta\mu\tau^2\sum y_i + \sigma^2(3\sum y_i^2 - n\sigma^2)$$

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

James G. Booth and James P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(1), 1999.

Brian S. Caffo, Wolfgang Jank, and Galin L. Jones. Ascent-based Monte Carlo expectation-maximization. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2), 2005.

K. S. Chan and Johannes Ledolter. Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association*, 90(429), 1995.

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.

Charles J. Geyer. On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1), 1994.

Thomas A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 44(2), 1982.

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

Ronald C. Neath. On convergence properties of the Monte Carlo EM algorithm. *Advances in Modern Statistical Theory and Applications*, 10, 2013.

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

David Oakes. Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(2), 1999.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Greg C. G. Wei and Martin A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411), 1990.

Index