Abstract

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

- Replace "conditional distribution of the missing data given the observed data" with "missing data distribution"
- Caffo et al. (2005) use M_k for the Monte Carlo size at iteration k. This is a useful definition for discussing other papers too.
- I have changed my iteration labels. $\hat{\theta}_{k-1}$ is now the maximizer of the current MCEM objective function, and $\hat{\theta}_{k-1}$ was used to construct that objective function. Watch for this when editing.

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 is given by

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

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 $\theta_k = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta_{k-1})$. 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 $M(\theta_{k-1}) = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta_{k-1})$. 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 to any point in Θ such that $Q(\theta_k|\theta_{k-1}) \geq Q(\theta_{k-1}|\theta_{k-1})$.

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.). Does it make sense to define $\mathcal{I}_c$ and $\mathcal{I}_m$ in the following proposition?

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

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

(ii) $I(\theta) = \mathcal{I}_c(\theta) - \mathcal{I}_m(\theta)$
Where $\mathcal{I}_c(\theta) := -\mathbb{E}_\theta\left[\nabla^2 \ell_c(\theta; y, X)|Y = y\right]$ *and* $\mathcal{I}_m(\theta) := -\mathbb{E}_\theta\left[\nabla^2 \ell_m(\theta; y, X)|Y = y\right]$

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*

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

In particular, if $\hat{\theta}$ is a stationary point of the observed data log-likelihood, then

$$I(\hat{\theta}) = \mathcal{I}_c(\hat{\theta}) - \mathbb{E}_{\hat{\theta}}[S_c(\hat{\theta})S_c(\hat{\theta})|Y = y] \tag{7}$$

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:

$$\nabla^2 \ell(\theta) = \nabla \left[\int_{\mathcal{X}(y)} \frac{\nabla f_c(\theta) dx}{f(\theta)} \right] \tag{8}$$

$$= \int_{\mathcal{X}(y)} \frac{\nabla^2 f_c(\theta)}{f(\theta)} dx - \frac{1}{f(\theta)^2} \left(\int_{\mathcal{X}(y)} \nabla f_c(\theta) dx \right) \left(\int_{\mathcal{X}(y)} \nabla f_c(\theta) dx \right)^T \tag{9}$$

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

$$= \mathbb{E}_\theta \left[\frac{\nabla^2 f_c(\theta)}{f_c(\theta)} \middle| Y = y \right] - S(\theta; y)S(\theta; y)^T \tag{11}$$

$$\nabla^2 \ell_c(\theta) = \nabla \left(\frac{\nabla f_c(\theta)}{f_c(\theta)} \right) \tag{12}$$

$$= \frac{\nabla^2 f_c(\theta)}{f_c(\theta)} - S_c(\theta)S_c(\theta)^T \tag{13}$$

Combining lines 11 and 13, we get

$$\nabla^2 \ell(\theta) = \mathbb{E}_\theta[\nabla^2 \ell_c(\theta)|Y = y] + \mathbb{E}_\theta[S_c(\theta)S_c(\theta)^T|Y = y] - S(\theta; y)S(\theta; y)^T \tag{14}$$

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). This method proceeds by replacing the conditional expectation in the E-step of the EM algorithm with a Monte Carlo average. More precisely, 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. Formally, at a given iteration of the MCEM algorithm, let $X_1, \ldots, X_M$ be a Monte Carlo sample from the law of $X|Y = y$ with θ set to the value from the previous iteration, say θ_0. Write

$$\begin{aligned} \hat{Q}(\theta|\theta_0) &= \sum_{i=1}^{M} w_i \ell_c(\theta; y, X_i) && (15) \\ &:= \hat{\mathbb{E}} \ell_c(\theta; y, X) && (16) \end{aligned}$$

Where the w_i are sampling weights. Confirm that this operator notation isn't contradicted elsewhere. Under iid sampling we simply get $w_i = M^{-1}$ for every i, but more involved sampling schemes may have more complicated weights. The estimate of θ is then the maximizer of the MCEM objective function: $\hat{\theta} = \mathrm{argmax}_\theta \, \hat{Q}(\theta|\theta_0)$. Write $\hat{\theta}_{k-1}$ for the kth MCEM estimate.

Provided that a valid sampling scheme is available for the missing data distribution, we can use Proposition 1.3 to estimate the observed data information matrix.

Proposition 2.1. *Under the conditions of Proposition 1.3, as well as any required for the sampler, we get*

$$-\hat{\mathbb{E}}_{\hat{\theta}} \nabla^2 \ell_c(\hat{\theta}) - \hat{\mathbb{E}}_{\hat{\theta}} S_c(\hat{\theta}) S_c(\hat{\theta})^T \to I(\theta) \tag{17}$$

Under stronger conditions, we also get asymptotic normality with variance obtained from importance sampling analysis.

The MCEM algorithm has the advantage of circumventing the challenge of computing potentially intractable conditional expectations for the EM algorithm. However, this analytical simplification does come at the cost of introducing some new computational problems. In this section, we outline the main problems faced by the MCEM algorithm and present various solutions which have been proposed in the literature. We focus primarily on practical aspects of the MCEM algorithm; see Neath (2013) for a survey of theoretical considerations.

Two problems which have received considerable attention in the literature are how to choose the Monte Carlo sample size at each iteration, and how to decide when to terminate the MCEM algorithm. These were identified as early as Wei and Tanner (1990), but did not receive systematic treatment until later. We here give a brief overview of different authors' approaches to solving these two problems, and spend the rest of this section going into more detail on each method individually. Wei and Tanner (1990) suggest examining a plot of the parameter estimates across iterations, and either terminating or increasing the Monte Carlo size when the plot appears to stabilize. Chan and Ledolter (1995) use a pilot study to choose the Monte Carlo sample size, and terminate when a confidence interval for the improvement of the observed data log-likelihood between successive iterations contains zero. Booth and Hobert (1999) frame each MCEM iteration as an M-estimation problem targeting the deterministic EM update. They increase the Monte Carlo size if an asymptotic confidence interval for the EM update contains the previous iteration's parameter estimate, and terminate when multiple successive iterations' estimates have sufficiently small relative error. Caffo et al. (2005) build confidence bounds for the increment in the EM objective function at each iteration of the MCEM algorithm. They increase the Monte Carlo size until the lower bound is positive and terminate when the upper bound is sufficiently small.

In the rest of this section, we give more detail on each of the implementations introduced above.

2.1 Early Work (Wei and Tanner, 1990)

In their seminal work, Wei and Tanner (1990) propose the MCEM algorithm and present a simple implementation. They illustrate that the complete data gradient and Hessian are easily obtained at each iteration from the Monte Carlo sample and, following Louis (1982), give an estimator for the observed data information matrix. Regarding convergence, Wei and Tanner recommend plotting the parameter estimates across iterations and stopping when the estimates appear to stabilize around some constant. When this stabilization is detected, one can either declare convergence and stop, or increase the Monte Carlo size and continue iterating until the estimate trajectory again stabilizes.

2.2 Running a Pilot Study

2.3 Uncertainty Quantification for the Parameter Estimate (Booth and Hobert, 1999)

Building on the ideas of Wei and Tanner, Booth and Hobert (1999) seek to start the MCEM algorithm with a small Monte Carlo size, and add more observations only when the parameter estimates are no longer changing discernibly across iterations. To this end, they recommend building a confidence interval for the EM update based on the Monte Carlo variability of the MCEM update at each iteration. If this interval contains the previous iteration's parameter estimate, then the parameter updates are too small relative to the amount of Monte Carlo variability and more samples are required. Similarly, Booth and Hobert recommend assessing convergence by checking for small relative error in the parameter updates. To account for the possibility of Monte Carlo variability leading to two consecutive estimates being similar before the algorithm has 'converged', the authors suggest waiting until the relative error is small for three consecutive iterations.

The confidence interval used to quantify Monte Carlo uncertainty within an iteration is obtained by framing the parameter update as the solution of an M-estimation problem. This allows us to inherit the desirable properties of M-estimators; specifically, asymptotic normality. See, e.g. van der Vaart (1998). Following the usual M-estimator construction and assuming that the relevant regularity conditions hold, we are able to estimate the asymptotic variance of the MCEM parameter estimator at each iteration. Note that this standard error is based on the Monte Carlo variability within an iteration; it does not measure sampling variability due to the observed data.

More formally, write $\tilde{\theta}_k$ for the EM update based on $\hat{\theta}_{k-1}$. Note that $\hat{\theta}_{k-1}$ is held fixed here and in the next subsection. Analysis of a single MCEM iteration is done conditional on the previous iteration (awk?). Unless stated otherwise, all expectations are taken with $\theta = \hat{\theta}_{k-1}$. Assuming sufficient smoothness and moment conditions, we get the following expression for the MCEM update:

$$\sqrt{M}(\hat{\theta}_k - \tilde{\theta}_k) = -\sqrt{M}\left[\nabla^2 Q(\tilde{\theta}_k|\hat{\theta}_{k-1})\right]^{-1}\left[\nabla \hat{Q}(\tilde{\theta}_k|\hat{\theta}_{k-1})\right] + o_p(1) \tag{18}$$

Where M is the Monte Carlo size and ∇ denotes differentiation with respect to the left argument of Q or $\hat{Q}$. Note that the first expression on the right-hand side is the inverse Hessian of the EM objective function (fixed) while the second is the gradient of the MCEM objective function (an average). Thus, $\hat{\theta}_k$ is asymptotically normal with asymptotic variance

$$\left[\nabla^2 Q(\tilde{\theta}_k|\hat{\theta}_{k-1})\right]^{-1} \mathbb{V}\left[S_c(\tilde{\theta}_k)|Y=y\right]\left[\nabla^2 Q(\tilde{\theta}_k|\hat{\theta}_{k-1})\right]^{-1} \tag{19}$$

$$\approx \left[\nabla^2 \hat{Q}(\hat{\theta}_k|\hat{\theta}_{k-1})\right]^{-1} \hat{\mathbb{E}}\left[S_c(\hat{\theta}_k)S_c(\hat{\theta}_k)^T|Y=y\right]\left[\nabla^2 \hat{Q}(\hat{\theta}_k|\hat{\theta}_{k-1})\right]^{-1} \tag{20}$$

Where S_c is the complete data score vector, and $\hat{\mathbb{E}}$ is the Monte Carlo average over the missing data, with $\hat{\theta}_k$ held fixed (remove comma?). Note that there is no first moment term in the conditional variance of S_c because $\hat{\theta}_k$ is a maximizer of $\hat{\mathbb{E}}[\ell_c|Y = y]$.

Based on the above discussion, we can build an asymptotic confidence interval for $\tilde{\theta}_k$, the EM update based on the MCEM estimate from iteration k. Booth and Hobert recommend checking whether this interval contains $\hat{\theta}_{k-1}$ and, if so, increasing M for the next iteration. Specifically, they suggest starting the next iteration with M/r more points, with $r = 3, 4$ or 5 working well in their examples (this whole discussion is pretty awkward).

To assess convergence of the MCEM algorithm, Booth and Hobert present two criteria. The first is a familiar measure of relative error in parameter estimates between consecutive iterations:

$$\max_j \left(\frac{\left|\hat{\theta}_{k,j} - \hat{\theta}_{k-1,j}\right|}{\left|\hat{\theta}_{k-1,j}\right| + \delta_1} \right) < \delta_2 \tag{21}$$

Where δ_1 and δ_2 are small positive constants, and the subscript j ranges over components of θ. Booth and Hobert suggest using $\delta_1 = 10^{-3}$ and δ_2 between $2 \cdot 10^{-3}$ and $5 \cdot 10^{-3}$. See (Citation Needed) (probably Searle et al., 2006, p. 436, or Marquardt, 1963) for why condition (21) has this particular form.

Alternatively, since Booth and Hobert apply their method to the analysis of generalized linear mixed models, where pathologies may arise due to parameter estimates being too close to a boundary, they propose a second stopping rule:

$$\max_j \left(\frac{\left|\hat{\theta}_{k,j} - \hat{\theta}_{k-1,j}\right|}{\sqrt{\mathbb{V}\hat{\theta}_{k-1,j} + \delta_1'}} \right) < \delta_2' \tag{22}$$

How do we estimate the variance here? The purpose of condition (22) is to detect when estimated variance components are very close to zero and the numerical precision needed to satisfy condition (21) requires a prohibitive amount of computation.

2.4 Uncertainty Quantification for the Objective Function (Caffo et al., 2005)

The approach of Caffo et al. (2005) is similar in spirit to that of Booth and Hobert (1999). The goal here is to quantify Monte Carlo uncertainty in the MCEM algorithm as an approximation to the EM algorithm. The difference is that where Booth and Hobert measure uncertainty in the parameter estimate, Caffo et al. focus on uncertainty in the objective function. Specifically, Caffo et al. base their analysis asymptotic normality of the MCEM increment:

Proposition 2.2. *Let $\Delta\hat{Q}(\theta_k|\theta_{k-1}) = \hat{Q}(\theta_k|\theta_{k-1}) - \hat{Q}(\theta_k|\theta_{k-1})$. Define $\Delta Q(\theta_k|\theta_{k-1})$ similarly. Let M_k be the Monte Carlo size at iteration k. Then*

$$\sqrt{M_k}\left[\Delta\hat{Q}(\theta_k|\theta_{k-1}) - \Delta Q(\theta_k|\theta_{k-1})\right] \rightsquigarrow N(0, \Sigma_k) \tag{23}$$

3 Simulation

A further obstacle to implementing the MCEM algorithm which was not addressed in the previous section, is how to generate the Monte Carlo sample. It is in general a hard problem to generate from arbitrary conditional distributions. In fact, much of the work in the Bayesian Computation literature centers on this problem (see, e.g., Gelman et al., 2013). In this section, we discuss several methods for simulating the necessary observations at each step of the MCEM algorithm.

4 Alternatives to the MCEM Algorithm

In this section, we outline a few alternatives to the MCEM algorithm for maximizing the likelihood of an incomplete dataset. Examples include the Monte Carlo Maximum Likelihood method of Geyer (1994), and Variational Inference (Blei et al., 2017; Tsikas et al., 2008).

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.

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mu\beta \\ \mu \end{pmatrix}, \begin{bmatrix} \sigma^2 + \tau^2\beta^2 & \tau^2\beta \\ \tau^2\beta & \tau^2 \end{bmatrix} \right) \tag{24}$$

Since our observed data, Y, is a marginal of the complete data, we can read off the distribution of Y from Expression (24). 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.

$$\begin{aligned} \ell(\theta; y) &= -\frac{n}{2}\log(2\pi) - n\log(\eta) - \sum_{i=1}^{n} \frac{(y_i - \mu\beta)^2}{2\eta^2} && (25) \\ &\equiv -n\log(\eta) - \sum_{i=1}^{n} \frac{(y_i - \mu\beta)^2}{2\eta^2} && (26) \end{aligned}$$

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

The score vector is given by

$$S(\theta; y) = \frac{1}{\eta^4} \begin{pmatrix} -n\beta^3\tau^4 - \beta^2\mu\tau^2 \sum y_i + \beta[\tau^2 \sum y_i^2 - n\sigma^2(\tau^2 + \sigma^2)] + \mu\sigma^2 \sum y_i \\ \sigma[n\beta^2(\mu^2 - \tau^2) - 2\beta\mu \sum y_i + \sum y_i^2 - n\sigma^2] \end{pmatrix} \tag{27}$$

The observed information matrix is given by

$$I(\theta; y) = \frac{1}{\eta^6} \begin{bmatrix} I^{(1,1)} & I^{(1,2)} \\ I^{(1,2)} & I^{(2,2)} \end{bmatrix} \tag{28}$$

where

$$\begin{aligned} I^{(1,1)} &= -n\beta^4\tau^6 - 2\beta^3\mu\tau^4\sum y_i + 3\beta^2\tau^2(3\tau^2\sum y_i^2 - n\sigma^2\mu^2) \qquad (29) \\ &\quad + 6\beta\sigma^2\tau^2\mu\sum y_i + \sigma^2[n\sigma^2(\tau^2+\mu^2) - \tau^2\sum y_i] \\ I^{(1,2)} &= 2n\beta^3\sigma\tau^2(\mu^2-\tau^2) - 6\beta^2\sigma\mu\tau^2\sum y_i + 2\beta\sigma[2\tau^2\sum y_i^2 - n\sigma^2(\mu^2+\tau^2)] \qquad (30) \\ &\quad + 2\mu\sigma^3\sum y_i \\ I^{(2,2)} &= n\beta^4\tau^2(\tau^2-\mu^2) + 2\beta^3\mu\tau^2\sum y_i + \beta^2(3n\mu^2\sigma^2 - \tau^2\sum y_i^2) \qquad (31) \\ &\quad - 6\beta\mu\tau^2\sum y_i + \sigma^2(3\sum y_i^2 - n\sigma^2) \end{aligned}$$

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

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518), 2017.

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.

Citation Needed.

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.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis.* CRC Press, third edition, 2013.

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.

Donald W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2), 1963.

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.

Shayle R. Searle, George Casella, and Charles E. McCulloch. *Variance Components.* Wiley Interscience, 2006.

Dimitris G. Tsikas, Aristidis C. Likas, and Nikolaos P. Galatsanos. The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine*, 25(6), 2008.

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