DISCUSSION ARTICLE

Discussion

Xiao-Li Meng

1. IT'S ALL IN THE NAME!

Of the several reasons for the popularity of the EM algorithm after the publication of Dempster, Laird, and Rubin (1977), one is its name. Almost at the instant of inquiring what EM stands for, the curious mind is already learning that the algorithm has two steps—the expectation step and the maximization step. Incidentally, the substance-oriented name also avoids the common distraction governed by Stigler's Law of Eponymy (Stigler 1980), and avoids awkward, "noninformative" acronyms such as the FHBSMDLR algorithm (for curious minds, see Meng and van Dyk 1997, sec. 1.1).

Since the authors hope their article will stimulate a nonnegligible amount of research activities compared to Dempster et al. (1977), a "sexier" name than *optimization transfer* seems in order, at least for statisticians. May I suggest *the SM algorithm*? Like EM, it immediately identifies that the algorithm has two steps (at iteration t) for *maximizing* an objective function $L(\theta)$ over $\theta \in \Theta$:

1. **Surrogate Step:** Substitute a surrogate function $Q\left(\theta|\theta^{(t)}\right)$ for $L(\theta)$ such that

$$H(\theta|\theta^{(t)}) \equiv Q\left(\theta|\theta^{(t)}\right) - L(\theta), \qquad \theta \in \Theta$$

   attains its maximum at $\theta = \theta^{(t)}$; and

2. **Maximization Step:** Maximize the surrogate function $Q\left(\theta|\theta^{(t)}\right)$ as a function of θ to determine the next iterate $\theta^{(t+1)}$.

Also like EM, this is really not an algorithm but rather a general principle (see the footnote on p. 6 of Dempster et al. 1977)—in fact, without further instruction on how to construct the surrogate function, it is really just a *principle*. But given that EM is now a household name, the new name SM may catch on simply because it rhymes (almost) with EM! (A physician once called me: "I heard about this cool stuff called EM. Can you tell me about it?" Now I can call him back: "I have this really cool stuff called SM. Do you want to hear about it?")

With this new spice, we can cook another alphabet soup. As a direct counterpart of GEM (Dempster et al. 1977), we have GSM (MSG in reverse!), which finds $\theta^{(t+1)}$

Xiao-Li Meng is Associate Professor, Department of Statistics, The University of Chicago, IL 60637 (E-mail: meng@galton.uchicago.edu).

Journal of Computational and Graphical Statistics, Volume 9, Number 1, Pages 35–43

such that $Q(\theta^{(t+1)}|\theta^{(t)}) > Q(\theta^{(t)}|\theta^{(t)})$, but does not necessarily maximize $Q\left(\theta|\theta^{(t)}\right)$. Similarly with the ECM algorithm (Meng and Rubin 1993), we can replace the M step by a set of conditional maximization steps, and hence the SCM algorithm. It is sometimes beneficial to use $L(\theta)$ as the surrogate function for itself in some of the CM steps, as in the ECME algorithm (Liu and Tubin 1994), which leads to SCME. Or more generally, we can have ASCM; that is, we can alternate the surrogate functions with the CM steps, as detailed in Meng and van Dyk (1997) in the AECM framework. In addition, in analogy to moving from EM to GEM, we can move from AECM to GAECM, which is the most general EM-type framework I am aware of. Consequently, we can move from ASCM to GASCM, which is likely to be currently the most fruitful framework for statisticians to construct intrinsically monotone optimization algorithms (i.e., the monotonicity is not "forced" by checking values of the objective function at each iteration). Furthermore, we can introduce a working parameter to index a set of surrogate functions for the purpose of optimizing speed (as in Meng and van Dyk 1997, 1999), or as in the PXEM algorithm (Liu, Rubin, and Wu 1998), we can maximize the working/expanded parameter in the SM iteration, and hence PXSM.

Finally, we may even try the Supplemented SM and SCM algorithms to mimic the SEM algorithm (Meng and Rubin 1991) and the SECM algorithm (van Dyk, Meng, and Rubin 1995) for computing the asymptotic variances, though these are less straightforward than the previous replacements because the rate of convergence of SM and SCM may not be directly related to the fraction of missing information. However, when directly differentiating the surrogate function $Q(\theta|\phi)$ is feasible with respect to *both* θ and ϕ, there is generally no need of a numerical algorithm for computing the second derivative of $L(\theta)$; see Section 4.

2. IS SM JUST EM?

Of course, the new alphabet soup will not be a really new delight if it is just the old soup presented in a new, perhaps larger, bowl. Could it be that SM, though apparently more general, is just a disguised or beautified version of EM? The answer is not completely obvious, especially if one starts the comparison with the most obvious construction of the surrogate function via linear minorization/majorization. As in the authors' Equation (3.1) (p. 9), assume our log-likelihood function $L(\theta|y)$ can be written as

$$L(\theta|y) = f_y(\theta) - g_y(\theta), \quad \theta \in \Theta \subset R^1, \tag{2.1}$$

where both f_y and g_y are concave functions and without loss of generality (when $|g_y(0)| < \infty$) we assume $g_y(0) = 0$ for all y. Now suppose $e^{-g_y(\theta)}$ is the moment-generating function of a conditional density $h(z|y)$, namely,

$$e^{-g_y(\theta)} = \int e^{\theta z} h(z|y)\mu(dz), \quad \theta \in \Theta. \tag{2.2}$$

Then if we augment $p(y|\theta) = e^{L(\theta|y)}$ to

$$p(z|y,\theta)p(y|\theta) \equiv \left[e^{\theta z + g_y(\theta)}h(z|y)\right]\left[e^{L(\theta|y)}\right] = e^{f_y(\theta)+\theta z}h(z|y), \tag{2.3}$$

we have, for the standard EM construction,

$$Q\left(\theta|\theta^{(t)}\right) = f_y(\theta) + \theta \mathrm{E}\left(Z|y,\theta^{(t)}\right) + \mathrm{E}\left[\log h(Z|y)|y,\theta^{(t)}\right]. \tag{2.4}$$

But this is equivalent to the proposed linear minorization surrogate function

$$Q\left(\theta|\theta^{(t)}\right) = f_y(\theta) - g_y'\left(\theta^{(t)}\right)\left(\theta - \theta^{(t)}\right), \tag{2.5}$$

because $\mathrm{E}(Z|y,\theta) = -g_y'(\theta)$ from differentiating both sides of (2.2). Incidentally, by differentiating both sides of (2.2) twice, we have $g_y''(\theta) = -\mathrm{V}(Z|y,\theta) \leq 0$, and thus the concavity of $g(\theta)$ is a necessary condition for this EM construction to be possible. (For multivariate θ we can construct the missing data Z with the same dimension and replace θz in (2.2) with $\theta^\top z$.)

Although *this* EM construction is not always possible (e.g., when $e^{-g_y(\theta)}$ may not be a moment-generating function), and even when it is possible it requires more brain power than the linear minorization method, it nevertheless suggests that a large class of SM algorithms based on (2.5) are also EM algorithms with augmentation $p(z,y|\theta)$ of (2.3).

So the question is, given $Q(\theta|\phi)$ from a particular SM construction, how do we know if there is a corresponding EM construction, regardless of how convoluted the latter might be? The practical relevance of this theoretically interesting question is that, if the EM class is as rich as the SM class, then the value of the new SM formulation is in providing a set of new tools for creative EM-type implementation. However, if the SM class is richer than the EM class, then it provides hope for solving problems that are difficult or even impossible to solve within the entire GAECM framework.

3. SO WHAT DOES IT TAKE TO BE AN EMer?

Let us call a surrogate function $Q(\theta|\phi)$ on $\Theta \times \Theta$ an *EMer* for an objective function $L(\theta), \theta \in \Theta$ if the following two conditions hold:

- **Condition 1:** There exists an *augmented* objective function $L(\theta; z)$, where z can be of any dimension, such that

$$p(z|\theta) \equiv e^{L(\theta;z)-L(\theta)} \tag{3.1}$$

  is a *proper* density with respect to some measure μ for any $\theta \in \Theta$; and
- **Condition 2:** The surrogate function $Q(\theta|\phi)$ can be expressed as

$$Q(\theta|\phi) = \mathrm{E}[L(\theta;Z)|\phi] + C(\phi) = \int L(\theta;z)p(z|\phi)\mu(dz) + C(\phi), \quad \text{for any } (\theta,\phi) \in \Theta \times \Theta, \tag{3.2}$$

  where $C(\phi)$ is a function of ϕ alone.

This definition is notationally more general than the one given in Dempster et al. (1977), because it explicitly allows $L(\theta)$ and $L(\theta; z)$ to be arbitrary objective functions as long as $p(z|\theta)$ of (3.1) is a proper density. A closer examination of the theory provided in Dempster et al. (1977) will reveal that it does not require $L(\theta)$ or $L(\theta; z)$ to be log-likelihood functions, as emphasized in the rejoinder of Meng and van Dyk (1997). Also note that in standard EM literature, $p(z|\theta)$ is expressed as $p(z|\theta, y)$, the conditional density of the missing variable Z given the observed data $Y = y$.

The following result provides a necessary and sufficient condition for a surrogate function $Q(\theta|\phi)$ to be an EMer.

Lemma 1. *A surrogate function $Q(\theta|\phi)$ is an EMer for $L(\theta), \theta \in \Theta$ iff there exists a probability family $\{p(z|\theta), \theta \in \Theta\}$ with respect to a measure μ such that*

$$H(\phi|\phi) - H(\theta|\phi) = \int \log \left[\frac{p(z|\phi)}{p(z|\theta)}\right] p(z|\phi)\mu(dz) \equiv KL(\phi : \theta), \tag{3.3}$$

where $H(\theta|\phi) = Q(\theta|\phi) - L(\theta)$ and $KL(\phi : \theta)$ is known as the Kullback–Leibler information, under family $\{p(z|\theta), \theta \in \Theta\}$, in favor of ϕ against θ when ϕ is true.

Proof: The necessity follows directly from (3.1) and (3.2), which imply that for any $(\theta, \phi) \in \Theta \times \Theta$,

$$\begin{aligned} H(\phi|\phi) - H(\theta|\phi) &= \mathrm{E}[L(\phi; Z) - L(\phi)|\phi] \\ &\quad -\mathrm{E}[L(\theta; Z) - L(\theta)|\phi] = \int \log \left[\frac{p(z|\phi)}{p(z|\theta)}\right] p(z|\phi)\mu(dz). \end{aligned} \tag{3.4}$$

To prove the sufficiency, we note that if (3.4) (ignoring the expression in the middle) holds for some $\{p(z|\theta), \theta \in \Theta\}$, then

$$H(\theta|\phi) = \int \log p(z|\theta) p(z|\phi)\mu(dz) + C(\phi), \tag{3.5}$$

where $C(\phi)$ is a function of ϕ only. Letting $L(\theta; z) = \log p(z|\theta) + L(\theta)$, which clearly satisfies Condition 1, we have from (3.5) that

$$Q(\theta|\phi) = H(\theta|\phi) + L(\theta) = \int L(\theta; z) p(z|\phi)\mu(dz) + C(\phi), \quad \text{for any} \quad (\theta, \phi) \in \Theta \times \Theta, \tag{3.6}$$

which is Condition 2. □

Lemma 1 says that to demonstrate that the SM class is more general than the EM class, all we need to do is to find a function $H(\theta|\phi)$ on $(\theta, \phi) \in \Theta \times \Theta$, where $\Theta \subset R^d$, such that

- **Requirement 1:** $H(\phi|\phi) - H(\theta|\phi) \geq 0$ for all $(\theta, \phi) \in \Theta \times \Theta$, as required by the S-Step; but
- **Requirement 2:** $H(\phi|\phi) - H(\theta|\phi)$ cannot be represented as a $KL(\phi : \theta)$, as required to leave the class of EMer.

To my amusement and frustration, this seemingly trivial task has doubled my headache from the Shanghai flu! The class of functions $H(\theta|\phi)$ that satisfy Requirement 1 is enormous, and the class of $KL(\phi:\theta)$ seems much more restrictive especially because of the separation of ϕ and θ inside the integrand, $\int \log p(z|\theta)p(z|\phi)\mu(dz)$. However, the class of *missing data* densities $p(z|\theta)$ is also enormous, especially because there is no restriction on the dimensionality of z. It is thus very difficult to prove Requirement 2 for any given $H(\theta|\phi)$ on a given $\Theta \times \Theta$. Pathological examples do exist when there is no restriction on Θ, for example, by taking Θ to be the *power set* of the set of all probability functions and let $H(\theta|\phi) = \delta_{\{\theta=\phi\}}$, an example constructed by my colleague Zhiyi Chi. Unfortunately, such examples do not shed much light on how one should proceed when $\Theta \subset R^d$, situations that are relevant for statistical applications.

When Θ is a differentiable manifold, an $H(\theta|\phi)$ satisfying Requirement 1 is a *yoke* if $H \in C^\infty(\Theta \times \Theta)$, and $\bar{H}(\theta|\phi) \equiv H(\theta|\phi) - H(\phi|\phi)$ is a *normalized/normed* yoke (Barndorff-Nielsen 1987; Barndorff-Nielsen and Cox 1994). One of the most important yokes in the differential-geometric approach to statistical asymptotics is the *expected (log-) likelihood yoke*, $\mathrm{E}[\log p(z|\theta) - \log p(z|\phi)|\phi]$, which is exactly the negative of $KL(\phi : \theta)$. So under the differentiability assumption, the mathematical questions that have doubled my headache are:

1. For a given Θ, is there a normed yoke on $C^\infty(\Theta \times \Theta)$ that cannot be represented as an expected likelihood yoke?
2. For a given normed yoke on $C^\infty(\Theta \times \Theta)$, how can one determine if it has an expected likelihood yoke representation?

Question 1 perhaps is not too hard to answer using the representation theory of yokes given by Barndorff-Nielsen and Jupp (1997), which is unfortunately too difficult for most statisticians even without headache. Question 2 perhaps is a lot harder to answer, but it is also a question of theoretical interest only because once a SM algorithm is constructed it does not really matter whether or not it is also an EM algorithm since the former also guarantees the celebrated monotone convergence property of EM. However, the theoretical results in the literature on yokes, especially those on how to generate new yokes from a given yoke (e.g., Barndorff-Nielsen and Jupp 1997), seem to me quite relevant for the SM algorithm, because for every yoke $H(\theta|\phi)$ there is a corresponding surrogate function $Q(\theta|\phi) = H(\theta|\phi) + L(\theta)$ for the SM implementation, at least in theory. Evidently, the more yokes we can choose from, the more likely we can construct algorithms that are simple, stable, and fast.

On the other hand, the formulation of the SM algorithm may call for generalizations of the theory of yoke beyond the one suggested in Blæsild (1991); namely, H is only required to be continuously differentiable for a finite number of terms. As emphasized by the authors, one advantage of the SM algorithm is its ability of transferring the optimization of a nondifferentiable objective function to that of a differentiable surrogate function, as demonstrated by the L_1 regression problem in Lange, Hunter, and Yang's Example 2 (p. 4). In such cases, the $H(\theta|\phi) = Q(\theta|\phi) - L(\theta)$ function is not differentiable, so we need to extend the theory of yoke to functions $H(\theta|\phi)$ that satisfy Requirement 1 but do not necessarily satisfy any differentiability assumption.

So although my attempt to cure my "EM flu" has not been successful, it is not without pleasant consequences (more will be reported in the next section). Furthermore, because the article's first author Lange is a leading statistical mathematician who can go back and forth between statistics and mathematics with great ease, I am very hopeful that he, together with his coauthors, will be able to provide a cure for my "EM flu."

4. MEETING AN OLD FRIEND: MR. BARTLETT

In the search for necessary conditions for a surrogate function to be an EMer, the form of $H(\theta|\phi)$ given in (3.5) initially suggested that I consider the well known Bartlett identities for the family $\{p(z|\theta), \theta \in \Theta\}$. Specifically, suppose θ is univariate and it is legitimate to interchange the differential and integration operators as needed. Denote $D^{u,v}F(\theta_1, \theta_2) = \frac{\partial^{u+v}F(\theta_1,\theta_2)}{\partial\theta_1^u \partial\theta_2^v}$. Then by differentiating the following identity k (≥ 0) times

$$D^{1,0}H(\theta|\theta) = \int \left[\frac{d \log p(z|\theta)}{d\theta}\right] p(z|\theta)\mu(dz) = 0, \quad \text{for any } \theta \in \Theta, \tag{4.1}$$

and by using the chain rule for differentiating the product of two functions, we obtain that for $H(\theta|\phi)$ of (3.5),

$$\sum_{j=0}^{k} \binom{k}{j} D^{j+1,k-j} H(\theta|\theta) = 0, \tag{4.2}$$

or equivalently

$$L^{(k+1)}(\theta) = \sum_{j=0}^{k} \binom{k}{j} D^{j+1,k-j} Q(\theta|\theta), \tag{4.3}$$

for any $k \geq 0$ such that all the derivatives involved exist.

Identity (4.3) is indeed a necessary condition for $Q(\theta|\phi)$ to be an EMer, but this is because it is actually a necessary condition for *any* surrogate function as defined by the S step, under the assumption of suitable differentiability of $L(\theta)$ and $Q(\theta|\phi)$. Given the important practical implication of this result (see, e.g., (4.9)), I will list it as a lemma, even though it is a direct consequence of $H(\theta|\phi)$ satisfying Requirement 1, a requirement that defines the surrogate function and is explicitly or implicitly assumed and used through out the authors' article.

Lemma 2. *Suppose $\theta = (\theta_1, \ldots, \theta_d)$. Denote*

$$D^{J,K}F(\theta|\phi) = \frac{\partial^{\sum_{\alpha=1}^{d}(j_\alpha + k_\alpha)} F(\theta|\phi)}{\partial\theta_1^{j_1} \cdots \partial\theta_d^{j_d} \partial\phi_1^{k_1} \cdots \partial\phi_d^{k_d}} \tag{4.4}$$

for a function $F(\theta|\phi)$, where $J = (j_1, \ldots, j_d)$ and $K = (k_1, \ldots, k_d)$. Denote

$$\binom{K}{J} = \binom{k_1}{j_1} \cdots \binom{k_d}{j_d} \quad \text{and} \quad \sum_{J=\mathbf{0}}^{K} f(J) = \sum_{j_1=0}^{k_1} \cdots \sum_{j_d=0}^{k_d} f(j_1, \ldots, j_d), \tag{4.5}$$

where $\mathbf{0}$ = (0, ..., 0), and let E_i be the row vector with 1 for its ith element and 0 elsewhere, for i = 1, ..., d. Suppose $Q(\theta|\phi)$ is a surrogate function for $L(\theta)$ such that for any fixed $\phi \in \Theta$, $\theta = \phi$ is a stationary point of $H(\theta|\phi) = Q(\theta|\phi) - L(\theta)$. Then $Q(\theta|\phi)$ must satisfy

$$\sum_{J=\mathbf{0}}^{K} \binom{K}{J} D^{J+E_i, K-J} Q(\theta|\theta) = D^{K+E_i} L(\theta), \qquad i = 1, \ldots, d, \tag{4.6}$$

for any $K = (k_1, \ldots, k_d)$, where k_α's are non-negative integers, such that all the derivatives in (4.6) exist.

Proof: Under the stationary-point assumption, for any $1 \le i \le d$,

$$D^{E_i} L(\theta) = D^{E_i, \mathbf{0}} Q(\theta|\theta). \tag{4.7}$$

Applying the $D^K \equiv D^{K,\mathbf{0}}$ operator to both sides of (4.7) yields (4.6) via the chain rule

$$D^K \not{F}(\theta) = \sum_{J=\mathbf{0}}^{K} \binom{K}{J} D^{J, K-J} F(\theta, \theta) \tag{4.8}$$

for $\not{F}(\theta) \equiv F(\theta, \theta)$. □

An important consequence of Lemma 2 is that the Hessian matrix for $L(\theta)$ is directly available from the second order derivatives of the surrogate function because

$$D^2 L(\theta) = D^{20} Q(\theta|\theta) + D^{11} Q(\theta|\theta), \tag{4.9}$$

using the notation of Dempster et al. (1977) (e.g., $D^{20} = D^{(2,\ldots,2),(0,\ldots,0)}$). For the EM algorithm, this result was the core of Oakes (1999), where it was proved via the indirect route (4.1). The direct approach (4.7)–(4.8) shows that the specific "product form" inside the integrand in (4.1) is inconsequential once we have $D^{10} H(\theta|\theta) = 0$, because the chain rule (4.8) has the same form as the chain rule for differentiating the product of two functions. Indeed, for general yokes, the indirect approach is not even relevant (see Barndorff-Nielsen and Cox 1994, chap. 5).

When $Q(\theta|\phi)$ is an EMer, the set of identities given by (4.6) are equivalent to the set of Bartlett identities for $p(z|\theta)$, typically presented in more compact tensor notation (e.g., McCullagh 1987; Mykland 1994). The fact that these identities hold for any surrogate function (assuming differentiability) reinforces the authors' key message that, the *missing data* aspect of EM, though responsible for the enormous success of the current EM methodology, is actually not at the core of the algorithm. Algorithmically, the core is that $H(\theta|\theta^{(t)})$ achieves the maximum at $\theta = \theta^{(t)}$. However, the same fact is also indicative of the difficulty in finding necessary conditions that are unique to EMers, or to put it differently, it is not indicative of the conjecture that the SM class is more general than the EM class.

5. A BIG S!

Regardless of the (remote?) mathematical possibility that the SM class is the same as the EM class for most practical purposes, the SM formulation provides a new set of tools for finding simple and stable algorithms for complicated statistical optimization problems. To me, the biggest advantage of SM is that it bypasses the E step, and thus it provides a methodological breakthrough in dealing with the fundamental difficulty within the GAECM framework, namely, the difficulty with the E step when it is not in closed form. Although a Monte Carlo or numerical E step is possible and can be very effective (e.g., Wei and Tanner 1990; Meng and Schilling 1996; Booth and Hobert 1999), any GAECM is less appealing when its E step requires numerical computation or approximation. Replacing E by S signifies this breakthrough, and for that reason I am very pleased to attribute a big "S" to the authors for a new ray of Sunshine on the EM empire!

And I definitely see myself indulged in a few SM sessions, once I have my "EM flu" cured!

ACKNOWLEDGMENTS

I thank the editor, Andreas Buja, for the invitation; Zhiyi Chi, Per Mykland, and Peter McCullagh for very helpful conversations; and Jay Servidea and David van Dyk for comments and proofreading. The research was supported in part by NSF grant DMS-9626691 and NSA grant MDA 904-99-1-0067.

[Received December 1999.]

REFERENCES

Barndorff-Nielsen, O. (1987), "Differential Geometry and Statistics: Some Mathematical Aspects," *Indian Journal of Mathematics*, Ramanujian Centenary Volume, 29, 335–350.

Barndorff-Nielsen, O., and Cox, D. (1994), *Inference and Asymptotics*, London: Chapman and Hall.

Barndorff-Nielsen, O., and Jupp, J. (1997), "Yokes and Symplectic Structures," *Journal of Statistical Planning and Inference*, 63, 133–146.

Blæsild, P. (1991), "Yokes and Tensors Derived From Yokes," *Annals of The Institute of Statistical Mathematics*, 43, 95–113.

Booth, J. G., and Hobert, J. P. (1999), "Maximizing Generalized Linear Mixed Model Likelihoods With an Automated Monte Carlo EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 61, 265–285.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Series B, 39, 1–37.

Liu, C., and Rubin, D. B. (1994), "The ECME Algorithm: A Simple Extension of EM and ECM With Faster Monotone Convergence," *Biometrika*, 81, 633–648.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998), "Parameter Expansion for EM Acceleration—The PXEM Algorithm," *Biometrika*, 75, 755–770.

McCullagh, P. (1987), *Tensor Methods in Statistics*, London: Chapman and Hall.

Meng, X.-L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909.

——— (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278.

Meng, X.-L., and Schilling, S. (1996), "Fitting Full-Information Item Factor Models and an Empirical Investigation of Bridge Sampling," *Journal of the American Statistical Association*, 91, 1254–1267.

Meng, X.-L., and van Dyk, D. A. (1997), "The EM Algorithm—An Old Folk Song Sung to a Fast New Tune" (with discussion), *Journal of the Royal Statistical Society*, Series B, 59, 511–567.

——— (1999), "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation," *Biometrika*, 86, 301–320.

Mykland, P. (1994), "Bartlett Type Identities for Martingales," *The Annals of Statistics*, 22, 21–38.

Oakes, D. (1999), "Direct Calculation of the Information Matrix via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 61, 479–482.

Stigler, S. M. (1980), "Stigler's Law of Eponymy," in *Transactions of the New York Academy of Sciences*, Ser. 2 (Merton festschrift volume), ed. T. Gieryn, vol. 39, pp. 147–158.

van Dyk, D. A., Meng, X.-L., and Rubin, D. B. (1995), "Maximum Likelihood Estimation via the ECM Algorithm: Computing the Asymptotic Variance," *Journal of Computational and Graphical Statistics*, 5, 55–75.

Wei, G., and Tanner, M. A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithm," *Journal of the American Statistical Association*, 85, 699–704.