Statistical Science
2010, Vol. 25, No. 4, 533–544
DOI: 10.1214/10-STS348

Parameter Expansion and Efficient Inference

Andrew Lewandowski, Chuanhai Liu and Scott Vander Wiel

Abstract. This EM review article focuses on parameter expansion, a simple technique introduced in the PX-EM algorithm to make EM converge faster while maintaining its simplicity and stability. The primary objective concerns the connection between parameter expansion and efficient inference. It reviews the statistical interpretation of the PX-EM algorithm, in terms of efficient inference via bias reduction, and further unfolds the PX-EM mystery by looking at PX-EM from different perspectives. In addition, it briefly discusses potential applications of parameter expansion to statistical inference and the broader impact of statistical thinking on understanding and developing other iterative optimization algorithms.

Key words and phrases: EM algorithm, PX-EM algorithm, robit regression, nonidentifiability.

1. INTRODUCTION

The expectation maximization (EM) algorithm of Dempster, Laird and Rubin (1977) has proven to be a popular computational method for optimization. While simple to implement and stable in its convergence, the EM algorithm can converge slowly. Many variants of the original EM algorithm have also been proposed in the last 30+ years in order to overcome shortcomings that are sometimes seen in implementations of the original method. Among these EM-type algorithms are the expectation-conditional maximization (ECM) algorithm of Meng and Rubin (1993), the expectation-conditional maximization either (ECME) algorithm of Liu and Rubin (1994), the alternating ECM (AECM) algorithm of Meng and van Dyk (1997) and, more recently, the dynamic ECME (DECME) algorithm of He and Liu (2009). This review article focuses on parameter expansion as a way of improving the performance of the EM algorithm through a discussion of the parameter expansion EM (PX-EM) algorithm proposed by Liu, Rubin and Wu (1998).

The EM algorithm is an iterative algorithm for maximum likelihood (ML) estimation from incomplete data. Let X_{obs} be the observed data and let $f(X_{\text{obs}};\theta)$ denote the observed-data model with unknown parameter θ, where $X_{\text{obs}} \in \mathcal{X}_{\text{obs}}$ and $\theta \in \Theta$. Suppose that the observed-data model can be obtained from a complete-data model, denoted by $g(X_{\text{obs}}, X_{\text{mis}};\theta)$, where $X_{\text{obs}} \in \mathcal{X}_{\text{obs}}$, $X_{\text{mis}} \in \mathcal{X}_{\text{mis}}$, and $\theta \in \Theta$. That is,

$$f(X_{\text{obs}};\theta) = \int_{\mathcal{X}_{\text{mis}}} g(X_{\text{obs}}, X_{\text{mis}};\theta)\,dX_{\text{mis}}.$$

Given a starting point $\theta^{(0)} \in \Theta$, the EM algorithm iterates for $t = 0, 1, \ldots$ between the expectation (E) step and maximization (M) step:

E step. Compute the expected complete-data log-likelihood

$$Q\bigl(\theta|\theta^{(t)}\bigr) = \mathrm{E}\bigl(\ln g(X_{\text{obs}}, X_{\text{mis}};\theta)|X_{\text{obs}}, \theta = \theta^{(t)}\bigr) \tag{1.1}$$

as a function of $\theta \in \Theta$; and

M step. Maximize $Q(\theta|\theta^{(t)})$ to obtain

$$\theta^{(t+1)} = \arg\max_{\theta\in\Theta} Q\bigl(\theta|\theta^{(t)}\bigr). \tag{1.2}$$

Andrew Lewandowski is Ph.D. Student, Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, Indiana 47907, USA (e-mail: alewand@purdue.edu). Chuanhai Liu is Professor of Statistics, Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, Indiana 47907, USA (e-mail: chuanhai@purdue.edu; URL: www.stat.purdue.edu). Scott Vander Wiel is Technical Staff Member, Statistical Sciences Group, MS F600, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA (e-mail: scottv@lanl.gov; URL: www.stat.lanl.gov).

Two EM examples are given in Section 2.

Roughly speaking, the E step can be viewed as creating a complete-data problem by imputing missing values, and the M step can be understood as conducting a maximum likelihood-based analysis. More exactly, for complete-data models belonging to the exponential family, the E step imputes the complete-data sufficient statistics with their conditional expectations given the observed data and the current estimate $\theta^{(t)}$ of the parameter θ. This to some extent explains the simplicity of EM. The particular choice of (1.1) together with Jensen's inequality implies monotone convergence of EM.

The PX-EM algorithm is essentially an EM algorithm, but it performs inference on a larger full model. This model is obtained by introducing extra parameters into the complete-data model while preserving the observed-data sampling model. Section 3.1 presents the structure used in PX-EM. The theoretical results established by Liu, Rubin and Wu (1998) show that PX-EM converges no slower than its parent EM. This is somewhat surprising, as it is commonly believed that optimization algorithms generally converge slower as the number of dimensions increases. To help understand the behavior of PX-EM, Liu, Rubin and Wu (1998) provided a statistical interpretation of the PX-M step in terms of covariance adjustment. This is reviewed in Section 3.2 in terms of bias reduction using the example of binary regression with a Student-t link (see Mudholkar and George, 1978; Albert and Chib, 1993; Liu, 2004), which serves as a simple robust alternative to logistic regression and is called robit regression by Liu (2004).

To help further understand why PX-EM can work so well, several relevant issues are discussed in Section 4. Section 4.1 provides additional motivation behind why PX-EM can improve upon EM or ECM. In Section 4.2 we argue that parameter expansion can also be used for efficient data augmentation in the E step. The resulting EM is effectively the PX-EM algorithm.

In addition to the models discussed here, parameter expansion has now been shown to have computational advantages in applications such as factor analysis (Liu, Rubin and Wu, 1998) and the analysis of both linear (Gelman et al., 2008) and nonlinear (Lavielle and Meza, 2007) hierarchical models. However, Gelman (2004) shows that parameter expansion offers more than a computational method to accelerate EM. He points out that parameter expansion can be viewed as part of a larger perspective on iterative simulation (see Liu and Wu, 1999; Meng and van Dyk, 1999; van Dyk and Meng, 2001; Liu, 2003) and that it suggests a new family of prior distributions in a Bayesian framework discussed by Gelman (2006). One example is the folded noncentral Student-t distribution for between-group variance parameters in hierarchical models. This method exploits a parameter expansion technique commonly used in hierarchical models, and Gelman (2006) shows that it can be more robust than the more common inverse-gamma prior. Inspired by Gelman (2004), we briefly discuss other potential applications of parameter expansion to statistical inference in Section 5.

2. TWO EM EXAMPLES

2.1 The Running Example: A Simple Poisson–Binomial Mixed-Effects Model

Consider the complete-data model for the observed data $X_{\text{obs}} = X$ and the missing data $X_{\text{mis}} = Z$:

$$Z|\lambda \sim \text{Poisson}(\lambda)$$

and

$$X|(Z,\lambda) \sim \text{Binomial}(Z,\pi),$$

where $\pi \in (0,1)$ is known and $\lambda > 0$ is the unknown parameter to be estimated.

The observed-data model $f(X;\lambda)$ is obtained from the joint sampling model of (X,Z):

$$g(X,Z;\lambda) = \frac{\lambda^Z}{Z!}e^{-\lambda}\frac{Z!}{X!(Z-X)!}\pi^X(1-\pi)^{Z-X}, \tag{2.1}$$

where $X = 0,1,\ldots,Z$, $Z = 0,1,\ldots,$ and $\lambda \geq 0$. That is, $f(X;\lambda)$ is derived from $g(X,Z;\lambda)$ by integrating out the missing data Z as follows:

$$\begin{aligned} f(X;\lambda) &= \sum_{z=X}^{\infty}\frac{\lambda^z}{z!}e^{-\lambda}\frac{z!}{X!(z-X)!}\pi^X(1-\pi)^{z-X} \\ &= \frac{\lambda^X\pi^X}{X!}e^{-\lambda}\sum_{z=X}^{\infty}\frac{\lambda^{z-X}}{(z-X)!}(1-\pi)^{z-X} \\ &\overset{k=z-X}{=} \frac{(\lambda\pi)^X}{X!}e^{-\lambda}\sum_{k=0}^{\infty}\frac{[\lambda(1-\pi)]^k}{k!} \\ &= \frac{(\lambda\pi)^X}{X!}e^{-\lambda}e^{\lambda(1-\pi)} \\ &= \frac{(\lambda\pi)^X}{X!}e^{-\lambda\pi}. \end{aligned}$$

Alternatively, one can get the result from the well-known fact related to the infinite divisibility of the

Poisson distribution; namely, if $X_1 = X$ and $X_2 = Z - X$ are independent Poisson random variables with rate $\lambda_1 = \lambda\pi$ and $\lambda_2 = \lambda(1-\pi)$, then $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$ and conditional on $X_1 + X_2$, $X_1 \sim \text{Binomial}(X_1 + X_2, \lambda_1/(\lambda_1 + \lambda_2))$.

It follows that the observed-data model is $X|\lambda \sim \text{Poisson}(\pi\lambda)$. Thus, the ML estimate of λ has a closed-form solution, $\hat{\lambda} = X/\pi$. This artificial example serves two purposes. First, it is easy to illustrate the general EM derivation. Second, we use this example in Section 3.3 to show an extreme case in which PX-EM can converge dramatically faster than its parent EM; PX-EM converges in one-step, whereas EM converges painfully slowly.

The complete-data likelihood is given by the joint sampling model of (X, Z) found in equation (2.1). It follows that the complete-data model belongs to the exponential family with sufficient statistic Z for λ. The complete-data ML estimate of λ is given by

$$\hat{\lambda}_{\text{com}} = Z. \tag{2.2}$$

To derive the E step of EM, the conditional distribution of the missing data Z given both the observed data and the current estimate of the parameter λ is used. It is determined as follows:

$$\begin{aligned} h(Z|X,\lambda) &= \frac{g(X,Z;\lambda)}{\sum_{z=X}^{\infty} g(X,z;\lambda)} \\ &= \frac{[\lambda(1-\pi)]^{Z-X}/(Z-X)!}{\sum_{z=X}^{\infty}([\lambda(1-\pi)]^{z-X}/(z-X)!)} \\ &= \frac{[\lambda(1-\pi)]^{Z-X}}{(Z-X)!} e^{\lambda(1-\pi)}. \end{aligned}$$

Thus, $Z|\{X,\lambda\} \sim X + \text{Poisson}(\lambda(1-\pi))$. This yields

$$E(Z|X,\lambda) = X + \lambda(1-\pi).$$

Thus, the EM algorithm follows from the discussion of Dempster, Laird and Rubin (1977) on exponential complete-data models. Specifically, given the updated estimate $\lambda^{(t)}$ at the tth iteration, EM follows these two steps:

E step. Compute $\hat{Z} = \text{E}(Z|X, \lambda = \lambda^{(t)}) = X + \lambda^{(t)} \times (1-\pi)$.

M step. Replace Z in (2.2) with $\hat{Z}$ to obtain $\lambda^{(t+1)} = \hat{Z}$.

It is clear that the EM sequence $\{\lambda^{(t)} : t = 0, 1, \ldots\}$ is given by

$$\lambda^{(t+1)} = X + \lambda^{(t)}(1-\pi) \quad (t = 0, 1, \ldots) \tag{2.3}$$

converging to the ML estimate

$$\hat{\lambda} = X/\pi.$$

Rewrite (2.3) as

$$\lambda^{(t+1)} - \hat{\lambda} = (1-\pi)\big(\lambda^{(t)} - \hat{\lambda}\big)$$

to produce a closed-form expression for the convergence rate of EM:

$$\frac{|\lambda^{(t+1)} - \hat{\lambda}|}{|\lambda^{(t)} - \hat{\lambda}|} = 1 - \pi.$$

This indicates that EM can be very slow when $\pi \approx 0$.

2.2 ML Estimation of Robit Regression via EM

Consider the observed data consisting of n observations $X_{\text{obs}} = \{(x_i, y_i) : i = 1, \ldots, n\}$ with a p-dimensional covariate vector x_i and binary response y_i that takes on values of 0 and 1. The binary regression model with Student-t link assumes that, given the covariates, the binary responses y_i's are independent with the marginal probability distributions specified by

$$\begin{aligned} \Pr(y_i = 1|x_i, \beta) &= 1 - \Pr(y_i = 0|x_i, \beta) \\ &= F_\nu(x_i'\beta) \quad (i = 1, \ldots, n), \end{aligned} \tag{2.4}$$

where $F_\nu(\cdot)$ denotes the c.d.f. of the Student-t distribution with center zero, unit scale and ν degrees of freedom. With $\nu \approx 7$, this model provides a robust approximation to the popular logistic regression model for binary data analysis. Here we consider the case with known ν.

The observed-data likelihood

$$\begin{aligned} &f(X_{\text{obs}};\beta) \\ &\quad = \prod_{i=1}^{n} [F_\nu(x_i'\beta)]^{y_i} [1 - F_\nu(x_i'\beta)]^{1-y_i} \quad (\beta \in \mathbb{R}^p) \end{aligned}$$

involves the product of the c.d.f. of the Student-t distribution $F_\nu(\cdot)$ evaluated at $x_i'\beta$ for $i = 1, \ldots, n$. The MLE of β does not appear to have a closed-form solution. Here we consider the EM algorithm for finding the MLE of β.

A complete-data model for implementing EM to find the ML estimate of β is specified by introducing the missing data consisting of independent latent variables (τ_i, z_i) for each $i = 1, \ldots, n$ with

$$\tau_i|\beta \sim \text{Gamma}(\nu/2, \nu/2) \tag{2.5}$$

and

$$z_i|(\tau_i, \beta) \sim \text{N}(x_i'\beta, 1/\tau_i). \tag{2.6}$$

Let

$$y_i = \begin{cases} 1, & \text{if } z_i > 0, \\ 0, & \text{if } z_i \le 0 \end{cases} \quad (i = 1, \ldots, n). \tag{2.7}$$

Then the marginal distribution of y_i is preserved and is given by (2.4). The complete-data model belongs to the exponential family and has the following sufficient statistics for β:

$$S_{\tau xx'} = \sum_{i=1}^{n} \tau_i x_i x_i' \quad \text{and} \quad S_{\tau xz} = \sum_{i=1}^{n} \tau_i x_i z_i'. \tag{2.8}$$

The complete-data ML estimate of β is given by

$$\hat{\beta}_{\text{com}} = S_{\tau xx'}^{-1} S_{\tau xz}, \tag{2.9}$$

leading to the following EM algorithm.

Starting with $\beta^{(0)}$, say, $\beta^{(0)} = (0, \ldots, 0)$, EM iterates for $t = 0, 1, \ldots$ with iteration $t+1$ consisting of the following E and M steps:

E step. Compute $\hat{S}_{\tau xx'} = \mathrm{E}(S_{\tau xx'} | \beta = \beta^{(t)}, X_{\text{obs}})$ and $\hat{S}_{\tau xz} = \mathrm{E}(S_{\tau xz} | \beta = \beta^{(t)}, X_{\text{obs}})$.

M step. Update the estimate of β to obtain $\beta^{(t+1)} = \hat{S}_{\tau xx'}^{-1} \hat{S}_{\tau xz}$.

Let $f_\nu(\cdot)$ denote the p.d.f. of $F_\nu(\cdot)$. The E step can be coded by using the following results derived in Liu (2004):

$$\begin{aligned} \hat{\tau}_i &= \mathrm{E}\big(\tau_i | \beta = \beta^{(t)}, X_{\text{obs}}\big) \\ &= \frac{y_i - (2y_i - 1) F_{\nu+2}(-(1 + 2/\nu)^{1/2} x_i' \beta^{(t)})}{y_i - (2y_i - 1) F_\nu(-x_i' \beta^{(t)})}, \end{aligned} \tag{2.10}$$

$$\tau_i \hat{z}_i = \mathrm{E}\big(\tau_i z_i | \beta = \beta^{(t)}, X_{\text{obs}}\big) = \hat{\tau}_i \hat{z}_i, \tag{2.11}$$

where

$$\hat{z}_i \equiv x_i' \beta^{(t)} + \frac{(2y_i - 1) f_\nu(x_i' \beta^{(t)})}{y_i - (2y_i - 1) F_{\nu+2}(-(1 + 2/\nu)^{1/2} x_i' \beta^{(t)})}$$

for $i = 1, \ldots, n$.

However, the EM algorithm can also converge slowly in this example. This is discussed in Section 3.2, where it is shown that PX-EM can greatly improve the convergence rate.

3. THE PX-EM ALGORITHM

3.1 The Algorithm

Suppose that the EM complete-data model can be embedded in a larger model $g_*(X_{\text{obs}}, X_{\text{mis}}; \theta_*, \alpha)$ with the expanded parameter $(\theta_*, \alpha) \in \Theta \times \mathcal{A}$. Assume that the observed-data model is preserved in the sense that, for every $(\theta_*, \alpha) \in \Theta \times \mathcal{A}$,

$$f(X_{\text{obs}}; \theta) = f_*(X_{\text{obs}}; \theta_*, \alpha) \tag{3.1}$$

holds for some $\theta \in \Theta$, where $f_*(X_{\text{obs}}; \theta_*, \alpha) = \int_{\mathcal{X}_{\text{mis}}} g_*(X_{\text{obs}}, X_{\text{mis}}; \theta_*, \alpha)\, dX_{\text{mis}}$. The condition (3.1) defines a mapping $\theta = R(\theta_*, \alpha)$, called the reduction function, from the expanded parameter space $\Theta \times \mathcal{A}$ to the original parameter space Θ. For convenience, assume that the expanded parameters are represented in such a way that the original complete-data and observed-data models are recovered by fixing α at α_0. Formally, assume that there exists a *null* value of α, denoted by α_0, such that $\theta = R(\theta, \alpha_0)$ for every $\theta \in \Theta$. When applied to the parameter-expanded complete-data model $g_*(X_{\text{obs}}, X_{\text{mis}}; \theta_*, \alpha)$, the EM algorithm, called the PX-EM algorithm, creates a sequence $\{(\theta_*^{(t)}, \alpha^{(t)})\}$ in $\Theta \times \mathcal{A}$. In the original parameter space Θ, PX-EM generates a sequence $\{\theta^{(t)} = R(\theta_*^{(t)}, \alpha^{(t)})\}$ and converges no slower than the corresponding EM based on $g(X_{\text{obs}}, X_{\text{mis}}; \theta)$; see Liu, Rubin and Wu (1998).

For simplicity and stability, Liu, Rubin and Wu (1998) use $(\theta^{(t)}, \alpha_0)$ instead of $(\theta_*^{(t)}, \alpha^{(t)})$ for the E step. As a result, PX-EM shares with EM its E step and modifies its M step by mapping $(\theta_*^{(t+1)}, \alpha^{(t+1)})$ to the original space $\theta^{(t+1)} = R(\theta_*^{(t+1)}, \alpha^{(t+1)})$. More precisely, the PX-EM algorithm is defined by replacing the E and M steps of EM with the following:

PX-E step. Compute

$$\begin{aligned} &Q(\theta_*, \alpha | \theta^{(t)}, \alpha_0) \\ &\quad = \mathrm{E}\big(\ln g_*(X_{\text{obs}}, X_{\text{mis}}; \theta_*, \alpha) | X_{\text{obs}}, \theta_* = \theta^{(t)}, \\ &\qquad\qquad \alpha = \alpha_0\big) \end{aligned}$$

as a function of $(\theta_*, \alpha) \in \Theta \times \mathcal{A}$.

PX-M step. Find

$$(\theta_*^{(t+1)}, \alpha^{(t+1)}) = \arg\max_{\theta_*, \alpha} Q(\theta_*, \alpha | \theta^{(t)}, \alpha_0)$$

and update

$$\theta^{(t+1)} = R(\theta_*^{(t+1)}, \alpha^{(t+1)}).$$

Since it is the ordinary EM applied to the parameter expanded complete-data model, PX-EM shares with EM its simplicity and stability. Liu, Rubin and Wu (1998) established theoretical results to show that PX-EM can converge no slower than EM. Section 3.2 uses the robit regression example to give the statistical interpretation of Liu, Rubin and Wu (1998) in terms of covariance adjustment. With the toy example, Section 3.3

demonstrates that PX-EM can be dramatically faster than its parent EM. A discussion of why PX-EM can perform better than EM is given in Section 4.

3.2 Efficient Analysis of Imputed Missing Data: Robit Regression

The E step of EM imputes the sufficient statistics $S_{\tau xx'}$ and $S_{\tau xz}$ with their expectations based on the predictive distribution of the missing (τ_i, z_i) data conditioned on the observed data X_{obs} and $\beta^{(t)}$, the current estimate of β at the tth iteration. Had the ML estimate of β, $\hat{\beta}$, been used to specify the predictive distribution, EM would have converged on the following M step, which in this case performs correct ML inference. We call the predictive distribution using $\hat{\beta}$ the correct imputation model. Before convergence, *that is*, $\beta^{(t)} \neq \hat{\beta}$, the E step imputes the sufficient statistics $S_{\tau xx'}$ and $S_{\tau xz}$ using an incorrect imputation model. The M step also uses a wrong model since it does not take into account that the data were incorrectly imputed based on an assumed parameter value $\beta^{(t)} \neq \hat{\beta}$. The M step moves the estimate $\beta^{(t+1)}$ toward $\hat{\beta}$, but the difference between $\beta^{(t+1)}$ and $\hat{\beta}$ can be regarded as bias due to the use of the $\beta^{(t)}$.

The bias induced by the E step can be reduced by making use of recognizable discrepancies between imputed statistics and their values under the correct imputation model. To capture such discrepancies, Liu, Rubin and Wu (1998) considered parameters that are statistically identified in the complete-data model but not in the observed-data model. These parameters are fixed at their default values to render the observed-data model identifiable. In the context of EM for robit regression, these parameters are the scale parameters of τ_i and z_i, denoted by α and σ. In the observed-data model, they take the default values $\alpha_0 = 1$ and $\sigma_0 = 1$.

When activated, the extra parameters are estimated by the M step and these estimates converge to the default values to produce ML parameter estimates for the observed data model. Thus, in the robit regression model, we identify the default values of the extra parameters as MLEs: $\alpha_0 = \hat{\alpha} = 1$ and $\sigma_0 = \hat{\sigma} = 1$. Denote the corresponding EM estimates by $\alpha^{(t+1)}$ and $\sigma^{(t+1)}$. The discrepancies between $(\alpha^{(t+1)}, \sigma^{(t+1)})$ and $(\hat{\alpha}, \hat{\sigma})$ reveal the existence of bias induced by the wrong imputation model. These discrepancies can be used to adjust the estimate of the parameter of interest, β, at each iteration. This is exactly what PX-EM is formulated to do, and the resulting algorithm converges faster than the original EM.

Formally, the extra parameter (α, σ) introduced to capture the bias in the imputed values of τ_i and z_i is called the *expansion parameter*. The complete-data model is thus both data-augmented as well as parameter-augmented. For correct inference at convergence, data augmentation is required to preserve the observed-data model after integrating out missing data. Likewise, parameter expansion needs to satisfy the observed-data model preservation condition (3.1). In the robit regression model, let $(\beta_*, \alpha, \sigma)$ be the expanded parameter with β_* playing the same role as β in the original model. The preservation condition states that for every expanded parameter value $(\beta_*, \alpha, \sigma)$, there exists a value of β such that the sampling model of the y_i's obtained from the parameter expanded model is the same as the original sampling model given β. This condition defines a mapping $\beta = R(\beta_*, \alpha, \sigma)$, the reduction function. This reduction function is used in PX-EM to adjust the value of $\beta^{(t+1)}$ produced by the M step.

The detailed implementation of PX-EM for robit regression is as follows. The parameter-expanded complete-data model is obtained by replacing (2.5) and (2.6) with

$$(\tau_i/\alpha)|(\beta_*, \alpha, \sigma) \sim \text{Gamma}(\nu/2, \nu/2) \tag{3.2}$$

and

$$z_i|(\tau_i, \beta_*, \alpha, \sigma) \sim \text{N}(x_i'\beta_*, \sigma^2/\tau_i) \tag{3.3}$$

for $i = 1, \ldots, n$. Routine algebraic operation yields the reduction function

$$\begin{aligned} \beta &= R(\beta_*, \alpha, \sigma) \\ &= (\alpha^{1/2}/\sigma)\beta_* \quad (\beta_* \in \mathcal{R}^p; \alpha > 0; \sigma > 0). \end{aligned} \tag{3.4}$$

The expanded parameterization in (3.2) and (3.3) is a natural choice if the missing data are viewed as real and a parameterization is sought that provides a model that is flexible while preserving the observed data model and allowing the original parameterization to be recovered through the reduction function. For example, if τ_i is treated as fixed, the model for z_i is a regression model with fixed variance. Adding σ^2 and α allows the variance of z_i and the scale of τ_i to be estimated freely in the expanded model.

The sufficient statistics for the expanded parameter $(\beta_*, \alpha, \sigma)$ now become

$$\begin{aligned} S_\tau &= \sum_{i=1}^n \tau_i, \qquad S_{\tau xx'} = \sum_{i=1}^n \tau_i x_i x_i', \\ S_{\tau z^2} &= \sum_{i=1}^n \tau_i z_i^2, \qquad S_{\tau xz} = \sum_{i=1}^n \tau_i x_i z_i'. \end{aligned} \tag{3.5}$$

The complete-data ML estimate of β_* is the same as that of β in the original complete-data model. The complete-data ML estimates of α and σ are given by

$$\begin{aligned}\hat{\alpha}_{\text{com}} &= \frac{1}{n}S_\tau \quad \text{and} \\ \hat{\sigma}^2_{\text{com}} &= \frac{1}{n}(S_{\tau z^2} - S_{\tau xz}S^{-1}_{\tau xx'}S_{\tau xz}).\end{aligned} \tag{3.6}$$

The PX-EM algorithm is simply an EM applied to the parameter expanded complete-data model with an M step followed by (or modified to contain) a reduction step. The reduction step uses the reduction function to map the estimate in the expanded parameter space to the original parameter space. For the robit example, PX-EM is obtained by modifying the E and M steps as follows.

PX-E step. This is the same as the E step of EM except for the evaluation of two additional expected sufficient statistics:

$$\hat{S}_\tau = \mathrm{E}\big(S_\tau|\beta = \beta^{(t)}, X_{\text{obs}}\big) = \sum_{i=1}^{n}\hat{\tau}_i$$

and

$$\begin{aligned}\hat{S}_{\tau z^2} &= \mathrm{E}\big(S_{\tau z^2}|\beta = \beta^{(t)}, X_{\text{obs}}\big) \\ &= n(\nu+1) \\ &\quad - \nu\sum_{i=1}^{n}\hat{\tau}_i + \sum_{i=1}^{n}\hat{\tau}_i x_i'\beta^{(t)}\big(2\hat{z}_i - x_i'\beta^{(t)}\big),\end{aligned}$$

where $\hat{\tau}_i$'s and $\hat{z}_i$'s are available from the E step of EM.

PX-M step. Compute $\hat{\beta}_* = \hat{S}^{-1}_{\tau xx'}\hat{S}_{\tau xz}$, $\hat{\sigma}^2_* = n^{-1}\times(\hat{S}_{\tau z^2} - \hat{S}_{\tau xz}\hat{S}^{-1}_{\tau xx'}\hat{S}_{\tau xz})$, and $\hat{\alpha}_* = n^{-1}\hat{S}_\tau$ and then use the reduction to obtain $\hat{\beta}^{(t+1)} = (\hat{\alpha}_*^{1/2}/\hat{\sigma}_*)\hat{\beta}_*$.

For a numerical example, consider the data of Finney (1947), which consist of 39 binary responses denoting the presence ($y = 1$) or absence ($y = 0$) of vasoconstriction of the skin of the subjects after inspiration of a volume V of air at inspiration rate R. The data were obtained from repeated measurements on three individual subjects, the numbers of observations per subject being 9, 8 and 22. Finney (1947) found no evidence of inter-subject variability, treated the data as 39 independent observations, and analyzed the data using the probit regression model with V and R in the logarithm scale as covariates. This data set was also analyzed by Liu (2004) to illustrate robit regression. Due to three outlying observations, the MLE of the degrees of freedom ν is very small, $\hat{\nu} = 0.11$.

Here we use this data set with $\ln(V)$ and $\ln(R)$ as the covariates and take the fixed $\nu = 2$ as a numerical example to compare EM and PX-EM. Numerical results comparing the rates of convergence of EM and PX-EM are displayed in Figure 1. PX-EM shows a clear and dramatic convergence gain over EM. For convenience we choose to report the detailed results over iterations. The algorithms were coded in R, which makes CPU comparison unreliable. Since extra computation for the PX-EM implementation is minor, we believe the same conclusion holds in terms of CPU times.

3.3 PX-EM with Fast Convergence: The Toy Example

The model $X|(Z,\lambda) \sim \text{Binomial}(Z,\pi)$ may not fit the imputed value of missing data Z very well in the sense that $X/\hat{Z}$ is quite different from π. This mismatch can be used to adjust $\lambda^{(t+1)}$. To adjust $\lambda^{(t+1)}$, we activate π and let α denote the activated parameter with $\alpha_0 = \pi$. Now the parameter-expanded complete-data model becomes

$$Z|(\lambda_*,\alpha) \sim \text{Poisson}(\lambda_*)$$

and

$$X|(Z,\lambda_*,\alpha) \sim \text{Binomial}(Z,\alpha),$$

where $\lambda_* > 0$ and $\alpha \in (0,1)$. If the missing data were treated as being observed, this model allows the mean parameters for both X and Z to be estimated. The two corresponding observed-data models are Poisson($\lambda\pi$) and Poisson($\lambda_*\alpha$), giving the reduction function

$$\lambda = R(\lambda_*,\alpha) = \frac{\alpha}{\pi}\lambda_*. \tag{3.7}$$

The complete-data sufficient statistics are Z and X. The complete-data ML estimates of λ_* and α are given by

$$\hat{\lambda}_{*,\text{com}} = Z \quad \text{and} \quad \hat{\alpha}_{\text{com}} = \frac{X}{Z}. \tag{3.8}$$

The resulting PX-EM has the following E and M steps:

PX-E step. This is the same as the E step of EM.

PX-M step. Replace Z in (3.8) with $\hat{Z}$ to obtain $\lambda_*^{(t+1)} = \hat{Z}$ and $\alpha^{(t+1)} = X/\hat{Z}$. Update λ using the reduction function and obtain

$$\lambda^{(t+1)} = \frac{X}{\pi\hat{Z}}\hat{Z} = \frac{X}{\pi}.$$

The PX-EM algorithm in this case converges in one step. Although artificial, this toy example shows again that PX-EM can converge dramatically faster than its parent EM.

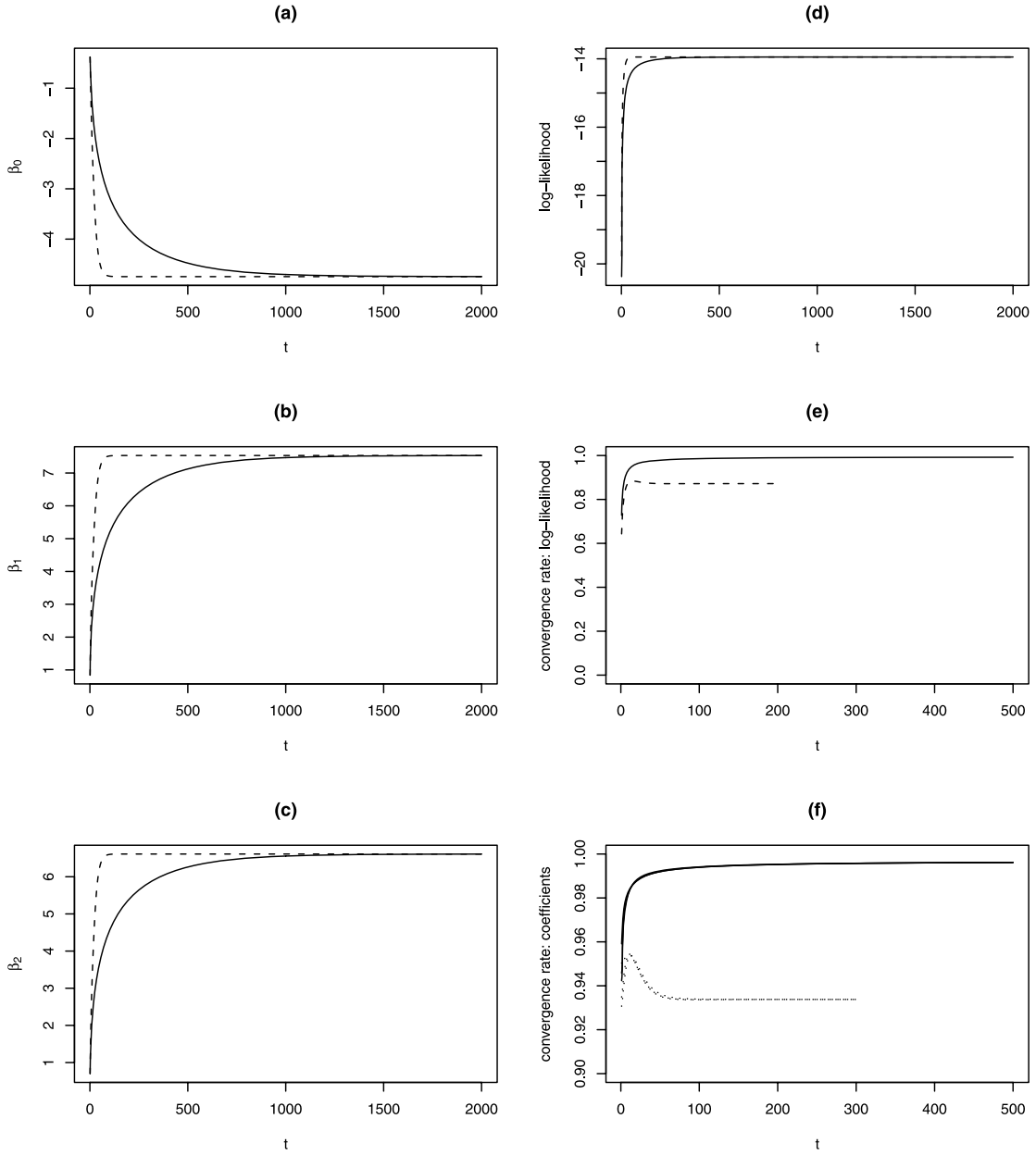

FIG. 1. *EM* (*solid*) *and PX-EM* (*dashed*) *sequences of the regression coefficients* β_0 (a), β_1 (b), β_2 (c), *and log-likelihood in the robit regression with* $x = (1, \ln(V), \ln(R))$. *The rates of convergence of EM* (*solid*) *and PX-EM* (*dashed*) *are shown in* (e) *by* $|\ell^{(t+1)} - \ell^{(\infty)}|/|\ell^{(t)} - \ell^{(\infty)}|$, *where* $\ell^{(t)}$ *denotes the log-likelihood value at the* t*th iteration, and in* (f) *by* $|\beta_j^{(t+1)} - \beta_j^{(\infty)}|/|\beta_j^{(t)} - \beta_j^{(\infty)}|$ *for* $j = 0, 1$ *and* 2.

4. UNFOLDING THE MYSTERY OF PX-EM

The statistical interpretation in terms of covariance adjustment, explained by the robit example above and the Student-t example in Liu, Rubin and Wu (1998), and the theoretical results of Liu, Rubin and Wu (1998) help reveal the PX-EM magic. To further unfold the mystery of PX-EM, we discuss the nonidentifiability of expanded parameters in the observed-data model in Section 4.1 and take a look at PX-EM from the point of view of efficient data augmentation in Section 4.2.

4.1 Nonidentifiability of Expanded Parameters and Applicability of PX-EM

It is often the case in PX-EM that, even though the expanded parameter (θ_*, α) is identifiable from $Q(\theta_*, \alpha|\theta^{(t)}, \alpha_0)$ (the expected parameter-expanded

complete-data log-likelihood), it is not identifiable from the corresponding observed-data loglikelihood

$$L_*(\theta_*, \alpha) = \ln f_*(X_{\text{obs}}; \theta_*, \alpha).$$

It is helpful to consider $L_*(\theta_*, \alpha)$ for understanding PX-EM, as the PX-M step directly increases $L_*(\theta_*, \alpha)$ through maximizing $Q(\theta_*, \alpha|\theta^{(t)}, \alpha_0)$. Naturally, from a mathematical point of view, unfolding the actual likelihood in the larger or expanded parameter space $\Theta \times \mathcal{A}$ shows how PX-EM steps can lead to increases in the likelihood function faster than can moves in the original space Θ.

4.1.1 *The observed-data log-likelihood surface over* $\Theta \times \mathcal{A}$. The observed-data log-likelihood, as a function of (θ_*, α), is determined by the actual log-likelihood $L(\theta) = \ln f(X_{\text{obs}}; \theta)$ with θ replaced by $\theta = R(\theta_*, \alpha)$ so that

$$L_*(\theta_*, \alpha) = L(R(\theta_*, \alpha)) \quad \big((\theta_*, \alpha) \in \Theta \times \mathcal{A}\big). \tag{4.1}$$

Thus, each point $\theta \in \Theta$ corresponds to a subspace $\{(\theta_*, \alpha) \in \Theta \times \mathcal{A}, R(\theta_*, \alpha) = \theta\}$, over which $L_*(\theta_*, \alpha)$ is constant.

For example, when θ and α are one-dimensional parameters, $L(\theta)$ can be represented by a curve in the two-dimensional space $\Theta \times L(\Theta)$, whereas $L_*(\theta_*, \alpha)$ is a family of curves indexed by α. The family of curves $L_*(\theta_*, \alpha)$ form a surface in the style of a mountain range in the three-dimensional space $\Theta \times \mathcal{A} \times L(\Theta)$. For the toy example, this is depicted in Figure 2 by the top panel 3-D perspective plot and in Figure 3 by the image with dashed contours or "elevation" lines. The mode of $L(\theta)$ now becomes a set of modes of the same "altitude," one for each fixed α. That is, the mode of $L(\theta)$ is expanded into the "ridge" shown, for example, by the thick line in Figure 3.

4.1.2 *Likelihood maximization in PX-EM*. The E step in PX-EM implicitly computes a family of expected complete-data log-likelihood functions, which are the Q-functions used in (1.1), over the original parameter space indexed by the expansion parameter α. This is because PX-EM introduces no additional or different missing data in the larger complete-data model. In other words, the parent E step effectively computes a surface over $\Theta \times \mathcal{A}$ that can be used as a Q-function to approximate the expanded loglikelihood $L_*(\theta_*, \alpha)$ defined in (4.1). This Q-function for the toy example is shown in Figure 2 by the bottom panel 3-D perspective plot and in Figure 3 by the nearly-elliptical contours. For this one-step convergence PX-EM example,

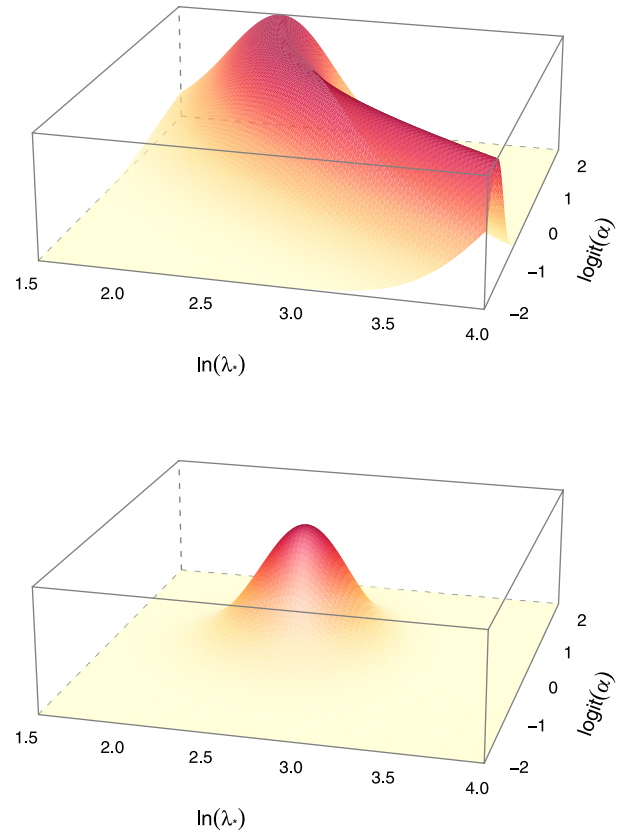

FIG. 2. *Perspective plots of the parameter-expanded observed-data log-likelihood* $L(\lambda_*, \alpha)$ (top) *and the parameter-expanded complete-data log-likelihood* $Q(\lambda_*, \alpha|\lambda^{(t)})$ (bottom) *in the toy example with* $X = 8$, $\pi = 0.25$, *and* $\lambda^{(t)} = 8$.

the mode of this Q-function is on the ridge of the expanded loglikelihood $L_*(\theta_*, \alpha)$. We note that this is typically not the case in more realistic examples. In the general case, the mode of the Q-function would typically be located on one elevation line that is higher than the elevation line where the update $(\theta^{(t)}, \alpha_0)$ found by EM is located.

Somewhat surprisingly, any such Q-function for each fixed α is EM-valid. By *EM-valid*, we mean that increasing the Q-function results in an increase of the actual likelihood in the expanded space and thereby in the original space after the reduction step. This is due to two facts: (i) the joint Q-function is EM-valid for $L_*(\theta_*, \alpha)$ and, thus, for $L(\theta)$ as well, and (ii) an M step with any fixed α, which finds

$$\theta_*^{(t+1)} = \arg\max_{\theta_*} Q\big(\theta_*, \alpha|\theta^{(t)}, \alpha_0\big),$$

followed by the reduction $\theta^{(t+1)} = R(\theta_*^{(t+1)}, \alpha)$ is simply a conditional maximization step. Additionally, in the context of the ECM algorithm of Meng and Rubin (1993), the parent EM is an incomplete ECM with

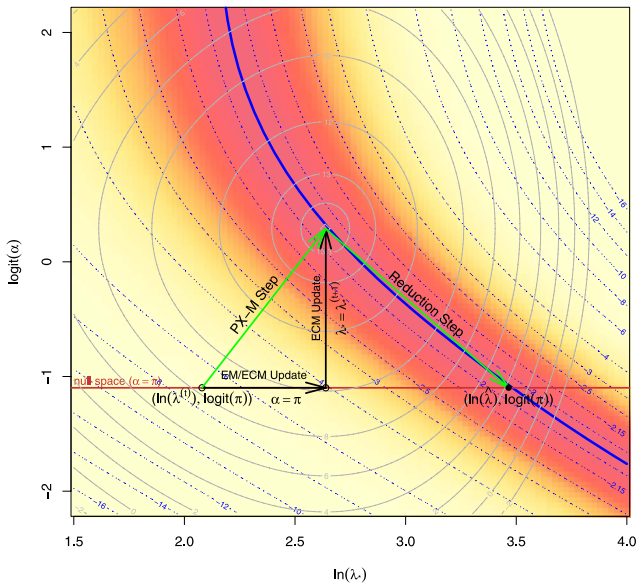

FIG. 3. *PX-EM for the toy example with* $X = 8$, $\pi = 0.25$, *and* $\lambda^{(t)} = 8$. *The parameter-expanded observed-data log-likelihood function* $L(\lambda_*, \alpha)$ *is shown by shading and dashed contours with a maximum along the ridge indicated by a solid thick line. The expected parameter-expanded complete-data log-likelihood* $Q(\lambda_*, \alpha|\lambda^{(t)})$ *is shown by the ellipse-like solid contours. In this example, maximization of* $Q(\lambda_*, \alpha|\lambda^{(t)})$ *over* (λ_*, α) *can be obtained in two conditional maximization steps, labeled as the two ECM updates. The PX-M step moves to a point on the ridge of* $L(\lambda_*, \alpha)$, *and the subsequent reduction-step moves this point along the the ridge of* $L(\lambda_*, \alpha)$ *to the point with* $\alpha = \pi$.

only one single CM step over $\Theta \times \mathcal{A}$. This relationship is explored in greater detail in the next section.

4.1.3 *PX-EM vs. (efficient) ECM over* $\Theta \times \mathcal{A}$. In theory, PX-EM has a single M step over the entire space $\Theta \times \mathcal{A}$. Note that

$$\max_{(\theta_*,\alpha)} Q(\theta_*, \alpha|\theta^{(t)}, \alpha_0) = \max_{\alpha} \max_{\theta_*} Q(\theta_*, \alpha|\theta^{(t)}, \alpha_0).$$

When

$$\hat{\theta}_*^{(t+1)} = \arg\max_{\theta_*} Q(\theta_*, \alpha|\theta^{(t)}, \alpha_0)$$

does not depends on α, as is often the case in many PX-EM examples, the PX-M step is equivalent to a cycle of two CM steps: one is the M step of EM, and the other updates α with θ_* fixed at $\theta_*^{(t+1)}$. This version of ECM for the toy example is illustrated in Figure 3. In this case, ECM is efficient for it generates the PX-EM update.

To summarize, denote by $\text{ECM}_{\{\alpha,\theta_*\}}$ the above version of ECM over $\Theta \times \mathcal{A}$. Typically, the algorithms can then be ordered in terms of performance as

$$\text{EM} \preceq \text{ECM}_{\{\alpha,\theta_*\}} \preceq \text{PX} - \text{EM} \tag{4.2}$$

over $\Theta \times \mathcal{A}$. It should be noted that by *typically*, we mean the conclusion is reached in an analogy with comparing the EM algorithm and the Generalized EM algorithm (GEM) (Dempster, Laird and Rubin, 1977), that is, EM typically converges faster than GEM, but counter examples exist; see, *for example*, Section 5.4 of van Dyk and Meng (2010) and the alternative explanation from an ECME point of view in Section 4.3 of Liu and Rubin (1998) on why ECM can be faster than EM. To elaborate it further with our robit example, it may be also interesting to note that when the reduction function (3.4) is modified by replacing the adjustment factor $(\alpha^{1/2}/\sigma)$ with (α/σ), a typo made in the earlier versions of the PX-EM for the robit regression model, the resulting (wrong) PX-EM converges actually faster than the (correct) PX-EM for the numerical example in Section 3.2. In general, more efficiency can be gained by replacing the CM step of ECM over α with a CM step maximizing the corresponding actual constrained likelihood in the parameter expanded space. This is effectively a parameter-expanded ECME algorithm; see such an example for the Student-t distribution given in Liu (1997). More discussion on ECME and other state-of-the-art methods for accelerating the EM algorithm can be found in He and Liu (2009). Their discussion on the method termed SOR provides a relevant explanation why the above wrong PX-EM and other wrong PX-EM versions, such as the one using the wrong reduction function $\beta = (\alpha/\sigma^2)\beta_*$ in the numerical robit example, can converge faster than the correct PX-EM.

Perhaps most importantly, the above discussion further explains why PX-EM can perform better than EM can, and unfolds the mystery of PX-EM, in addition to the covariance adjustment interpretation.

4.2 Efficient Data Augmentation via Parameter Expansion

Meng and van Dyk (1997) consider efficient data augmentation for creating fast converging algorithms. They search for efficient augmenting schemes by working with the fraction of missing-data information. Here we show that PX-EM can also be viewed as an alternative way of doing efficient data augmentation. Unlike Meng and van Dyk (1997), who find a fixed augmenting scheme that works for all EM iterations, the following procedure is a way to choose an adaptive augmenting scheme for each EM iteration. Rather than control the fraction of missing-data information, this procedure reduces bias through the expansion parameter. For the sake of clarity, we use the artificial example of Section 2.1 to make our argument.

Consider the parameter-expanded complete-data likelihood obtained from (2.1) by activating $\alpha_0 = \pi$, *that is*,

$$\frac{\lambda_*^Z}{Z!}e^{-\lambda_*}\frac{Z!}{X!(Z-X)!}\alpha^X(1-\alpha)^{Z-X} \qquad (\lambda_* > 0; 0 < \alpha < 1),$$

which has the canonical representation

$$h(X,Z)c(\lambda_*,\alpha)e^{Z\ln[\lambda_*(1-\alpha)]+X\ln\alpha/(1-\alpha)} \qquad (\lambda_* > 0; 0 < \alpha < 1).$$

Thus, when fixed at the given value, π, for identifiability, the complete-data ML estimate $\hat{\alpha} = X/Z$ plays the role of an ancillary statistic; see Ghosh, Reid and Fraser (2010) for an introduction to ancillary statistics. With the correct imputation model, or at convergence, the imputed value $\hat{Z}$ satisfies

$$\pi = \frac{X}{\hat{Z}} \quad \text{or} \quad \hat{Z} = \frac{X}{\pi}. \tag{4.3}$$

Thus, we can consider modifying the E step of EM to produce an imputed statistic $\hat{Z}$ that satisfies (4.3).

In the context of PX-EM, the current estimate $\lambda^{(t)}$ corresponds to the following subset of the expanded parameter space:

$$\begin{aligned}\Omega_*^{(t)} &\equiv \{(\lambda_*,\alpha) : R(\lambda_*,\alpha) = R(\lambda^{(t)},\alpha_0)\} \\ &= \{(\lambda_*,\alpha) : \lambda^{(t)}\pi = \alpha\lambda_*\}.\end{aligned} \tag{4.4}$$

Thus, we can use the imputation model defined by the parameter-expanded complete-data model conditioned on an arbitrary point $(\tilde{\lambda}_*,\tilde{\alpha}) \in \Omega_*^{(t)}$. For efficient data augmentation, we choose a particular point $(\tilde{\lambda}_*,\tilde{\alpha}) \in \Omega_*^{(t)}$, if it exists, so that (4.3) holds. Since

$$\hat{Z} = \mathrm{E}(Z|X,\lambda_*,\alpha) = X + \lambda_*(1-\alpha),$$

to obtain the desired imputation model, we solve

$$\begin{aligned}X + \tilde{\lambda}_*(1-\tilde{\alpha}) &= \frac{X}{\pi}, \\ \lambda^{(t)}\pi &= \tilde{\alpha}\tilde{\lambda}_*\end{aligned}$$

for $(\tilde{\lambda}_*,\tilde{\alpha})$. This system of equations has the solution

$$\tilde{\lambda}_* = X\frac{1-\pi}{\pi} + \lambda^{(t)}\pi$$

and

$$\tilde{\alpha} = \frac{\lambda^{(t)}\pi}{X((1-\pi)/\pi) + \lambda^{(t)}\pi}.$$

The E step of the EM algorithm based on the corresponding imputation model produces $\hat{Z} = X/\pi$. The following M step of EM gives $\lambda^{(t+1)} = \hat{Z} = X/\pi$.

The resulting EM algorithm is effectively the PX-EM algorithm. This implies that PX-EM can be understood from the perspective of efficient data augmentation via parameter expansion. Similar arguments can be made for other PX-EM examples having imputed ancillary statistics. In the general case, such an efficient data augmentation amounts to modifying imputed complete-data sufficient statistics and can be viewed as re-imputation of missing sufficient statistics.

5. DISCUSSION

Gelman (2004) notes that "progress in statistical computation often leads to advances in statistical modeling," which opens our eyes to the broader picture. Statistical interpretations of EM and PX-EM reveal that statistical thinking can aid in understanding and developing iterative algorithms. It seems natural to apply fundamental concepts from statistical inference to address statistical problems such as ML estimation and Bayesian estimation (see, e.g., Liu and Wu, 1999; van Dyk and Meng, 2001; Qi and Jaakkola, 2007; Hobert and Marchev, 2008). A recent example is the work of Yu and Meng (2008, 2010), which uses relationships motivated by the concepts of ancillarity and sufficiency in order to find optimal parameterizations for data augmentation algorithms used in Bayesian inference. However, statistical thinking can also be helpful for general-purpose optimization algorithms such as in the improvements to the quasi-Newton algorithm developed by Liu and Vander Wiel (2007).

Thinking outside the box, here we briefly discuss other potential applications of parameter expansion to statistical inference. The fundamental idea of PX-EM—the use of expanded parameters to capture information in data—leads immediately to a possible application of parameter expansion for "dimension-matching" in Fisher's conditional inference and fiducial inference (see, e.g., Fisher, 1973), where difficulties arise when the dimensionality of the minimal sufficient statistics is larger than the number of free parameters to be inferred. It is well known that, while attempting to build a solid foundation for statistical inference, the ideas behind Fisher's fiducial inference have not been well developed. Nevertheless, it is expected that parameter expansion can be useful in developing new ideas for statistical inference. For example, a Dempster–Shafer or fiducial-like method, called

the inferential model (IM) framework, has been proposed by Zhang and Liu (2011) and Martin, Zhang and Liu (2010). Of particular interest is the parameter expansion technique proposed by Martin, Hwang and Liu (2010) for what they call weak marginal inference using IMs. Using this parameter expansion technique, they provide satisfactory resolutions to the famous Stein's paradox and the Behrens–Fisher problem.

Although brief, the above discussion shows that parameter expansion has the potential to contribute to a variety of applications in computation and statistical inference. To conclude this review article, we speculate on one possible application of parameter expansion to the method of maximum likelihood for which the EM algorithm has proven to be a useful computational tool. The prospect of applying general statistical ideas to computational problems has also led us to thinking about model checking or goodness of fit to solve the unbounded likelihood problem in fitting Student-t and mixture models, for which EM is often the first choice. In the case with unbounded likelihood functions, for example, a high-likelihood model may not fit the observed data well and then inferential results can be nonsensical. It would be interesting to see if the general idea of parameter expansion for efficient inference can be extended for "valid inference" as well. However, it is not our intention here to discuss these open problems in depth. Based on the past success in this area, it can be expected that parameter expansion methods will continue to aid computation and inference.

ACKNOWLEDGMENTS

The authors thank the editors and their reviewers for their helpful comments and suggestions on earlier versions of this article. Chuanhai Liu was partially supported by NSF Grant DMS-10-07678.

REFERENCES

ALBERT, J. H. and CHIB. S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

FISHER, R. A. (1973). *Statistical Methods for Scientific Induction*, 3rd ed. Hafner, New York. MR0346954

FINNEY, D. J. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika* **34** 320–334.

GELMAN, A. (2004). Parametrization and Bayesian modeling. *J. Amer. Statist. Assoc.* **99** 537–545. MR2109315

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.* **1** 515–533. MR2221284

GELMAN, A., VAN DYK, A. A., HUANG, Z. and BOSCARDIN, J. W. (2008). Using redundant parameters to fit hierarchical models. *J. Comput. Graph. Statist.* **17** 95–122. MR2424797

GHOSH, M., REID, N. and FRASER, D. A. S. (2010). Ancillary statistics: A review. *Statist. Sinica* **20** 1309–1332.

HE, Y. and LIU, C. (2009). The dynamic ECME algorithm. Technical report, Dept. Statistics, Purdue Univ.

HOBERT, J. P. and MARCHEV, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *Ann. Statist.* **36** 532–554. MR2396806

LAVIELLE, M. and MEZA, C. (2007). A parameter expansion version of the SAEM algorithm. *Statist. Comput.* **17** 121–130. MR2380641

LIU, C. (1997). ML estimation of the multivariate t distribution and the EM algorithm. *J. Multivariate Anal.* **63** 296–312. MR1484317

LIU, C. (2003). Alternating subspace-spanning resampling to accelerate Markov chain Monte Carlo simulation. *J. Amer. Statist. Assoc.* **98** 110–117. MR1965678

LIU, C. (2004). Robit regression: A simple robust alternative to logistic and probit. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (A. Gelman and X. L. Meng, eds.) 227–238. Wiley, London. MR2138259

LIU, C. and RUBIN, D. B. (1994). The ECME algorithm: An simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81** 633–648. MR1326414

LIU, C. and RUBIN, D. B. (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika* **85** 673–688. MR1665830

LIU, C., RUBIN, D. B. and WU, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85** 755–770. MR1666758

LIU, C. and VANDER WIEL, S. A. (2007). Statistical quasi-Newton: A new look at least change. *SIAM J. Optim.* **18** 1266–1285. MR2373301

LIU, J. S. and WU, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.* **94** 1264–1274. MR1731488

MARTIN, R., ZHANG, J. and LIU, C. (2010). Dempster–Shafer theory and statistical inference with weak beliefs. *Statist. Sci.* **25** 72–87.

MARTIN, R., HWANG, J.-S. and LIU, C. (2010). General theory of inferential models II. Marginal inference. Technical report, Dept. Statistics, Purdue Univ.

MENG, X. L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503

MENG, X. L. and VAN DYK, D. (1997). The EM algorithm—an old folk-song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 511–567. MR1452025

MENG, X. L. and VAN DYK, D. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrica* **86** 301–320.

MUDHOLKAR, G. S. and GEORGE E. O. (1978). A remark on the shape of the logistic distribution. *Biometrika* **65** 667–668.

QI, A. and JAAKKOLA, T. S. (2007). Parameter expanded variational Bayesian methods. In *Adv. Neural Info. Proc. Syst.* **19**. MIT Press, Cambridge.

VAN DYK, D. A. and MENG, X. L. (2001). The art of data augmentation (with discussion). *J. Comput. Graph. Statist.* **10** 1–111. MR1936358

VAN DYK, D. A. and MENG, X. L. (2010). Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: A graphical guide book. *Statist. Sci.* To appear.

YU, Y. and MENG, X. L. (2008). Espousing classical statistics with modern computation: Sufficiency, ancillarity and and interweacing generation of MCMC. Technical report, Dept. Statistics, Univ. California, Irvine.

YU, Y. and MENG, X. L. (2010). To center or not to center, that is not the question: An ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *J. Comput. Graph. Statist.* To appear.

ZHANG, J. and LIU, C. (2011). Dempster–Shafer inference with weak beliefs. *Statist. Sinica*. To appear.