Submitted to the Statistical Science

# Parameter Expansion and Efficient Inference *

**Andrew Lewandowski** † **Chuanhai Liu**† **and Scott Vander Wiel**‡

Purdue University and Los Alamos National Laboratory

*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 explores the potential of parameter expansion for constructing ancillary statistics for conditional inference. In addition, it briefly discusses a few relevant ideas motivated by EM and PX-EM, including an alternative view of PX-EM from the perspective of efficient data augmentation, the broader impact of the statistical thinking on understanding and developing other iterative optimization algorithms.

*Key words and phrases:* Ancillary statistics, Conditional inference, Robit regression, The EM algorithm, The PX-EM algorithm.

## 1. INTRODUCTION

The EM algorithm of Dempster, Laird, and Rubin (1977) has proven to be a popular computational method for optimization. Many variants of the EM algorithms have also proposed in the last 30+ years. Among these EM-type algorithms are the ECM algorithm of Meng and Rubin (1993), the ECME algorithm of Liu and Rubin (1994), the AECM algorithm of Meng and van Dyk (1997), and the PX-EM algorithm of Liu, Rubin, and Wu (1998), to name a few. This review article focuses on parameter expansion and explores the connection between parameter expansion and efficient inference.

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

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

*This is an invited contribution to the Special EM issue for Statistical Science. The authors thank Xiao-Li Meng and David van Dyk for their invitation.

†Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 47907. (e-mail: alewand@purdue.edu, chuanhai@purdue.edu; url: www.stat.purdue.edu).

‡Statistical Sciences Group, MS F600, Los Alamos National Laboratory, Los Alamos, NM 87545. (e-mail: scottv@lanl.gov; url: www.stat.lanl.gov).

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

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

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

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

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

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

Two EM examples are given in Sections 2.2 and 3.

Suppose that the complete-data model can be embedded in a larger model $g_*(X_{\mathrm{obs}}, X_{\mathrm{mis}}; \theta_*, \alpha)$ 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 $(\theta_*, \alpha) \in \Theta \times \mathcal{A}$,

$$f(X_{\mathrm{obs}}; \theta) = f_*(X_{\mathrm{obs}}; \theta_*, \alpha), \tag{1.3}$$

holds for some $\theta \in \Theta$, where $f_*(X_{\mathrm{obs}}; \theta_*, \alpha) = \int_{\mathcal{X}_{\mathrm{mis}}} g_*(X_{\mathrm{obs}}, X_{\mathrm{mis}}; \theta_*, \alpha) dX_{\mathrm{mis}}$. The condition (1.3) 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 $\Theta$. 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 $\alpha$ at $\alpha_0$. Formally, assume that there exists a *null* value of $\alpha$, denoted by $\alpha_0$, such that $\theta = R(\theta, \alpha_0)$ for every $\theta \in \Theta$. When applied to the parameter-expanded complete-data model $g_*(X_{\mathrm{obs}}, X_{\mathrm{mis}}; \theta_*, \alpha)$, 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 $\Theta$, PX-EM generates a sequence $\{\theta^{(t)} = R(\theta_*^{(t)}, \alpha^{(t)})\}$ and converges no slower than the corresponding EM based on $g(X_{\mathrm{obs}}, X_{\mathrm{mis}}; \theta)$; see Liu, Rubin, and Wu (1998).

For simplicity and stability, Liu, Rubin, and Wu (1998) uses $(\theta^{(t)}, \alpha_0)$ 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: $\theta^{(t)} = R(\theta_*^{(t+1)}, \alpha^{(t+1)})$. More precisely, the PX-EM algorithm is defined by replacing the E and M steps of EM with the following:

*PX-E step.* Compute

$$Q(\theta_*, \alpha|\theta^{(t)}, \alpha_0) = \mathrm{E}\left(\ln g_*(X_{\mathrm{obs}}, X_{\mathrm{mis}}; \theta_*, \alpha)|X_{\mathrm{obs}}, \theta_* = \theta^{(t)}, \alpha = \alpha_0\right)$$

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

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

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 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); and Liu (2004)), which serves as a simple robust alternative to logistic regression and is called robit regression by Liu (2004). In Section 3.4, we argue that parameter expansion can also be used for efficient data augmentation in the E step. However, the resulting EM is effectively the PX-EM algorithm.

Gelman (2004; see also Gelman et al. 2008) showed that parameter expansion offers more than a computational method to accelerate EM. He pointed 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; and Liu 2003) and that it suggests a new family of prior distributions in a Bayesian framework. Inspired by Gelman (2004), we also consider other potential applications of parameter expansion. The selected topics include a stochastic version of PX-EM for Bayesian posterior simulation in Section 4, identification of ancillary statistics for conditional inference in Section 5, and a few remarks in Section 6.

## 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 $X_{\text{mis}} = Z$:

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

where $\pi \in (0,1)$ is known and $\lambda > 0$ is the unknown parameter to be estimated. The observed-data model is $X|\lambda \sim \text{Poisson}(\pi\lambda)$, giving the ML estimate of $\lambda$, $\hat{\lambda} = X/\pi$. We use this example to show an extreme case that PX-EM can converge dramatically faster than its parent EM: PX-EM converges in one-step whereas EM converges painfully slow.

The complete-data likelihood is given by

$$\frac{\lambda^Z}{Z!}e^{-\lambda}\frac{Z!}{X!(Z-X)!}\pi^X(1-\pi)^{Z-X} \qquad (\lambda > 0) \tag{2.1}$$

It follows that the complete-data model belongs to the exponential family with sufficient statistic $Z$ for $\lambda$. The complete-data ML estimate of $\lambda$ is given by

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

With the updated estimate $\lambda^{(t)}$ at the $t$-th iteration, the $t$-th iteration of EM steps follows these two steps:

*E step.* Compute $\hat{Z} = \text{E}(Z|X,\lambda = \lambda^{(t)}) = X + \lambda^{(t)}(1-\pi)$.
*M step.* Replace $Z$ in (2.2) with $\hat{Z}$ to obtained $\lambda^{(t+1)} = \hat{Z}$.

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

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

converging to the ML estimate

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

We rewrite (2.3) to give a simple expression for the convergence rate of EM

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

indicating that EM is very slow when $\pi \approx 0$.

### 2.2 ML estimation of Robit regression via EM

Consider the observed data consisting of $n$ observations $X_{\text{obs}} = \{(x_i, y_i) : i = 1, ..., n\}$ with a $p$-dimensional covariates 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

$$\Pr(y_i = 1|x_i, \beta) = 1 - \Pr(y_i = 0|x_i, \beta) = F_\nu(x_i'\beta) \qquad (i = 1, ..., n) \tag{2.4}$$

where $F_\nu(.)$ denotes the cdf of the student-t distribution with center zero, unit scale, and $\nu$ 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 $\nu$.

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

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

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 $\beta$:

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

The complete-data ML estimate of $\beta$ 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, $\beta^{(0)} = (0, ..., 0)$, EM iterates for $t = 0, 1, ...$ with the $(t+1)$-th iteration consisting of the following E and M steps.

*E step.* Compute $\hat{S}_{\tau xx'} = \text{E}\left(S_{\tau xx'}|\beta = \beta^{(t)}, X_{\text{obs}}\right)$ and $\hat{S}_{\tau xz} = \text{E}\left(S_{\tau xz}|\beta = \beta^{(t)}, X_{\text{obs}}\right)$.
*M step.* Update the estimate of $\beta$ to obtain $\beta^{(t+1)} = \hat{S}_{\tau xx'}^{-1} \hat{S}_{\tau xz}$.

Let $f_\nu(.)$ denote the pdf of $F_\nu(.)$. The E-step is coded by using the following results derived in Liu (2004):

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

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

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 $i=1,...,n$.

## 3. THE PX-EM ALGORITHM

### 3.1 Robit regression: efficient analysis of imputed missing data

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 $(\tau_i, z_i)$ data conditioned on the observed data $X_{\mathrm{obs}}$ and $\beta^{(t)}$, the current estimate of $\beta$ at the $t$-th iteration. Had the ML estimate of $\beta$, $\hat{\beta}$, 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, *i.e.*, $\beta^{(t)} \neq \hat{\beta}$, 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 $\beta^{(t+1)}$ towards $\hat{\beta}$ 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 $\tau_i$ and $z_i$, denoted by $\alpha$ and $\sigma$. 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 (see Lemma 3.1) 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, $\beta$, 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 $(\alpha, \sigma)$ introduced to capture the bias in the imputed values of $\tau_i$ 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 (1.3). In the robit regression model, let $(\beta_*, \alpha, \sigma)$ be the expanded parameter with $\beta_*$ playing the same role as $\beta$ in the original model. The preservation condition states that for every expanded parameter value $(\beta_*, \alpha, \sigma)$, there exists a value of $\beta$ 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 $\beta$. 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

$$(\tau_i/\alpha)|(\beta_*, \alpha, \sigma) \sim \text{Gamma}(\nu/2, \nu/2) \tag{3.1}$$

and

$$z_i|(\tau_i, \beta_*, \alpha, \sigma) \sim \text{N}(x_i'\beta_*, \sigma^2/\tau_i) \tag{3.2}$$

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

$$\beta = R(\beta_*, \alpha, \sigma) = (\alpha/\sigma)\beta_* \qquad (\beta_* \in \mathcal{R}^p; \alpha > 0; \sigma > 0) \tag{3.3}$$

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

$$S_\tau = \sum_{i=1}^n \tau_i, \quad S_{\tau x x'} = \sum_{i=1}^n \tau_i x_i x_i', \quad S_{\tau z^2} = \sum_{i=1}^n \tau_i z_i^2, \quad S_{\tau x z} = \sum_{i=1}^n \tau_i x_i z_i'. \tag{3.4}$$

The complete-data ML estimate of $\beta_*$ is the same as that of $\beta$ in the original complete-data model. The complete-data ML estimates of $\alpha$ and $\sigma$ are given by

$$\hat{\alpha}_{\text{com}} = \frac{1}{n} S_\tau \quad \text{and} \quad \hat{\sigma}^2_{\text{com}} = \frac{1}{n}\left(S_{\tau z^2} - S_{\tau x z} S^{-1}_{\tau x x'} S_{\tau x z}\right). \tag{3.5}$$

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:

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

and

$$\hat{S}_{\tau z^2} = \text{E}\left(S_{\tau z^2} | \beta = \beta^{(t)}, X_{\text{obs}}\right) = n(\nu+1) - \nu \sum_{i=1}^n \hat{\tau}_i + \sum_{i=1}^n \hat{\tau}_i x_i' \beta^{(t)} (2\hat{z}_i - x_i'\beta^{(t)}),$$

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

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

### 3.2 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 $\pi$. This mismatch can be used to adjust $\lambda^{(t+1)}$. To adjust $\lambda^{(t+1)}$, we activate $\pi$ and let $\alpha$ denote the activated parameter with $\alpha_0 = \pi$. Now the parameter-expanded complete-data model becomes

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

where $\lambda_* > 0$ and $\alpha \in (0,1)$. The two corresponding observed-data models are Poisson($\lambda\pi$) and Poisson($\lambda_*\alpha$), giving the reduction function

$$\lambda = R(\lambda_*,\alpha) = \frac{\alpha}{\pi}\lambda_*. \tag{3.6}$$

The complete-data sufficient statistics are $Z$ and $X$. The complete-data ML estimates of $\lambda_*$ and $\alpha$ are given by

$$\hat{\lambda}_{*,\text{com}} = Z \quad \text{and} \quad \hat{\alpha}_{\text{com}} = \frac{X}{Z}. \tag{3.7}$$

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 (3.7) with $\hat{Z}$ to obtained $\lambda_*^{(t+1)} = \hat{Z}$ and $\alpha^{(t+1)} = X/\hat{Z}$. Update $\lambda$ 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 does show that PX-EM can converge dramatically faster than its parent EM.

### 3.3 A relevant theoretical result

To present a general result supporting the statistical explanation made in Section 3, we consider the following hypothetical EM algorithm based on the parameter-augmented complete-data $g_*(X_{obs}, X_{mis}; \theta_*, \alpha)$ without updating the estimate of $\alpha$.

*H-E step.* This is the E step of EM but with parameter fixed at $(\theta_*^{(t)}, \alpha_0)$, instead of $(\theta_*^{(t)}, \alpha^{(t)})$.

*H-M step.* This is the M step of EM that estimates both $\theta_*$ and $\alpha$.

For this hypothetical EM, we have the following theorem in parallel to the standard results for EM-type algorithms. The proof can be established using the standard method.

THEOREM 3.1. *Suppose that the parameter-expanded complete-data model satisfies the condition that for all* $\alpha \in \mathcal{A}$ *and* $t$,

$$\begin{aligned} & E\left[\ln g_*(X_{obs}, X_{mis}; \theta_*^{(t+1)}, \alpha) | X_{\text{obs}}, \theta_*^{(t)}, \alpha_0\right] \\ = & \max_{\theta_* \in \Theta} E\left[\ln g_*(X_{obs}, X_{mis}; \theta_*, \alpha) | X_{\text{obs}}, \theta_*^{(t)}, \alpha_0\right]. \end{aligned}$$

*The hypothetical EM algorithm increases the loglikelihood of the observed-data model at each iteration, that is,* $\ln f(X_{obs}; \theta^{(t+1)}) \geq \ln f(X_{obs}; \theta^{(t)})$ *for all* $t$. *If* $\ln f(X_{obs}|\theta)$ *is bounded, then* $\ln f(X_{obs}|\theta^{(t)})$ *converges.*

The main result of interest here concerns the convergence property of $\alpha^{(t)}$ to $\alpha_0$, to support the statistical interpretation in terms of bias reduction based on observed discrepancies captured by $\alpha^{(t)}$. The result can be easily proved by contradiction using the "standard" arguments for convergence of EM-type algorithms or by using the PX-EM theory.

LEMMA 3.1. *Suppose that* $(i)$ *for every fixed* $\alpha \in \mathcal{A}$, *the reduction function* $\theta = R(\theta_*, \alpha)$ *is one-to-one from* $\Theta$ *to* $\Theta$, $(ii)$ *both complete-data and observed-data likelihood functions are unimodal, and* $(iii)$ *the hypothetical EM converges to* $\hat{\theta} = \max_\theta \ln f(X_{\text{obs}}; \theta)$. *Under the condition of Theorem 3.1, the sequence* $\alpha^{(t)}$ *produced by the hypothetical EM converges to* $\hat{\alpha} = \alpha_0$.

### 3.4 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 augmenting scheme for each EM iteration. Rather than control the fraction of missing-data information, this procedure eliminates 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 $\alpha_0 = \pi$, *i.e.*,

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

which has the canonical representation

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

Thus, when fixed at the given value, $\pi$, for identifiability, the complete-data ML estimate $\hat{\alpha} = X/Z$ plays a role of ancillary statistic; see Section 5. With the correct imputation model, or at convergence, the imputed value $\hat{Z}$ satisfies

$$\pi = \frac{X}{\hat{Z}} \quad \text{or} \quad \hat{Z} = \frac{X}{\pi}. \tag{3.8}$$

Thus, we can consider modifying the E step of EM to produce imputed statistics $\hat{Z}$ that satisfies (3.8).

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

$$\Omega_*^{(t)} \equiv \{(\lambda_*, \alpha) : R(\lambda_*, \alpha) = R(\lambda^{(t)}, \alpha_0)\} = \{(\lambda_*, \alpha) : \lambda^{(t)}\pi = \alpha\lambda_*\}. \tag{3.9}$$

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 (3.8) 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

$$\tilde{\lambda}_* = X\frac{1-\pi}{\pi} + \lambda^{(t)}\pi \quad \text{and} \quad \tilde{\alpha} = \frac{\lambda^{(t)}\pi}{X\frac{(1-\pi)}{\pi} + \lambda^{(t)}\pi}.$$

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 in a certain sense can be viewed as re-imputation of missing sufficient statistics.

## 4. STOCHASTIC VERSIONS OF PX-EM FOR BAYESIAN INFERENCE

The EM algorithm influenced the creation of the Data Augmentation (DA) algorithm of Tanner and Wong (1987). DA can be thought of as a Bayesian version of EM, and as with EM, its performance can be improved through parameter-expansion. DA requires the introduction of a data augmentation scheme, which will continue to be denoted by $X_{\mathrm{mis}}$, that allows the posterior of interest $p(\theta|X_{\mathrm{obs}})$ to be recovered through the relationship

$$p(\theta|X_{\mathrm{obs}}) = \int_{\mathcal{X}_{\mathrm{mis}}} g(X_{\mathrm{obs}}, X_{\mathrm{mis}}; \theta)\pi(\theta)dX_{\mathrm{mis}}.$$

After the specification of $\pi(\theta)$, a prior distribution for $\theta$, the DA algorithm alternates between the following two steps, which are based on draws from the conditional posterior distributions:

*Imputation (I)-step.* Draw $X_{\mathrm{mis}}$ from its conditional distribution, given $\theta$ and $X_{\mathrm{obs}}$; and

*Posterior (P)-step.* Draw $\theta$ from its conditional distribution, given $X_{\text{mis}}$ and $X_{\text{obs}}$.

The Gibbs sampler (Gelfand and Smith, 1990) generalizes this method. When $\theta$ is one-dimensional, DA is essentially the same as a two-stage Gibbs sampler. As with the Gibbs sampler, DA creates a Markov chain which asymptotically converges to $\pi(X_{\text{mis}}, \theta|X_{\text{obs}})$, the joint posterior of $X_{\text{mis}}$ and $\theta$.

The I- and M-steps parallel the E- and M-steps in the EM algorithm. However, DA also resembles EM in that the convergence rate is sometimes notoriously slow. As with other Markov Chain Monte Carlo (MCMC) methods, high levels of autocorrelation in the Markov chain produced by DA slow the convergence rate. For a discussion of the convergence behavior of DA and the Gibbs sampler, see Liu, Wong, and Kong (1994, 1995) and Liu (2001). The latter also discusses various methods that have been proposed to improve the convergence rate of the Gibbs sampler and DA.

In particular, reparametrization techniques have long been used to improve the performance of the Gibbs sampler. Some examples are discussed in Gelman (2004) and Gelman et al. (2008). However, the use of parameter expansion in DA was not formalized until Meng and van Dyk (1999) and Liu and Wu (1999) independently published similar methods that resemble PX-EM. These methods are called marginal augmentation and parameter-expanded DA (PX-DA), respectively, and differ in the choice of prior and the sequence in which the missing data, model parameters, and expansion parameter are drawn. Imputing the expansion parameter weakens the dependence between the the imputed missing data and the parameters drawn in the P-step. Both methods have been shown to have a better convergence rate than DA methods that hold $\alpha$ fixed. Related theoretical discussions can be found in Section 4.2 of Liu and Wu (1999) and Section 3 of van Dyk and Meng (2001).

Instead of a reduction function, both methods use the expansion parameter $\alpha$ to determine a one-to-one and differentiable missing-data transformation $t_\alpha$ of the missing data in the parameter-expanded model. This transformation allows the missing data in the original augmented model to be recovered from the expanded model. For example, a common transformation is $t_\alpha(\tilde{X}_{\text{mis}}) = \alpha\tilde{X}_{\text{mis}}$, where $\tilde{X}_{\text{mis}}$ is the missing data in the parameter-expanded space.

The exact form of these methods depends on the choice of prior on $\alpha$, which can also be conditioned on $\theta$. The prior for $\theta$ will not change when a parameter-expanded model is used. In fact, this is a condition for the posterior of $\theta$ to remain unchanged in the new model (see Proposition 1 in Liu and Wu (1999)). Meng and van Dyk (1999) present several standard DA procedures to simulate from the posterior when the prior on $\alpha$ is proper. One simple method is the following:

*Step 1* Draw $\alpha$ and $\tilde{X}_{\text{mis}}$ jointly from their conditional distribution, given $\theta$ and $X_{\text{obs}}$; and

*Step 2* Draw $\theta$ and $\alpha$ jointly from

$$p(\theta, \alpha|X_{\text{obs}}, \tilde{X}_{\text{mis}}) \propto g_*(X_{\text{obs}}, \tilde{X}_{\text{mis}}|\theta, \alpha)p(\alpha|\theta)f(\theta),$$

where $f(\theta) = \int p(\theta, \alpha)d\alpha$ is the marginal prior for $\theta$. The proper prior on $\alpha$ ensures that $p(\alpha|\tilde{X}_{\text{mis}}) = p(\alpha)$ is a proper distribution Then, $\tilde{X}_{\text{mis}}$ and $\alpha$ can be drawn sequentially in Step 1. This algorithm does not require the specification of

the missing data transformation $t_\alpha$. However, if $X_{\text{mis}}$ is drawn from its conditional distribution, the inverse transformation $t_\alpha^{-1}(X_{\text{mis}})$ then allows $\tilde{X}_{\text{mis}}$ to be found once $\alpha$ is simulated. When $|J_\alpha(\tilde{X}_{\text{mis}})| = det\{dt_\alpha(\tilde{X}_{\text{mis}})/d\tilde{X}_{\text{mis}}\}$ is the Jacobian for the transformation induced by $t_\alpha$, the parameter-expanded complete data likelihood needed in Step 2 can be written as

$$g_*(X_{\text{obs}}, \tilde{X}_{\text{mis}}|\theta, \alpha) = f(X_{\text{obs}}, t_\alpha(\tilde{X}_{\text{mis}})|\theta)|J_\alpha(\tilde{X}_{\text{mis}})|. \tag{4.1}$$

However, a proper prior on the expansion parameter is typically not desirable. A less informative prior will allow $\alpha$ to vary with more freedom, which should reduce autocorrelation in the chain. Several options are available in this case. When the improper prior on $\alpha$ is the limit of a sequence of improper priors, both Liu and Wu (1999) an Meng and van Dyk (1999) show that this prior can be used in a valid DA scheme. Additionally, Liu and Wu (1999) show that a prior based on a Haar measure produces a DA method that is optimal in the sense that no other DA method with a proper prior can have a better convergence rate. This convergence result is the main motivation behind using a prior based on the Haar measure here.

We note that the formulation of PX-DA using data-transformation limits the applications of these results in that not all missing data problems have such a transformation. For example, in the toy Poisson-Binomial example from Section 2.1, $t_\alpha(\tilde{X}_{\text{mis}}) = \frac{\alpha}{\pi}\tilde{X}_{\text{mis}}$ is the missing-data transformation version of the reduction function in (3.6), but this does not produce a valid distribution for $t_\alpha(\tilde{X}_{\text{mis}})$. Thus, at least to some extent, constructing an exact DA version of PX-EM remains an open problem. Nevertheless, work in this area has led to efficient DA algorithms with ideas that can be easily generalized to the Gibbs sampler.

More recently, methods have been developed that avoid the need to specify the prior for an expansion parameter. In the Covariance-Adjusted DA (CA-DA) method (Liu 2003), a pair of transformations $A(X_{\text{mis}})$ and $S(X_{\text{mis}})$ replace the transformation induced by the expansion parameter. The $S$ transformation captures essential information from the imputed data. The $A$ transformation serves as a covariance adjustment and is thus similar in function to the expansion parameter in parameter-expansion methods. Then, CA-DA adds a step that takes advantage of these transformations:

*CA P-step* Draw $\theta$ and $A(X_{\text{mis}})$ from their conditional distribution given $X_{\text{obs}}$ and $S(X_{\text{mis}})$.

In practice, the $S$ and $A$ transformations essentially create sufficient and ancillary statistics for $\theta$. This is seen in Example 1 of Liu (2003). Yu and Meng (2007) have developed a DA scheme that further exploits the role that transformations can play in reducing the influence of $X_{\text{mis}}$ on the the simulated parameters. To clarify the similarities between their method and both CA-DA and the related Alternate Subspace-Spanning Resampling (ASSR) algorithm (Liu 2003), let $S_\theta(X_{\text{mis}}) = A_\theta^{-1}(X_{\text{mis}})$, although this relationship does not have to hold in general. Even if the distribution of $A_\theta$ is free of the parameter, the subscripts are needed to show that the transformation depends on the value of $\theta$. Then, they expand the CA P-step so that

*Ancillary (A)-step* Draw $\hat{\theta}$ from its conditional distribution given $A_\theta(X_{\text{mis}})$ and $X_{\text{obs}}$,

*Sufficient (S)-step* Draw $\theta$ from its conditional distribution given $S_{\hat{\theta}}(X_{\text{mis}})$ and $X_{\text{obs}}$.

Although the ancillary and sufficient transformations may be difficult to find in practice, these methods use statistical principles to reduce the dependence between steps in a DA chain, and Yu and Meng (2007) present promising convergence results that suggest that this scheme will perform as well as PX-DA.

## 5. CONDITIONAL INFERENCE

Ancillary statistics were introduced in Fisher (1925; see also Fisher (1973)) for efficient inference. R. A. Fisher did not provide a formal definition of ancillarity, but the usual definition is that a statistic is ancillary if its distribution does not depend on the parameters in the model; see Ghosh, Reid, and Fraser (2010) for a recent nice review of ancillary statistics. It is sensible, at least for simplicity, to demand in addition that ancillary statistics should be functions of minimal sufficient statistics; see Owen (1948) and Lehmann and Scholz (1992). But this is not necessary as far as the objective of using ancillary statistics is concerned.

The objective of using ancillary statistics is to provide efficient maximum likelihood-based inference. A way of doing this is to adjust for the observed ancillary statistic. Thus, the ancillary statistic to be adjusted for plays the role of covariate. Accordingly, efficient inference about the parameters in the assumed model is then obtained via the general method of covariance/covariate adjustment by conditioning on the ancillary statistic. A peculiar feature of such a (generalized) covariance adjustment for ancillary statistics is not on the "location", the MLE of the parameter, but on the variance of the MLE of the parameter. This point of view leads us to consider a parameter expansion approach to identification of ancillary statistics, which has proven to be a challenging problem.

To motivate our discussion, consider the problem of the Nile (see, *e.g.*, Fisher, 1973), a classic example of ancillary statistics.

EXAMPLE 5.1. Suppose that $(x_i, y_i)$ are iid random variables such that $x_i$ and $y_i$ follow exponential distributions with means $\theta^{-1}$ and $\theta$, respectively, for $i = 1, \ldots, n$. Then, for $\theta > 0$, their pdf can be written as

$$f(x, y; \theta) = \exp(-\theta x - y/\theta) \qquad (x > 0; y > 0) \tag{5.1}$$

The log-likelihood function

$$-[\theta S_x + S_y/\theta]$$

is maximized at

$$\hat{\theta} = T = \sqrt{\frac{S_y}{S_x}}$$

the ML estimate of $\theta$, where

$$S_x = \sum_{i=1}^{n} x_i \quad \text{and} \quad S_y = \sum_{i=1}^{n} y_i$$

denote the minimal sufficient statistics. It is shown, for example, in Fisher (1973) that the distribution of the statistic

$$U = n^{-1}\sqrt{S_x S_y}$$

is independent of $\theta$, but inference about $\theta$ conditional on $U$, an ancillary, is more efficient than the inference based on $T$ alone.

An important consideration is the identification of ancillary statistics for a given sampling model for the observed data. We refer to Ghosh, Reid, and Fraser (2010) for discussion of the various methods. The approach closely related to parameter expansion is *to embed the assumed model into a parameter expanded model where the expanded parameter and the minimal sufficient statistic have the same dimensionality*; see Reid (2003). However, as a formal way of specifying ancillary statistics, this idea has not been fully explored. For example, defining the mapping from the parameter expanded space to the original space has been mostly based on intuition for a small number of examples. While a full exploration needs to be carried out in the future, the potential of using parameter expansion to identify ancillary statistics is discussed here with the problem of the Nile.

Note that the minimal sufficient statistics $S_x$ and $S_y$ allow two unknown scalar parameters to be estimated. This suggests to consider parameter expanded models. The fact that $S_x$ and $S_y$ are independent with

$$\theta S_x \sim \text{Gamma}(n) \quad \text{and} \quad \theta^{-1} S_y \sim \text{Gamma}(n)$$

leads to the simple but otherwise arbitrary parameter expanded model

$$S_x/\lambda_x \sim \text{Gamma}(n) \quad \text{and} \quad S_y/\lambda_y \sim \text{Gamma}(n) \qquad (\lambda_x > 0; \lambda_y > 0) \tag{5.2}$$

where $(\lambda_x, \lambda_y)$ is the expanded parameter. While the expanded model is extended from the original model based on simplicity, a problem of our interest is how to specify a one-to-one mapping

$$\theta = \theta(\lambda_x, \lambda_y) \quad \text{and} \quad \alpha = \alpha(\lambda_x, \lambda_y)$$

with $\alpha$ being the expansion parameter and $\theta$ playing the role of the original parameter when $\alpha$ is fixed at a default value $\alpha_0$. More conditions are to imposed below on this mapping.

Let $\hat{\lambda}_x$ and $\hat{\lambda}_y$ denote the ML estimators of $\lambda_x$ and $\lambda_y$, respectively. Let $\hat{\theta}_\alpha$ denote the conditional ML estimator of $\theta$ given $\alpha$. Since the use of ancillary statistics is a "variance" adjustment, we require

$$\hat{\theta}_{\alpha_0} = \theta(\hat{\lambda}_x, \hat{\lambda}_y)$$

to hold for all $(\hat{\lambda}_x, \hat{\lambda}_y)$. This condition effectively determines the mapping $\theta = \theta(\lambda_x, \lambda_y)$, which is obtained by representing the minimal sufficient statistics as functions of $\hat{\lambda}_x$ and $\hat{\lambda}_y$. More precisely, let $\ell(\theta|\hat{\lambda}_x, \hat{\lambda}_y)$ be the log-likelihood function. Then the mapping $\theta = \theta(\lambda_x, \lambda_y)$ is given by

$$\arg\max_\theta \ell(\theta|\lambda_x, \lambda_y) = \theta, \tag{5.3}$$

that is, $\partial \ell(\theta|\lambda_x, \lambda_y)/\partial\theta = 0$. Note that

$$\hat{\lambda}_x = \frac{S_x}{n} \quad \text{and} \quad \hat{\lambda}_y = \frac{S_y}{n}.$$

Then, $\ell(\theta|\hat{\lambda}_x, \hat{\lambda}_y) = -\theta S_x - S_y/\theta = -n\lambda_x\theta - n\lambda_y/\theta$ and, thereby,

$$\arg\max_{\theta} \ell(\theta|\lambda_x, \lambda_y) = \sqrt{\lambda_y/\lambda_x}.$$

The condition (5.3) gives the explicit expression

$$\theta = \theta(\lambda_x, \lambda_y) = \sqrt{\frac{\lambda_y}{\lambda_x}}. \tag{5.4}$$

It appears to be challenging to determine $\alpha = \alpha(\lambda_x, \lambda_y)$ in such a way that $\hat{\alpha} = \alpha(\hat{\lambda}_x, \hat{\lambda}_y)$ is ancillary. Here we consider two possible approaches. The first is based on the consideration of the effect of the expansion parameter on the maximum likelihood. The parameter expanded model fits the data better than the original model in the sense of the maximum likelihoods. Thus the gain in the maximum log-likelihood

$$\Delta(\hat{\lambda}_x, \hat{\lambda}_y) = \max_{\lambda_x, \lambda_y} \ln f(S_x, S_y|\lambda_x, \lambda_y) - \max_{\theta} \ln f(S_x, S_y|\theta)$$

is what is captured by the expansion parameter $\alpha$. This suggests to specify $\alpha(\lambda_x, \lambda_y)$ based on $\Delta(\lambda_x, \lambda_y)$. We note that this is essentially the idea discussed by Barndorff-Nielsen and Cox (1994) who call the signed squared-root likelihood ratio statistic

$$A_\Delta = \text{sign}(\hat{\alpha} - \alpha_0)\sqrt{2\Delta(\hat{\lambda}_x, \hat{\lambda}_y)}$$

a directed likelihood ancillary. The asymptotic validity of $A_\Delta$ as an ancillary statistic is evident because the asymptotic distribution of $2\Delta(\hat{\lambda}_x, \hat{\lambda}_y)$ is $\chi^2_1$, independent of $\theta$. Here in this example, $\Delta(\hat{\lambda}_x, \hat{\lambda}_y)$ has the explicit expression

$$\Delta(\hat{\lambda}_x, \hat{\lambda}_y) = 2n\left[\sqrt{\hat{\lambda}_y\hat{\lambda}_y} - 1 - \ln\sqrt{\hat{\lambda}_y\hat{\lambda}_y}\right] \approx n\hat{\lambda}_x\hat{\lambda}_y,$$

when $\hat{\lambda}_x\hat{\lambda}_y \approx 1$, the default value of $\lambda_x\lambda_y$ at which the expanded model recovers the original model. Thus, $\Delta(\hat{\lambda}_x, \hat{\lambda}_y)$ suggests to construct the ancillary as a simple function of $\hat{\lambda}_x\hat{\lambda}_y$. As $\hat{\lambda}_x\hat{\lambda}_y$ is indeed an exact ancillary, we can set

$$\alpha = \alpha(\lambda_x, \lambda_y) = \lambda_x\lambda_y$$

or any monotone function of $\lambda_x\lambda_y$.

It is clear to see that the signed likelihood ratio approach is attractive because it does not need to have the explicit expressions for the mapping $\theta = \theta(\lambda_x, \lambda_y)$ and $\alpha = \alpha(\lambda_x, \lambda_y)$. The use of the likelihood ratio provides an efficient way of assessing the effect of the expansion parameter. Thus we can seek for some function $\alpha(\hat{\lambda}_x, \hat{\lambda}_y)$ in such a way that it is either exactly or approximately ancillary and that $\Delta(\hat{\lambda}_x, \hat{\lambda}_y)$ is (approximately) a function of $\alpha(\hat{\lambda}_x, \hat{\lambda}_y)$. Technically, this approach appears to be difficult in general, especially for multivariate ancillary statistics.

The second approach is based on the fact that the distribution of $\alpha(\hat{\lambda}_x, \hat{\lambda}_y)$ under the original model is independent of $\theta$. In terms of its characteristic function, this implies that $\alpha(\hat{\lambda}_x, \hat{\lambda}_y)$ satisfies the equation

$$\frac{\partial}{\partial\theta}\text{E}\left[e^{it\alpha(\hat{\lambda}_x, \hat{\lambda}_y)} f(\hat{\lambda}_x, \hat{\lambda}_y; \theta) d\hat{\lambda}_x d\hat{\lambda}_y\right] = 0 \qquad (\theta > 0; -\infty < t < \infty) \tag{5.5}$$

where $f(\hat{\lambda}_x, \hat{\lambda}_y; \theta)$ denotes the density function of $(\hat{\lambda}_x, \hat{\lambda}_y)$ in the original model. Let $Z_x = n\theta\hat{\lambda}_x$ and $Z_y = n\hat{\lambda}_y/\theta$. Then $Z_x$ and $Z_y$ are iid with Gamma$(n)$. Thus, (5.5) can be written as

$$\frac{\partial}{\partial\theta}\mathrm{E}\left[e^{it\alpha(Z_x/(n\theta),\theta Z_y/n)}\right] = 0.$$

Since the distribution of $(Z_x, Z_y)$ is independent of $\theta$, we have

$$it\mathrm{E}\left[\left(-\frac{\partial\alpha(\hat{\lambda}_x,\hat{\lambda}_y)}{\partial\hat{\lambda}_x}\frac{Z_x}{n\theta^2} + \frac{\partial\alpha(\hat{\lambda}_x,\hat{\lambda}_y)}{\partial\hat{\lambda}_y}\frac{Z_y}{n}\right)e^{it\alpha(Z_x/(n\theta),\theta Z_y/n)}\right] = 0,$$

that is,

$$\mathrm{E}\left[\frac{1}{\theta}\left(-\frac{\partial\alpha(\hat{\lambda}_x,\hat{\lambda}_y)}{\partial\hat{\lambda}_x}\hat{\lambda}_x + \frac{\partial\alpha(\hat{\lambda}_x,\hat{\lambda}_y)}{\partial\hat{\lambda}_y}\hat{\lambda}_y\right)e^{it\alpha(Z_x/(n\theta),\theta Z_y/n)}\right] = 0. \tag{5.6}$$

It is clear that (5.6) holds, provided that

$$\frac{\partial\alpha(\hat{\lambda}_x,\hat{\lambda}_y)}{\partial\hat{\lambda}_y}\hat{\lambda}_y - \frac{\partial\alpha(\hat{\lambda}_x,\hat{\lambda}_y)}{\partial\hat{\lambda}_x}\hat{\lambda}_x = 0. \tag{5.7}$$

Thus, the solution $\alpha(\hat{\lambda}_x, \hat{\lambda}_y)$ to the differential equation (5.7), a differentiable function of $\hat{\lambda}_x\hat{\lambda}_y$, *e.g.*, $\alpha(\hat{\lambda}_x, \hat{\lambda}_y) = \sqrt{\hat{\lambda}_x\hat{\lambda}_y}$, is an ancillary statistic.

With the problem of the Nile, we have demonstrated that parameter expansion has the potential to identify ancillary statistics in condition inference. The generality of these ideas remain to be seen in the future. For example, the second approach to finding $\alpha(\lambda_x, \lambda_y)$ is technically different from but conceptually related to Owen (1948). In the context of fiducial inference, Owen (1948) works directly with sufficient statistics for identification of ancillary statistics. Thus, examination of general situations is needed to evaluate the benefits of parameter expansion over that based on sufficient statistics for maximum likelihood-based inference.

## 6. DISCUSSION

Statistical interpretations of EM and PX-EM reveal that statistical thinking can be very helpful for understanding and developing iterative algorithms. While it seems natural for statistical problems such as ML and Bayesian estimation, statistical thinking can also be helpful for general-purpose optimization algorithms. For example, two of the authors, Liu and Vander Wiel (2007), took a statistical approach to two fundamental problems in the well-known Quasi-Newton method: (*i*) how to choose what is called the Broyden parameter for Hessian updates, and (*ii*) how to set the initial step size for line search algorithms. Thinking statistically led them to a significantly improved Quasi-Newton algorithm compared to existing Quasi-Newton methods, such as BFGS.

Gelman (2004) calls attention to his observation that "progress in statistical computation often leads to advances in statistical modeling", opening our eyes to look at the broader picture. In this article, we considered the potential application of parameter expansion in conditional inference using ancillary statistics.

It 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. It is not our intention here to have lengthy discussions on open problems. However, reparametrization and parameter expansion will continue to aid computation, particularly in the analysis of hierarchical models. Additionally, the concepts of sufficiency and ancillarity may lead to further computational advancements, even if their role is still being understood.

## ACKNOWLEDGEMENTS

The authors thank the reviewers for their helpful comments and suggestions.

## REFERENCES

Albert, J. H. and Chib. S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.*, **88**, 669-679.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics.* Chapman & Hall, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with Discussion). *J. Royal. Statist. Soc.* B, **39**, 1-38.

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700-725.

Fisher, R. A. (1973). *Statistical methods for scientific induction*, 3rd Ed. Hafner Publishing Company, New York.

Gelfand, A.E. and Smith, A.F.M. (1990). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.*, **82**, 528-550.

Gelman, A. (2004). Parametrization and Bayesian modeling. *J. Amer. Statist. Assoc.*, **99**, 537-545.

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.

Ghosh, M., Reid, N., and Fraser, D.A.S. (2010). Ancillary Statistics: A Review. *To appear in Statistica Sinica*

Lehmann, E. L. and Scholz F. W. (1992). Ancillarity. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu.* Eds. M. Ghosh and P. K. Pathak. IMS Lecture Notes and Monograph Series **17**, 32-51.

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

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

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.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, **85**, 755-770.

Liu, C. and Vander Wiel, S. A. (2007). Statistical Quasi-Newton: A new look at least change. *SIAM J. Optim.*, **18** 1266-1285.

Liu, J. S. (2001). *Monte Carlo strategies in statistical computing.* Springer, New York.

Liu, J. S., Wong, W.H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to comparisions of estimators and augmentation schemes. *Biometrika*, **81**, 27-40.

Liu, J. S., Wong, W.H., and Kong, A. (1995). Correlation structure and convergence rate of the Gibbs sampler for various scans. *J. R. Statist. Soc.* B, **57**, 157-169.

Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.*, **94**, 1264-1274.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267-278.

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.* B, **59**, 511-567.

Meng, X. L. and van Dyk, D. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation, *Biometrika*, **86**, 301-320.

Mudholkar, G. S. and George E. O. (1978). A remark on the shape of the logistic distribution. *Biometrika*, **65**, 667-668.

Owen, A. R. G. (1948). Ancillary statistics and fiducial inference. *Sankhya*, **9**, 1-18.

Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.*, **31**, 1695-1731.

Tanner, M. and Wong, W. (2001). The calculation of posterior distributions by data augmentation. (with discussion) *J. Amer. Statist. Assoc.*, **82**, 528-550.

van Dyk, D. A. and Meng, X. L. (2001). The art of data augmentation (with discussion). *J. Comput. Graph. Statist.*, **10**, 1-111.

Yu, Y. and Meng, X. L. (2008). Espousing classical statistics with modern computation: sufficiency, ancillarity and and interweacing generation of MCMC. *Technical Report*, Dept. of Statistics, University of California, Irvine.