Biometrika (1998), **85**, 4, *pp.* 755–770
Printed in Great Britain

Parameter expansion to accelerate EM: The PX-EM algorithm

BY CHUANHAI LIU
Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, U.S.A.
liu@research.bell-labs.com

DONALD B. RUBIN
Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A.
rubin@hustat.harvard.edu

AND YING NIAN WU
Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
yingnian@umich.edu

SUMMARY

The EM algorithm and its extensions are popular tools for modal estimation but are often criticised for their slow convergence. We propose a new method that can often make EM much faster. The intuitive idea is to use a 'covariance adjustment' to correct the analysis of the M step, capitalising on extra information captured in the imputed complete data. The way we accomplish this is by parameter expansion; we expand the complete-data model while preserving the observed-data model and use the expanded complete-data model to generate EM. This parameter-expanded EM, PX-EM, algorithm shares the simplicity and stability of ordinary EM, but has a faster rate of convergence since its M step performs a more efficient analysis. The PX-EM algorithm is illustrated for the multivariate t distribution, a random effects model, factor analysis, probit regression and a Poisson imaging model.

Some key words: AECM; Algorithms; Covariance adjustment; ECM; ECME; Factor analysis; Multivariate t distribution; Parameter expansion; Poisson imaging model; Probit regression; Random effects model.

1. INTRODUCTION

The EM algorithm (Dempster, Laird & Rubin, 1977) and many of its extensions and variations, such as the ECM algorithm (Meng & Rubin, 1993), the ECME algorithm (Liu & Rubin, 1994a), the SAGE algorithm (Fessler & Hero, 1994), the AECM algorithm (Meng & van Dyk, 1997) and efficient augmentation (Meng & van Dyk, 1997), which we shall generically call EM-type algorithms, are very popular tools for modal inference in a wide variety of statistical models in the physical, medical, biological and social sciences. Besides being conceptually attractive, EM-type algorithms are simple to implement and converge monotonically in terms of the loglikelihood or log-posterior of the observed-data model, under very general conditions (Dempster et al., 1977; Wu, 1983). An often-voiced criticism, however, is the slow convergence in some situations. We propose a method, called parameter expansion, which can often make EM dramatically faster. Our method, PX-EM, is based

on a statistical principle that is different from the principles underlying other EM acceleration methods, such as the variations mentioned above, as well as Aitken acceleration (Laird, Lange & Stram, 1987), conjugate gradient acceleration (Jamshidian & Jennrich, 1993) and quasi-Newtonian acceleration (Lange, 1995a, b). Besides often achieving significant improvements in rates of convergence, key features of PX-EM are that (i) it can be applied to many EM-type algorithms with only simple modifications, and (ii) it maintains the stability of EM-type algorithms with their monotone convergence.

The underlying statistical principle of PX-EM is to perform a 'covariance adjustment' to correct the M step, capitalising on extra information captured in the imputed complete data, in a manner analogous to the way a covariance adjustment captures extra information in the observed difference in treatment and control group covariate means in a randomised experiment. More specifically, we find an expanded complete-data model that has a larger set of parameters, but leads to the original observed-data model with the original parameters determined from the expanded parameters via a reduction function. Then PX-EM iteratively maximises the expected loglikelihood of the expanded complete-data model, with its expanded parameters and corresponding expanded sufficient statistics. The rate of convergence of PX-EM is at least as fast as the parent EM because its M step performs a more efficient analysis by fitting the expanded model.

Section 2 illustrates PX-EM for the multivariate t distribution. Section 3 provides the basic theory of PX-EM, including its definition, proof of both its monotone convergence and its superior rate of convergence relative to its parent EM, and an explicit interpretation of the M-step of PX-EM as covariance adjustment. Section 4 describes more examples, including a random effects model, factor analysis, probit regression and a Poisson imaging model. Finally, § 5 concludes with a short discussion.

2. EXAMPLE: MULTIVARIATE t DISTRIBUTION

The multivariate t distribution is a useful model for data analysis, especially for robust estimation, e.g. Rubin (1983), Lange, Little & Taylor (1989). Let $t_p(\mu, \Psi, \nu)$ denote a p-dimensional t random variable with centre μ, scatter matrix Ψ and known degrees of freedom ν, and let $\theta = (\mu, \Psi)$. For observed data $Y_{\text{obs}} = \{Y_1, \ldots, Y_N\}$, where $Y_i|\theta \sim t_p(\mu, \Psi, \nu)$, independently, the maximum likelihood estimate is known to have no closed form, but the EM algorithm is simple to apply by augmenting Y_{obs} to $Y_{\text{com}} = \{(Y_1, \tau_1), \ldots, (Y_N, \tau_N)\}$. The model for Y_{com} is

Model O:
$$Y_i|\tau_i, \theta \sim N_p(\mu, \Psi/\tau_i), \tag{1}$$
$$\tau_i|\theta \sim \chi^2_\nu/\nu, \tag{2}$$

ν known and positive, with $(Y_1, \tau_1), \ldots, (Y_N, \tau_N)$ independent; the $\tau_1, \ldots, \tau_N$ are augmented weights for the observations, where ν controls the uniformity of these weights.

Let $\theta^{(t)} = (\mu^{(t)}, \Psi^{(t)})$ be the parameter estimate at the tth iteration of the EM implementation generated by Model O. Then, at the $(t+1)$st iteration, we have the following.

E *step.* Assuming $\theta = \theta^{(t)}$, impute the expected values of the weights as a function of the dimension of Y_i, p, the known parameter, ν, and the Mahanobis distance between Y_i and its expectation, $d(Y, \mu, \Psi) = (Y-\mu)'\Psi^{-1}(Y-\mu)$:

$$\tau_i^{(t+1)} = E(\tau_i|\theta^{(t)}, Y_{\text{obs}}) = \frac{\nu + p}{\nu + d(Y_i, \mu^{(t)}, \Psi^{(t)})}. \tag{3}$$

M *step.* Maximise the expected loglikelihood of Model O by weighted least squares, giving

$$\mu^{(t+1)} = \sum_{i=1}^{N} \tau_i^{(t+1)} Y_i \Big/ \sum_{i=1}^{N} \tau_i^{(t+1)},$$

$$\Psi^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \tau_i^{(t+1)} (Y_i - \mu^{(t+1)})(Y_i - \mu^{(t+1)})'. \tag{4}$$

Kent, Tyler & Vardi (1994) proposed a modified algorithm, which changes (4) to

$$\Psi^{(t+1)} = \sum_{i=1}^{N} \tau_i^{(t+1)} (Y_i - \mu^{(t+1)})(Y_i - \mu^{(t+1)})' \Big/ \sum_{i=1}^{N} \tau_i^{(t+1)}.$$

This modification does not change the limit of the algorithm because

$$\sum_{i=1}^{N} \tau_i^{(t+1)}/N = 1, \tag{5}$$

when $\theta^{(t)}$ is its maximum likelihood estimate, θ_{MLE}, as proved by Kent et al. (1994). The modified EM, however, converges faster than the conventional EM, as reported by Kent et al. (1994), Arslan, Constable & Kent (1995) and Meng & van Dyk (1997), without incurring any extra computational cost. Meng & van Dyk (1997) also show that this modified EM is optimal among the EM implementations generated from a class of data augmentation schemes.

We now derive this modified EM using PX-EM. Note that, because ν is a fixed constant, there is no unknown parameter in (2) of Model O. We can, however, expand Model O to

Model X:
$$Y_i | \tau_i, \Theta \sim N_p(\mu_*, \Psi_*/\tau_i) \tag{6}$$

$$\tau_i | \Theta \sim \alpha \chi_\nu^2/\nu, \tag{7}$$

ν known, with $(Y_1, \tau_1), \ldots, (Y_N, \tau_N)$ independent, which adds an auxiliary scale parameter α in (7), where $\Theta = (\mu_*, \Psi_*, \alpha)$ is identifiable from the complete data. We use the notation μ_* and Ψ_* because we reserve μ and Ψ for the original parameter, $\theta = (\mu, \Psi)$. The auxiliary parameter α is 'hidden' at 1 in Model O in that Model X reduces to Model O when $\mu_* = \mu$, $\Psi_* = \Psi$ and $\alpha = 1$. Under Model X, the observed-data model becomes $Y_i | \Theta \sim t_p(\mu_*, \Psi_*/\alpha)$, independently, in which Ψ_*/α corresponds to Ψ, and μ_* to μ; that is, $\theta = R(\Theta)$ and

$$(\mu, \Psi) = R\{(\mu_*, \Psi_*, \alpha)\} = (\mu_*, \Psi_*/\alpha),$$

where R is the reduction function from the expanded parameter space to the original parameter space.

The PX-EM algorithm is implemented over the expanded parameter space as follows. Let $\Theta^{(t)} = (\mu^{(t)}, \Psi^{(t)}, \alpha_0)$ be the estimate of the expanded parameter with $\alpha^{(t)} = \alpha_0$ from the tth iteration. Then, by analogy with the E and M steps of EM, at the $(t+1)$st iteration we have the following.

PX-E *step.* Assuming $\Theta = \Theta^{(t)}$ in Model X, impute the weights as a function of p, ν and the distance between Y_i and its expectation:

$$\tau_i^{(t+1)} = E(\tau_i | \Theta^{(t)}, Y_{\text{obs}}) = \alpha^{(t)} \frac{\nu + p}{\nu + d(Y_i, \mu_*^{(t)}, \Psi_*^{(t)}/\alpha^{(t)})}. \tag{8}$$

PX-M *step.* Maximise the expected complete-data loglikelihood of Model X, giving

$$\mu_*^{(t+1)} = \sum_{i=1}^{N} \tau_i^{(t+1)} Y_i \Big/ \sum_{i=1}^{N} \tau_i^{(t+1)}, \quad \Psi_*^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \tau_i^{(t+1)} (Y_i - \mu^{(t+1)})(Y_i - \mu^{(t+1)})',$$

$$\alpha^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \tau_i^{(t+1)};$$

then apply the reduction function $R(\Theta)$ to obtain $\mu^{(t+1)} = \mu_*^{(t+1)}$ and $\Psi^{(t+1)} = \Psi_*^{(t+1)}/\alpha^{(t+1)}$, which are the same as those proposed by Kent et al. (1994).

The statistical principle underlying this PX-EM is to adjust the M step for the observed deviation between the imputed value of the statistic $\sum_i \tau_i$, which is the sufficient statistic associated with α under Model X, and its expectation under Model O, which is N. Although this deviation disappears at θ_{MLE} as shown by (5), this may not be the case before EM converges, and therefore the M step can be corrected for this observed deviation.

3. Basic theory of PX-EM

3·1. *Notation and background*

Let Y_{obs} be the set of observed data following the observed-data model $p(y_{\text{obs}}|\theta)$, with parameter θ to be estimated by its maximum likelihood estimate, θ_{MLE}. When $\log p(Y_{\text{obs}}|\theta)$ is difficult to maximise over θ directly, the EM algorithm augments Y_{obs} to a set of complete data Y_{com}, which can be reduced to Y_{obs} via a many-to-one mapping ρ, that is $Y_{\text{obs}} = \rho(Y_{\text{com}})$, where Y_{com} follows the complete-data model $p(y_{\text{com}}|\theta)$.

The EM algorithm maximises $\log p(y_{\text{obs}}|\theta)$ by iteratively maximising the expected $\log p(y_{\text{com}}|\theta)$ using an expectation (E) step and a maximisation (M) step at each iteration. If $\theta^{(t)}$ is the estimate of θ at the tth iteration, then, at the $(t+1)$st iteration, the E step computes the expected loglikelihood of the complete-data model,

$$Q(\theta|\theta^{(t)}) = E_{Y_{\text{com}}}\{\log p(Y_{\text{com}}|\theta)|Y_{\text{obs}}, \theta^{(t)}\},$$

where $E_{Y_{\text{com}}}(.|Y_{\text{obs}}, \theta^{(t)})$ denotes the expectation with respect to the conditional distribution $p(y_{\text{com}}|Y_{\text{obs}}, \theta^{(t)})$. Then the M step finds $\theta^{(t+1)}$ by maximising $Q(\theta|\theta^{(t)})$ over θ. When $p(Y_{\text{com}}|\theta)$ is from an exponential family, the E step computes the conditional expectations of its sufficient statistics, and the M step fits the complete-data model using these expected statistics.

Each iteration of EM increases $\log p(Y_{\text{obs}}|\theta)$ because

$$\log p(Y_{\text{obs}}|\theta) = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)}),$$

where

$$H(\theta|\theta^{(t)}) = E_{Y_{\text{com}}}\{\log p(Y_{\text{com}}|Y_{\text{obs}}, \theta)|Y_{\text{obs}}, \theta^{(t)}\},$$

as a function of θ, is maximised at $\theta = \theta^{(t)}$.

3·2. *Formal definition of* PX-EM

In all existing EM-type algorithms, the complete-data model $p(y_{\text{com}}|\theta)$ and the observed-data model $p(y_{\text{obs}}|\theta)$ share the same set of parameters. The PX-EM algorithm expands $p(y_{\text{com}}|\theta)$ to a larger model, $p_X(y_{\text{com}}|\Theta)$, with $\Theta = (\theta_*, \alpha)$, where θ_* plays the same role

in $p_X(y_{com}|\Theta)$ that θ plays in $p(y_{com}|\theta)$, and α is the auxiliary parameter whose value is fixed at α_0 in the original model. Formally, two conditions must be satisfied. First, the observed-data model is preserved in the sense that, for all Θ, there is a common many-to-one reduction function R, such that $Y_{obs}|\Theta \sim p\{y_{obs}|\theta = R(\Theta)\}$. Secondly, the complete-data model is preserved at the null value of α, α_0, in the sense that, for all θ,

$$p_X\{y_{com}|\Theta = (\theta_*, \alpha_0)\} = p(y_{com}|\theta = \theta_*).$$

These conditions imply that if $\theta_1 \neq \theta_2$ then $\Theta_1 \neq \Theta_2$, and that, for all θ, there exists at least one Θ such that $Y_{obs}|\Theta \sim p\{y_{obs}|\theta = R(\Theta)\}$.

The PX-EM algorithm uses $p_X(y_{com}|\Theta)$ to generate EM by iteratively maximising the expected loglikelihood of $p_X(y_{com}|\Theta)$. Specifically, let $\Theta^{(t)} = (\theta^{(t)}, \alpha_0)$ be the estimate of Θ with $\alpha^{(t)} = \alpha_0$ from the tth iteration. Then, at the $(t+1)$st iteration we have the following.

PX-E *step.* Compute $Q_X(\Theta|\Theta^{(t)}) = E_{Y_{com}}\{\log p_X(Y_{com}|\Theta)|Y_{obs}, \Theta^{(t)}\}$.

PX-M *step.* Find $\Theta^{(t+1)} = \arg\max_\Theta Q_X(\Theta|\Theta^{(t)})$; then apply the reduction function $R(\theta)$ to obtain $\theta^{(t+1)} = R(\Theta^{(t+1)})$.

Each iteration of PX-EM increases $\log p(Y_{obs}|\theta)$ because, by the definition of parameter expansion, at $\theta = R(\Theta)$,

$$\log p(Y_{obs}|\theta) = Q_X(\Theta|\Theta^{(t)}) - H_X(\Theta|\Theta^{(t)}),$$

where

$$H_X(\Theta|\Theta^{(t)}) = E_{Y_{com}}\{\log p(Y_{com}|Y_{obs}, \Theta)|Y_{obs}, \Theta^{(t)}\},$$

as a function of Θ, is maximised at $\Theta^{(t)}$. A sequence of $\Theta^{(t)}$ generates a sequence of $\theta^{(t)} = R(\Theta^{(t)})$. Therefore we have the following theorem in parallel to the standard results for EM-type algorithms.

THEOREM 1. *The* PX-EM *algorithm increases the loglikelihood of the observed-data model at each iteration, that is* $\log p(Y_{obs}|\theta^{(t+1)}) \geqslant \log p(Y_{obs}|\theta^{(t)})$ *for all* t. *If* $\log p(Y_{obs}|\theta)$ *is bounded, then* $\log p(Y_{obs}|\theta^{(t)}) \to L^*$ *for some* L^*.

Since the monotone convergence of EM is maintained, conditions for the convergence of PX-EM iterates to a stationary point or a local maximum, θ_{MLE}, can be obtained following Wu (1983). As a result of the generality of the above results, parameter expansion can also be applied to the extensions and variations of the EM algorithm, such as those mentioned in § 1. For the multivariate t distribution, for example, the idea underlying ECME (Liu & Rubin, 1994a, b; 1995) can be applied to PX-EM to update α by maximising the observed-data loglikelihood, and thereby obtain an even faster converging algorithm. Moreover, Liu & Rubin (1995) described EM and ECME for the multivariate t with unknown degrees of freedom and missing values, which can be extended to PX-EM with the inclusion of the scale parameter α. Liu (1997) provides details of these extensions.

When the complete-data models are in the exponential family, the programming for implementing PX-EM beyond its parent EM algorithm only involves, first, adding to the original E step the calculation of the expected sufficient statistics related to the expanded parameterisation, which may have already been done in the original E step, and, secondly, modifying the maximisation over θ in the original M step to include maximisation over α, which is typically simple.

3·3. *Rate of convergence of* PX-EM

We now study the rate of convergence of PX-EM generated by $p(y_{\text{com}}|\Theta)$ as compared to the parent EM generated by $p(y_{\text{com}}|\theta)$. Since additional arithmetic operations introduced by parameter expansion are usually minor, the comparison of the rates of convergence typically corresponds realistically to the comparison of CPU times.

Each iteration of EM defines a mapping M, $\theta^{(t+1)} = M(\theta^{(t)})$, and a Taylor expansion at θ_{MLE} gives $\theta^{(t+1)} - \theta_{\text{MLE}} \doteq DM(\theta^{(t)} - \theta_{\text{MLE}})$, where DM is the gradient of M evaluated at θ_{MLE} and is called the matrix rate of convergence. The largest eigenvalue of DM, r, is known as the global rate of convergence. The speed matrix is $S = I - DM$, whose smallest eigenvalue $s = 1 - r$ is called the global speed.

As proved by Dempster et al. (1977), the speed matrix of EM is the matrix fraction of the observed information, $S = i_{\text{com}}^{-1} i_{\text{obs}}$, where

$$i_{\text{obs}} = -\left.\frac{\partial^2 \log p(Y_{\text{obs}}|\theta)}{\partial\theta \cdot \partial\theta'}\right|_{\theta=\theta_{\text{MLE}}}$$

is the observed-data information matrix, and

$$i_{\text{com}} = i_{\theta\theta} = -E_{Y_{\text{com}}}\left\{\left.\frac{\partial^2 \log p(Y_{\text{com}}|\theta)}{\partial\theta \cdot \partial\theta'}\right| Y_{\text{obs}}, \theta\right\}\Bigg|_{\theta=\theta_{\text{MLE}}}$$

is the complete-data information matrix.

It is easy to verify that $\Theta_{\text{MLE}} = (\theta_{\text{MLE}}, \alpha_0)$ is a fixed point of the mapping M_{X} induced by the PX-E and PX-M steps. Here we use the parameterisation $\Theta = (\theta, \alpha) = (R(\theta_*, \alpha), \alpha)$ so that θ is the parameter being estimated by the output of each PX-EM iteration. A Taylor expansion of M_{X} at Θ_{MLE} gives $\Theta^{(t+1)} - \Theta_{\text{MLE}} \doteq DM_{\text{X}}(\Theta^{(t)} - \Theta_{\text{MLE}})$, where $\Theta^{(t)} = (\theta^{(t)}, \alpha_0)$. Similarly to the result of Dempster et al. (1977), it can be shown that, under mild conditions including $I_{\text{com}} > 0$, $DM_{\text{X}} = I - I_{\text{com}}^{-1} I_{\text{obs}}$, where I is the identity matrix of appropriate dimension, and I_{obs} and I_{com} are the observed-data information matrix and the complete-data information matrix for $p_{\text{X}}(y_{\text{com}}|\theta, \alpha)$, respectively, i.e.

$$I_{\text{obs}} = \begin{pmatrix} i_{\text{obs}} & 0 \\ 0 & 0 \end{pmatrix}, \quad I_{\text{com}} = \begin{pmatrix} i_{\text{com}} & i_{\theta\alpha} \\ i_{\alpha\theta} & i_{\alpha\alpha} \end{pmatrix},$$

where

$$i_{\theta\alpha} = -E_{Y_{\text{com}}}\left\{\left.\frac{\partial^2 \log p_{\text{X}}(Y_{\text{com}}|\theta, \alpha)}{\partial\theta \cdot \partial\alpha'}\right| Y_{\text{obs}}, \theta, \alpha\right\}\Bigg|_{(\theta_{\text{MLE}}, \alpha_0)},$$

$i_{\alpha\theta} = i'_{\theta\alpha}$, and

$$i_{\alpha\alpha} = -E_{Y_{\text{com}}}\left\{\left.\frac{\partial^2 \log p_{\text{X}}(Y_{\text{com}}|\theta, \alpha)}{\partial\alpha \cdot \partial\alpha'}\right| Y_{\text{obs}}, \theta, \alpha\right\}\Bigg|_{(\theta_{\text{MLE}}, \alpha_0)}.$$

For notational convenience, also let $v_{\text{com}} = i_{\text{com}}^{-1} = i_{\theta\theta}^{-1}$ and $V_{\text{com}} = I_{\text{com}}^{-1}$ with

$$V_{\text{com}} = \begin{pmatrix} V_{\theta\theta} & V_{\theta\alpha} \\ V_{\alpha\theta} & V_{\alpha\alpha} \end{pmatrix} = \begin{pmatrix} i_{\theta\theta} & i_{\theta\alpha} \\ i_{\alpha\theta} & i_{\alpha\alpha} \end{pmatrix}^{-1},$$

so that

$$V_{\theta\theta} = v_{\text{com}} + V_{\theta\alpha} V_{\alpha\alpha}^{-1} V_{\alpha\theta}. \tag{9}$$

Therefore, we have

$$DM_X = I - \begin{pmatrix} V_{\theta\theta} i_{obs} & 0 \\ V_{\alpha\theta} i_{obs} & 0 \end{pmatrix} = \begin{pmatrix} I - V_{\theta\theta} i_{obs} & 0 \\ -V_{\alpha\theta} i_{obs} & I \end{pmatrix},$$

whence $\theta^{(t+1)} - \theta_{MLE} \doteq (I - V_{\theta\theta} i_{obs})(\theta^{(t)} - \theta_{MLE})$ and, because $\alpha^{(t)} = \alpha_0$,

$$\alpha^{(t+1)} - \alpha_0 \doteq -V_{\alpha\theta} i_{obs}(\theta^{(t)} - \theta_{MLE}). \tag{10}$$

Thus, the convergence of $\theta^{(t)}$ determines the convergence of PX-EM, and Θ has $S_X = V_{\theta\theta} i_{obs}$ as its speed matrix. For EM, the speed matrix is $S = v_{com} i_{obs}$. Since $V_{\theta\theta} \geq v_{com}$ in semipositive definite order, the smallest eigenvalue s_X of S_X is at least as large as the smallest eigenvalue s of S. Thus, we have Theorem 2.

THEOREM 2. *Given that* PX-EM *converges to* (θ_{MLE}, α_0), *and the derivatives and inverses used in the above derivations exist,* $s_X \geq s$, *that is* PX-EM *dominates* EM *in global rate of convergence.*

From a Bayesian perspective, the fraction of missing information (Rubin, 1987, pp. 93–4) is the variance of the parameter θ given the complete data relative to the variance of θ given the observed data: $V_{\theta\theta}$ is the complete-data variance of θ in PX-EM, which is larger than v_{com}, the complete-data variance of θ in the parent EM. The difference between them, $V_{\theta\alpha} V_{\alpha\alpha}^{-1} V_{\alpha\theta}$, is the extra variance due to the auxiliary parameter α, which reduces the fraction of missing information and makes PX-EM faster than EM.

For an original complete-data model, there can be various expansions that lead to different PX-EM implementations. The above derivations suggest that, to generate the fastest PX-EM, we should expand the complete-data model as much as we can, provided that the extra computational cost is negligible. This conclusion is easy to understand in light of covariance adjustment since, the more covariates we adjust for, the more efficient is our analysis, at least in large samples.

3·4. *The* PX-M *step as covariance adjustment*

Let $\theta^{(t)}$ be the current estimate of θ with $\alpha^{(t)} = \alpha_0$, and let $\theta_{EM}^{(t+1)}$ and $\theta_X^{(t+1)}$ be the estimates of θ updated by EM and PX-EM, respectively:

$$\theta_{EM}^{(t+1)} - \theta_{MLE} \doteq (I - v_{com} i_{obs})(\theta^{(t)} - \theta_{MLE}), \quad \theta_X^{(t+1)} - \theta_{MLE} \doteq (I - V_{com} i_{obs})(\theta^{(t)} - \theta_{MLE}).$$

Thus, from (9), $\theta_X^{(t+1)} - \theta_{EM}^{(t+1)} \doteq -V_{\theta\alpha} V_{\alpha\alpha}^{-1} V_{\alpha\theta} i_{obs}(\theta^{(t)} - \theta_{MLE})$, which, from (10), can be written as

$$\theta_X^{(t+1)} \doteq \theta_{EM}^{(t+1)} + V_{\theta\alpha} V_{\alpha\alpha}^{-1}(\alpha^{(t+1)} - \alpha_0). \tag{11}$$

Expression (11) justifies an explicit interpretation of PX-EM as covariance adjustment. The E-steps of both EM and PX-EM effectively impute missing data under the wrong model with $\theta^{(t)} \neq \theta_{MLE}$. The M-step of EM ignores this effect, whereas the M-step of PX-EM uses the extra parameter α as a covariate: it uses the difference between the imputed value of α and its true value, $\alpha^{(t+1)} - \alpha_0$, and the regression coefficient of θ on α, $V_{\theta\alpha} V_{\alpha\alpha}^{-1}$, to correct the unadjusted estimate of θ, $\theta_{EM}^{(t+1)}$, to produce the adjusted estimate $\theta_X^{(t+1)}$.

More explicitly, covariance adjustment can be viewed in general as the formulation of an adjusted estimate of θ, $\hat{\theta}_X$, created by adding to a naive estimate of θ, $\hat{\theta}_0$, obtained without the use of the covariate, an adjustment orthogonal to $\hat{\theta}_0$. The naive estimate, $\hat{\theta}_0$, can be viewed as the estimate of θ obtained when the covariate parameter, α, is fixed at

its known true value, α_0. The adjustment to $\hat{\theta}_0$ is the scaled difference between (i) the estimate, $\hat{\alpha}_X$, of the covariate parameter, α, obtained jointly with $\hat{\theta}_X$ and (ii) its true value, α_0. The resulting adjusted estimate is

$$\hat{\theta}_X = \hat{\theta}_0 + B(\hat{\alpha}_X - \alpha_0). \tag{12}$$

Since the naive estimate of θ and the adjustment are orthogonal, that is $\text{cov}\{\hat{\theta}_0, B(\hat{\alpha}_X - \alpha_0)\} = 0$, treating B as fixed in (12) gives $\text{cov}\{\hat{\theta}_X, (\hat{\alpha}_X - \alpha_0)\} = B\,\text{var}(\hat{\alpha}_X - \alpha_0)$. If we let $\hat{\theta}_X = \theta_X^{(t+1)}$, $\hat{\theta}_0 = \theta_{EM}^{(t+1)}$ and $\hat{\alpha}_X = \alpha^{(t+1)}$, (12) is identical to (11).

We illustrate the above results with a simple example. Consider the EM generated by the complete-data model $Y_{obs}|(Y_{mis}, \theta) \sim N(Y_{mis}, 1)$ and $Y_{mis}|\theta \sim N(\theta, \sigma^2)$, where θ is the unknown parameter and $\sigma^2 > 0$ is known. Consider the PX-EM generated by the expanded model $Y_{obs}|(Y_{mis}, \theta_*, \alpha) \sim N(Y_{mis} + \alpha, 1)$ and $Y_{mis}|(\theta_*, \alpha) \sim N(\theta_*, \sigma^2)$, where $\Theta = (\theta_*, \alpha)$ is the expanded parameter with $\alpha_0 = 0$ and the reduction function $\theta = R(\Theta) = \theta_* + \alpha$:

$$Y_{obs}|\theta \sim N(\theta, 1 + \sigma^2), \quad Y_{obs}|\Theta \sim N(\theta_* + \alpha, 1 + \sigma^2). \tag{13}$$

The maximum likelihood estimate of θ is $\theta_{MLE} = Y_{obs}$. Given $\theta^{(t)}$, the E-steps of both EM and PX-EM impute $\hat{Y}_{mis} = (\theta^{(t)} + \sigma^2 Y_{obs})/(1 + \sigma^2)$; the M-step of EM gives $\theta_{EM}^{(t+1)} = \hat{Y}_{mis}$, that is $\theta_{EM}^{(t+1)} - \theta_{MLE} = (\theta^{(t)} - Y_{obs})/(1 + \sigma^2)$, whereas the M-step of PX-EM gives $\theta_{EM}^{(t+1)} = \hat{Y}_{mis}$, that is $\alpha^{(t+1)} = Y_{obs} - \hat{Y}_{mis} = (Y_{obs} - \theta^{(t)})/(1 + \sigma^2)$. Thus, with reduction, $\theta_X^{(t+1)} = \theta_{EM}^{(t+1)} + (\alpha^{(t+1)} - \alpha_0) = \theta_{MLE}$. When $\sigma^2 \doteq 0$, EM is extremely slow, whereas PX-EM converges in one iteration for all $\sigma^2 > 0$. To verify (11), we have

$$\begin{pmatrix} \text{var}(\theta_{EM}^{(t+1)}) & \text{cov}(\theta_{EM}^{(t+1)}, \alpha^{(t+1)}) \\ \text{cov}(\alpha^{(t+1)}, \theta_{EM}^{(t+1)}) & \text{var}(\alpha^{(t+1)}) \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} V_{\theta\theta} & V_{\theta\alpha} \\ V_{\alpha\theta} & V_{\alpha\alpha} \end{pmatrix} = \begin{pmatrix} \sigma^2 + 1 & 1 \\ 1 & 1 \end{pmatrix},$$

which gives $V_{\theta\alpha} V_{\alpha\alpha}^{-1} = 1$, where we note that the naive estimate of θ and the estimate of α are orthogonal, as required for the covariance-adjustment interpretation.

4. MORE EXAMPLES

4·1. *Random effects model*

The random effects, i.e. variance components or repeated measures or mixed, model is extremely useful in applied statistics. Consider the following general linear mixed model, e.g. Hartley & Rao (1967), Laird & Ware (1982), Laird et al. (1987).

Model O:
$$Y_i|\theta, b_i \sim X_i'\beta + Z_i' b_i + e_i, \tag{14}$$

$$b_i|\theta \sim N_q(0, \Psi), \quad e_i|\theta \sim N(0, \sigma^2), \quad b_i \perp e_i, \tag{15}$$

for $i = 1, \ldots, N$, where $Y_{obs} = (Y_1, \ldots, Y_N)$ are the observed scalar responses, X_i $(p \times 1)$ and $Z_i = (Z_{i1}, \ldots, Z_{iq})'$ $(q \times 1)$ are known covariates, β $(p \times 1)$ are the fixed effects, $b_i = (b_{i1}, \ldots, b_{iq})'$ $(q \times 1)$ are the random effects, $b = (b_1, \ldots, b_N)$ and $\theta = (\beta, \Psi, \sigma^2)$. The joint distribution of (Y_i, b_i) is

$$\begin{bmatrix} Y_i \\ b_i \end{bmatrix} \sim N_{1+q}\left(\begin{bmatrix} X_i'\beta \\ 0 \end{bmatrix}, \begin{bmatrix} Z_i'\Psi Z_i + \sigma^2 & Z_i'\Psi \\ \Psi Z_i & \Psi \end{bmatrix}\right), \tag{16}$$

where the row for Y_i is the observed-data model.

The EM algorithm generated by Model O is as follows.

E *step.* Impute the random effects and their cross-products by regression of b_i on Y_i,

$$b_i^{(t+1)} = E(b_i | Y_{\text{obs}}, \theta^{(t)}) = \frac{Y_i - X_i'\beta^{(t)}}{(\sigma^2)^{(t)} + Z_i'\Psi^{(t)}Z_i}\Psi^{(t)}Z_i, \tag{17}$$

$$\Psi_i^{(t+1)} = E(b_i b_i' | Y_{\text{obs}}, \theta^{(t)}) = b_i^{(t+1)}(b_i^{(t+1)})' + \Psi^{(t)} - \frac{\Psi^{(t)}Z_i Z_i'\Psi^{(t)}}{(\sigma^2)^{(t)} + Z_i'\Psi^{(t)}Z_i}. \tag{18}$$

M *step.* Let $D_i = (X_i', Y_i - Z_i'b_i^{(t+1)})$ be the current data, i.e. covariates and working response deviations for the ith observation, and let $C = \sum_{i=1}^{N} E(D_i'D_i | \theta^{(t)}, Y_{\text{obs}})$ be the expected cross-product matrix, which can be calculated according to the results in the E step. Then $(\beta^{(t+1)}, (\sigma^2)^{(t+1)})$ can be found from the last column or row of SWEEP$[1, \ldots, p]C$, where SWEEP is the sweep operator, e.g. Little & Rubin (1987, pp. 112–9). Also, update Ψ with

$$\Psi^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N}\Psi_i^{(t+1)}. \tag{19}$$

Since this implementation is often criticised for its slow convergence, we consider parameter expansion for acceleration. Note that, in the above M step, the updated parameter $\Psi^{(t+1)}$ is the average of the set of imputed $b_i b_i'$ in (18), and the model covariance between Y_i and b_i is fixed at $\Psi^{(t+1)}Z_i$. This covariance does not in general reflect the relationship between the imputed random effects $b_i^{(t+1)}$ and the known Y_i. To adjust for the deviations, we expand Model O to

Model X:
$$Y_i | \Theta,\ b_i \sim X_i'\beta_* + Z_i'\alpha b_i + e_i = X_i'\beta_* + \sum_{j=1}^{q}\sum_{k=1}^{q}\alpha_{jk}Z_{ij}b_{ik} + e_i, \tag{20}$$

$$b_i | \Theta \sim N_q(0, \Psi_*), \quad e_i | \Theta \sim N(0, \sigma_*^2), \quad b_i \perp e_i, \tag{21}$$

with parameter $\Theta = (\beta_*, \sigma_*^2, \Psi_*, \alpha)$; the value of α in the original model is the $(q \times q)$ identity matrix I. Under Model X, we have $Y_i | \Theta \sim X_i'\beta_* + N(0, Z_i'\alpha\Psi_*\alpha'Z_i + \sigma^2)$, so θ is identified as $\theta = (\beta, \sigma^2, \Psi) = R(\Theta) = (\beta_*, \sigma_*^2, \alpha\Psi_*\alpha')$. The PX-EM algorithm is as follows.

PX-E *step.* This is unchanged from EM.

PX-M *step.* Write (20) in Model X in the following standard linear regression form:

$$Y_i \sim X_i'\beta_* + \{\text{vec}(Z_i b_i')\}'\,\text{vec}(\alpha) + e_i.$$

Let $D_i = [X_i', \{\text{vec}(Z_i b_i')\}', Y_i]$ be the current data for the ith observation, and let $C = \sum_i E(D_i'D_i | \Theta^{(t)}, Y_{\text{obs}})$ be the expected cross-product matrix, which can be calculated according to the results in the E step. Then $\beta_*^{(t+1)}$, $\text{vec}(\alpha^{(t+1)})$ and $(\sigma_*^2)^{(t+1)}$ can be found from the last column of SWEEP$[1, \ldots, p+q^2]C$. Also update Ψ_* as in (19). Reduction to the original parameters gives $\beta^{(t+1)} = \beta_*^{(t+1)}$, $(\sigma^2)^{(t+1)} = (\sigma_*^2)^{(t+1)}$ and $\Psi^{(t+1)} = \alpha^{(t+1)}\Psi_*^{(t+1)}[\alpha^{(t+1)}]'$.

Parameter expansion uses α to collect the information in the extra regression structure, uses Ψ_* to collect the information in the sample covariance matrix of the random effects,

and then combines the two pieces of information to find the new estimate for Ψ. As suggested by the numerical study to be described, when the residual variance σ^2 is much larger than the random effects variability $\sum Z_i' \Psi Z_i / N$, the original EM can be extremely slow, but PX-EM can be very fast.

Alternatively, Meng & van Dyk (1998) propose the following efficient augmentation scheme, where the $c_i = \mathrm{Chol}(\Psi) b_i$ are used as missing data, Chol(.) denotes the Cholesky decomposition, and the unexpanded complete-data model is

$$\text{Model U:} \qquad Y_i | \theta, c_i \sim X_i' \beta + Z_i \,\mathrm{Chol}(\Psi) c_i + N(0, \sigma^2), \tag{22}$$

$$c_i | \theta \sim N_q(0, I), \quad e_i | \theta \sim N(0, \sigma^2), \quad c_i \perp e_i. \tag{23}$$

They report that the EM implementation generated by this model usually, but not always, converges faster than the one generated by Model O, and the gain can be dramatic when the residual variance σ^2 is large. Our PX-EM scheme can also be applied to Model U by adding an auxiliary covariance matrix parameter in $c_i | \theta \sim N_q(0, I)$ and auxiliary parameters for the components in the upper triangular part of $\mathrm{Chol}(\Psi)$. The resulting PX-EM is identical to the implementation generated by Model X. Hence, we expect our PX-EM to be superior to EM on either Model O or Model U.

Consider the following simple numerical example, where we simulate data with $N = 100$, $p = 2$, $q = 2$, $X_i = (1, i)'$, $\beta' = (0, 0)$, $z_i \sim N(0, I_2)$, $\Psi = I$ and various values for σ^2. The starting point is chosen at $\beta' = (0, 0)$, $\Psi = I_2$ and $\sigma^2 = 1{\cdot}0$, and the convergence criterion is $\| \theta^{(t+1)} - \theta^{(t)} \|_\infty < 10^{-6}$. Table 1 displays the comparison, in terms of the numbers of iterations required to converge for various values of σ^2, for the EM generated by Model O, $\mathrm{EM_O}$, the EM generated by Model U, $\mathrm{EM_U}$, and the EM generated by Model X, PX-EM. We report one dataset for each σ^2; based on our limited experience, the results change little for different simulations with the same σ^2 or for different starting values.

Table 1. *Comparison of numbers of iterations to convergence for three algorithms*

	$\sigma^2 = \frac{1}{25}$	$\sigma^2 = \frac{1}{4}$	$\sigma^2 = \frac{1}{2}$	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 25$
$\mathrm{EM_O}$	197	136	243	830	6657	12935
$\mathrm{EM_U}$	260	173	239	452	547	466
PX-EM	190	88	59	140	140	93

Table 1 shows that, in these simulations, PX-EM is faster than both $\mathrm{EM_U}$ and $\mathrm{EM_O}$. When σ^2 is small, PX-EM behaves like $\mathrm{EM_O}$, but $\mathrm{EM_U}$ is slower than both $\mathrm{EM_O}$ and PX-EM. When σ^2 is large, PX-EM and $\mathrm{EM_U}$ are much faster than $\mathrm{EM_O}$.

The extension to multivariate responses is straightforward. The method of Laird et al. (1987) for updating β, which is also discussed by Liu & Rubin (1994a, b), can be incorporated. Also, ECME can be incorporated with an M step that maximises the constrained actual likelihood over σ^2. As is noticed by J. Schafer in an unpublished report, see also Lindstrom & Bates (1988), this M step has a closed-form solution with the reparameterisation $\Psi = \sigma^2 \Phi$. Some components in the mean parameter of $b_i | \theta \sim N_q(0, \Psi)$ can also be activated depending on $\{X_i, Z_i\}$; such over-parameterisation is related to the reparameterisation introduced by Gelfand, Sahu & Carlin (1995).

4·2. *Factor analysis*

Factor analysis is a standard tool in multivariate analysis. It can be viewed as the normal linear regression analysis of an observed p-dimensional random variable Y on an

unobserved variable Z consisting of $q < p$ factors that are themselves normal; the key assumption allowing estimation despite all Z being missing is that the components of Y are conditionally independent given Z. To be more specific, let $Y_{\text{obs}} = \{Y_1, \ldots, Y_N\}$ be the observed data, and let $Z = \{Z_1, \ldots, Z_N\}$ be the unknown factors; then the complete-data model is

Model O:
$$Y_i | Z_i, \theta \sim N_p(\beta Z_i, \Sigma),$$
$$Z_i | \theta \sim N_q(0, I),$$

for $i = 1, \ldots, N$, where β $(p \times q)$ is called the factor-loading matrix, I is the $(q \times q)$ identity matrix, $\Sigma = \text{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$ is called the uniquenesses matrix, and $\theta = (\beta, \Sigma)$. Under Model O, the observed-data model, after integrating out the unobserved factors $\{Z_i, i = 1, \ldots, N\}$, is $Y_i | \beta, \Sigma \sim N_p(0, \beta\beta' + \Sigma)$. The following is a description of the EM algorithm generated by Model O; also see Rubin & Thayer (1982). Let

$$C_{yy} = \sum_{i=1}^{N} (Y_i - \bar{Y})(Y_i - \bar{Y})'/N, \quad C_{yz} = \sum_{i=1}^{N} (Y_i - \bar{Y})Z_i'/N, \quad C_{zz} = \sum_{i=1}^{N} Z_i Z_i'/N$$

be the three sufficient statistics for the two parameters β and Σ.

E *step.* Calculate the expected sufficient statistics

$$C_{yz}^{(t+1)} = E(C_{yz} | Y_{\text{obs}}, \theta) = C_{yy}\gamma, \quad C_{zz}^{(t+1)} = E(C_{zz} | Y_{\text{obs}}, \theta) = \gamma' C_{yy} \gamma + \Delta,$$

where γ and Δ are the regression coefficients and the residual covariance matrix of Z on Y given $\theta^{(t)}$. More precisely, let

$$B = \begin{pmatrix} \beta^{(t)'}\beta^{(t)} + \Sigma^{(t)} & \beta^{(t)'} \\ \beta^{(t)} & I \end{pmatrix}$$

be the current variance-covariance matrix of (Y, Z); then γ and Δ are obtained from the last q columns of SWEEP$[1, \ldots, p]B$.

M *step.* Define the cross-product matrix

$$C = \begin{pmatrix} C_{yy} & C_{yz}^{(t+1)} \\ C_{yz}^{(t+1)'} & C_{zz}^{(t+1)} \end{pmatrix}.$$

Then $\beta^{(t+1)}$ and $\Sigma^{(t+1)}$ are obtained from the last q columns of SWEEP$[1, \ldots, p]C$.

Model O can be expanded to

Model X:
$$Y_i | Z_i, \Theta \sim N_p(\beta_* Z_i, \Sigma_*),$$
$$Z_i | \Theta \sim N_q(0, \alpha),$$

where $\Theta = (\beta_*, \sigma_*, \alpha)$, with α being the auxiliary parameter, which has C_{zz} as its natural sufficient statistic. Under Model X, $Y_i | \Theta \sim N(0, \beta_* \alpha \beta_*' + \Sigma_*)$, so

$$\theta = (\beta, \Sigma) = (\beta_* \text{ Chol}(\alpha), \Sigma_*).$$

The PX-EM algorithm is as follows.

PX-E *step.* This is unchanged from EM.

PX-M *step.* The computations for $\beta_*^{(t+1)}$ and $\Sigma_*^{(t+1)}$ are the same as those for $\beta^{(t+1)}$ and $(\sigma^2)^{(t+1)}$ in the M step of EM, and $\alpha^{(t+1)} = C_{zz}^{(t+1)}$. Reduction to the original parameters gives $\Sigma^{(t+1)} = \Sigma_*^{(t+1)}$, $\beta^{(t+1)} = \beta_*^{(t+1)} \text{ Chol}(\alpha^{(t+1)})$.

The idea underlying the PX-M step is to adjust the estimation for the deviations between $C_{zz}^{(t+1)}$ and its expectation under Model O, I. Model X and its EM implementation were also considered by Rubin & Thayer (1982), but without the benefit of our general perspective.

In a simple simulation study, we chose $p = 2$, $q = 1$, with $\Sigma = I$ and various β. We started both the original EM and PX-EM from the true values of the parameters, with the convergence criterion being $\|\theta^{(t+1)} - \theta^{(t)}\|_\infty < 10^{-10}$. When $\beta = (1, 1)'$, both algorithms converged very quickly; 26 iterations for PX-EM and 37 iterations for the original EM. When $\beta = (10, 10)'$, PX-EM converged in 26 iterations, but the original EM took 1765 iterations. When $\beta = (100, 100)'$, PX-EM still converged in only 26 iterations, but the original EM took 151 322 iterations. This suggests that the gain can be very large when the magnitude of the factor loading matrix β is large relative to the magnitude of the uniqueness matrix Σ, which is also confirmed by calculations for some simple cases using results in § 3·3. The PX-EM algorithm can also be applied to accelerate the ECME algorithm for factor analysis with missing data (Liu & Rubin, 1998).

4·3. *Probit regression*

Let $Y_{\text{obs}} = \{Y_1, \ldots, Y_N\}$ be the set of observed 0/1 random variables, which follow the model $Y_i | \theta \sim \text{Bernoulli}\{\Phi(X_i'\theta)\}$, independently, where X_i ($p \times 1$) are the covariates, and Φ is the standard normal cumulative distribution function. In the conventional EM implementation, Y_{obs} is augmented to $Y_{\text{com}} = \{(Y_1, Z_1), \ldots, (Y_N, Z_N)\}$, which follows

Model C:
$$Y_i = \text{sgn}(Z_i), \tag{24}$$
$$Z_i | \theta \sim X_i'\theta + N(0, 1), \tag{25}$$

where $\text{sgn}(z) = 1$ if $z \geqslant 0$, and $\text{sgn}(z) = 0$ otherwise.

The following is a description of the EM implementation generated by Model O.

E *step.* Impute Z_i according to a truncated normal;

$$Z_i^{(t+1)} = E(Z_i | Y_{\text{obs}}, \theta^{(t)}) = \begin{cases} X_i'\theta^{(t)} + \phi(X_i'\theta^{(t)})/\{1 - \Phi(-X_i'\theta^{(t)})\} & \text{if } Y_i = 1, \\ X_i'\theta^{(t)} - \phi(X_i'\theta^{(t)})/\Phi(-X_i'\theta^{(t)}) & \text{if } Y_i = 0, \end{cases}$$

where ϕ is the probability density function of the standard normal distribution.

M *step.* Regress $\{Z_i^{(t+1)}, i = 1, \ldots, N\}$ on $\{X_i, i = 1, \ldots, N\}$ to obtain $\theta^{(t+1)}$, where the regression can be accomplished by SWEEP.

Note that only the imputed first-moment statistic $\sum Z_i^{(t+1)}$ is used in EM, because the residual variance in (25) in Model O is a hidden parameter frozen at 1, but we can activate the variance structure by the following expanded model:

Model X:
$$Y_i = \text{sgn}(Z_i),$$
$$Z_i | \Theta \sim X_i'\theta_* + N(0, \alpha^2),$$

where $\Theta = (\theta_*, \alpha^2)$, and the second-moment statistic becomes a sufficient statistic of Model X. Under Model X, the observed data follow $Z_i | \Theta \sim \text{Bernoulli}\{\Phi(X_i'\theta_*/\alpha)\}$, so $\theta = \theta_*/\alpha$. The PX-EM algorithm is as follows.

PX-E *step.* Compute $Z_i^{(t+1)} = E\{Z_i | Y_{\text{obs}}, \Theta^{(t)} = (\theta^{(t)}, \alpha_0)\}$ in the same way as in the E step of the conventional EM described above, and

$$E(Z_i^2 | Y_{\text{obs}}, \Theta^{(t)}) = (Z_i^{(t+1)})^2 + 1 - X_i'\theta^{(t)}(Z_i^{(t+1)} - X_i'\theta^{(t)}).$$

PX-M *step.* Let $D_i = (X_i', Z_i)$, and compute the cross-product matrix of the data D_i, $C = \sum_i E(D_i' D_i | Y_{obs}, \Theta^{(t)})$ according to the results in the E step. Then $\theta_*^{(t+1)}$ and $(\alpha^2)^{(t+1)}$ are obtained from the last column of SWEEP$[1, \ldots, p]C$. Reduction to the original parameter gives $\theta^{(t+1)} = \theta_*^{(t+1)}/\alpha^{(t+1)}$.

We use the Kyphosis data in S (Chambers & Hastie, 1992, p. 200) for a numerical example, where the outcome is the presence or absence of a postoperative deformity and there are three continuous covariates, Age, Number and Start. We fit a probit model to this dataset using the original EM and PX-EM starting from all coefficients equal to zero and with convergence criterion being $\|\theta^{(t+1)} - \theta^{(t)}\|_\infty < 10^{-10}$. It takes the original EM 106 iterations to converge but PX-EM only 63.

Calculations for simple cases suggest that the gain in the rate of convergence for PX-EM can be significant when the β coefficients are large relative to $(X'X)^{-1}$. We therefore simulate a modified dataset, where the covariates remain unchanged but the outcome variable is simulated by setting the coefficients for intercept, Age, Number and Start at 0, 0, 3, -1 respectively. We use 3 and -1 for the coefficients of Number and Start because both Number and Start are positive, with the magnitude of the latter three times the magnitude of the former. The original EM took 173 988 iterations to converge, whereas PX-EM took only 2301 iterations.

4·4. *Poisson imaging model*

The EM algorithm has become an important computational method in image reconstruction problems since Shepp & Vardi (1982), Vardi, Shepp & Kaufman (1985) and Lange & Carson (1984). Recently, Fessler & Hero (1994) proposed a method to accelerate EM significantly using efficient data augmentation. The following is a version of the simplified model.

Let $Y_{obs} = \{Y_1, \ldots, Y_N\}$ be the observed data, which follow the model $Y_i | \theta \sim \text{Po}(k_i\theta + r_i)$, independently, for $i = 1, \ldots, N$, where θ is an unknown positive scalar, and $\{k_1, \ldots, k_N\}$ and $\{r_1, \ldots, r_N\}$ are known positive constants. The conventional EM algorithm augments Y_{obs} to $Y_{com} = \{(Z_1, R_1), \ldots, (Z_N, R_N)\}$, where $Y_i = Z_i + R_i$, and the complete-data model is

Model O:
$$Y_i = Z_i + R_i, \quad Z_i \perp R_i,$$
$$Z_i | \theta \sim \text{Po}(k_i\theta), \quad R_i | \theta \sim \text{Po}(r_i),$$

where Z_i can be considered as signal and R_i as residual noise.

The EM generated by Model O can often be very slow. For acceleration, Fessler & Hero (1994) propose the following set of unexpanded complete-data models indexed by a_0.

Model U(a_0):
$$Y_i = Z_i + R_i, \quad Z_i \perp R_i$$
$$Z_i | \theta \sim \text{Po}\{k_i(\theta + a_0)\}, \quad R_i | \theta \sim \text{Po}(r_i - k_i a_0),$$

where a_0 is a positive constant in $[0, \min\{r_i/k_i,\}]$. For fixed a_0, Model U(a_0) leads to the following EM implementation.

E *step.* For $i = 1, \ldots, N$,
$$Z_i^{(t+1)} = \frac{k_i(\theta^{(t)} + a_0)}{k_i\theta^{(t)} + r_i} Y_i.$$

M *step.* Fit Model U(a_0) under the constraint that θ is positive:
$$\theta^{(t+1)} = \max\left(\frac{\sum_{i=1}^N Z_i^{(t+1)}}{\sum_{i=1}^N k_i} - a_0, 0\right).$$

The conventional EM corresponds to $a_0 = 0$. Fessler & Hero (1994) show that the rate of convergence of EM generated by Model U(a_0) is a monotone function of a_0, which achieves its maximum at the largest possible value of a_0, $\min\{r_i/k_i\}$.

The EM generated by Model U(a_0) can be understood as a PX-EM. First, we rewrite Model O as

Model O.1:
$$Y_i = Z_i + U_i + R_i, \quad Z_i \perp U_i \perp R_i,$$
$$Z_i|\theta \sim \text{Po}(k_i\theta), \quad U_i|\theta \sim \text{Po}(k_i a_0), \quad R_i|\theta \sim \text{Po}(r_i - a_0 k_i),$$

which is equivalent to Model O in the sense that they lead to identical EM implementations. However, Model O.1 splits a piece U_i from the original residual R_i in Model O, thus allowing us to expand Model O.1 by activating the hidden fixed parameter a_0 in the model structure for U_i:

Model X:
$$Y_i = Z_i + U_i + R_i, \quad Z_i \perp U_i \perp R_i,$$
$$Z_i|\Theta \sim \text{Po}(k_i\theta_*), \quad U_i|\Theta \sim \text{Po}(k_i\alpha), \quad R_i|\Theta \sim \text{Po}(r_i - a_0 k_i),$$

where $\Theta = (\theta_*, \alpha)$; α is the hidden parameter in the original model with sufficient statistic $\sum_i U_i$, where $\theta = \theta_* + \alpha - a_0$. We therefore have the following PX-EM.

PX-E *step.* For $i = 1, \ldots, N$,

$$Z_i^{(t+1)} = \frac{k_i\theta^{(t)}}{k_i\theta^{(t)} + r_i} Y_i, \quad U_i^{(t+1)} = \frac{a_0\theta^{(t)}}{k_i\theta^{(t)} + r_i} Y_i.$$

PX-M *step.* Let

$$\bar{Z} = \frac{\sum_{i=1}^N Z_i^{(t+1)}}{\sum_{i=1}^N k_i}, \quad \bar{U} = \frac{\sum_{i=1}^N U_i^{(t+1)}}{\sum_{i=1}^N k_i}.$$

Then we have that $(\theta_*^{(t+1)}, \alpha^{(t+1)}) = (\bar{Z}, \bar{U})$ if $\bar{Z} + \bar{U} - a_0 \geq 0$; otherwise $(\theta_*^{(t+1)}, \alpha^{(t+1)})$ lies on the line $\theta_* + \alpha - a_0 = 0$. The reduction to the original parameter gives $\theta^{(t+1)} = \theta_*^{(t+1)} + \alpha^{(t+1)} - a_0$.

It is easy to verify that this implementation is equivalent to the one generated by Model U(a_0). Statistically, PX-EM makes use of an auxiliary statistic $\sum_i U_i$ out of the residuals, and adjusts for the deviation between $\sum_i U_i^{(t+1)}/\sum_i k_i$ and its expectation under Model O.1, a_0. If $a_0 < \min\{r_i/k_i\}$, there is still room for us to split one more piece V_i besides U_i from the residual R_i for more adjustment. Therefore, the optimal a_0 is $\min\{r_i/k_i\}$.

5. DISCUSSION

When the observed data are augmented to the complete data, the estimation of the original parameters in the M step is simpler than maximum likelihood estimation from the observed data. Moreover, this augmentation often provides extra information that can be used for adjustment to improve the M step estimation. The basic machinery of PX-EM expands the complete-data model to allow efficient use of the imputed data to find the maximum likelihood estimate in the observed-data model. The PX-EM algorithm is simple to design and also simple to program, and maintains the stable convergence of all EM-type algorithms, often with significantly accelerated speed.

The PX-EM algorithm is not, however, without its limitations. In some EM implementations, the best expanded model can be the original EM model itself. For instance, we have yet to find a parameter expansion scheme for mixture models, e.g. Titterington, Smith & Makov (1985), which is an important application of EM. Also, finding the parameter expansion scheme is still a matter of art, just like implementing EM itself. A vague guideline is to look for hidden mean parameters and scale parameters, but we are unable to provide a general rule.

Philosophically, our parameter expansion and Meng & van Dyk's (1997) efficient augmentation are two complementary ideas in the art of implementing EM-type algorithms: PX-EM works along the analysis dimension, in the sense that the altered maximisation takes place at each M step, whereas efficient augmentation works along the design dimension in the sense that an a priori maximisation is used to design a more efficient EM. Technically, however, efficient augmentation can often be viewed as a special case of PX-EM, where the PX-M step is performed under the constraint $\alpha = f_a(\theta)$, for some function f indexed by a working parameter a, with both f and a chosen before the algorithm starts. That is, the reparameterisation in efficient augmentation can often be derived from the over-parameterisation in PX-EM. Finding optimal f and a is impossible in general because the fraction of missing information can depend on the unknown maximum likelihood estimate. In fact, according to the result in § 3·3, the unconstrained PX-M step can always lead to a faster algorithm than fixing a at any value.

Since the Gibbs sampler and the data augmentation algorithm (Gelfand & Smith, 1990; Tanner & Wong, 1987) can typically be considered as Bayesian/stochastic versions of EM-type algorithms, the idea of parameter expansion also applies to them by placing a prior distribution on the auxiliary parameter α. As established by Liu, Wong & Kong (1994), the rate of convergence of the data augmentation algorithm can often be characterised by the autocorrelation of the simulated parameters, and this autocorrelation is actually the Bayesian fraction of missing information mentioned in § 3. As a result of the extra variation brought by the auxiliary parameter, the PX-EM version of the data augmentation algorithm will have a smaller fraction of missing information, and therefore a smaller autocorrelation and thus a faster rate of convergence under quite general conditions.

Acknowledgement

The authors wish to thank Professor A. P. Dempster, Professor X.-L. Meng and the editor and reviewers for their thoughtful comments. Ying Nian Wu is grateful to W. S. Cleveland and D. Lambert for inviting him to Lucent Bell Labs in the summers of 1996 and 1997, where part of the work was done. The work of Donald B. Rubin and Ying Nian Wu was partially supported by National Science Foundation and National Institute of Health grants, and the U.S. Census Bureau.

References

Arslan, O., Constable, P. D. L. & Kent, J. (1995). Convergence behavior of the EM algorithm for the multivariate t-distribution. *Commun. Statist.* A **24**, 2981–3000.

Chambers, J. M. & Hastie, T. J. (1992). *Statistical Models in S.* Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

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

Fessler, J. A. & Hero, A. O. (1994). Space-alternating generalized expectation-maximization algorithm. *IEEE Trans. Sig. Proces.* **42**, 2664–77.

GELFAND, A. E., SAHU, S. K. & CARLIN, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika* **82**, 479–88.

GELFAND, A. E. & SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85**, 398–409.

HARTLEY, H. O. & RAO, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**, 93–108.

JAMSHIDIAN, M. & JENNRICH, R. I. (1993). Conjugate gradient acceleration of the EM algorithm. *J. Am. Statist. Assoc.* **88**, 221–8.

KENT, J. T., TYLER, D. E. & VARDI, Y. (1994). A curious likelihood identity for the multivariate t distribution. *Commun. Statist.* B, **23**, 441–53.

LAIRD, N., LANGE, N. & STRAM, D. (1987). Maximizing likelihood computations with repeated measures: application of the EM algorithm. *J. Am. Statist. Assoc.* **82**, 97–105.

LAIRD, N. M. & WARE, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–74.

LANGE, K. (1995a). A gradient algorithm locally equivalent to the EM algorithm. *J. R. Statist. Soc.* B **57**, 425–38.

LANGE, K. (1995b). A quasi-Newtonian acceleration of the EM algorithm. *Statist. Sinica* **5**, 1–18.

LANGE, K. & CARSON, R. (1984). EM reconstruction for emission and transmission tomography. *J. Comp. Assist. Tomog.* **8**, 306–16.

LANGE, K., LITTLE, R. J. A. & TAYLOR, J. M. G. (1989). Robust statistical modeling using the t-distribution. *J. Am. Statist. Assoc.* **84**, 881–96.

LINDSTROM, M. J. & BATES, D. M. (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Am. Statist. Assoc.* **83**, 1014–22.

LITTLE, R. J. A. & RUBIN, D. B. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley.

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

LIU, C. & RUBIN, D. B. (1994a). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**, 633–48.

LIU, C. & RUBIN, D. B. (1994b). Application of the ECME algorithm and the Gibbs sampler to general linear mixed models. In *Proceedings of the XVIIth International Biometric Conference*, Hamilton, Ontario, **1** (invited papers), Ed. IBC '94 Local Organizing Committee, pp. 97–107. Dept of Mathematics and Statistics, McMaster's University.

LIU, C. & RUBIN, D. B. (1995). ML estimation of the multivariate t distribution. *Statist. Sinica*, **5**, 19–39.

LIU, C. & RUBIN, D. B. (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statist. Sinica* **8**. To appear.

LIU, J. S., WONG, W. H. & KONG, A. (1995). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

MENG, X. L. & RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267–78.

MENG, X. L. & VAN DYK, D. A. (1997). The EM algorithm — an old folk song sung to a fast new tune (with Discussion). *J. R. Statist. Soc.* B **59**, 511–67.

MENG, X. L. & VAN DYK, D. A. (1998). Fast EM-type implementation for mixed effects models. *J. R. Statist. Soc.* B **60**, 559–78.

RUBIN, D. B. (1983). Iteratively reweighted least squares. In *Encyclopedia of Statistical Sciences* **4**, Ed. S. Kotz, N. L. Johnson and C. B. Read, pp. 272–5. New York: John Wiley.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

RUBIN, D. B. & THAYER, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47**, 69–76.

SHEPP, L. A. & VARDI, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Image. Proces.* **2**, 113–22.

TANNER, M. A. & WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with Discussion). *J. Am. Statist. Assoc.* **82**, 805–11.

TITTERINGTON, D. M., SMITH, A. F. M. & MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: John Wiley.

VARDI, Y., SHEPP, L. A. & KAUFMAN, L. (1985). A statistical model for positron emission tomography. *J. Am. Statist. Assoc.* **80**, 8–19.

WU, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95–103.

[*Received February* 1997. *Revised February* 1998]