Comment on the article by Meng and Van Dyk

Andrew Gelman
Department of Statistics
Columbia University
New York, NY 10027

December 9, 1996

This and the accompanying Meng and Van Dyk (1996) paper have changed how I think about computation for mixed effects regression models, the simplest of which have the form:

$$y \sim \mathrm{N}(X\beta + W\alpha, \Sigma), \quad \alpha \sim \mathrm{N}(0, \tau^2 I), \tag{1}$$

where y, α and β are vectors. The joint ML estimate of (α, β, τ) has poor properties if the dimension of α is high, and the EM algorithm can be used to obtain the marginal ML estimate of (β, τ); see Laird and Ware (1982). When the estimate of τ is near zero, the parameters τ and $||\alpha||$ tend to be highly correlated, with the unfortunate result that EM (and the related Gibbs sampler) can move very slowly.

The present article would suggest breaking the correlation using the parameterization

$$y \sim \mathrm{N}(X\beta + \tau^a W\gamma, \Sigma), \quad \gamma \sim \mathrm{N}(0, \tau^{2-2a} I),$$

where the vector $\gamma = \tau^{-a}\alpha$ is treated as "missing data" in the EM algorithm, and a is a scalar "working parameter" that can be set by the user. The E-step can be performed in closed form for any value of a, but the M-step is simple only if $a = 0$ or 1. The usual parameterization (1) corresponds to $a = 0$.

The PX-EM algorithm (Liu, Rubin, and Wu, 1996) uses the more general parameterization,

$$y \sim \mathrm{N}(X\beta + \theta W\gamma, \Sigma), \quad \gamma \sim \mathrm{N}(0, \phi^2 I),$$

where θ and ϕ are scalars and, in the notation of model (1), $\alpha = \theta\gamma$ and $\tau = \theta\phi$. In PX-EM, the vector γ is treated as "missing data," and the M-step maximizes over β, θ, and ϕ.

We compared the old ($a = 0$), new ($a = 1$), and PX-EM algorithms for a hierarchical logistic regression problem from forestry with 379 data points, 3 fixed effects, and 15 random effects. (In this case, the algorithms are approximate EM using, at each step, the local linear approximation to the glm based on the current parameter estimates.) Figure 1 shows the number of iterations required until convergence (defined as when the relative error for τ is less than 10^{-4}) for each algorithm, as a function of the starting value for τ. The old algorithm has the well-known problem that if τ is started too low, the new algorithm corrects this, and PX-EM dominates both.

These ideas can easily be extended to multiple variance components and to the Gibbs sampler. It would also be interesting to relate to the work of Gelfand, Sahu, and Carlin (1994).

References

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1994). Efficient parameterizations for generalized linear mixed models. Technical report, Department of Statistics, University of Connecticut.

Laird, N. M., and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Liu, C., Rubin, D. B., and Wu, Y. (1996). Parameter expansion for EM acceleration—the PX-EM algorithm. Technical report, Bell Laboratories.

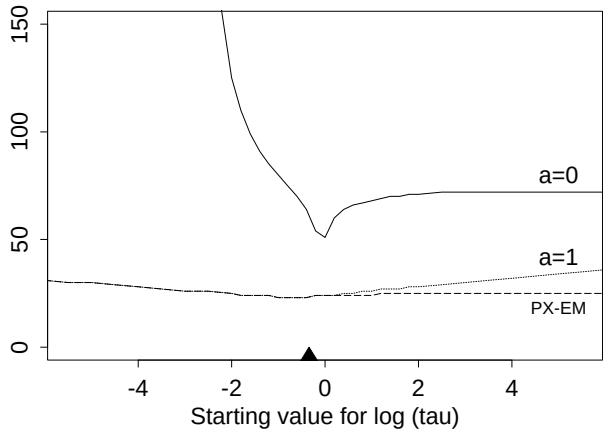

Figure 1: Number of iterations until convergence of τ, as a function of the starting point, in the approximate EM algorithm for a hierarchical logistic regression problem. Solid, dotted, and dashed lines show old ($a = 0$), new ($a = 1$), and PX-EM algorithms, respectively. The value of $\log \tau$ at convergence is indicated the bottom of the graph. Even when τ is started at the right value, the time to convergence is nonzero because the starting values for β are not perfect.