A Correlated Probit Model for Joint Modeling of Clustered Binary and Continuous Responses

Ralitza V. GUEORGUIEVA and Alan AGRESTI

A difficulty in joint modeling of continuous and discrete response variables is the lack of a natural multivariate distribution. For joint modeling of clustered observations on binary and continuous responses, we study a correlated probit model that has an underlying normal latent variable for the binary responses. Catalano and Ryan have factored the model into a marginal and a conditional component and used generalized estimating equations methodology to estimate the effects. We propose a Monte Carlo expectation–conditional maximization algorithm for finding maximum likelihood estimates of the mixed model itself, extending and accelerating an algorithm for models with binary responses. We demonstrate the methodology with a developmental toxicity study measuring fetal weight and a binary malformation status for several litters of mice. A simulation study suggests that efficiency gains of joint fittings over separate fittings of the response variables occur mainly for small datasets with strong correlations between the responses within cluster.

KEY WORDS: Generalized linear mixed model; Latent variable; Monte Carlo EM algorithm; Random effect; Teratology.

1. INTRODUCTION

The modeling of various forms of clustered data, such as repeated measurements in a longitudinal study, has received much attention in recent years. Most research has concentrated on a single response variable, but many studies have measured multiple response variables for each subject. In this article we consider modeling a clustered multivariate response with binary and continuous components.

Models for multivariate clustered data are necessarily complex, because they must consider two types of correlations: between measurements on different variables for each cluster and between measurements on different subjects within a cluster. A difficulty in constructing parametric models for joint modeling of continuous and discrete responses is the lack of a natural multivariate distribution. For joint modeling of binary and continuous responses, we study a correlated probit model that applies with underlying latent normal variables for the binary responses. Catalano and Ryan (1992) considered such a model and used generalized estimating equations (GEE) methodology to estimate the effects. Here we propose a Monte Carlo EM algorithm for finding maximum likelihood (ML) estimates for the subject-specific model. We extend an algorithm introduced by Chan and Kuk (1997) for models with binary responses to binary and normal variables with correlated errors. Because the algorithm is slow, we also provide an acceleration of it.

We illustrate the methods with a dataset from a developmental toxicity study of ethylene glycol in mice conducted through the National Toxicology Program (Price, Kimmel, Tyl, and Marr 1985). The experiment assigned pregnant mice randomly to four groups, one group serving as a control and the other three groups exposed to different levels of ethylene glycol during major organogenesis. Following sacrifice, measurements were taken on each fetus in the uterus. The two outcome measures on each live fetus were fetal weight (continuous) and whether the fetus was malformed (binary). Table 1 gives descriptive statistics for the data. Mean fetal weight decreases monotonically with increasing dose, whereas the malformation rate increases with dose. The goal of the analysis was to study the joint effects of ethylene glycol dose on fetal weight and on the probability of malformation.

Other authors who have studied joint modeling of discrete and continuous responses include Catalano and Ryan (1992), Fitzmaurice and Laird (1995), Regan and Catalano (1999), and Rochon (1996). But the focus of their work was on marginal modeling, rather than on the joint, subject-specific effects on the models in this article. Dunson (2000) considered Bayesian latent variable models. Catalano and Ryan (1992) used the correlated probit model for this developmental toxicity study, but factored the joint distribution into the marginal distribution of the continuous response (fetal weight) and the conditional distribution of the binary response (malformation) given the continuous response. Then they used GEE to fit consecutively the marginal and conditional models. Because of the reparameterization, some parameters in the original model were not estimable. Fitzmaurice and Laird (1995) also considered a two-stage model but reversed the conditioning order, specifying a marginal logit model for malformation and a conditional model for fetal weight given malformation.

Regan and Catalano (1999) considered joint estimation of all marginal and correlation parameters through ML. But their approach works only for an exchangeable correlation structure between the continuous and the latent continuous responses, an assumption that may be restrictive in this example. The correlated probit model that we consider does not make this assumption and allows simultaneous estimation of all parameters, but at the cost of greater computational complexity.

Section 2 defines the correlated probit model. Section 3 develops Monte Carlo EM algorithms for ML estimation, and Section 4 applies them to the developmental toxicity study. Section 5 discusses a simulation study to assess efficiency gains of joint over separate fitting of the response variables. Section 6 describes extensions to joint modeling of continuous, truncated continuous, binary, and ordinal data.

Ralitza V. Gueorguieva is Associate Research Scientist, Division of Biostatistics, Department of Epidemiology and Public Health, Yale University, New Haven, CT 06520 (E-mail: *ralitza.gueorguieva@yale.edu*). Alan Agresti is Professor, Department of Statistics, University of Florida, Gainesville, FL 32611 (E-mail: *aa@stat.ufl.edu*). Agresti's research was partially supported by grants from the National Institutes of Health and the National Science Foundation. The authors appreciate helpful comments from the referees.

Table 1. Descriptive Statistics for the Developmental Toxicity Example

Dose			Fetal weight (g)		Malformation	
(g/kg)	Dams	Live Fetuses	Mean	SD	Number	Percent
0	25	297	.97	.10	1	.3
.75	24	276	.88	.10	26	9.4
1.50	22	229	.76	.11	89	38.9
3.00	23	226	.70	.12	126	57.1

2. CORRELATED PROBIT MODEL DEFINITION

We now formulate the correlated probit model for a single binary response and a single continuous response. Section 6 discusses extensions to more than two responses.

Let $\{y_{i1j}\}$ denote the continuous response at the jth observation for the ith subject or cluster, $i = 1, \ldots n$, $j = 1, \ldots n_i$. We assume that the binary response results from dichotomizing a latent normal response. Let $\{y^*_{i2j}\}$ denote the latent measurement, such that the observed binary response at the jth observation for the ith subject or cluster is the indicator $y_{i2j} = I\{y^*_{i2j} > 0\}$. The underlying linear mixed model is defined as

$$\begin{aligned} y_{i1j} &= \mathbf{x}^T_{i1j}\boldsymbol{\beta}_1 + \mathbf{z}^T_{i1j}\mathbf{b}_{i1} + \epsilon_{i1j}, \\ y^*_{i2j} &= \mathbf{x}^T_{i2j}\boldsymbol{\beta}_2 + \mathbf{z}^T_{i2j}\mathbf{b}_{i2} + \epsilon_{i2j}, \end{aligned} \tag{1}$$

where $\mathbf{x}_{i1j}$, $\mathbf{x}_{i2j}$, $\mathbf{z}_{i1j}$, and $\mathbf{z}_{i2j}$ are known $p_1 \times 1$, $p_2 \times 1$, $q_1 \times 1$, and $q_2 \times 1$ (column) vectors and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unknown $p_1 \times 1$ and $p_2 \times 1$ parameter vectors. The random effects and the random errors are assumed to be normally distributed,

$$\mathbf{b}_i = \begin{pmatrix} \mathbf{b}_{i1} \\ \mathbf{b}_{i2} \end{pmatrix} \sim \text{iid } \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}) = \mathbf{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}^T_{12} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right) \tag{2}$$

and

$$\boldsymbol{\epsilon}_{ij} = \begin{pmatrix} \epsilon_{i1j} \\ \epsilon_{i2j} \end{pmatrix} \sim \text{iid } \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_e) = \mathbf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{e1} & \sigma_{e12} \\ \sigma_{e12} & \sigma^2_{e2} \end{bmatrix}\right). \tag{3}$$

This model for the complete data $\{y_{i1j}\}$ and $\{y^*_{i2j}\}$ translates into the following model for the observed data $\{y_{i1j}\}$ and $\{y_{i2j}\}$: Conditional on $\{\mathbf{b}_{i1}\}$ and $\{\mathbf{b}_{i2}\}$,

$$\mu_{i1j} = \mathbf{x}^T_{i1j}\boldsymbol{\beta}_1 + \mathbf{z}^T_{i1j}\mathbf{b}_{i1} \quad \text{and} \quad \Phi^{-1}(\mu_{i2j}) = \mathbf{x}^T_{i2j}\boldsymbol{\beta}_2 + \mathbf{z}^T_{i2j}\mathbf{b}_{i2},$$

where μ_{i1j} and μ_{i2j} are the conditional means for the two observed variables and Φ denotes a normal cumulative distribution function.

Several special cases exist. For the binary outcome alone, the model is a multivariate probit model (Lessafre and Molenberghs 1991) with equicorrelated structure. For the continuous outcome alone, the model is a general linear model with equicorrelated structure. If $\sigma_{e12} = 0$, then the model is a bivariate generalized linear mixed model (GLMM), with a probit link for the Bernoulli response and an identity link for the normal response (Gueorguieva 1999). If $\sigma_{e12} = 0$ and the random effects structure consists only of intercepts for each variable, then the model is a special case of the Regan and Catalano (1999) model but with no modeling of the covariance parameters. If $\sigma_{e12} = 0$ and $\sigma_{12} = 0$, then the model is equivalent to specifying separate GLMMs for the two response variables.

3. MAXIMUM LIKELIHOOD MODEL FITTING

We use modifications of the EM algorithm to obtain ML estimates for the correlated probit model. We first extend an approach of Chan and Kuk (1997) using a Monte Carlo expectation–conditional maximization (ECM) algorithm. We then formulate a stochastic approximation approach to increase the speed. Finally, we discuss standard error approximation.

3.1 Monte Carlo Expectation–Conditional Maximization Algorithm

The complete dataset consists of $\mathbf{b}_i$ and $\{\mathbf{y}^*_{i\_j}\} = \{(y_{i1j}, y^*_{i2j})^T\}$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. Let $\mathbf{x}_{i\_j} = \begin{pmatrix} \mathbf{x}^T_{i1j} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}^T_{i2j} \end{pmatrix}$, $\mathbf{z}_{i\_j} = \begin{pmatrix} \mathbf{z}^T_{i1j} & \mathbf{0} \\ \mathbf{0} & \mathbf{z}^T_{i2j} \end{pmatrix}$, and $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$. Then the complete data log-likelihood is

$$\begin{aligned} \log L = &-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\log|\boldsymbol{\Sigma}_e| - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\mathbf{y}^*_{i\_j} - \mathbf{x}_{i\_j}\boldsymbol{\beta} - \mathbf{z}_{i\_j}\mathbf{b}_i)^T \\ &\times \boldsymbol{\Sigma}^{-1}_e(\mathbf{y}^*_{i\_j} - \mathbf{x}_{i\_j}\boldsymbol{\beta} - \mathbf{z}_{i\_j}\mathbf{b}_i) - \frac{1}{2}\sum_{i=1}^{n}\log|\boldsymbol{\Sigma}| \\ &- \frac{1}{2}\sum_{i=1}^{n}\mathbf{b}^T_i\boldsymbol{\Sigma}^{-1}\mathbf{b}_i. \end{aligned}$$

The complete data ML estimates result from closed-form expressions. Let

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{b}_i\mathbf{b}^T_i. \tag{4}$$

For a fixed $\boldsymbol{\Sigma}_e$, let

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mathbf{x}^T_{i\_j}\boldsymbol{\Sigma}^{-1}_e\mathbf{x}_{i\_j}\right)^{-1}\left(\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mathbf{x}^T_{i\_j}\boldsymbol{\Sigma}^{-1}_e(\mathbf{y}^*_{i\_j} - \mathbf{z}_{i\_j}\mathbf{b}_i)\right), \tag{5}$$

and for a fixed $\boldsymbol{\beta}$, let

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_e = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}&(\mathbf{y}^*_{i\_j} - \mathbf{x}_{i\_j}\boldsymbol{\beta} - \mathbf{z}_{i\_j}\mathbf{b}_i) \\ &\times (\mathbf{y}^*_{i\_j} - \mathbf{x}_{i\_j}\boldsymbol{\beta} - \mathbf{z}_{i\_j}\mathbf{b}_i)^T. \end{aligned} \tag{6}$$

Meng and Rubin (1993) showed that iterating between the last two equations in the EM algorithm provides convergence to the true ML estimates. At each step, the new estimate of $\boldsymbol{\Sigma}_e$ uses the previous value of $\hat{\boldsymbol{\beta}}$, and then the new value of $\hat{\boldsymbol{\Sigma}}_e$ is used to update $\hat{\boldsymbol{\beta}}$.

In the E step of the Monte Carlo ECM algorithm, (4), (5), and (6) are replaced by conditional versions depending only on the mean and covariance matrices of the conditional distributions of the latent continuous responses $\{\mathbf{y}^*_{i2\_}|\mathbf{y}_{i1\_}, \mathbf{y}_{i2\_}, \hat{\boldsymbol{\psi}}^{(r)}\}$. The Appendix gives the derivation, which is tedious, and describes the Gibbs sampler used to simulate values from the truncated multivariate normal distributions.

The steps of the Monte Carlo ECM algorithm are then as follows:

1. Select an initial estimate $\hat{\boldsymbol{\psi}}^{(0)}$ of the parameter vector. Set $r = 1$.
2. Increase r by 1. **E step:** For each subject i, $i = 1, \ldots, n$, generate m^* random samples from the conditional distribution of $\{\mathbf{y}^*_{i2\_}|\mathbf{y}_{i1\_}, \mathbf{y}_{i2\_}; \hat{\boldsymbol{\psi}}^{(r-1)}\}$ and compute approximations to the mean and the covariance matrix [see (A.5) in the Appendix].
3. **M step**: Update the estimate of the parameter vector $\hat{\boldsymbol{\psi}}^{(r)}$ using the conditional versions of (4), (6), and (5); (A.2), (A.3), and (A.4).
4. Iterate between steps 2 and 3 until convergence is achieved.

Now σ^2_{e2} is estimable from the complete data, but not from the observed data. With the usual approach of setting $\sigma^2_{e2} = 1$, the maximization step becomes more complicated because there is no closed-form expression for $\hat{\boldsymbol{\Sigma}}_e$. On the other hand, with σ^2_{e2} unrestricted, the ECM algorithm converges to unique estimates of the fully identifiable ratios $\boldsymbol{\beta}/\sigma_{e2}$, $\boldsymbol{\Sigma}_{12}/\sigma_{e2}$, and $\boldsymbol{\Sigma}_{22}/\sigma^2_{e2}$. The resulting ECM algorithm is also a parameter expanded (PX-EM) algorithm (Liu, Rubin, and Wu 1999).

Both the ECM and the PX-EM are extensions of the original EM algorithm. The ECM algorithm usually takes more iterations to converge than the EM algorithm but can be faster computationally because it simplifies the maximization step. The PX-EM also aims to reduce the total computing time by simplifying the maximization step and/or reducing the number of iterations. For the correlated probit model, the E step is computationally intensive. Adding numerical maximization for the M step, the algorithm is extremely slow. Even using both ECM and PX-EM to speed the algorithm, this approach was still slow. Thus we also developed an alternative fitting procedure by adapting an accelerated Monte Carlo EM algorithm proposed by Liao (1999) and by Delyon, Lavielle, and Moulines (1999). The latter article called the algorithm *stochastic approximation EM* (SAEM), because it replaced each expectation step of the EM algorithm by one iteration of a stochastic approximation procedure.

3.2 Stochastic Approximation Expectation–Conditional Maximization Algorithm

The convergence results hold for a complete data log-likelihood of the form

$$\log\, L_{\mathbf{u}}(\boldsymbol{\psi}) = [a(\boldsymbol{\psi})]^T \mathbf{z}(\mathbf{u}) - b(\mathbf{y}, \boldsymbol{\psi}),$$

where $\mathbf{z}(\mathbf{u})$ is a vector function of the complete data $\mathbf{u}$. Liao (1999) and Delyon et al. (1999) proposed replacing the usual Monte Carlo EM E step by calculating

$$\bar{\mathbf{z}}^{(r)} = (1 - w_r)\bar{\mathbf{z}}^{(r-1)} + w_r \left[\frac{1}{m_r}\sum_{k=1}^{m_r} \mathbf{z}(\mathbf{u}^{(k)})\right],$$

where $\mathbf{z}(\mathbf{u}^{(k)})$, $k = 1, \ldots, m_r$ are generated values from $\{\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\psi}}^{(r-1)}\}$ and $\{w_r\}$ are chosen weights that satisfy $\sum w_r = +\infty$ and $\sum w_r^2 < +\infty$. The E step is followed by the usual M step after an initial "stabilization period" of length r_0 for $\bar{\mathbf{z}}_r$. We apply this algorithm to the correlated probit model as follows:

1. Select an initial estimate $\hat{\boldsymbol{\psi}}^{(0)}$ of the parameter vector. Generate r_0 samples from the distribution of $\{\mathbf{y}^*_{i2\_}|\mathbf{y}_{i1\_}, \mathbf{y}_{i2\_}; \hat{\boldsymbol{\psi}}^{(0)}\}$ and compute the approximations

$$\bar{\mathbf{z}}_1^{(r_0)} = \frac{1}{r_0}\sum_{k=1}^{r_0} \mathbf{y}^{*\,(k)}_{i2\_} \quad \text{and} \quad \bar{\mathbf{z}}_2^{(r_0)} = \frac{1}{r_0}\sum_{k=1}^{r_0} \mathbf{y}^{*\,(k)}_{i2\_}\mathbf{y}^{*\,(k)^T}_{i2\_}.$$

   Set $r = r_0$.
2. Increase r by 1. **E step:** For $i = 1, \ldots, n$, generate a random sample $\mathbf{y}^{*\,(r)}_{i2\_}$ from the distribution of $(\mathbf{y}^*_{i2\_}|\mathbf{y}_{i1\_}, \mathbf{y}_{i2\_}; \hat{\boldsymbol{\psi}}^{(r-1)})$ using multivariate rejection sampling (or the Gibbs sampler with starting values $\mathbf{y}^{*\,(r-1)}_{i2\_}$) and compute the approximations

$$\bar{\mathbf{z}}_1^{(r)} = (1 - w_r)\bar{\mathbf{z}}_1^{(r-1)} + w_r \mathbf{y}^{*\,(r)}_{i2\_} \quad \text{and}$$
$$\bar{\mathbf{z}}_2^{(r)} = (1 - w_r)\bar{\mathbf{z}}_2^{(r-1)} + w_r \mathbf{y}^{*\,(r)}_{i2\_}\mathbf{y}^{*\,(r)^T}_{i2\_} \qquad (7)$$

   to $E(\mathbf{y}^*_{i2\_}|\mathbf{y}_{i1\_}, \mathbf{y}_{i2\_}; \hat{\boldsymbol{\psi}}^{(r-1)})$ and $E(\mathbf{y}^*_{i2\_}\mathbf{y}^{*\,T}_{i2\_}|\mathbf{y}_{i1\_}, \mathbf{y}_{i2\_}; \hat{\boldsymbol{\psi}}^{(r-1)})$. (Here m_r is chosen as 1, because the E step is much more computationally intensive than the M step.)
3. **M step:** Update the estimate of the parameter vector $\hat{\boldsymbol{\psi}}^{(r)}$ by substituting $\bar{\mathbf{z}}_1^{(r-1)}$ and $\bar{\mathbf{z}}_2^{(r-1)}$ in (A.2), (A.3), and (A.4).
4. Iterate between steps 2 and 3 until convergence is achieved.

There are two versions of this algorithm, according to whether one uses multivariate rejection sampling or the Gibbs sampler in the E step. Compared to the Monte Carlo ECM algorithm, the stochastic approximation ECM algorithm is much faster because of the closed-form expressions for the complete-data ML estimates. It also has the advantage of easily approximating the standard errors within the algorithm using the same type of shrinkage estimators as for the parameters. No simulated values are wasted in the process.

3.3 Standard Error Approximation

Based on work of Louis (1982), the observed data information matrix has representation

$$\begin{aligned}
&-\frac{\partial^2 l(\mathbf{y}, \boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T} \\
&\quad = -E\left[\frac{\partial^2 \log L_u(\mathbf{b}, \mathbf{y}^*, \boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}|\mathbf{y}\right] - \operatorname{var}\left[\frac{\partial \log L_u(\mathbf{b}, \mathbf{y}^*, \boldsymbol{\psi})}{\partial\boldsymbol{\psi}}|\mathbf{y}\right] \\
&\quad = -E\left[\frac{\partial^2 \log L_u(\mathbf{b}, \mathbf{y}^*, \boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T} + \frac{\partial \log L_u(\mathbf{b}, \mathbf{y}^*, \boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right. \\
&\qquad \left.\times \frac{\partial \log L_u(\mathbf{b}, \mathbf{y}^*, \boldsymbol{\psi})}{\partial\boldsymbol{\psi}^T}|\mathbf{y}\right] + E\left[\frac{\partial \log L_u(\mathbf{b}, \mathbf{y}^*, \boldsymbol{\psi})}{\partial\boldsymbol{\psi}}|\mathbf{y}\right] \\
&\qquad \times E\left[\frac{\partial \log L_u(\mathbf{b}, \mathbf{y}^*, \boldsymbol{\psi})}{\partial\boldsymbol{\psi}^T}|\mathbf{y}\right] \doteq \mathbf{G} + \boldsymbol{\Delta}\boldsymbol{\Delta}^T.
\end{aligned}$$

By simulating values from the conditional distribution of $\{(\mathbf{b}, \mathbf{y}^*)|\mathbf{y}\}$, both $\mathbf{G}$ and $\boldsymbol{\Delta}$ can be approximated at each step of the algorithm:

$$\mathbf{G}^{(r)} = (1 - w_r)\mathbf{G}^{(r-1)} + w_r \mathbf{N}_r,$$

where

$$\mathbf{N}_r = \frac{\partial^2 \log L_u(\mathbf{b}^{(r)}, \mathbf{y}^{*(r)}, \boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} + \frac{\partial \log L_u(\mathbf{b}^{(r)}, \mathbf{y}^{*(r)}, \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \times \frac{\partial \log L_u(\mathbf{b}^{(r)}, \mathbf{y}^{*(r)}, \boldsymbol{\psi})}{\partial \boldsymbol{\psi}^T}$$

and

$$\boldsymbol{\Delta}_r = (1 - w_r)\boldsymbol{\Delta}_{r-1} + w_r \frac{\partial \log L_u(\mathbf{b}^{(r)}, \mathbf{y}^{*(r)}, \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}.$$

Because

$$f(\mathbf{y}^*, \mathbf{b}|\mathbf{y}) = f(\mathbf{y}^*|\mathbf{y}) f(\mathbf{b}|\mathbf{y}, \mathbf{y}^*) = f(\mathbf{y}^*|\mathbf{y}) f(\mathbf{b}|\mathbf{y}^*),$$

one can first simulate $\mathbf{y}_2^*$ as outlined for the parameter estimation, and then simulate $\mathbf{b}$ from $(\mathbf{b}|\mathbf{y}^*)$ using the multivariate normal.

4. APPLICATION TO THE DEVELOPMENTAL TOXICITY STUDY

For the developmental toxicity example, for the jth live fetus in litter i, let y_{i1j} = fetal weight, y_{i2j}^* = latent malformation, $y_{i2j} = I\{y_{i2j}^* > 0\}$ = observed malformation status, and x_i = ethylene glycol dosage level. The correlated probit model is

$$y_{i1j} = \beta_{10} + x_i \beta_{11} + b_{i1} + \epsilon_{i1j},$$
$$y_{i2j}^* = \beta_{20}^* + x_i \beta_{21}^* + b_{i2} + \epsilon_{i2j},$$

where

$$\mathbf{b}_i = \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim \text{iid}\mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \text{and} \quad \boldsymbol{\epsilon}_{i\_j} = \begin{pmatrix} \epsilon_{i1j} \\ \epsilon_{i2j} \end{pmatrix} \sim \text{iid}\mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_e).$$

We also considered a quadratic term for dose in the linear predictor for fetal weight. It was not significant and is not included in the models discussed here.

The identifiable parameters in this specification are β_{10}, β_{11}, $\beta_{20} = \beta_{20}^*/\sigma_{e2}$, $\beta_{21} = \beta_{21}^*/\sigma_{e2}$, $\sigma_{b1} = \sigma_1$, $\sigma_{b2} = \sigma_2/\sigma_{e2}$, $\rho_b = \sigma_{12}/(\sigma_1\sigma_2)$, σ_{e1}, and $\rho_e = \sigma_{e12}/(\sigma_{e2}\sigma_{e1})$. Table 2 contains the approximate ML estimates obtained using the adaptation of Chan and Kuk's algorithm (after 500 iterations) and Liao's algorithm with Gibbs sampling (after 10,000 iterations). Multivariate rejection sampling was also considered, but it was very inefficient and thus results are not shown here. All estimates were the same up to two significant digits after the decimal point.

Figure 1 compares the convergence of the parameter estimates from the two methods, measured in actual minutes, using the same Sun Ultra Enterprise 450 workgroup server for 24 hours. This corresponded to 353 iterations of the Chan and Kuk algorithm and to 68,160 iterations for the Liao algorithm using the Gibbs sampler. (Even after 24 hours, some estimates in the algorithm using multivariate rejection sampling had not yet converged.) The Liao algorithm using the Gibbs sampler is much faster than the Chan and Kuk algorithm. A closer look at Liao's Gibbs sampler shows that the estimates appeared to have converged (or nearly converged) after only about 2,000 iterations and in less than an hour. The Chan and Kuk algorithm took about 8 hours. Both algorithms had convergence problems when initial estimates were far from the true values.

General results for Robbins and Monroe types of stochastic approximations suggest that the best rate of convergence will occur for weights with $O(l^{-1})$ (Ruppert 1991). To investigate this in the SAECM algorithm, we fitted the model using three different weighting schemes satisfying the conditions in 3.2: $w_l = \frac{2}{2+l^c}$ for $c = 1, .75, .6$. All three schemes converged, but smaller c showed higher fluctuations in the parameter estimates, especially in early iterations. Figure 2 illustrates the behavior of the slope estimate for malformation at three stages of the algorithm (beginning, sometime in the middle, and toward the end). Of the three weighting schemes considered, the one with $c = 1$ converged the fastest and showed the least fluctuation. The results with weighting schemes of the form $w_l = j/j + l$, $j = 2, 5, 10$ were very similar. Such weights have the same asymptotic speed of convergence, but bigger j's may be more appropriate if the initial values are far away from the ML estimates. Such weighting schemes may also have a substantial effect in small samples (Ruppert 1991).

We incorporated standard error computations into Liao's Gibbs sampler algorithm. Table 3 shows the results (under "full model") after 10,000 iterations. The standard error estimates converged more slowly than the parameter estimates. Figure 3 illustrates a typical situation, showing the intercept parameter for malformation and its standard error.

When $\rho_e = 0$, the correlated probit model implies the exchangeable correlation structure between the continuous and the latent continuous responses used by Regan and Catalano (1999). The correlation between fetal weight and latent mal-

Table 2. Approximate ML Estimates of Parameters in the Correlated Probit Model for the Developmental Toxicity Example Using the Chan and Kuk Method and Liao Method Based on the Gibbs Sampler

Parameter	Description	Chan and Kuk	Liao
β_{10}	Fetal weight intercept	.952	.952
β_{11}	Dosage slope on fetal weight	−.087	−.087
β_{20}	Malformation intercept	−2.398	−2.396
β_{21}	Dosage slope on malformation	.970	.969
σ_{b1}	Standard deviation for fetal weight random effect	.086	.086
σ_{b2}	Standard deviation for malformation random effect	.842	.837
ρ_b	Correlation between random effects	−.641	−.642
σ_{e1}	Error standard deviation for fetal weight	.075	.075
ρ_e	Error correlation	−.214	−.210

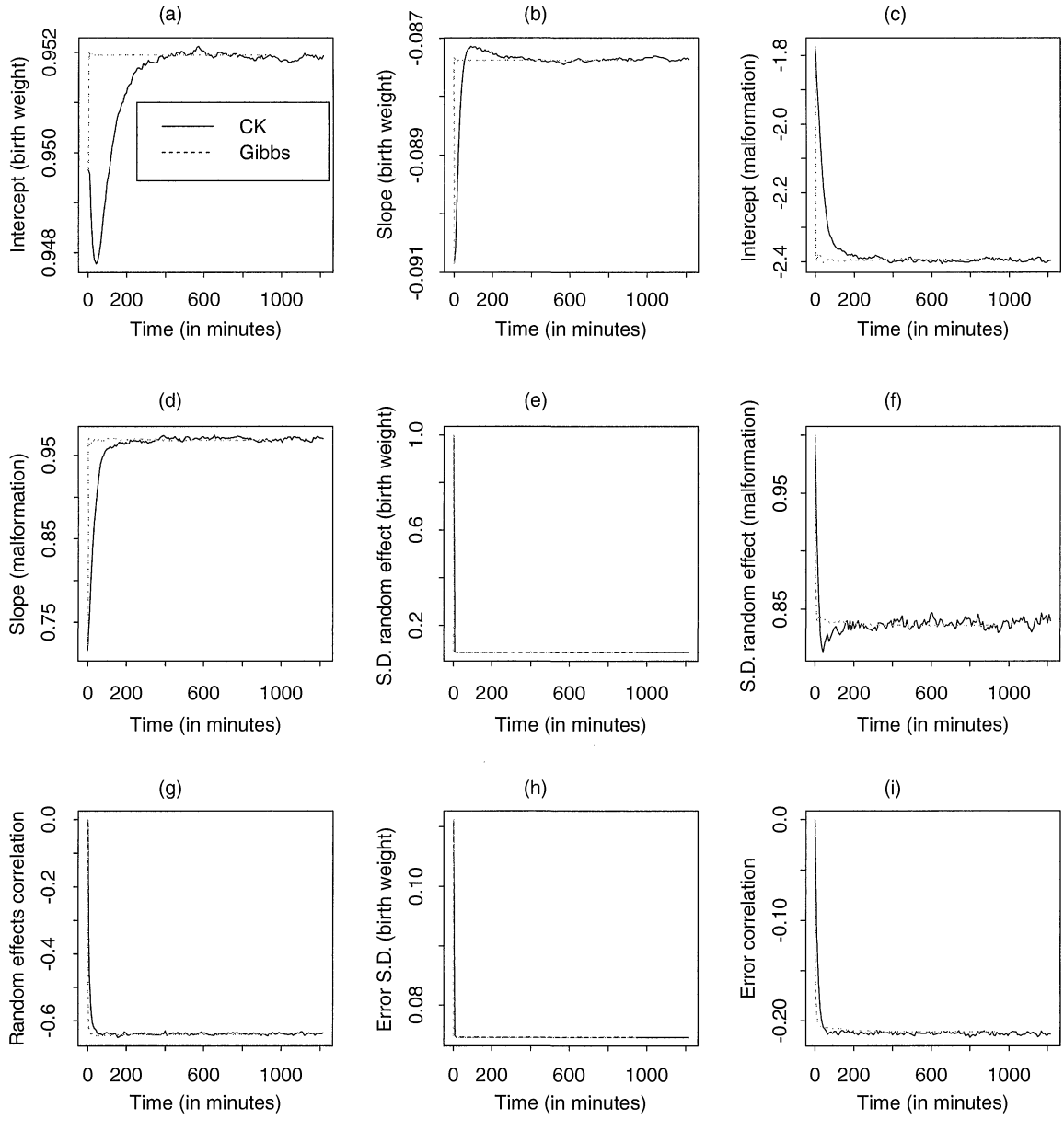

Figure 1. Convergence of the Parameter Estimates in the Development Toxicity Example. (a) Intercept (birth weight); (b) slope (birth weight); (c) intercept (malformation); (d) slope (malformation); (e) standard deviation random effect (birth weight); (f) standard deviation random effect (malformation); (g) random effect correlations; (h) error standard deviation (birth weight); (i) error correlation. (—— Chan and Kuk; - - - - Liao Gibbs).

formation within a fetus is assumed to be the same as the correlation between fetal weight and latent malformation measured on two different fetuses within a litter. When $\rho_e = 0$ and $\rho_b = 0$, this corresponds to specifying two separate GLMMs for the two response variables. When this reduced model holds, one can analyze the two responses by fitting the component models separately. Table 3 also contains the ML estimates and standard errors for these two reduced models, using Liao's Gibbs sampler. Table 3 shows a weak, but significant intrafetus correlation between malformation and fetal weight ($\hat{\rho}_e/\widehat{\mathrm{SE}}(\hat{\rho}_e) = -.211/.055 = -3.8$), suggesting that the exchangeable correlation structure is not appropriate. Table 3 also shows a moderately strong negative correlation between the random effects, so even if $\rho_e = 0$, there could be some advantage to fitting the models jointly. However, Table 3 shows that for the three models, the regression parameter estimates and standard errors are very similar. This is somewhat surprising, because one expects efficiency gains from joint fitting of the responses using a more realistic correlation structure. The next section studies this issue.

5. SIMULATION STUDY OF POTENTIAL EFFICIENCY GAINS

We performed a simulation study to investigate efficiency gains in fitting the full correlated probit model (denoted by FM) instead of fitting reduced model 1 (i.e., $\rho_e = 0$, denoted by RM1) or reduced model 2 of separate univariate GLMMs for the individual response variables (i.e., $\rho_e = \rho_b = 0$, denoted

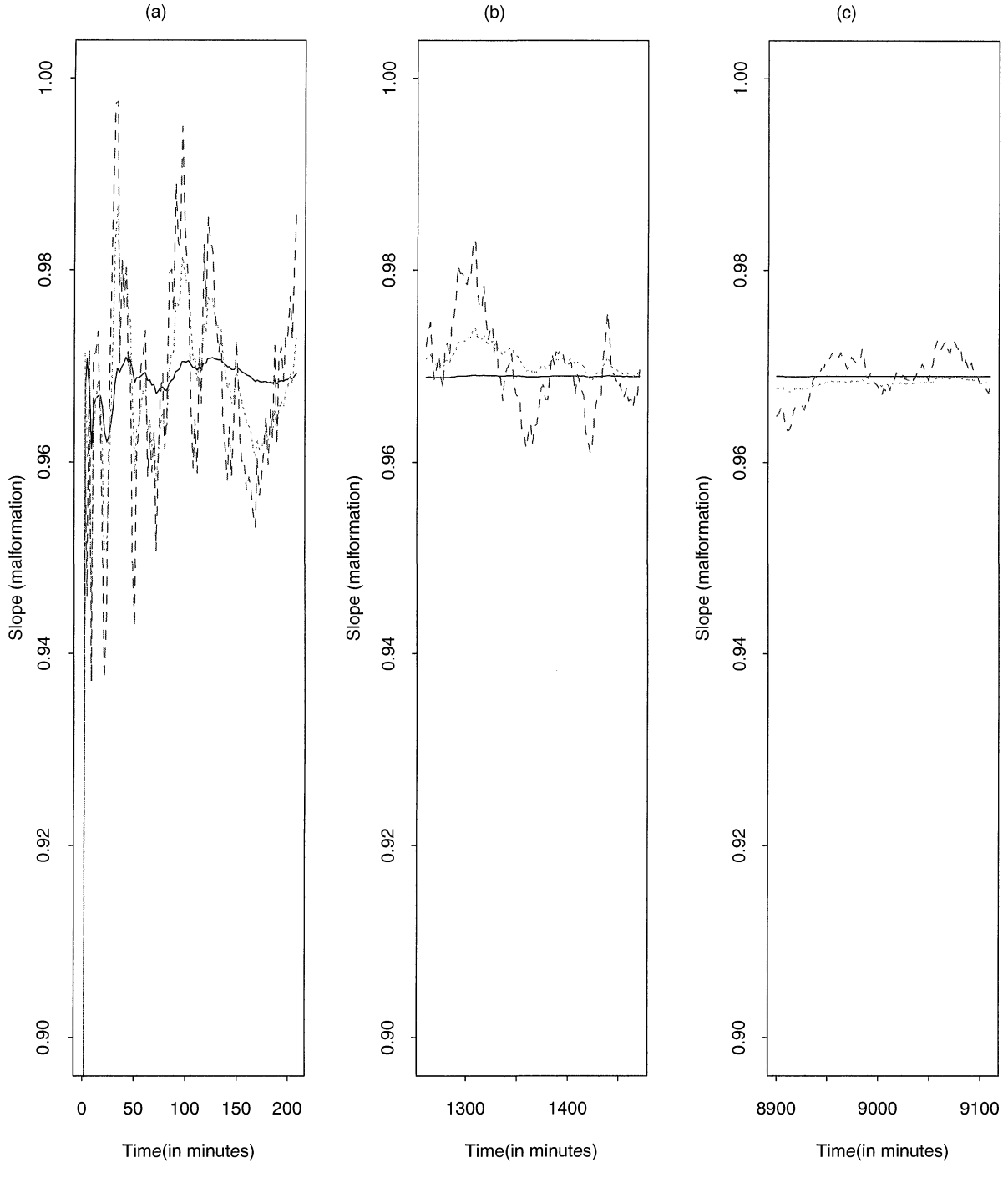

Figure 2. Convergence of the Slope Estimate for Malformation at Early (a), Middle (b), and Late (c) Stages of the Fitting Process Using Different Weights (—— 1, ... 3/4, - - - - - 3/5).

by RM2). We used the same data structure as in the developmental toxicity example. The parameter values, except for ρ_e, were set equal to the ML estimates from Liao's Gibbs sampler. We used four settings, corresponding to the combinations of size of dataset (large, small) and strength of intrafetus correlation (strong, weak). The large dataset had the same numbers of clusters and observations as in the example (94 clusters and an average of 11 observations per cluster), whereas the small dataset had 24 clusters with 6 observations per cluster. The correlation settings were $\rho_e = -.80$ and $\rho_e = -.20$. A total of 50 samples were generated at each of the four settings, and all three models were fitted using Liao's Gibbs sampler. (The restriction to 50 samples reflected the nearly 1 month of computing time needed to run the cases of 94 clusters.)

Tables 4 and 5 show results for two of the four scenarios: the large dataset with weak correlation and the small dataset with strong correlation. The tables report the average parameter estimates and their average standard errors and standard deviations. Empirical standard errors were also computed for the standard error estimates, to assess whether efficiency gains

Table 3. ML Estimates and Standard Errors (Obtained Using Liao's Gibbs Sampler) With Linear Dose Effects for Both Variables

	Full model		Reduced model 1		Reduced model 2	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
β_{10}	.952	.014	.952	.014	0.952	0.014
β_{11}	−.087	.008	−.087	.008	−.087	.008
β_{20}	−2.396	.216	−2.401	.216	−2.416	.217
β_{21}	.971	.110	.972	.110	0.988	0.112
σ_{b1}	.086	.007	.086	.007	0.086	0.007
σ_{b2}	.837	.106	.839	.107	.873	.113
ρ_b	−.640	.091	−.664	.091		
σ_{e1}	.075	.002	.075	.002	.075	.002
ρ_e	−.211	.055				

were important. By comparing the average of the standard error estimates and the standard deviations for the parameter estimates, Monte Carlo error can be judged. However, differences between the two may also reflect inadequacy of asymptotic standard error estimates for the sample sizes used.

When the sample size is large and ρ_e is weak, the average standard errors for all parameter estimates were similar for all three models. Hardly any efficiency gain resulted from fitting the responses jointly. Similarly, hardly any efficiency gain occurred with the (large n, strong correlation) and (small n, weak correlation) cases not shown here. For small sample size and strong intrafetus correlation ($\rho_e = -.804$), the efficiency gains were more pronounced; for instance, the average standard error estimate for β_{21} increased from .175 to .255 between FM and RM2. The difference was real, as suggested by the empirical standard errors associated with the average standard error estimates (.010 and .012).

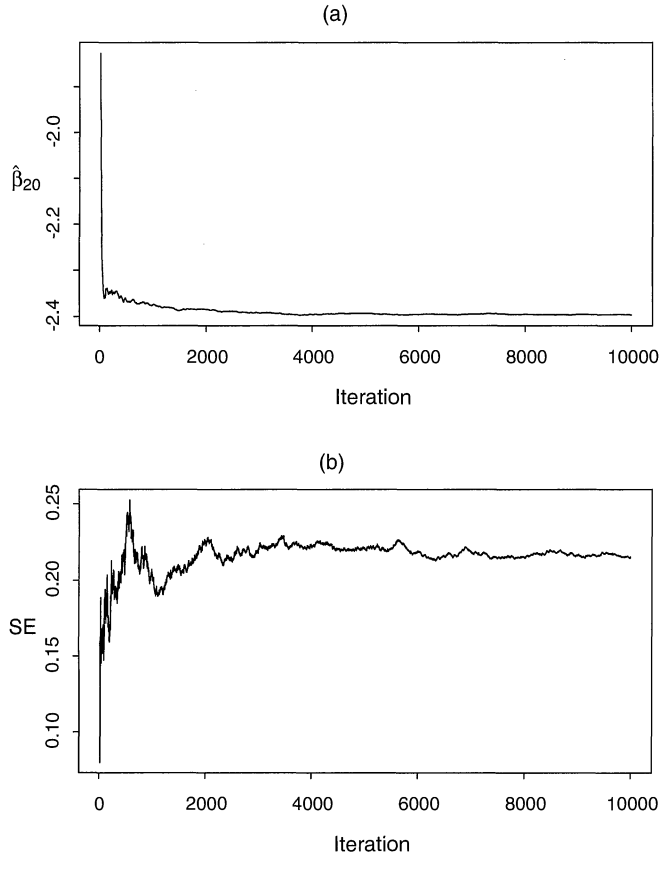

Figure 3. Convergence of the Intercept Estimate for Malformation (a) and Its Estimated Standard Error (b), Using the Gibbs Sampler.

It was not surprising that noticeable efficiency gains occurred for small datasets with strong correlations between the responses within cluster and that they were the greatest for parameters for the binary response. For small studies, there may not be enough information in the binary response by itself to estimate its effects precisely; when a continuous response is strongly correlated with the binary outcome, one may gain efficiency by using the information in the continuous response to help estimate parameters for the binary response. We also performed an additional simulation study of two binary outcomes with small sample sizes and large correlations. There also we observed some efficiency gains with the multivariate analysis.

Although the notion that main efficiency gains occur with large intrafetus correlation is intuitively appealing, it seems curious that they could diminish with large n. To investigate further, for paired binary and continuous responses we considered a simple model for which it is possible to compute the asymptotic relative efficiency (ARE) of the ML estimator of the binary parameter from the joint model with respect to its ML estimator for the binary marginal model. The model, which contains only intercepts and no random effects, is defined as

$$\begin{pmatrix} y_{i1} \\ y_{i2}^* \end{pmatrix} \sim \text{iid}\,\mathbf{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{bmatrix}\right), \tag{8}$$

where y_{i1} and $y_{i2} = I(y_{i2}^* > 0)$ are the observed responses. The asymptotic variance of the marginal ML estimator of μ_2 is $V_M = \Phi(\mu_2)(1-\Phi(\mu_2))/\phi^2(\mu_2)$, where $\phi(\cdot)$ denotes standard normal density function. The asymptotic variance of the joint ML estimator of μ_2 has no closed-form expression, equaling

$$V_J = (1-\rho^2)\left[\frac{1}{E_0} + \frac{(\rho\mu_2 E_0 + E_1)^2}{E_0(E_0E_2 - E_1^2)} + \frac{\rho^2}{1-\rho^2}\left(1 + \frac{(\rho\mu_2)^2}{2}\right)\right],$$

Table 4. Results From the Simulation Study: Average Estimates, Standard Deviations (SD), and Average Estimated Standard Errors (SE) for Large Sample and Weak Correlation

		Full model			Reduced model 1			Reduced model 2		
Parameter	True value	Average estimates	SD	Average SE	Average estimates	SD	Average SE	Average estimates	SD	Average SE
β_{10}	.952	.955	.014	.014	.955	.014	.014	.955	.014	.014
β_{11}	−.087	−.089	.008	.008	−.089	.008	.008	−.089	.008	.008
β_{20}	−2.369	−2.471	.211	.197	−2.462	.209	.194	−2.466	.216	.202
β_{21}	.966	1.016	.108	.115	1.012	.106	.114	1.014	.109	.119
σ_{b1}	.086	.086	.007	.007	.086	.007	.007	.086	.007	.007
σ_{b2}	.822	.804	.114	.102	.799	.114	.100	.796	.116	.102
ρ_b	−.608	−.620	.098	.098	−.642	.098	.097			
σ_{e1}	.075	.075	.002	.002	.075	.002	.002	.075	.002	.002
ρ_e	−.201	−.195	.059	.059						

where

$$E_0 = E\frac{\phi^2(\mu^*)}{\Phi(\mu^*)(1-\Phi(\mu^*))}, \qquad E_1 = E\frac{\phi^2(\mu^*)(\frac{y_1-\mu_1}{\sigma_1})}{\Phi(\mu^*)(1-\Phi(\mu^*))},$$

$$E_2 = E\frac{\phi^2(\mu^*)(\frac{y_1-\mu_1}{\sigma_1})^2}{\Phi(\mu^*)(1-\Phi(\mu^*))},$$

and

$$\mu^* = \frac{\mu_2 + \rho\frac{y_1-\mu_1}{\sigma_1}}{\sqrt{1-\rho^2}}.$$

Then ARE $= V_M/V_J$, which is identically equal to 1 when $\rho = 0$. For $\rho \neq 0$, the ARE depends on ρ and μ_2 such that the formulas are the same for (ρ, μ_2), $(\rho, -\mu_2)$, $(-\rho, \mu_2)$, and $(-\rho, -\mu_2)$.

We considered values of ρ between 0 and .9 in increments of .1 and values of μ_2 between 0 and 3 in increments of .2 (with 0 corresponding to a marginal probability of .5 and 3 corresponding to a marginal probability of .999). Without loss of generality, we set $\mu_1 = 0$ and $\sigma_1 = 1$. The expectations in V_J were closely approximated with Monte Carlo means using simulation sample sizes of 1 million. Results were double-checked for selected settings with a simulation sample size of 10 million. Table 6 shows results for $\rho = .5$, .6, .7, .8, and .9 for selected values of μ_2. For smaller ρ in absolute value, the ARE were uniformly 1.00 to two decimal places. This is intriguing, suggesting that when n is sufficiently large, no advantage results from using joint fitting over separate fitting when ρ is at best moderate. For large ρ, the gains were also minimal, except when $P(y_{i2} = 1)$ was close to 0 or 1; this implies a sample in which the binary data are highly unbalanced, with relatively few observations in one category. These results corroborate the findings from our larger simulation study that in larger samples, the efficiency gains are not substantial for moderately large ρ and not very unbalanced data.

We are unaware of other simulation studies addressing efficiency gains for multivariate mixed outcomes, but a number of authors have reported results from both multivariate and univariate analyses of different examples. Our finding that there are no efficiency gains in large samples is consistent with observations of Lesaffre and Molenberghs (1991) and Fitzmaurice and Laird (1997), who considered probit analysis of bivariate binary outcomes and marginal models for mixed discrete and continuous outcomes. Fitzmaurice and Laird (1997) showed that in a simple paired data model, efficiency gains could be observed when different covariates are included in the linear predictors for the two responses. With two continuous variables, Matsuyama and Ohashi (1997) reported small efficiency gains, whereas Heitjan and Sharma (1997) indicated that ignoring the correlation in a repeated series context may lead to underestimation of standard errors. When missing data are present, the multivariate analysis could

Table 5. Results From the Simulation Study: Average Estimates, Standard Deviations (SD), and Average Estimated Standard Errors (SE) for Small Sample and Strong Correlation

		Full model			Reduced model 1			Reduced model 2		
Parameter	True value	Average estimate	SD	Average SE	Average estimate	SD	Average SE	Average estimate	SD	Average SE
β_{10}	.952	.950	.025	.026	.950	.028	.026	.950	.028	.026
β_{11}	−.087	−.086	.015	.016	−.086	.016	.016	−.086	.016	.016
β_{20}	−2.369	−2.316	.323	.442	−2.377	.455	.482	−2.401	.533	.561
β_{21}	.966	.950	.175	.213	.975	.229	.229	.987	.255	.256
σ_{b1}	.086	.082	.014	.015	.082	.013	.015	.082	.014	.015
σ_{b2}	.822	.721	.197	.216	.761	.197	.223	.766	.293	.272
ρ_b	−.608	−.637	.185	.235	−.788	.130	.219			
σ_{e1}	.075	.076	.005	.004	.076	.005	.004	.076	.005	.004
ρ_e	−.804	−.767	.058	.061						

Table 6. ARE at Different Values of the Correlation Coefficient ρ in the Simple Intercept Model

	μ_2								
ρ	0	.40	.80	1.20	1.60	2.00	2.40	2.80	3.00
0.50	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.01
0.60	1.01	1.01	1.00	1.00	1.01	1.01	1.02	1.02	1.02
0.70	1.02	1.02	1.01	1.01	1.02	1.03	1.04	1.04	1.04
0.80	1.04	1.04	1.02	1.02	1.04	1.07	1.10	1.11	1.11
0.90	1.10	1.09	1.07	1.07	1.11	1.19	1.27	1.31	1.32

lead to even higher efficiency gains (Fitzmaurice and Laird 1997), but if the missing mechanism is not random, the standard error estimates from the univariate analysis might be underestimated (Matsuyama and Ohashi 1997).

We have considered only a few cases thus far, and so our recommendations are tentative. However, it seems that unless strong intracluster correlations exist between the responses and either the sample is small or the binary data are highly unbalanced, it may not be worth the extra effort to fit the responses jointly. Joint fitting may still be necessary to answer multivariate questions, but the efficiency gains may not be great.

6. EXTENSIONS AND DISCUSSION

The correlated probit model generalizes to incorporate any number and combination of binary and continuous response variables. In fact, it also extends to accommodate continuous censored data and ordinal variables for which cutoff points for an underlying continuous random variable are known.

To demonstrate these extensions, we consider the underlying linear mixed model defined in Section 2 in (1)–(3) and denote the first variable by y^*_{i1j} rather than by y_{i1j}. We may observe either $(y^*_{i1j}, y^*_{i2j})^T$, or $(I\{y^*_{i1j} > 0\}, I\{y^*_{i2j} > 0\})$, or $(I\{y^*_{i1j} > \tau_{11}\}, I\{y^*_{i1j} > \tau_{12}\}, \ldots, I\{y^*_{i1j} > \tau_{1,p_1}\}, I\{y^*_{i2j} > \tau_{21}\}, I\{y^*_{i2j} > \tau_{22}\}, \ldots, I\{y^*_{i2j} > \tau_{2,p_2}\})$, where $\tau_{11}, \ldots, \tau_{1,p_1}, \tau_{21}, \ldots, \tau_{2,p_2}$ are known; or

$$y^c_{i1j} = \begin{cases} y^*_{i1j} & \text{if } y^*_{i1j} > \gamma_1 \\ \gamma_{y1} & \text{if } y^*_{i1j} \le \gamma_1 \end{cases}$$

and

$$y^c_{i2j} = \begin{cases} y^*_{i2j} & \text{if } y^*_{i2j} > \gamma_2 \\ \gamma_{y2} & \text{if } y^*_{i2j} \le \gamma_2 \end{cases}$$

where γ_1, γ_2, γ_{y1}, and γ_{y2} are known; or any combination of these. For all of such cases, the complete-data ML estimate is the ML estimate for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\Sigma}_b$, and $\boldsymbol{\Sigma}_e$ from the mixed model in Section 3. Therefore, the M step in the ECM algorithm is the same. Each E step requires expressions or approximations of $E(\mathbf{y}^*_i|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)})$ and $E(\mathbf{y}^*_i\mathbf{y}^{*T}_i|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)})$, where $\mathbf{y}_i$ is the observed response vector for subject i. If $\mathbf{y}_{i1\_} = \mathbf{y}^*_{i1\_}$, then we do not need to generate values for this response and simply use $E(\mathbf{y}^*_{i1\_}|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}) = \mathbf{y}_{i1\_}$ and $E(\mathbf{y}^*_{i1\_}\mathbf{y}^{*T}_{i1\_}|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}) = \mathbf{y}_{i1\_}\mathbf{y}^T_{i1\_}$, and similarly for the other response. Otherwise, we do need to generate values.

The ordinal case is handled similarly to the binary case, except that generated values for the truncated multivariate normal distribution must fall in the region specified by the ordinary response. In the censored-data case, the response is kept unchanged if it corresponds to an uncensored observation; that is, $E(y^*_{i1j}|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}) = y^c_{i1j}$ and $E(y^*_{i1j}y^{*T}_{i1j}|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}) = y^c_{i1j}y^{cT}_{i1j}$, if $y^c_{i1j} \neq \gamma_{y1}$. But if $y^c_{i1j} = \gamma_{y1}$, then values are generated from the truncated normal distribution of $(y^*_{i1j}|y^{*(-)}_i, y_{i1j}) = \gamma_{y1}$ as in the binary case. (The truncation in this particular example is from above at γ_1.) Then $E(y^*_{i1j}|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}) = \frac{1}{m}\sum_{k=1}^{m} y^{*(k)}_{i1j}$, and the variance is handled similarly.

Another possible extension of the correlated probit model incorporates more general correlation structures at both the random effects and the random error levels. One possibility is an autoregressive structure, which allows modeling of various longitudinal datasets.

A drawback of using the Gibbs sampler instead of multivariate rejection sampling is the lack of assurance about convergence of the algorithm. Chan and Kuk (1997) proposed using several runs, but this is usually too slow for problems such as the current one. A key to solving this problem is the choice of weights to yield faster convergence of the Monte Carlo ECM algorithm. We noticed that the algorithms had convergence problems when the initial estimates were far from the ML estimates, and this can probably be remedied if the weights are wisely chosen. Generally, computational issues remain a concern if the correlated probit model is to be used for datasets more complex than the example given in this article.

There are other possible approaches to speed up the Monte Carlo ECM algorithm; for example, the working parameter approach of Meng and van Dyk (1998). But we doubt that the latter would outperform the SAECM approach. Although it seems to improve on the EM and ECM approaches by reducing the number of iterations to convergence, it does not simplify the E step, the most computationally intensive part of our algorithm.

The problem of checking the assumptions for the correlated probit model is also important. The effects of incorrectly specifying the random-effects distribution are of special interest, because the problem of joint versus separate fitting of the response variables can be addressed from that perspective. Neuhaus, Hauck, and Kalbfleisch (1992) investigated the effects of misspecified random-effects distribution in mixed-effects logistic models and found that the bias in the regression parameter estimates is usually small. It will not be surprising if this is the case for the correlated probit model, but the issue remains to be studied.

APPENDIX: MONTE CARLO ECM ALGORITHM

For the E step of the Monte Carlo ECM algorithm, we let

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i\_1} \\ \vdots \\ \mathbf{x}_{i\_n_i} \end{pmatrix}, \quad \mathbf{Z}_i = \begin{pmatrix} \mathbf{z}_{i\_1} \\ \vdots \\ \mathbf{z}_{i\_n_i} \end{pmatrix}, \quad \mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{i\_1} \\ \vdots \\ \mathbf{y}_{i\_n_i} \end{pmatrix},$$

$$\mathbf{y}^*_i = \begin{pmatrix} \mathbf{y}^*_{i\_1} \\ \vdots \\ \mathbf{y}^*_{i\_n_i} \end{pmatrix}, \quad \mathbf{b}_i = \begin{pmatrix} \mathbf{b}_{i1} \\ \mathbf{b}_{i2} \end{pmatrix}.$$

From the model definition, the joint distribution of the complete data is

$$\begin{pmatrix}\mathbf{y}_i^*\\ \mathbf{b}_i\end{pmatrix} \sim N\left[\begin{pmatrix}\mathbf{X}_i\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i^T+\mathbf{I}_{n_i}\otimes\boldsymbol{\Sigma}_e & \mathbf{Z}_i\boldsymbol{\Sigma}\\ \boldsymbol{\Sigma}\mathbf{Z}_i^T & \boldsymbol{\Sigma}\end{pmatrix}\right]. \quad \text{(A.1)}$$

The conditional distribution of $\mathbf{b}_i$ given the complete response $\mathbf{y}_i^*$ is $N[\boldsymbol{\Sigma}_{B_i}(\mathbf{y}_i^*-\mathbf{X}_i\boldsymbol{\beta}), \boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{B_i}\mathbf{Z}_i\boldsymbol{\Sigma}]$, where $\boldsymbol{\Sigma}_{E_i}=\mathbf{I}_{n_i}\otimes\boldsymbol{\Sigma}_e$ and $\boldsymbol{\Sigma}_{B_i}=\boldsymbol{\Sigma}\mathbf{Z}_i^T(\mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i^T+\boldsymbol{\Sigma}_{E_i})^{-1}$. Therefore,

$$\hat{\boldsymbol{\Sigma}}^{(r+1)}=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\Sigma}}^{(r)}-\hat{\boldsymbol{\Sigma}}_{B_i}^{(r)}\mathbf{Z}_i\hat{\boldsymbol{\Sigma}}^{(r)}+\hat{\boldsymbol{\Sigma}}_{B_i}^{(r)}\hat{\mathbf{V}}_i^{(r)}\boldsymbol{\Sigma}_{B_i}^{(r)T}\right), \quad \text{(A.2)}$$

where $\hat{\mathbf{V}}_i^{(r)}=E[(\mathbf{y}_i^*-\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(r)})(\mathbf{y}_i^*-\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(r)})^T|\mathbf{y}_i,\hat{\boldsymbol{\psi}}^{(r)}]$,

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_e^{(r+1)}=&\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}E\left(\left(\mathbf{y}_{i\_j}^*-\mathbf{x}_{i\_j}\hat{\boldsymbol{\beta}}^{(r)}\right)\right.\\ &\left.\times\left(\mathbf{y}_{i\_j}^*-\mathbf{x}_{i\_j}\hat{\boldsymbol{\beta}}^{(r)}\right)|\mathbf{y}_{i\_j},\hat{\boldsymbol{\psi}}^{(r)}\right)\\ &-\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}E\left(\left(\mathbf{y}_{i\_j}^*-\mathbf{x}_{i\_j}\hat{\boldsymbol{\beta}}^{(r)}\right)\mathbf{b}_i^T\mathbf{z}_{i\_j}^T|\mathbf{y}_{i\_j},\hat{\boldsymbol{\psi}}^{(r)}\right)\\ &-\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}E\left(\mathbf{z}_{i\_j}\mathbf{b}_i\left(\mathbf{y}_{i\_j}^*-\mathbf{x}_{i\_j}\hat{\boldsymbol{\beta}}^{(r)}\right)^T|\mathbf{y}_{i\_j},\hat{\boldsymbol{\psi}}^{(r)}\right)\\ &+\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}E\left(\mathbf{z}_{i\_j}\mathbf{b}_i\mathbf{b}_i^T\mathbf{z}_{i\_j}|\mathbf{y}_{i\_j},\hat{\boldsymbol{\psi}}^{(r)}\right), \quad \text{(A.3)}\end{aligned}$$

and

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(r+1)}=&\left(\sum_{i=1}^{n}\mathbf{X}_i\left(\hat{\boldsymbol{\Sigma}}_{E_i}^{(r+1)}\right)^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i^T\left(\hat{\boldsymbol{\Sigma}}_{E_i}^{(r+1)}\right)^{-1}\right.\\ &\times\left(E(\mathbf{y}_i^*|\mathbf{y}_i)-\mathbf{Z}_i\hat{\boldsymbol{\Sigma}}_{B_i}^{(r+1)}\right.\\ &\left.\left.\times\left(E\left(\mathbf{y}_i^*|\mathbf{y}_i,\hat{\boldsymbol{\psi}}^{(r)}\right)-\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(r)}\right)\right)\right). \quad \text{(A.4)}\end{aligned}$$

Because

$$\begin{aligned}E(\mathbf{y}_{i\_j}^*\mathbf{b}_i^T|\mathbf{y}_{i\_j})&=E[E(\mathbf{y}_{i\_j}^*\mathbf{b}_i^T|\mathbf{y}_{i\_j}^*)|\mathbf{y}_{i\_j}]\\ &=E\{\mathbf{y}_{i\_j}^*[E(\mathbf{b}_i^T|\mathbf{y}_{i\_j}^*)|\mathbf{y}_{i\_j}]\}\\ &=E[\mathbf{y}_{i\_j}^*(\mathbf{y}_i^*-\mathbf{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_{B_i}^T|\mathbf{y}_{i\_j}]\end{aligned}$$

all conditional expectations depend only on $E(\mathbf{y}_i^*|\mathbf{y}_i,\hat{\boldsymbol{\psi}}^{(r)})$ and $\text{var}\,(\mathbf{y}_i^*|\mathbf{y}_i,\hat{\boldsymbol{\psi}}^{(r)})$. Note that $E(y_{i1j}^*|\mathbf{y}_i,\hat{\boldsymbol{\psi}}^{(r)})=y_{i1j}$ for $j=1,\ldots n_i$, and so we only need to approximate $E(\mathbf{y}_{i2\_}^*|\mathbf{y}_i,\hat{\boldsymbol{\psi}}^{(r)})$ and $\text{var}\,(\mathbf{y}_{i2\_}^*|\mathbf{y}_i,\hat{\boldsymbol{\psi}}^{(r)})$.

Because

$$f(\mathbf{y}_{i2\_}^*|\mathbf{y}_i)=\begin{cases}\dfrac{f(\mathbf{y}_{i2\_}^*|\mathbf{y}_{i1\_})}{c_i} & \text{if } \mathbf{y}_{i2\_}\in\mathbf{A}_i\\ 0 & \text{otherwise,}\end{cases}$$

where $c_i=P(\mathbf{y}_{i2\_}^*\in\mathbf{A}_i)$ and $\mathbf{A}_i=\{\mathbf{y}_{i2\_}^*: y_{i2j}^*>0 \text{ if } y_{i2j}=1 \;\&\; y_{i2j}^*<0 \text{ if } y_{i2j}=0\}$,

$$E(\mathbf{y}_{i2\_}^*|\mathbf{y}_{i1\_},\mathbf{y}_{i2\_},\hat{\boldsymbol{\psi}}^{(r)})=\frac{1}{c_i}E\left[\mathbf{y}_{i2\_}^*I\{\mathbf{y}_{i2\_}^*\in\mathbf{A}_i\}|\mathbf{y}_{i1\_},\hat{\boldsymbol{\psi}}^{(r)}\right].$$

Monte Carlo approximations of $E(\mathbf{y}_{i2\_}^*|\mathbf{y}_{i1\_},\mathbf{y}_{i2\_},\hat{\boldsymbol{\psi}}^{(r)})$ and $\text{var}(\mathbf{y}_{i2\_}^*|\mathbf{y}_{i1\_},\mathbf{y}_{i2\_},\hat{\boldsymbol{\psi}}^{(r)})$ are

$$\frac{1}{m^*}\sum_{l=1}^{m^*}\mathbf{y}_{i2\_}^{*(l)} \quad \text{and}$$

$$\frac{1}{m^*}\sum_{l=1}^{m^*}\mathbf{y}_{i2\_}^{*(l)}\mathbf{y}_{i2\_}^{*(l)T}-\left(\frac{1}{m^*}\sum_{l=1}^{m^*}\mathbf{y}_{i2\_}^{*(l)}\right)\left(\frac{1}{m^*}\sum_{l=1}^{m^*}\mathbf{y}_{i2\_}^{*(l)T}\right), \quad \text{(A.5)}$$

where $\mathbf{y}_{i2\_}^{*(l)}$ are simulated values from the distribution of $\{\mathbf{y}_{i2\_}^*|\mathbf{y}_{i1\_},\hat{\boldsymbol{\psi}}^{(r)})\}$.

To simulate values from the distribution of $\{\mathbf{y}_{i2\_}^*|\mathbf{y}_{i1\_},\hat{\boldsymbol{\psi}}^{(r)}\}$ that fall in $\mathbf{A}_i=\{\mathbf{y}_{i2\_}^*: y_{i2j}^*>0 \text{ if } y_{i2j}=1 \;\&\; y_{i2j}^*<0 \text{ if } y_{i2j}=0\}$, multivariate rejection sampling is not feasible. A more practical alternative uses Gibbs sampling, as proposed by Chan and Kuk (1997). Let $\mathbf{y}_{i2j}^{*-}$ denote $\mathbf{y}_{i2\_}^*$ with y_{i2j}^* omitted. To obtain the $(r+1)$st sample from the distribution of $\{\mathbf{y}_{i2\_}^*|\mathbf{y}_{i1\_},y_{i2j},\hat{\boldsymbol{\psi}}^{(r)}\}$, one iteratively generates values from

$$\begin{aligned}&f(y_{i21}^{*(r+1)}|\mathbf{y}_{i1\_},y_{i21},\mathbf{y}_{i21}^{*(r)-},\hat{\boldsymbol{\psi}}^{(r)}),\\ &f(y_{i22}^{*(r+1)}|\mathbf{y}_{i1\_},y_{i22},y_{i21}^{*(r+1)},y_{i23}^{*(r)},\ldots,y_{i2n_i}^{*(r)},\hat{\boldsymbol{\psi}}^{(r)}),\\ &\ldots\\ &f(y_{i2n_i}^{*(r+1)}|\mathbf{y}_{i1\_},y_{i2n_i},y_{i2n_i}^{*(r+1)},\hat{\boldsymbol{\psi}}^{(r)}).\end{aligned}$$

[Received February 2000. Revised November 2000.]

REFERENCES

Catalano, P. J., and Ryan, L. M. (1992), "Bivariate Latent Variable Models for Clustered Discrete and Continuous Outcomes," *Journal of the American Statistical Association,* 87, 651–658.

Chan, J. S. K., and Kuk, A. Y. C. (1997), "Maximum Likelihood Estimation for Probit-Linear Mixed Models With Correlated Random Effects," *Biometrics,* 53, 86–97.

Delyon, B., Lavielle M., and Moulines, E. (1999), "On a Stochastic Approximation Version of the EM Algorithm," *The Annals of Statistics,* 27, 94–128.

Dunson, D. B. (2000), "Bayesian Latent Variable Models for Clustered Mixed Outcomes," *Journal of the Royal Statistical Society,* Series B, 62, 355–366.

Fitzmaurice, G. M., and Laird, N. M. (1995), "Regression Models for a Bivariate Discrete and Continuous Outcome With Clustering," *Journal of the American Statistical Association,* 90, 845–852.

——— (1997), "Regression Models for Mixed Discrete and Continuous Responses With Potentially Missing Values," *Biometrics,* 53, 110–122.

Gueorguieva, R. V. (1999), "Models for Repeated Measures of a Multivariate Response," Doctoral dissertation, University of Florida, Gainesville, Dept. of Statistics.

Heitjan, D. F., and Sharma, D. (1997), "Modelling Repeated-Series Longitudinal Data," *Statistics in Medicine,* 16, 347–355.

Lesaffre, E., and Molenberghs, G. (1991), "Multivariate Probit Analysis: A Neglected Procedure in Medical Statistics," *Statistics in Medicine,* 10, 1391–1403.

Liao, J. (1999), "A Simplified and Accelerated Monte Carlo EM Algorithm With Application to a Hierarchical Mixture Model," technical report.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998), "Parameter Expansion to Accelerate EM: The PX-EM Algorithm," *Biometrika,* 85, 755–770.

Matsuyama, Y., and Ohashi, Y. (1997), "Mixed Models for Bivariate Response Repeated Measures Data Using Gibbs Sampling," *Statistics in Medicine,* 16, 1587–1601.

Meng, X. L., and Rubin, D. B. (1993), "Maximum Likelihood Estimation via the ECM Algorithm. A General Framework," *Biometrika,* 80, 267–78.

Meng, X. L., and van Dyk, D. (1998), "Fast EM-Type Implementations for Mixed Effects Models," *Journal of the Royal Statistical Society,* Series B, 60, 559–578.

Neuhaus, J. M., Hauck, W. W., and Kalbfleisch, J. D. (1992), "The Effects of Mixture Distribution Misspecification When Fitting Mixed Effects Logistic Models," *Biometrika,* 79, 755–762.

Price, C. J., Kimmel, C. A., Tyl, R. W., and Marr, M. C. (1985), "The Developmental Toxicity of Ethylene Glycol in Rats and Mice," *Toxicological Applications in Pharmacology,* 81, 825–839.

Regan, M. M., and Catalano, P. J. (1999), "Likelihood Models for Clustered Binary and Continuous Outcomes: Application to Developmental Toxicology," *Biometrics,* 55, 760–768.

Rochon, J. (1996), "Analyzing Bivariate Repeated Measures for Discrete and Continuous Outcome Variables," *Biometrics,* 52, 740–750.

Ruppert, D. (1991), "Stochastic Approximation," in *Handbook of Sequential Analysis,* New York: Marcel Dekker, pp. 503–529.