Maximum Likelihood Estimation for Probit-Linear Mixed Models with Correlated Random Effects

Jennifer S. K. Chan; Anthony Y. C. Kuk

Biometrics, Vol. 53, No. 1. (Mar., 1997), pp. 86-97.

BIOMETRICS 53, 86–97
March 1997

Maximum Likelihood Estimation for Probit-Linear Mixed Models with Correlated Random Effects

Jennifer S. K. Chan and Anthony Y. C. Kuk
Department of Statistics, University of New South Wales, Sydney 2052, Australia

SUMMARY

The probit-normal model for binary data (McCulloch, 1994, *Journal of the American Statistical Association* **89**, 330–335) is extended to allow correlated random effects. To obtain maximum likelihood estimates, we use the EM algorithm with its M-step greatly simplified under the assumption of a probit link and its E-step made feasible by Gibbs sampling. Standard errors are calculated by inverting a Monte Carlo approximation of the information matrix rather than via the SEM algorithm. A method is also suggested that accounts for the Monte Carlo variation explicitly. As an illustration, we present a new analysis of the famous salamander mating data. Unlike previous analyses, we find it necessary to introduce different variance components for different species of animals. Finally, we consider models with correlated errors as well as correlated random effects.

1. Introduction

The analysis of correlated binary data has received a lot of attention in recent years. A number of authors have proposed the use of a mixed effects generalized linear model under which the probability of a positive response, when transformed by a suitable link function, is a linear function of fixed as well as random effects. The incorporation of random effects into the model is useful in accounting for population heterogeneity, overdispersion, and intracluster correlation. It also enables the pooling of information across different subjects to result in better subject-specific inference as opposed to population-averaged inference. The presence of random effects, however, complicates the estimation problem considerably. To obtain the marginal likelihood function, one has to integrate out the random effects, which, except for a few special cases, cannot be performed analytically. The intractability of the likelihood function has led various authors to propose a host of alternative estimation methods rather than carrying out maximum likelihood estimation exactly. These include the approximate maximum likelihood and approximate residual maximum likelihood estimators proposed by Schall (1991), McGilchrist (1994), and Drum and McCullagh (1993); the penalized quasi-likelihood approach of Breslow and Clayton (1993); the Gibbs sampling Bayesian approach of Zeger and Karim (1991); the estimating function approach of Waclawiw and Liang (1993); and the iterative bias correction approach of Kuk (1995).

McCulloch (1994) pointed out several advantages of using the probit link instead of the customary logit link. In his paper McCulloch considers only independent random effects. We extend McCulloch's model by allowing correlated random effects. This extension widens the applicability of the model considerably and is described in Section 2. In Section 3 we describe the EM algorithm for obtaining maximum likelihood estimates and highlight the simplification made possible by the probit link assumption. The derivation is similar to that of McCulloch (1994), but we end up with a slightly different formula for the M-step. A Monte Carlo implementation of the E-step of the algorithm via Gibbs sampling is given in Section 4. In Section 5 we consider the estimation of standard errors, a topic not dealt with by McCulloch (1994) in any detail. When the E-step requires a Monte Carlo method, the SEM algorithm (Meng and Rubin, 1991) for calculating standard errors is numerically unstable and extremely computer intensive, and so we resort to inverting a Monte

---

Key words: EM algorithm; Gibbs sampling; Monte Carlo method; Probit-normal models; Random effects; Threshold models.

Carlo estimate of the information matrix. In Section 6 we describe a method of accounting for the Monte Carlo variation explicity. In Section 7 we illustrate the flexibility and feasibility of our methods by fitting two models to the salamander mating data reported in McCullagh and Nelder (1989, pp. 439–450). Both models assume a probit link and include the male species, the female species, their interaction, and season (fall versus summer) as fixed effects. Model 1 includes male and female animal effects as random effects. As the same animals were used in the first two mating experiments, the effects of the same animal over the two occasions are correlated. In model 2 the random effects are classified by species as well as by gender. Finally, in Section 8 we extend our model to allow correlated errors as well as correlated random effects. Applications to longitudinal binary data and multivariate clustered binary data are considered briefly.

2. The Model

Let $W_1, \dots, W_n$ denote the observed binary variables. Following McCulloch (1994), we assume that the probabilities $P_i = \Pr(W_i = 1)$ are probit-linear. In matrix form

$$\Phi^{-1}(\mathbf{P}) = \left(\Phi^{-1}(P_1), \dots, \Phi^{-1}(P_n)\right)^{\mathrm{T}} = \mathbf{X}\beta + \sum_{r=1}^{R} \mathbf{Z}_r \mathbf{u}_r,$$

where $\mathbf{X}$ is an $n \times p$ design matrix, β is a $p \times 1$ vector of fixed effects, $\mathbf{u}_r$ is a $q_r k_r \times 1$ vector of random effects with corresponding $n \times q_r k_r$ design matrix $\mathbf{Z}_r$. We assume that $\mathbf{u}_1, \dots, \mathbf{u}_R$ are independent. For each $\mathbf{u}_r$ we have

$$\mathbf{u}_r = \begin{pmatrix} \mathbf{u}_{r1} \\ \vdots \\ \mathbf{u}_{rq_r} \end{pmatrix} \sim \mathcal{N}_{q_r \times k_r} \left\{ \mathbf{0}, \begin{pmatrix} \Sigma_r & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \Sigma_r \end{pmatrix} \right\} = \mathcal{N}_{q_r \times k_r}(\mathbf{0}, \mathbf{I}_{q_r} \otimes \Sigma_r). \tag{1}$$

In other words, $\mathbf{u}_r$ is made up of q_r i.i.d. random vectors of dimension k_r each.

For the purpose of estimation, it is useful to view the above probit-linear mixed model as a threshold model that results from dichotomizing the observations from a Gaussian mixed model. In other words, $W_i = I_{(Y_i > 0)}$ and

$$\mathbf{Y} = \mathbf{X}\beta + \sum_{r=1}^{R} \mathbf{Z}_r \mathbf{u}_r + \varepsilon, \tag{2}$$

where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ independently of the $\mathbf{u}_r$.

The above model allows correlated random effects and so is an extension of the model proposed by McCulloch (1994), which considers only independent random effects. The extended model is quite general and includes the following useful models as special cases.

McCulloch's model. This model is obtained by setting $k_r \equiv 1$.

Random coefficient regression model for clustered binary data. Let $W_{ij}, i = 1, \dots, m$ and $j = 1, \dots, n_i$, represent the jth observation from the ith cluster. We assume $W_{ij} = I_{(Y_{ij} > 0)}$, where

$$Y_{ij} = \mathbf{x}_{ij} \mathbf{b}_i + \varepsilon_{ij},$$

$\mathbf{x}_{ij}$ denotes a $1 \times p$ vector of covariates, and $\mathbf{b}_i$ denotes the corresponding cluster-specific $p \times 1$ vector of regression coefficients. If we assume that the m cluster-specific vectors of regression coefficients $\mathbf{b}_1, \dots, \mathbf{b}_m$ are i.i.d. $\mathcal{N}(\beta, \Sigma)$, we can rewrite $\mathbf{b}_i$ as $\beta + \mathbf{u}_i$ and we get

$$Y_{ij} = \mathbf{x}_{ij}\beta + \mathbf{x}_{ij}\mathbf{u}_i + \varepsilon_{ij}.$$

Putting it in vector form, we have

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{X}\mathbf{u} + \varepsilon.$$

This model corresponds to (2) with $R = 1$, $\mathbf{Z}_1 = \mathbf{X}$, $q_1 = m$, and $k_1 = p$.

Random intercept model for bivariate clustered binary data. Let W_{1ij}, W_{2ij} denote the two binary variables observed for unit j of cluster i. We assume $W_{kij} = I_{(Y_{kij} > 0)}$, where

$$Y_{1ij} = u_{1i} + x_{ij}\beta_1 + \varepsilon_{1ij},$$
$$Y_{2ij} = u_{2i} + x_{ij}\beta_2 + \varepsilon_{2ij}.$$

In other words, we have a random intercept model for each variable and the random intercepts $(u_{1i}, u_{2i})^{\mathrm{T}}$ are i.i.d. according to a bivariate normal distribution with zero mean. This model is a special case of (2) with $R = 1, q_1 = m$, and $k_1 = 2$.

Models involving crossed design of correlated random effects. Two complicated models of crossed design are illustrated in Section 7 through the study of the famous salamander data.

3. Maximum Likelihood Estimation

In this section we describe the maximum likelihood estimation of the fixed effects β and the variance components $\Sigma_r, r = 1, \ldots, R$, by the EM algorithm. To apply the EM algorithm we treat the complete data as $\mathbf{Y}, \mathbf{u}_1, \ldots, \mathbf{u}_R$ and regard the observed data $\mathbf{W}$ as the incomplete data. The EM algorithm is particularly suited for the probit-normal model given by (1) and (2) as closed form formulae for the complete-data MLE exist, and they are

$$\hat{\beta}_c = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\left(\mathbf{Y} - \sum_{r=1}^{R}\mathbf{Z}_r\mathbf{u}_r\right) \tag{3}$$

and

$$\hat{\Sigma}_{rc} = \left(\sum_{j=1}^{q_r}\mathbf{u}_{rj}\mathbf{u}_{rj}^{\mathrm{T}}\right)/q_r. \tag{4}$$

Instead of (3), McCulloch (1994) used $\hat{\beta}(\mathbf{Y}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Y}$, where $\mathbf{V}$ is the covariance matrix of $\mathbf{Y}$. Note that $\hat{\beta}(\mathbf{Y})$ is the MLE for β based on $\mathbf{Y}$ alone rather than based on the complete data $\mathbf{Y}, \mathbf{u}_1, \ldots, \mathbf{u}_R$. As pointed out by the referees, McCulloch's procedure is closely related to version 1 of the ECME algorithm proposed by Liu and Rubin (1994, p. 641).

Given the current estimates $\beta^{(k)}$ and $\Sigma_r^{(k)}$, $r = 1, \ldots, R$, we update our parameter estimates by replacing $\mathbf{Y}$, $\mathbf{u}_r$, and $\mathbf{u}_{rj}\mathbf{u}_{rj}^{\mathrm{T}}$ in (3) and (4) by their conditional expectations evaluated at the current parameter estimates to get

$$\beta^{(k+1)} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\left(\mathrm{E}[\mathbf{Y} \mid \mathbf{W}] - \sum_{r=1}^{R}\mathbf{Z}_r\mathrm{E}[\mathbf{u}_r \mid \mathbf{W}]\right) \tag{5}$$

and

$$\Sigma_r^{(k+1)} = \left(\sum_{j=1}^{q_r}\mathrm{E}[\mathbf{u}_{rj}\mathbf{u}_{rj}^{\mathrm{T}} \mid \mathbf{W}]\right)/q_r. \tag{6}$$

The steps are then iterated until convergence is achieved. Now

$$\mathrm{E}[\mathbf{u}_r \mid \mathbf{W}] = \mathrm{E}[\mathrm{E}[\mathbf{u}_r \mid \mathbf{Y}] \mid \mathbf{W}], \tag{7}$$

$$\mathrm{E}[\mathbf{u}_{rj}\mathbf{u}_{rj}^{\mathrm{T}} \mid \mathbf{W}] = \mathrm{E}[\mathrm{E}[\mathbf{u}_{rj}\mathbf{u}_{rj}^{\mathrm{T}} \mid \mathbf{Y}] \mid \mathbf{W}]. \tag{8}$$

To find the inner expectation, we use the fact that the joint distribution of $\mathbf{Y}, \mathbf{u}_1, \ldots, \mathbf{u}_R$ are

$$\begin{pmatrix} Y \\ u_1 \\ \vdots \\ u_R \end{pmatrix} \sim \mathcal{N}_{n+Q}\left\{\begin{pmatrix} X\beta \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} V & Z_1(I_{q_1} \otimes \Sigma_1) & \cdots & \cdots & Z_R(I_{q_R} \otimes \Sigma_R) \\ (I_{q_1} \otimes \Sigma_1)Z_1^{\mathrm{T}} & (I_{q_1} \otimes \Sigma_1) & 0 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ (I_{q_R} \otimes \Sigma_R)Z_R^{\mathrm{T}} & 0 & \cdots & 0 & (I_{q_R} \otimes \Sigma_R) \end{pmatrix}\right\},$$

where

$$\mathbf{V} = \mathbf{I}_n + \sum_{r=1}^{R}\mathbf{Z}_r(\mathbf{I}_{q_r} \otimes \Sigma_r)\mathbf{Z}_r^{\mathrm{T}} \tag{9}$$

and $Q = \Sigma_{r=1}^{R} q_r k_r$. It follows that

$$\mathrm{E}[\mathbf{u}_r \mid \mathbf{Y}] = (\mathbf{I}_{q_r} \otimes \Sigma_r)\mathbf{Z}_r^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\beta) \tag{10}$$

and

$$\mathrm{E}[\mathbf{u}_{rj}\mathbf{u}_{rj}^{\mathrm{T}} \mid \mathbf{Y}] = \boldsymbol{\Sigma}_r\mathbf{Z}_{rj}^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Z}_{rj}\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_r - \boldsymbol{\Sigma}_r\mathbf{Z}_{rj}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Z}_{rj}\boldsymbol{\Sigma}_r, \tag{11}$$

where $\mathbf{Z}_{rj}$ is the $n \times k_r$ submatrix of $\mathbf{Z}_r = (\mathbf{Z}_{r1}, \ldots, \mathbf{Z}_{rq_r})$ that corresponds to $\mathbf{u}_{rj}$. Substituting (10) and (11) into (7) and (8), we obtain

$$\mathrm{E}[\mathbf{u}_r \mid \mathbf{W}] = (\mathbf{I}_{q_r} \otimes \boldsymbol{\Sigma}_r)\mathbf{Z}_r^{\mathrm{T}}\mathbf{V}^{-1}(\mathrm{E}[\mathbf{Y} \mid \mathbf{W}] - \mathbf{X}\boldsymbol{\beta}) \tag{12}$$

and

$$\begin{aligned}\mathrm{E}[\mathbf{u}_{rj}\mathbf{u}_{rj}^{\mathrm{T}} \mid \mathbf{W}] = \boldsymbol{\Sigma}_r\mathbf{Z}_{rj}^{\mathrm{T}}\mathbf{V}^{-1}\{&\mathrm{E}[\mathbf{Y} \mid \mathbf{W}]\mathrm{E}[\mathbf{Y} \mid \mathbf{W}]^{\mathrm{T}} + \mathrm{cov}[\mathbf{Y} \mid \mathbf{W}] - \mathrm{E}[\mathbf{Y} \mid \mathbf{W}](\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} \\ &- (\mathbf{X}\boldsymbol{\beta})\mathrm{E}[\mathbf{Y} \mid \mathbf{W}]^{\mathrm{T}} + (\mathbf{X}\boldsymbol{\beta})(\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\} \\ &\cdot \mathbf{V}^{-1}\mathbf{Z}_{rj}\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_r - \boldsymbol{\Sigma}_r\mathbf{Z}_{rj}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Z}_{rj}\boldsymbol{\Sigma}_r.\end{aligned} \tag{13}$$

We can see from (12) and (13) that $\mathrm{E}[\mathbf{u}_r \mid \mathbf{W}]$ and $\mathrm{E}[\mathbf{u}_{rj}\mathbf{u}_{rj}^{\mathrm{T}} \mid \mathbf{W}]$ can be expressed entirely in terms of the conditional mean and the conditional covariance matrix of $\mathbf{Y}$ given $\mathbf{W}$. Thus, $\mathrm{E}[\mathbf{Y} \mid \mathbf{W}]$ and $\mathrm{cov}[\mathbf{Y} \mid \mathbf{W}]$ are all that we need to carry out the EM iterations in (5) and (6).

4. The Monte Carlo EM Algorithm

Unfortunately, $\mathrm{E}[\mathbf{Y} \mid \mathbf{W}]$ and $\mathrm{cov}[\mathbf{Y} \mid \mathbf{W}]$ which are required for the EM algorithm cannot be expressed in closed form. In such circumstances, Wei and Tanner (1990) and McCulloch (1994) suggest the so-called Monte Carlo EM algorithm where $\mathrm{E}[\mathbf{Y} \mid \mathbf{W}]$ and $\mathrm{cov}[\mathbf{Y} \mid \mathbf{W}]$ are approximated by simulations. Since it is difficult to simulate directly from the conditional distribution of $\mathbf{Y}$ given $\mathbf{W}$, we use the method of Gibbs sampling (Smith and Roberts, 1993). The details are given as follows.

Let us first denote the vector obtained by taking the diagonal and lower triangular elements of $\boldsymbol{\Sigma}_r$ by $\mathrm{vec}\boldsymbol{\Sigma}_r$ and the $P \times 1$ vector of all parameters $(\boldsymbol{\beta}^{\mathrm{T}}, \mathrm{vec}\boldsymbol{\Sigma}_1{}^{\mathrm{T}}, \ldots, \mathrm{vec}\boldsymbol{\Sigma}_R{}^{\mathrm{T}})^{\mathrm{T}}$ by $\boldsymbol{\theta}$, where $P = p + \Sigma_{r=1}^{R} q_r(q_r+1)/2$. Starting with some initial values $\mathbf{Y}^{(0)} = (Y_1^{(0)}, \ldots, Y_n^{(0)})^{\mathrm{T}}$ consistent in signs with the observed data $\mathbf{W}$, we proceed to generate $\mathbf{Y}^{(b)} = (Y_1^{(b)}, \ldots, Y_n^{(b)})^{\mathrm{T}}$, $b = 1, 2, \ldots,$ sequentially in the following manner.

Given $\mathbf{Y}^{(b)}$ and the current estimates $\boldsymbol{\theta}^{(k)}$, we simulate

$$\begin{aligned} &Y_1^{(b+1)} \text{ from } f(y_1 \mid y_2^{(b)}, y_3^{(b)}, \ldots, y_n^{(b)}, \mathbf{W}; \boldsymbol{\theta}^{(k)}) \\ &\vdots \\ &Y_i^{(b+1)} \text{ from } f(y_i \mid y_1^{(b+1)}, y_2^{(b+1)}, \ldots, y_{i-1}^{(b+1)}, y_{i+1}^{(b)}, \ldots, y_n^{(b)}, \mathbf{W}; \boldsymbol{\theta}^{(k)}) \\ &\vdots \\ &Y_n^{(b+1)} \text{ from } f(y_n \mid y_1^{(b+1)}, y_2^{(b+1)}, \ldots, y_{n-1}^{(b+1)}, \mathbf{W}; \boldsymbol{\theta}^{(k)}). \end{aligned} \tag{14}$$

To carry out the above simulations, we use the fact that

$$f(y_i \mid y_j, j \neq i; w_1, \ldots, w_n) = f(y_i \mid y_j, j \neq i; w_i).$$

Since $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is $\mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ with $\mathbf{V}$ given by (9), it follows that $f(y_i \mid y_j, j \neq i)$ is also normal. Thus, $f(y_i \mid y_j, j \neq i; w_i)$ is a truncated (above 0 if $w_i = 1$; below 0 if $w_i = 0$) normal distribution. We use the fast acceptance–rejection method of Marsaglia (1964) to simulate from such a truncated normal distribution. In this way, we can carry out the Gibbs sampling cycles (14) to result in $\mathbf{Y}^{(b)} = (Y_1^{(b)}, \ldots, Y_n^{(b)})^{\mathrm{T}}$, $b \geq 1$. According to the theory of Gibbs sampling, $\{\mathbf{Y}^{(b)}\}$ is a Markov chain whose stationary distribution is $f(\mathbf{y} \mid \mathbf{w})$ from which we wish to sample. In our analysis we discard the first $T = 200$ elements of the Markov chain $\{\mathbf{Y}^{(b)}\}$ as transient values and treat the following $B = 1000$ elements of $\{\mathbf{Y}^{(b)}\}$ as realizations from the conditional distribution of $\mathbf{Y}$ given $\mathbf{W}$. We use their sample mean and sample covariance matrix to approximate $\mathrm{E}[\mathbf{Y} \mid \mathbf{W}]$ and $\mathrm{cov}[\mathbf{Y} \mid \mathbf{W}]$. To be precise, we have

$$\hat{\mathrm{E}}[\mathbf{Y} \mid \mathbf{W}] = \bar{\mathbf{Y}} = \frac{1}{B}\sum_{b=T+1}^{T+B} \mathbf{Y}^{(b)} \tag{15}$$

and

$$\widehat{\text{cov}}[\mathbf{Y} \mid \mathbf{W}] = \frac{1}{B} \sum_{b=T+1}^{T+B} (\mathbf{Y}^{(b)} - \bar{\mathbf{Y}})(\mathbf{Y}^{(b)} - \bar{\mathbf{Y}})^{\mathrm{T}}. \tag{16}$$

Therefore, by substituting (12), (13), (15), and (16) back to (5) and (6), we are able to update the current estimates to $\boldsymbol{\beta}^{(k+1)}$ and $\boldsymbol{\Sigma}_r^{(k+1)}$, $r = 1, \ldots, R$, and we continue to iterate until convergence is reached to result in the final estimates $\hat{\boldsymbol{\theta}}$. Note that we are able to work out the estimates of the random effects as a by-product from (12).

5. Monte Carlo Approximation of the Observed Information Matrix

McCulloch (1994) did not report standard errors in his examples but he did mention the supplemented EM (SEM) algorithm of Meng and Rubin (1991) as a possible method. The basic idea of the SEM algorithm is to use the fact that the fraction of missing information is related to the rate of convergence of the EM. By running a sequence of supplementary EM iterations, we can approximate the rate of convergence of the EM algorithm by using finite differences. In this way, we can estimate the increased variability due to missing information, which can then be added to the complete-data variance–covariance matrix. We do not recommend the SEM algorithm for problems requiring Monte Carlo E-steps. At each iteration of the SEM procedure, we need to consider the P sets of parameter values that result from perturbing the P components of $\boldsymbol{\theta}$ one at a time. For each of these P sets of parameter values, we need to run one iteration of the EM algorithm via Gibbs sampling. In other words, we have to carry out Gibbs sampling P times at each step of the SEM algorithm. This is very time consuming if P is large. For the models we propose in Section 7, $P = 13$ and 21. Thus, to run 100 SEM steps, we have to carry out Gibbs sampling 1300 times for model 1 and 2100 times for model 2. Moreover, our experience suggests that a Monte Carlo implementation of the SEM algorithm is numerically unstable, has convergence problems, and sometimes leads to negative variance estimates. This is somewhat surprising as no such problems are reported in the literature for the SEM algorithm. A possible explanation is that the finite difference method of approximating the rate of convergence matrix DM (Meng and Rubin, 1991, p. 902) is adversely affected by the extra variation due to Monte Carlo sampling. In conclusion, the Monte Carlo SEM algorithm is undesirable in terms of both computing time and numerical stability.

Instead of using the SEM algorithm, we will use simulations to approximate the observed information matrix directly. The details are as follows. Let $l(\boldsymbol{\theta}; \mathbf{W}) = \log f(\mathbf{w}; \boldsymbol{\theta})$ denote the log-likelihood function based on the observed data $\mathbf{W}$. Louis (1982) expressed $l''(\boldsymbol{\theta}; \mathbf{W})$ in terms of certain conditional expectations of the derivatives of the complete-data log-likelihood $l(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{U})$ given the observed $\mathbf{W}$. Specifically,

$$\begin{aligned} l''(\boldsymbol{\theta}; \mathbf{W}) = {} & \mathrm{E}[l''(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{U}) \mid \mathbf{W}; \boldsymbol{\theta}] + \mathrm{E}[l'(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{U}) l'^{\mathrm{T}}(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{U}) \mid \mathbf{W}; \boldsymbol{\theta}] \\ & - \mathrm{E}[l'(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{U}) \mid \mathbf{W}; \boldsymbol{\theta}] \mathrm{E}[l'(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{U}) \mid \mathbf{W}; \boldsymbol{\theta}]^{\mathrm{T}}, \end{aligned} \tag{17}$$

where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_R)$. By simulating $(\mathbf{Y}_m, \mathbf{U}_m)$, $m = 1, \ldots, M$, from the conditional distribution of $(\mathbf{Y}, \mathbf{U})$ given $\mathbf{W}$, we can approximate the conditional expectations involved in (17) by the corresponding sample means to obtain

$$\begin{aligned} l''_M = {} & \frac{1}{M} \sum_{m=1}^{M} l''(\hat{\boldsymbol{\theta}}; \mathbf{Y}_m, \mathbf{U}_m) + \frac{1}{M} \sum_{m=1}^{M} l'(\hat{\boldsymbol{\theta}}; \mathbf{Y}_m, \mathbf{U}_m) l'^{\mathrm{T}}(\hat{\boldsymbol{\theta}}; \mathbf{Y}_m, \mathbf{U}_m) \\ & - \left\{ \frac{1}{M} \sum_{m=1}^{M} l'(\hat{\boldsymbol{\theta}}; \mathbf{Y}_m, \mathbf{U}_m) \right\} \left\{ \frac{1}{M} \sum_{m=1}^{M} l'(\hat{\boldsymbol{\theta}}; \mathbf{Y}_m, \mathbf{U}_m) \right\}^{\mathrm{T}} \end{aligned} \tag{18}$$

as a Monte Carlo estimate of $l''(\hat{\boldsymbol{\theta}}; \mathbf{W})$ and the variance–covariance matrix of $\hat{\boldsymbol{\theta}}$ is estimated by $-{l''_M}^{-1}$. In our example, M is increased gradually until the estimated standard errors become stable.

We now describe how to simulate from the conditional distribution of $(\mathbf{Y}, \mathbf{U})$ given $\mathbf{W}$. Since

$$f(\mathbf{y}, \mathbf{u} \mid \mathbf{w}) = f(\mathbf{y} \mid \mathbf{w}) f(\mathbf{u} \mid \mathbf{y}, \mathbf{w}) = f(\mathbf{y} \mid \mathbf{w}) f(\mathbf{u} \mid \mathbf{y}),$$

we can first simulate $\mathbf{Y}$ from $f(\mathbf{y} \mid \mathbf{w})$ using the Gibbs sampling technique described in Section 4 and then simulate $\mathbf{U}$ from $f(\mathbf{u} \mid \mathbf{y})$, which is normal with mean $\boldsymbol{\mu}_{U|Y}$ and covariance matrix $\boldsymbol{\Sigma}_{U|Y}$ obtainable from standard formulae.

6. Accounting for Monte Carlo Variation

Since Gibbs sampling is used to approximate the various conditional expectations required at the E-step of the algorithm, we need to check whether the Gibbs sampler has converged. While a lot of stopping criteria have been proposed in the literature, they are too microscopic in nature and are not designed with Monte Carlo maximum likelihood estimation in mind. Specifically, the existing criteria are primarily concerned with simulations from one target distribution to approximate an expectation. In contrast, we need to approximate a lot of expectations and the distribution that we wish to sample from changes with each iteration as the parameter estimates are updated. Furthermore, our primary interest is not in the expectations themselves but in the parameter estimates they eventually lead to. In view of the above, we decide to adopt a more macroscopic strategy that consists of L independent runs of the Monto Carlo EM algorithm. Letting $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_L$ denote the estimates of $\boldsymbol{\theta}$ that result from the L runs, we can assess the extent of Monte Carlo variation by calculating the sample variance–covariance matrix $\mathbf{S}$ based on $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_L$. As a by-product, we can also compute

$$\bar{\boldsymbol{\theta}} = \frac{1}{L} \sum_{l=1}^{L} \hat{\boldsymbol{\theta}}_l$$

as a more precise estimate of $\boldsymbol{\theta}$. An estimate of the asymptotic covariance matrix of $\bar{\boldsymbol{\theta}}$ that explicitly accounts for Monte Carlo variation (Kuk and Chen, 1992) is

$$V = V_1 + V_2,$$

where

$$\begin{aligned} V_1 &= -\left\{ l''_M(\bar{\boldsymbol{\theta}}) \right\}^{-1}, \\ V_2 &= \mathbf{S}/L, \end{aligned}$$

and $l''_M(\bar{\boldsymbol{\theta}})$ given by (18) is a Monte Carlo approximation of the observed information evaluated at $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$. If V_2 contributes negligibly to the total V, we can conclude that the Monte Carlo variation is insignificant.

7. Example

The salamander mating data reported by McCullagh and Nelder (1989, pp. 439–450) have been extensively analyzed (Schall, 1991; Breslow and Clayton, 1993; Karim and Zeger, 1992). The data were recorded from experiments involving two geographically isolated populations of salamanders, Rough Butt (R) and Whiteside (W). The scientific question addressed in the study is whether the geographically isolated species of salamanders develop barriers to successful mating. Ten R males and 10 W males were sequestered as pairs with 10 R females and 10 W females on six occasions according to the design given in Table 14.3 of McCullagh and Nelder (1989). For each pair, it was recorded whether mating occurred and there are $n = 360$ such records altogether, 120 records from each experiment. The first experiment was conducted in the summer of 1986, while the other two experiments were conducted in fall of the same year. The same animals were used in the first two experiments, while a new set of animals was used for the third experiment. Our main objective in this analysis is to estimate the probability of a successful mating for each of the four types of cross in mating, RR (R female with R male), RW, WR, WW, as well as the seasonal effect. We are also interested in knowing whether there exists heterogeneity among animals and, if so, whether it is greater for females or males and for which species. To answer this question we model the animal effects as random effects. As the same animals were used in the first two experiments, those random effects corresponding to the same animal are correlated.

Previous analyses of the data (Schall, 1991; Breslow and Clayton, 1993; Karim and Zeger, 1992) used a generalized linear model with random effects and a logit link function. McCulloch (1994) used probit link and a Monte Carlo ECME algorithm to analyze the three experiments separately. We obtain similar results using the Monte Carlo EM algorithm described in this paper. For example, for experiment 1, we obtain $\hat{\beta}_{RR} = 0.791$ (SE $= 0.393$), $\hat{\beta}_{RW} = 0.531$ (0.374), $\hat{\beta}_{WR} = -0.954$ (0.412), $\hat{\beta}_{WW} = 0.699$ (0.385), $\hat{\sigma}^2_F = 0.592$ (0.363), and $\hat{\sigma}^2_M = 0.019$ (0.153) compared with the estimates 0.819, 0.538, −0.978, 0.707, 0.600, and 0.067 obtained by McCulloch.

7.1 *A Probit Linear Model with Correlated Random Effects*

To analyze the combined data set we use the model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}_{F12}\mathbf{u}_{F12} + \mathbf{Z}_{M12}\mathbf{u}_{M12} + \mathbf{Z}_{F3}\mathbf{u}_{F3} + \mathbf{Z}_{M3}\mathbf{u}_{M3} + \varepsilon \tag{19}$$

and $\mathbf{W} = I_{(\mathbf{Y}>\mathbf{0})}$. The design matrix $\mathbf{X}$ consists of the indicator variables for the four types of cross, RR, RW, WR, and WW, as well as an indicator variable for the season (fall = 1; summer = 0). The vector $\beta = (\beta_{RR}, \beta_{RW}, \beta_{WR}, \beta_{WW}, \beta_{FALL})^{\mathrm{T}}$ consists of the corresponding fixed effects. The random effects are

$$\mathbf{u}_{F12} \sim \mathcal{N}_{40}(\mathbf{0}, \mathbf{I}_{20} \otimes \Sigma_{F12}), \quad \text{where } \Sigma_{F12} = \begin{pmatrix} \sigma_{F1}^2 & \sigma_{F12} \\ \sigma_{F12} & \sigma_{F2}^2 \end{pmatrix},$$

$$\mathbf{u}_{M12} \sim \mathcal{N}_{40}(\mathbf{0}, \mathbf{I}_{20} \otimes \Sigma_{M12}), \quad \text{where } \Sigma_{M12} = \begin{pmatrix} \sigma_{M1}^2 & \sigma_{M12} \\ \sigma_{M12} & \sigma_{M2}^2 \end{pmatrix},$$

$$\mathbf{u}_{F3} \sim \mathcal{N}_{20}(\mathbf{0}, \sigma_{F3}^2\mathbf{I}_{20}),$$

$$\mathbf{u}_{M3} \sim \mathcal{N}_{20}(\mathbf{0}, \sigma_{M3}^2\mathbf{I}_{20}),$$

where $\mathbf{u}_{F12}(\mathbf{u}_{M12})$ is made up of 20 2×1 random vectors corresponding to the effects of 20 female(male) animals over the two occasions, experiments 1 and 2, and $\mathbf{u}_{F3}(\mathbf{u}_{M3})$ represents the effects of the new set of animals used in experiment 3. The design matrices corresponding to these random effects are $\mathbf{Z}_{F12}, \mathbf{Z}_{M12}, \mathbf{Z}_{F3}$, and $\mathbf{Z}_{M3}$. Finally, we assume that the error vector $\varepsilon \sim \mathcal{N}_{360}(\mathbf{0}, \mathbf{I})$.

In fitting the model we set the starting values to be zero for all beta parameters, 0.01 for all variance parameters, and 0.001 for all covariance parameters. Using (5), (6), (15), and (16), we iterate 300 times to obtain the parameter estimates. The results of five runs of the algorithm are given in Table 1. It is clear that we obtain more or less the same estimates from each run. In fact, V_2 contributes only negligibly to $V = V_1 + V_2$, where $V_1 = -\{l''_M(\bar{\theta})\}^{-1}$ is stabilized at M = 50,000.

The following conclusions are drawn. For the fixed effects, we find that the mating rate of the WR cross type is lowest, whereas the mating rates of RR and WW are highest and of similar

Table 1

Parameter estimates and standard errors (in italic) of model 1 *for the salamander data*

	Fixed effects				
	β_{RR}	β_{RW}	β_{WR}	β_{WW}	β_{FALL}
	0.874	0.443	−0.948	0.837	−0.353
	0.879	0.446	−0.955	0.838	−0.357
	0.881	0.448	−0.952	0.843	−0.361
	0.874	0.442	−0.954	0.836	−0.352
	0.876	0.445	−0.951	0.841	−0.355
Average	0.877	0.445	−0.952	0.839	−0.356
V_2	*2.01E-6*	*1.26E-6*	*1.17E-6*	*1.71E-6*	*2.55E-6*
$V = V_1 + V_2$	*0.111*	*0.100*	*0.109*	*0.105*	*0.088*
SE = $V^{1/2}$	*0.333*	*0.316*	*0.330*	*0.324*	*0.297*

	Variance and covariance of random effects							
	σ_{F1}^2	σ_{F2}^2	σ_{F12}	σ_{M1}^2	σ_{M2}^2	σ_{M12}	σ_{F3}^2	σ_{M3}^2
	0.630	0.703	−0.0419	0.268	0.478	0.356	0.146	0.676
	0.634	0.714	−0.0396	0.271	0.483	0.359	0.144	0.675
	0.634	0.721	−0.0446	0.267	0.476	0.354	0.149	0.662
	0.631	0.699	−0.0453	0.270	0.482	0.359	0.146	0.680
	0.644	0.713	−0.0511	0.270	0.481	0.358	0.148	0.679
Average	0.635	0.710	−0.0445	0.269	0.480	0.357	0.146	0.674
V_2	*6.53E-6*	*1.53E-5*	*3.73E-6*	*4.68E-7*	*1.59E-6*	*8.48E-7*	*6.87E-7*	*1.08E-5*
$V = V_1 + V_2$	*0.172*	*0.207*	*0.074*	*0.025*	*0.082*	*0.044*	*0.039*	*0.153*
SE = $V^{1/2}$	*0.414*	*0.455*	*0.271*	*0.157*	*0.287*	*0.210*	*0.198*	*0.391*

magnitude. The seasonal effect is in the direction of less successful mating in fall. An estimate of the contrast of primary interest $\beta_{RW} - \beta_{WR}$ is 1.397 (SE = 0.364), which is significantly different from zero. Finally, we examine the random effects and find that the female random effects have a higher variability than the male random effects in the first two experiments, whereas in the last experiment the relation is reversed. Furthermore, the female random effects in experiments 1 and 2 are apparently not correlated, while for the males they appear to be positively correlated. These findings are in reasonable agreement with the results from previous analyses using logit link.

In assessing the goodness-of-fit, we calculate the estimated probability of successful mating for RR, RW, WR, and WW mating types, denoted by $\pi_{RR}, \pi_{RW}, \pi_{WR}, \pi_{WW}$ in experiments 1, 2, and 3. For example,

$$\pi_{RR} = \Phi(\beta_{RR}/(\sigma_{F1}^2 + \sigma_{M1}^2 + 1)^{1/2})$$

for experiment 1 (Zeger, Liang, and Albert, 1988). Then we compare the estimated probabilities with the observed proportions in Table 2. We can see that all the estimated probabilities match quite well with the observed proportions. The sample variances of the animal-specific totals for the female R, female W, male R, and male W salamanders, denoted by $S_{FR}^2, S_{FW}^2, S_{MR}^2, S_{MW}^2$, are calculated for each experiment and their expected values are approximated on the basis of 5000 samples drawn from the model at the estimated parameter values. The results, also given in Table 2, reveal the inadequacy of the model as we find that, for each gender, the ordering of the expected variances for species R and W is often opposite to that of the corresponding observed variances. For example, the observed S_{FR}^2 (1.733) is smaller than the observed S_{FW}^2 (3.789) in experiment 2, but the expected S_{FR}^2 (2.898) is larger than that of S_{FW}^2 (2.138) under model 1. Therefore, we revise our model to allow the variance parameters of different species to be different. To our knowledge, this is the first attempt of fitting a model to the salamander data with species-specific random effects.

7.2 *A Probit Linear Model with Species-Specific Random Effects*

The revised model, model 2, has five fixed effects and eight random effects,

$$\begin{aligned} \mathbf{Y} = \mathbf{X}\beta &+ \mathbf{Z}_{FR12}\mathbf{u}_{FR12} + \mathbf{Z}_{FW12}\mathbf{u}_{FW12} + \mathbf{Z}_{MR12}\mathbf{u}_{MR12} + \mathbf{Z}_{MW12}\mathbf{u}_{MW12} \\ &+ \mathbf{Z}_{FR3}\mathbf{u}_{FR3} + \mathbf{Z}_{FW3}\mathbf{u}_{FW3} + \mathbf{Z}_{MR3}\mathbf{u}_{MR3} + \mathbf{Z}_{MW3}\mathbf{u}_{MW3} + \varepsilon. \end{aligned} \tag{20}$$

This model is an extension of model 1 by further subdividing $\mathbf{u}_{F12}, \mathbf{u}_{M12}, \mathbf{u}_{F3}$, and $\mathbf{u}_{M3}$ in (19) into two species-specific parts. Thus,

$$\mathbf{u}_{FR12} \sim \mathcal{N}_{20}(\mathbf{0}, \mathbf{I}_{10} \otimes \Sigma_{FR12}),$$
$$\mathbf{u}_{FW12} \sim \mathcal{N}_{20}(\mathbf{0}, \mathbf{I}_{10} \otimes \Sigma_{FW12}),$$

Table 2

Observed and expected values for various statistics

				Expected proportion					
	Observed proportion			Model 1			Model 2		
	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3
π_{RR}	0.733	0.600	0.667	0.737	0.638	0.650	0.745	0.636	0.632
π_{RW}	0.667	0.467	0.533	0.626	0.524	0.526	0.631	0.511	0.508
π_{WR}	0.233	0.233	0.167	0.245	0.188	0.166	0.250	0.236	0.144
π_{WW}	0.700	0.667	0.633	0.728	0.628	0.640	0.731	0.610	0.639
				Expected variance					
	Observed variance			Model 1			Model 2		
	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3
S_{FR}^2	2.844	1.733	2.711	2.574	2.898	1.714	2.283	1.944	2.406
S_{FW}^2	2.844	3.789	0.711	2.159	2.138	1.285	2.399	3.282	1.001
S_{MR}^2	2.100	2.944	2.278	1.523	1.788	2.148	1.652	2.398	2.021
S_{MW}^2	1.878	1.378	3.833	1.801	2.412	3.137	1.497	1.725	3.319

$$\begin{aligned}\mathbf{u}_{MR12} &\sim \mathcal{N}_{20}(\mathbf{0}, \mathbf{I}_{10} \otimes \Sigma_{MR12}),\\ \mathbf{u}_{MW12} &\sim \mathcal{N}_{20}(\mathbf{0}, \mathbf{I}_{10} \otimes \Sigma_{MW12}),\\ \mathbf{u}_{FR3} &\sim \mathcal{N}_{10}(\mathbf{0}, \sigma^2_{FR3}\mathbf{I}_{10}),\\ \mathbf{u}_{FW3} &\sim \mathcal{N}_{10}(\mathbf{0}, \sigma^2_{FW3}\mathbf{I}_{10}),\\ \mathbf{u}_{MR3} &\sim \mathcal{N}_{10}(\mathbf{0}, \sigma^2_{MR3}\mathbf{I}_{10}),\\ \mathbf{u}_{MW3} &\sim \mathcal{N}_{10}(\mathbf{0}, \sigma^2_{MW3}\mathbf{I}_{10}),\end{aligned}$$

where $\mathbf{u}_{FR12}(\mathbf{u}_{MR12})$ is made up of 10 2×1 random vectors corresponding to the effects of the 10 female(male) animals of species R over the two occasions, experiments 1 and 2, and $\mathbf{u}_{FW12}(\mathbf{u}_{MW12})$ is defined similarily for species W. Again, $\mathbf{u}_{FR3}(\mathbf{u}_{MR3})$ represents the effects of the new set of species R animals used in experiment 3 and $\mathbf{u}_{FW3}(\mathbf{u}_{MW3})$ is similarily defined for species W. The design matrices corresponding to these random effects are $\mathbf{Z}_{FR12}, \mathbf{Z}_{FW12}, \mathbf{Z}_{MR12}, \mathbf{Z}_{MW12}, \mathbf{Z}_{FR3}, \mathbf{Z}_{FW3}, \mathbf{Z}_{MR3}$, and $\mathbf{Z}_{MW3}$.

We use the estimates of model 1 as the starting values and iterate 400 times to obtain a new set of estimates. As our estimate of σ^2_{FW3} (0.0026) is extremely close to zero, we iterate 100 times more subject to the constraint $\sigma^2_{FW3} = 0$. We rerun the algorithm 4 times under the constraint $\sigma^2_{FW3} = 0$. The five runs of the Monte Carlo EM algorithm give similar results and will not be shown individually. The final estimates obtained by averaging over the five runs are reported in Table 3. For this example, $V_1 = -\{l''_M(\bar{\theta})\}^{-1}$ is stabilized at $M = 30{,}000$ and the Monte Carlo variance V_2 is negligible relative to V_1. To assess the goodness-of-fit of model 2, we use similar procedures as for model 1 and the results are also given in Table 2. We can see that there is better agreement between the observed variances $S^2_{FR}, S^2_{FW}, S^2_{MR}, S^2_{MW}$ and their expected values under model 2 than under model 1.

8. Extension

To extend model (2), we allow correlation in the error vector ε so that $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$, where $\mathbf{\Psi} \neq \mathbf{I}$. The extended model has potential applications in the analysis of longitudinal binary data and multivariate clustered binary data.

For longitudinal data, it is common to assume autocorrelated errors within subjects. If the time points are equally spaced, we may consider an AR(1) correlation matrix

$$\Psi_{n_i}(\rho) = \begin{pmatrix} 1 & \rho & \dots & \rho^{n_i-1} \\ \rho & 1 & \dots & \rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \dots & 1 \end{pmatrix}$$

for subject i. Assuming independence across subjects, the correlation matrix for ε is block diagonal

$$\mathbf{\Psi}(\rho) = \operatorname{diag}\left(\Psi_{n_1}(\rho), \ \dots, \Psi_{n_m}(\rho)\right),$$

where m is the number of subjects and n_i is the number of observations for subject i. Under this correlation structure, the M-step of the EM algorithm becomes slightly more complicated than that

Table 3

Parameter estimates and standard errors (in italics) of model 2 for the salamander data

	Fixed effects				
	β_{RR}	β_{RW}	β_{WR}	β_{WW}	β_{FALL}
	0.892	0.418	−1.007	0.869	−0.387
	0.329	*0.288*	*0.384*	*0.326*	*0.308*

	Variance and covariance of random effects							
Species	σ^2_{F1}	σ^2_{F2}	σ_{F12}	σ^2_{M1}	σ^2_{M2}	σ_{M12}	σ^2_{F3}	σ^2_{M3}
R	0.454	0.200	−0.195	0.377	0.916	0.586	0.517	0.728
	0.467	*0.194*	*0.237*	*0.282*	*0.658*	*0.424*	*0.476*	*0.606*
W	0.861	1.845	0.221	0.120	0.139	0.128	0.000	0.841
	0.667	*1.551*	*0.594*	*0.103*	*0.165*	*0.142*	*0.000*	*0.676*

in Section 3. The complete-data MLE of Σ_r is still given by (4), but the complete-data MLE of β and ρ have no closed form. For a fixed value of ρ, the MLE of β is the generalized least squares estimate

$$\hat{\beta}(\rho) = (\mathbf{X}^{\mathrm{T}}\ \boldsymbol{\Psi}(\rho)^{-1}\ \mathbf{X})^{-1}\ \mathbf{X}^{\mathrm{T}}\boldsymbol{\Psi}(\rho)^{-1}\ \left(\mathbf{Y}\ -\ \sum_{r=1}^{R}\mathbf{Z}_r\mathbf{u}_r\right).$$

The complete-data MLE of $\hat{\rho}_c$ of ρ can be obtained by maximizing the profile log-likelihood

$$\begin{aligned} l(\rho) = &\ -\frac{N-m}{2}\log(1-\rho^2) \\ &-\frac{1}{2}\left(\mathbf{Y}\ -\ \sum_{r=1}^{R}\mathbf{Z}_r\mathbf{u}_r\ -\ \mathbf{X}\hat{\beta}(\rho)\right)^{\mathrm{T}}\ \boldsymbol{\Psi}(\rho)^{-1}\left(\mathbf{Y}\ -\ \sum_{r=1}^{R}\mathbf{Z}_r\mathbf{u}_r\ -\ \mathbf{X}\hat{\beta}(\rho)\right)\end{aligned}$$

and $\hat{\beta}(\hat{\rho_c})$ is the complete-data MLE for β. Thus, the EM algorithm works as follows. Given the current estimates $\Sigma_r^{(k)}$, $\beta^{(k)}$, and $\rho^{(k)}$, we obtain the updated estimate $\rho^{(k+1)}$ by taking the conditional expectation of the above profile log-likelihood given the observed data $\mathbf{W} = I_{(\mathbf{Y}>\mathbf{0})}$ and maximizing it with respect to ρ. The updated estimates of Σ_r and β are $\Sigma_r^{(k+1)} = \mathrm{E}(\hat{\Sigma}_{rc} \mid \mathbf{W})$ and $\beta^{(k+1)} = \mathrm{E}(\hat{\beta}(\rho^{(k+1)}) \mid \mathbf{W})$, where the expectations are evaluated at the current parameter estimates.

Next we consider multivariate clustered binary data which arise, for example, in the study of multiple binary traits in animal breeding (Foulley, Gianola, and Im, 1990). Let $h = 1, \ldots, H$ index the H traits. We can model each trait by a threshold model like (2) so that $\mathbf{W}_h = I_{(\mathbf{Y}_h>\mathbf{0})}$ and

$$\mathbf{Y}_h = \mathbf{X}_h\beta_h + \sum_{r=1}^{R}\mathbf{Z}_{hr}\mathbf{u}_{hr} + \varepsilon_h, \qquad h = 1, \ldots, H,$$

where β_h and $\mathbf{u}_{hr}$ denote the trait-specific fixed and random effects. Combining, we have

$$\left(\mathbf{Y}_1 - \sum_{r=1}^{R}\mathbf{Z}_{1r}\mathbf{u}_{1r}, \ldots, \mathbf{Y}_H - \sum_{r=1}^{R}\mathbf{Z}_{Hr}\mathbf{u}_{Hr}\right) = (\mathbf{X}_1\beta_1, \ldots, \mathbf{X}_H\beta_H) + (\varepsilon_1, \ldots, \varepsilon_H). \tag{21}$$

The row vectors of the error matrix $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_H)$ are assumed to be independent and identically distributed as $\mathcal{N}(\mathbf{0}, \mathbf{C})$, where $\mathbf{C}$ denotes the residual covariance matrix. Note that this model is more realistic than a model which assumes that $\mathbf{Y}_1, \ldots, \mathbf{Y}_H$ are conditionally independent given the random effects $\mathbf{u}_{hr}, h = 1, \ldots, H; r = 1, \ldots, R$. Note also that the right-hand side of (21) is a multivariate linear model. If we define the complete data as $(\mathbf{Y}_h, \mathbf{u}_{hr}), h = 1, \ldots, H; r = 1, \ldots, R$, the complete-data MLE of β_h and Σ_{hr} are just trait-specific analogs of (3) and (4). The complete-data MLE of the covariance matrix $\mathbf{C}$ is given by the sample covariance matrix

$$\hat{\mathbf{C}} = \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}/n, \tag{22}$$

where $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_H)$ is the matrix of residuals with trait-specific residual vector $\hat{\varepsilon}_h = \mathbf{Y}_h - \mathbf{X}_h\hat{\beta}_h - \Sigma_{r=1}^{R}\mathbf{Z}_{hr}\mathbf{u}_{hr}$. A subtle point is the following. Since only $\mathbf{W} = I_{(\mathbf{Y}>\mathbf{0})}$ is observed, the variances $c_{11}, \ldots, c_{HH}$ of the error matrix ε are not estimable from the observed data. As the scale parameter $\sqrt{c_{hh}}$ can be absorbed into the regression parameters β_h, we can only estimate the ratios $\beta_1/\sqrt{c_{11}}, \ldots, \beta_H/\sqrt{c_{HH}}$. The common way to overcome this identifiability problem for a probit model is to assume $c_{11} = \cdots = c_{HH} = 1$. This constraint, however, complicates the estimation of $\mathbf{C}$ because the constrained MLE of $\mathbf{C}$ is no longer given by (22). For example, if $H = 2$, it is well known that the MLE of c_{12} subject to $c_{11} = c_{22} = 1$ is the solution to a cubic equation. We are now in an interesting situation where all the parameters are estimable from the complete data $(\mathbf{Y}_h, \mathbf{u}_{hr}), h = 1, \ldots, H; r = 1, \ldots, R$, but only the ratios $\beta_1/\sqrt{c_{11}}, \ldots, \beta_H/\sqrt{c_{HH}}$ are estimable from the observed data $\mathbf{W} = I_{(\mathbf{Y}>\mathbf{0})}$ and we have to impose the constraint $c_{11} = \cdots = c_{HH} = 1$. To implement the EM algorithm, we think it is easier to obtain the unconstrained complete-data MLE of β_h, Σ_{hr}, and $\mathbf{C}$. The updating formula now consists of taking conditional expectations of the unconstrained MLEs given the observed data and evaluating them at the current parameter estimates. We conjecture that we can obtain convergent estimates of the estimable parameters $\beta_1\sqrt{c_{11}}, \ldots, \beta_H\sqrt{c_{HH}}$ by using the above unconstrained maximization version of the EM algorithm. To our knowledge, problems where the complete-data model has more parameters than the observed data model have not been addressed before and deserve further investigation.

ACKNOWLEDGEMENTS

The work of Jennifer Chan was supported by an Australian National Drug Strategy Research Scholarship. Anthony Kuk's research was supported by the Australian Research Council. The authors wish to thank the referees for pointing out the connection between McCulloch's procedure and the ECME algorithm.

RÉSUMÉ

Le modèle normal-probit pour données binaires (McCulloch, 1994) est généralisé à des effets aléatoires corrélés. Pour obtenir les estimations du maximum de vraisemblance nous utilisons l'algorithme EM avec une étape *M* grandement simplifiée par la supposition d'une fonction de lien probit et avec une étape *E* réalisée par échantillonnage de Gibbs. Les erreurs-standards sont calculées en inversant l'approximation de Monte-Carlo de la matrice d'information plutôt que via l'algorithme SEM. Une méthode est égalemment proposée qui tient compte explicitement de la variation de Monte-Carlo. Nous proposons comme illustration une nouvelle analyse des fameuses données d'accouplement des salamandres. Contrairement aux analyses précédentes, nous constatons qu'il est nécessaire d'introduire différentes composantes de variance pour les différentes espèces d'animaux. Finalement nous considérons des modèles avec des erreurs corrélées ainsi que des effets aléatoires corrélés.

REFERENCES

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88,** 9–25.

Drum, M. L. and McCullagh, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrics* **49,** 677–689.

Foulley, J. L., Gianola, D., and Im, S. (1990). Genetic evaluation for discrete polygenic traits in animal breeding. In *Advances in Statistical Methods for Genetic Improvement of Livestock,* D. Gianola and K. Hammond (eds), 361–411. New York: Springer-Verlag.

Karim, M. R. and Zeger, S. L. (1992). Generalized linear models with random effects; salamander data revisited. *Biometrics* **48,** 631–644.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalised linear models with random effects. *Journal of the Royal Statistical Society, Series B* **57,** 395–407.

Kuk, A. Y. C. and Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79,** 531–541.

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

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44,** 226–233.

Marsaglia, G. (1964). Generating a variable from the tail of the normal distribution. *Technometrics* **6,** 101–102.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Model,* 2nd edition. London: Chapman and Hall.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association* **89,** 330–335.

McGilchrist, C. A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society, Series B* **56,** 61–69.

Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance–covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86,** 899–909.

Schall, R. (1991). Estimation in generalised linear models with random effects. *Biometrika* **78,** 719–727.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* **55,** 3–23.

Waclawiw, M. A. and Liang, K. Y. (1993). Prediction of random effects in the generalized linear model. *Journal of the American Statistical Association* **88,** 171–178.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85,** 699–704.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association* **86,** 79–86.

Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44,** 1049–1060.

Received February 1995; *revised April and July* 1996; *accepted July* 1996.