Maximum Likelihood Variance Components Estimation for Binary Data

BU-1037-MD April, 1993

by

Charles E. McCulloch

Biometrics Unit and Statistics Center, Cornell University, Ithaca, NY 14853

1. INTRODUCTION

Models for continuous data which incorporate both fixed and random effects (mixed models) are commonly used in a variety of disciplines, from ecology and medicine to the physical sciences. However, the same is not true for binary data (Stram, Wei and Ware, 1988). Usage has been limited to a large extent by the intractability of the computations involved in fitting many of the models. In this paper, we consider a class of probit-normal models. We describe maximum likelihood (ML) and restricted maximum likelihood (REML) estimation of the parameters in the model by use of the EM algorithm (Dempster, Laird and Rubin, 1977). Our version of the EM algorithm is very similar to that for the continuous, normal linear model and offers a framework for computation of the ML and REML estimates. We demonstrate through two examples that the computations are feasible for any number and structure of random effects and an arbitrary number of fixed effects. This has not previously been possible; ML estimation has only been described in models with nested random effects.

Our focus will be on variance components estimation in mixed models and the analogs of best linear unbiased prediction (BLUP) of the observed values of the random effects. Thus our concentration differs somewhat from the usual one of repeated measures models, which is to treat the fixed effects as the primary quantities of interest, with the random effects introducing a "nuisance" correlation. We do not consider covariance components models.

A number of models for correlated binary data have been proposed. The beta-binomial distribution is a natural model to use (Williams 1975, Crowder 1978) that hypothesizes a mixing distribution directly on the probability of success. However, it does not generalize easily to multiple random effects. Zeger and Liang (1986) and Liang and Zeger (1986) have proposed generalizations of quasi-likelihood methods but their methods focus on the fixed effects and only estimate the variances and covariances as nuisance parameters. Prentice (1988) has considered extensions of the Zeger and Liang (1986) estimating equation approach, explicitly estimating the covariances also. However, like the beta-binomial models, his models are also difficult to generalize to multiple random effects. For these reasons we consider in this paper correlated probit models which are generalizations of those of Ochi and Prentice (1984). These are similar to the logit normal models of Pierce and Sands (1975), Wong and Mason (1984) and Stiratelli, Laird and Ware (1984), though Stiratelli, Laird and Ware's models are intended only for the longitudinal data setting. Our model is essentially a simplified version of the threshold model considered in Harville and Mee (1984), but for their general model the computations were deemed "insurmountable" (p.397) and they were forced to resort to ad hoc estimation methods. Zeger, Liang, and Albert (1988), Liang, Zeger, and Qaqish (1992), and Anderson, Gilmour and Rae (1985) consider a generalized estimating equation approach and Anderson and Aitken (1985) considered an iterative, weighted logit analysis approach with models similar to ours. Other papers which consider related models are Preisler (1989), Im and Gianola (1988), Gianola (1980), Quaas and Van Vleck (1980), and Manski and McFadden (1981).

2. THE MODEL

Our model is a threshold model where $\mathbf{Y}$ represents an unobserved, continuous variable and we observe only $W_i = I_{\{Y_i > 0\}}$, i.e., whether W_i exceeds a threshold of zero. A flexible class of binary data models can be generated by assuming

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}\ , \tag{2.1}$$

$$W_i = I_{\{Y_i > 0\}} \qquad i = 1,2,\cdots,n\ ,$$

where $\mathbf{X}$ and $\mathbf{Z}$ are known matrices, $\mathbf{u} \sim \mathcal{N}(0,\mathbf{D})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\mathbf{I})$, independently of $\mathbf{u}$. It is unimportant whether we actually believe in the threshold model and the unobserved variable $\mathbf{Y}$ or if we merely use it as a device to obtain estimates for the model. We will be primarily interested in estimating the elements of $\mathbf{D}$, the variances of the random effects.

By taking $\mathbf{u} \equiv \mathbf{0}$ the model simplifies to the usual probit analysis model. If we set $\mathbf{X} = \text{diag}\left\{1_{m_i}\right\}$, $i = 1,2,\cdots,G$, $\mathbf{Z} = \text{diag}\left\{1_{n_{ij}}\right\}$, $j = 1,2,\cdots,m_i$, $\boldsymbol{\beta} = \mu$, this reduces to the Ochi and Prentice (1984) model, with the restriction that negative correlations cannot be modeled. Model (2.1) has the advantage over the Ochi and Prentice model that it does not require the mean to be constant within levels of the random effect.

This model is closely related to those of Pierce and Sands (1975) and Stiratelli, Laird and Ware (1984). If ϵ is assumed to have a logistic distribution instead of a normal distribution, then generalizations of their models are obtained.

Advantages of the probit-normal model (2.1) over the logit-normal models of Pierce and Sands and Stiratelli, Laird and Ware are threefold:

1. With a single random effect and only one observation per level of the random effect it reduces to the usual probit model, except with a different error term for ϵ. The logit-normal models do not reduce to the usual logit models. It is conceptually distasteful for a generalization of a simple model (the logit) not to reduce to the simple model when analyzing a dataset appropriate for that model.
2. The marginal mean of W_i has a simple representation (Zeger, Liang and Albert, 1988):

$$E[W_i] = \Phi\left(\mathbf{x}_i'\beta(\mathbf{z}_i'D\mathbf{z}_i+1)^{-\frac{1}{2}}\right). \tag{2.2}$$

3. The EM algorithm (Section 3) takes a form nearly identical to the continuous, normal linear model.

Point 3. is perhaps the most important because we exploit it using a Gibbs sampling approach (see Section 4) to find ML and REML estimates for arbitrarily complicated models of the form (2.3) below.

In what follows, we will assume the standard ANOVA model for variance components estimation, i.e.,

$$\begin{gathered} Y = X\beta + \sum_{i=1}^{r} Z_i u_i + \epsilon, \\ u_i \sim \text{independently } \mathcal{N}_{q_i}(\mathbf{0}, \theta_i \mathbf{I}). \end{gathered} \tag{2.3}$$

These, along with $W_i = I_{\{Y_i > 0\}}$, define our basic model.

3. MAXIMUM LIKELIHOOD ESTIMATION

In this section, we describe estimation of the fixed effects parameters and variance components via the EM algorithm and prediction of the realized values of the random effects. The EM algorithm is used for four reasons: it offers a framework for estimation which is similar to the normal theory case, it automatically constrains iterates to be in the parameter space, it offers a natural extension for REML estimation, and we have found in practice that for simple problems it tends to converge from a wider range of starting values than a quasi-Newton algorithm (see Section 4). To use the EM algorithm we regard the complete data as **Y** and $\mathbf{u}_i$ (i=1,2,... r) as is typically done for the continuous, normal linear model (Laird, 1982). The advantage of the threshold model approach is that we can now appeal to standard results for normally distributed data.

The maximization step is quite simple as shown by Laird (1982). The maximum likelihood estimates for the θ_i are $\hat{\theta}_i = \mathbf{u}_i'\mathbf{u}_i/q_i$ and, given estimates of the θ_i, the maximum likelihood estimate of β is $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}$, where $\hat{\mathbf{V}}^{-1}$ is Var(**Y**) with θ_i replaced by $\hat{\theta}_i$ and

$$\mathrm{Var}(\mathbf{Y}) = \mathbf{I} + \sum_{i=1}^{c} \theta_i \mathbf{Z}_i \mathbf{Z}_i' \ .$$

The expectation step is also conceptually simple. We need to calculate E[**Y**|**W**] and $E[\ \mathbf{u}_i'\mathbf{u}_i|\mathbf{W}]$. This latter expectation can be calculated in two steps, as done by Pettit (1986) for censored data:

$$E[\mathbf{u}_i'\mathbf{u}_i|\mathbf{W}] = E\Big[E[\ \mathbf{u}_i'\mathbf{u}_i|\mathbf{Y}] \mid \mathbf{W}\Big] \ . \tag{3.1}$$

To calculate the inner expectation, we can use the usual multivariate normal results:

$$E[\mathbf{u}_i'\mathbf{u}_i|\mathbf{Y}] = \theta_i^2(\mathbf{Y}-\mathbf{X}\beta)'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\beta) + \mathrm{tr}(\theta_i\mathbf{I}-\theta_i^2\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_i) \ . \tag{3.2}$$

Using (3.1) we therefore have,

$$\begin{aligned}
E[\mathbf{u}_i'\mathbf{u}_i \mid \mathbf{W}] &= \theta_i^2 E[(\mathbf{Y}-\mathbf{X}\beta)'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\beta) \mid \mathbf{W}] + \mathrm{tr}(\theta_i\mathbf{I}-\theta_i^2\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_i) \\
&= \theta_i^2 \mathrm{tr}E[\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\beta)(\mathbf{Y}-\mathbf{X}\beta)' \mid \mathbf{W}] + \mathrm{tr}(\theta_i\mathbf{I}-\theta_i^2\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_i) \\
&= \theta_i^2 \mathrm{tr}\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\Big(\mathbf{V}_{\mathbf{Y}|\mathbf{W}} + (\mu_{\mathbf{Y}|\mathbf{W}}-\mathbf{X}\beta)(\mu_{\mathbf{Y}|\mathbf{W}}-\mathbf{X}\beta)'\Big) + \mathrm{tr}(\theta_i\mathbf{I}-\theta_i^2\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_i) \\
&= \theta_i^2 \mathrm{tr}\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{V}_{\mathbf{Y}|\mathbf{W}} + \theta_i^2(\mu_{\mathbf{Y}|\mathbf{W}}-\mathbf{X}\beta)'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}(\mu_{\mathbf{Y}|\mathbf{W}}-\mathbf{X}\beta) \\
&\qquad + \mathrm{tr}(\theta_i\mathbf{I}-\theta_i^2\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_i) \ ,
\end{aligned}$$

where $\mathbf{V}_{\mathbf{Y}|\mathbf{W}} = \mathrm{Var}(\mathbf{Y}|\mathbf{W})$ and $\mu_{\mathbf{Y}|\mathbf{W}} = E[\mathbf{Y}|\mathbf{W}]$.

This shows that the only extra computations needed for maximum likelihood estimation for discrete data are the computation of $\mathbf{V}_{\mathbf{Y}|\mathbf{W}}$ and $\mu_{\mathbf{Y}|\mathbf{W}}$. By demonstrating that only the conditional mean and variance of **Y** are needed, the EM algorithm offers a framework for relatively unrestricted computation of complicated mixed models for binary data. In Section 4, using both numerical integration and a Gibbs sampling approaches, we show that the computations are feasible in practice.

We are now prepared to make a formal statement of the EM algorithm for maximum likelihood estimation. In the statement of the algorithm, superscripts in parentheses on **V**, $\mathbf{V}_{\mathbf{Y}|\mathbf{W}}$, and $\mu_{\mathbf{Y}|\mathbf{W}}$ indicate that the current values of the parameters have been substituted.

<u>EM Algorithm for ML Estimation</u>

0. Obtain starting values $\beta^{(0)}$ and $\theta^{(0)}$. Set m = 0.

1. (E-Step) Calculate

$$\hat{t}_i^{(m)} = E[u_i'u_i|\mathbf{W}, \beta=\beta^{(m)}, \theta=\theta^{(m)}]$$

$$= \theta_i^{(m)^2}\mathrm{tr}\mathbf{V}^{(m)^{-1}}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{(m)^{-1}}\mathbf{V}^{(m)}_{\mathbf{Y}|\mathbf{W}}$$

$$+ \theta_i^{(m)^2}(\mu^{(m)}_{\mathbf{Y}|\mathbf{W}}-\mathbf{X}\beta^{(m)})'\mathbf{V}^{(m)^{-1}}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{(m)^{-1}}(\mu^{(m)}_{\mathbf{Y}|\mathbf{W}}-\mathbf{X}\beta^{(m)})$$

$$+ \mathrm{tr}(\theta_i^{(m)}\mathbf{I}-\theta_i^{(m)^2}\mathbf{Z}_i'\mathbf{V}^{(m)^{-1}}\mathbf{Z}_i)\ .$$

2. (M-step) Set

$$\theta_i^{(m+1)} = \hat{t}_i^{(m)}/q_i$$

$$\beta^{(m+1)} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mu^{(m)}_{\mathbf{Y}|\mathbf{W}}\ .$$

3. If convergence is reached, set $\hat{\theta} = \theta^{(m+1)}$ and $\hat{\beta} = \beta^{(m+1)}$, otherwise increase m by one and return to step 1.

This implementation of the EM algorithm is identical to the continuous case, except that (Y - Xβ)(Y - Xβ)' and **Y** are replaced by their expected values given **W**. Most of the computational effort is expended in the calculation of $\mathbf{V}_{\mathbf{Y}|\mathbf{W}}$ and $\mu_{\mathbf{Y}|\mathbf{W}}$. This will be discussed in more detail in Section 4.

We are now in a position to give a version of restricted maximum likelihood estimation

(REML). The basic idea behind REML is to maximize a portion of the likelihood which depends only on the variance components and not on the fixed effects. See Searle, Casella and McCulloch (1992, Sections 6.6 and 9.2b) for further details. We use the approach of Laird (1982) and obtain REML estimators by treating the fixed effects as random effects whose variance tends to infinity. This approach is motivated by adopting a Bayesian viewpoint and letting the prior information about the fixed effects tend to zero (variance tends to infinity); see Harville (1974). Using the same device, equation (3.1), as for ML estimation, we calculate

$$\begin{aligned}
\mathrm{E}[\mathbf{u}_i'\mathbf{u}_i|\mathbf{W}] &= \mathrm{E}\Big[\mathrm{E}[\mathbf{u}_i'\mathbf{u}_i|\mathbf{Y}] \mid \mathbf{W}\Big]\,. \\
&= \theta_i^2 \mathrm{E}[\mathbf{Y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Y} \mid \mathbf{W}] + \mathrm{tr}(\theta_i\mathbf{I} - \theta_i^2\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_i) \\
&= \theta_i^2 \mathrm{tr}\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}(\mathbf{V}_{\mathbf{Y}|\mathbf{W}} + \mu_{\mathbf{Y}|\mathbf{W}}\mu_{\mathbf{Y}|\mathbf{W}}') + \mathrm{tr}(\theta_i\mathbf{I} - \theta_i^2\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_i) \\
&= \theta_i^2 \mathrm{tr}\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{V}_{\mathbf{Y}|\mathbf{W}} + \theta_i^2\mu_{\mathbf{Y}|\mathbf{W}}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mu_{\mathbf{Y}|\mathbf{W}} + \mathrm{tr}(\theta_i\mathbf{I} - \theta_i^2\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_i)\,,
\end{aligned}$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$. An analog of REML can be defined for discrete data, using an EM algorithm as follows:

EM Algorithm for REML Estimation

0. Obtain starting values $\theta^{(0)}$. Set m = 0.

1. (E-Step) Calculate

$$\begin{aligned}
\hat{t}_i^{(m)} &= \mathrm{E}[\mathbf{u}_i'\mathbf{u}_i|\mathbf{W},\ \theta = \theta^{(m)}] \\
&= \theta_i^{(m)^2}\mathrm{tr}\mathbf{P}^{(m)}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}^{(m)}\mathbf{V}_{\mathbf{Y}|\mathbf{W}}^{(m)} + \theta_i^{(m)^2}\mu_{\mathbf{Y}|\mathbf{W}}^{(m)}{}'\mathbf{P}^{(m)}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}^{(m)}\mu_{\mathbf{Y}|\mathbf{W}}^{(m)} \\
&\qquad + \mathrm{tr}(\theta_i^{(m)}\mathbf{I} - \theta_i^{(m)^2}\mathbf{Z}_i'\mathbf{P}^{(m)}\mathbf{Z}_i)\,.
\end{aligned}$$

2. (M-step) Set

$$\theta_i^{(m+1)} = \hat{t}_i^{(m)}/q_i\,.$$

3. If convergence is reached, set $\hat{\theta} = \theta^{(m+1)}$ otherwise increase m by one and return to step 1.

A major difference between ML and REML estimation is that for REML the limiting values of $\mathbf{V}_{\mathbf{Y}|\mathbf{W}}$

and $\mu_{\mathbf{Y}|\mathbf{W}}$ as the variance of the fixed effects tends to infinity must be used.

The prediction of the observed values of the random effects, $\mathbf{u}_i$, is often of interest in applied work (Mabry *et al.*, 1987). For continuous data, the Best Linear Unbiased Prediction (BLUP) methodology is often used, giving rise to $\hat{\mathbf{u}}_i = \hat{\theta}_i^2\mathbf{Z}_i'\hat{\mathbf{V}}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\beta}) = \hat{\theta}_i^2\mathbf{Z}_i'\hat{\mathbf{P}}\mathbf{Y}$, which is an estimate of $\mathrm{E}[\mathbf{u}_i|\mathbf{Y}]$. The corresponding calculation for discrete data is $\hat{\mathbf{u}}_i = \hat{\theta}_i^2\mathbf{Z}_i'\hat{\mathbf{V}}^{-1}(\hat{\mu}_{\mathbf{Y}|\mathbf{W}}-\mathbf{X}\hat{\beta}) = \hat{\theta}_i^2\mathbf{Z}_i'\hat{\mathbf{P}}\hat{\mu}_{\mathbf{Y}|\mathbf{W}}$, which is an estimate of $\mathrm{E}[\mathbf{u}_i|\mathbf{W}]$. The form of the estimator is the same whether we use ML or REML estimation though the estimates will, in general, be different due to different values for the variance components and $\hat{\mu}_{\mathbf{Y}|\mathbf{W}}$.

As pointed out by Wu (1983), EM is not guaranteed to converge to a global maximum. Our experience has shown that multimodal likelihoods are possible for models such as these; so the best we can hope for in this setting is that EM will converge to a local maximum. Unfortunately the regularity conditions of Wu (1983) do not apply; a realistic compactification of the parameter space by including infinite variance components leads to identifiability problems. Truncation of the parameter space to exclude extremely large values would allow the regularity conditions to be met. Then, since $\mathrm{Q}\left((\theta^*, \beta^*)\,|\,(\theta, \beta)\right) = \mathrm{E}\left[\ell\mathrm{og}\,\mathrm{f}\,(\mathbf{Y}\,|\,(\theta^*, \beta^*))\,|\,\mathbf{W}, (\theta, \beta)\right]$ is continuous in both (θ^*, β^*) and (θ, β), Theorem 2 of Wu (1983) applies and EM is guaranteed to converge to a stationary point. For any particular dataset a local maximum would need to be verified by numerically calculating the second derivative matrix (via numerical integration or techniques like Meng and Rubin, 1991).

4. EXAMPLES

We applied the methods derived in Section 3 to the data analyzed by Ochi and Prentice (1984), from Weil (1970), and to the salamander data from McCullagh and Nelder (1989, Section 14.5).

4.1 The Weil data

The Weil dataset has a treatment and control group and a single, nested random effect. The response is survival of rats and the random effect is litter. The model would be

$$\mathrm{Y}_{ij} = \mu_i + \mathrm{u}_{ij} + \epsilon_{ijk}$$

$$\mathrm{W}_{ij} = \mathrm{I}_{\{\mathrm{Y}_{ij}>0\}}\ ,$$

where i indexes treatment/control, j indexes litter and k indexes rat within litter, so μ_i is the treatment mean on the latent scale and the u_{ij} are the random, litter effects. To find ML estimates for the Weil data set it is most efficient to numerically evaluate integrals of the form

$$\int_{-\infty}^{\infty} \alpha\, \Psi(\alpha;n,s,\mu,\sigma)\phi(\alpha)d\alpha ,$$

where

$$\Psi(\alpha;n,s,\mu,\sigma) = \frac{\Phi(\mu+\sigma\alpha)^s[1-\Phi(\mu+\sigma\alpha)]^{n-s}}{\int_{-\infty}^{\infty}\Phi(\mu+\sigma x)^s[1-\Phi(\mu+\sigma x)]^{n-s}\phi(x)dx},$$

and $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal c.d.f. and p.d.f. As pointed out by Pettit (1986), integrals such as these have been well studied and are relatively easy to evaluate numerically. We used Hermite quadrature with 20 evaluation points (Abramowitz and Stegun, 1964, Table 25.10, n=10). We noticed none of the accuracy problems reported in Ochi and Prentice (1984) and, in fact, were able to reproduce the true values in their Table 1 exactly.

We used the matrix language GAUSS (Aptech Systems, 1990) on an IBM PC and fit each group separately (previous analyses have shown unequal values of the variance in the two groups). For the treated group (i=2), the algorithm converged in 28 iterations and under 1 minute. For the control group (i=1), the EM algorithm required 207 iterations about 1 1/2 minutes. The maximum likelihood estimates and their standard errors (calculated from the observed information matrix) were $\hat{\mu}_1 = 1.306$ (standard error .169), $\widehat{\text{stddev}}(u_{1j}) \equiv \hat{\sigma}_1 = .240$ (standard error .301), $\hat{\mu}_2 = 0.946$ (standard error .319) and $\widehat{\text{stddev}}(u_{2j}) \equiv \hat{\sigma}_2 = 1.023$ (standard error .291). These estimates agree substantially with those of Ochi and Prentice; slight differences are to be expected because they used the approximation due to Mendell and Elston (1974). For example in group 2, Ochi and Prentice obtain $\hat{\gamma}_2 \equiv \frac{\hat{\mu}_2}{\sqrt{1+\hat{\sigma}_2^2}} = .651$, whereas our estimates give $\hat{\gamma}_2 = .661$. The large number of iterations required by the EM algorithm for the control group is typical of problems for which the estimates lie near the boundary of the parameter space. When the likelihood can be evaluated numerically, as in this example, it is straightforward to conduct likelihood ratio tests and to evaluate derivatives of the likelihood function for calculating standard errors.

We also fitted this data set using a quasi-Newton algorithm (Aptech Systems, 1990, Applications Manual, p. 207). Convergence was achieved to essentially the same parameter values and each group was fitted in less than a minute. A small amount of experimentation with the starting values showed that the EM algorithm converged from a wider range of starting values than the quasi-Newton algorithm.

4.2 The Gibbs Sampler and the salamander data

In a design with a more complicated random effects structure, for example crossed effects, the computations become too burdensome for direct numerical calculation (e.g., the algorithm of Leppard and Tallis (1989) only works for small dimensions). To illustrate the flexibility of the framework of Section 3 we employ a Gibbs sampling approach (Gelfand and Smith, 1990) to calculate E[**Y**|**W**] and Var(**Y**|**W**). Tanner (1991) suggests a similar Monte Carlo EM algorithm. By using the Gibbs sampler, arbitrarily complicated designs can be easily accommodated. We apply this approach to the salamander data of McCullagh and Nelder (1989, Section 14.5) which has two crossed random effects and four fixed effects.

We now outline the use of the Gibbs sampler. It rests on a result of Robert (1992) for sampling from a truncated multivariate normal and is similar to the treatment of Gelfand, Smith and Lee (1992). The basic idea is that fast acceptance-rejection methods exist (e.g. Marsaglia, 1964) for sampling from a truncated univariate normal. By cycling through the conditional distributions of $Y_i|Y_j$, $j \neq i$ we only ever need to simulate truncated univariate normals. Here is an outline of how the Gibbs sampler is used to generate a sample of **Y**'s from the conditional distribution of **Y** | **W**.

1. For each i calculate

$$\sigma^2_{i|(i)} = \text{Var}\,(Y_i|Y_j\, j \neq i) \quad \text{and} \quad \beta_{i|(i)} = \text{Cov}(Y_i, Y_{(i)})\ ,$$

where $Y_{(i)} = (Y_1, Y_2 \ldots, Y_{i-1}, Y_{i+1}, \cdots Y_n)'$.

2. For each i calculate

$$\mu_{i|(i)} = E[Y_i|Y_j, j \neq i] = \mathbf{x}_i\beta + \beta'_{i|(i)}(\mathbf{Y}_{(i)} - \mathbf{X}_{(i)}\beta)\ ,$$

where $\mathbf{X}_{(i)} = \mathbf{X}$ with row i deleted, and $\mathbf{x}_i$ is the i^{th} row of **X**.

3. Simulate Y_i from a truncated normal distribution with mean $\mu_{i\,|\,(i)}$ and standard deviation $\sigma_{i|(i)}$. If $W_i = 1$, simulate Y_i truncated above 0. If $W_i = 0$ simulate Y_i to be truncated below 0.

Repeat steps 2 and 3 a large number of times, k, to obtain $\mathbf{Y}^{(1)}$, $\mathbf{Y}^{(2)}$ ..., $\mathbf{Y}^{(k)}$. Discard a suitable number of the $Y^{(j)}$ from the beginning of the sequence (the burn in period), after that use every m^{th} one to estimate E[**Y**|**W**] and Var (**Y**|**W**). Because of the iterative nature of the EM algorithm and the desire to take as few Gibbs samples as possible (especially at the beginning of EM) we settled on a burn in period of i, skipped integer $(i/10) + 1$ samples and used $i + 1$ replications, where i is the iteration in the EM algorithm. These numbers are small in relation to those recommended in the literature, but we noticed no problems. We tried larger values with no improvement. For simpler cases where quasi-Newton estimation was possible we compared EM-Gibbs and quasi-Newton for a number of simulated data sets and had success with the smaller number of Gibbs samples in each case.

In the Gibbs sampler, most of the computational effort is expended in repeating steps 2 and 3 a sufficiently large number of times. Thus, complicated random effects structures have little impact on the computational time since they only affect step 1.

The salamander data consists of three experiments, each with n = 120 matings. $W_i = 1$ if the i^{th} mating is successful and zero otherwise. There were 20 males and 20 females, ten of each of two species. There were four types of crosses in the matings: species R female - species R male, species R female - species W male, species W female - species R male, species W female - species W male. For each experiment the model is

$$\mathbf{Y} = \mathbf{X}\beta + \mathrm{Z_f}\mathbf{U}_f + \mathrm{Z_m}\mathbf{U}_m + \epsilon,$$

$$W_i = I_{\{Y_i > 0\}},$$

$$\mathbf{U}_f \sim \mathcal{N}_{20}(0, \theta_f\, \mathbf{I}), \text{ the female effects,}$$

$$\mathbf{U}_m \sim \mathcal{N}_{20}(0, \theta_m\, \mathbf{I}), \text{ the male effects,}$$

$$\epsilon \sim \mathcal{N}_{120}(0, \mathbf{I}),$$

$\mathbf{X}$ = indicator matrix for the type of cross,

Z_f = indicator matrix for the females,

$\mathbf{Z}_m$ = indicator matrix for the males,

$\beta = (\beta_{R/R}, \beta_{R/W}, \beta_{W/R}, \beta_{W/W})'$ effect for type of cross.

$\mathbf{U}_f$ and $\mathbf{U}_m$ represent the consistent effect individual males and females have across matings on the latent variable Y, which governs mating success. They are the random effects which are assumed to be i.i.d. with variances given by, respectively, θ_f and θ_m. Figure 1 shows the convergence of the parameter estimates for experiment 1. The final estimates were $\hat{\beta}_{R/R} = 0.819$, $\hat{\beta}_{R/W} = 0.538$, $\hat{\beta}_{W/R} = -0.978$, $\hat{\beta}_{W/W} = 0.707$, $\hat{\theta}_f = 0.600$ and $\hat{\theta}_m = 0.067$. These are relatively similar to the estimates Karim and Zeger (1992) obtained in a Bayesian analysis using the Gibbs sampler and a logit-normal model. Table 1 shows the Bayesian and ML estimates of the variance components for the three experiments. When the likelihood is not directly evaluated, as in this example, it is much more complicated to calculate standard errors. Techniques based directly on EM, e.g., Meng and Rubin (1991) are necessary.

The estimates of the marginal probabilities (see 2.2) are almost exactly equal to the observed proportions:

Cross	Estimated marginal proportion $\Phi\left(\hat{\beta}/(\hat{\theta}_f + \hat{\theta}_m + 1)^{1/2}\right)$	Observed proportion
R/R	$\Phi\left(.819/(.6 + .067 + 1)^{1/2}\right) = 0.737$	22/30 = 0.733
R/W	0.661	20/30 = 0.667
W/R	0.224	7/30 = 0.233
W/W	0.708	21/30 = 0.7

While this approach is computationally intensive, it is not prohibitive. On a fast (33 MH, 486) IBM PC compatible using the language GAUSS (Aptech Systems, 1990), 50 iterations were completed in 90 minutes and 80 iterations were completed in 250 minutes. (Later iterations do more Gibbs sampling — see Appendix). And these times could undoubtedly be improved by more efficient programming and computational techniques.

This application of the Gibbs sampler is unusual in that it is used to solve directly for maximum likelihood estimates rather than utilizing a Bayesian framework. It would seem to be of broad utility for models which contain a latent, multivariate normal component.

5. CONCLUSIONS

We have developed a framework for ML and REML estimation of variance components from binary data using the EM algorithm. This is very similar to the EM algorithm for the continuous, normal linear model. For simple settings the ML computations can be performed by numerical integration. For more complicated problems this framework can be used with Gibbs sampling approach to calculate ML and REML estimates. This has not been previously possible in designs with complicated (e.g. crossed) random effects.

REFERENCES

Abramowitz, M. and Stegun, I. (1964). Handbook of Mathematical Functions. National Bureau of Standards, Washington, D.C.

Anderson, D. A. and Aitken, M. (1985). Variance components models with binary response: Interviewer variability. *Journal of the Royal Statistical Society, Series B,* **47**: 203-210.

Aptech Systems (1990). *GAUSS 2.1 Users' Manuals.* Kent, Washington.

Crowder, M. J. (1978). Beta-binomial ANOVA for proportions. *Applied Statistics* **27**: 34-37.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete observations. *Journal of the Royal Statistical Society, Series B*, **39**: 1-38.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 105-114.

Gelfand, A.E., Smith, A.F.M. and Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association* **87**: 523-532.

Gianola, D. (1980). Genetic evaluation of animals for traits with categorical responses. *Journal of Animal Science* **51**: 1272-1276.

Gilmour, A. R., Anderson, R. D., and Rae, A. L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72**: 593-599.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**: 383-385.

Harville, D. A. and Mee, R. W. (1984). A mixed model procedure for analyzing ordered categorical data. *Biometrics* **40**: 393-408.

Im, S. and Gianola, D. (1988). Mixed models for binomial data with an application to lamb mortality. *Applied Statistics* **37**: 196-204.

Karim, M.R. and Zeger, S.L. (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics* **48**: 681-694.

Laird, N.L. (1982). Computation of variance components using the EM algorithm. *Journal of Statistical Computation and Simulation* **14**: 295-303.

Leppard, P. and Tallis, G.M. (1989). Evaluation of the mean and covariance of the truncated multinormal distribution. *Applied Statistics* **38**: 543-553.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13-22.

Liang, K.-Y., Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B,* **54**: 3-40.

Mabry, J.W., Benyskek, L.L., Johnson, M.H., and Little, D.E. (1987). A comparison of methods for ranking boars from different central test stations. *Journal of Animal Science* **65**: 56-62.

Manski, C. F. and McFadden, D. (1981). Structural analysis of discrete data with econometric applications. MIT Press, Cambridge, Massachusetts.

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

Mendell, N. R. and Elston, R. C. (1974). Multifactorial qualitative traits: Genetic analysis and prediction of recurrence risks. *Biometrics* **30**: 41-57.

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.

Ochi, Y. and Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* 71: 531-543.

Pettit, A. N. (1986). Censored observations, repeated measures and mixed effects models: An approach using the EM algorithm and normal errors. *Biometrika* **73**: 634-643.

Pierce, D. A. and Sands, B. R. (1975) Extra-Bernoulli variation in binary data. Technical Report 46. Department of Statistics, Oregon State University.

Preisler, H.K. (1989). Analysis of a Toxicological Experiment using a Generalized Linear Model with Nested Random Effects. *International Statistical Review* 57: 145-159.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* 4: 1033-1048.

Robert, C.R. (1992). Simulation of truncated normal variables. Technical Report No. 161, LSTA, University Paris 6.

Quaas, D. L. and Van Vleck, L. D. (1980). Categorical trait sire evaluation by best linear unbiased prediction of future progeny category frequency. *Biometrics* **36**: 117-122.

Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance Components.* Wiley.

Stiratelli, R., Laird, N. M., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics* 40:961-971.

Stram, D. O., Wei, L. J., and Ware, J. H. (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *Journal of the American Statistical Association* **83**: 631-637.

Tanner, M.A. (1991). *Tools for statistical inference: observed data and data augmentation methods.* Springer-Verlag. Berlin.

Tierney, L. (1991). Markov chains for exploring posterior distributions. Technical Report No. 560, School of Statistics, University of Minnesota.

Weil, C. S. (1970). Selection of the valid number of sampling units and consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food and Cosmetic Toxicology* **8**: 177-182.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**: 949-952.

Wong, G. Y. and Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association* **80**: 513-524.

Wu, C.-F. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**: 95-103.

Zeger, S.L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**: 121-130.

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

Figure 1. EM iterations using a Gibbs sampler for the salamander data set 1.

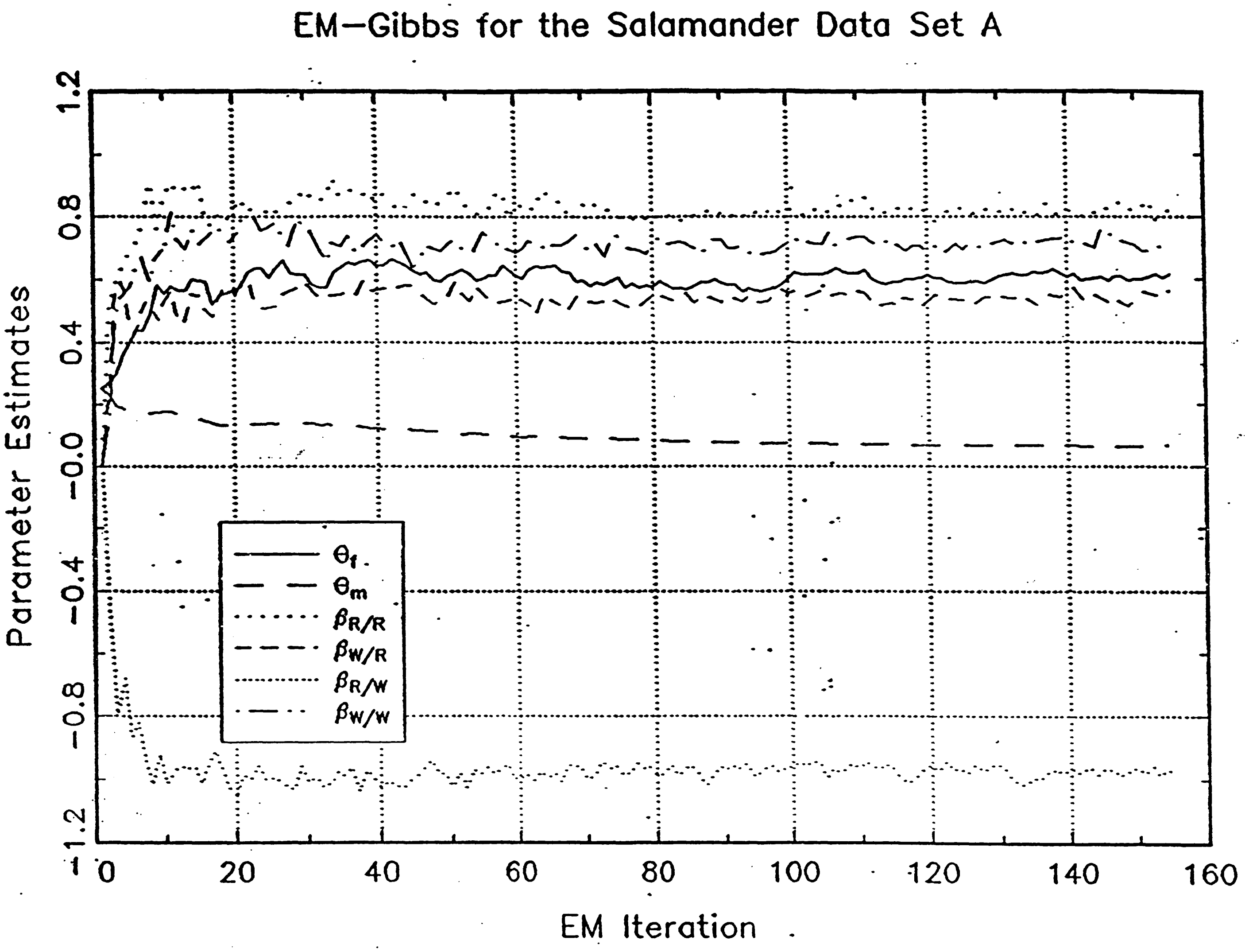

Table 1. Comparison of ML and Bayes estimates of the female (θ_f) and male (θ_m) variance components from the three salamander data sets (McCullagh and Nelder, 1989). The Bayesian estimates are taken from Karim and Zeger (1992, Table 4.) and are divided by $\left(\frac{\pi}{\sqrt{3}}\frac{15}{16}\right)^2$ for comparability (Johnson and Kotz, 1970, p. 6).

		Data Set					
		1		2		3	
Estimate	Variance	θ_f	θ_m	θ_f	θ_m	θ_f	θ_m
ML		.60	.06	.49	.45	.10	.44
Bayes		.81	.05	1.03	.49	.11	1.00