Discussiones Mathematicae
Probability and Statistics 24 (2004) 255–280

APPLICATION OF THE RASCH MODEL IN CATEGORICAL PEDIGREE ANALYSIS USING MCEM: I BINARY DATA

Guoqi Qian, Richard M. Huggins

Department of Statistical Science, La Trobe University
VIC, 3086, Australia

e-mail: g.qian@latrobe.edu.au
e-mail: r.huggins@latrobe.edu.au

and

Danuta Z. Loesch

School of Psychological Science, La Trobe University
VIC, 3086, Australia

e-mail: d.loesch@latrobe.edu.au

Abstract

An extension of the Rasch model with correlated latent variables is proposed to model correlated binary data within families. The latent variables have the classical correlation structure of Fisher (1918) and the model parameters thus have genetic interpretations. The proposed model is fitted to data using a hybrid of the Metropolis-Hastings algorithm and the MCEM modification of the EM-algorithm and is illustrated using genotype-phenotype data on a psychological subtest in families where some members are affected by the genetic disorder fragile X. In addition, hypothesis testing and model selection methods based on the Wald statistic are discussed.

Keywords: pedigree analysis, binary data, MCEM algorithm, Metropolis-Hastings algorithm.

1991 Mathematics Subject Classification: 62F10, 62F03, 92D30.

1. Introduction

In the study of genetic disorders and genotype-phenotype relationships categorical outcomes are often of interest. In particular, many psychological or clinical test items have categorical outcomes and an analysis of these data is important in understanding how genetic disorders affect cognitive and clinical status. However, the analysis of categorical outcomes from family data is not well developed and is less well understood than that of quantitative data. The analysis of quantitative family data under the multivariate normal model is well established, has been extensively applied and the parameters are readily interpreted. In particular, the arguments of Fisher (1918) and approach of Lange *et al.* (1976), Hopper & Mathews (1982), Hopper (1993) allow the estimation of genetic and environmental variance components of a polygenic trait from family data, as well as the effect of a genetic disorder on trait mean. Here we propose a latent variable approach to model the correlations between individuals within the same family so that the standard, readily interpretable models commonly used in quantitative genetics may be applied to binary family data.

1.1. Motivation

We are motivated by correlated binary data arising in a large of families affected by the fragile X syndrome. This disorder is one of the most common inherited forms of intellectual disability and results in a deficit of a protein (FMRP) in affected individuals. We illustrate our approach by considering aspects of the Behaviour Dyscontrol Scale (BDS), which is a measure of the capacity to use intentions to guide the performance of purposeful behaviour. We consider item 2, "Tap twice with non-dominant hand and once with dominant hand in a series" was rated on a scale of 0-3, with 3 representing no errors and 0 poor performance. Here we combine the categories 2 & 3 as "good performance" and 0 & 1 as "poor performance". The data on this variable consisted of 218 observations from 46 families affected with the fragile X condition. The biological interest is in whether there is any effect of FMRP on high motor control represented by item 2 of BDS after adjusting for FSIQ (full scale IQ), as this implies that this protein deficit affects some specific cognitive processes disproportionately to overall intellectual impairment, and in estimation of the genetic and environmental

correlations between relatives. A boxplot of FMRP levels for the two categories (good and poor performance, respectively) are given in Figure 1.

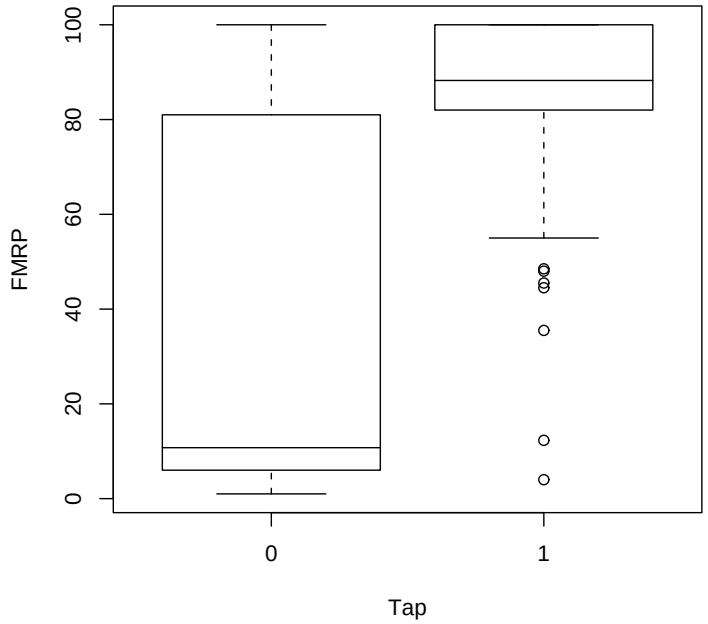

Figure 1. Box plot of FMRP for poor performing (0) and better performing groups (1) on the BDS subtest.

It is clear from the plot that FMRP can predict performance on this test. However, we will also include FSIQ and sex as fixed effects in the mean model, and simultaneously take the performance of other family members into account through a covariance model.

1.2. Modelling correlated binary data

There are several possible approaches to the modelling of categorical family data. The models of Bonney (1986) for binary data have the advantage of being easily fitted to data using standard logistic regression methods. However, Bonney (1986) only models relationships between family members and does not include genetic and environmental factors. As noted in FitzGerald and Knuiman (1998), there are problems in interpretation of the Bonney

models if there are missing data. An alternate approach using latent variables and a threshold is possible for uncorrelated binary data (eg. Albert & Chib 1993). In these models, values of the latent variable above a threshold are associated with the occurrence of the event. Albert & Chib (1993) propose a Gibb's sampling approach to these models and a review of the methods are contained in Chib (2000).

Latent variable models for the correlations between relatives have many attractions, particulary if the categorical outcome is the manifestation of an underlying polygenic effect. The Rasch model (Rasch 1980) is commonly used to model random effects via latent variable, see Fischer & Molenaar (1995) for an extensive review of the model in the item response setting and Albert & Ghosh (2000) for a Bayesian approach. We prefer this model to the threshold models as it associates a large value of the latent variable with a high probability of the occurrence of the event rather than the certain occurrence of the event. In its simplest form, if $X \sim f(x|\theta)$ represents the latent variable, and y is a binary outcome taking the value 1 if the event occurs and zero otherwise, then $\text{logit}\left(P(Y = 1|X = x)\right) = x$, and interest is in the estimation of θ from observations on Y. However, when data arises from family studies, the outcomes within a family may be correlated due to both genetic and common environment effects. That is, the latent variables X for different individuals within a family may be correlated, with the correlation being due to both shared genes and shared environment. The interest is then in estimating these genetic and environmental effects.

Here we adapt well known quantitative models for genetic and environmental correlation and use the Rasch model to relate latent variables to the binary outcome Y. The model is fitted using a variation of the EM algorithm where the E-step is implemented by the Metropolis-Hastings (MH) algorithm, a Markov chain Monte Carlo (MCMC) sampling technique, and the M-step by the Newton-Raphson algorithm. The idea of using an MCMC approximation in the E-step was motivated by Wei and Tanner (1990) who used a Monte Carlo approximation and called it the MCEM algorithm. Guo and Thompson (1994) used the MCEM, implemented by the Gibbs sampler in the E-step, in fitting a linear mixed model for complex pedigree data. Applications of the MCEM can also be found in the analyses of time series involving small counts (Chan and Ledolter 1995) and of grouped survival data (Sinha *et al.* 1994).

2. THE RASCH MODEL FOR PEDIGREE DATA

We suppose a binary outcome is dependent on both observed covariates, z, and unobserved latent variables, x. Let $j = 1, \cdots, J$ denote the families in the study, with family j containing n_j members. Let $Y_{ji} = 1$ if individual i in family j has a positive outcome and 0 otherwise, $\mathbf{z}_{ji}$ denote a $p \times 1$ vector of covariates associated with individual i in family j and $\mathbf{Z}_j = (\mathbf{z}_{j1}, \cdots, \mathbf{z}_{jn_j})^T$ the corresponding design matrix for family j. The conditional dependence between Y_{ij} and the covariates $\mathbf{z}_{ji}$ given the latent variable $x_{\cdot ji}$, corresponding to individual i in family j, is assumed to follow the logistic model

$$P(Y_{ji} = 1 | x_{\cdot ji}) = \frac{e^{x_{\cdot ji} + \mathbf{z}_{ji}^T \beta}}{1 + e^{x_{\cdot ji} + \mathbf{z}_{ij}^T \beta}} \equiv \pi_{ji}(\beta; \mathbf{z}_{ji}, x_{\cdot ji}). \tag{1}$$

The latent variables associated with an individual are taken to depend on polygenic effects that contribute to correlations between relatives. Two important elements of the genetic correlations between relatives are the additive and dominance (non-additive) genetic components, which are modelled using the kinship matrix, $\mathbf{\Phi}_j \equiv (\phi_{ik})_j$, and Jacquard's condensed coefficient of identity matrix, $\mathbf{\Delta}_j \equiv (\Delta_{ik})_j$, defined in Lange *et al.* (1976) for example. We omit assortive mating models as they would be more appropriate to a variable like FSIQ than a specific executive ability. Thus, $\phi_{ii} = \frac{1}{2}$ for each individual i; and for each pair of relatives i and k, $\phi_{ik} = \frac{1}{4}$ for first degree relatives, $\frac{1}{8}$ for second degree relatives, and so on. In the pedigrees considered, $\Delta_{ii} = 1$ and $\Delta_{ik} = 0.25$ if i and k are full siblings and is zero otherwise. The environmental correlation consists of individual and common environment components. Therefore we introduce 4 independent latent vectors $\mathbf{x}_{gj} = (x_{gj1}, \cdots, x_{gjn_j})^T$ ($g = 1, 2, 3, 4$ and $j = 1, \cdots, J$) for each family corresponding to the four components: additive genetic ($g = 1$), dominance genetic ($g = 2$), individual environment ($g = 3$) and common environment ($g = 4$). In the model (1) we take $x_{\cdot ji} = x_{1ji} + \cdots + x_{4ji}$. The latent vectors for the different families are assumed to be independent. Each latent vector $\mathbf{x}_{gj}$ is assumed to follow a multivariate normal distribution $\text{MVN}(\mathbf{0}, \sigma_g^2 \mathbf{V}_{gj})$ where $\sigma_1^2 \equiv \sigma_a^2$, $\sigma_2^2 \equiv \sigma_d^2$, $\sigma_3^2 \equiv \sigma_e^2$, $\sigma_4^2 \equiv \sigma_c^2$, $\mathbf{V}_{1j} \equiv \mathbf{\Phi}_j$, $\mathbf{V}_{2j} \equiv \mathbf{\Delta}_j$, $\mathbf{V}_{3j} \equiv \mathbf{I}_j$, $\mathbf{V}_{4j} \equiv \mathbf{C}_j$. Here $\mathbf{I}_j$ is the identity matrix and $\mathbf{C}_j$ is the identity matrix plus a matrix of ones. Denoting $\mathbf{x}_{\bullet j} \equiv (x_{\cdot j1}, \cdots, x_{\cdot jn_j})^T = \sum_{g=1}^4 \mathbf{x}_{gj}$, the covariance matrix of $\mathbf{x}_{\bullet j}$ is then $\mathbf{\Omega}_j = \sigma_a^2 \mathbf{\Phi}_j + \sigma_d^2 \mathbf{\Delta}_j + \sigma_e^2 \boldsymbol{I}_j + \sigma_c^2 \boldsymbol{C}_j$, which summarizes the overall genetic and environmental variability of family j.

(This differs from the usual definition as the additive component is typically taken as $2\sigma_a^2\mathbf{\Phi}_j$ and for technical reasons in our calculations below, $\mathbf{C}_j$ also differs from what is usual. Thus the usual additive genetic variance is $\sigma_a^{*2} = \sigma_a^2/2$ and the usual individual environment variance is $\sigma_e^{*2} = \sigma_e^2 + \sigma_c^2$.) If the $\mathbf{x}_{gj}$ or $\mathbf{x}_{\bullet j}$ were observable, then the variance components $\boldsymbol{\sigma}^2 = (\sigma_1^2, \cdots, \sigma_4^2)^T$ could be estimated using maximum likelihood (Lange *et al.* 1976), or robust methods (Huggins 1993).

3. Fitting the model

Our interest is in the parameter vector $\theta = (\beta^T, \sigma_a^2, \sigma_d^2, \sigma_e^2, \sigma_c^2)^T$. We suppose that given the $\mathbf{x}_{gj}$ and the $\mathbf{Z}_j$, the Y_{ji} are independent. Hence, the conditional probability of $\mathbf{Y}_j = (Y_{j1}, \cdots, Y_{jn_j})^T$ given the $\mathbf{x}_{gj}$ is $L_j(\mathbf{Y}_j|\mathbf{x}_{1j}, \cdots, \mathbf{x}_{4j}; \theta) = \prod_{i=1}^{n_j} e^{Y_{ji}(x_{\cdot ji} + \mathbf{z}_{ji}^T\beta)}(1 + e^{x_{\cdot ji} + \mathbf{z}_{ji}^T\beta})^{-1}$ and the joint probability function of $\mathbf{Y}_j$ and $(\mathbf{x}_{1j}, \cdots, \mathbf{x}_{4j})$ is

$$L_j\Big(\mathbf{Y}_j, \mathbf{x}_{1j}, \cdots, \mathbf{x}_{4j}; \theta\Big) = \prod_{i=1}^{n_j} \frac{e^{Y_{ji}(x_{\cdot ji} + \mathbf{z}_{ji}^T\beta)}}{1 + e^{x_{\cdot ji} + \mathbf{z}_{ji}^T\beta}}$$

$$\times \prod_{g=1}^{4} \Big(2\pi\sigma_g^2\Big)^{-\frac{1}{2}n_j} |\mathbf{V}_{gj}|^{-\frac{1}{2}} e^{\{-\frac{1}{2}\sigma_g^{-2}\mathbf{x}_{gj}^T\mathbf{V}_{gj}^{-1}\mathbf{x}_{gj}\}}.$$

Hence the unconditional distribution of $\mathbf{Y}_j$ is

$$L_j(\theta) = \int \cdots \int L_j\Big(\mathbf{Y}_j, \mathbf{x}_{1j}, \cdots, \mathbf{x}_{4j}; \theta\Big) d\mathbf{x}_{1j} \cdots d\mathbf{x}_{4j}. \tag{2}$$

Denote by $\mathbf{Y} = (\mathbf{Y}_1^T, \cdots, \mathbf{Y}_J^T)^T$ the response vector for all the J families. The likelihood function for $\mathbf{Y}$ is then $L_y(\theta) = \prod_{j=1}^{J} L_j(\theta)$. Directly maximizing $L_y(\theta)$ to find the MLE of θ is computationally difficult as many multiple integrals are involved in $L_y(\theta)$.

To overcome these computational difficulties we apply the EM algorithm together with an MCMC sampling technique and the Newton-Raphson algorithm to find the MLE of θ. This is motivated by the following facts: Firstly, the complete-data likelihood involves only the logistic and the multivariate normal density functions but not any multiple integration.

Hence it would be relatively easy to estimate θ using the Newton-Raphson algorithm if $\mathbf{Y} = (\mathbf{Y}_1^T, \cdots, \mathbf{Y}_J^T)^T$, the covariate matrices $\{\mathbf{Z}_1, \cdots, \mathbf{Z}_J\}$ and the latent variables $\mathbf{X}_g = (\mathbf{x}_{g1}^T, \cdots, \mathbf{x}_{gJ}^T)^T$ $(g = 1, 2, 3, 4)$ were observed. Secondly, the conditional distribution of $\mathbf{x}_{gj}$ given $\mathbf{Y}_j$, although complicated, may be simulated by the MH algorithm, as the distributions of $\mathbf{x}_{gj}$ and $\mathbf{Y}_j$ given $\mathbf{x}_{\bullet j}$ are both well known.

The complete-data log-likelihood of θ for given $\mathbf{Y}$ and $\{\mathbf{X}_g, g = 1, \cdots, 4\}$ is

$$\ell_{yx}\Big(\theta; \mathbf{Y}, \mathbf{X}_1, \cdots, \mathbf{X}_4\Big) = \ell_{y|x}\Big(\beta; \mathbf{Y}|\mathbf{X}_1, \cdots, \mathbf{X}_4\Big) + \sum_{g=1}^{4} \ell_{xg}\Big(\sigma_g^2; \mathbf{X}_g\Big) \tag{3}$$

where

$$\begin{aligned} &\ell_{y|x}\Big(\beta; \mathbf{Y}|\mathbf{X}_1, \cdots, \mathbf{X}_4\Big) \\ &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} \log\left\{ e^{Y_{ji}(x_{\cdot ji} + \mathbf{z}_{ji}^T \beta)} \left(1 + e^{x_{\cdot ji} + \mathbf{z}_{ji}^T \beta}\right)^{-1} \right\} \end{aligned} \tag{4}$$

is the conditional log-likelihood of $\mathbf{Y}$ given $\{\mathbf{X}_g, g = 1, \cdots, 4\}$, and

$$\begin{aligned} &\ell_{xg}\Big(\sigma_g^2; \mathbf{X}_g\Big) \\ &= -\frac{1}{2} \sum_{j=1}^{J} n_j \log 2\pi\sigma_g^2 - \frac{1}{2} \sum_{j=1}^{J} \log |\mathbf{V}_{gj}| - \frac{1}{2} \sigma_g^{-2} \sum_{j=1}^{J} \mathbf{x}_{gj}^T \mathbf{V}_{gj}^{-1} \mathbf{x}_{gj} \end{aligned} \tag{5}$$

is the marginal log-likelihood of $\mathbf{X}_g$. In order to use the EM algorithm we let $\theta' = (\beta'^T, \boldsymbol{\sigma}^{2\prime T})^T$ and define

$$\begin{aligned} Q\Big(\theta, \theta'\Big) &= E\Big(\ell_{y|x}\Big(\beta; \mathbf{Y}|\mathbf{X}_1, \cdots, \mathbf{X}_4)|\mathbf{Y}, \theta'\Big) \\ &\quad + \sum_{g=1}^{4} E\Big(\ell_{xg}\Big(\sigma_g^2; \mathbf{X}_g)|\mathbf{Y}, \theta'\Big) \end{aligned} \tag{6}$$

which is the conditional expectation of the complete-data log-likelihood with respect to the conditional distribution of $\{\mathbf{X}_g, g = 1, \cdots, 4\}$ given $\mathbf{Y}$ and $\theta = \theta'$. Then (see e.g. Dempster *et al.* 1977) the MLE of θ is obtained by iteratively updating the maximizer of $Q(\theta, \theta')$ for the current estimate θ' until convergence is attained.

From (4) and (5) we see that β appears only in the first term of (6) while σ_g^2 appears only in one of the summation terms in (6). Moreover, it is easy to show that (6) is second order differentiable with respect to θ and has a unique maximizer. Therefore, suppose $\hat{\theta}(r) = (\hat{\beta}^T(r), \hat{\boldsymbol{\sigma}}^{2T}(r))^T$ is the r-th step estimate of θ arising from the EM algorithm, the $(r+1)$-th step estimate of θ can be obtained by solving

$$\frac{\partial}{\partial \beta} Q(\theta, \hat{\theta}(r)) = \frac{\partial}{\partial \beta} E(\ell_{y|x}(\beta; \mathbf{Y}|\mathbf{X}_1, \cdots, \mathbf{X}_4)|\mathbf{Y}, \hat{\theta}(r)) = 0 \tag{7}$$

$$\frac{\partial}{\partial \sigma_g^2} Q(\theta, \hat{\theta}(r)) = \frac{\partial}{\partial \sigma_g^2} E(\ell_{xg}(\sigma_g^2; \mathbf{X}_g)|\mathbf{Y}, \hat{\theta}(r)) = 0, \quad g = 1, 2, 3, 4. \tag{8}$$

The equations (7) and (8) can be solved by either Newton-Raphson or Fisher scoring method if the conditional expectations involved are easily evaluated (see Appendix 2).

Note that from (2) the conditional distribution of $\{\mathbf{X}_g, g = 1, \cdots, 4\}$ given $\mathbf{Y}$ is

$$\prod_{j=1}^{J} L_j(\mathbf{x}_{1j}, \cdots, \mathbf{x}_{4j}|\mathbf{Y}_j, \theta) = \prod_{j=1}^{J} (L_j(\theta))^{-1} L_j(\mathbf{Y}_j, \mathbf{x}_{1j}, \cdots, \mathbf{x}_{4j}; \theta), \tag{9}$$

which according to (2) involves $4 \times \sum_{j=1}^{J} n_j$ integrals. Thus an exact evaluation of the conditional expectations, and hence $Q(\theta, \theta')$, is infeasible because of the complexity of this conditional distribution. To overcome this, one can apply the MH algorithm (see e.g. Chib and Greenberg 1995) or related MCMC techniques to generate a sample of $\{\mathbf{X}_g, g = 1, \cdots, 4\}$ values from the conditional distribution (9). Then the conditional expectations in (7), (8) and $Q(\theta, \theta')$ can be approximated by conditional averages based on the generated sample. For the conditional distribution (9), it is particularly convenient to use an MH algorithm where we choose the

multivariate normal as the operating transition density and an easily verified acceptance-rejection condition. The between-within variance criterion of Gelman and Rubin (1992) can be used to monitor the convergence of the MH algorithm. Details of the algorithm are given in the appendix. Implementation of the MCEM algorithm is also discussed in Appendix 1.

3.1. Convergence in the MCEM algorithm

An important difference between MCEM and EM algorithms is that the observed-data likelihood always increases along an EM sequence while this is not guaranteed in MCEM (e.g. Tanner 1996, Section 4.5). In solving (7) and (8) we replace $Q(\theta, \hat{\theta}(r))$ by its MCMC approximation $\tilde{Q}(\theta, \hat{\theta}(r))$. Thus the numerical solution $\hat{\theta}(r+1)$ is only the maximizer of $\tilde{Q}(\theta, \hat{\theta}(r))$ but not of $Q(\theta, \hat{\theta}(r))$. Consequently, the estimated likelihood sequence $\{\tilde{L}_y(\hat{\theta}(r)\}$ — a Monte Carlo approximation of $\{L_y(\hat{\theta}(r)\}$ — is not increasing and $\hat{\theta}(r)$ does not necessarily converge as $r \to \infty$. Thus, unlike the EM algorithm, there is no precise rule for determining the convergence of the MCEM algorithm. However, the sequence $\{\tilde{L}_y(\hat{\theta}(r)\}$ often exhibits an increasing trend; and provided that the approximation error between $\tilde{Q}(\theta, \hat{\theta}(r))$ and $Q(\theta, \hat{\theta}(r))$ is small enough, the sequence will fluctuate around the value of the maximized observed-data likelihood once r becomes sufficiently large. The sequence $\{\hat{\theta}(r)\}$ would also fluctuate around the MLE $\hat{\theta}$ when r is sufficiently large. Chan and Ledolter (1995) showed that under suitable regularity conditions the MCEM sequence of the parameter estimate updates becomes close to the MLE with high probability. Therefore, to monitor the convergence of MCEM algorithm we can plot $\tilde{L}_y(\hat{\theta}(r))$ as well as $\hat{\theta}(r)$ against iteration number r. We terminate the algorithm when the sequence becomes stationary. Otherwise, we continue by increasing the Monte Carlo precision in the E-step provided that the required computation is computationally feasible. The final MLE $\hat{\theta}$ and $\tilde{L}_y(\hat{\theta})$ are approximated by the corresponding averages of the $\{\hat{\theta}(r)\}$ and $\{\tilde{L}_y(\hat{\theta}(r))\}$ sub-sequences in the stationarity state. In Appendix 1 we will show that for our model, monitoring the convergence of MCEM using $\{\tilde{L}_y(\hat{\theta}(r))\}$ may be more reliable than using $\hat{\theta}(r)$.

3.2. The standard error of $\hat{\theta}$

Large sample theory yields that under regularity conditions asymptotically the MLE $\hat{\theta}$ has a normal distribution $\text{MVN}(\theta, \mathbf{I}^{-1}(\theta))$. We estimate

$\mathbf{I}(\hat{\theta})$ by the observed information matrix $\mathbf{J}_{\hat{\theta}}(\mathbf{Y}) = -\partial^2 \log L_y(\theta)/\partial\theta\partial\theta^T|_{\theta=\hat{\theta}}$. Applying a basic result of Louis (1982) and Tanner (1996, Section 4.4.3) one can show that $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ is

$$(10)\quad \left(-\frac{\partial^2 Q(\theta,\theta')}{\partial\theta\partial\theta^T} - \mathrm{Var}\{\sum_{j=1}^{J}\frac{\partial}{\partial\theta}\log L_j(\mathbf{Y}_j,\mathbf{x}_{1j},\cdots,\mathbf{x}_{4j};\theta)|\mathbf{Y},\theta'\}\right)_{\theta=\theta'=\hat{\theta}}$$

which can be represented as *Observed information = Complete information − Missing information*, implying that the unobserved latent variables $\mathbf{x}_{gj}$ increase the variability in the MLE $\hat{\theta}$. The first term in the right hand side of (10) is approximately $-\partial^2\tilde{Q}(\theta,\theta')/\partial\theta\partial\theta^T|_{\theta=\theta'=\hat{\theta}}$ and is available as a byproduct of computing $\hat{\theta}$. The conditional variance in the second term of (10) can be estimated by the sample variance after samples of $\{\mathbf{x}_{gj}, j = 1,\cdots,J; g = 1,\cdots,4\}$ are generated from the conditional distribution (9) using the MH algorithm. We provide the details of the estimation of $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ in Appendix 4 where it is noted that the estimate may not be positive definite. An empirical approach to obtain a positive definite estimate of $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ with low degree of singularity is to multiply the sample variance term, which estimates the missing information, by a positive scalar $\lambda \leq 1$. This approach is similar to that used in ridge regression for handling collinearity and near singularity of the design matrix. The optimal value of λ is then chosen in such a way that the resulting estimate of $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ is positive definite and has the same degree of non-singularity (defined by the condition number of the matrix) as the estimate $-\partial^2\tilde{Q}(\theta,\theta')/\partial\theta\partial\theta^T|_{\theta=\theta'=\hat{\theta}}$ of the complete information. This will be illustrated in Section 4 and further discussed in Appendix 4.

3.3. Inference for the variance components by the wald test

We illustrate the approach by testing for the effects of the non-additive genetic and common environment variance components in the model and focus on testing $H_0 : \sigma_d^2 = \sigma_c^2 = 0$ against H_1 : *no restrictions on* σ_d^2 *and* σ_c^2. Three possible methods for testing these hypotheses are the Wald test, the score test and the likelihood ratio test. We found that the Fisher information matrix associated with the score test was not definite, the Monte Carlo approximations possibly resulted in negative values of the log likelihood ratio statistic so it was not applicable, and the distribution of the predictive likelihood ratio statistic was intractable.

Therefore, we use the Wald test. Let $\theta(H)$ and $\hat{\theta}(H)$ denote the values of θ and its MLE respectively under a hypothesis H. The Wald test statistic is $W = (\hat{\sigma}_d^2(H_1), \hat{\sigma}_c^2(H_1))\{\mathrm{Var}(\hat{\sigma}_d^2(H_1), \hat{\sigma}_c^2(H_1))\}^{-1}(\hat{\sigma}_d^2(H_1), \hat{\sigma}_c^2(H_1))^T$ which asymptotically has a χ_2^2 distribution under H_0 and does not require any significant extra computation beyond that involved in finding the MLE $\hat{\boldsymbol{\sigma}}^2(H_1)$ and its estimated covariance matrix under H_1. The precision of W is dependent on that of $(\hat{\sigma}_d^2(H_1), \hat{\sigma}_c^2(H_1))$ and their estimated covariance, which may be constrained by the high computational cost of achieving a close approximation in the MCEM algorithm.

Sommer & Huggins (1996) introduced a variable selection procedure based on the Wald test that is equivalent to Mallow's C_p in linear regression. Let S denote the number of parameters in the full model, write $\theta^T = (\theta_1^T, \theta_2^T)$ where θ_1 is the p-dimensional vector of parameters in the sub-model of interest and $\theta_2 = 0$ for this model. Similarly partition the covariance matrix $\Sigma = \mathbf{J}_{\hat{\theta}}^{-1}(\mathbf{Y})$ as

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Let $\hat{\theta}^T = (\hat{\theta}_1^T, \hat{\theta}_2^T)$ be the estimate of θ arising from the full model. Then the Wald test of $H_0 : \theta_2 = 0$ is based on the statistic $W_p = \hat{\theta}_2^T \Sigma_{22}^{-1} \hat{\theta}_2$. Sommer & Huggins (1996) then compute $T_p = W_p - S + 2p$ and plot this against p. Models with T_p values close to or less than p yield reasonable models for the data.

4. Application

The models for good or poor performance on item 2 of the BDS were fitted by the hybrid of MH and MCEM algorithms. We consider the Rasch model: $\mathrm{logit}P(Y_{ji}=1) = x_{1ji} + x_{2ji} + x_{3ji} + x_{4ji} + \beta_0 + \beta_1 \mathrm{sex} + \beta_2 \mathrm{FMRP} + \beta_3 \mathrm{FSIQ}$ where x_{gji}'s are latent variables defined in Section 2 and sex takes the value 1 for females and 0 for males. In applying the MCEM algorithm, we generated a sequence of 500 replicates of $\hat{\theta}(r)$ to find the MLE of θ. The parameters K and B in the MCE-step were chosen to be 1,000 for the first 100 $\hat{\theta}(r)$s; 5,000 for the next 100 $\hat{\theta}(r)$s; 10,000 for the next 40 $\hat{\theta}(r)$ and 20,000 for the final 260 $\hat{\theta}(r)$s. When using the MH algorithm to generate the conditional distribution in the MCE-step, the first

1,000 in the sequence was discarded, which was sufficient for the remaining to be stationary. (This was tested using the Gelman-Rubin between-within variance criterion.) The performance is summarized in Figure 2.

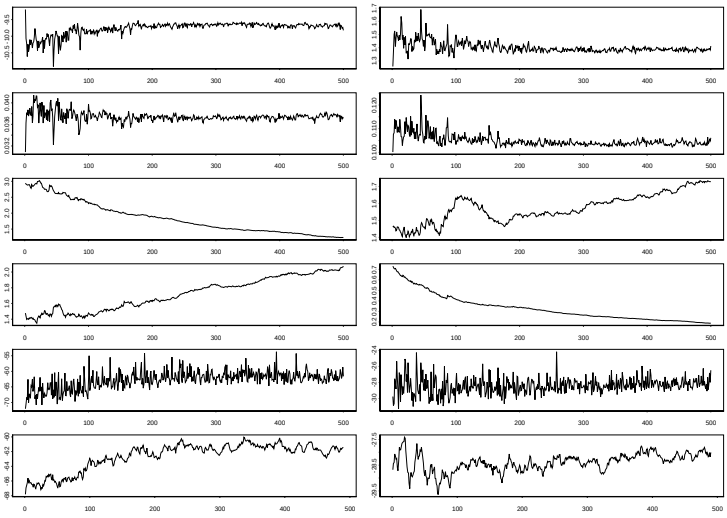

Figure 2. Plotted row-by-row are the MCEM sequences of $\hat{\theta}(r)$, $\tilde{L}_y(\hat{\theta}(r))$, $\tilde{L}_y(\hat{\theta}(r)|\mathbf{Y})$, and moving averages (in span 10) of $\tilde{L}_y(\hat{\theta}(r))$ and $\tilde{L}_y(\hat{\theta}(r)|\mathbf{Y})$ for BDS data.

Plotted from left to right then from top to bottom are the MCEM sequences of $\hat{\beta}_0(r)$, $\hat{\beta}_1(r)$, $\hat{\beta}_2(r)$,$\hat{\beta}_3(r)$, $\hat{\sigma}_a^2(r)$, $\hat{\sigma}_d^2(r)$, $\hat{\sigma}_e^2(r)$, $\hat{\sigma}_c^2(r)$, observed-data log-likelihood $\log \tilde{L}_y(\hat{\theta}(r))$, predictive posterior log-likelihood $\log \tilde{L}_y(\hat{\theta}(r)|\mathbf{Y})$ (defined in Appendix 5), and moving averages (of span 10) of $\log \tilde{L}_y(\hat{\theta}(r))$ and $\log \tilde{L}_y(\hat{\theta}(r)|\mathbf{Y})$. Plots of the moving averages of $\log \tilde{L}_y(\hat{\theta}(r))$ make the increasing trend of $\log \tilde{L}_y(\hat{\theta}(r))$ more stand-out. As explained in Appendix 1, some sequences of the variance components estimates do not converge. But by examining the sequences of $\hat{\beta}(r)$'s and $\log \tilde{L}_y(\hat{\theta}(r))$, one may regard the MCEM process becoming stationary after about 250 iterations. Therefore, the MLE of θ was taken to be the average over the last

250 iterations, $\hat{\theta} = 250^{-1} \sum_{r=251}^{500} \hat{\theta}(r)$. Then the method of Section 3.2 is used to compute $\mathrm{Var}(\hat{\theta})$ and the standard errors. Table 1 lists the values of $\hat{\theta}$ and their standard errors for some candidate values of λ (see Appendix 4).

Table 1. Estimates of θ for BDS data. The p-values for covariate effects are calculated based on the approximate two-sided z-test while the p-values for variance components are obtained based on the Wald test. ($\sigma_a^{2*} = \sigma_a^2/2$ which is the usual parameterization of the model.)

Term	Estimate	Standard Error at different values of λ							p-value
		0.50	0.95	*0.96*	0.97	0.98	0.99	1.0	
Intercept	-9.667	1.646	1.875	*1.912*	2.034	1.732	1.911	2.099	0.0000
Sex	1.384	0.683	0.769	*0.779*	0.816	0.708	0.756	0.782	0.0755
FMRP	0.037	0.015	0.018	*0.018*	0.018	0.017	0.018	0.019	0.0364
FSIQ	0.104	0.026	0.029	*0.029*	0.030	0.028	0.029	0.030	0.0004
σ_a^{2*}	0.721	0.097	0.277	*0.318*	0.441	NA	0.239	0.329	0.0235
σ_d^2	1.632	0.220	0.775	*0.945*	1.477	NA	0.160	1.035	0.0843
σ_e^2	1.902	0.254	0.673	*0.725*	0.794	0.881	1.024	1.317	0.0087
σ_c^2	0.235	0.032	0.117	*0.143*	0.226	NA	NA	0.151	0.1007

The selected λ and standard errors are printed in italics in the table and were used to compute the p-values. Also the NAs are the results of non-positive definite estimate of $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ and $\mathrm{Var}(\hat{\theta})$ computed at certain λ values. Shown in Figure 3 are the plots of the estimated variances for every $\hat{\theta}$ component and the condition numbers of estimates of $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ (or equivalently $\mathrm{Var}(\hat{\theta})$) against values of λ in [0,1]. From the plots of condition numbers and the discussion in Appendix 4, we see the optimal λ value 0.96 should be used for estimating $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ and $\mathrm{Var}(\hat{\theta})$.

We tested the effects of the variance components σ_d^2 and σ_c^2, by the Wald test. The test statistic is $W = 3.8446$ which follows a χ_2^2 distribution.

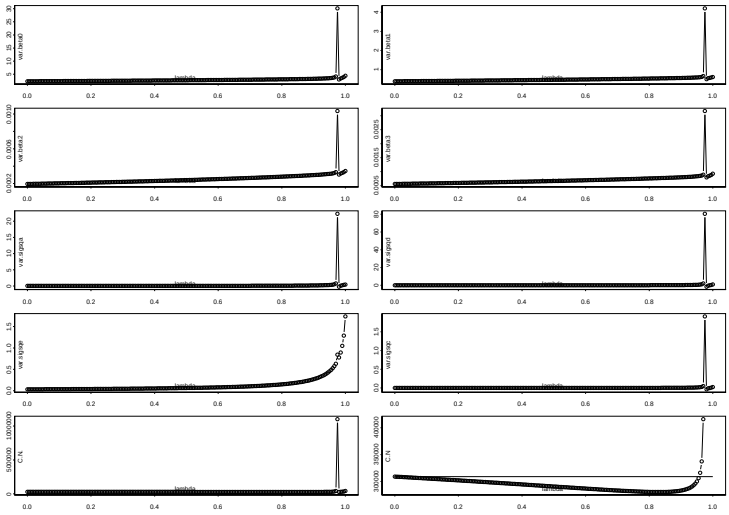

Figure 3. Plots for BDS data of variances of $\hat{\theta}$ and condition number of $\tilde{\mathbf{J}}_{\hat{\theta}}(\mathbf{Y})$ against λ. The bottom-right plot is the enlargement of part of the bottom-left plot ($\lambda \in [0, 0.97]$).

Hence the p-value=0.1463 implying there is not strong statistical evidence of jointly significant effects of σ_d^2 and σ_c^2. The T_p plot of Figure 4 suggests a model (123457) containing all the fixed effects and additive and individual environmental variance components as most appropriate. The next two closest models (1234567, 1234578) both contain all the fixed effects and additive and individual environment but one contains non-additive genetic and the other common environment components. The next model (1345678) contains all the fixed effects except sex and all four variance components. In the interests of parsimony the model 123457 seems preferable, and this model was also one of the better models for other values of λ near 0.96. In the full model, the main effects of FMRP and FSIQ are seen to be related to the probability of a good performance on the BDS subtest examined and even after adjusting for these variables there is a significant genetic correlation.

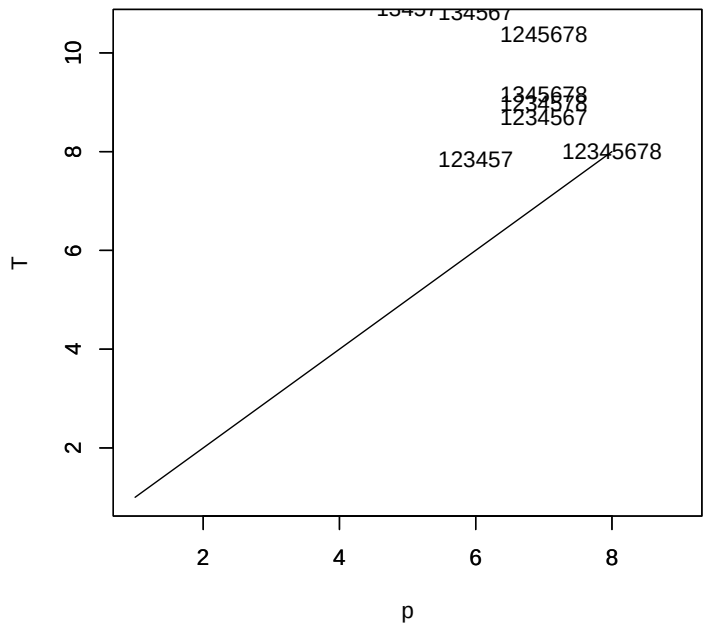

Figure 4. Plot of T_p against p for BDS data.

We conclude there is an effect of FMRP deficit (and hence fragile X) on BDS sub-test 2 beyond the effect on IQ and that after adjusting for FSIQ there is still a significant (or close to significant) genetic correlation between family members. The heritability in the full model $((0.5\sigma_a^2 + \sigma_d^2)/(0.5\sigma_a^2 + \sigma_d^2 + \sigma_e^2 + 2\sigma_c^2))$ was 0.498.

5. Discussion

We have demonstrated that MCEM methods together with MCMC sampling techniques may be applied to Rasch models for binary outcomes in pedigree data to obtain estimates of genetic and environmental variance

components that have an interpretation similar to those obtained in the more conventional analyses of quantitative traits. With minor adjustments, the method can be extended to models for categorical and ordinal data, and this shall be considered elsewhere. Moreover, in the practical application of the method, a more sophisticated treatment of missing values and refinements of the programs to enable the more usual parameterization of the variance components could also be considered.

The results obtained in fitting the model is biologically sensible, although further models could be fitted to properly elucidate the correlation structure. However, the results demonstrated a significant effect of shared genes on familiar correlations for Item 2 of BDS, as well as a significant effect of FMRP on the mean of this trait after adjustment for FSIQ. This result is consistent with our earlier data obtained on the total BDS score by analysis of quantitative family data under the multivariate normal model (Loesch *et al.* 2002).

The major difficulty in the method is the approximations used in computing the asymptotic covariance matrix that require an adaptive correction to obtain estimates of the standard error. The observed information $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ is in theory positive definite. But its computed estimate $\tilde{\mathbf{J}}_{\hat{\theta}}(\mathbf{Y})$ may not be so, even when the MCEM algorithm attains stationarity, due to the Monte Carlo approximation errors in estimating the variance term in (10) and the estimation of $Q(\theta, \theta')$ by $\tilde{Q}(\theta, \theta')$. This is more likely to happen when the observed-data likelihood is very flat along some directions, which implies that the observed information $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ could be quite ill-conditioned and the missing information is very close to the complete information. An alternative approach of estimating $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ would be to use the supplemented EM (SEM), properly modified, proposed by Meng and Rubin (1991) in which $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ is estimated by the estimated complete information multiplying a matrix $(I - DM)$ with DM being determined by the rate of convergence of EM. The difficulty with this approach is that rate of convergence of MCEM does not have the same behaviour pattern as that of EM. So there is an issue of how to get a convergent estimate of $I - DM$ being positive definite and not near-singular. In addition, as with the SEM algorithm, the modified SEM would require roughly $(\dim \theta + 1)/2$ times as much computational time as MCEM itself. In our approach, however, the computational time of estimating $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ is only about that for computing one iteration in MCEM.

A. APPENDIX

A.1. Implementing the MCEM algorithm

The implementation of the MCEM algorithm used was:

1. Choose an initial estimate $\hat{\theta}(0) = (\hat{\beta}^T(0), \hat{\sigma}_1^2(0), \cdots, \hat{\sigma}_4^2(0))^T$ for θ.
2. Repeat for $r = 0, 1, \cdots$ the following MCE- and M-steps until the Monte Carlo approximation $\tilde{L}_y(\hat{\theta}(r))$ of the likelihood $L_y(\hat{\theta}(r))$ does not show an increasing trend and attains stationarity. An empirical stopping rule is based on plotting $\tilde{L}_y(\hat{\theta}(r))$ against r.

MCE-step: Generate a sample $\{(\mathbf{X}_{11}^{(r)}, \cdots, \mathbf{X}_{41}^{(r)}), \cdots, (\mathbf{X}_{1K}^{(r)}, \cdots, \mathbf{X}_{4K}^{(r)})\}$ of size K from the conditional distribution (9) of $\{\mathbf{X}_g, g = 1, \cdots, 4\}$ given $\mathbf{Y}$ and $\theta = \hat{\theta}(r)$, using the MH algorithm of Appendix 3. The value K is determined by a trade-off between the accuracy of $\hat{\theta}(r)$ and the computational intensity. Then replace $Q(\theta, \hat{\theta}(r))$ in (7) and (8) by its MCMC approximation $\tilde{Q}(\theta, \hat{\theta}(r)) = K^{-1} \sum_{k=1}^{K} \ell_{yx}(\theta; \mathbf{Y}, \mathbf{X}_{1k}^{(r)}, \cdots, \mathbf{X}_{4k}^{(r)})$.

M-step: Compute $\hat{\theta}(r+1) = (\hat{\beta}(r+1)^T, \hat{\sigma}_1^2(r+1), \cdots, \hat{\sigma}_4^2(r+1))^T$ by solving the renewed (7) and (8) using the Newton-Raphson algorithm. Estimate the observed-data likelihood $L_y(\hat{\theta}(r+1))$ by its Monte Carlo approximation $\tilde{L}_y(\hat{\theta}(r+1)) = B^{-1} \sum_{b=1}^{B} \prod_{j=1}^{J} L_j(\mathbf{Y}_j | \hat{\mathbf{x}}_{1jb}(r+1), \cdots, \hat{\mathbf{x}}_{4jb}(r+1), \hat{\theta}(r+1))$ where $\hat{\mathbf{x}}_{gjb}(r+1) \sim \mathrm{MVN}(\mathbf{0}, \hat{\sigma}_g^2(r+1)\mathbf{V}_{gj})$ $(b = 1, \cdots, B; g = 1, \cdots, 4)$.

The most intensive computation lies in the generation of $\{\mathbf{X}_{g1}^{(r)}, \cdots, \mathbf{X}_{gK}^{(r)}, g = 1, \cdots, 4\}$ for each update $\hat{\theta}(r+1)$. The sample size K usually needs to be very large in order to have good precision in the Monte Carlo approximation. We found that the updates $\hat{\beta}(r)$ and $\tilde{L}_y(\hat{\theta}(r))$ eventually became stationary but updates of the variance components $\hat{\sigma}_g^2(r)$ sometimes only showed an asymptotic tendency. There are two possible reasons for this. Firstly, the observed-data likelihood $L_y(\theta)$ may be very flat near the MLE of σ_g^2 so that the change of $L_y(\theta)$ about σ_g^2 is masked by the Monte Carlo approximation error between $L_y(\theta)$ and $\tilde{L}_y(\theta)$. The second is that $\sigma_g^2 = 0$ is an absorbing state in generating $\{\mathbf{X}_{g1}^{(r)}, \cdots, \mathbf{X}_{gK}^{(r)}, g = 1, \cdots, 4\}$ by the MH algorithm. Once $\hat{\sigma}_g^2(r)$ is very close to zero, $\hat{\sigma}_g^2(r+1)$ will tend to remain close to zero whether or not 0 is the MLE of σ_g^2. In response to this, we may regard the above MCEM process as having attained stationarity once $\tilde{L}(\hat{\theta}(r))$ has.

A.2. Computing $\hat{\theta}(r+1)$ by the Newton-Raphson Algorithm

Equations (4) to (8) yield:

$$\frac{\partial}{\partial\beta}Q\Big(\theta,\hat{\theta}(r)\Big)=\sum_{j=1}^{J}\mathbf{Z}_j^T\Big(\mathbf{Y}_j-E\Big\{\boldsymbol{\pi}_j\Big(\beta;\mathbf{Z}_j,\mathbf{x}_{\bullet j}\Big)|\mathbf{Y}_j,\hat{\theta}(r)\Big\}\Big) \tag{11}$$

where $\boldsymbol{\pi}_j(\beta;\mathbf{Z}_j,\mathbf{x}_{\bullet j})=(\pi_{j1}(\beta;\mathbf{z}_{j1},x_{\cdot j1}),\cdots,\pi_{jn_j}(\beta;\mathbf{z}_{jn_j},x_{\cdot jn_j}))^T$ (without causing confusion, this is abbreviated as $\boldsymbol{\pi}_j(\beta;\cdot)=(\pi_{j1}(\beta;\cdot),\cdots,\pi_{jn_j}(\beta;\cdot))^T$);

$$\begin{aligned}&\frac{\partial^2}{\partial\beta\partial\beta^T}Q\Big(\theta,\hat{\theta}(r)\Big)\\&=-\sum_{j=1}^{J}\mathbf{Z}_j^T\mathrm{diag}\Big(E\Big\{\boldsymbol{\pi}_j(\beta;\cdot)(1-\boldsymbol{\pi}_j(\beta;\cdot))|\mathbf{Y}_j,\hat{\theta}(r)\Big\}\Big)\mathbf{Z}_j\end{aligned} \tag{12}$$

where $\mathrm{diag}(E\{\boldsymbol{\pi}_j(\beta;\cdot)(1-\boldsymbol{\pi}_j(\beta;\cdot))|\mathbf{Y}_j,\hat{\theta}(r)\})$ is the diagonal matrix generated by $E\{\pi_{ji}(\beta;\cdot)(1-\pi_{ji}(\beta;\cdot))|\mathbf{Y}_j,\hat{\theta}(r)\}$, $i=1,\cdots,n_j$;

$$\begin{aligned}&\frac{\partial}{\partial\sigma_g^2}Q\Big(\theta,\hat{\theta}(r)\Big)\\&=-\frac{1}{2}\sigma_g^{-2}\sum_{j=1}^{J}n_j+\frac{1}{2}\sigma_g^{-4}\sum_{j=1}^{J}E\Big\{\mathbf{x}_{gj}^T\mathbf{V}_{gj}^{-1}\mathbf{x}_{gj}|\mathbf{Y}_j,\hat{\theta}(r)\Big\},\end{aligned} \tag{13}$$

$g=1,\cdots,4$; and

$$\begin{aligned}&\frac{\partial^2}{\partial\sigma_g^2\partial\sigma_{g'}^2}Q\Big(\theta,\hat{\theta}(r)\Big)\\&=\begin{cases}\dfrac{1}{2}\sigma_g^{-4}\displaystyle\sum_{j=1}^{J}n_j-\sigma_g^{-6}\sum_{j=1}^{J}E\left\{\mathbf{x}_{gj}^T\mathbf{V}_{gj}^{-1}\mathbf{x}_{gj}|\mathbf{Y}_j,\hat{\theta}(r)\right\}, & \text{if } g=g';\\ 0, & \text{if } g\neq g'.\end{cases}\end{aligned} \tag{14}$$

Based on (11) and (12) one can solve the equation (7) by Newton-Raphson algorithm which starts from an initial value $\hat{\beta}(r,0) = \hat{\beta}(r)$ and calculates $\hat{\beta}(r,s+1)$ by taking this quantity to be

$$\begin{aligned}&\hat{\beta}(r,s)\\ &+\left(\sum_{j=1}^{J}\mathbf{Z}_j^T\operatorname{diag}\Big(E\Big\{\boldsymbol{\pi}_j\Big(\hat{\beta}(r,s);\cdot\Big)\Big(1-\boldsymbol{\pi}_j\Big(\hat{\beta}(r,s);\cdot\Big)\Big)|\mathbf{Y}_j,\hat{\theta}(r)\Big\}\Big)\mathbf{Z}_j\right)^{-1}\\ &\times\sum_{j=1}^{J}\mathbf{Z}_j^T\Big(\mathbf{Y}_j-E\Big\{\boldsymbol{\pi}_j\Big(\hat{\beta}(r,s);\cdot\Big)|\mathbf{Y}_j,\hat{\theta}(r)\Big\}\Big).\end{aligned}\tag{15}$$

When s is sufficiently large, $\hat{\beta}(r,s+1)$ will converge and the limit is $\hat{\beta}(r+1)$. Following from (13) and (14) the solution of (8) is

$$\hat{\sigma}_g^2(r+1)=\left(1/\sum_{j=1}^{J}n_j\right)\sum_{j=1}^{J}E\left\{\mathbf{x}_{gj}^T\mathbf{V}_{gj}^{-1}\mathbf{x}_{gj}|\mathbf{Y}_j,\hat{\theta}(r)\right\}.\tag{16}$$

In the actual computation, the conditional expections in (15)and (16) will be replaced by the corresponding sample statistics given in Appendix 3.

A.3. Generating $\{\mathbf{X}_{g1}^{(r)},\cdots,\mathbf{X}_{gK}^{(r)};g=1,\cdots,4\}$ by MH algorithm

To generate a sample of size K from (9) for given $\mathbf{Y}$ and $\theta=\hat{\theta}(r)$, one can use the following MH algorithm:

1. For each $g=1,\cdots,4$ and $j=1,\cdots,J$, generate an $\mathbf{x}_{gj0}^{(r)}$ from $\mathrm{MVN}(\mathbf{0},\hat{\sigma}_g^2(r)\mathbf{V}_{gj})$.
2. Repeat for $t=1,\cdots,K$ and for each $j=1,\cdots,J$. To get $(\mathbf{x}_{1jt}^{(r)},\cdots,\mathbf{x}_{4jt}^{(r)})$ first generate an $n_j\times 1$ vector $\tilde{\mathbf{x}}_g=(\tilde{x}_{g1},\cdots,\tilde{x}_{gn_j})^T$ from $\mathrm{MVN}(\mathbf{0},\hat{\sigma}_g^2(r)\mathbf{V}_{gj})$ for each g and a u from $\mathrm{Unif}(0,1)$. Then set $(\mathbf{x}_{1jt}^{(r)},\cdots,\mathbf{x}_{4jt}^{(r)})=(\tilde{\mathbf{x}}_1,\cdots,\tilde{\mathbf{x}}_4)$ if

$$u \le \alpha\left(\mathbf{x}^{(r)}_{\bullet j(t-1)}, \tilde{\mathbf{x}}_\bullet\right) = \min\left\{\prod_{i=1}^{n_j} \frac{e^{Y_{ji}(\tilde{x}_{\cdot i} - \mathbf{x}^{(r)}_{\cdot j(t-1),i})}\left(1 + e^{\mathbf{x}^{(r)}_{\cdot j(t-1),i} + \mathbf{z}^T_{ji}\hat{\beta}(r)}\right)}{1 + e^{\tilde{x}_{\cdot i} + \mathbf{z}^T_{ji}\hat{\beta}(r)}}, 1\right\};$$

otherwise set $(\mathbf{x}^{(r)}_{1jt}, \cdots, \mathbf{x}^{(r)}_{4jt}) = (\mathbf{x}^{(r)}_{1j(t-1)}, \cdots, \mathbf{x}^{(r)}_{4j(t-1)})$. Here $\mathbf{x}^{(r)}_{\bullet j(t-1)} = \sum_{g=1}^4 \mathbf{x}^{(r)}_{gj(t-1)} = (\mathbf{x}^{(r)}_{\cdot j(t-1),1}, \cdots, \mathbf{x}^{(r)}_{\cdot j(t-1),n_j})^T$ and $\tilde{\mathbf{x}}_\bullet = \sum_{g=1}^4 \tilde{\mathbf{x}}_g = (\tilde{x}_{\cdot 1}, \cdots, \tilde{x}_{\cdot n_j})^T$.

3. Return $\{\mathbf{x}^{(r)}_{gj1}, \cdots, \mathbf{x}^{(r)}_{gjK}; g = 1, \cdots, 4; j = 1, \cdots, J\}$ which is $\{\mathbf{X}^{(r)}_{g1}, \cdots, \mathbf{X}^{(r)}_{gK}; g = 1, \cdots, 4\}$.

In practice one may generate a sufficiently long sequence and take only the last K items as the sample $\{\mathbf{X}^{(r)}_{g1}, \cdots, \mathbf{X}^{(r)}_{gK}; g = 1, \cdots, 4\}$. When this sample is obtained, the conditional expectations in (15) and (16) will be approximated by the following sampling statistics:

$$\tilde{E}\left\{\pi_{ji}(\hat{\beta}(r,\ s);\cdot)|\mathbf{Y}_j, \hat{\theta}(r)\right\} = \frac{1}{K}\sum_{k=1}^K \pi_{ji}\left(\hat{\beta}(r,s); \mathbf{z}_{ji}, \mathbf{x}^{(r)}_{\cdot jk,i}\right)$$

$$= \frac{1}{K}\sum_{k=1}^K e^{\mathbf{x}^{(r)}_{\cdot jk,i} + \mathbf{z}^T_{ji}\hat{\beta}(r,s)}\left(1 + e^{\mathbf{x}^{(r)}_{\cdot jk,i} + \mathbf{z}^T_{ji}\hat{\beta}(r,s)}\right)^{-1},$$

(17) $$\tilde{E}\Big\{\pi_{ji}\Big(\hat{\beta}(r,\ s);\cdot\Big)\Big(1 - \pi_{ji}\Big(\hat{\beta}(r,s);\cdot\Big)\Big)|\mathbf{Y}_j, \hat{\theta}(r)\Big\}$$

$$= \frac{1}{K}\sum_{k=1}^K \pi_{ji}\Big(\hat{\beta}(r,s); \mathbf{z}_{ji}, \mathbf{x}^{(r)}_{\cdot jk,i}\Big)\Big(1 - \pi_{ji}\left(\hat{\beta}(r,s); \mathbf{z}_{ji}, \mathbf{x}^{(r)}_{\cdot jk,i}\right)\Big),$$

$$\tilde{E}\Big\{\mathbf{x}^T_{gj}\mathbf{V}^{-1}_{gj}\ \mathbf{x}_{gj}|\mathbf{Y}_j,\ \hat{\theta}(r)\Big\} = \frac{1}{K}\sum_{k=1}^K \mathbf{x}^{(r)T}_{gjk}\mathbf{V}^{-1}_{gj}\mathbf{x}^{(r)}_{gjk},$$

where $g = 1, \cdots, 4; i = 1, \cdots, n_j$ and $j = 1, \cdots, J$.

A.4. Estimating $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$

Equation (6) yields

$$\frac{\partial^2 Q(\theta,\theta')}{\partial\beta\partial\boldsymbol{\sigma}^{2T}} = \frac{\partial^2 Q(\theta,\theta')}{\partial\boldsymbol{\sigma}^2\partial\beta^T} = 0. \tag{18}$$

Now the first term $\partial^2 Q(\theta,\theta')/\partial\theta\partial\theta^T|_{\theta=\theta'=\hat{\theta}}$ in (10) can be estimated using (12), (14), (17) and (18). To estimate the second term in (10) write $\frac{\partial}{\partial\theta^T}\log L_j(\mathbf{Y}_j,\mathbf{x}_{1j},\cdots,\mathbf{x}_{4j};\theta) = (\mathbf{h}_{j1}(\beta,\mathbf{x}_{\bullet j})^T, \mathbf{h}_{j2}(\boldsymbol{\sigma}^2,\mathbf{x}_{1j},\cdots,\mathbf{x}_{4j})^T)$ where

$$\begin{aligned}\mathbf{h}_{j1}(\beta,\mathbf{x}_{\bullet j}) &= \frac{\partial}{\partial\beta}\log L_j(\mathbf{Y}_j,\mathbf{x}_{1j},\cdots,\mathbf{x}_{4j};\theta)\\ &= \mathbf{Z}_j^T\{\mathbf{Y}_j - \boldsymbol{\pi}_j(\beta;\mathbf{Z}_j,\mathbf{x}_{\bullet j})\} = \mathbf{Z}_j^T\left\{\mathbf{Y}_j - \frac{e^{\mathbf{x}_{\bullet j}+\mathbf{Z}_j\beta}}{1+e^{\mathbf{x}_{\bullet j}+\mathbf{Z}_j\beta}}\right\};\end{aligned} \tag{19}$$

and $\mathbf{h}_{j2}(\boldsymbol{\sigma}^2,\mathbf{x}_{1j},\cdots,\mathbf{x}_{4j})^T = \{\frac{\partial}{\partial\sigma_g^2}\log L_j(\mathbf{Y}_j,\mathbf{x}_{1j},\cdots,\mathbf{x}_{4j};\theta);\ \ g=1,\cdots,4\}$ with

$$\frac{\partial}{\partial\sigma_g^2}\log L_j\Big(\mathbf{Y}_j,\mathbf{x}_{1j},\cdots,\mathbf{x}_{4j};\theta\Big) = -\frac{1}{2}\sigma_g^{-2}n_j + \frac{1}{2}\sigma_g^{-4}\mathbf{x}_{gj}^T\mathbf{V}_{gj}^{-1}\mathbf{x}_{gj}. \tag{20}$$

When a sample $\{\mathbf{x}_{gj1}^{(r)},\cdots,\mathbf{x}_{gjK}^{(r)}; g=1,\cdots,4; j=1,\cdots,J\}$, which is actually a Markov chain, is generated by the MH algorithm from the conditional distribution (9) at $\theta = \hat{\theta}(r) = \hat{\theta}$, the sample variance of $\{(\mathbf{h}_{j1}(\hat{\beta}(r),\mathbf{x}_{\bullet jk}^{(r)})^T, \mathbf{h}_{j2}(\hat{\boldsymbol{\sigma}}^2(r),\mathbf{x}_{1jk}^{(r)},\cdots,\mathbf{x}_{4jk}^{(r)})^T);\ \ k=1,\cdots,K\}$ may be used to estimate the second term in (10). The estimation error of this sample variance may cause the estimate of $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ to be not positive definite. For this reason, we multiply the sample variance by a scalar λ, with $0<\lambda\le 1$, in estimating $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ to obtain a positive definite estimate.

Writing

$$\mathbf{J}_{\hat{\theta}}(\mathbf{Y}) = -\frac{\partial^2 \log L_y(\theta)}{\partial\theta\partial\theta^T}\bigg|_{\theta=\hat{\theta}}$$

$$\equiv \begin{pmatrix} -\dfrac{\partial^2 \log L_y(\theta)}{\partial\beta\partial\beta^T} & -\dfrac{\partial^2 \log L_y(\theta)}{\partial\beta\partial\boldsymbol{\sigma}^{2T}} \\ -\dfrac{\partial^2 \log L_y(\theta)}{\partial\boldsymbol{\sigma}^2\partial\beta^T} & -\dfrac{\partial^2 \log L_y(\theta)}{\partial\boldsymbol{\sigma}^2\partial\boldsymbol{\sigma}^{2T}} \end{pmatrix}\bigg|_{\theta=\hat{\theta}} \overset{\text{denote}}{=} \begin{pmatrix} \mathbf{J}_{\hat{\theta}}(\mathbf{Y})_{11} & \mathbf{J}_{\hat{\theta}}(\mathbf{Y})_{12} \\ \mathbf{J}_{\hat{\theta}}(\mathbf{Y})_{21} & \mathbf{J}_{\hat{\theta}}(\mathbf{Y})_{22} \end{pmatrix},$$

we can estimate the four partitions of $\mathbf{J}_{\hat{\theta}}(\mathbf{Y})$ as follows:

$$\tilde{\mathbf{J}}_{\hat{\theta}}(\mathbf{Y})_{11} = \sum_{j=1}^{J} \mathbf{Z}_j^T \operatorname{diag}\left\{ \frac{1}{K}\sum_{k=1}^{K} \frac{e^{\mathbf{x}_{\bullet jk}^{(r)} + \mathbf{Z}_j\hat{\beta}(r)}}{\left(1 + e^{\mathbf{x}_{\bullet jk}^{(r)} + \mathbf{Z}_j\hat{\beta}(r)}\right)^2} \right\} \mathbf{Z}_j$$
$$-\frac{\lambda}{K}\sum_{j=1}^{J}\sum_{k=1}^{K} \mathbf{h}_{j1}^*\left(\hat{\beta}(r), \mathbf{x}_{\bullet jk}^{(r)}\right)\mathbf{h}_{j1}^*\left(\hat{\beta}(r), \mathbf{x}_{\bullet jk}^{(r)}\right)^T,$$

$$\tilde{\mathbf{J}}_{\hat{\theta}}(\mathbf{Y})_{12} = \tilde{\mathbf{J}}_{\hat{\theta}}(\mathbf{Y})_{21}^T = -\frac{\lambda}{K}\sum_{j=1}^{J}\sum_{k=1}^{K} \mathbf{h}_{j1}^*\left(\hat{\beta}(r), \mathbf{x}_{\bullet jk}^{(r)}\right)\mathbf{h}_{j2}^*\left(\hat{\boldsymbol{\sigma}}^2(r), \mathbf{x}_{1jk}^{(r)}, \cdots, \mathbf{x}_{4jk}^{(r)}\right)^T,$$

$$\tilde{\mathbf{J}}_{\hat{\theta}}(\mathbf{Y})_{22} = \operatorname{diag}\left\{ -\frac{1}{2}\hat{\sigma}_g^{-4}\sum_{j=1}^{J} n_j + \hat{\sigma}_g^{-6}\frac{1}{K}\sum_{j=1}^{J}\sum_{k=1}^{K} \mathbf{x}_{gjk}^{(r)T}\mathbf{V}_{gj}^{-1}\mathbf{x}_{gjk}^{(r)};\ g = 1, \cdots, 4 \right\}$$
$$-\frac{\lambda}{K}\sum_{j=1}^{J}\sum_{k=1}^{K} \mathbf{h}_{j2}^*\left(\hat{\boldsymbol{\sigma}}^2(r), \mathbf{x}_{1jk}^{(r)}, \cdots, \mathbf{x}_{4jk}^{(r)}\right)\mathbf{h}_{j2}^*\left(\hat{\boldsymbol{\sigma}}^2(r), \mathbf{x}_{1jk}^{(r)}, \cdots, \mathbf{x}_{4jk}^{(r)}\right)^T,$$

where

$$\mathbf{h}^*_{j1}\Big(\hat{\beta}(r), \mathbf{x}^{(r)}_{\bullet jk}\Big) = \mathbf{h}_{j1}\Big(\hat{\beta}(r), \mathbf{x}^{(r)}_{\bullet jk}\Big) - K^{-1}\sum_{k'=1}^{K} \mathbf{h}_{j1}\Big(\hat{\beta}(r), \mathbf{x}^{(r)}_{\bullet jk'}\Big),$$

and

$$\mathbf{h}^*_{j2}\Big(\hat{\boldsymbol{\sigma}}^2(r), \mathbf{x}^{(r)}_{1jk}, \cdots, \mathbf{x}^{(r)}_{4jk}\Big) = \mathbf{h}_{j2}\Big(\hat{\boldsymbol{\sigma}}^2(r), \mathbf{x}^{(r)}_{1jk}, \cdots, \mathbf{x}^{(r)}_{4jk}\Big)$$

$$-K^{-1}\sum_{k'=1}^{K} \mathbf{h}_{j2}\Big(\hat{\boldsymbol{\sigma}}^2(r), \mathbf{x}^{(r)}_{1jk'}, \cdots, \mathbf{x}^{(r)}_{4jk\prime}\Big).$$

Denote λ^* be the smallest value in [0,1] at which $\tilde{\mathbf{J}}_{\hat{\theta}}(\mathbf{Y})$ is singular. If λ^* exists, then $\tilde{\mathbf{J}}_{\hat{\theta}}(\mathbf{Y})$ computed at $\lambda = 1$ must not be positive definite. So a valid $\tilde{\mathbf{J}}_{\hat{\theta}}(\mathbf{Y})$ has to be computed at some $\lambda < \lambda^*$, but not too close to λ^* to avoid near-singularity. If we define CN(λ) as the condition number (i.e, maximum absolute eigenvalue/minimum absolute eigenvalue) of $\tilde{\mathbf{J}}_{\hat{\theta}}(\mathbf{Y})$ computed at λ. Then typically that CN(λ) decreases as λ increases away from 0 and then increases as λ approaches λ^*. An optimal value of λ for computing $\tilde{\mathbf{J}}_{\hat{\theta}}(\mathbf{Y})$ would be the maximum $\lambda < \lambda^*$ satisfying CN(λ)=CN(0). Choosing this λ would ensure $\tilde{\mathbf{J}}_{\hat{\theta}}(\mathbf{Y})$ is positive definite, has the same degree of non-singularity as the complete information matrix and takes into account the maximum proportion of missing information.

A.5. The predictive posterior likelihood

The predictive posterior likelihood is

$$L_y(\theta|\mathbf{Y}) = \int\cdots\int \prod_{j=1}^{J} L_j(\mathbf{Y}_j|\mathbf{x}_{1j},\cdots,\mathbf{x}_{4j},\theta) L_j(\mathbf{x}_{1j},\cdots,\mathbf{x}_{4j}|\mathbf{Y}_j,\theta) \prod_{j=1}^{J} d\mathbf{x}_{1j}\cdots d\mathbf{x}_{4j}$$

$$= \int\cdots\int \prod_{j=1}^{J}\prod_{i=1}^{n_j} \frac{e^{Y_{ij}(x_{\cdot ji}+\mathbf{z}^T_{ji}\beta)}}{\Big(1+e^{x_{\cdot ji}+\mathbf{z}^T_{ji}\beta}\Big)} \cdot \prod_{j=1}^{J} L_j(\mathbf{x}_{1j},\cdots,\mathbf{x}_{4j}|\mathbf{Y}_j,\theta) \prod_{j=1}^{J} d\mathbf{x}_{1j}\cdots d\mathbf{x}_{4j}.$$

Under hypothesis H at $\theta = \hat{\theta}(H)$, $L_y(\theta|\mathbf{Y})$ is estimated by $\tilde{L}_y(\hat{\theta}(H)|\mathbf{Y})$ which is equal to

$$\frac{1}{K}\sum_{k=1}^{K}\prod_{j=1}^{J} L_j\left(\mathbf{Y}_j|\tilde{\mathbf{x}}_{1jk}(H),\cdots,\tilde{\mathbf{x}}_{4jk}(H),\hat{\theta}(H)\right)$$

$$=\frac{1}{K}\sum_{k=1}^{K}\prod_{j=1}^{J}\prod_{i=1}^{n_j}\frac{e^{Y_{ji}(\tilde{x}_{\cdot jk,i}(H)+\mathbf{z}_{ji}^T\hat{\beta}(H))}}{1+e^{\tilde{x}_{\cdot jk,i}(H)+\mathbf{z}_{ji}^T\hat{\beta}(H)}}$$

where $\tilde{\mathbf{x}}_{\bullet jk}(H) = (\tilde{x}_{\cdot jk,1}(H),\cdots,\tilde{x}_{\cdot jk,n_j}(H))^T = \sum_{g=1}^{4}\tilde{\mathbf{x}}_{gjk}(H)$; and $\{\tilde{\mathbf{x}}_{gjk}(H), j=1,\cdots,J; g=1,\cdots,4\}$ is generated from the conditional distribution $\prod_{j=1}^{J} L_j(\mathbf{x}_{1j},\cdots,\mathbf{x}_{4j}|\mathbf{Y}_j, \hat{\theta}(H))$ given by (9) $(k=1,\cdots,K)$.

Acknowledgement

This research was supported by NICHID grant # HD36071 and an Australian Research Council large grant. The FMRP results were obtained by Annette Taylor of Kimball Genetics, Inc (Denver,CO).

References

[1] J.H. Albert and S. Chib, *Bayesian analysis of binary and polychotomous response data*, Journal of the American Statistical Association, **88** (1993), 669–679.

[2] J. Albert and M. Ghosh, *Item respose modelling*, in: Generalized Linear Models A Bayesian Perspective Ed. Dey, D.K. Ghosh, S.K. Mallick, B.K. Marcel Dekker, New York (2000), 173–193.

[3] G.E. Bonney, *Regressive logistic models for familial disease and other binary traits*, Biometrics **42** (1986), 611–625.

[4] K.S. Chan and J. Ledholter, *Monte Carlo EM estimation for time series models involving counts*, J. Amer. Stat. Assoc. **90** (1995), 242–252.

[5] S. Chib, *Bayesian methods for correlated binary data*, in: Generalized Linear Models, A Bayesian Perspective, Ed. Dey, D.K., Ghosh, S.K., Mallick, B.K. Marcel Dekker, New York (2000), 113–131.

[6] S. Chib and E. Greenberg, *Understanding the Metropolis-Hastings algorithm*, American Statistician **49** (1995), 327–335.

[7] A.P. Dempster, N. Laird and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Royal Stat. Soc. B **39** (1977), 1–38.

[8] G.H. Fischer and I.W. Molenaar, *Rasch Models*, Foundations, Recent Developments, and Applications, Springer-Verlag, New York 1995.

[9] R.A. Fisher, *The correlation between relatives on the supposition of Mendelian inheritance*, Trans. of the Royal Society of Edinburgh **52** (1918), 399–433.

[10] P.E.B. FitzGerald and M.W. Knuiman, *Interpretation of regressive logistic regression coefficients in analyses of familial data*, Biometrics **54** (1998), 909–920.

[11] A. Gelman and D.B. Rubin, *Inference from iterative simulation using multiple sequences*, Statistical Science **7** (1992), 457–472.

[12] S.W. Guo and E.A. Thompson, *Monte Carlo estimation of mixed models for large complex pedigrees*, Biometrics **50** (1994), 417–432.

[13] J.L. Hopper, *Variance components for statistical genetics: applications in medical research to characteristics related to human diseases and health*, Statistical Methods in Medical Research **2** (1993), 199–223.

[14] J.L. Hopper and J.D. Mathews, *Extensions to multivariate normal models for pedigree analysis*, Ann. Hum. Genet. **46** (1982), 373–383.

[15] R.M. Huggins, *On robust analysis of pedigree data*, Aust J. Stat. **35** (1993), 43–57.

[16] K.L. Lange, J. Westlake and M.A. Spence, *Extensions to pedigree analysis, III, Variance components by the scoring method*, Ann. Hum. Genet. **39** (1976), 485–491.

[17] D.Z. Loesch, Q.M. Bui, J. Grigsby, E. Butler, J. Epstein, RM. Huggins and AK. Taylor, *Effect of the fragile X status categories and the FMRP levels on executive functioning in fragile X males and females*, Neuropsychology (2002) (in press).

[18] T.A. Louis, *Finding observed information using the EM algorithm*, J. Royal Stat. Soc. B **44** (1982), 226–233.

[19] X.L. Meng and D.B. Rubin, *Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm*, J. Amer. Stat. Assoc. **86** (1991), 899–909.

[20] G. Rasch, *Probabilistic Models for some Intelligence and Attainment Tests*, University of Chicago Press, Chicago 1980.

[21] D. Sinha, M.A. Tanner and W.J. Hall, *Maximization of the marginal likelihood of grouped survival data*, Biometrika **81** (1994), 53–60.

[22] S. Sommer and R.M. Huggins, *Variable selection using the Wald test and a robust C_p*, Applied Statistics **45** (1996), 15–29.

[23] M.A. Tanner, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd Ed., Springer, New York 1996.

[24] G.C.G. Wei and M.A. Tanner, *A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm*, J. Amer. Stat. Assoc. **85** (1990), 699–704.

Received 14 March 2004