2010

Multivariate Random Effect Models with complete and incomplete data

James O. Chipperfield
University of Wollongong

David G. Steel
University of Wollongong, dsteel@uow.edu.au

Recommended Citation
Chipperfield, James O. and Steel, David G., Multivariate Random Effect Models with complete and incomplete data, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 16-10, 2010, 29p.
https://ro.uow.edu.au/cssmwp/66

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

16-10

Multivariate Random Effect Models with complete and incomplete data

James O. Chipperfield and David G. Steel

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Multivariate Random Effect Models with complete and incomplete data

James O. Chipperfield and David G. Steel [1]

Abstract

This paper considers the problem of estimating fixed effects, random effects and variance components for the multi-variate random effects model with complete and incomplete data. It also considers making inference about the fixed and random effects, a problem which requires careful consideration of the choice of degrees of freedom to use in confidence intervals. This paper uses the EM algorithm to maximise the hierachical likelihood (HL). The HL estimates are often the same as the REML and Bayesian-justified estimates in Shah, Laird, and Schoenfeld (1997). A key benefit of the h-likelihood approach is its simplicity- it doesn't require integrating over the random effects or use of priors for its justification. Another benefit is that all inference can be made within a single framework. Extensive simulations show: that the h-likelihood approach is significantly more accurate than the well-known ANOVA approach; the h-likelihood approach often recovers a lot of the information lost through missing data; the h-likelihood approach has good coverage properties for fixed and random effects that are estimated using small samples.

Key words: maximum likelihood, hierachical likelihood, EM algorithm, missing data

1 Introduction

The multivariate random effects model (MVEM) is a common way to analyse group-level and individual or observation-level effects. For example, the variance components of the MVEM give an insight into the relative importance of institution and individual on examination performance (e.g. Yang, Goldstein, Browne,

[1] James Chipperfield is Assistant Director, Methodology and Data Management Division, Australian Bureau of Statistics. David Steel is Director, Centre for Statistical and Survey Methodology, University of Wollongong. The authors would like to thank the Australian Bureau of Statistics for supporting this research.

& Woodhouse, 2002). While the fixed effects are often of primary interest Lee, Nelder, and Pawitan (2006) (pp.148) notes, there are an increasing number of applications in which the random effects themselves are of interest. Some examples include ranking school performance and improvement in breeding programs. The MVEM distinuguishes itself from the more commonly used 1-way or 2-way random effects models by the fact that the MVEM allows the variance components to be unstructred. It is also this very reason that distinguishes the MVEM from generalised linear mixed models (see McCulloch & Searle, 2001).

With complete or missing data, Maximum Likelihood (ML) treatment of the MVEM (see Shah et al., 1997) focuses on making inferences about the fixed effects: the random effects are treated as nuisance parameters to be integrated out of the likelihood. Estimates of random effects and their measures of accuracy can then be obtained as a Best Linear Unbiased Predictor (BLUP) (see McCulloch & Searle, 2001, pp 170). A much more convenient approach of making inference for the present problem is to use the Hierachical Likelihood (HL), as it provides a single framework to making inference about both the fixed and random effects. As Lee et al. (2006) (pp. 133) notes, with the HL framework *standard error estimates are easily obtained* whereas for the ML approach *other methods are necessary to obtain them.*

The h-likelihood (HL) was initially proposed by Lee and Nelder J. (1996), and expanded upon by Lee et al. (2006), as a more general and tractable framework than the ML framework, particularly for mixed models. The HL approach to the missing data problems for generalised linear mixed models were subsequently explored by Yun, Lee, and Kenward (2007) . The HL approach in Yun et al. (2007) characterises the missing data *and* the random effects to be parameters to be estimated, while using the profile likelihood to make a REML-type adjustment to account for the number of parameters in estimates of the variance components. They do not consider the MVEM, which is the focus of this paper.

For the MVEM, we show that the HL estimates have the same form as the REML estimates of the fixed effects and the between-group variance as well as the BLUP of the random effects. When accounting for missing data within the HL framework, an EM algorithm is used to replace the missing data with their expection conditional on the observed data and the loss of accuracy is accounted for using a method typically applied in the context of ML; this approach is interesting in that it combines features of both ML and HL, whereas they are often seen as alternatives in the literature. In addition, this paper shows that inferences about the fixed and random effects using the HL approach (and so the REML estimates of the fixed effects) are theoretically valid if the probability that an observation is missing only depends upon the observation's group-level effects (e.g. if the probability depends on the observation's non-missing values, inferences are theoretically invalid).

This paper also evaluates the accuracy and coverage of estimates of fixed and random effects; this paper pays particular attention to the degrees of freedom used to construct confidence intervals, which is particularly important in small samples.

Sections 2 and 3 consider the multivariate random effects model for the complete and incomplete data cases, respectively. Section 4 evaluates the HL approach in a simulation study. Section 5 makes some concluding remarks.

2 Multivariate Random Effects Model with Complete Data

2.1 Fixed and Random Effects

Define $\mathbf{y}_{ij} = (y_{ij1}, \ldots y_{ijk}, \ldots y_{ijK})'$ to be the complete data about K variables from observation i in group j, where $k = 1 \ldots, K$, $i = 1, \ldots, n_j$, $j = 1, \ldots, J$ and $n = \Sigma_j n_j$. Let $\mathbf{y}^* = (\mathbf{y}'_{11}, \mathbf{y}'_{21}, \ldots, \mathbf{y}'_{ij}, \ldots, \mathbf{y}'_{n_J J})'$ be the M column vector obtained by stacking the $\mathbf{y}_{ij}$ s. Here we denote the complete data by d_c. Throughout this paper we assume the sampling process that lead to $\mathbf{y}^*$ can be ignored (see Chambers and Skinner (2003)). Assume the data follow the model

$$\mathbf{y}^* = \mathbf{q}\boldsymbol{\mu} + \mathbf{Z}^*\mathbf{b} + \mathbf{e}^* \tag{1}$$

where $\mathbf{q}$ is an MxK design matrix, $\boldsymbol{\mu}$ is the K column vector of means with element μ_k (allowing for an unequal number of variables, say K_i, per observation is straight-forward). Define $\mathbf{b}_j = (b_{j1}, b_{j2}, \ldots b_{jk}, \ldots, b_{jK})'$ to be a vector of random effects for group j and therefore that $\mathbf{b} = (\mathbf{b}'_1, \mathbf{b}'_2, \ldots \mathbf{b}'_j, \ldots, \mathbf{b}'_J)'$ is a Tx1 column vector, where $T = JK$. The design matrix for the random effects is given by $\mathbf{Z}^*$, an MxT matrix with element (m, t) equal to 1 if the mth element of $\mathbf{y}^*$ is subject to random effect j and zero otherwise, and $t = 1, \ldots, T$. In terms of a randomised trial, for example, $\mathbf{q}$ could indicate different experimental conditions and $\mathbf{b}$ could indicate measurement errors associated with different clinics used in the trial. The vector of residuals is $\mathbf{e}^* = (\mathbf{e}'_{11}, \mathbf{e}'_{21}, \ldots, \mathbf{e}'_{ij}, \ldots, \mathbf{e}'_{n_J J})'$, where $\mathbf{e}_{ij} = (e_{ij1}, e_{ij2}, \ldots e_{ijK})'$ and $e_{ijk} = y_{ijk} - \mu_k - b_{jk}$.

We assume the random effects, $\mathbf{b}_j$ to be $N(\mathbf{0}_K, \mathbf{\Sigma}_b)$, where $\mathbf{0}_K$ is a K column vector of zeros and we denote $\mathbf{\Sigma}_b = (\sigma_{b,kk'})$. Given $\mathbf{b}_j$ s are assumed independent it follows that $\mathbf{b}$ is $N(\mathbf{0}_T, \mathbf{V}_b)$ where $\mathbf{V}_b = \mathbf{I}_J \otimes \mathbf{\Sigma}_b$. We also assume the residuals, $\mathbf{e}_{ij}$, are $N(\mathbf{0}_K, \mathbf{\Sigma}_w)$ and we denote $\mathbf{\Sigma}_w = (\sigma_{w,kk'})$. Given the $\mathbf{e}_{ij}$ s are independent $\mathbf{e}^*$ is $N(\mathbf{0}_M, \mathbf{V}_w)$ where $\mathbf{V}_w = \mathbf{I}_n \otimes \mathbf{\Sigma}_w$. It then follows that $V = Var(\mathbf{y}^*)$ has block-wise elements

$$\begin{array}{lllll} Cov(\mathbf{y}_{ij}, \mathbf{y}_{i'j'}) & = \mathbf{\Sigma}_w + \mathbf{\Sigma}_b & if & i = i' \;\; and \;\; j = j' \\ & = \mathbf{\Sigma}_b & if & i \neq i' \;\; and \;\; j = j' \\ & = \mathbf{0}_{KK} & if & i \neq i' \;\; and \;\; j \neq j' \end{array} \tag{2}$$

where $\mathbf{0}_{KK}$ is a KxK matrix of zeros. Other variance structures for (2) can be considered, say by replacing $\mathbf{0}_{KK}$ by a parameter of some sort (see Shah et al., 1997). The joint distribution of $\mathbf{y}^*$ and $\mathbf{b}$ (see Lee & Nelder J., 1996 and Robinson, 1991) can be factorised as

$$p(\mathbf{y}^* \mid \mathbf{b}; \mathbf{V}_w)p(\mathbf{b}; \mathbf{V}_b) \tag{3}$$

with HL

$$h_c = -1/2\mathbf{b}'\mathbf{V}_b^{-1}\mathbf{b} - 1/2log \mid \mathbf{V}_b \mid -1/2(\mathbf{y}^* - \mathbf{q}\boldsymbol{\mu} - \mathbf{Z}^*\mathbf{b})'\mathbf{V}_w^{-1}(\mathbf{y}^* - \mathbf{q}\boldsymbol{\mu} - \mathbf{Z}^*\mathbf{b}) - 1/2log \mid \mathbf{V}_w \mid \tag{4}$$

The corresponding score equation for $\boldsymbol{\Gamma} = (\boldsymbol{\mu}, \mathbf{b})$, obtained by differentiating (4) with respect to $\boldsymbol{\Gamma}$, is

$$Sc(\boldsymbol{\Gamma}; d_c) = \begin{pmatrix} \mathbf{q}'\mathbf{V}_w^{-1}(\mathbf{y}^* - \mathbf{q}\boldsymbol{\mu} - \mathbf{Z}^*\mathbf{b}) \\ \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}(\mathbf{y}^* - \mathbf{q}\boldsymbol{\mu}) - \mathbf{V}_b^{-1}\mathbf{b} - \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{Z}^*\mathbf{b} \end{pmatrix} \tag{5}$$

The HL estimate of $\boldsymbol{\Gamma}$, denoted by $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\mu}}', \hat{\mathbf{b}}')'$, is obtained by solving $Sc(\boldsymbol{\Gamma}; d_c) = 0$. The solution is

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \left[\mathbf{q}'(\mathbf{V}_w + \mathbf{Z}^*\mathbf{V}_b\mathbf{Z}^{*\prime})^{-1}\mathbf{q}\right]^{-1} \mathbf{q}'(\mathbf{V}_w + \mathbf{Z}^*\mathbf{V}_b\mathbf{Z}^{*\prime})^{-1}\mathbf{y}^* \\ \hat{\mathbf{b}} &= \left(\mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{Z}^* + \mathbf{V}_b^{-1}\right)^{-1} \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}(\mathbf{y}^* - \mathbf{q}\hat{\boldsymbol{\mu}}) \end{aligned} \tag{6}$$

The expected information referred to as *hinfo*, matrix of $\boldsymbol{\Gamma}$ using d_c, obtained by twice differentiating (4) with respect to $\boldsymbol{\Gamma}$, is given by

$$\mathbf{H}_c = hinfo(\boldsymbol{\Gamma}; d_c) = \begin{pmatrix} \mathbf{q}'\mathbf{V}_w^{-1}\mathbf{q} & \mathbf{q}'\mathbf{V}_w^{-1}\mathbf{Z}^* \\ \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{q} & \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{Z}^* + \mathbf{V}_b^{-1} \end{pmatrix} \tag{7}$$

It is easy to show, essentially using the same argument in Lee et al. (2006) (see pp. 157-8) that $\mathbf{H}_c^{-1}$ in (7) gives a valid estimate of the variance of $\boldsymbol{\Gamma}$. The estimators in (6) are the same as in Shah et al. (1997).

The next section discusses estimating $\mathbf{V}_w$ and $\mathbf{V}_b$.

2.2 Dispersion Parameters

Let $\mathbf{\Sigma} = (\mathbf{\Sigma}_w, \mathbf{\Sigma}_b)$. For estimation of $\mathbf{\Sigma}$, consider the adjusted likelihood

$$h_{A,c} = h_c + log\{det(\mathbf{H}_c^{-1})\}. \tag{8}$$

The second term in (8) is essentially a degrees of freedom adjustment for the estimation of $\mathbf{\Sigma}$ that accounts for the fact that $\mathbf{\Gamma}$, which includes the fixes and random effects, are parameters that must be estimated. The adjusted profile likelihood (Patterson & Thompson, 1971, Cox & Reid, 1987 and Lee & Nelder J., 1996) is

$$h_{P,c} = h_{A,c} \mid_{\mathbf{\Gamma}=\hat{\mathbf{\Gamma}}} \tag{9}$$

Patterson and Thompson (1971) shows that use of (9) requires that $\hat{\Sigma}$ and $\hat{\mathbf{\Gamma}}$ are orthogonal. This requirement is met by noting that $\partial^2 h_{P,c}/\Big(\partial\mathbf{\Gamma}\partial\mathbf{\Sigma}\Big) = 0$.

Let $\mathbf{\Sigma}_w$ have elements ϕ_r and $\mathbf{\Sigma}_b$ have elements α_s. The score equation for ϕ_r is

$$\begin{aligned} Sc(\phi_r; d_c) &= \partial h_{P,c}/\partial\phi_r \\ &= -\{(\mathbf{y}^* - \mathbf{q}\hat{\boldsymbol{\mu}} - \mathbf{Z}^*\hat{\mathbf{b}})'\mathbf{V}_{w(r)}^{-1}(\mathbf{y}^* - \mathbf{q}\hat{\boldsymbol{\mu}} - \mathbf{Z}^*\hat{\mathbf{b}})\} - tr(\mathbf{H}_c^{-1}\mathbf{H}_{c(r)}) - tr[\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}] \end{aligned} \tag{10}$$

where $\mathbf{V}_{w(r)} = \partial \mathbf{V}_w/\partial \phi_r$, $\mathbf{V}^{-1}_{w(r)} = \partial \mathbf{V}^{-1}_w/\partial \phi_r = \mathbf{V}^{-1}_w \mathbf{V}_{w(r)} \mathbf{V}^{-1}_w$,

$$\mathbf{H}_{c(r)} = \partial \mathbf{H}_c/\partial \phi_r = \begin{pmatrix} \mathbf{q}'\mathbf{V}^{-1}_{w(r)}\mathbf{q} & \mathbf{q}'\mathbf{V}^{-1}_{w(r)}\mathbf{Z}^* \\ \mathbf{Z}^{*}{}'\mathbf{V}^{-1}_{w(r)}\mathbf{q} & \mathbf{Z}^{*}{}'\mathbf{V}^{-1}_{w(r)}\mathbf{Z}^* \end{pmatrix}.$$

The score equation for α_s is

$$\begin{aligned} Sc(\alpha_s; d_c) &= \partial h_{P,c}/\partial \alpha_s \\ &= -tr\{\hat{\mathbf{b}}'\mathbf{V}^{-1}_{b(s)}\hat{\mathbf{b}} - \mathbf{K}_c \mathbf{V}^{-1}_{b(s)} - tr[\mathbf{V}^{-1}_b \mathbf{V}_{b(s)}\} \end{aligned} \tag{11}$$

where $\mathbf{K}_c$ is submatrix of $\mathbf{H}^{-1}_c$ corresponding to $\hat{\mathbf{b}}$, $\mathbf{V}_{b(s)} = \partial \mathbf{V}_b/\partial \alpha_s$ and $\mathbf{V}^{-1}_{b(s)} = \partial \mathbf{V}^{-1}_b/\partial \alpha_s = -\mathbf{V}^{-1}_b \mathbf{V}_{b(s)} \mathbf{V}^{-1}_b$.

We now introduce notation. Define $\{_r m_j\}^J_{j=1}$ to be a J-length vector with elements m_j; replacing the subscript r with c or d simialrly defines the elements of a column vector or a diagonal matrix respectively. The HL estimators of $\mathbf{\Sigma}_b$ and $\mathbf{\Sigma}_w$ are the solutions for $\mathbf{\Sigma}_b$ and $\mathbf{\Sigma}_w$ after equating (10) and (11) to zero for all r and s, respectively. It is shown in A.1 and A.2 that the HL estimators of $\mathbf{\Sigma}_b$ and $\mathbf{\Sigma}_w$, respectively, are

$$\hat{\mathbf{\Sigma}}_b = \Sigma_j[\hat{\mathbf{b}}'_j\hat{\mathbf{b}}_j + \mathbf{K}_{c,j}]/J \tag{12}$$

where $\mathbf{K}_{c,j} = Var[\hat{\mathbf{b}}_j \mid d_c]$ is the j th diagonal block of $\mathbf{K}_c$ corresponding to the random effects in group j and

$$\hat{\mathbf{\Sigma}}_w = (n\mathbf{I}_K - \Sigma^{J+1}_j \hat{\mathbf{g}}_j)^{-1}\Sigma_{ij}\hat{\mathbf{e}}'_{ij}\hat{\mathbf{e}}_{ij} \tag{13}$$

respectively, where $\hat{\mathbf{g}} = \hat{\mathbf{B}}^{-1}\mathbf{A}$ with j th diagonal block denoted $\hat{\mathbf{g}}_j$ of dimension KxK,

$$\mathbf{A} = \begin{pmatrix} n\mathbf{I}_K & \{_r n_j \mathbf{I}_K\}_{j=1}^J \\ \\ \{_c n_j \mathbf{I}_K\}_{j=1}^J & \{_d n_j \mathbf{I}_K\}_{j=1}^J \end{pmatrix}, \hat{\mathbf{B}} = \begin{pmatrix} n\mathbf{I}_K & \{_r n_j \mathbf{I}_K\}_{j=1}^J \\ \\ \{_c n_j \mathbf{I}_K\}_{j=1}^J & \{_d n_j \mathbf{I}_K + \hat{\mathbf{\Sigma}}_w \hat{\mathbf{\Sigma}}_b^{-1}\}_{j=1}^J \end{pmatrix}$$

and $\hat{e}_{ijk} = y_{ijk} - \hat{\mu}_k - \hat{b}_{jk}$. Since $\hat{\mathbf{\Sigma}}_b$ and $\hat{\mathbf{\Sigma}}_w$ are clearly functions of themselves, estimates must be calculated by iteration (see section 2.3). As n_j increases, and $\hat{\mathbf{\Sigma}}_w \hat{\mathbf{\Sigma}}_b^{-1}$ makes less of a contribution to $\hat{\mathbf{g}}$, then $\Sigma_j \hat{\mathbf{g}}_j \approx J + 1$. The estimate $\hat{\mathbf{\Sigma}}_b$ is the same as in Shah et al. (1997).

An alternative method for estimating $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_b$ is ANOVA (see Chambers & Skinner, 2003, chapter 20). The ANOVA estimators in the balanced case ($n_j = \bar{n}$ for all j) are

$$\begin{aligned} \hat{\mathbf{\Sigma}}_w^{AN} &= (n - J)^{-1} \Sigma_{ij} (\mathbf{y}_{ij} - \mathbf{m}_j)'(\mathbf{y}_{ij} - \mathbf{m}_j) \\ \hat{\mathbf{\Sigma}}_b^{AN} &= \bar{n}^{-1} (\mathbf{S} - \hat{\mathbf{\Sigma}}_w^{AN}) \end{aligned} \quad (14)$$

where $\mathbf{m}_j = \bar{n}^{-1} \Sigma_{i=1}^{\bar{n}} \mathbf{y}_{ij}$, $\mathbf{S} = (J - 1)^{-1} \Sigma_{j=1}^{J} \bar{n} (\mathbf{m}_j - \mathbf{m})'(\mathbf{m}_j - \mathbf{m})$, and $\mathbf{m} = n^{-1} \Sigma_{ij}^{n} \mathbf{y}_{ij}$. We show in simulations that the HL approach is clearly preferred to the ANOVA approach with complete data.

2.3 Estimation

The estimation procedure based on d_c involves:

1. Initialising $\hat{\mathbf{\Sigma}}$, denoted by $\hat{\mathbf{\Sigma}}^{(0)}$

2. Calculating $\hat{\mathbf{\Gamma}}^{(t)}$ from (6) using $\hat{\mathbf{\Sigma}}^{(t-1)}$

3. Calculating $\hat{\mathbf{\Sigma}}^{(t)}$ from (12) and (13) using $\hat{\mathbf{\Gamma}}^{(t)}$

4. Repeating 2 - 3 until convergence.

5. Calculating $\mathbf{H}_c$.

3 Multivariate Random Effects Model with Incomplete Data

Define a Kxn matrix $\mathbf{M}$ with elements indicating whether the kth variable is missing for the i th observation in group j. Let $\mathbf{M}$ be some function of a parameter $\boldsymbol{\zeta}$. We define the data to be Missing at Random Within Groups (MARWG) (also called the selection model of Diggle & Kenward, 1994) if

$$p(\mathbf{y}^*, \mathbf{b}, \mathbf{M}; \mathbf{V}_b, \mathbf{V}_w, \boldsymbol{\zeta}) = p(\mathbf{y}^* \mid \mathbf{b}; \mathbf{V}_w) p(\mathbf{b}; \mathbf{V}_b) p(\mathbf{M} \mid \mathbf{y}^*_{obs}, \mathbf{b}; \boldsymbol{\zeta})$$

where $\mathbf{y}^*_{obs}$ are the observed elements of $\mathbf{y}^*$. This means the probability that an observation's variable is missing depends upon its observed variables and its group effects. The data are Missing Completely at Random Within Groups (MCARWG) (a special case of the selection model)

$$p(\mathbf{y}^*, \mathbf{b}, \mathbf{M}; \mathbf{V}_w, \mathbf{V}_b, \boldsymbol{\zeta}) = p(\mathbf{y}^* \mid \mathbf{b}; \mathbf{V}_w) p(\mathbf{b}; \mathbf{V}_b) p(\mathbf{M} \mid \mathbf{b}; \boldsymbol{\zeta}).$$

This means the probability that an observation's variable is missing depends on its group effects. The data are Missing Completely at Random (MCAR) (see Rubin & Little, 1987) if

$$p(\mathbf{y}^*, \mathbf{b}, \mathbf{M}; \mathbf{V}_w, \mathbf{V}_b, \boldsymbol{\zeta}) = p(\mathbf{y}^* \mid \mathbf{b}; \mathbf{V}_w) p(\mathbf{b}; \mathbf{V}_b) p(\mathbf{M}; \boldsymbol{\zeta}).$$

Under MCAR analysis using only the complete cases (i.e. observations for which there are no missing variables) leads to unbiased estimation and inference.

If the data are MCARWG or MARWG, using only complete cases leads to biased estimation and inference. The MCAR, MCARWG and MARWG factorisations mean we can ignore the factors $p(\mathbf{M};\boldsymbol{\zeta})$, $p(\mathbf{M} \mid \mathbf{b};\boldsymbol{\zeta})$ and $p(\mathbf{M} \mid \mathbf{y}^*_{obs},\mathbf{b};\boldsymbol{\zeta})$ respectively and we are essentially still maximising (3).

3.1 Fixed and Random Effects

Consider an observed sample set, d_o, which arises from subjecting d_c to a missing data mechanism. One key result of Breckling, Chambers, Dorfman, Tam, and Welsh (1994) is that the ML estimate of $\boldsymbol{\theta}$ based on d_o is obtained by solving

$$E_{d_c|d_o}[Sc\,(\boldsymbol{\theta};d_c) \mid d_o] = 0 \tag{15}$$

where $E_{d_c|d_o}$ is the expectation with respect to the complete data d_c conditional on the incomplete data d_o and $Sc\,(\boldsymbol{\theta};d_c)$ is the score function for $\boldsymbol{\theta}$ based on d_c. Here we assume the distribution of the data is defined by (1). In fact we only need assume that the distribution of the missing data given the observed data follows a normal distribution (see below). The result (15) for the likelihood is applied here for the HL, in line with assersion of Lee and Nelder J. (1996) that the *the h-likelihood is the fundamental likelihood.*

It follows that the HL estimate of $\mathbf{\Gamma}$ based on d_o, denoted by $\tilde{\mathbf{\Gamma}}$, is given by (6) except that y_{ijk} is replaced by $\tilde{y}_{ijk} = E_{d_c|d_o}(y_{ijk} \mid d_o)$, where

$$
\begin{aligned}
\tilde{y}_{ijk} &= y_{ijk} && if\ \ y_{ijk}\ \ is\ \ observed \\
&= E_{d_c|d_o}(\mu_k + b_{jk} + e_{ijk} \mid d_o) && otherwise \\
&= \mu_k + b_{jk} + E_{d_c|d_o}(e_{ijk} \mid d_o) \\
&= \mu_k + b_{jk} + \mathbf{e}_{obs,ij}\boldsymbol{\beta}^w_{ki},
\end{aligned} \tag{16}
$$

where $E_{d_c|d_o}(e_{ijk} \mid d_o) = \mathbf{e}_{obs,ij}\boldsymbol{\beta}^w_{ki}$ follows from the multivariate assumption for the residuals in (1), $\boldsymbol{\beta}^w_{ki} = \boldsymbol{\Sigma}^{-1}_{w\cdot ij}\boldsymbol{\Sigma}_{w\cdot ij}(k)$, $\boldsymbol{\Sigma}_{w\cdot ij}$ is $\boldsymbol{\Sigma}_w$ after removing the rows and columns corresponding to the missing data items for observation i in group j, $\boldsymbol{\Sigma}_{w\cdot ij}(k)$ is the k th column vector of $\boldsymbol{\Sigma}_{w\cdot ij}$, and $\mathbf{e}_{obs,ij}$ is subset of $\mathbf{e}_{ij}$ corresponding to the observed elements of $\mathbf{y}_{ij}$.

Another key result of Breckling et al. (1994) is that the observed information for the ML estimate of a parameter $\boldsymbol{\theta}$ under d_o, and adopted here for the hierachical estimate of $\boldsymbol{\theta}$, is

$$
hinfo_o(\boldsymbol{\theta}; d_o) = hinfo_c(\boldsymbol{\theta}; d_c) \mid d_o - var\left[Sc\left(\boldsymbol{\theta}; d_c\right) \mid d_o\right] \tag{17}
$$

The second term in (17) represents the loss of information due to observing d_o rather than d_c. Using (17), as well as (5) and (7), the observed information of $\tilde{\boldsymbol{\Gamma}}$, denoted by $\mathbf{H}_o = hinfo_o(\tilde{\boldsymbol{\Gamma}}; d_o)$, is

$$
\begin{aligned}
\mathbf{H}_o &= \mathbf{H}_c - var\left[Sc\left(\boldsymbol{\Gamma}; d_c\right) \mid d_o\right] \\
&= \begin{pmatrix} \mathbf{q}'\mathbf{V}_w^{-1}\mathbf{q} - \mathbf{q}'\mathbf{V}_w^{-1}\mathbf{V}^o\mathbf{V}_w^{-1}\mathbf{q} & \mathbf{q}'\mathbf{V}_w^{-1}\mathbf{Z}^* - \mathbf{q}'\mathbf{V}_w^{-1}\mathbf{V}^o\mathbf{V}_w^{-1}\mathbf{Z}^* \\ \mathbf{Z}^{*\,\prime}\mathbf{V}_w^{-1}\mathbf{q} - \mathbf{Z}^{*\,\prime}\mathbf{V}_w^{-1}\mathbf{V}^o\mathbf{V}_w^{-1}\mathbf{q} & \mathbf{Z}^{*\,\prime}\mathbf{V}_w^{-1}\mathbf{Z}^* + \mathbf{V}_b^{-1} - \mathbf{Z}^{*\,\prime}\mathbf{V}_w^{-1}\mathbf{V}^o\mathbf{V}_w^{-1}\mathbf{Z}^* \end{pmatrix}
\end{aligned} \tag{18}
$$

$$= \begin{pmatrix} \mathbf{q}'\mathbf{V}_w^{-1}(\mathbf{I}_M - \mathbf{V}^o\mathbf{V}_w^{-1})\mathbf{q} & \mathbf{q}'\mathbf{V}_w^{-1}(\mathbf{I}_M - \mathbf{V}^o\mathbf{V}_w^{-1})\mathbf{Z}^* \\ \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}(\mathbf{I}_M - \mathbf{V}^o\mathbf{V}_w^{-1})\mathbf{q} & \mathbf{V}_b^{-1} + \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}(\mathbf{I}_M - \mathbf{V}^o\mathbf{V}_w^{-1})\mathbf{Z}^* \end{pmatrix}$$

where $\mathbf{I}_M$ is the identity matrix of order M, $\mathbf{V}^o = Var(\mathbf{y}^* \mid d_o) = \left\{_d \left\{_d \mathbf{\Sigma}_{w \cdot ij} \right\}_{i=1}^{n_j} \right\}_{j=1}^{J}$ and $\mathbf{\Sigma}_{w \cdot ij}$ is obtained by sweeping the observed variables for observation i in group j from $\mathbf{\Sigma}_w$, since

$$\begin{aligned} Cov(y_{ijk}, y_{i'j'k'} \mid d_o) &= Cov(\mu_k + b_{jk} + e_{ijk}, \\ &\qquad \mu_{k'} + b_{j'k'} + e_{i'j'k'} \mid d_o) \\ &= Cov(e_{ijk}, e_{i'j'k'} \mid \mathbf{e}_{obs}) \\ &= \sigma^2_{wkk' \cdot ij} \quad if\ i = i'\ and\ j = j' \\ &= 0 \qquad otherwise \end{aligned} \tag{19}$$

For example, if y_{ijk} or $y_{ijk'}$ is observed then $\sigma^2_{wkk' \cdot ij} = 0$. The negative terms in (18) reflect the information loss due to the missing data. The term $\mathbf{H}_o$ in (18) also appears in Shah et al. (1997), though in a slightly different form.

Above the missing data are treated as unobserved variables, as is the case with the ML approach, not as parameters to be estimated. This is why only the Hessian matrix for the fixed and random effects appear in the second term of the profile likelihood of (8). As the hierachical approach in Yun et al. (2007) treats missing observations as parameters to be estimated, the missing data also appear in the Hessian matrix. For the MVEM this is really only a minor point of difference.

3.2 Dispersion Parameters

The HL estimates of the dispersion parameters from the observed data, denoted by $\tilde{\boldsymbol{\Sigma}} = (\tilde{\boldsymbol{\Sigma}}_w, \tilde{\boldsymbol{\Sigma}}_b)$, are constructed so that $E_{d_c|d_o}[\hat{\boldsymbol{\Sigma}}] = \tilde{\boldsymbol{\Sigma}}$.

The HL estimate of $\boldsymbol{\Sigma}_w$ under d_o is then

$$\tilde{\boldsymbol{\Sigma}}_w = (n\mathbf{I}_K - \Sigma_j \tilde{\mathbf{g}}_j)^{-1}[\Sigma_{ij}\tilde{\mathbf{e}}_{ij}\tilde{\mathbf{e}}'_{ij} + \boldsymbol{\Sigma}_{w\cdot ij}] \tag{20}$$

where $\tilde{\mathbf{e}}_{ij}$ is a vector with kth element $\tilde{y}_{ijk} - \tilde{\mu}_k - \tilde{b}_{jk}$, $\tilde{\mathbf{b}} = (\tilde{b}_{jk})$ has the same form as $\hat{\mathbf{b}}$ except that y_{ijk} is replaced by $\tilde{y}_{ijk}$ and $\tilde{\mathbf{g}}_j$ has the same form as $\hat{\mathbf{g}}_j$ except that $\hat{\boldsymbol{\Sigma}}_w$ and $\hat{\boldsymbol{\Sigma}}_b$ are replaced with $\tilde{\boldsymbol{\Sigma}}_w$ and $\tilde{\boldsymbol{\Sigma}}_b$ ($\tilde{\boldsymbol{\Sigma}}_b$ is defined below). This is justified since, from (13), $E_{d_c|d_o}[\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}'_{ij}] = \tilde{\mathbf{e}}_{ij}\tilde{\mathbf{e}}'_{ij} + \tilde{\boldsymbol{\Sigma}}_{w\cdot ij}$ and $E_{d_c|d_o}[\hat{\mathbf{g}}_j] = \tilde{\mathbf{g}}_j$.

Similarly, an estimate of $\boldsymbol{\Sigma}_b$ under d_o is

$$\tilde{\boldsymbol{\Sigma}}_b = \Sigma_j[\tilde{\mathbf{b}}'_j\tilde{\mathbf{b}}_j + \tilde{\mathbf{K}}_{o,j}]/J \tag{21}$$

where $\tilde{\mathbf{K}}_{o,j}$ is an estimate of $\mathbf{K}_{o,j}$, $\mathbf{K}_{o,j} = Var[\tilde{\mathbf{b}}_j \mid d_o]$ is the jth diagonal block of dimension KxK of $\mathbf{K}_o$, $\mathbf{K}_o$ is submatrix of $\mathbf{H}_o^{-1}$ corresponding to $\mathbf{b}$, and $\tilde{\mathbf{K}}_{c,j}$ has the same form as $\mathbf{K}_{c,j}$ except that $\boldsymbol{\Sigma}$ is replaced with $\tilde{\boldsymbol{\Sigma}}$. This is justified since, from (12), $E_{d_c|d_o}[\hat{\mathbf{b}}'_j\hat{\mathbf{b}}_j + var_{d_c|d_o}[\hat{\mathbf{b}}_j \mid d_o]] = \tilde{\mathbf{b}}'_j\tilde{\mathbf{b}}_j + var_{d_c|d_o}[\tilde{\mathbf{b}}_j \mid d_o]$. The estimate $\tilde{\boldsymbol{\Sigma}}_b$ has the same form as Shah et al. (1997).

Use of the adjusted profile likelihood under d_o requires that $\hat{\boldsymbol{\Sigma}}$ is orthogonal to $\tilde{\boldsymbol{\Gamma}}$ under d_o, which means that $hinfo(\tilde{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{\Sigma}}; d_o)$ must be block diagonal. From (17) this requirement is met by noting that: (i) $\mathbf{H}_c(\boldsymbol{\Gamma}, \boldsymbol{\Sigma}; d_c)) =$

$diag\{\mathbf{H}_c(\mathbf{\Gamma}; d_c)), \mathbf{H}_c(\mathbf{\Sigma}; d_c))\}$ (see Section 2.2) and; (ii) $Cov\,[Sc(\mathbf{\Gamma}; d_c), Sc(\mathbf{\Sigma}; d_c) \mid d_o]$ is block diagonal if the data are MCARWG. If the data are MARWG, the off-diagonals of $Cov\,[Sc(\mathbf{\Gamma}; d_c), Sc(\mathbf{\Sigma}; d_c) \mid d_o]$ will be non-zero. However, we show in simulations that the HL estimates work well even when the data are MARWG.

3.3 Estimation

The estimation procedure based on d_o involves:

1. Initialising $\mathbf{\Sigma}$, denoted by $\mathbf{\Sigma}^{(0)}$, by the identify matrix.

2. Calculating $\tilde{\mathbf{\Gamma}}^{(t)}$ from (21) and (20) using $\tilde{\mathbf{\Sigma}}^{(t)}$

3. Calculating $\tilde{\mathbf{\Sigma}}^{(t+1)}$ from (6)after replacing the missing values by their conditional expectation (see 16) and using $\tilde{\mathbf{\Gamma}}^{(t)}$

4. Repeating 2 - 3 until convergence.

5. Calculating $\mathbf{H}_o$.

4 Simulation Study

4.1 Data

The simulation study involved creating the complete data from (1) for the case of three variables ($K = 3$), $\boldsymbol{\mu} = (5, 3, 1)$ and 10 groups ($J = 10$). This study considered $\bar{n} = 6, 10$, $\mathbf{\Sigma}_w = \boldsymbol{\rho}$, $\mathbf{\Sigma}_b = v\boldsymbol{\rho}$, $v = 0.1, 1$,

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0.83 & 0.88 \\ & 1 & 0.81 \\ & & 1 \end{pmatrix}$$

This study considered each of the 4 possible combinations of $\bar{n}$ and v to generate complete data. For each of these 4 combinations, 1200 complete data sets were randomly generated. From each set of complete data, the data were simulated to be either MCARWG and MARWG, as described below.

Data were simulated to be missing so that when $\bar{n} = 6\,(10)$, only 3 (4) of the 6 (10) observations in each group were complete.

When the data were MCARWG and $\bar{n} = 10$, the six incomplete observations per group were missing y_1, y_2, y_3, (y_1, y_2), (y_1, y_3), and (y_2, y_3). When $\bar{n} = 6$, the three incomplete observations were missing y_1, y_2, and (y_2, y_3). The observations selected to be incomplete were made completely at random.

When the data were MARWG the incomplete observations per group were missing either y_2 or y_3 (but not both). The probability that observation i in group j was incomplete was proportional to $|y_{ij1}|/|\Sigma_i^{\bar{n}} y_{ij1}|$. If an observation was determined to be incomplete, y_{2i} or y_{3i} (but not both) was randomly chosen to be missing.

With complete data we estimate $\mathbf{\Sigma}$ using ANOVA (see 14) and HL (see (12) and (13)). With incomplete data we estimate $\mathbf{\Sigma}$ by the ANOVA method using only the complete cases (i.e. observations for which all variables are observed) and by the HL method with complete and incomplete cases (see section 3.3). Each estimate of $\mathbf{\Sigma}$ just mentioned is substituted into (6) to give a corresponding estimate of $\mathbf{\Gamma}$ for the ANOVA and HL methods.

The MSE of the estimator $\hat{\boldsymbol{\theta}}$, is $MSE(\hat{\boldsymbol{\theta}}) = 1200^{-1}\Sigma_{g=1}^{1200}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta})^2$ where $\boldsymbol{\theta}$ is known and $\hat{\boldsymbol{\theta}}_g$ is the estimate of $\boldsymbol{\theta}$ from the g th simulated data set, where $g = 1, \ldots, 1200$.

Define the Relative MSE (RMSE) of $\hat{\boldsymbol{\theta}}$ by

$$100 \; MSE(\hat{\boldsymbol{\theta}})/MSE(\hat{\boldsymbol{\theta}}_{AC}).$$

where $MSE(\hat{\boldsymbol{\theta}}_{AC})$ is the MSE of the ANOVA estimator with complete data (AC). Tables 1 and 2 give the RMSE for HL with complete and incomplete data and ANOVA with complete cases (ACC).

It is important to note that ANOVA gives unbiased estimates of $\boldsymbol{\Sigma}_b$ only if the probability that it gives infeasible values (e.g. negative diagonals) is zero (McCulloch & Searle, 2001, see p 172). For the AC estimator of $\boldsymbol{\Sigma}_b$ with v=0.1 this was not the case, with up to 70% of the 1200 simulated samples resulting in infeasible values. When there are infeasible values, the estimate of $\boldsymbol{\Sigma}_b$ is set to $\mathbf{0}_{KK}$ (see McCulloch & Searle, 2001, see p 172). Doing so made AC biased: if AC gives infeasible values 70% of the time its bias would be 70%- assuming it is unbiased when it gives feasible values. This situation was more severe for ACC than for AC (see tables for details). This means, as a general approach, ANOVA performed poorly. Nevertheless, to make ANOVA competitive, AC and ACC estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ from the g th simulated data set were only included in their coverage and MSE calculations if the estimate of $\boldsymbol{\Sigma}_b$ was feasible. This should be kept in mind when analysing the tables. We note that HL estimates of the diagonals of $\mathbf{\Sigma}_b$ were always positive and so estimates from all 1200 simulated data sets were used its MSE calculation.

With complete data, the RMSE of estimates of $\mathbf{\Sigma}$ from HL are close to 100 when v=1. This means the MSEs for HL and ANOVA estimates of $\mathbf{\Sigma}$ are close in this case. When v=0.1, the HL is slightly more efficient than ANOVA when estimating $\mathbf{\Sigma}_w$, but can be significantly more efficient when estimating $\mathbf{\Sigma}_b$. In particular, the MSE of HL can be half that of AC.

With incomplete data, the results show that ACC has the highest RMSEs. This is especially the case when the data are MARWG, in which case ACC is biased. The RMSEs for HL are substantially smaller than ACC. Despite the considerable amount of missing data, the RMSEs for HL with incomplete data is often not much larger than HL with complete data. The RMSEs for HL did not depend greatly upon whether the data were MCARWG or MARWG.

Tables 3 and 4 give the coverage properties for $\mathbf{\Gamma}$. Whether for ACC , AC or HL, the coverage of the confidence intervals based on the t-distribution were very sensitive to the choice of the degrees of freedom, v and n_j. A range of options were considered for the degree of freedom (e.g. $df(\mu_k) = J-1$ and $df(b_{jk}) = \bar{n}-1$) but most performed poorly. The most promising choices for the degrees of freedom, based on trial and error, are discussed below.

The degrees of freedom for the t-distribution used to construct confidence

Table 1: RMSEs for $(\mathbf{\Gamma}, \Sigma)$ when $n_j = 10$

	$v = 1$					$v = 0.1$				
	Complete	MCARWG		MARWG		Complete	MCARWG		MARWG	
	HL	HL	ACC	HL	ACC	HL	HL	ACC	HL	ACC
μ_1	100	100	113	101	219	101	105	173	108	1150
μ_2	100	102	112	100	187	100	107	175	104	805
μ_3	100	100	112	100	244	103	108	172	102	1472
b_{j1}	100	104	169	105	224	93	97	200	97	273
b_{j2}	100	107	167	106	220	91	97	195	96	264
b_{j3}	100	105	167	100	226	92	96	195	93	277
$\sigma_{w,11}$	100	122	304	128	804	99	114	285	119	504
$\sigma_{w,22}$	100	133	304	122	666	102	131	272	126	435
$\sigma_{w,33}$	100	129	295	100	954	97	118	276	100	495
$\sigma_{w,12}$	100	119	294	119	744	100	114	266	116	505
$\sigma_{w,13}$	100	107	294	108	894	97	111	278	105	505
$\sigma_{w,23}$	100	122	293	110	843	99	119	273	113	546
$\sigma_{b,11}$	100	101	131	100	354	74	83	470	81	1150
$\sigma_{b,22}$	100	101	129	103	265	79	96	565	90	113
$\sigma_{b,33}$	100	101	128	100	422	78	89	47	80	1411
$\sigma_{b,12}$	100	101	135	101	322	74	83	585	78	1314
$\sigma_{b,13}$	100	100	131	100	412	74	80	600	76	1425
$\sigma_{b,23}$	100	101	131	100	366	72	78	528	75	1328

Notes on Convergence

-AC did not give positive values for the diagonals of $\mathbf{\Sigma}_b$ in 5% and 50% of the 1200 simulated samples when v=1 and v=0.1, respectively

-When the data were MCARWG, ACC did not give positive values for the diagonals of $\mathbf{\Sigma}_b$ in 5% and 30% of the 1200 simulated samples when v=1 and v=0.1, respectively

-When the data were MARWG, ACC did not give positive values for the diagonals of $\mathbf{\Sigma}_b$ in 8% and 74% of the 1200 simulated samples when v=1 and v=0.1, respectively

Table 2: RMSE for $(\mathbf{\Gamma}, \Sigma)$ when $n_j = 6$

	$v = 1$					$v = 0.1$				
	Complete	MCARWG		MARWG		Complete	MCARWG		MARWG	
	HL	HL	ACC	HL	ACC	HL	HL	ACC	HL	ACC
μ_1	100	101	109	101	201	90	96	154	98	567
μ_2	100	101	109	101	171	95	109	132	100	320
μ_3	100	101	109	100	211	90	94	145	90	638
b_{j1}	100	107	159	103	200	84	89	180	88	190
b_{j2}	100	111	157	103	200	85	89	180	87	191
b_{j3}	100	105	157	100	100	84	88	180	85	197
$\sigma_{w,11}$	100	127	251	117	900	94	116	210	105	264
$\sigma_{w,22}$	100	146	245	113	666	101	143	248	111	240
$\sigma_{w,33}$	100	116	251	100	1030	93	110	205	94	292
$\sigma_{w,12}$	100	115	250	102	747	97	123	240	106	260
$\sigma_{w,13}$	100	118	245	105	940	92	112	206	96	284
$\sigma_{w,23}$	100	128	238	106	770	94	117	232	96	260
$\sigma_{b,11}$	100	106	133	103	205	56	65	357	62	450
$\sigma_{b,22}$	100	105	128	112	163	54	74	422	64	466
$\sigma_{b,33}$	100	104	133	100	238	60	66	420	61	502
$\sigma_{b,12}$	100	105	131	106	191	52	56	377	54	461
$\sigma_{b,13}$	100	104	133	102	230	61	62	394	63	505
$\sigma_{b,23}$	100	103	133	106	216	50	53	418	52	506

Notes on Convergence

-AC did not give positive values for the diagonals of $\mathbf{\Sigma}_b$ in 4% and 70% of the 1200 simulated samples when v=1 and v=0.1, respectively

-When the data were MCARWG, ACC did not give positive values for the diagonals of $\mathbf{\Sigma}_b$ in 12% and 76% of the 1200 simulated samples when v=1 and v=0.1, respectively

-When the data were MARWG, ACC did not give positive values for the diagonals of $\mathbf{\Sigma}_b$ in 10% and 75% of the 1200 simulated samples when v=1 and v=0.1, respectively intervals for estimates $\hat{\mu}_k$ is $df(\hat{\mu}_k) = n\Big[\hat{\sigma}^2_{w,kk} n^{-1}\{Var(\hat{\mu}_k)\}^{-1}\Big]$. The term in the square brackets is often referred to as the *design effect* in survey sampling (see Chambers & Skinner, 2003). The design effect measures the increase in variance of an estimate, or the equivalently decrease in sample size, due to the fact that each observation is not independent. If the sample was selected by Simple Random Sampling or if $\hat{\mathbf{\Sigma}}_b = \mathbf{0}_{KK}$ then the term in the square brackets would be 1 and $df(\hat{\mu}_k) = n$; this effectively means the n observations are independent. In the present case, the design effect will be greater than 1 meaning $df(\hat{\mu}_k) < n$. The degrees of freedom for HL, $df(\tilde{\mu}_k)$, is the same as above except that $Var(\hat{\mu}_k)$ is replaced by $Var(\tilde{\mu}_k)$.

A general expression for the degrees of freedom associated with an estimate of $\boldsymbol{\theta}$ is $trace(\mathbf{H})$, where $\hat{y}(\boldsymbol{\theta}) = \mathbf{H}y$, where $\hat{y}(\boldsymbol{\theta})$ are the fitted values of y which are functions of the parameter $\boldsymbol{\theta}$, and y are the observed values. The estimator of b_{jk} in (6) is already in this form. This justified setting $df(\hat{b}_{jk}) = max\Big\{1, n_j\hat{\sigma}^2_{bkk}(\hat{\sigma}^2_{bkk} + \hat{\sigma}^2_{wkk} n_j^{-1})\Big\}$, where the second term in the curly brackets is equal to n_j multiplied by the shrinkage factor for the random effect $\hat{b}_{jk}$. The minimum of 1 provided robustness against the variability in the estimates of the variance components. The shrinkage factor can also be thought of as effectively reducing the effective sample size, by down-weighting the contribution of the n_j observations in the estimate of b_{jk}. For the same reason, $df(\tilde{b}_{jk}) = max\Big\{1, n_j\tilde{\sigma}^2_{bkk}(\tilde{\sigma}^2_{bkk} + \tilde{\sigma}^2_{wkk} n_j^{-1})\Big\}$ From the form of $df(\tilde{b}_{jk})$ it is apparent that no explicit attempt is made to account

Table 3: Coverage (95%) for $\mathbf{\Gamma}$ when $n_j = 10$

	$v = 1$						$v = 0.1$					
	Complete		MCARWG		MARWG		Complete		MCARWG		MARWG	
	HL	AC	HL	ACC	HL	ACC	HL	AC	HL	ACC	HL	ACC
μ_1	94.7	94.9	94.9	96.3	94.6	93.2	94.5	96.0	93.5	98.3	94.2	69.9
μ_2	95.5	95.5	95.6	96.2	95.2	94.6	94.0	95.8	94.2	97.5	94.8	78.7
μ_3	94.9	94.9	94.5	95.4	94.8	90.4	95.6	97.0	94.8	98.3	94.4	64.5
b_{j1}	96.1	96.1	95.9	98.6	95.9	100.0	96.7	95.7	95.5	98.7	96.5	97.0
b_{j2}	96.3	96.5	96.1	98.7	96.0	99.0	96.8	94.3	94.8	97.8	96.9	97.0
b_{j3}	96.0	96.0	95.7	98.4	95.9	100.0	96.7	94.7	94.9	97.2	96.6	97.9

Table 4: Coverage (95%) for $\mathbf{\Gamma}$ when $n_j = 6$

	$v = 1$						$v = 0.1$					
	Complete		MCARWG		MARWG		Complete		MCARWG		MARWG	
	HL	AC	HL	ACC	HL	ACC	HL	AC	HL	ACC	HL	ACC
μ_1	95.8	95.9	95.6	97.2	93.9	94.5	93.9	94.7	94.2	97.5	93.8	81.1
μ_2	95.4	95.5	95.8	96.7	94.2	94.1	94.2	95.5	95.0	97.5	94.2	86.6
μ_3	95.2	95.5	95.4	96.7	94.0	93.4	94.1	94.8	95.6	97.9	94.0	76.3
b_{j1}	97.8	97.8	97.3	99.0	98.6	99.4	98.9	97.9	96.5	96.9	98.0	92.7
b_{j2}	97.6	97.6	97.1	98.7	98.7	97.9	98.7	97.5	96.8	96.5	98.7	96.7
b_{j3}	97.6	97.5	97.4	98.7	98.6	99.2	98.7	98.2	96.5	98.5	98.5	99.4

for the loss in the degrees of freedom due to missing data.

The coverage for the AC and HL were reasonably close to the nominal value of 95%. When the data are MARWG, ACC estimates are biased, leading to coverage rates varying far from their nominal values.

5 Discussion and Future Work

This paper proposes a method for estimating the fixed effects, random effects and the variance components for both a multi-variate random effects model with complete and incomplete data. This paper uses the EM algorithm to maximise the hierachical likelihood and shows that it equivalent to the REML approach of Shah et al. (1997). A key benefit of the h-likelihood approach is its simplicity- it doesn't require integrating over the random effects or use of priors for its justification. Simulations show the h-likelihood approach is significantly more efficient than the well-known ANOVA approach at estimating the variance components. The ANOVA is unstable in that it often gives values for the between-group variance, especially when the the between-group variance is a tenth the size of the between-observation (or individual) variance. Even when ignoring this major draw-back, ANOVA is inefficient compared with the HL approach, particularly when estimating the between-group variation and the random effects. Allowing for missing data is straight-forward and avoids the complexities associated with integration, commonly used to handle missing data in mixed models. The paper suggests a way of choosing the degrees of freedom to support good coverage rates in small samples.

A Appendix

A.1 Estimate of $\boldsymbol{\Sigma}_w$

We look at the three terms in $Sc(\alpha_s; d_c)$ given by (10). Let $\hat{\mathbf{e}} = (\mathbf{y}^* - \mathbf{q}\hat{\boldsymbol{\mu}} - \mathbf{Z}^*\hat{\mathbf{b}})$ and $\hat{\mathbf{e}}_{ij}$ be the K subvector of $\hat{\mathbf{e}}$ corresponding to the (i, j) th observation. Since $\mathbf{V}_{w(r)}$ is block diagonal, from the first term note that $-\hat{\mathbf{e}}'\mathbf{V}_{w(r)}^{-1}\hat{\mathbf{e}} = tr[\hat{\mathbf{e}}\hat{\mathbf{e}}'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}] = tr[\Sigma_{ij}\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}'_{ij}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}]$, where $\boldsymbol{\Sigma}_{w(r)} = \partial\boldsymbol{\Sigma}_w/\partial\phi_r$. Looking at the third term $-tr[\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}] = -ntr[\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}]$. The second term is $tr[\mathbf{H}_c^{-1}\mathbf{H}_{c(r)}]$, where $\mathbf{H}_{c(r)} = \partial\mathbf{H}_c/\partial\phi_r$. As $\mathbf{q}$ is formed by stacking copies of $\mathbf{I}_K$,

$$\begin{aligned}
\mathbf{H}_{c(r)} &= \begin{pmatrix} -\mathbf{q}'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{q} & -\mathbf{q}'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{Z}^* \\ -\mathbf{Z}^{*\,\prime}\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{q} & -\mathbf{Z}^{*\,\prime}\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{Z}^* \end{pmatrix} \\
&= \begin{pmatrix} -n\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} & \left\{_r - n_j\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}\right\}_{j=1}^{J} \\ \left\{_c - n_j\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}\right\}_{j=1}^{J} & \left\{_d - n_j\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}\right\}_{j=1}^{J} \end{pmatrix} \\
&= -\left\{_d\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}\right\}_{j=1}^{J} \begin{pmatrix} n\mathbf{I}_K & \left\{_r n_j\mathbf{I}_K\right\}_{j=1}^{J} \\ \left\{_c n_j\mathbf{I}_K\right\}_{j=1}^{J} & \left\{_d n_j\mathbf{I}_K\right\}_{j=1}^{J} \end{pmatrix} \\
&= -\left\{_d\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}\right\}_{u=1}^{J+1}\mathbf{a}
\end{aligned}$$

where

$$\mathbf{a} = \begin{pmatrix} n\mathbf{I}_K & \left\{_r n_j\mathbf{I}_K\right\}_{j=1}^{J} \\ \left\{_c n_j\mathbf{I}_K\right\}_{j=1}^{J} & \left\{_d n_j\mathbf{I}_K\right\}_{j=1}^{J} \end{pmatrix}$$

Similarly we may write

$$\mathbf{H}_c = \begin{pmatrix} \boldsymbol{\Sigma}_w^{-1}n\mathbf{I}_K & \left\{_r n_j\boldsymbol{\Sigma}_w^{-1}\right\}_{j=1}^{J} \\ \left\{_c n_j\boldsymbol{\Sigma}_w^{-1}\right\}_{j=1}^{J} & \left\{_d n_j\boldsymbol{\Sigma}_w^{-1} + \boldsymbol{\Sigma}_b^{-1}\right\}_{j=1}^{J} \end{pmatrix}$$

$$= \Big\{_d \mathbf{\Sigma}_w^{-1}\Big\}_{u=1}^{J+1}\mathbf{b}$$

where

$$\mathbf{b} = \begin{pmatrix} n\mathbf{I}_K & \Big\{_r n_j\mathbf{I}_K\Big\}_{j=1}^{J} \\ & \\ \Big\{_c n_j\mathbf{I}_K\Big\}_{j=1}^{J} & \Big\{_d n_j\mathbf{I}_K + \mathbf{\Sigma}_w\mathbf{\Sigma}_b^{-1}\Big\}_{j=1}^{J} \end{pmatrix}$$

It follows that

$$\begin{aligned} \mathbf{H}_c^{-1}\mathbf{H}_{c(r)} &= \mathbf{b}^{-1}\Big\{_d \mathbf{\Sigma}_w\Big\}_{u=1}^{J+1}\Big\{_d \mathbf{\Sigma}_w^{-1}\mathbf{\Sigma}_{w(r)}\mathbf{\Sigma}_w^{-1}\Big\}_{u=1}^{J+1}\mathbf{a} \\ &= \mathbf{b}^{-1}\mathbf{a}\Big\{_d \mathbf{\Sigma}_w\Big\}_{u=1}^{J+1}\Big\{_d \mathbf{\Sigma}_w^{-1}\mathbf{\Sigma}_{w(r)}\mathbf{\Sigma}_w^{-1}\Big\}_{u=1}^{J+1} \\ &= \mathbf{g}\Big\{_d \mathbf{\Sigma}_{w(r)}\mathbf{\Sigma}_w^{-1}\Big\}_{u=1}^{J+1}, \end{aligned}$$

noting that swapping the order of the matrices is permissable since all matrices are symmetric.

Substituting these three terms into the equation $Sc(\phi_r; d_c) = 0$, letting $\mathbf{g} = \mathbf{b}^{-1}\mathbf{a}$ and $\mathbf{g}_j$ be the diagonal blocks of $\mathbf{g}$ of dimension KxK we obtain

$$tr\Big[\Sigma_{ij}\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}_{ij}'\mathbf{\Sigma}_w^{-1}\mathbf{\Sigma}_{w(r)}\mathbf{\Sigma}_w^{-1}\Big] + tr\Big[\mathbf{g}\Big\{_d \mathbf{\Sigma}_{w(r)}\mathbf{\Sigma}_w^{-1}\Big\}_{j=1}^{J+1}\Big] - ntr\Big[\mathbf{\Sigma}_w^{-1}\mathbf{\Sigma}_{w(r)}\Big] = 0$$

which implies

$$tr\Big[\Sigma_{ij}\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}_{ij}'\mathbf{\Sigma}_w^{-1}\mathbf{\Sigma}_{w(r)}\mathbf{\Sigma}_w^{-1}\Big] + tr\Big[\Sigma_{u=1}^{J+1}\mathbf{g}_u\mathbf{\Sigma}_{w(r)}\mathbf{\Sigma}_w^{-1}\Big] - ntr\Big[\mathbf{\Sigma}_w^{-1}\mathbf{\Sigma}_{w(r)}\Big] = 0 \quad (22)$$

A solution to this equation for all ϕ_r requires that

$$\Sigma_{ij}\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}_{ij}'\mathbf{\Sigma}_w^{-1} + \Sigma_u^{J+1}\mathbf{g}_u - n\mathbf{I}_K = 0$$

After rearranging we obtain an estimate of Σ_w from d_c given by

$$\hat{\mathbf{\Sigma}}_w = (n\mathbf{I}_K - \Sigma_j^{J+1}\mathbf{g}_j)^{-1}\Sigma_{ij}\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}_{ij}'$$

A.2 Estimate of $\mathbf{\Sigma}_b$

From the first term in (11),

$$tr[\hat{\mathbf{b}}'\mathbf{V}_{b(s)}^{-1}\hat{\mathbf{b}}] = tr[\hat{\mathbf{b}}\hat{\mathbf{b}}'\mathbf{V}_{b(s)}^{-1}] = tr[\hat{\mathbf{b}}\hat{\mathbf{b}}'\mathbf{V}_b^{-1}\mathbf{V}_{b(s)}\mathbf{V}_b^{-1}],$$

$$\mathbf{V}_{b(s)}^{-1} = -\partial\mathbf{V}_b^{-1}/\partial\alpha_s = \mathbf{V}_b^{-1}\mathbf{V}_{b(s)}\mathbf{V}_b^{-1}$$

and

$$\mathbf{V}_{b(s)} = \partial\mathbf{V}_b/\partial\alpha_s.$$

Making these substitutions into $Sc(\alpha_s; d_c)) = 0$ and solving results in

$$\begin{aligned}
tr[\hat{\mathbf{b}}\hat{\mathbf{b}}'\mathbf{V}_b^{-1}\mathbf{V}_{b(s)}\mathbf{V}_b^{-1}] + tr[\mathbf{K}_c\mathbf{V}_b^{-1}\mathbf{V}_{b(s)}\mathbf{V}_b^{-1}] - tr[\mathbf{V}_b^{-1}\mathbf{V}_{b(s)}] &= 0 \\
A\ solution\ for\ \alpha_s\ for\ all\ s\ is\ then \\
tr[\hat{\mathbf{b}}\hat{\mathbf{b}}'\mathbf{V}_b^{-1}] + tr[\mathbf{K}_c\mathbf{V}_b^{-1}] - tr[\mathbf{I}_{KJ}] &= 0 \\
tr\left[\Sigma_j\hat{\mathbf{b}}_j\hat{\mathbf{b}}_j'\mathbf{\Sigma}_b^{-1} + \Sigma_j\mathbf{K}_{c,j}\mathbf{\Sigma}_b^{-1} - J\mathbf{I}_K\right] &= 0 \\
Noting\ that\ tr(\mathbf{A}) = tr(\mathbf{B})\ if\ \ \mathbf{A} = \mathbf{B} &\quad it\ follows\ that \\
\Sigma_j\hat{\mathbf{b}}_j'\hat{\mathbf{b}}_j\mathbf{\Sigma}_b^{-1} + \Sigma_j\mathbf{K}_{c,j}\mathbf{\Sigma}_b^{-1} - J\mathbf{I}_K &= \mathbf{0}_{KK} \\
\mathbf{\Sigma}_b^{-1} &= [\Sigma_j\hat{\mathbf{b}}_j'\hat{\mathbf{b}}_j + \Sigma_j\mathbf{K}_{c,j}]^{-1}J \\
\mathbf{\Sigma}_b &= [\Sigma_j\hat{\mathbf{b}}_j'\hat{\mathbf{b}}_j + \Sigma_j\mathbf{K}_{c,j}]J^{-1}
\end{aligned}$$

Therefore an estimate of $\mathbf{\Sigma}_b$ based on d_c is $\hat{\mathbf{\Sigma}}_b = \Sigma_j[\hat{\mathbf{b}}_j'\hat{\mathbf{b}}_j + \mathbf{K}_{c,j}]J^{-1}$.

References

Breckling, J. U., Chambers, R. L., Dorfman, A. H., Tam, S. M., & Welsh, A. H. (1994). Maximum likelihood inference from sample survey data. *Internatoinal Statistical Review*, *62*, 349-63.

Chambers, R. L., & Skinner, C. J. (2003). *Analysis of survey data.* John Wiley and Sons.

Cox, D. R., & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society Series B*, *49*, 1-39.

Diggle, P. J., & Kenward, M. G. (1994). Informative drop-out in longitudinal analysis (with discussion). *Applied Statistics*, *43*, 49-93.

Lee, Y., Nelder, J. A., & Pawitan, Y. (2006). *Generalized linear models with randome effects: Unified analysis via h-likelihood.* Chapman and Hall.

Lee, Y., & Nelder J., A. (1996). Heirachical generalized linear models. *Journal of the Royal Statistical Society. Series B.*, *58*, 619-678.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized linear and mixed models.* John Wiley and Sons.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*, 545-554.

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science*, *6*, 15-51.

Rubin, D. B., & Little, R. J. A. (1987). *Statistical analysis of missing data.* John Wiley and Sons.

Shah, A., Laird, N., & Schoenfeld, D. (1997). A randome effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association*, *92*, 775-779.

Yang, M., Goldstein, H., Browne, W., & Woodhouse, G. (2002). Multivariate multilevel analyses of examination results. *Journal of the Royal Statistical Society Series A*, *165*, 137-153.

Yun, S., Lee, Y., & Kenward, M. G. (2007). Using hierachical likelihood for missing data problems. *Biometrika*, *94*, 905-919.