STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:1397–1412
Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2384

A robust approach to *t* linear mixed models applied to multiple sclerosis data

Tsung I. Lin[1,∗,†] and Jack C. Lee[2]

[1]*Department of Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan*
[2]*Institute of Statistics and Graduate Institute of Finance, National Chiao Tung University, Hsinchu 30050, Taiwan*

SUMMARY

We discuss a robust extension of linear mixed models based on the multivariate *t* distribution. Since longitudinal data are successively collected over time and typically tend to be autocorrelated, we employ a parsimonious first-order autoregressive dependence structure for the within-subject errors. A score test statistic for testing the existence of autocorrelation among the within-subject errors is derived. Moreover, we develop an explicit scoring procedure for the maximum likelihood estimation with standard errors as a by-product. The technique for predicting future responses of a subject given past measurements is also investigated. Results are illustrated with real data from a multiple sclerosis clinical trial. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: Fisher scoring; longitudinal data; prediction; random effects; *t*-REML

1. INTRODUCTION

Multiple sclerosis (MS), one of the most common chronic diseases of the central nervous system in young adults, occurs when the myelin around the nerve fibres in the brain becomes damaged. As yet, the precise causes of MS remain unknown, though abundant research suggests MS may be an autoimmune disease in which the immune system attacks its own myelin, causing disruptions to the nerve transmissions. There are no drugs to cure MS, but some treatments are available to ease the symptom. For example, interferon beta-1b (INFB) was approved by the US Food and Drug Administration in mid-1993 for use in early stage relapsing-remitting MS (RRMS) patients. For diagnosis, cranial magnetic resonance imaging

∗Correspondence to: Tsung I. Lin, Department of Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan.
†E-mail: tilin@amath.nchu.edu.tw

Contract/grant sponsor: National Science Council of Taiwan; contract/grant numbers: NSC93-2118-M-029-004, NSC93-2118-M-009-003

Received 23 June 2004
Accepted 6 July 2005

(MRI) is the most preferred tool for monitoring MS evolution in both natural history studies and treatment trials.

Gill [1] presents a robust approach based on Huber's ρ function to a linear mixed model for the analysis of a data set, called the MS data throughout this paper, from a cohort study of 52 patients with RRMS. The study was a placebo-controlled trial of interferon beta-1b (INFB) in which patients were randomized to either a placebo (PL), a low-dose (LD), or a high-dose (HD) treatment. The LD and HD treatments correspond to doses of 1.6 and 8 million international units (MIU) of IFNB every other day, respectively. Each patient had a baseline cranial MRI and subsequent MRIs once every 6 weeks over two years. The 6-weekly serial MRI data were collected from June 1988 to May 1990 at the University of British Columbia site.

The use of the t distribution in place of the normal for robust regression has been investigated by a number of authors, including West [2], Lange *et al.* [3] and James *et al.* [4]. The linear mixed model with multivariate t distributed responses, called the *t linear mixed model* hereafter, was considered by Welsh and Richardson [5], however, they do not explicitly discuss or derive the distributions of the random effects as well as the error terms. More recently, Pinheiro *et al.* [6] incorporated multivariate t distributed random effects and error terms to formulate a normal–normal–gamma hierarchy for the t linear mixed model. They provide several efficient EM-type algorithms for maximum likelihood (ML) estimation and illustrate the robustness with respect to outlying observations using a real example and some simulation results.

In this paper, we develop additional tools for a simplified version of the Pinheiro *et al.* [6] model and use these tools to analyse the MS data. The model considered here is

$$\begin{aligned} \mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i, \quad \mathbf{b}_i \mid \tau_i \sim N_{m_2}\left(\mathbf{0}, \frac{\sigma^2}{\tau_i}\boldsymbol{\Gamma}\right) \\ \epsilon_i \mid \tau_i \sim N_{p_i}\left(\mathbf{0}, \frac{\sigma^2}{\tau_i}\mathbf{C}_i\right), \quad \tau_i \sim Ga(v/2, v/2), \ (i = 1, \ldots, N) \end{aligned} \tag{1}$$

where i is the subject index, $\mathbf{Y}_i$ is a p_i-dimensional observed response vector, N is the number of subjects, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are, respectively, known $p_i \times m_1$ and $p_i \times m_2$ design matrices, $\boldsymbol{\beta}$ is an $m_1 \times 1$ vector of fixed effects, $\mathbf{b}_i$ is an $m_2 \times 1$ vector of unobservable random effects, τ_i is an unknown scale assumed to be distributed as gamma with mean 1 and variance $2/v$, and $\mathbf{b}_i \mid \tau_i$ and $\epsilon_i \mid \tau_i$ are assumed to be independent. Furthermore, $\boldsymbol{\Gamma}$ is an $m_2 \times m_2$ matrix, which may be unstructured or structured, and $\mathbf{C}_i$ is a $p_i \times p_i$ correlation matrix.

Pinheiro *et al.* [6] consider a general model where $\mathbf{C}_i$ is allowed to depend upon a vector of parameters and the parameter v is allowed to vary across subgroups of subjects. In this paper, we exploit the widely used autoregressive structure to model the dependence for the within-subject errors. As an illustration, we concentrate on the simple case where $\mathbf{C}_i$ has an AR(1) dependence structure that is common to all subjects, i.e.

$$\mathbf{C}_i = \mathbf{C}_i(\rho) = [\rho^{|r-s|}], \quad r, s = 1, 2, \ldots, p_i \tag{2}$$

The dependence structure of $\mathbf{C}_i$ can be extended to a high order autoregressive moving average (ARMA) dependence as provided by Rochon [7], Lin and Lee [8] and Lee *et al.* [9].

In model (1), the marginal distribution of the response $\mathbf{Y}_i$, after integrating over $\mathbf{b}_i$ and τ_i, can be expressed as

$$\mathbf{Y}_i \sim t_{p_i}(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{\Lambda}_i, v) \tag{3}$$

where $\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_i(\boldsymbol{\Gamma}, \rho) = \mathbf{Z}_i\boldsymbol{\Gamma}\mathbf{Z}_i' + \mathbf{C}_i(\rho)$ and $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, v)$ denotes the p-dimensional multivariate t distribution with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$ and degrees-of-freedom (d.f.) v.

In Section 2, we describe computational aspects of both the ML estimation and restricted maximum likelihood (REML) estimation based on the marginal likelihood of (3). Section 3 describes how to obtain the score test statistic for testing $H_0 : \rho = 0$ for $\mathbf{C}_i$ in (2) against the alternative hypothesis $\rho \neq 0$. Section 4 discusses inferences of random effects and prediction problems. In Section 5, we illustrate the proposed methodologies with the MS data. Finally, Section 6 offers some concluding remarks.

2. ESTIMATION

For notational convenience, let $\mathbf{e}_i = \mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}$, $\boldsymbol{\Delta}_i(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \rho) = \mathbf{e}_i^\top \boldsymbol{\Lambda}_i(\boldsymbol{\Gamma}, \rho)\mathbf{e}_i$ and $n = \sum_{i=1}^N p_i$ denote the total number of observations. Given independent observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$, write the log-likelihood function as

$$\ell = \sum_{i=1}^N \left\{ \log\left(\Gamma\left(\frac{v+p_i}{2}\right)\right) - \log\left(\Gamma\left(\frac{v}{2}\right)\right)\right\} - \frac{n}{2}\log(v\sigma^2)$$
$$-\frac{1}{2}\sum_{i=1}^N \log|\boldsymbol{\Lambda}_i(\boldsymbol{\Gamma}, \rho)| - \frac{1}{2}\sum_{i=1}^N (v+p_i)\log\left(1 + \frac{\boldsymbol{\Delta}_i(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \rho)}{\sigma^2 v}\right) \tag{4}$$

To ensure non-negative definiteness of $\boldsymbol{\Gamma}$, we reparameterize $\boldsymbol{\Gamma} = \mathbf{F}^\top\mathbf{F}$ by the Cholesky decomposition, where $\mathbf{F}$ is an upper triangular matrix. Let $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\sigma^2, \mathrm{vech}(\mathbf{F}), \rho, v)$ is the vector of unknown model parameters excluding the fixed effects $\boldsymbol{\beta}$. Explicit expressions for the score vector $\mathbf{s}_{\boldsymbol{\alpha}} = (\mathbf{s}_{\boldsymbol{\beta}}^\top, \mathbf{s}_{\boldsymbol{\theta}}^\top)^\top$ and the Fisher information matrix $\mathbf{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}$ are derived in Appendix A. We employ the Fisher scoring algorithm to obtain the ML estimate. Under some regularity conditions, the asymptotic variance–covariance estimates can be computed by plugging the ML estimate $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ into the inverse of the Fisher information matrix. The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ can be obtained by

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{I}}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} = \hat{\sigma}^2\left(\sum_{i=1}^N \frac{\hat{v}+p_i}{\hat{v}+p_i+2}\mathbf{X}_i^\top\hat{\boldsymbol{\Lambda}}_i^{-1}\mathbf{X}_i\right)^{-1}$$
$$\mathrm{var}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{I}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \tag{5}$$

A disadvantage of the ML estimates of variance components is that they are biased downward in finite samples. REML corrects for the loss of degrees-of-freedom incurred in estimating the fixed effects and produces unbiased estimating equations for the variance components. As pointed out by Harville [10], REML can be viewed as the Bayesian principle of marginal inference. Adopting the prior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\theta}) \propto 1$ and Laplace's method as in Welsh and Richardson [5], the t-REML likelihood function can be approximated by $L_R(\boldsymbol{\theta}) = \int L(\boldsymbol{\beta}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\beta} \approx L_R^*(\boldsymbol{\theta})$, where

$$L_R^*(\boldsymbol{\theta}) = (\sigma^2 \pi v)^{-n/2} \left| \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{H}_i \mathbf{X}_i \right|^{-1/2} \prod_{i=1}^{N} \frac{\boldsymbol{\Gamma}((v+p_i)/2)}{\boldsymbol{\Gamma}(v/2)} \times |\boldsymbol{\Lambda}_i|^{-1/2} \left(1 + \frac{\boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\Gamma}, \rho)}{v\sigma^2} \right)^{-(v+p_i)/2} \quad (6)$$

$$\mathbf{H}_i = (v + p_i) \left\{ \frac{\boldsymbol{\Lambda}_i^{-1}}{\sigma^2 v + \boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\Gamma}, \rho)} - \frac{2\boldsymbol{\Lambda}_i^{-1} \hat{\mathbf{e}}_i(\boldsymbol{\theta}) \hat{\mathbf{e}}_i^\top(\boldsymbol{\theta}) \boldsymbol{\Lambda}_i^{-1}}{(\sigma^2 v + \boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\Gamma}, \rho))^2} \right\}$$

$\hat{\mathbf{e}}_i(\boldsymbol{\theta}) = \mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, and $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is obtained by solving the following estimating equation:

$$\sum_{i=1}^{N} (v + p_i) \frac{\mathbf{X}_i^\top \boldsymbol{\Lambda}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})}{\sigma^2 v + \boldsymbol{\Delta}_i(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \rho)} = 0 \quad (7)$$

The resulting (approximately) REML estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_R$, can be obtained by implementing the Newton–Raphson (NR) algorithm with the ML estimate as the initial value. In the NR algorithm, the empirical Bayes estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}_R)$, must be computed at each iteration. It can be easily obtained by solving the estimating equation (7) with $\boldsymbol{\theta}$ replaced by current estimate $\hat{\boldsymbol{\theta}}_R$.

Appendix B presents the necessary first partial derivatives of $\ell_R^*(\boldsymbol{\theta})$, the logarithm of $L_R^*(\boldsymbol{\theta})$ in (6), for the NR algorithm. However, the second partial derivatives of $\ell_R^*(\boldsymbol{\theta})$ are tedious. The entries of the Hessian matrix can instead be approximated numerically.

3. THE SCORE TEST FOR AUTOCORRELATION

It is of interest to test whether autocorrelation exists among the within-subject errors. We derive a score test statistic which is asymptotically a chi-squared random variable with 1 d.f. and can be easily computed. In this context, the score test statistic is based on the score vector and information matrix evaluated under $H_0 : \rho = 0$. The advantage of the score test over other testing procedures such as the likelihood ratio or Wald tests is that it does not require comparison with the alternative model. Rejection of the null model does not indicate that the AR(1) dependence structure is appropriate, however, it provides a simple check for the presence of possible autocorrelation among the within-subject errors.

Let $\boldsymbol{\eta} = (\sigma^2, \mathrm{vech}(\mathbf{F}), v)$, so that $\boldsymbol{\theta} = (\boldsymbol{\eta}, \rho)$. Because $\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\theta}} = \mathbf{0}$, the information matrix $\mathbf{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\eta}, \rho)$ can be reexpressed as the block partitioned matrix

$$\mathbf{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} = \begin{bmatrix} \mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^\top & \mathbf{I}_{\boldsymbol{\eta}\boldsymbol{\eta}} & \mathbf{I}_{\boldsymbol{\eta}\rho} \\ \mathbf{0}^\top & \mathbf{I}_{\boldsymbol{\eta}\rho}^\top & \mathbf{I}_{\rho\rho} \end{bmatrix} \quad (8)$$

Let $\hat{\boldsymbol{\alpha}}_0 = (\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\eta}}_0, 0)$ denote the ML estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ under $H_0 : \rho = 0$. The score vector $[\partial\ell/\partial\boldsymbol{\alpha}]_{\hat{\boldsymbol{\alpha}}_0}$ has all components equal to 0 except the derivative with respect to ρ. The score test statistic λ_s is

$$\lambda_s = \left[\frac{\partial \ell}{\partial \boldsymbol{\alpha}}\right]_{\hat{\boldsymbol{\alpha}}_0}^{\top} [\mathbf{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}]_{\hat{\boldsymbol{\alpha}}_0}^{-1} \left[\frac{\partial \ell}{\partial \boldsymbol{\alpha}}\right]_{\hat{\boldsymbol{\alpha}}_0} = \frac{[\partial\ell/\partial\rho]_{\hat{\boldsymbol{\alpha}}_0}^2}{[\mathbf{I}_{\rho\rho\cdot\boldsymbol{\eta}}]_{\hat{\boldsymbol{\alpha}}_0}} \tag{9}$$

where

$$\frac{\partial \ell}{\partial \rho} = -\frac{1}{2}\sum_{i=1}^{N} \operatorname{tr}\left(\boldsymbol{\Lambda}_i^{-1}\frac{\partial \mathbf{C}_i}{\partial \rho}\right) + \frac{1}{2}\sum_{i=1}^{N}(v + p_i)\frac{\mathbf{e}_i^{\top}\boldsymbol{\Lambda}_i^{-1}\dfrac{\partial \mathbf{C}_i}{\partial \rho}\boldsymbol{\Lambda}_i^{-1}\mathbf{e}_i}{\sigma^2 v + \boldsymbol{\Delta}_i(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \rho)} \tag{10}$$

$$\mathbf{I}_{\rho\rho\cdot\boldsymbol{\eta}} = \mathbf{I}_{\rho\rho} - \mathbf{I}_{\boldsymbol{\eta}\rho}^{\top}\mathbf{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}\mathbf{I}_{\boldsymbol{\eta}\rho}$$

The detailed expressions for $\mathbf{I}_{\rho\rho}$, $\mathbf{I}_{\boldsymbol{\eta}\rho}$ and $\mathbf{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}$ are given in Appendix A.

When evaluated at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}_0$, we obtain $\hat{\mathbf{u}}_i = [\boldsymbol{\Lambda}_i^{-1}\mathbf{e}_i]_{\hat{\boldsymbol{\alpha}}_0} = \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\mathbf{b}}_i$, where $\hat{\mathbf{b}}_i = (\mathbf{Z}_i^{\top}\mathbf{Z}_i + \hat{\boldsymbol{\Gamma}}^{-1})^{-1}\mathbf{Z}_i^{\top}(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})$ and $\hat{\mathbf{u}}_i$ can be viewed as the vector of residuals from the model for subject i. In addition, $[\mathbf{C}_i^{-1}]_{\hat{\boldsymbol{\alpha}}_0} = \mathbf{I}_{p_i}$ and $[\partial\mathbf{C}_i/\partial\rho]_{\hat{\boldsymbol{\alpha}}_0} = \mathbf{L}_i^{\top} + \mathbf{L}_i$, where $\mathbf{L}_i^{\top}$ is a $p_i \times p_i$ matrix with the entries of 1 on the first super- and sub-diagonals and 0 elsewhere. Applying

$$\boldsymbol{\Lambda}_i^{-1} = (\mathbf{Z}_i\boldsymbol{\Gamma}\mathbf{Z}_i^{\top} + \mathbf{C}_i)^{-1} = \mathbf{C}_i^{-1} - \mathbf{C}_i^{-1}\mathbf{Z}_i(\mathbf{Z}_i^{\top}\mathbf{C}_i^{-1}\mathbf{Z}_i + \boldsymbol{\Gamma}^{-1})^{-1}\mathbf{Z}_i^{\top}\mathbf{C}_i^{-1}$$

we obtain

$$\left[\operatorname{tr}\left(\boldsymbol{\Lambda}_i^{-1}\frac{\partial \mathbf{C}_i}{\partial \rho}\right)\right]_{\hat{\boldsymbol{\alpha}}_0} = -2\operatorname{tr}((\mathbf{Z}_i^{\top}\mathbf{Z}_i + \hat{\boldsymbol{\Gamma}}^{-1})^{-1}\mathbf{Z}_i^{\top}\mathbf{L}_i\mathbf{Z}_i) \tag{11}$$

The score statistic λ_s can be calculated, from (11), as

$$\lambda_s = \frac{\left(\sum_{i=1}^{N}\operatorname{tr}((\mathbf{Z}_i^{\top}\mathbf{Z}_i + \hat{\boldsymbol{\Gamma}}^{-1})^{-1}\mathbf{Z}_i^{\top}\mathbf{L}_i\mathbf{Z}_i) + \sum_{i=1}^{N}(\hat{v} + p_i)\dfrac{\hat{\mathbf{u}}_i^{\top}\mathbf{L}_i\hat{\mathbf{u}}_i}{\hat{\sigma}^2\hat{v} + \boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Gamma}}, 0)}\right)^2}{[\mathbf{I}_{\rho\rho\cdot\boldsymbol{\eta}}]_{\hat{\boldsymbol{\alpha}}_0}}$$

where $\boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Gamma}}, 0) = (\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\mathbf{b}}_i)^{\top}\hat{\mathbf{u}}_i$.

4. INFERENCES FOR RANDOM EFFECTS AND PREDICTION

We consider empirical Bayes estimates of the random effects, which are useful in explaining subject-specific deviations and helpful in predicting future measurements. If values of $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Gamma}, \rho, v)$ were known, the conditional mean of $\mathbf{b}_i$ given $\mathbf{Y}$ is

$$\hat{\mathbf{b}}_i(\boldsymbol{\alpha}) = [\mathbf{I}_{m_2} - \mathbf{W}_i(\mathbf{W}_i + \boldsymbol{\Gamma})^{-1}]\mathbf{b}_i^*(\boldsymbol{\alpha}) \tag{12}$$

where $\mathbf{W}_i = (\mathbf{Z}_i\mathbf{C}_i^{-1}\mathbf{Z}_i^{\top})^{-1}$ and $\mathbf{b}_i^*(\boldsymbol{\alpha}) = \mathbf{W}_i(\rho)\mathbf{Z}_i^{\top}\mathbf{C}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})$. The resulting error covariance matrix is

$$E((\hat{\mathbf{b}}_i(\boldsymbol{\alpha}) - \mathbf{b}_i)(\hat{\mathbf{b}}_i(\boldsymbol{\alpha}) - \mathbf{b}_i)^{\top}) = \frac{v\sigma^2}{v-2}[\mathbf{W}_i - \mathbf{W}_i(\mathbf{W}_i + \boldsymbol{\Gamma})^{-1}\mathbf{W}_i] \tag{13}$$

see Appendix C. As in Reference [11], substituting the ML estimate $\hat{\boldsymbol{\alpha}}$ into (12) leads to the empirical Bayes estimate $\hat{\mathbf{b}}_i = \hat{\mathbf{b}}_i(\hat{\boldsymbol{\alpha}})$.

Furthermore, we are interested in the prediction of $\mathbf{y}_i$, a future $q \times 1$ vector of measurements of $\mathbf{Y}_i$, given the observed measurements $\mathbf{Y} = (\mathbf{Y}_{(i)}^\top, \mathbf{Y}_i^\top)^\top$, where $\mathbf{Y}_{(i)} = (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_{i-1}^\top, \mathbf{Y}_{i+1}^\top, \ldots, \mathbf{Y}_N^\top)^\top$. Given $\boldsymbol{\alpha}$, the vectors $(\mathbf{Y}_i^\top, \mathbf{y}_i^\top)^\top$ $(i = 1, \ldots, N)$ are independent, so it is only necessary to consider the joint distribution of $\mathbf{Y}_i$ and $\mathbf{y}_i$ in predicting $\mathbf{y}_i$.

Let $\mathbf{x}_i$ and $\mathbf{z}_i$ denote $q \times m_1$ and $q \times m_2$ matrices of prediction regressors corresponding to $\mathbf{y}_i$. We thus have

$$\begin{bmatrix} \mathbf{Y}_i \\ \mathbf{y}_i \end{bmatrix} \sim t_{p_i+q}(\mathbf{X}_i^* \boldsymbol{\beta}, \sigma^2 \boldsymbol{\Omega}, v)$$

where $\mathbf{X}_i^* = (\mathbf{X}_i^\top, \mathbf{x}_i^\top)^\top$, $\mathbf{Z}_i^* = (\mathbf{Z}_i^\top, \mathbf{z}_i^\top)^\top$, $\boldsymbol{\Omega} = \mathbf{Z}_i^* \boldsymbol{\Gamma} \mathbf{Z}_i^{*\top} + \mathbf{C}_i^*$ and $\mathbf{C}_i^* = [\rho^{|r-s|}]$ for $r, s = 1, \ldots, p_i + q$. Let $\mathbf{C}_i^*$ and $\boldsymbol{\Omega}_i$ be partitioned conformably with $(\mathbf{Y}_i^\top, \mathbf{y}_i^\top)^\top$, i.e.

$$\mathbf{C}_i^* = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}, \quad \boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_i \boldsymbol{\Gamma} \mathbf{Z}_i^\top + \mathbf{C}_{11} & \mathbf{Z}_i \boldsymbol{\Gamma} \mathbf{z}_i^\top + \mathbf{C}_{12} \\ \mathbf{z}_i \boldsymbol{\Gamma} \mathbf{Z}_i^\top + \mathbf{C}_{21} & \mathbf{z}_i \boldsymbol{\Gamma} \mathbf{z}_i^\top + \mathbf{C}_{22} \end{bmatrix}$$

where $\mathbf{C}_{11} = \mathbf{C}_i$ and $\mathbf{C}_{21} = \mathbf{C}_{12}^\top$. The use of

$$f(\mathbf{y}_i | \mathbf{Y}_i) \propto \int f(\mathbf{y}_i | \mathbf{Y}_i, \tau_i) f(\mathbf{Y}_i | \tau_i) f(\tau_i) \, \mathrm{d}\tau_i$$

leads to

$$\mathbf{y}_i | \mathbf{Y}_i \sim t_q(\boldsymbol{\mu}_{i,2\cdot1}, \omega_i \boldsymbol{\Omega}_{22\cdot1}, v + p_i)$$

where $\boldsymbol{\mu}_{i,2\cdot1} = \mathbf{x}_i \boldsymbol{\beta} + \boldsymbol{\Omega}_{21} \boldsymbol{\Omega}_{11}^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})$, $\boldsymbol{\Omega}_{22\cdot1} = \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21} \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\Omega}_{12}$ and $\omega_i = (\sigma^2 v + (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})^\top \boldsymbol{\Omega}_{11}^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}))/(v + p_i)$. The minimum MSE predictor of $\mathbf{y}_i$ is the conditional expectation of $\mathbf{y}_i$ given $\mathbf{Y}_i$, i.e.

$$\begin{aligned} \hat{\mathbf{y}}_i(\boldsymbol{\alpha}) &= \mathbf{x}_i \boldsymbol{\beta} + \boldsymbol{\Omega}_{21} \boldsymbol{\Omega}_{11}^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}) \\ &= \mathbf{x}_i \boldsymbol{\beta} + \mathbf{z}_i \hat{\mathbf{b}}_i(\boldsymbol{\alpha}) + \mathbf{C}_{21} \mathbf{C}_{11}^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{Z}_i \hat{\mathbf{b}}_i(\boldsymbol{\alpha})) \end{aligned} \quad (14)$$

where $\hat{\mathbf{b}}_i(\boldsymbol{\alpha})$ is given in (12). Similarly, the error covariance matrix for the predictor (14) is given by

$$\mathrm{E}((\hat{\mathbf{y}}_i(\boldsymbol{\alpha}) - \mathbf{y}_i)(\hat{\mathbf{y}}_i(\boldsymbol{\alpha}) - \mathbf{y}_i)^\top) = \frac{v + p_i}{v + p_i - 2} \omega_i \boldsymbol{\Omega}_{22\cdot1}$$

where $\boldsymbol{\Omega}_{22\cdot1}$ can be rewritten as

$$\boldsymbol{\Omega}_{22\cdot1} = \mathbf{C}_{22\cdot1} + (\mathbf{z}_i - \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{Z}_i)(\mathbf{W}_{11} - \mathbf{W}_{11}(\boldsymbol{\Gamma} + \mathbf{W}_{11})^{-1} \mathbf{W}_{11})(\mathbf{z}_i - \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{Z}_i)^\top \quad (15)$$

with $\mathbf{C}_{22\cdot1} = \mathbf{C}_{22} - \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{C}_{12}$ and $\mathbf{W}_{11} = (\mathbf{Z}_i^\top \mathbf{C}_{11}^{-1} \mathbf{Z}_i)^{-1}$; see Appendix D. The prediction of $\mathbf{y}_i$ is obtained by substituting the ML estimate $\hat{\boldsymbol{\alpha}}$ into (14), leading to $\hat{\mathbf{y}} = \hat{\mathbf{y}}(\hat{\boldsymbol{\alpha}})$.

5. EXAMPLE

To illustrate the methodology, we next analyse the MS data. More details of the clinical trial leading to the MS data is described by D'yachkova *et al.* [12]. The PL, LD and HD treatment groups involve 17, 18, and 17 patients, respectively. Of the 52 patients, 3 patients were not included in the analysis since two of them (one in each of groups LD and HD) dropped out very early and one in group LD had 3 measurements of zero on MRI scans. All but 5 of the remaining patients have a complete set of 17 scans: one dropped out from PL after completing 14 visits, two dropped out from LD after completing 13 visits, and two dropped out from HD after completing 12 visits. We will analyse the data of the 49 patients. Our analyses will assume these early dropouts are *ignorable* [13]. Of these 49 patients, six patients have one or two isolated MRI scans missing. We impute these missing observations using the mean of two adjacent values as in Reference [1].

The disease burden, the total area (in mm^2) of MS lesions on all slices of an MRI scan, for the ith patient at time point j is denoted by $\text{Area}(i, j)$, with $j = 0$ as the baseline time point. The corresponding log relative burden (LRB) is $\text{LRB}(i, j) = \log(\text{Area}(i, j)/\text{Area}(i, 0))$, which is used as the response variable Y_{ij} due to strong skewness of the untransformed burden measurements. Figure 1 depicts the LRB evolution of the 49 patients from various groups. Apparently, the MS data involve many outlying observations, especially for PL and LD.

We model the average evolution of LRB as a linear function of time and carry out the analyses separately for the 3 treatment groups. For the fixed effects, we set $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ with the corresponding design matrix $\mathbf{X}_i = [\mathbf{1}_{p_i} \;\; \mathbf{k}_i]$, where $\mathbf{1}_{p_i} = (1, 1, \ldots, 1)^\top$ and $\mathbf{k}_i = (1, 2, \ldots, p_i)^\top$. To explore the autocorrelation among the within-subject errors, we start by fitting model (1) with random intercepts ($\mathbf{Z}_i = \mathbf{1}_{p_i}$) and white noise errors ($\rho = 0$), denoted by $\mathcal{M}_1$.

Table I lists the resulting ML estimates with their standard errors in parentheses obtained from (5), Akaike's Information Criteria ($\text{AIC} = -2 \times \text{maximized log-likelihood} + 2m$, where m is the number of model parameters), along with the values of the score test statistic λ_s for the three groups. The score tests are all highly significant compared with the χ^2_1 distribution, indicating that there exists autocorrelation among the within-subjects errors. Moreover, the estimates of the d.f. v's are quite small, which justify the modelling effort via the t distribution.

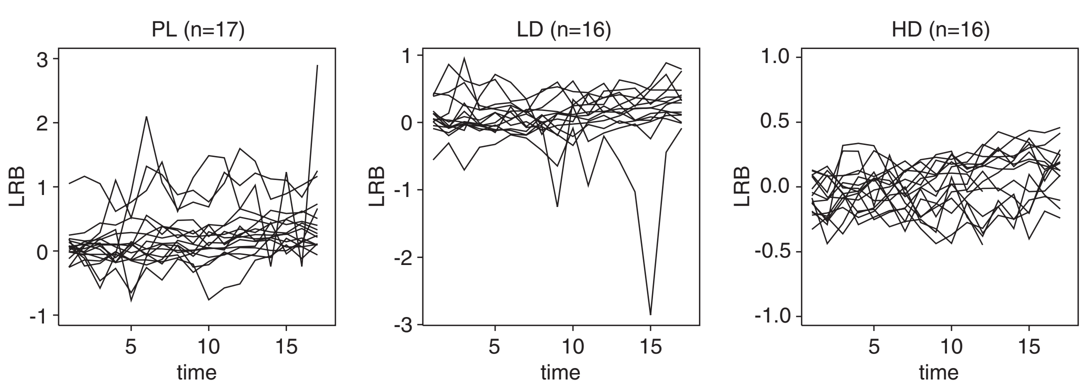

Figure 1. Longitudinal trends in LRB of 49 patients from three treatment groups.

Table I. ML estimation results and score test statistics for $\mathscr{M}_1$, where f is such that $\Gamma = f^2$.

Group	β_0	β_1	σ^2	f	v	AIC	λ_s
PL	−0.0076 (0.0460)	0.0185 (0.0015)	0.0143 (0.0042)	1.8290 (0.272)	1.82 (0.59)	−99.02	16.91
LD	−0.0048 (0.0325)	0.0122 (0.0015)	0.0132 (0.0039)	0.9462 (0.194)	2.08 (0.71)	−145.58	21.30
HD	−0.0846 (0.0356)	0.0143 (0.0015)	0.0128 (0.0027)	1.0856 (0.2185)	4.58 (1.87)	−251.44	49.58

Table II. ML estimation results for $\mathscr{M}_2$, where f is such that $\Gamma = f^2$.

Group	β_0	β_1	σ^2	f	ρ	v	AIC
PL	−0.0036 (0.0476)	0.0176 (0.0020)	0.0154 (0.0046)	1.8290 (0.7322)	0.3164 (0.0685)	1.81 (0.59)	−116.26
LD	−0.0040 (0.0343)	0.0125 (0.0021)	0.0137 (0.0043)	0.8734 (0.1994)	0.3584 (0.0712)	1.97 (0.67)	−167.10
HD	−0.0814 (0.0400)	0.0137 (0.0027)	0.0174 (0.0040)	0.8348 (0.2192)	0.5587 (0.0689)	6.79 (3.15)	−310.38

As a note in passing, the REML estimates are somewhat similar to the ML estimates and hence are omitted for the rest of the paper.

We further fit an alternate model with random intercepts and an AR(1) dependence structure for the within-subject errors, denoted by $\mathscr{M}_2$. The corresponding ML estimation results and AICs are shown in Table II. Based on smaller AIC values, $\mathscr{M}_2$ is preferred to $\mathscr{M}_1$. Figure 2 displays the profile likelihood function of ρ and f, where f satisfies $\Gamma = f^2$. Obviously, all three plots are unimodal and exhibit significant serial correlations, indicating that the AR(1) model is a possible dependence structure for modelling the MS data.

In some cases, however, a more general random effects model may be useful for the interpretation of autocorrelation among the within-subjects errors. For comparison purposes, we fit a model, denoted by $\mathscr{M}_3$, with random intercepts and slopes ($\mathbf{Z}_i = \mathbf{X}_i$) and white noise errors. The associated ML estimation results are given in Table III. The AIC values did not improve over $\mathscr{M}_2$. In addition, all elements of F_{12} and F_{22} are relatively small (compared with their standard errors) with the exception of F_{22} for HD, indicating negligible variation in slopes except for HD. Based on this, we check the expanded model that combines the random intercepts, random slopes and an AR(1) dependence structure for HD. The value of AIC is −309.13, slightly higher than AIC = − 310.38 for $\mathscr{M}_2$. Thus, among the models considered,

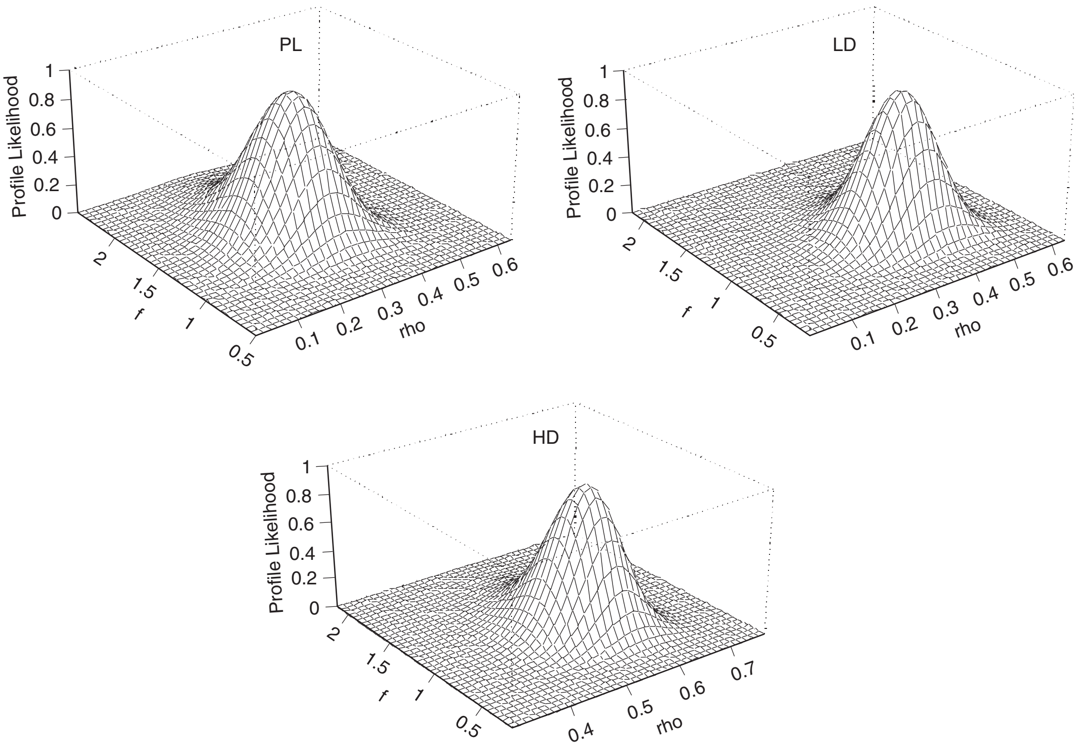

Figure 2. Profile likelihood functions of (ρ, f) from three treatment groups.

Table III. ML estimation results for $\mathscr{M}_3$, where $\mathbf{F}$ is the Cholesky decomposition of $\mathbf{\Gamma}$.

Group	β_0	β_1	σ^2	F_{11}, F_{12}, F_{22}	v	AIC
PL	−0.0139	0.0190	0.0129	1.1255, 0.0352, 0.0480	1.76	−108.20
	(0.0358)	(0.0023)	(0.0038)	(0.2629, 0.1022, 0.0323)	(0.57)	
LD	−0.0070	0.0125	0.0114	1.0115, −0.0264, 0.0663	1.97	−153.68
	(0.0319)	(0.0025)	(0.0040)	(0.5310, 0.2712, 0.0344)	(0.67)	
HD	−0.0857	0.0140	0.0113	0.9412, −0.0136, 0.0989	5.49	−273.26
	(0.0299)	(0.0032)	(0.0023)	(0.2665, 0.1216, 0.0253)	(2.39)	

$\mathscr{M}_2$ is our preferred model for the MS data since it has the smallest AIC. Furthermore, it incorporates the autocorrelation and is parsimonious.

Based on the analysis so far, we found that the ML estimates of the d.f. v's for the MS data are all relatively small, especially for PL and LD. To assess further the adequacy of normal

Table IV. ML estimation results for $\mathscr{M}_4$, where f is such that $\Gamma = f^2$.

Group	β_0	β_1	σ^2	f	ρ	AIC
PL	0.0368	0.0244	0.0826	1.1329	0.2208	141.34
	(0.0897)	(0.0042)	(0.0078)	(0.2163)	(0.0651)	
LD	0.0326	0.0089	0.0619	0.8953	0.2717	40.26
	(0.0685)	(0.0039)	(0.0063)	(0.1868)	(0.0682)	
HD	−0.0040	0.0125	0.0137	0.8734	0.3584	−299.30
	(0.0416)	(0.0030)	(0.0034)	(0.1945)	(0.0642)	

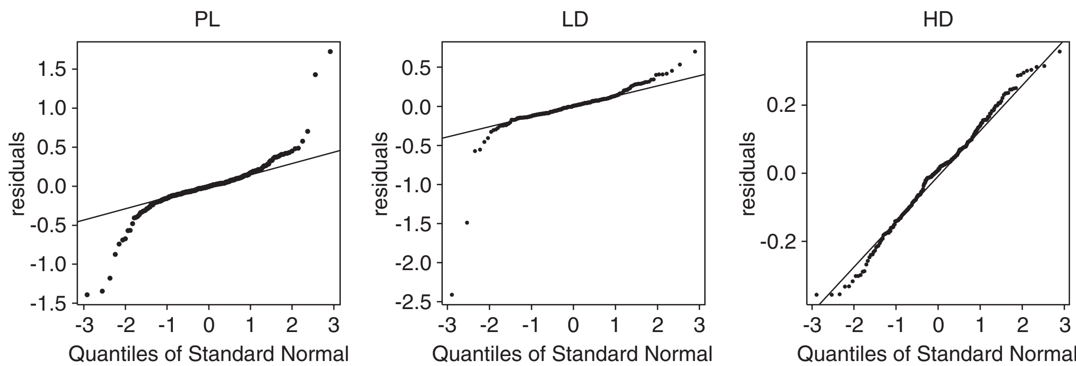

Figure 3. Normal quantile plots for residuals from fitting $\mathscr{M}_4$.

modelling for the MS data, we fit a normal version of $\mathscr{M}_2$ (obtained by setting $v = \infty$), denoted by $\mathscr{M}_4$. The ML estimation results and AICs are reported in Table IV. Compared to Table II, we found that the fixed effects are rather different and have consistently larger standard errors. The substantially larger AICs indicate that $\mathscr{M}_4$ is not suitable for the PL and LD groups, although the AICs are comparable for the HD group.

Figure 3 displays the corresponding normal quantile plots for the residuals from $\mathscr{M}_4$. Obviously, the residuals of PL and LD seriously deviate from normality, confirming the presence of longer-than-normal tails. In contrast, the departure from normality for HD is minor. Based on these findings, it appears that $\mathscr{M}_4$ might be adequate for HD.

We next compare the prediction accuracy of $\mathscr{M}_2$ and $\mathscr{M}_4$. We use the predictive sample reuse procedure of Geisser [14] sequentially by removing the last few points of each response vector as the true values to be predicted. As a measure of precision we use the MARD, which is defined as the mean of absolute relative deviations $|(y_{jp} - \hat{y}_{jp})/y_{jp}|$, where p is the time point being forcast.

We restrict our attention to the one-step-ahead forecasts by setting $p = 13$–17. To predict Y_{ip}, we use $Y_{i1}, \ldots, Y_{i,p-1}$ and $\mathbf{Y}_{(i)}$ as the sample to obtain the ML estimate $\hat{\boldsymbol{\alpha}}$ and plug it into

Table V. Comparison of one-step-ahead forecast accuracy in terms of MARD.

Group	The time point being forecast	$\mathscr{M}_4$	$\mathscr{M}_2$
PL	13	0.532	0.414
	14	1.186	1.020
	15	1.825	1.610
	16	0.818	0.702
	17	0.717	0.675
	Average	1.016	0.884
LD	13	0.528	0.504
	14	1.044	1.010
	15	0.580	0.591
	16	0.716	0.484
	17	0.302	0.287
	Average	0.634	0.575
HD	13	0.803	0.817
	14	0.245	0.220
	15	0.414	0.382
	16	0.765	0.741
	17	0.412	0.397
	Average	0.528	0.511
	Overall average	0.726	0.657

the predictor (14). Since the predictions are done sequentially for each subject and for each p, the procedure is termed *prequential* by Dawid [15]. Table V shows prediction results from these two models. The t-based model has a much smaller MARD than the normal model for PL and LD; the relative improvement percentages are 13 and 9.3 per cent, respectively. On the contrary, the prediction performance for HD is only slightly better (3.2 per cent). The t-based model not only provides better model fitting, it also yields smaller forecast errors for the MS data; the overall improvement is 9.5 per cent.

6. CONCLUDING REMARKS

There are rather extensive approaches to robustifying linear mixed models, see for example, References [16–19]. The t linear mixed model provides an alternative robust way of dealing with longitudinal data when some outlying observations are present. Moreover, the explicit derivative-based estimation and score testing procedures developed in this paper can be easily implemented with low computational burden.

As shown by Lee [20] and Chi and Reinsel [21], inclusion of the simple AR(1) dependence could lead to an appropriate representation of dependence structure and may reduce the need for including complex random effects in the model. From our illustrated MS data it is encouraging that the use of t linear mixed model coupled with AR(1) structure offers better fitting as well as better prediction performance than the normal counterpart. It may be worthwhile comparing with other dependence structures, such as higher order AR, MA or ARMA.

Finally, as one referee pointed out, the current model does not allow simultaneous fitting for all three treatment groups of the MS data set and it may be of interest to extend the model in that direction.

APPENDIX A: THE SCORE FUNCTION AND FISHER INFORMATION MATRIX

The score vector $\mathbf{s}_\alpha$ is the vector of the first derivatives of (4) with respect to $\boldsymbol{\alpha}=(\boldsymbol{\beta},\sigma^2,v,\boldsymbol{\omega})$, where $\boldsymbol{\omega}=(\mathrm{vech}(\mathbf{F}),\rho)$ and $\mathbf{F}$ is the Cholesky decomposition of $\boldsymbol{\Gamma}$

$$\mathbf{s}_{\boldsymbol{\beta}} = \sum_{i=1}^{N}(v+p_i)\frac{\mathbf{X}_i^\top \boldsymbol{\Lambda}_i^{-1}\mathbf{e}_i}{\sigma^2 v+\boldsymbol{\Delta}_i(\boldsymbol{\beta},\boldsymbol{\Gamma},\rho)}$$

$$\mathbf{s}_{\sigma^2} = -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^2}\sum_{i=1}^{N}(v+p_i)\left(\frac{\boldsymbol{\Delta}_i(\boldsymbol{\beta},\boldsymbol{\Gamma},\rho)}{\sigma^2 v+\boldsymbol{\Delta}_i(\boldsymbol{\beta},\boldsymbol{\Gamma},\rho)}\right)$$

$$\mathbf{s}_v = \frac{1}{2}\sum_{i=1}^{N}\left\{\phi\left(\frac{v+p_i}{2}\right)-\phi\left(\frac{v}{2}\right)-\frac{p_i}{v}-\log\left(1+\frac{\boldsymbol{\Delta}_i(\boldsymbol{\beta},\boldsymbol{\Gamma},\rho)}{\sigma^2 v}\right)\right.$$
$$\left.+\frac{(v+p_i)}{v}\,\frac{\boldsymbol{\Delta}_i(\boldsymbol{\beta},\boldsymbol{\Gamma},\rho)}{\sigma^2 v+\boldsymbol{\Delta}_i(\boldsymbol{\beta},\boldsymbol{\Gamma},\rho)}\right\}$$

$$[\mathbf{s}_{\boldsymbol{\omega}}]_r = -\frac{1}{2}\sum_{i=1}^{N}\left\{\mathrm{tr}(\boldsymbol{\Lambda}_i^{-1}\dot{\boldsymbol{\Lambda}}_{ir})-(v+p_i)\left(\frac{\mathbf{e}_i^\top\boldsymbol{\Lambda}_i^{-1}\dot{\boldsymbol{\Lambda}}_{ir}\boldsymbol{\Lambda}_i^{-1}\mathbf{e}_i}{\sigma^2 v+\boldsymbol{\Delta}_i(\boldsymbol{\beta},\boldsymbol{\Gamma},\rho)}\right)\right\}$$

where $\dot{\boldsymbol{\Lambda}}_{ir}=\partial\boldsymbol{\Lambda}_i/\partial\omega_r$, for $r=1,\ldots,g$; $g=(m_2^2+m_2+2)/2$ and $\phi(x)=\frac{\mathrm{d}}{\mathrm{d}x}\log(\Gamma(x))$ denotes the digamma function.

The Fisher information, obtained by the negative expectation of the second derivative of (4), has the following components:

$$\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sum_{i=1}^{N}\frac{v+p_i}{\sigma^2(v+p_i+2)}\mathbf{X}_i^\top\boldsymbol{\Lambda}_i^{-1}\mathbf{X}_i,\quad \mathbf{I}_{\boldsymbol{\beta}\sigma}=\mathbf{0}_{m_1\times 1},\quad \mathbf{I}_{\boldsymbol{\beta}v}=\mathbf{0}_{m_1\times 1},\quad \mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\omega}}=\mathbf{0}_{m_1\times g}$$

$$\mathbf{I}_{\sigma^2\sigma^2} = \frac{v}{2\sigma^4}\sum_{i=1}^{N}\frac{p_i}{v+p_i+2},\quad \mathbf{I}_{\sigma^2 v}=\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(\frac{p_i}{v+p_i+2}-\frac{p_i}{v+p_i}\right)$$

$$[\mathbf{I}_{\sigma^2\boldsymbol{\omega}}]_r = \frac{v}{2\sigma^2}\sum_{i=1}^{N}\frac{1}{v+p_i+2}\,\mathrm{tr}(\boldsymbol{\Lambda}_i^{-1}\dot{\boldsymbol{\Lambda}}_{ir})$$

$$\mathbf{I}_{vv} = \frac{1}{4}\sum_{i=1}^{N}\left\{\psi\left(\frac{v}{2}\right)-\psi\left(\frac{v+p_i}{2}\right)-\frac{2(v+2)}{v(v+p_i+2)}-\frac{2}{v}+\frac{4}{v+p_i}\right\}$$

$$[\mathbf{I}_{v\boldsymbol{\omega}}]_r = -\sum_{i=1}^{N} \frac{1}{(v+p_i)(v+p_i+2)} \operatorname{tr}(\boldsymbol{\Lambda}_i^{-1}\dot{\boldsymbol{\Lambda}}_{ir})$$

$$[\mathbf{I}_{\boldsymbol{\omega}\boldsymbol{\omega}}]_{rs} = \frac{1}{2}\sum_{i=1}^{N} \frac{1}{v+p_i+2}\{(v+p_i)\operatorname{tr}(\boldsymbol{\Lambda}_i^{-1}\dot{\boldsymbol{\Lambda}}_{ir}\boldsymbol{\Lambda}_i^{-1}\dot{\boldsymbol{\Lambda}}_{is}) - \operatorname{tr}(\boldsymbol{\Lambda}_i^{-1}\dot{\boldsymbol{\Lambda}}_{ir})\operatorname{tr}(\boldsymbol{\Lambda}_i^{-1}\dot{\boldsymbol{\Lambda}}_{is})\}$$

for $r, s = 1, \ldots, g$, where $\psi(x) = \frac{d^2}{dx^2}\log(\Gamma(x))$ denotes the trigamma function.

APPENDIX B: THE FIRST PARTIAL DERIVATIVES OF ℓ_R^*

The first partial derivatives of ℓ_R^* with respect to σ^2, v, and $\boldsymbol{\omega}$ are

$$\begin{aligned}
\frac{\partial \ell_R^*}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} - \frac{1}{2}\operatorname{tr}\left\{\left(\sum_{i=1}^{N}\mathbf{X}_i^\top\mathbf{H}_i\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i^\top\frac{\partial \mathbf{H}_i}{\partial \sigma^2}\mathbf{X}_i\right)\right\} \\
&\quad + \frac{1}{2\sigma^2}\sum_{i=1}^{N}\frac{(v+p_i)\boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\Gamma},\rho)}{\sigma^2 v + \boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\Gamma},\rho)} \\
\frac{\partial \ell_R^*}{\partial v} &= -\frac{1}{2}\operatorname{tr}\left\{\left(\sum_{i=1}^{N}\mathbf{X}_i^\top\mathbf{H}_i\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i^\top\frac{\partial \mathbf{H}_i}{\partial v}\mathbf{X}_i\right)\right\} \\
&\quad + \frac{1}{2}\sum_{i=1}^{N}\left\{\phi\left(\frac{v+p_i}{2}\right) - \phi\left(\frac{v}{2}\right) - \frac{1}{v} - \log\left(1 + \frac{\boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\Gamma},\rho)}{\sigma^2 v}\right)\right. \\
&\quad \left. + \frac{(v+p_i)}{v}\frac{\boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\Gamma},\rho)}{\sigma^2 v + \boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\Gamma},\rho)}\right\} \\
\left[\frac{\partial \ell_R^*}{\partial \omega}\right]_r &= -\frac{1}{2}\operatorname{tr}\left\{\left(\sum_{i=1}^{N}\mathbf{X}_i^\top\mathbf{H}_i\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i^\top\frac{\partial \mathbf{H}_i}{\partial \omega_r}\mathbf{X}_i\right)\right\} \\
&\quad - \frac{1}{2}\sum_{i=1}^{N}\left\{\operatorname{tr}(\boldsymbol{\Lambda}_i^{-1}\dot{\boldsymbol{\Lambda}}_{ir}^{-1}) - (v+p_i)\left(\frac{\hat{\mathbf{e}}_i^\top(\boldsymbol{\theta})\boldsymbol{\Lambda}_i^{-1}\dot{\boldsymbol{\Lambda}}_{ir}^{-1}\boldsymbol{\Lambda}_i^{-1}\hat{\mathbf{e}}_i(\boldsymbol{\theta})}{\sigma^2 v + \boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\Gamma},\rho)}\right)\right\}
\end{aligned}$$

where

$$\frac{\partial \mathbf{H}_i}{\partial \sigma^2} = -v(v+p_i)\left\{\frac{\boldsymbol{\Lambda}_i^{-1}}{(\sigma^2 v + \boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\Gamma},\rho))^2} - \frac{4\boldsymbol{\Lambda}_i^{-1}\hat{\mathbf{e}}_i(\boldsymbol{\theta})\hat{\mathbf{e}}_i^\top(\boldsymbol{\theta})\boldsymbol{\Lambda}_i^{-1}}{(\sigma^2 v + \boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\Gamma},\rho))^3}\right\}$$

$$\frac{\partial \mathbf{H}_i}{\partial v} = \frac{\boldsymbol{\Lambda}_i^{-1}}{\sigma^2 v + \boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\Gamma},\rho)} - \frac{2\boldsymbol{\Lambda}_i^{-1}\hat{\mathbf{e}}_i(\boldsymbol{\theta})\hat{\mathbf{e}}_i^\top(\boldsymbol{\theta})\boldsymbol{\Lambda}_i^{-1}}{(\sigma^2 v + \boldsymbol{\Delta}_i(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\Gamma},\rho))^2}$$

$$-\sigma^2(v+p_i)\left\{\frac{\mathbf{\Lambda}_i^{-1}}{(\sigma^2 v+\mathbf{\Delta}_i(\hat{\boldsymbol\beta}(\boldsymbol\theta),\mathbf{\Gamma},\rho))^2}-\frac{4\mathbf{\Lambda}_i^{-1}\hat{\mathbf{e}}_i(\boldsymbol\theta)\hat{\mathbf{e}}_i^\top(\boldsymbol\theta)\mathbf{\Lambda}_i^{-1}}{(\sigma^2 v+\mathbf{\Delta}_i(\hat{\boldsymbol\beta}(\boldsymbol\theta),\mathbf{\Gamma},\rho))^3}\right\}$$

$$\begin{aligned}\frac{\partial \mathbf{H}_i}{\partial\omega_r}=(v+p_i)\Bigg\{&\frac{-\mathbf{\Lambda}_i^{-1}\dot{\mathbf{\Lambda}}_{ir}\mathbf{\Lambda}_i^{-1}}{\sigma^2 v+\mathbf{\Delta}_i(\hat{\boldsymbol\beta}(\boldsymbol\theta),\mathbf{\Gamma},\rho)}+\frac{\mathbf{\Lambda}_i^{-1}\hat{\mathbf{e}}_i^\top(\boldsymbol\theta)\mathbf{\Lambda}_i^{-1}\dot{\mathbf{\Lambda}}_{ir}\mathbf{\Lambda}_i^{-1}\hat{\mathbf{e}}_i(\boldsymbol\theta)}{(\sigma^2 v+\mathbf{\Delta}_i(\hat{\boldsymbol\beta}(\boldsymbol\theta),\mathbf{\Gamma},\rho))^2}\\&+\frac{2\mathbf{\Lambda}_i^{-1}\dot{\mathbf{\Lambda}}_{ir}\mathbf{\Lambda}_i^{-1}\hat{\mathbf{e}}_i(\boldsymbol\theta)\hat{\mathbf{e}}_i^\top(\boldsymbol\theta)\mathbf{\Lambda}_i^{-1}+2\mathbf{\Lambda}_i^{-1}\hat{\mathbf{e}}_i(\boldsymbol\theta)\hat{\mathbf{e}}_i^\top(\boldsymbol\theta)\mathbf{\Lambda}_i^{-1}\dot{\mathbf{\Lambda}}_{ir}\mathbf{\Lambda}_i^{-1}}{(\sigma^2 v+\mathbf{\Delta}_i(\hat{\boldsymbol\beta}(\boldsymbol\theta),\mathbf{\Gamma},\rho))^2}\\&-\frac{4\mathbf{\Lambda}_i^{-1}\hat{\mathbf{e}}_i(\boldsymbol\theta)\hat{\mathbf{e}}_i^\top(\boldsymbol\theta)\mathbf{\Lambda}_i^{-1}\hat{\mathbf{e}}_i^\top(\boldsymbol\theta)\mathbf{\Lambda}_i^{-1}\dot{\mathbf{\Lambda}}_{ir}\mathbf{\Lambda}_i^{-1}\hat{\mathbf{e}}_i(\boldsymbol\theta)}{(\sigma^2 v+\mathbf{\Delta}_i(\hat{\boldsymbol\beta}(\boldsymbol\theta),\mathbf{\Gamma},\rho))^3}\Bigg\}\end{aligned}$$

APPENDIX C: PROOFS OF (12) AND (13)

Recall that $\mathbf{\Lambda}_i=\mathbf{Z}_i\mathbf{\Gamma}\mathbf{Z}_i^\top+\mathbf{C}_i$. With some algebraic manipulations, we can get

$$\mathbf{\Lambda}_i^{-1}=\mathbf{C}_i^{-1}-\mathbf{C}_i^{-1}\mathbf{Z}_i\mathbf{\Gamma}(\mathbf{W}_i+\mathbf{\Gamma})^{-1}(\mathbf{Z}_i^\top\mathbf{C}_i^{-1}\mathbf{Z}_i)^{-1}\mathbf{Z}_i^\top\mathbf{C}_i^{-1}$$

It follows that

$$\begin{aligned}\hat{\mathbf{b}}_i(\boldsymbol\alpha)&=\mathbf{\Gamma}\mathbf{Z}_i^\top\mathbf{\Lambda}_i^{-1}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol\beta)\\&=\mathbf{\Gamma}\mathbf{Z}_i^\top\mathbf{C}_i^{-1}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol\beta)\\&\quad-\mathbf{\Gamma}\mathbf{Z}_i^\top\mathbf{C}_i^{-1}\mathbf{Z}_i\mathbf{\Gamma}(\mathbf{W}_i+\mathbf{\Gamma})^{-1}(\mathbf{Z}_i^\top\mathbf{C}_i^{-1}\mathbf{Z}_i)^{-1}\mathbf{Z}_i^\top\mathbf{C}_i^{-1}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol\beta)\mathbf{Z}_i^\top\mathbf{C}_i^{-1}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol\beta)\\&=\mathbf{\Gamma}\mathbf{W}_i^{-1}\mathbf{b}_i^*(\boldsymbol\alpha)-\mathbf{\Gamma}\mathbf{W}_i^{-1}\mathbf{\Gamma}(\mathbf{W}_i+\mathbf{\Gamma})^{-1}\mathbf{b}_i^*(\boldsymbol\alpha)\\&=\mathbf{b}_i^*(\boldsymbol\alpha)-\mathbf{W}_i(\mathbf{W}_i+\mathbf{\Gamma})^{-1}\mathbf{b}_i^*(\boldsymbol\alpha)\end{aligned}$$

where $\mathbf{b}_i^*(\boldsymbol\alpha)=(\mathbf{Z}_i^\top\mathbf{C}_i^{-1}\mathbf{Z}_i^\top)^{-1}\mathbf{Z}_i^\top\mathbf{C}_i^{-1}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol\beta)$, and equation (12) holds.

The error covariance matrix of $\hat{\mathbf{b}}_i(\boldsymbol\theta)$ is

$$\begin{aligned}&E((\hat{\mathbf{b}}_i(\boldsymbol\alpha)-\mathbf{b}_i)(\hat{\mathbf{b}}_i(\boldsymbol\alpha)-\mathbf{b}_i)^\top)\\&\quad=\frac{v}{v-2}\sigma^2\mathbf{\Gamma}+\mathbf{\Gamma}\mathbf{Z}_i^\top\mathbf{\Lambda}_i^{-1}\left(\frac{v}{v-2}\sigma^2\mathbf{\Lambda}_i\right)\mathbf{\Lambda}_i^{-1}\mathbf{Z}_i\mathbf{\Gamma}-2\mathbf{\Gamma}\mathbf{Z}_i^\top\mathbf{\Lambda}_i^{-1}\mathbf{Z}_i\left(\frac{v}{v-2}\sigma^2\right)\mathbf{\Gamma}\\&\quad=\frac{v}{v-2}\sigma^2(\mathbf{\Gamma}-\mathbf{\Gamma}\mathbf{Z}_i^\top(\mathbf{C}_i^{-1}-\mathbf{C}_i^{-1}\mathbf{Z}_i\mathbf{\Gamma}(\mathbf{W}_i+\mathbf{\Gamma})^{-1}(\mathbf{Z}_i^\top\mathbf{C}_i^{-1}\mathbf{Z}_i)^{-1}\mathbf{Z}_i^\top\mathbf{C}_i^{-1})\mathbf{Z}_i\mathbf{\Gamma})\\&\quad=\frac{v}{v-2}\sigma^2(\mathbf{W}_i-\mathbf{W}_i(\mathbf{W}_i+\mathbf{\Gamma})^{-1}\mathbf{W}_i)\end{aligned}$$

APPENDIX D: PROOF OF (15)

Recall that $\mathbf{\Omega}_{11} = \mathbf{Z}_i \mathbf{\Gamma} \mathbf{Z}_i^\top + \mathbf{C}_{11}$, then

$$\mathbf{\Omega}_{11}^{-1} = \mathbf{C}_{11}^{-1} - \mathbf{C}_{11}^{-1}\mathbf{Z}_i(\mathbf{W}_{11} - \mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11})\mathbf{Z}_i^\top \mathbf{C}_{11}^{-1}$$

It suffices to show that

$$\begin{aligned}
\mathbf{\Omega}_{22\cdot 1} &= \mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12} \\
&= \mathbf{z}_i\mathbf{\Gamma}\mathbf{z}_i^\top + \mathbf{C}_{22} - (\mathbf{z}_i\mathbf{\Gamma}\mathbf{Z}_i^\top + \mathbf{C}_{21}) \\
&\quad \times (\mathbf{C}_{11}^{-1} - \mathbf{C}_{11}^{-1}\mathbf{Z}_i(\mathbf{W}_{11} - \mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11})\mathbf{Z}_i^\top \mathbf{C}_{11}^{-1})(\mathbf{Z}_i\mathbf{\Gamma}\mathbf{z}_i^\top + \mathbf{C}_{12}) \\
&= \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12} + \mathbf{z}_i\mathbf{\Gamma}\mathbf{z}_i^\top - \mathbf{z}_i\mathbf{\Gamma}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{\Gamma}\mathbf{z}_i^\top \\
&\quad - \mathbf{z}_i\mathbf{\Gamma}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11}\mathbf{Z}_i^\top \mathbf{C}_{11}^{-1}\mathbf{C}_{12} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}_i\mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{\Gamma}\mathbf{z}_i^\top \\
&\quad + \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}_i(\mathbf{W}_{11} - \mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11})\mathbf{Z}_i^\top \mathbf{C}_{11}^{-1}\mathbf{C}_{12}
\end{aligned}$$

Since

$$\begin{aligned}
\mathbf{z}_i\mathbf{\Gamma}\mathbf{z}_i^\top - \mathbf{z}_i\mathbf{\Gamma}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{\Gamma}\mathbf{z}_i^\top &= \mathbf{z}_i(\mathbf{W}_{11} - \mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11})\mathbf{z}_i^\top \\
\mathbf{z}_i\mathbf{\Gamma}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11}\mathbf{Z}_i^\top \mathbf{C}_{11}^{-1}\mathbf{C}_{12} &= \mathbf{z}_i(\mathbf{W}_{11} - \mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11})(\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}_i)^\top
\end{aligned}$$

and

$$\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}_i\mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{\Gamma}\mathbf{z}_i^\top = \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}_i(\mathbf{W}_{11} - \mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11})\mathbf{z}_i^\top$$

we have

$$\begin{aligned}
\mathbf{\Omega}_{22\cdot 1} &= \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12} + \mathbf{z}_i(\mathbf{W}_{11} - \mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11})\mathbf{z}_i^\top \\
&\quad - \mathbf{z}_i(\mathbf{W}_{11} - \mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11})(\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}_i)^\top \\
&\quad - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}_i(\mathbf{W}_{11} - \mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11})\mathbf{z}_i^\top \\
&\quad + \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}_i(\mathbf{W}_{11} - \mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11})\mathbf{Z}_i^\top \mathbf{C}_{11}^{-1}\mathbf{C}_{12} \\
&= \mathbf{C}_{22\cdot 1} + (\mathbf{z}_i - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}_i)(\mathbf{W}_{11} - \mathbf{W}_{11}(\mathbf{\Gamma} + \mathbf{W}_{11})^{-1}\mathbf{W}_{11})(\mathbf{z}_i - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}_i)^\top
\end{aligned}$$

This completes the proof.

ACKNOWLEDGEMENTS

The authors thank Prof. Paramjit S. Gill for providing the MS data at an early stage of this work. We also thank the editor and two anonymous referees for their helpful comments and suggestions that substantially improved the presentation. This research was supported by the National Science Council of Taiwan (Grant No. NSC93-2118-M-029-004 and NSC93-2118-M-009-003).

REFERENCES

1. Gill PS. A robust mixed linear model analysis for longitudinal data. *Statistics in Medicine* 2000; **19**:975–987.
2. West M. Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society*, *Series B* 1984; **46**:431–438.
3. Lange KL, Little RJA, Taylor JMG. Robust statistical modelling using the t distribution. *Journal of the American Statistical Association* 1989; **84**:881–896.
4. James AT, Wiskich JT, Conyers RAJ. t-REML for robust heteroscedastic regression analysis of mitochondrial power. *Biometrics* 1993; **49**:339–356.
5. Welsh AH, Richardson AM. Approaches to the robust estimation of mixed models. In *Handbooks of Statistics*, Maddlal GS, Rao CR (eds). vol. 15. Elsevier Science: Amsterdam, 1997; 343–383.
6. Pinheiro JC, Liu CH, Wu YN. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* 2001; **10**:249–276.
7. Rochon J. ARMA covariance structures with time heteroscedasticity for repeated measures experiments. *Journal of the American Statistical Association* 1992; **87**:777–784.
8. Lin TI, Lee JC. On modelling data from degradation sample paths over time. *Australian and New Zealand Journal of Statistics* 2003; **45**:257–270.
9. Lee JC, Lin TI, Lee KJ, Hsu YL. Bayesian analysis of Box-Cox transformed linear mixed models with ARMA(p,q) dependence. *Journal of Statistical Planning and Inference* 2005; **133**:435–451.
10. Harville DA. Bayesian inference for variance components using only error contracts. *Biometrika* 1977; **61**: 383–385.
11. Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
12. D'yachkova YD, Petkau J, White R. Longitudinal analysis for magnetic resonance imaging outcomes in multiple sclerosis clinical trials. *Journal of Biopharmaceutical Statistics* 1997; **7**:501–531.
13. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**:581–592.
14. Geisser S. The predictive sample reuse method with applications. *Journal of the American Statistical Association* 1975; **70**:320–328.
15. Dawid AP. Present position and potential developments: some personal views. *Journal of the Royal Statistical Society*, *Series A* 1984; **147**:278–292.
16. Fellner WH. Robust estimation of variance components. *Technometrics* 1986; **28**:51–60.
17. Huggins RM. A robust approach to the analysis of repeated measures. *Biometrics* 1993; **49**:715–720.
18. Richardson AM. Bounded influence estimation in the mixed linear model. *Journal of the American Statistical Association* 1997; **92**:154–161.
19. Richardson AM, Welsh AH. Robust restricted maximum likelihood in mixed linear models. *Biometrics* 1995; **51**:1429–1439.
20. Lee JC. Prediction and estimation of growth curve with special covariance structures. *Journal of the American Statistical Association* 1988; **83**:432–440.
21. Chi EM, Reinsel GC. Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association* 1989; **84**:452–459.

