

Bayesian Estimation With Flexible Prior for the Covariance Structure of Linear Mixed Effects Models

Marick S. Sinay¹, Chi-Wen Hsu² & John S. J. Hsu³

¹ Zest Finance, Los Angeles, California, USA

² Center for Teacher Education, National Taiwan Sport University, Taoyuan, Taiwan, ROC

³ Department of Statistics and Applied Probability, University of California, Santa Barbara, California, USA

Correspondence: John S. J. Hsu, Department of Statistics and Applied Probability, University of California, Santa Barbara, California, USA. Tel: 1-805-893-4055. E-mail: hsu@pstat.ucsb.edu

Received: February 21, 2013 Accepted: August 25, 2013 Online Published: September 9, 2013

doi:10.5539/ijsp.v2n4p29

URL: <http://dx.doi.org/10.5539/ijsp.v2n4p29>

Abstract

Linear mixed effects models arise quite naturally in a number of settings. Two of the more prominent uses are in experimental designs and multilevel models. Furthermore, Bayesian analysis has also been utilized with respect to such models. Here we will consider such an approach with emphasis placed on estimation of the covariance matrix for the random effects. With respect to the covariance structure, however, we depart from the traditional Bayesian prior usage of the Inverse Wishart distribution. The rationale for such a departure is that this distribution is somewhat constraining. Instead, we employ a multivariate Normal approximation procedure for the likelihood of the matrix logarithm of the random effects covariance matrix. Such an approximation allows us to use a multivariate Normal prior for the logarithm of the random effects covariance matrix and still maintain the tractability of conjugacy, at least in an approximate sense. All posterior moments are calculated via Markov Chain Monte Carlo (MCMC) techniques. The Metropolis–Hastings accept reject algorithm is utilized to appropriately account for the approximation procedures. As a particular application we consider a multilevel model where student grade point average relate to a number of standardized test scores.

Keywords: multivariate normal distribution, Markov chain Monte Carlo, hierarchical models, inverted Wishart prior, matrix logarithm, multilevel models, Metropolis-Hastings algorithm

1. Introduction

The goal here in this research is estimation of the covariance matrix of a multivariate response variable. This allows practitioners to appropriately model correlation structures. It is well known that within Bayesian statistics that the Inverse Wishart is a conjugate prior for the classic multivariate Normal model (Evans, 1965; Dickey et al., 1985; Chen, 1979). The Inverse Wishart prior is completely characterized by a scalar degree of freedom hyperparameter and a location matrix hyperparameter, that is equal in dimension to the underlying covariance matrix. As the name implies, the prior location matrix contains information concerning the location of each element of the original covariance matrix. The degree of freedom hyperparameter possesses all of the information surrounding the strength of prior beliefs in the location values.

This choice of prior distribution is mainly driven by ease of analytical exposition. However, the Inverse Wishart is restrictive in two key ways. First, the scalar degree of freedom hyperparameter contains all the information surrounding strength of prior beliefs. This single value does not allow for heterogeneity across the elements of the prior location matrix. If a diffuse prior is desired for certain elements of the location matrix hyperparameter, while a more a concentrated prior is desired for other elements, then the Inverse Wishart cannot accommodate this. The second is that since the prior location matrix only models locational information, we cannot model any correlation structure in the underlying covariance. Please note some noninformative priors for a covariance matrix, such as the Jeffreys prior, developed by Jeffreys (1961), Geisser (1965) and Villegas (1969) and the reference prior, proposed by Yang and Berger (1994). Shrinkage priors based on scale mixtures of uniform distributions were proposed by Wang and Pillai (2013). Please also note, Zeithammer and Lenk (2006) discussed a related work in a special case when dimensions are absent.

In their original work Leonard and Hsu (1992), illustrated another method of Bayesian analysis for a covariance matrix that solves both of the problems delineated above. Berger (1985, p. 400) outlines that if we consider the logarithm of the variance of a Normal distribution, then a Normal can be utilized as a prior. As a multivariate extension to this technique, Leonard and Hsu use an eigenvalue decomposition to take the logarithm of a covariance matrix. This matrix logarithm transformation is then coupled with a result from Bellman (1970, p. 171). In particular, Leonard and Hsu show that the exponential terms can be written as a Volterra integral. Bellman's iterative solution to the Volterra integral can then be utilized to approximate the likelihood. Quite conveniently, it turns out that the resulting approximate likelihood is in the form of a multivariate Normal. Thus, the Normal will serve as a conjugate. The approximate posterior distribution will also be of this same form. It should be obvious that a multivariate Normal distribution overcomes the main shortcomings outlined above of the Inverse Wishart prior. Varying degrees of confidence can now be expressed. Additionally, interdependency amongst the covariance parameters can also be model *a priori*.

Throughout we will make use of such a novel Bayesian procedure for estimation purposes of the covariance matrix of a linear mixed effects model. The general outline of the article is to begin by introducing the linear mixed effects model and performing the requisite Bayesian analysis. We then will outline a Markov Chain Monte Carlo method with a Metropolis-Hastings algorithm. We will consider one particular application to the High School and Beyond survey, which was administered by the National Center for Education Statistics (NCES, 1980). Since the study is representative of a multilevel model a mixed effects model is appropriate. Standardized test scores will serve as explanatory variables and the response variable will be student grade point average. We will also produce the best linear unbiased predictor (BLUP) for the grade point average of a hypothetical student.

2. Linear Mixed Effects Model

We begin our analysis by considering the linear mixed effects model. Consider the i - j th observation which is given by $y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{b}_i + \epsilon_{ij}$, where in general $\mathbf{x}_{ij} = [x_{ij0}, \dots, x_{ijk}]^T$ is a $[(k+1) \times 1]$ fixed effect vector of explanatory variables and $\mathbf{z}_{ij} = [z_{ij0}, \dots, z_{ijp}]^T$ is a $[(p+1) \times 1]$ random effect vector of explanatory variables. $\boldsymbol{\beta} = [\beta_0, \dots, \beta_k]^T$ is a $[(k+1) \times 1]$ fixed effects vector and $\mathbf{b}_i = [b_{i0}, \dots, b_{ip}]^T$ is a $[(p+1) \times 1]$ random effects vector for the i th group.

Assume that each group has n_i observations for $i = 1, \dots, N$. Thus, we are able to express the model for the i th group in the following matrix vector form.

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}_{(n_i \times 1)} = \begin{bmatrix} x_{i10} & \cdots & x_{i1k} \\ \vdots & \ddots & \vdots \\ x_{in_i0} & \cdots & x_{in_ik} \end{bmatrix}_{[n_i \times (k+1)]} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}_{[(k+1) \times 1]} + \begin{bmatrix} z_{i10} & \cdots & z_{i1p} \\ \vdots & \ddots & \vdots \\ z_{in_i0} & \cdots & z_{in_ip} \end{bmatrix}_{[n_i \times (p+1)]} \begin{bmatrix} b_{i0} \\ \vdots \\ b_{ip} \end{bmatrix}_{[(p+1) \times 1]} + \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{in_i} \end{bmatrix}_{(n_i \times 1)}$$

This can be succinctly written in matrix vector notation as $\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i$. We are then able to *stack* all the observations in the following fashion

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix}_{(n \times 1)} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{bmatrix}_{[n \times (k+1)]} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_N \end{bmatrix}_{[n \times (p+1)N]} \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_N \end{bmatrix}_{[(p+1)N \times 1]} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_N \end{bmatrix}_{(n \times 1)}$$

where $n = \sum_{i=1}^N n_i$. The entire model can thus be succinctly written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$.

2.1 Likelihood Function

The first distributional assumption we make is with respect to the random error vector $\boldsymbol{\epsilon}$. In particular, the distribution of the random error vector conditional on the parameter ϕ is multivariate Normal such that $\boldsymbol{\epsilon} | \phi \sim N_n(\mathbf{0}, \phi \mathbf{I}_n)$. Based upon this distributional assumption we can specify the functional form of the likelihood function.

$$l(\boldsymbol{\beta}, \mathbf{b}, \phi | \mathbf{y}) = (2\pi\phi)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2\phi} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}) \right\} \quad (1)$$

The likelihood function as presented in Equation (1) will be used extensively throughout the Bayesian analysis in

the subsequent sections.

2.2 Conditional Distribution for Fixed Effects

Since our main concern lies with the random effects covariance matrix we make the simplifying assumption of vague prior information for the fixed effects. In particular, we assume $\pi(\beta) \propto 1$. If there is greater interest in modeling the fixed effects one can use a multivariate Normal as a prior for β as in Searle et al. (1992, p. 319). This will only slightly increase the complexity of the analysis. Now let $\mathbf{u} = \mathbf{y} - \mathbf{Zb}$. Then the full conditional posterior for β is given by

$$\begin{aligned}\pi(\beta | \mathbf{y}, \mathbf{b}, \phi) &\propto \exp \left\{ -\frac{1}{2\phi} (\mathbf{u} - \mathbf{X}\beta)^T (\mathbf{u} - \mathbf{X}\beta) \right\} \\ &\propto \exp \left\{ -\frac{1}{2\phi} (\beta - \widehat{\beta})^T \mathbf{X}^T \mathbf{X} (\beta - \widehat{\beta}) \right\}\end{aligned}$$

where $\widehat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{u}$. This last result is easily recognizable as the well known multivariate Normal probability density function.

$$\beta | \mathbf{y}, \mathbf{b}, \phi \sim N_{(k+1)} \left((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \mathbf{Zb}), \phi (\mathbf{X}^T \mathbf{X})^{-1} \right) \quad (2)$$

From a computational standpoint the fact that the posterior distribution for the fixed effects is of a multivariate Normal form greatly facilitates the numerical procedures.

Incidentally, had we assumed a Normal prior for β we would have maintained a multivariate Normal posterior distribution due to the multivariate Normal distribution's conjugacy. However, the mean vector and covariance matrix would be slightly modified. The main drawback would be the need to further model the additional hyperparameters associated with the multivariate Normal prior distribution for β .

2.3 Conditional Posterior Distribution for Random Effects

It is fairly common practice to model the random effects with a multivariate Normal distribution (Pinheiro & Bates, 2000, p. 58). We do the same here, however, with a slight modification. In experimental design problems a random effect intercept term is usually not included. However, in mixed effects multilevel models it is not uncommon to include a random effect intercept term. Since our particular application will be with respect to a multilevel model we will in fact include a random effect intercept term. The inclusion of such a random effect intercept term accounts for the group level variability. Additionally, the Normal prior for the random effects covariance matrix requires that an exchangeability condition be met. For this reason we will model the random effect intercept terms separately from the random effect *slope* terms. However, we stress that the Bayesian analysis and subsequent computational procedures are fully generalizable to include models both with and without random effect intercept terms. Here we will assume that $\mathbf{b}_1, \dots, \mathbf{b}_N | \Sigma^* \stackrel{iid}{\sim} N_{(p+1)}(\mathbf{0}, \Sigma^*)$ where Σ^* is a $[(p+1) \times (p+1)]$ block diagonal matrix given by

$$\Sigma^* = \begin{bmatrix} \tau & \mathbf{0} \\ \mathbf{0} & \Sigma \end{bmatrix}. \quad (3)$$

Here the scalar parameter τ is the variance of the random effect intercept terms and Σ is the $(p \times p)$ covariance for the random effect slope terms.

Note that if we let $\mathbf{w}_i = \mathbf{y}_i - \mathbf{X}_i \beta$ then the likelihood for the *i*th group is

$$\begin{aligned}l(\beta, \mathbf{b}_i, \phi | \mathbf{y}_i) &\propto \exp \left\{ -\frac{1}{2\phi} (\mathbf{w}_i - \mathbf{Z}_i \mathbf{b}_i)^T (\mathbf{w}_i - \mathbf{Z}_i \mathbf{b}_i) \right\} \\ &\propto \exp \left\{ -\frac{1}{2\phi} (\mathbf{b}_i - \widehat{\mathbf{b}}_i)^T \mathbf{Z}_i^T \mathbf{Z}_i (\mathbf{b}_i - \widehat{\mathbf{b}}_i) \right\}\end{aligned}$$

where $\widehat{\mathbf{b}}_i = (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \mathbf{w}_i$. Then the posterior distribution for the random effects of the *i*th group is given by the

following.

$$\begin{aligned}\pi(\mathbf{b}_i | \mathbf{y}_i, \boldsymbol{\beta}, \phi, \boldsymbol{\Sigma}^*) &\propto \exp \left\{ -\frac{1}{2\phi} (\mathbf{b}_i - \widehat{\mathbf{b}}_i)^T \mathbf{Z}_i^T \mathbf{Z}_i (\mathbf{b}_i - \widehat{\mathbf{b}}_i) - \frac{1}{2} \mathbf{b}_i^T \boldsymbol{\Sigma}^{*-1} \mathbf{b}_i \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left[\mathbf{b}_i - (\mathbf{Z}_i^T \mathbf{Z}_i + \phi \boldsymbol{\Sigma}^{*-1})^{-1} \mathbf{Z}_i^T \mathbf{w}_i \right]^T \left[\frac{1}{\phi} \mathbf{Z}_i^T \mathbf{Z}_i + \boldsymbol{\Sigma}^{*-1} \right] \right. \\ &\quad \left. \times \left[\mathbf{b}_i - (\mathbf{Z}_i^T \mathbf{Z}_i + \phi \boldsymbol{\Sigma}^{*-1})^{-1} \mathbf{Z}_i^T \mathbf{w}_i \right] \right\}\end{aligned}$$

In this last step here we have made use of a multivariate technique for completing the square (Leonard & Hsu, 1999, p. 245). We recognize the following posterior distribution for the random effects

$$\mathbf{b}_i | \mathbf{y}, \boldsymbol{\beta}, \phi, \boldsymbol{\Sigma}^* \sim N_{(p+1)} \left(\left[\mathbf{Z}_i^T \mathbf{Z}_i + \phi \boldsymbol{\Sigma}^{*-1} \right]^{-1} \mathbf{Z}_i^T \mathbf{w}_i, \left[\frac{1}{\phi} \mathbf{Z}_i^T \mathbf{Z}_i + \boldsymbol{\Sigma}^{*-1} \right]^{-1} \right) \quad (4)$$

2.4 Conditional Posterior Distribution for ϕ

For the variance parameter of the error terms we utilize a diffuse prior distribution, that is $\pi(\phi) \propto \phi^{-1}$. We can then derive the following posterior for the variance of the random error.

$$\pi(\phi | \mathbf{y}, \boldsymbol{\beta}, \mathbf{b}) \propto \phi^{-\left(\frac{n}{2}+1\right)} \exp \left\{ -\frac{1}{2\phi} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}) \right\}$$

From this last result we recognize the Inverse Gamma probability density function. Therefore, we have following posterior distribution for the random error variance parameter.

$$\phi | \mathbf{y}, \boldsymbol{\beta}, \mathbf{b} \sim \text{Inverse Gamma} \left(\frac{n}{2}, \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}) \right) \quad (5)$$

Again, we have a highly tractable posterior distribution in the Inverse Gamma, which will facilitate the computational procedures.

2.5 Conditional Posterior Distribution for τ

Denote the set of random effect intercept terms as $\mathbf{b}_0 = [b_{10}, \dots, b_{N0}]^T$, that is $\mathbf{b}_0$ is the $(N \times 1)$ vector of all the random effect intercept terms for all N groups. Then we have $\mathbf{b}_0 | \tau \sim N_N(\mathbf{0}, \tau \mathbf{I}_N)$ where $\mathbf{I}_N$ is an $(N \times N)$ identity matrix. We assume a vague prior specification for τ , that is $\pi(\tau) \propto \tau^{-1}$. Combining this prior for τ with the distribution of the random effect intercept terms we have the following posterior

$$\pi(\tau | \mathbf{b}_0) \propto \tau^{-\left(\frac{N}{2}+1\right)} \exp \left\{ -\frac{1}{2\tau} \mathbf{b}_0^T \mathbf{b}_0 \right\}.$$

Therefore, we recognize the following posterior distribution for the variance of the intercept random effect regression coefficients.

$$\tau | \mathbf{b}_0 \sim \text{Inverse Gamma} \left(\frac{N}{2}, \frac{1}{2} \mathbf{b}_0^T \mathbf{b}_0 \right) \quad (6)$$

In the subsequent subsections we derive a novel Bayesian estimation technique for the covariance matrix of the random effect slope terms.

2.6 Conditional Posterior Distribution for $\mathbf{A} = \log(\boldsymbol{\Sigma})$

2.6.1 Likelihood Function of $\mathbf{A}$

Here we make use of an approximation technique first detailed by Leonard and Hsu (1992). A more detailed exposition can be found in Hsu et al. (2012). Here we merely outline the procedure. Begin by defining $\widehat{\mathbf{b}}_i = [b_{i1}, \dots, b_{ip}]^T$, for all $i = 1, \dots, N$, as the $(p \times 1)$ vector of just the random effect slope terms for the i th group. That is, $\widehat{\mathbf{b}}_i$ is the vector of all the random effect terms for the i th group except for the random effect intercept term

b_{i0} . Furthermore, define the $(p \times p)$ sample covariance matrix of the random effect slope terms as

$$\mathbf{S}_{(p \times p)} = \frac{1}{N} \sum_{i=1}^N \tilde{\mathbf{b}}_i \tilde{\mathbf{b}}_i^T.$$

Based on the eigen-decomposition we can define the matrix logarithm transformation

$$\begin{aligned} \mathbf{A}_{(p \times p)} &= \log(\mathbf{\Sigma}) = \mathbf{E} [\log(\mathbf{D})] \mathbf{E}^T \\ \mathbf{A}_{(p \times p)} &= \log(\mathbf{S}) = \mathbf{E}_0 [\log(\mathbf{D}_0)] \mathbf{E}_0^T \end{aligned}$$

where the $(p \times p)$ matrix of orthonormal eigenvectors is given by $\mathbf{E}$ and the diagonal $\mathbf{D}$ matrix contains the normalized eigenvalues for $\mathbf{\Sigma}$. Analogously, we define $\mathbf{E}_0$ and $\mathbf{D}_0$ for $\mathbf{S}$.

Noting that $|\mathbf{\Sigma}| = \exp \{ \text{tr} [\mathbf{A}] \}$ we can then write,

$$\begin{aligned} l(\mathbf{\Sigma} | \tilde{\mathbf{b}}_1, \dots, \tilde{\mathbf{b}}_N) &= (2\pi)^{-\frac{Np}{2}} |\mathbf{\Sigma}|^{-\frac{N}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^N \tilde{\mathbf{b}}_i^T \mathbf{\Sigma}^{-1} \tilde{\mathbf{b}}_i \right\} \\ &\propto \exp \left\{ -\frac{1}{2} N \text{tr} [\mathbf{A} + \mathbf{S} \exp \{-\mathbf{A}\}] \right\}. \end{aligned} \quad (7)$$

Now, we define the following function to obtain the upper triangular elements of a matrix in a very specific order, starting with the main diagonal and progressively moving upward. If a_{ij} is the (i, j) th element of $\mathbf{A}$, then,

$$\boldsymbol{\alpha}_{(q \times 1)} = \text{Vec}^*(\mathbf{A}) = \begin{bmatrix} a_{11}, a_{22}, \dots, a_{pp} & a_{12}, a_{23}, \dots, a_{p-1,p} & \dots & a_{1,p-1}, a_{2p} & a_{1p} \end{bmatrix}^T$$

where $q = \frac{1}{2}p(p+1)$. Analogously, $\boldsymbol{\lambda} = \text{Vec}^*(\mathbf{\Lambda})$.

Using this definition and a result for a Volterra integral from (Bellman, 1970, p. 175), Leonard and Hsu (1992) show how the following approximation can be made to Equation (7),

$$l^*(\boldsymbol{\alpha} | \tilde{\mathbf{b}}) = (2\pi e)^{-\frac{np}{2}} |\mathbf{S}|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\lambda})^T \mathbf{Q} (\boldsymbol{\alpha} - \boldsymbol{\lambda}) \right\} \quad (8)$$

where $\tilde{\mathbf{b}} = [\tilde{\mathbf{b}}_1, \dots, \tilde{\mathbf{b}}_N]^T$. The $(q \times q)$ symmetric almost surely positive definite matrix $\mathbf{Q}$ is the so called information matrix of $\boldsymbol{\alpha}$ and is a function of the normalized eigenvalues and normalized eigenvectors of $\mathbf{S}$. In particular,

$$\mathbf{Q}_{(q \times q)} = \frac{N}{2} \sum_{i=1}^p \mathbf{f}_{ii} \mathbf{f}_{ii}^T + N \sum_{i < j}^p \sum_{i < j}^p \xi_{ij} \mathbf{f}_{ij} \mathbf{f}_{ij}^T$$

where

$$\xi_{ij} = \frac{(d_i - d_j)^2}{d_i d_j [\log(d_i) - \log(d_j)]^2}$$

and d_j is the j th normalized eigenvalue of $\mathbf{S}$ for $j = 1, \dots, p$. $\mathbf{f}_{ij} = \mathbf{e}_i * \mathbf{e}_j$ is a $(q \times 1)$ vector where $\boldsymbol{\alpha}^T (\mathbf{e}_i * \mathbf{e}_j) = \mathbf{e}_i^T \mathbf{A} \mathbf{e}_j$.

The approximate probability density function for $\boldsymbol{\alpha}$, given in (8) is Normal. Specifically, when viewed as a function of $\boldsymbol{\alpha}$ the approximate probability density function (8) is a q dimensional Normal with mean $\boldsymbol{\lambda}$ and covariance $\mathbf{Q}^{-1}$. Equation (8) will be used for the Bayesian analysis for $\boldsymbol{\alpha}$. Refer to Leonard and Hsu (1992) for the details on the approximation.

2.6.2 Hierarchical Prior Specification for $\boldsymbol{\alpha}$

As stated above, the approximate likelihood for $\boldsymbol{\alpha}$ is Normal. Therefore, we can use a multivariate Normal prior along with the approximate likelihood to obtain a multivariate Normal approximate posterior. Formally, we assume

$\alpha \mid \eta, \Upsilon$ follows a q dimensional multivariate normal distribution with a $(q \times 1)$ prior mean location hyperparameter vector η and a $(q \times q)$ covariance hyperparameter matrix Υ .

In terms of capturing *a priori* information, the multivariate Normal prior provides an extremely rich class of prior specifications. Here we opt to only utilize a portion of the flexibility afforded us by the multivariate Normal. Specifically, we consider $\eta = \eta(\mu)$ and $\Upsilon = \Upsilon(\sigma)$, where μ and σ are of smaller dimension than η and Υ , respectively. In certain situations, *a priori* we may want to only model a subset of the parameters. In particular, suppose we consider the variance terms separate from the covariance terms.

To this end, we utilize the intra-class matrix model. Specifically, we model the first p elements of α versus from the remaining $(q - p)$ terms. Interested readers are referred to Hsu et al. (2012) for more on this specification.

Formally, *a priori* $\alpha \mid \mu, \Delta \sim N_q(\mathbf{J}\mu, \Delta)$. Then,

$$\pi(\alpha \mid \mu, \Delta) \propto |\Delta|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\alpha - \mathbf{J}\mu)^T \Delta^{-1} (\alpha - \mathbf{J}\mu) \right\} \quad (9)$$

where the (2×1) vector $\mu = [\mu_1, \mu_2]^T$ and

$$\mathbf{J}_{(q \times 2)} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix}^T \quad \Delta_{(q \times q)} = \begin{bmatrix} \sigma_1^2 \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \mathbf{I}_{q-p} \end{bmatrix}.$$

To be clear, the first p elements of the first column of $\mathbf{J}$ are all equal to one and the remaining $(q - p)$ terms are equal to zero. Analogously, the first p elements of the second column of $\mathbf{J}$ are all zero and the remaining $(q - p)$ elements are all one. Thus, μ_1 and σ_1^2 are the location and variance hyperparameters, respectively, for the variance components of α , and μ_2 and σ_2^2 are the location and variance hyperparameters for the covariance components of α . By choosing to model the random effect intercept terms apart from the slope terms we satisfy an exchangeability condition for $\mathbf{b}_1, \dots, \mathbf{b}_N$.

2.6.3 Approximate Posterior Distribution for α

For the hyperparameters $\mu = [\mu_1, \mu_2]^T$ and $\Delta = h(\sigma_1^2, \sigma_2^2)$ we will assume the following Inverse Gamma prior distribution

$$\pi(\mu, \Delta) = \pi(\mu) \pi(\Delta) \propto (\sigma_1^2)^{-(\delta_1+1)} (\sigma_2^2)^{-(\delta_2+1)} \exp \left\{ -\frac{\gamma_1}{\sigma_1^2} - \frac{\gamma_2}{\sigma_2^2} \right\} \quad (10)$$

where $\delta_1, \delta_2, \gamma_1$ and γ_2 are fixed constants to be chosen by the practitioner in order to implement a sensitivity analysis. This topic will be discussed in greater detail in the computational section below. Note that a vague prior specification for μ and Δ will be investigated and can be obtained by letting $\delta_1 = \delta_2 = \gamma_1 = \gamma_2 = 0$.

The joint prior distribution for α, μ and Δ is given by the product of (9) and (10) the prior distribution for μ and Δ . In order to facilitate the Markov Chain Monte Carlo procedure we consider the joint prior distribution for just α and Δ . After integrating over the joint prior distribution for α, μ and Δ with respect to μ we have

$$\pi(\alpha, \Delta) \propto (\sigma_1^2)^{-\left(\frac{p+1}{2} + \delta_1\right)} (\sigma_2^2)^{-\left(\frac{q-p+1}{2} + \delta_1\right)} \exp \left\{ -\frac{1}{2} \alpha^T \mathbf{G}^* \alpha - \frac{\gamma_1}{\sigma_1^2} - \frac{\gamma_2}{\sigma_2^2} \right\} \quad (11)$$

where

$$\mathbf{G}^*_{(q \times q)} = \left[\mathbf{I}_q - \mathbf{J} (\mathbf{J}^T \Delta^{-1} \mathbf{J})^{-1} \mathbf{J}^T \Delta^{-1} \right]^T \Delta^{-1} \left[\mathbf{I}_q - \mathbf{J} (\mathbf{J}^T \Delta^{-1} \mathbf{J})^{-1} \mathbf{J}^T \Delta^{-1} \right]$$

and $\mathbf{I}_q$ is a $(q \times q)$ identity matrix. We make heavy use of (11) throughout §2.6.4 and §2.7.

The approximate posterior for α and Δ will be proportional to the product of (8) and (11).

$$\pi^*(\alpha, \Delta \mid \mathbf{b}) \propto (\sigma_1^2)^{-\left(\frac{p+1}{2} + \delta_1\right)} (\sigma_2^2)^{-\left(\frac{q-p+1}{2} + \delta_1\right)} \exp \left\{ -\frac{1}{2} [(\alpha - \lambda)^T \mathbf{Q} (\alpha - \lambda) + \alpha^T \mathbf{G}^* \alpha] - \frac{\gamma_1}{\sigma_1^2} - \frac{\gamma_2}{\sigma_2^2} \right\} \quad (12)$$

We can then apply the matrix version of completing the square to (12) with respect to the terms in the exponent

involving α . After completing the square, we ignore all terms that do not involve α .

$$\pi^*(\alpha | \mathbf{b}, \Delta) \propto \exp \left\{ -\frac{1}{2} (\alpha - \alpha^*)^T (\mathbf{Q} + \mathbf{G}^*) (\alpha - \alpha^*) \right\} \quad (13)$$

where the $(q \times 1)$ vector $\alpha^* = (\mathbf{Q} + \mathbf{G}^*)^{-1} \mathbf{Q}\lambda$. In summary,

$$\alpha | \mathbf{b}, \Delta \sim N_q(\alpha^*, (\mathbf{Q} + \mathbf{G}^*)^{-1}). \quad (14)$$

This demonstrates the conjugacy of utilizing the approximate likelihood function (8), and numerical methods such as MCMC can readily be implemented by making use of (14). In short, we have developed a highly flexible while at the same time tractable Bayesian methodology for the covariance structure.

2.6.4 Exact Posterior Distribution for α

In order to carry out the Metropolis-Hastings step we require the exact posterior for α , which will be the product of (7) and (11). Here we will use $\alpha = \text{Vec}^*(\mathbf{A})$ interchangeably.

$$\pi(\alpha | \mathbf{b}, \Delta) \propto \exp \left\{ -\frac{N}{2} \text{tr} [\mathbf{A} + \mathbf{S} \exp \{-\mathbf{A}\}] - \frac{1}{2} \alpha^T \mathbf{G}^* \alpha \right\} \quad (15)$$

2.7 Posterior Distribution for Δ

Conditional on α , the posterior for $\Delta = h(\sigma_1^2, \sigma_2^2)$ will be proportional to (11),

$$\pi(\Delta | \alpha) \propto (\sigma_1^2)^{-\left(\frac{p+1}{2} + \delta_1\right)} (\sigma_2^2)^{-\left(\frac{q-p+1}{2} + \delta_2\right)} \exp \left\{ -\frac{1}{\sigma_1^2} \left[\frac{1}{2} \sum_{i=1}^p (\alpha_i - \bar{\alpha}_v)^2 + \gamma_1 \right] - \frac{1}{\sigma_2^2} \left[\frac{1}{2} \sum_{i=p+1}^q (\alpha_i - \bar{\alpha}_c)^2 + \gamma_2 \right] \right\}$$

where $\bar{\alpha}_v = \frac{1}{p} \sum_{i=1}^p \alpha_i$ and $\bar{\alpha}_c = \frac{1}{q-p} \sum_{i=p+1}^q \alpha_i$ are the means of the variance and covariance terms of α , respectively. It can be observed that the posteriors for σ_1^2 and σ_2^2 , conditional on α , are Inverse Gamma and independent of one another.

$$\sigma_1^2 | \alpha \sim \text{Inverse Gamma} \left(\frac{p-1}{2} + \delta_1, \frac{1}{2} \sum_{i=1}^p (\alpha_i - \bar{\alpha}_v)^2 + \gamma_1 \right) \quad (16)$$

$$\sigma_2^2 | \alpha \sim \text{Inverse Gamma} \left(\frac{q-p-1}{2} + \delta_2, \frac{1}{2} \sum_{i=p+1}^q (\alpha_i - \bar{\alpha}_c)^2 + \gamma_2 \right) \quad (17)$$

We should note that with respect to the hyperparameters $\delta_1, \delta_2, \gamma_1$ and γ_2 we will not conduct the usual Bayesian analysis in terms of prior distribution specification and subsequent derivation of the associated posterior distributions. Rather, the purpose of these hyperparameters will be to serve in the sensitivity analysis. Specifically, we wish to investigate how sensitive the posterior distribution and subsequent Bayesian estimates of α will be for various choices of $\delta_1, \delta_2, \gamma_1$ and γ_2 . That is, $\delta_1, \delta_2, \gamma_1$ and γ_2 will be predetermined to be equal to specific numerical values in order to reflect the strength of belief in the prior specification for α .

This completes the analytical portion of the both the Bayesian and hierarchical Bayesian derivations. We now turn to the computational and numerical analysis related to the estimation procedures.

2.8 Markov Chain Monte Carlo Procedures

Equations (2), (4), (5), (6), (14), (15), (16) and (17) provide the necessary distributions for a Markov Chain Monte Carlo procedure with a Metropolis-Hastings step. The goal is to produce posterior means for all relevant model parameters. We now outline how to implement such a procedure.

Initial starting values for β and $\mathbf{b}$ can be obtained by a number of methods. Maximum likelihood estimation or restricted maximum likelihood estimation are two widely used estimation techniques (Pinheiro & Bates, 2000, p. 75).

Alternatively, we obtained initial starting values by iteratively applying ordinary least square estimation. In the first stage, we applied ordinary least square estimation on the entire data set using the original response variables

and the fixed effects explanatory variables to obtain initial starting values for β (please refer to Table 1). In the second stage, we again applied ordinary least square estimation, however now in this case on the group level data, using the residuals from the first stage as the response variables and the random effect explanatory variables. In this fashion we obtained initial starting values for each of the $\mathbf{b}_i$ for all $i = 1, \dots, N$. One of the chief advantages of Bayesian estimation via Markov Chain Monte Carlo techniques is the robustness with respect to initial starting values. Coupling this feature with the analytical and computational ease of ordinary least square estimation offers a secondary advantage.

Table 1. GPA regressed on standardized test scores

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.4001	0.0537	7.45	0.0000
VOCAB1	0.0056	0.0013	4.27	0.0000
VOCAB2	0.0044	0.0013	3.47	0.0005
READ	0.0112	0.0013	8.27	0.0000
MATH1	0.0123	0.0013	9.17	0.0000
MATH2	0.0065	0.0011	5.83	0.0000

Below we outline the specific Markov Chain Monte Carlo procedure invoked in our estimation algorithm. As is the case with Markov Chain Monte Carlo techniques the ordering of draws is somewhat arbitrary, however, it does seem logical to simulate α last since the Metropolis-Hastings accept reject algorithm is employed at this step. Initial starting values can be obtained in a number ways as mentioned above.

- 1) Simulate $\tau^{(t)}$ according to

$$\tau | \mathbf{b}_0 \sim \text{Inverse Gamma} \left(\frac{N}{2}, \frac{1}{2} \mathbf{b}_0^T \mathbf{b}_0 \right).$$

- 2) Simulate $\phi^{(t)}$ according to

$$\phi | \mathbf{y}, \beta, \mathbf{b} \sim \text{Inverse Gamma} \left(\frac{n}{2}, \frac{1}{2} (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b}) \right).$$

- 3) Simulate $\beta^{(t)}$ according to

$$\beta^{(t)} | \mathbf{y}, \mathbf{b}, \phi \sim N_{(k+1)} \left(\left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T (\mathbf{y} - \mathbf{Z}\mathbf{b}), \phi \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \right).$$

- 4) Simulate $\mathbf{b}_i^{(t)}$ for all $i = 1, \dots, N$ according to

$$\mathbf{b}_i | \mathbf{y}, \beta, \phi, \Sigma^* \stackrel{\text{ind}}{\sim} N_{(p+1)} \left(\left[\mathbf{Z}_i^T \mathbf{Z}_i + \phi \Sigma^{*-1} \right]^{-1} \mathbf{Z}_i^T \mathbf{w}_i, \left[\frac{1}{\phi} \mathbf{Z}_i^T \mathbf{Z}_i + \Sigma^{*-1} \right]^{-1} \right).$$

- 5) Simulate $\sigma_1^{2(t)}$ and $\sigma_2^{2(t)}$ according to

$$\sigma_1^2 | \alpha \sim \text{Inverse Gamma} \left(\frac{p-1}{2} + \delta_1, \frac{1}{2} \sum_{i=1}^p (\alpha_i - \bar{\alpha}_v)^2 + \gamma_1 \right)$$

$$\sigma_2^2 | \alpha \sim \text{Inverse Gamma} \left(\frac{q-p-1}{2} + \delta_2, \frac{1}{2} \sum_{i=p+1}^q (\alpha_i - \bar{\alpha}_c)^2 + \gamma_2 \right).$$

- 6) Simulate a candidate value $\hat{\alpha}$ according to

$$\alpha | \mathbf{b}, \Delta \sim N_q \left(\alpha^*, (\mathbf{Q} + \mathbf{G}^*)^{-1} \right).$$

Then let

$$\alpha^{(t)} = \begin{cases} \hat{\alpha} & \text{with probability } \min(\rho_\alpha, 1) \\ \alpha^{(t-1)} & \text{otherwise} \end{cases}$$

$$\text{where } \rho_\alpha = \frac{\pi(\hat{\alpha} | \mathbf{b}^{(t)}, \Delta^{(t)})}{\pi^*(\hat{\alpha} | \mathbf{b}^{(t)}, \Delta^{(t)})} \bigg/ \frac{\pi(\alpha^{(t-1)} | \mathbf{b}^{(t)}, \Delta^{(t)})}{\pi^*(\alpha^{(t-1)} | \mathbf{b}^{(t)}, \Delta^{(t)})}.$$

This last procedure in step (6) is of course the well known Metropolis-Hastings accept reject algorithm. We employ this procedure since we are utilizing an approximation to the exact posterior distribution (Gelman et al., 2005, p. 325). Having laid out the groundwork for the Markov Chain Monte Carlo we now move on to how the posterior moments were produced.

2.9 Posterior Means and Standard Errors

Bayesian estimates of the posterior means and standard errors can be calculated quite easily based upon the Markov Chain Monte Carlo results. We define l to be the number of iterations used as a *burn in*. Furthermore, we define there to be a total of T iterations of the Markov Chain Monte Carlo algorithm and simply for notational simplicity let $M = T - l$. Then for a general parameter vector θ we have the following posterior mean and variance.

$$\widehat{\mathbf{E}} = \frac{1}{M} \sum_{t=l}^T \theta^{(t)}, \quad \widehat{\mathbf{V}} = \frac{1}{M} \sum_{t=l}^T \theta^{(t)^2} - \widehat{\mathbf{E}}^2$$

Standard errors will simply be calculated as the square root of $\widehat{\mathbf{V}}$. In the case of any matrix estimates it is understood that the *vectorized* matrix will be estimated. In practice, we found that a total iteration size of $T = 200000$ was quite sufficient to achieve stability and we used $l = 2000$.

2.10 Sensitivity Analysis

According to the intra-class prior specification all the variance terms of α have the same mean and variance a priori. Similarly, all the covariance terms of α have the same mean and variance a priori. Depending upon the strength of the information contained in the data the posterior distribution for α will reflect this fact more or less to some degree. In practice, we found that with a vague prior for both σ_1^2 and σ_2^2 the Bayesian estimates for these hyperparameters converged to zero. In turn, this led to all the variance terms of α being nearly identical as well as all the covariance terms also being nearly numerically equivalent. That is, the posterior covariance matrix for the slope random effect regression coefficients assumed the classic intra-class form from Dickey et al. (1985).

In order to better investigate and understand this phenomenon we implemented a sensitivity analysis wherein we chose various values for δ_1 , δ_2 , γ_1 and γ_2 . We did so in manner that is consistent with attempting to counteract the degree of strength of the information present in the data. Specifically, we always chose a common value for the four hyperparameters. In fact, to simplify the notation of this section we shall henceforth let $\delta_1 = \delta_2 = \gamma_1 = \gamma_2 = \delta$, a common value. By choosing values for δ in this fashion we maintain the posterior mean for σ_1^2 and σ_2^2 close to unity. This procedure will also lead to a richer class of posterior means for α .

We considered two different cases $\delta = 0$ and $\rho = 0$, which is the limiting case representative of vague prior information with respect to σ_1^2 and σ_2^2 . We now move on to address some more of the specifics of the computational procedures outlined above.

2.11 Other Computational Concerns

As mentioned above in §2.8 starting values for β and $\mathbf{b}$ can be obtained through a number of methods. In our particular case we opted to employ an iterative ordinary least square procedure. This produced very reasonable initial starting values for β . In fact, the ordinary least square estimates for the fixed effect regression coefficients were not too different from the estimates produced using more sophisticated techniques such as restricted maximum likelihood.

Despite the robustness of Markov Chain Monte Carlo techniques there are cases when certain parameter simulations exist in some sense in the tails of the respective probability density functions. This was the case with the α parameter vector. Specifically, the Metropolis-Hastings accept reject algorithm rejected all initial candidate values of α . In order to combat this we employed a training set methodology. That is, the first handful of the α candi-

dates were accepted with certitude. Essentially, the Metropolis-Hastings accept reject algorithm was overridden. In practice, we found that accepting the first fifty α candidates with certitude was more than adequate to allow the Markov Chain to relocate to a space with higher probability of occurrence.

Along these same lines we also found that under the case when $\delta = 0$ the Markov Chain with respect to α also had a tendency to migrate to a state space of very low probability and was unable to recover. To alleviate this we employed a very similar procedure as outlined above for the initial first fifty iterations. That is, when the value of ρ from the Metropolis-Hastings accept reject algorithm fell below the threshold value of 10^{-15} we employed a so called subroutine. Within the subroutine the exact same Markov Chain Monte Carlo procedure as outlined above was invoked, however, the Metropolis-Hastings accept reject algorithm was overridden and in addition none of the simulated values for any parameters were included in the estimates. Upon completion of the subroutine the program returned to the main Markov Chain Monte Carlo procedure with the last simulated values of all parameters being passed forward as initial starting values. We found that this proved to be quite an effective technique for allowing the Markov Chain to migrate away from areas of the probability density function for α of very low probability.

3. High School and Beyond Survey

In 1980 the National Education Longitudinal Studies program of the National Center for Education Statistics administered the High School and Beyond (HSB) survey NCES (1980). The HSB study contains both a 1980 senior class cohort and a 1980 sophomore class cohort. Within the senior class cohort we focused on schools with six or more students of which there were a total of 679. The maximum number of students any one school had was twenty eight. The total number of seniors at all 679 schools investigated was 6671.

The HSB study contains a myriad of data and variables. In particular, for the senior class cohort two exams each in the areas of vocabulary and mathematics and one exam in the area reading were administered. The test scores were standardized to have a mean of fifty and a standard deviation equal to ten. We will use these standardized test scores for the senior class cohort as our explanatory variables. Additionally, grade point average data was obtained from official transcripts that were included in the survey. The grade point average data will serve as our response variable. Below is the summary Table 1 of the initial ordinary least square fitted regression. The adjusted R^2 value was a modest 0.187, however, we feel that in this context this value is reasonable.

Table 2. Posterior mean and standard deviation of β when $\delta = 0$

	β_0	β_1	β_2	β_3	β_4	β_5
Est	0.405783	0.005603	0.004375	0.011219	0.012150	0.006487
Std	0.054666	0.001328	0.001273	0.001348	0.001339	0.001114

Table 3. Posterior mean of $\mathbf{A}$ when $\delta = 0$

	VOCAB1	VACAB2	READ	MATH1	MATH2
VOCAB1	-21.144221	0.409187	0.409187	0.409187	0.409187
VACAB2	0.409187	-21.144221	0.409187	0.409187	0.409187
READ	0.409187	0.409187	-21.144221	0.409187	0.409187
MATH1	0.409187	0.409187	0.409187	-21.144221	0.409187
MATH2	0.409187	0.409187	0.409187	0.409187	-21.144221

Table 4. Posterior standard deviation of $\mathbf{A}$ when $\delta = 0$

	VOCAB1	VACAB2	READ	MATH1	MATH2
VOCAB1	3.589164	1.296743	1.296743	1.296743	1.296743
VACAB2	1.296743	3.589164	1.296743	1.296743	1.296743
READ	1.296743	1.296743	3.589164	1.296743	1.296743
MATH1	1.296743	1.296743	1.296743	3.589164	1.296743
MATH2	1.296743	1.296743	1.296743	1.296743	3.589164

Table 5. Posterior mean and standard deviation of β when $\delta = 50$

	β_0	β_1	β_2	β_3	β_4	β_5
Est	0.405946	0.005609	0.004366	0.011221	0.012140	0.006491
Std	0.054551	0.001325	0.001277	0.001353	0.001336	0.001113

Table 6. Posterior mean of $\mathbf{A}$ when $\delta = 50$

	VOCAB1	VACAB2	READ	MATH1	MATH2
VOCAB1	-21.379099	0.344532	0.524354	0.034575	0.058290
VACAB2	0.344532	-21.178984	0.479396	0.151186	0.514279
READ	0.524354	0.479396	-21.188721	0.124281	0.019755
MATH1	0.034575	0.151186	0.124281	-21.394447	-0.277725
MATH2	0.058290	0.514279	0.019755	-0.277725	-21.356358

Table 7. Posterior standard deviation of $\mathbf{A}$ when $\delta = 50$

	VOCAB1	VACAB2	READ	MATH1	MATH2
VOCAB1	2.947028	1.573193	1.590529	1.537621	1.627801
VACAB2	1.573193	2.708715	1.815628	1.670122	1.569375
READ	1.590529	1.815628	2.732740	1.763906	1.781201
MATH1	1.537621	1.670122	1.763906	2.674138	1.576600
MATH2	1.627801	1.569375	1.781201	1.576600	2.894790

As can be expected with educational data there was some moderate degree of missing data. In particular, out of a total sample size of 6671 students 1661 did have a recorded grade point average, 841 did not have a recorded VOCAB1 test score, 853 did not have a recorded VOCAB2 test score, 845 did not have a recorded READ test score, 867 did not have a recorded MATH1 test score and 963 did not have a MATH2 test score.

Of course working with a complete data set is preferable to simply omitting any students who have one or more missing variables. To this end we employed a data augmentation technique. The specific procedure was invoked using the statistical and mathematical software program R using the “mice” library (Buurin & Oudshoorn, 2000). This particular data augmentation procedure fits nicely within the context of our research here since it employs a multivariate imputation by chained equations technique. That is, it uses a multivariate Gibbs sampler procedure to augment the missing data set. In particular, incomplete columns of data, that is variables with missing data, are augmented by generating appropriate values of data given the values of the other columns of variables. In this way, the other columns act as predictors.

Tables 2-7 present Bayesian estimates and the associated standard errors for all relevant parameters of interest. As discussed above a sensitivity analysis was implementing for investigating the impact of the prior specification for σ_1^2 and σ_2^2 . The sensitivity analysis varies from the extreme limiting case when $\delta = 0$ which is representative of vague prior information for σ_1^2 and σ_2^2 up to when $\delta = 50$ representing greater prior information strength. Recall that $\delta = \delta_1 = \delta_2 = \gamma_1 = \gamma_2$ is the common hyperparameter for both σ_1^2 and σ_2^2 .

Notice from Table 3 that in the limiting case when $\delta = 0$ the matrix $\mathbf{A}$ assumes the intra-class form as described above and in Dickey et al. (1985). Furthermore, for nonzero values of δ the Bayesian estimates of σ_1^2 and σ_2^2 are roughly centered around one. This is due to the mean of the Inverse Gamma posterior distributions for σ_1^2 and σ_2^2 which is on the order of the ratio of the scale parameter divided by the shape parameter. Moreover, as δ increases in magnitude we observe the departure away from the intra-class form for $\mathbf{A}$ as we should expect to occur.

It is common practice to produce a best linear unbiased predictor (BLUP) when working with mixed effects models (Pinheiro & Bates, 2000, p. 94). From a classical frequentist perspective such a quantity can be computed quite easily and is given by $\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T \hat{\mathbf{b}}_i$. However, under our Bayesian framework where posterior means are calculated via Markov Chain Monte Carlo procedures we require a slightly different computation. Specifically, for the posterior best linear unbiased predictor we calculate

$$\text{BLUP}_E = \frac{1}{M} \sum_{t=l}^T \mathbf{x}_{ij}^T \boldsymbol{\beta}^{(t)} + \mathbf{z}_{ij}^T \mathbf{b}_i^{(t)}$$

where as before T is the total number of iterations of the Markov Chain Monte Carlo algorithm, l is the number of burn in trials and for notational simplicity $M = T - l$. Furthermore, the posterior variance of the best unbiased linear predictor can be computed as

$$\text{BLUP}_V = \frac{1}{M} \sum_{t=l}^T \left(\mathbf{x}_{ij}^T \boldsymbol{\beta}^{(t)} + \mathbf{z}_{ij}^T \mathbf{b}_i^{(t)} \right)^2 - \text{BLUP}_E^2$$

As a particular illustration of such a calculation we consider a student at a given school whose explanatory standardized exam scores were equal to the mean test scores. We calculated a posterior best linear unbiased predicted grade point average value of 2.4312 with a standard error equal to 0.0697393.

4. Conclusions

We investigated a linear mixed effects model. In particular, we modeled the covariance matrix of the random effect slope coefficients via the flexible prior paradigm that has been developed in Leonard and Hsu (1992), and further discussed in Hsu et al. (2012). Additionally, we implemented a sensitivity analysis in order to examine the behavior of the hyperparameters as they relate to the matrix intra-class prior for the covariance structure. All posterior means were calculated via a Markov Chain Monte Carlo procedure. As we have done throughout the text we utilized a Metropolis-Hastings step.

Our application here was made with respect to the High School and Beyond survey. Students were nested in a multilevel model via the various schools included in the survey. Due the multilevel nature of the data set a mixed effects model is an appropriate approach. Specifically, we considered grade point average as a response variable and a number of standardized test scores as the explanatory variables. We concluded by calculating the best linear unbiased predictor for a student at a given school with average exam scores.

References

- Bellman, R. (1970). *Introduction to Matrix Analysis*. New York: McGraw Hill.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). New York: Springer. <http://dx.doi.org/10.1007/978-1-4757-4286-2>
- Chen, C. F. (1979). Bayesian inference for a normal dispersion matrix and its applications to stochastic multiple regression analysis. *Journal of the Royal Statistical Society: Series B*, 41, 235-248.
- Dickey, J., Lindley, D. V., & Press, S. (1985). Bayesian estimation of the dispersion matrix of a multivariate normal distribution. *Communications in Statistics Theory and Methods*, 14(5), 1019-1034. <http://dx.doi.org/10.1080/03610928508828960>
- Evans, I. G. (1965). Bayesian estimation of parameters of a multivariate normal distribution. *Journal of the Royal Statistical Society: Series B*, 27, 279-283.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2005). *Bayesian Data Analysis* (2nd ed.). New York: Chapman & Hall.
- Geisser, S. (1965). Bayesian Estimation in Multivariate Analysis. *Annals of Mathematical Statistics*, 36, 150-159. <http://dx.doi.org/10.1214/aoms/1177700279>
- Hsu, C., Sinay, M., & Hsu, J. S. J. (2012). Bayesian Estimation of a Covariance Matrix with Flexible Prior Specification. *The Annals of the Institute of Statistical Mathematics*, 64, 319-342. <http://dx.doi.org/10.1007/s10463-010-0314-5>
- Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press.
- Leonard, T., & Hsu, J. S. J. (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics*, 20(4), 1669-1696. <http://dx.doi.org/10.1214/aos/1176348885>
- Leonard, T., & Hsu, J. S. J. (1999). *Bayesian Methods*. New York: Cambridge University Press.
- National Center for Education Statistics. High School and Beyond, 1980: A Longitudinal Survey of Students in the United States [Computer file]. 2nd ed. Chicago, IL: National Opinion Research Center [producer], 1980. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1986.
- Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus*. New York: Springer. <http://dx.doi.org/10.1007/978-1-4419-0318-1>
- Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components*. New York: John Wiley and Sons, Inc. <http://dx.doi.org/10.1002/9780470316856>
- Van Buuren, S., & Oudshoorn, K. (2000). *Multivariate Imputation by Chained Equations: MICE V 1.0 User's Manual*. Leiden, England: TNO Prevention and Health.
- Villegas, C. (1969). On the a Priori Distribution of the Covariance Matrix. *Annals of Mathematical Statistics*, 40,

1098-1099. <http://dx.doi.org/10.1214/aoms/1177697617>

Wang, H., & Pilliai, N. S. (2013). On a Class of Shrinkage Priors for Covariance Matrix Estimation. *Journal of Computational and Graphical Statistics*, Forthcoming. <http://dx.doi.org/10.1080/10618600.2013.785732>

Yang, R., & Berger, J. O. (1994). Estimation of a Covariance Matrix Using the Reference Prior. *The Annals of Statistics*, 22(3), 1195-1211. <http://dx.doi.org/10.1214/aos/1176325625>

Zeithammer, R., & Lenk, P. (2006). Bayesian Estimation of Multivariate Normal Models when Dimensions are Absent. *Quantitative Marketing and Economics*, 4(3), 241-265. <http://dx.doi.org/10.1007/s11129-005-9006-5>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).