Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2014, Article ID 673657, 13 pages
http://dx.doi.org/10.1155/2014/673657

Research Article

Bayesian Inference of a Multivariate Regression Model

Marick S. Sinay[1] and John S. J. Hsu[2]

[1] *ZestFinance, 6636 Hollywood Boulevard, Los Angeles, CA 90028, USA*
[2] *Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA*

Correspondence should be addressed to John S. J. Hsu; hsu@pstat.ucsb.edu

Received 11 June 2014; Revised 14 October 2014; Accepted 28 October 2014; Published 24 November 2014

Academic Editor: Z. D. Bai

We explore Bayesian inference of a multivariate linear regression model with use of a flexible prior for the covariance structure. The commonly adopted Bayesian setup involves the conjugate prior, multivariate normal distribution for the regression coefficients and inverse Wishart specification for the covariance matrix. Here we depart from this approach and propose a novel Bayesian estimator for the covariance. A multivariate normal prior for the unique elements of the matrix logarithm of the covariance matrix is considered. Such structure allows for a richer class of prior distributions for the covariance, with respect to strength of beliefs in prior location hyperparameters, as well as the added ability, to model potential correlation amongst the covariance structure. The posterior moments of all relevant parameters of interest are calculated based upon numerical results via a Markov chain Monte Carlo procedure. The Metropolis-Hastings-within-Gibbs algorithm is invoked to account for the construction of a proposal density that closely matches the shape of the target posterior distribution. As an application of the proposed technique, we investigate a multiple regression based upon the 1980 High School and Beyond Survey.

1. Introduction

The multivariate multiple regression model is a natural extension of the univariate multiple regression model. The key difference, as the name implies, is that the univariate response variable is instead a multivariate response vector. By utilizing the multivariate multiple regression model the covariance of the response vector can be modeled. From an estimation standpoint van der Merwe and Zidek [1] suggest an intrinsic role to be played by the covariance structure, whereas, in the case of separate univariate multiple regression models, the covariance of the distinct response variables cannot be modeled. Although optimal point estimates of any linear combination of the means of the various response variables can still be obtained, an appropriate estimate of the variance of said estimator cannot be obtained without fully incorporating the covariance amongst the multivariate response vector. Under this framework multivariate analysis is required to most appropriately produce an estimate of the standard error.

As a particular example, in educational testing data when multiple subject area exams are administered it is common practice to simply report the sum of the individual exam scores as a total score. For instance, the metric most commonly associated with the Scholastic Aptitude Test (SAT) is simply the sum of the student's verbal score and the mathematical score. Other exams, such as the ACT exam, have even more than two subject areas and report a composite score which is the arithmetic average of the individual subject area scores. In these instances, multivariate analysis, by capturing the covariance amongst the various subject exams, is required to properly estimate the standard error of the final score.

Formal Bayesian analysis has long been used for multivariate multiple regression models [2]. The inverse Wishart is widely used in this respect, since it is a conjugate prior distribution for the multivariate normal covariance matrix [3–5]. Also see Dawid [6] for a general discussion of the inverse Wishart and Wishart distributions. However, in contrast to traditional Bayesian methods we will not make use of the standard inverse Wishart conjugate prior for the covariance matrix. The reason is that the inverse Wishart is a rather restrictive distribution in its ability to capture prior information that may be available to the practitioner.

See Leonard and Hsu [7], Hsu et al. [8], and Sinay et al. [9] for a more detailed explanation of the disadvantages of the inverse Wishart as a prior distribution for the covariance matrix.

Leonard and Hsu [7] presented an alternative approach that remedies the shortcomings of the inverse Wishart and allows for greater flexibility in the prior specification. In a univariate normal model setting, the normal distribution has been used as a prior for the logarithm of the variance parameter. In this same vein, Leonard and Hsu [7] consider the matrix logarithm transformation of the covariance matrix for the multivariate case. Making use of a result from Bellman [10, page 171], it can be demonstrated that the exponential terms of a multivariate normal likelihood function can be expressed in the form of a Volterra linear integral equation. An approximation to a function, that is, proportional to the likelihood, can then be obtained via Bellman's iterative solution to the Volterra integral equation. The resulting approximation has a multivariate normal form, with respect to the unique elements of the matrix logarithm of the covariance matrix. This allows a normal prior specification to act as a conjugate prior distribution, thereby yielding an approximate normal posterior for the covariance structure.

One of the primary benefits of such a technique is the ability to specify varying degrees of confidence in each element of the normal prior hyperparameter mean vector via the variance terms of the prior hyperparameter covariance matrix. Obviously, larger variance terms in the prior hyperparameter covariance matrix indicate a lack of confidence in the corresponding prior location hyperparameter. Another chief advantage of this method is the ability to model beliefs about any possible interdependency between the covariance parameters. This can be accomplished by specifying the covariance terms of the prior hyperparameter covariance matrix. Note that in this way both the interrelationships and the strength of prior beliefs with respect to the covariance parameters can be modeled.

Bayesian estimates of all relevant parameters of interest are calculated using Markov chain Monte Carlo (MCMC) techniques. With respect to the covariance structure, since an approximate posterior distribution is used as a proposal density, we employ the Metropolis-Hastings-within-Gibbs (MHWG) algorithm [11, page 291], to correctly estimate the true target posterior density.

Having laid out the general outline we now move on to the body of the paper. We begin by introducing and defining the standard multivariate multiple regression model. We follow that by making a distributional assumption about the error matrix of the multivariate regression model. This provides us with a mechanism to state the likelihood function for the model. In turn, we then go through the formal analytical Bayesian derivations. Subsequent to the Bayesian analysis we outline the MCMC procedure and discuss how the posterior means and standard errors are numerically calculated. We conclude with an application to the High School and Beyond Survey [12].

2. Multivariate Multiple Regression Model

We consider the standard multivariate multiple regression model:

$$\underset{(n\times p)}{\begin{bmatrix} Y_{11} & \cdots & Y_{1p} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \cdots & Y_{np} \end{bmatrix}} = \underset{(n\times k)}{\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}} \underset{(k\times p)}{\begin{bmatrix} \beta_{11} & \cdots & \beta_{1p} \\ \vdots & \ddots & \vdots \\ \beta_{k1} & \cdots & \beta_{kp} \end{bmatrix}} + \underset{(n\times p)}{\begin{bmatrix} \epsilon_{11} & \cdots & \epsilon_{1p} \\ \vdots & \ddots & \vdots \\ \epsilon_{n1} & \cdots & \epsilon_{np} \end{bmatrix}}. \tag{1}$$

Notationally in matrix form we can succinctly write

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{Y}$ is the $(n \times p)$ matrix of response variables, $\mathbf{X}$ is the $(n \times k)$ matrix of explanatory variables, $\boldsymbol{\beta}$ is the $(k \times p)$ matrix of unknown regression coefficients, and $\boldsymbol{\epsilon}$ is the $(n \times p)$ matrix of random errors. In Section 2.1 we introduce the matrix normal distribution and make a general assumption about the distribution of the $(n \times p)$ random error matrix $\boldsymbol{\epsilon}$ in (2). The matrix normal representation will greatly facilitate the Baysian analysis that follows. Based upon the matrix normal, we proceed in Section 2.2 to develop the joint likelihood function for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Hierarchical prior specifications are discussed in Section 2.3 and the joint posterior distribution is reported in Section 2.4.

2.1. Distributional Assumptions and Matrix Normal Distribution. The matrix normal distribution is closely related to, and is a generalization of the multivariate normal. In particular, the $(n \times p)$ random matrix $\mathbf{M} \sim \mathrm{MN}_{(n\times p)}(\boldsymbol{\Phi}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$, if and only if, the $(np \times 1)$ random vector $\mathrm{Vec}(\mathbf{M}) \sim \mathrm{N}_{np}(\mathrm{Vec}(\boldsymbol{\Phi}), \boldsymbol{\Sigma} \otimes \boldsymbol{\Omega})$, where $\mathrm{MN}_{(n\times p)}$ denotes the $(n \times p)$ dimensional matrix normal distribution, $\boldsymbol{\Phi}$ is a $(n \times p)$ location matrix, $\boldsymbol{\Sigma}$ is a $(p \times p)$ *first* covariance matrix, and $\boldsymbol{\Omega}$ is a $(n \times n)$ *second* covariance matrix [13, page 54]. $\mathrm{Vec}(\cdot)$ and $\otimes$ are the standard vector operator and Kronecker product, respectively.

We make the distributional assumption that, conditional on the $(p \times p)$ covariance matrix $\boldsymbol{\Sigma}$, the $(n \times p)$ random error matrix $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \boldsymbol{\epsilon}_2^T, \ldots, \boldsymbol{\epsilon}_n^T)^T$, in (2), follows a matrix normal with $(n\times p)$ zero mean matrix and covariance matrices given by $\boldsymbol{\Sigma}$ and $\mathbf{I}_n$, where $\mathbf{I}_n$ is a $(n \times n)$ identity matrix. Formally, we have $\boldsymbol{\epsilon} \mid \boldsymbol{\Sigma} \sim \mathrm{MN}_{(n\times p)}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_n)$, or equivalently, the error terms $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_n$ are independent and identically distributed normal random vectors each with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The probability density function for the error matrix is given by

$$f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \boldsymbol{\Sigma}) = (2\pi)^{-np/2} |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\boldsymbol{\Sigma}^{-1}\right]\right\}, \tag{3}$$

where $\mathrm{tr}(\cdot)$ is the standard trace function.

2.2. Likelihood Function for $\boldsymbol{\Sigma}$ *Conditional on* $\boldsymbol{\beta}$. From the multivariate multiple regression model (2) we can write $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$. Therefore, the joint likelihood function for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ is given by the following:

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \mathbf{Y}) = (2\pi)^{-np/2} |\boldsymbol{\Sigma}|^{-n/2} \times \exp\left\{-\frac{1}{2} \operatorname{tr}\left[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \boldsymbol{\Sigma}^{-1}\right]\right\}. \quad (4)$$

For a given value of $\boldsymbol{\beta}$ let $\mathbf{S} = n^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$. Note that $\mathbf{S}$ is a $(p \times p)$ symmetric almost surely positive definite matrix. Then the likelihood function (4) for $\boldsymbol{\Sigma}$ conditional on $\boldsymbol{\beta}$ can be written as

$$L(\boldsymbol{\Sigma} \mid \mathbf{Y}, \boldsymbol{\beta}) = (2\pi)^{-np/2} |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{-\frac{n}{2} \operatorname{tr}\left[\mathbf{S}\boldsymbol{\Sigma}^{-1}\right]\right\}. \quad (5)$$

In Bayesian analysis for a univariate normal model, the logarithm of the variance parameter has been modeled by a univariate normal prior distribution. In a multivariate setting the matrix logarithm of a covariance matrix has also been investigated by Chiu et al. [14]. Along these same lines, we consider the matrix logarithm of $\boldsymbol{\Sigma}$ and $\mathbf{S}$:

$$\underset{(p\times p)}{\mathbf{A}} = \log(\boldsymbol{\Sigma}) = \mathbf{E}\left[\log(\mathbf{D})\right]\mathbf{E}^T,$$
$$\underset{(p\times p)}{\boldsymbol{\Lambda}} = \log(\mathbf{S}) = \mathbf{E}_0\left[\log(\mathbf{D}_0)\right]\mathbf{E}_0^T, \quad (6)$$

where $\mathbf{E}$ is a $(p \times p)$ orthonormal matrix whose columns are normalized eigenvectors and $\mathbf{D}$ is a $(p\times p)$ diagonal matrix of the corresponding normalized eigenvalues associated with $\boldsymbol{\Sigma}$. $\mathbf{E}_0$ and $\mathbf{D}_0$ are defined analogously for $\mathbf{S}$. Using the fact that $\mathbf{A} = \log(\boldsymbol{\Sigma})$ from (6) and noting that $|\boldsymbol{\Sigma}| = \exp\{\operatorname{tr}[\mathbf{A}]\}$, we can express the exact likelihood function (5) in the following equivalent fashion:

$$L(\mathbf{A} \mid \mathbf{Y}, \boldsymbol{\beta}) = (2\pi)^{-np/2} \exp\left\{-\frac{n}{2} \operatorname{tr}\left[\mathbf{A} + \mathbf{S}\exp\{-\mathbf{A}\}\right]\right\}. \quad (7)$$

We now define the following unconventional matrix operator $\operatorname{Vec}^*(\cdot)$. Let a_{ij} be the element in the ith row and jth column of the matrix $\mathbf{A}$, and then

$$\underset{(q\times 1)}{\boldsymbol{\alpha}} = \operatorname{Vec}^*(\mathbf{A}) = \left[a_{11}, a_{22}, \ldots, a_{pp} \mid a_{12}, a_{23}, \ldots a_{p-1,p} \mid \ldots, a_{1,p-1}, a_{2p} \mid a_{1p}\right]^T, \quad (8)$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_q]^T$ is a $(q\times 1)$ vector and $q = (1/2)p(p+1)$. We analogously define $\boldsymbol{\lambda} = \operatorname{Vec}^*(\boldsymbol{\Lambda})$, which will appear again in Section 3.4. Moving forward we will use $\boldsymbol{\alpha}$, which captures the unique elements of the matrix logarithm of the covariance matrix, to model the covariance structure.

2.3. Prior Distributional Specfications. We will assume *a priori* that $\boldsymbol{\beta}$ is independent of $\boldsymbol{\Sigma}$. More specifically, with respect to $\boldsymbol{\beta}$ we make the assumption of a uniform prior distribution:

$$\pi(\boldsymbol{\beta}) \propto 1. \quad (9)$$

Note that the uniform prior assumption for $\boldsymbol{\beta}$ can be viewed as a limiting case of the informative multivariate normal prior specification, which this modeling approach could accommodate.

We will assume *a priori* that, given $\boldsymbol{\eta}$ and $\boldsymbol{\Upsilon}$, $\boldsymbol{\alpha}$ follows a q dimensional normal distribution with mean location hyperparameter vector $\boldsymbol{\eta}$ and covariance hyperparameter matrix $\boldsymbol{\Upsilon}$. The multivariate normal provides a very rich and flexible family of prior distributions for the matrix logarithm of the covariance structure. This adds far greater flexibility than the conventional inverse Wishart prior specification. Since the multivariate normal is fully parameterized by a mean vector and covariance matrix, we have the ability to model more complex prior information. In particular, we can specify different prior mean values for each element of $\boldsymbol{\alpha}$ via the elements of the location hyperparameter $\boldsymbol{\eta}$. Moreover, we have the ability to model varying degrees of strength of the prior belief in each of the q elements of $\boldsymbol{\eta}$ through the q diagonal elements of the covariance hyperparameter matrix $\boldsymbol{\Upsilon}$. Additionally, with the multivariate normal prior we are able to model potential interdependency among the elements of $\boldsymbol{\alpha}$ because we can specify nontrivial covariance terms in the covariance hyperparameter matrix. That is, the off diagonal elements of $\boldsymbol{\Upsilon}$ can be used to specify any potential correlations amongst the elements of $\boldsymbol{\alpha}$.

We are now able to craft a more complex and accurate prior specification for the covariance structure. A subjective Bayesian may in fact wish to specify all $q+(1/2)q(q+1)$ hyperparameters. In this way, the practitioner can fully take advantage of any relevant prior information through use of the flexible multivariate normal prior specification for the covariance structure. Alternatively, we can opt to model $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\mu})$ and $\boldsymbol{\Upsilon} = \boldsymbol{\Upsilon}(\boldsymbol{\sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are of smaller order than $\boldsymbol{\eta}$ and $\boldsymbol{\Upsilon}$, respectively. That is, *a priori* we may wish to only model certain subsets of the covariance structure. An obvious choice is to consider the variance components as one subset and the covariance components as another. However, we stress the point that the fully general multivariate normal prior specification can be utilized in its totality.

Here we will consider the intraclass matrix form for the prior specification as an example of the fully generalized multivariate normal prior distribution. Specifically, we will consider the first p elements of $\boldsymbol{\alpha}$ separate from the remaining $(q - p)$ terms. That is, we wish to model the variance components separately from the covariance components of $\boldsymbol{\alpha}$. Formally, we assume $\boldsymbol{\alpha} \mid \boldsymbol{\mu}, \Delta \sim \mathrm{N}_q(\mathbf{J}\boldsymbol{\mu}, \Delta)$ for the prior distribution. We have the following prior distributional form:

$$\pi(\boldsymbol{\alpha} \mid \boldsymbol{\mu}, \Delta) \propto |\Delta|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\alpha} - \mathbf{J}\boldsymbol{\mu})^T \Delta^{-1} (\boldsymbol{\alpha} - \mathbf{J}\boldsymbol{\mu})\right\}, \quad (10)$$

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

$$\underset{(q\times 2)}{\mathbf{J}} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix}^T, \qquad \underset{(q\times q)}{\boldsymbol{\Delta}} = \begin{bmatrix} \sigma_1^2 \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \mathbf{I}_{q-p} \end{bmatrix}. \tag{11}$$

Note that $\mathbf{J}$ is a $(q\times 2)$ matrix, whose first p elements of the first column are equal to one and the remaining $(q-p)$ terms of the first column are equal to zero. The second column of $\mathbf{J}$ consists of the first p elements equal to zero and the remaining $(q-p)$ elements equal to one. In $\boldsymbol{\Delta}$, $\mathbf{I}_p$ and $\mathbf{I}_{q-p}$ are indicator matrices of dimension p and $q-p$, respectively.

Thus, μ_1 and σ_1^2 are the location and variance hyperparameters, respectively, for the variance components of $\boldsymbol{\alpha}$. Analogously, μ_2 and σ_2^2 are the location and variance hyperparameters for the covariance components of $\boldsymbol{\alpha}$. In this way we can specify two different location hyperparameters and two levels of confidence.

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

$$\pi(\boldsymbol{\mu}, \boldsymbol{\Delta}) \propto 1. \tag{12}$$

Note here that the uniform prior specification can be viewed as a limiting case of a multivariate normal and inverse Wishart prior specification for $\boldsymbol{\mu}$ and $\boldsymbol{\Delta}$, respectively. Furthermore, the analysis could in fact accommodate such nontrivial specifications quite easily. Having stated all the prior distributional assumptions we now turn to the posterior Bayesian analysis. We begin this by first examining the exact joint posterior distribution.

2.4. Exact Joint Posterior Distribution. The exact joint posterior distribution for all parameters and hyperparameters will be proportional to the product of (4) the exact likelihood function, (9) the prior distribution for $\boldsymbol{\beta}$, (10) the prior distribution for $\boldsymbol{\alpha}$, and (12) the prior distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\Delta}$. Note that here we will use $\boldsymbol{\alpha}$ interchangeably with $\boldsymbol{\Sigma}$:

$$\begin{aligned} &\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Delta} \mid \mathbf{Y}) \\ &\quad \propto |\boldsymbol{\Delta}|^{-1/2} |\boldsymbol{\Sigma}|^{-1/2} \\ &\qquad \times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}\right]\right\} \\ &\qquad \times \exp\left\{-\frac{1}{2}(\boldsymbol{\alpha}-\mathbf{J}\boldsymbol{\mu})^T\boldsymbol{\Delta}^{-1}(\boldsymbol{\alpha}-\mathbf{J}\boldsymbol{\mu})\right\}. \end{aligned} \tag{13}$$

The $\boldsymbol{\mu}$ term in (13) can be integrated out, so that

$$\begin{aligned} &\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Delta} \mid \mathbf{Y}) \\ &\quad \propto |\boldsymbol{\Delta}|^{-1/2} |\boldsymbol{\Sigma}|^{-1/2} \\ &\qquad \times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}\right]\right\} \\ &\qquad \times \left|\mathbf{J}^T\boldsymbol{\Delta}^{-1}\mathbf{J}\right| \exp\left\{-\frac{1}{2}(\boldsymbol{\alpha}-\mathbf{J}\widehat{\boldsymbol{\mu}})^T\boldsymbol{\Delta}^{-1}(\boldsymbol{\alpha}-\mathbf{J}\widehat{\boldsymbol{\mu}})\right\}, \end{aligned} \tag{14}$$

where $\widehat{\boldsymbol{\mu}} = (\mathbf{X}^T\boldsymbol{\Delta}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}$.

We clearly see that the exact joint posterior distribution is in fact not tractable. This is the driving motivation behind the implementation of the numerical MHWG sampling techniques. Rather than working with the cumbersome exact joint posterior distribution it is much easier to consider the so-called full conditional distributions for each of the parameters and hyperparameters.

3. Markov Chain Monte Carlo Approach

As already noted, the joint posterior distribution in (14) is not analytically tractable with respect to drawing inference for the relevant parameters and hyperparameters. This will give rise to our consideration in the subsequent subsections of the full conditional posterior distributions. The conditional posterior distributions we derive below will provide the framework for the MHWG sampling techniques.

3.1. Exact Conditional Posterior Distribution for $\boldsymbol{\beta}$. Recall that $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$. Furthermore, we can rewrite the exact likelihood function (4) as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \mathbf{Y}) &= (2\pi)^{-np/2} |\boldsymbol{\Sigma}|^{-n/2} \\ &\quad \times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)\boldsymbol{\Sigma}^{-1}\right.\right. \\ &\qquad \left.\left. + \left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T\mathbf{X}^T\mathbf{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\boldsymbol{\Sigma}^{-1}\right]\right\}. \end{aligned} \tag{15}$$

Therefore, the posterior distribution for $\boldsymbol{\beta}$ conditional on $\boldsymbol{\Sigma}$ will be proportional, with respect to only the terms involving $\boldsymbol{\beta}$, to the product of (15) the exact likelihood function multiplied by (9) the prior distribution for $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta} \mid \mathbf{Y}, \boldsymbol{\Sigma}) \propto \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T\mathbf{X}^T\mathbf{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\boldsymbol{\Sigma}^{-1}\right]\right\}. \tag{16}$$

We recognize that the posterior distribution of $\boldsymbol{\beta}$ conditional on $\boldsymbol{\Sigma}$ is of a matrix normal form. Making use of the relationship between the matrix normal and the multivariate normal distributions as stated above in Section 2.1, we have

$$\operatorname{Vec}(\boldsymbol{\beta}) \mid \mathbf{Y}, \boldsymbol{\Sigma} \sim \mathrm{N}_{kp}\left(\operatorname{Vec}\left(\widehat{\boldsymbol{\beta}}\right), \boldsymbol{\Sigma} \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right). \tag{17}$$

3.2. Exact Conditional Posterior Distribution for $\boldsymbol{\alpha}$. The joint prior distribution for $\boldsymbol{\alpha}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Delta}$ is given by the product of (10) the prior distribution for $\boldsymbol{\alpha}$ and (12) the joint prior distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\Delta}$. We can derive the joint prior distribution for just $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$ by integrating over the joint prior distribution for $\boldsymbol{\alpha}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Delta}$ with respect to $\boldsymbol{\mu}$. Upon completion of the integration we have the following joint prior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$**:**

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\Delta}) \propto |\boldsymbol{\Delta}|^{-1/2}\left|\mathbf{J}^T\boldsymbol{\Delta}^{-1}\mathbf{J}\right|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{G}^*\boldsymbol{\alpha}\right\}, \tag{18}$$

where

$$\underset{(q\times q)}{\mathbf{G}^{*}} = \left[\mathbf{I}_q - \mathbf{J}\left(\mathbf{J}^T\boldsymbol{\Delta}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^T\boldsymbol{\Delta}^{-1}\right]^T \times \boldsymbol{\Delta}^{-1}\left[\mathbf{I}_q - \mathbf{J}\left(\mathbf{J}^T\boldsymbol{\Delta}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^T\boldsymbol{\Delta}^{-1}\right] \qquad (19)$$

and $\mathbf{I}_q$ is a $(q\times q)$ identity matrix. By integrating $\boldsymbol{\mu}$ out we will help to facilitate the MCMC procedure both in terms of speed and simplification of the algorithm.

The exact posterior distribution will be proportional to the product of (7) the exact likelihood function, multiplied by (18) the joint prior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$. Note that here we will use $\boldsymbol{\alpha}$ interchangeably with $\mathbf{A}$:

$$\pi\left(\boldsymbol{\alpha}\mid\mathbf{Y},\boldsymbol{\beta},\boldsymbol{\Delta}\right) \propto \exp\left\{-\frac{n}{2}\operatorname{tr}\left[\mathbf{A}+\mathbf{S}\exp\left\{-\mathbf{A}\right\}\right]-\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{G}^{*}\boldsymbol{\alpha}\right\}. \qquad (20)$$

Note that the above exact posterior distribution is not of a known form and cannot be directly simulated from in an easy fashion. This will motivate us to use a proposal density that closely matches this target density in a MHWG sampling routine.

3.3. *Exact Conditional Posterior Distribution for* $\boldsymbol{\Delta}$.

The exact posterior distribution for $\boldsymbol{\Delta} = h(\sigma_1^2,\sigma_2^2)$ conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ will be proportional to (18) the joint prior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$. Note that the exact likelihood function (7) does not depend upon $\boldsymbol{\Delta}$ and thus can be omitted entirely:

$$\begin{aligned}\pi\left(\boldsymbol{\Delta}\mid\mathbf{Y},\boldsymbol{\alpha},\boldsymbol{\beta}\right) &\propto \sigma_1^{-(p-1)/2}\sigma_2^{-(q-p-1)/2} \\ &\quad\times\exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{p}\left(\alpha_i-\overline{\alpha}_v\right)^2-\frac{1}{2\sigma_2^2}\sum_{i=p+1}^{q}\left(\alpha_i-\overline{\alpha}_c\right)^2\right\},\end{aligned} \qquad (21)$$

where $\overline{\alpha}_v = (1/p)\sum_{i=1}^{p}\alpha_i$ and $\overline{\alpha}_c = (1/(q-p))\sum_{i=p+1}^{q}\alpha_i$ are the arithmetic means of the variance and covariance components of $\boldsymbol{\alpha}$, respectively. We recognize that the posterior distributions for σ_1^2 and σ_2^2 conditional on $\boldsymbol{\alpha}$ are independent Inverse Gamma random variables:

$$\sigma_1^2\mid\mathbf{Y},\boldsymbol{\alpha},\boldsymbol{\beta}\sim\text{Inverse Gamma}\left(\frac{p-3}{2},\frac{1}{2}\sum_{i=1}^{p}\left(\alpha_i-\overline{\alpha}_v\right)^2\right), \qquad (22)$$

$$\sigma_2^2\mid\mathbf{Y},\boldsymbol{\alpha},\boldsymbol{\beta}\sim\text{Inverse Gamma}\left(\frac{q-p-3}{2},\frac{1}{2}\sum_{i=p+1}^{q}\left(\alpha_i-\overline{\alpha}_c\right)^2\right). \qquad (23)$$

This result is intuitive and theoretically appealing in that the posterior distribution for σ_1^2, the variance hyperparameter for the variance components of $\boldsymbol{\alpha}$, depends only on the number of variance terms p and $\alpha_1,\ldots,\alpha_p$, whereas the posterior distribution for σ_2^2, the variance hyperparameter for the covariance components of $\boldsymbol{\alpha}$, depends only on the number of covariance terms $q-p$ and $\alpha_{p+1},\ldots,\alpha_q$. This draws out the point of modeling the variance components separate from the covariance components. In addition, the Inverse Gamma is highly tractable and lends itself to the numerical procedures in the subsequent section.

3.4. *Approximate Conditional Posterior Distribution for* $\boldsymbol{\alpha}$.

In order to implement the MHWG sampling algorithm, we have derived all the full conditional posterior distributions $\pi(\boldsymbol{\beta}\mid\mathbf{Y},\boldsymbol{\Sigma})$, $\pi(\boldsymbol{\alpha}\mid\mathbf{Y},\boldsymbol{\beta},\boldsymbol{\Delta})$, and $\pi(\boldsymbol{\Delta}\mid\mathbf{Y},\boldsymbol{\alpha},\boldsymbol{\beta})$ in (17), (20), (22), and (23), respectively. However, we clearly see that the simulation of $\boldsymbol{\alpha}$ based upon the true conditional posterior distribution in (20), the target distribution, is not tractable. Therefore, the Metropolis-Hastings algorithm is employed and a proposal density needs to be constructed. The algorithm works best if the proposal density closely matches the shape of the target distribution.

We can construct an approximation to a function, that is, proportional to the likelihood function by utilizing the linear Volterra integral. The key advantage of this is that the approximation can be written as a multivariate normal with respect to the unique elements of the matrix logarithm of the covariance matrix. This allows for a multivariate normal to act as conjugate prior. Hence, we have a multivariate normal posterior, that is, a good proxy for the true posterior, and the proposal can be easily simulated from. Interested readers should refer to Leonard and Hsu [7] and Hsu et al. [8] for a detailed exposition of how this is performed.

Leonard and Hsu [7] show how we can use Bellman's solution of a Volterra integral equation [10, page 171] to derive the following approximation, that is, proportioanl to the likelihood function of $\boldsymbol{\alpha}$ given $\mathbf{Y}$ and $\boldsymbol{\beta}$:

$$L^{*}\left(\boldsymbol{\alpha}\mid\mathbf{Y},\boldsymbol{\beta}\right)\propto\exp\left\{-\frac{1}{2}\left(\boldsymbol{\alpha}-\boldsymbol{\lambda}\right)^T\mathbf{Q}\left(\boldsymbol{\alpha}-\boldsymbol{\lambda}\right)\right\}. \qquad (24)$$

Recall from Section 2.2 that $\boldsymbol{\lambda} = \operatorname{Vec}^{*}(\boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda}$ is as defined in (6). The $(q\times q)$ symmetric almost surely positive definite matrix $\mathbf{Q}$ is the likelihood information matrix of $\boldsymbol{\alpha}$ and is a function of the normalized eigenvalues and normalized eigenvectors of $\mathbf{S}$. In particular,

$$\underset{(q\times q)}{\mathbf{Q}} = \frac{n}{2}\sum_{i=1}^{p}\mathbf{f}_{ii}\mathbf{f}_{ii}^T + n\sum_{i<j}^{p}\sum^{p}\xi_{ij}\mathbf{f}_{ij}\mathbf{f}_{ij}^T, \qquad (25)$$

where $\underset{(q\times 1)}{\mathbf{f}_{ij}} = \mathbf{e}_i * \mathbf{e}_j$ and

$$\xi_{ij} = \frac{\left(d_i-d_j\right)^2}{d_id_j\left[\log\left(d_i\right)-\log\left(d_j\right)\right]^2} \qquad (26)$$

where d_j and $\mathbf{e}_j$ for $j = 1,\ldots,p$ are the jth normalized eigenvalue and eigenvector, respectively, of $\mathbf{S}$. $\mathbf{f}_{ij}$ denotes the $(q\times 1)$ vector that satisfies the condition $\boldsymbol{\alpha}^T(\mathbf{e}_i * \mathbf{e}_j) = \mathbf{e}_i{}^T\mathbf{A}\mathbf{e}_j$. We see that the approximate likelihood function (24) is a multivariate normal form with respect to $\boldsymbol{\alpha}$. This functional form of the approximate likelihood function in (24) will be the driving mechanism in the Bayesian analysis for $\boldsymbol{\alpha}$. Again, for details of the derivation of the approximate likelihood function please refer to Leonard and Hsu [7] and Hsu et al. [8].

Equation (24) provides an excellent approximation to a function, that is, proportional to the exact likelihood (7), when the sample size n is large. To illustrate the effects of n, the sample size, and p, the dimension of the covariance matrix $\boldsymbol{\Sigma}$, we conduct the following exercise. Without loss of generality, we consider a simplified model, when $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is a random sample from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. In our illustrative example, three dimensional sizes ($p = 3, 5$, and 10), and four sample sizes ($n = 20, 100, 500$, and 5000) were considered for comparison. For a fair comparison, the same sample covariance is used for all four different sample sizes, for each p. The sample covariance matrix $\mathbf{S}$ is assumed to consist of elements s_{ij} for the ith row and jth column, where $s_{ij} = 1.0 - |i - j| \times 0.1$. For example, when $p = 3$, then

$$\mathbf{S} = \begin{bmatrix} 1.0 & 0.9 & 0.8 \\ 0.9 & 1.0 & 0.9 \\ 0.8 & 0.9 & 1.0 \end{bmatrix}. \tag{27}$$

The histograms, in Figure 1, represent the exact likelihood from (7) of the (1, 1)th element, a_{11}, of $\mathbf{A}$, where the histogram is normalized for comparison purposes, and the dotted curves represent the univariate normal densities based on approximation (24). Please note that these histograms are computed according to an importance sampling method using (24) as the importance function. For an overview of importance sampling methods, please see, for example, Rubinstein [15] and Leonard et al. [16]. It can be seen from Figure 1 that the approximation is better when the sample size n is bigger and the dimension size p is smaller. Similar patterns were found for other variance and covariance elements of the covariance matrix $\mathbf{A}$. The approximation is fairly accurate when n is 500 or greater.

The approximate joint posterior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$ conditional on $\boldsymbol{\beta}$ will be proportional to the product of the approximate likelihood function (24) and the joint prior distribution (18):

$$\pi^* (\boldsymbol{\alpha}, \boldsymbol{\Delta} \mid \mathbf{Y}, \boldsymbol{\beta}) \propto \sigma_1^{-(p-1)/2} \sigma_2^{-(q-p-1)/2} \times \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\alpha} - \boldsymbol{\lambda})^T \mathbf{Q} (\boldsymbol{\alpha} - \boldsymbol{\lambda}) + \boldsymbol{\alpha}^T \mathbf{G}^* \boldsymbol{\alpha}\right]\right\}. \tag{28}$$

Completing the square for the two terms in the exponent of (28) yields the following approximate posterior distribution for $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\Delta}$:

$$\pi^* (\boldsymbol{\alpha} \mid \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Delta}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T (\mathbf{Q} + \mathbf{G}^*)(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)\right\}, \tag{29}$$

where the $(q \times 1)$ vector $\boldsymbol{\alpha}^* = (\mathbf{Q} + \mathbf{G}^*)^{-1}\mathbf{Q}\boldsymbol{\lambda}$. Recall that $\mathbf{Q}$ and $\mathbf{G}^*$ are as defined in (25) and (19), respectively. Thus, we have the following approximate posterior distribution for $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\Delta}$:

$$\boldsymbol{\alpha} \mid \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Delta} \sim \mathrm{N}_q\left(\boldsymbol{\alpha}^*, (\mathbf{Q} + \mathbf{G}^*)^{-1}\right). \tag{30}$$

This demonstrates the conjugacy of utilizing the approximate likelihood function. Equation (30) provides an efficient proposal distribution for implementing a MHWG algorithm. In short, we have developed a highly flexible while at the same time tractable Bayesian methodology for the covariance structure.

3.5. Metropolis-Hastings within Gibbs Sampling Procedure. Based upon the theoretical results derived above we outline the following procedure for implementing the MHWG algorithm. Specifically, from (17), (30), (20), (22), and (23) we have a formal setup for implementing a MCMC procedure with a Metropolis-Hastings step. Below we outline the specific steps involved in implementing the MHWG algorithm.

(1) Simulate $\sigma_1^{2(t)}$ and $\sigma_2^{2(t)}$ from (22) and (23), respectively. Initial starting values for $\boldsymbol{\alpha}$ may be set equal to $\widehat{\boldsymbol{\lambda}} = \mathrm{Vec}^*(\log(\widehat{\mathbf{S}}))$, where

$$\widehat{\mathbf{S}} = n^{-1}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^T \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \tag{31}$$

and $\widehat{\boldsymbol{\beta}}$ is the maximum likelihood estimator for $\boldsymbol{\beta}$. Subsequent simulations of $\sigma_1^{2(t)}$ and $\sigma_2^{2(t)}$ will be based upon simulated values of $\boldsymbol{\alpha}^{(t)}$.

(2) Simulate $\boldsymbol{\beta}^{(t)}$ from (17). Initial starting values for $\boldsymbol{\Sigma}$ may be based upon $\widehat{\mathbf{S}}$. Subsequent simulations of $\boldsymbol{\beta}^{(t)}$ will be based upon simulated values of $\boldsymbol{\alpha}^{(t)}$.

(3) Simulate a candidate value $\widetilde{\boldsymbol{\alpha}}$ from (30) based upon $\sigma_1^{2(t)}$ and $\sigma_2^{2(t)}$ and $\boldsymbol{\beta}^{(t)}$ from steps (1) and (2), respectively. Then let

$$\boldsymbol{\alpha}^{(t+1)} = \begin{cases} \widetilde{\boldsymbol{\alpha}}, & \text{with probability } \min(\rho, 1), \\ \boldsymbol{\alpha}^{(t)}, & \text{otherwise}, \end{cases} \tag{32}$$

where ρ is given by following expression:

$$\rho = \frac{\pi\left(\widetilde{\boldsymbol{\alpha}} \mid \mathbf{Y}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\Delta}^{(t)}\right)}{\pi^*\left(\widetilde{\boldsymbol{\alpha}} \mid \mathbf{Y}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\Delta}^{(t)}\right)} \frac{\pi^*\left(\boldsymbol{\alpha}^{(t)} \mid \mathbf{Y}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\Delta}^{(t)}\right)}{\pi\left(\boldsymbol{\alpha}^{(t)} \mid \mathbf{Y}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\Delta}^{(t)}\right)} \tag{33}$$

and $\pi^*(\cdot \mid \cdot)$ and $\pi(\cdot \mid \cdot)$ are as defined in (29) and (20), respectively.

The last procedure in step (3) is the Metropolis-Hastings algorithm. We employ this procedure since we are utilizing an approximation to the exact posterior distribution [11, page 291].

Posterior moments for any parameter of interest can be calculated easily upon the MHWG results. It is usually the case in MHWG estimation procedures that the first l simulations are not included in the estimates, due to the fact

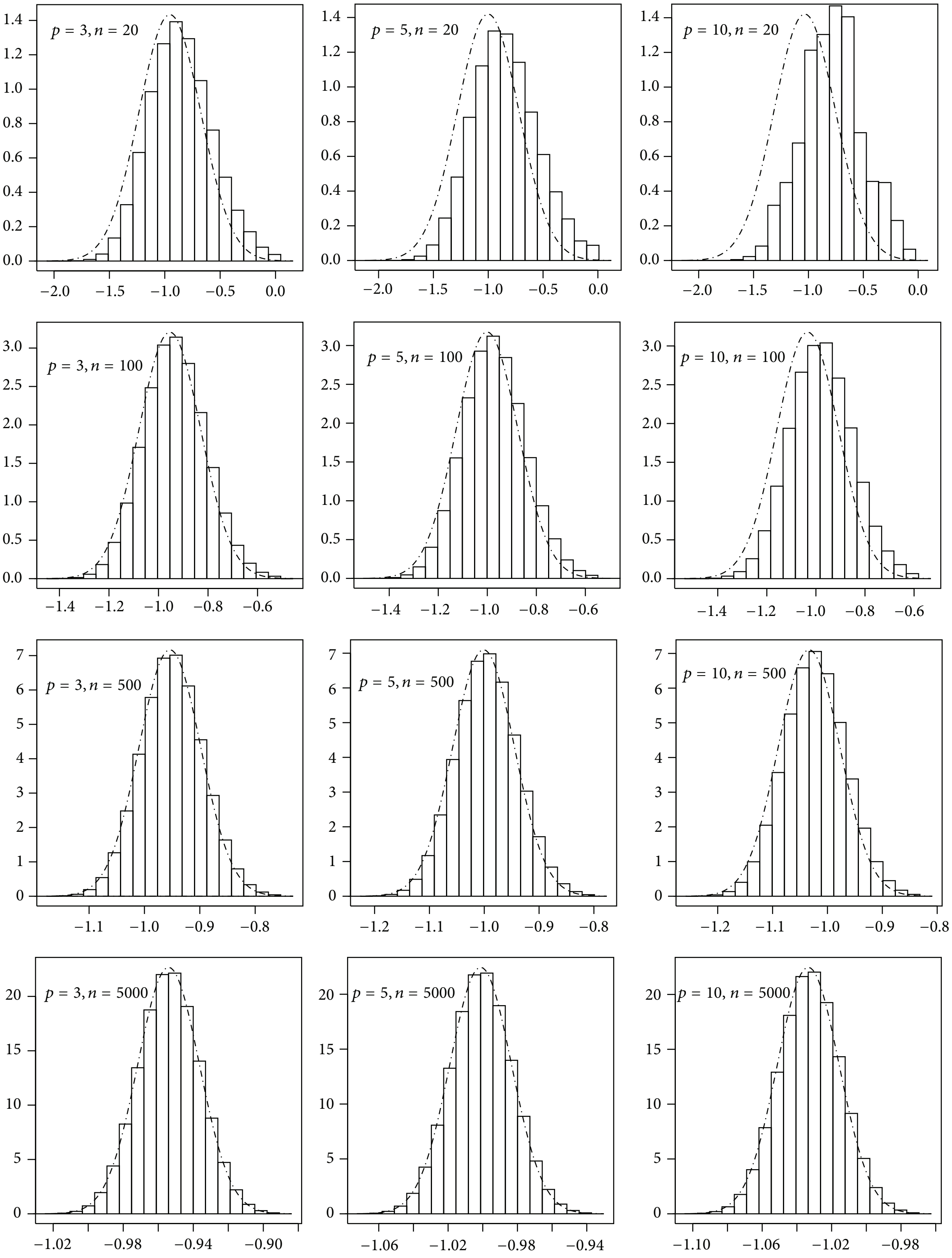


Figure 1: Comparison of the approximate likelihood (normalized) with the exact likelihood (normalized) for covariance matrices of dimensions $p = 3, 5$, and 10, and sample sizes $n = 20, 100, 500$, and 5000. The histogram is the normalized likelihood and the dashed curve is the approximate normalized likelihood.

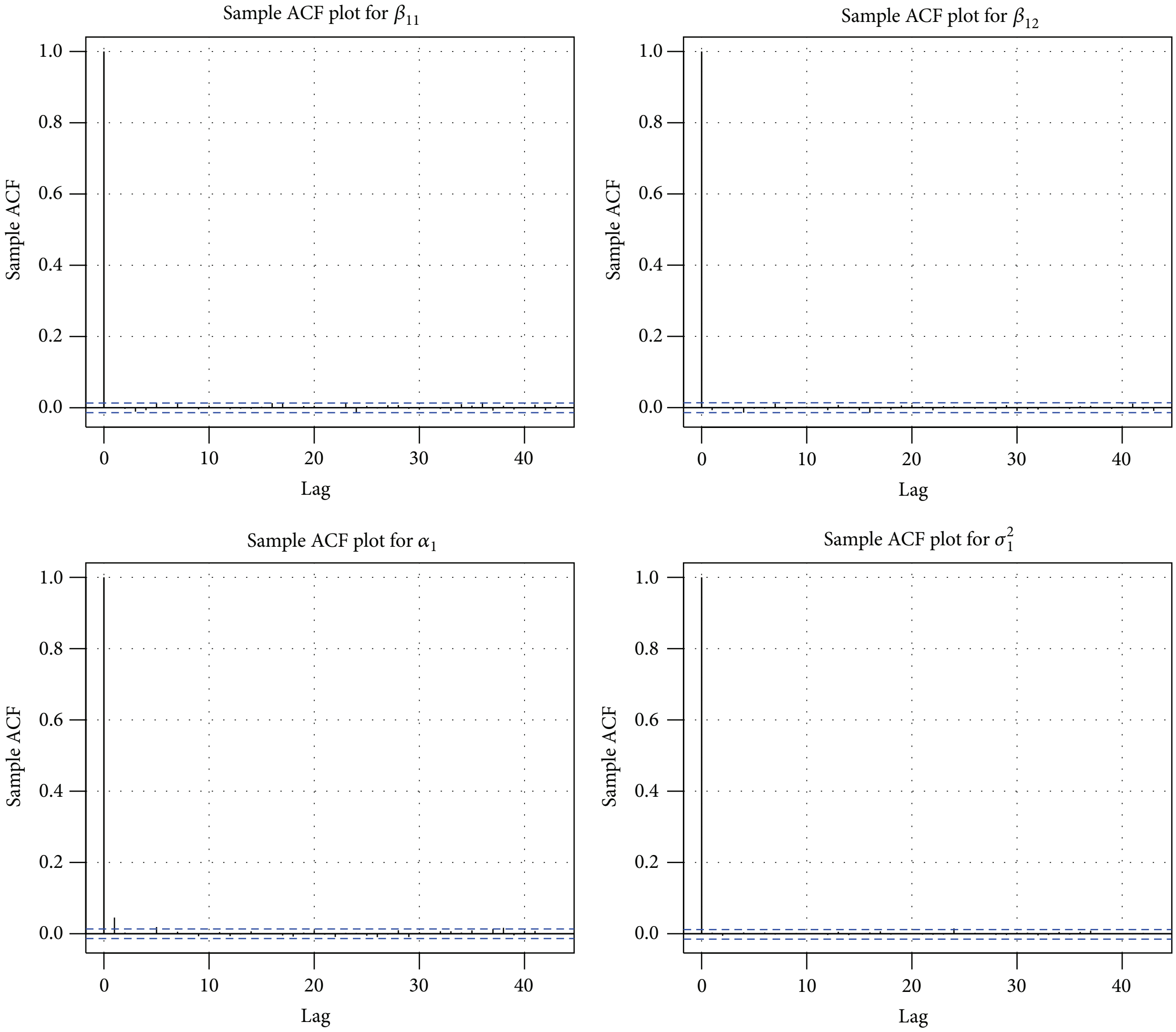


Figure 2: The sample ACF plots demonstrate that in our application a thinning procedure was not necessary. Other sample ACF plots for various parameters looked quite similar. Practitioners should check for this in their applications.

that the Markov chain has not yet reached a steady state. In the parlance of numerical procedures, l is usually referred to as the number of *burn in* iterations.

In addition to the burn in value, we also explored the potential need to perform a so-called *thinning* procedure. A thinning step entails only retaining every rth simulated value of the MHWG sampling, where r is chosen large enough so that any autocorrelation is removed. By examining the sample autocorrelation function (ACF) plots, practitioners can decide if thinning is necessary. We investigated several plots for numerous parameters. For illustrative purposes, in Figure 2, we present the sample ACF plots for β_{11}, β_{12}, α_1, and σ_1^2. We can see that autocorrelation is not a significant concern in our particular application. Other sample ACF plots looked quite similar to numerous other parameters. Practitioners should explore the need to use a thinning procedure in their specific applications.

4. Application: 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 [12]. The HSB study contains both a 1980 senior class cohort and a 1980 sophomore class cohort. Within the sophomore class cohort we have a total sample size of $n = 14{,}667$ students. The HSB study has been analyzed extensively by many, for example, Astone and McLanahan [17], Grogger [18], St. John [19], and Zwick and Sklar [20].

4.1. Description of Data. The HSB study contains a myriad of data and variables. In particular, for the sophomore class cohort a total of $p = 7$ exams were administered in the areas of vocabulary, reading, two exams in mathematics, science, writing, and civics. As is often the case with educational testing data 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 sophomore class cohort as our multivariate response variables.

Grade point average data was also obtained from official transcripts that were included in the survey. In addition, a number of demographic variables were collected on both the school and individual student levels. These variables included school type, relative urbanization of school environment, geographic region of the country, gender, and race. All of these variables will serve as categorical or qualitative explanatory variables in the multivariate multiple regression. Table 1 provides a description of the categorical explanatory variables as well as the associated number of students per category. Note that the original HSB study did not include a school type four.

As can be expected with educational data there was some moderate degree of missing data. We employed a data augmentation technique for missing data imputation. The specific procedure was invoked using the `mice` library [21] in R. 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.

4.2. Treatment Contrast for Categorical Variables. Characterization of the explanatory data or design matrix $\mathbf{X}$ to account for categorical explanatory variables is not unique [22, page 173]. There are actually several *contrast* methods. In our analysis, school type zero, which corresponds to a regular public school, will serve as the base level and will not in fact have its own regressor. In the same fashion, urban will serve as the base level for the relative urbanization categorical explanatory variable. Also, the New England region will act as the base level for the geographic variable. Hispanic or Spanish is the base level for the race identifier variable. Finally, we will designate male as the base gender level. Thus, after properly accounting for the base levels in our particular application $k = 26$.

4.3. Classical Multivariate Multiple Regression Results. We consider the standard multivatiate regression model described in (2) for the HSB survey data within the Sophomore class cohort. A total of $n = 14{,}667$ students in that cohort participated in the study. We estimated the model for the seven standardized exams VOCAB, READ, MATH 1, MATH 2, SCI, WRITE, and CIVIC regressed on GPA and the other explanatory variables described in the previous section. Most of the regression coefficient estimates are highly significant. In particular, GPA is highly significant for all seven exams. Although, certain particular levels for various categorical variables are not significant for some of the response variables. All of the associated F statistics are highly significant for all response variables. Moreover, based upon the analysis of variance that was performed we can conclude that nearly all the explanatory variables are highly significant for all seven subject area exams.

Table 1: Categorical explanatory variable descriptions.

Categorical variables	Levels	Level description	Students per level
School type	School type 0	Regular public	9534
	School type 1	Alternative	437
	School type 2	Cuban/Hispanic public	156
	School type 3	Other Hispanic public	1483
	School type 5	Regular catholic	1341
	School type 6	Black catholic	758
	School type 7	Cuban/Hispanic catholic	216
	School type 8	Private elite	294
	School type 9	Private nonelite	448
Urbanization	Urban 1	Urban	3451
	Urban 2	Suburban	7325
	Urban 3	Rural	3891
Geographic region	Region 1	New England	736
	Region 2	Middle Atlantic	2518
	Region 3	South Atlantic	2178
	Region 4	East South Central	713
	Region 5	West South Central	1750
	Region 6	East North Central	2931
	Region 7	West North Central	1128
	Region 8	Mountain	696
	Region 9	Pacific	2017
Race	Race 1	Hispanic/Spanish	3122
	Race 2	American Indian/Alaskan native	224
	Race 3	Asian/Pacific Islander	345
	Race 4	Black	2050
	Race 5	White	8880
	Race 6	Other	46
Gender	Gender 1	Male	7265
	Gender 2	Female	7402

4.4. Posterior Estimates of the Model Parameters. In Tables 2, 3, 4, 5, and 6 we present the Bayesian posterior means and the associated standard errors for the matrices $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. In practice, for the MCMC procedure we found that a total iteration size of $T = 500{,}000$ was quite sufficient to establish convergence and we used $l = 2{,}000$ as the burn in value.

From Table 4 we observe that the posterior means for the variance and covariance components are each slightly *pulled* towards central quantities, respectively. To better illustrate

Table 2: Posterior mean for $\boldsymbol{\beta}$.

	VOCAB	READ	MATH 1	MATH 2	SCI	WRITE	CIVIC
(Intercept)	37.0582	35.6555	33.6377	37.7119	37.3335	31.2303	37.9283
GPA	4.7810	5.2711	5.9424	4.8139	4.4674	5.7163	4.3733
School type 1	−1.1971	−0.3382	−1.0521	−0.5196	−1.1388	−0.9455	−0.5326
School type 2	−0.8747	−1.9121	−0.6682	−0.4325	−2.4736	−1.4395	−2.4039
School type 3	−1.0713	−0.9146	−0.3638	−0.3816	−1.1779	−0.9105	−0.7742
School type 5	2.7359	2.1862	2.3550	1.5275	0.7704	2.5247	2.1433
School type 6	1.8095	1.8877	0.9502	0.2953	0.2920	2.0570	2.1941
School type 7	3.0532	1.6308	1.5283	0.0003	0.5107	2.1525	1.7716
School type 8	10.3236	8.7564	8.8743	8.7145	5.8165	7.0452	5.5020
School type 9	2.9475	1.9583	2.9381	2.9743	2.1760	1.9975	1.7927
Urban 2	0.5558	0.0821	0.6002	0.9824	0.7407	0.5115	0.2248
Urban 3	−0.8006	−0.5359	−0.3611	−0.2156	0.4344	0.0087	−0.3363
Region 2	−1.0396	−0.7905	−0.3661	−1.0051	−0.4256	−0.6938	−0.9750
Region 3	−2.2086	−1.1256	−1.6328	−1.7664	−1.4603	−1.1176	−1.7760
Region 4	−3.9350	−1.8827	−2.4900	−2.8335	−2.0910	−1.3902	−2.3088
Region 5	−2.6954	−1.5560	−1.5315	−2.1349	−1.1494	−0.8252	−1.6827
Region 6	−2.1648	−1.5579	−0.9690	−1.7886	−0.6775	−0.5598	−1.5524
Region 7	−1.9171	−0.4627	0.1163	−0.3902	0.2802	0.2396	−0.3970
Region 8	−1.6480	−0.9278	−1.6167	−2.4756	−0.6947	−0.4478	−1.4824
Region 9	−0.4866	−0.7586	−0.5338	−1.0931	−0.6208	−0.0413	−1.5015
Gender 2	−1.5605	−1.2414	−1.8966	−1.5490	−3.0722	3.0841	0.2048
Race 2	−0.9813	0.5513	−0.8029	−0.2627	0.5303	0.8640	−0.9432
Race 3	1.7286	1.7006	4.9363	3.9146	2.4132	3.5702	1.3489
Race 4	−0.8534	0.0337	−0.6765	−0.5625	−1.3975	−0.3582	0.4884
Race 5	4.5403	3.7077	3.8777	2.3674	4.5873	3.9272	2.7858
Race 6	−0.5276	0.3656	−0.0162	−0.1581	0.4371	0.3194	0.9498

this shrinkage property of the posterior mean we present the classical frequentist estimate of $\boldsymbol{\Sigma}$ in Table 5. If we make the element-wise comparison between the posterior mean in Table 4 and classical frequentist estimate in Table 5 we see that, among all diagonal elements (variances), relatively smaller elements of the classical frequentist estimate are pulled up and relatively larger elements of the classical frequentist estimate are pulled down. For example, the estimated variance for MATH 1 moved up from 64.9922 in Table 5 to 65.0307 in Table 4 and the estimated variance for CIVIC moved down from 80.8225 in Table 5 to 80.8041 in Table 4. Similar phenomenon appeared for the off-diagonal elements (covariances). This is due to the fact that we have assumed the intraclass matrix form for the prior specification of $\boldsymbol{\alpha}$.

To further investigate the shrinkage property we considered an informative prior distribution for σ_1^2 and σ_2^2 instead of the vague prior specification of (12). In particular, we assumed *a priori* that $\sigma_1^2, \sigma_2^2 \overset{iid}{\sim}$ Inverse Gamma (5000, 1). Table 7 presents the posterior mean for $\boldsymbol{\Sigma}$ with such informative conjugate prior specification. Under this informative prior specification the variance and covariance components are each pulled more towards central quantities, respectively, in comparison to the elements of Table 4. For example, the estimated variance for MATH 1 moved further up from 65.0307 in Table 4 to 65.7643 in Table 7 and the estimated variance for CIVIC moved further down from 80.8041 in Table 4 to 79.0518 in Table 7. Under the informative prior specification for σ_1^2 and σ_2^2 we observe that the shrinkage property is more pronounced.

Specifically, observe that the posterior estimates for the variance of the MATH 2 and CIVIC exams are less than their associated classical frequentist estimates, whereas, all others on diagonal elements of the posterior mean estimate for $\boldsymbol{\Sigma}$ are greater than their respective classical frequentist estimates. An analogous statement concerning the covariance terms can also be made. This draws our attention to the notion of the Bayesian posterior mean as a compromise between the prior information and the information contained in the data.

4.5. Posterior Estimates of the Sum Total. Of particular interest to educational testing data is some estimate of the overall summary or composite score of the individual subject area exams for a given set of explanatory variables. An obvious choice is to estimate the sum total of the individual exam scores. In a Bayesian framework, this can easily be accomplished by simply summing the individual posterior estimates of the seven subject area exams. However, the standard error associated with this estimate of the sum total

Table 3: Posterior standard deviations for $\boldsymbol{\beta}$.

	VOCAB	READ	MATH 1	MATH 2	SCI	WRITE	CIVIC
(Intercept)	0.4421	0.4482	0.4195	0.4585	0.4383	0.4201	0.4678
GPA	0.0920	0.0930	0.0872	0.0951	0.0913	0.0874	0.0972
School type 1	0.4327	0.4379	0.4102	0.4483	0.4297	0.4110	0.4568
School type 2	0.7111	0.7199	0.6736	0.7352	0.7047	0.6761	0.7518
School type 3	0.2797	0.2823	0.2651	0.2902	0.2773	0.2662	0.2954
School type 5	0.2561	0.2589	0.2428	0.2659	0.2540	0.2438	0.2707
School type 6	0.3358	0.3392	0.3185	0.3480	0.3327	0.3193	0.3553
School type 7	0.6116	0.6187	0.5786	0.6331	0.6054	0.5806	0.6452
School type 8	0.5164	0.5229	0.4901	0.5359	0.5125	0.4912	0.5457
School type 9	0.4171	0.4222	0.3963	0.4325	0.4138	0.3972	0.4414
Urban 2	0.1864	0.1882	0.1764	0.1927	0.1844	0.1769	0.1964
Urban 3	0.2135	0.2158	0.2020	0.2206	0.2114	0.2027	0.2253
Region 2	0.3623	0.3672	0.3441	0.3759	0.3593	0.3438	0.3825
Region 3	0.3706	0.3759	0.3520	0.3848	0.3681	0.3524	0.3915
Region 4	0.4512	0.4574	0.4286	0.4678	0.4480	0.4290	0.4774
Region 5	0.3877	0.3928	0.3681	0.4020	0.3850	0.3686	0.4097
Region 6	0.3538	0.3587	0.3358	0.3669	0.3513	0.3365	0.3743
Region 7	0.4056	0.4103	0.3846	0.4208	0.4022	0.3858	0.4286
Region 8	0.4647	0.4709	0.4411	0.4819	0.4614	0.4426	0.4917
Region 9	0.3783	0.3830	0.3586	0.3925	0.3753	0.3598	0.3998
Gender 2	0.1421	0.1436	0.1347	0.1470	0.1411	0.1351	0.1500
Race 2	0.5944	0.6018	0.5630	0.6159	0.5897	0.5652	0.6273
Race 3	0.4938	0.4996	0.4679	0.5114	0.4903	0.4692	0.5216
Race 4	0.2599	0.2639	0.2467	0.2697	0.2579	0.2470	0.2752
Race 5	0.2061	0.2085	0.1952	0.2136	0.2042	0.1959	0.2175
Race 6	1.2666	1.2806	1.1995	1.3091	1.2548	1.2040	1.3359

Table 4: Posterior mean for $\boldsymbol{\Sigma}$.

	VOCAB	READ	MATH 1	MATH 2	SCI	WRITE	CIVIC
VOCAB	72.3406	45.3069	34.3079	25.0339	39.1280	35.7151	32.6781
READ	45.3069	74.0616	37.4238	29.2657	40.8084	36.8560	34.3226
MATH 1	34.3079	37.4238	65.0307	39.9597	36.9083	34.2467	27.4186
MATH 2	25.0339	29.2657	39.9597	77.6078	28.6053	25.5057	19.7419
SCI	39.1280	40.8084	36.9083	28.6053	71.1835	34.7257	31.7008
WRITE	35.7151	36.8560	34.2467	25.5057	34.7257	65.3904	31.9590
CIVIC	32.6781	34.3226	27.4186	19.7419	31.7008	31.9590	80.8041

Table 5: Classical frequentist estimate for $\boldsymbol{\Sigma}$.

	VOCAB	READ	MATH 1	MATH 2	SCI	WRITE	CIVIC
VOCAB	72.3308	45.3567	34.2973	24.9684	39.1487	35.7217	32.6627
READ	45.3567	74.0529	37.4264	29.2202	40.8335	36.8651	34.3098
MATH 1	34.2973	37.4264	64.9922	39.9886	36.9119	34.2393	27.3645
MATH 2	24.9684	29.2202	39.9886	77.6275	28.5646	25.4523	19.6396
SCI	39.1487	40.8335	36.9119	28.5646	71.1592	34.7254	31.6765
WRITE	35.7217	36.8651	34.2393	25.4523	34.7254	65.3498	31.9434
CIVIC	32.6627	34.3098	27.3645	19.6396	31.6765	31.9434	80.8225

Table 6: Posterior standard deviations for $\mathbf{\Sigma}$.

	VOCAB	READ	MATH 1	MATH 2	SCI	WRITE	CIVIC
VOCAB	0.8456	0.7119	0.6329	0.6507	0.6759	0.6399	0.6872
READ	0.7119	0.8657	0.6514	0.6705	0.6886	0.6513	0.6990
MATH1	0.6329	0.6514	0.7604	0.6731	0.6389	0.6082	0.6393
MATH2	0.6507	0.6705	0.6731	0.9069	0.6567	0.6251	0.6742
SCI	0.6759	0.6886	0.6389	0.6567	0.8316	0.6326	0.6779
WRITE	0.6399	0.6513	0.6082	0.6251	0.6326	0.7647	0.6556
CIVIC	0.6872	0.6990	0.6393	0.6742	0.6779	0.6556	0.9425

Table 7: Posterior mean for $\mathbf{\Sigma}$ with an informative prior.

	VOCAB	READ	MATH 1	MATH 2	SCI	WRITE	CIVIC
VOCAB	72.3394	44.8244	34.4191	25.5068	38.8961	35.6451	32.5834
READ	44.8244	74.1193	37.4280	29.5260	40.5447	36.7774	34.1974
MATH 1	34.4191	37.4280	65.7643	39.5284	36.9010	34.3695	27.7634
MATH 2	25.5068	29.5260	39.5284	76.2078	28.8305	25.8843	20.4709
SCI	38.8961	40.5447	36.9010	28.8305	71.2920	34.7172	31.6945
WRITE	35.6451	36.7774	34.3695	25.8843	34.7172	65.8718	31.8856
CIVIC	32.5834	34.1974	27.7634	20.4709	31.6945	31.8856	79.0518

cannot simply be calculated as the square root of the sum of the variance of the individual estimates. This would fail to incorporate the obvious covariance terms that exist amongst the subject area exams. Additionally, we would be overlooking any potential parameter uncertainty.

Suppose for an individual we have the $(p \times 1)$ vector of response variables $\mathbf{Y}_h = [Y_{h1}, \ldots, Y_{hp}]^T$ and the associated $(k \times 1)$ vector of explanatory variables $\mathbf{x}_h = [x_{h1}, \ldots, x_{hk}]^T$. Note that the linear model for a single observation can be expressed as $\mathbf{Y}_h = \boldsymbol{\beta}^T \mathbf{x}_h + \boldsymbol{\epsilon}_h$. Furthermore, the sum total score is $W = \mathbf{1}^T \mathbf{Y}_h = \sum_{j=1}^{p} Y_{hj}$, where $\mathbf{1} = [1, \ldots, 1]^T$ is a $(p \times 1)$ vector of ones. Then, given $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$, the total score W follows a univariate normal distribution with mean $\mathbf{1}^T \boldsymbol{\beta}^T \mathbf{x}_h$ and variance $\mathbf{1}^T \mathbf{\Sigma} \mathbf{1}$.

In order to more fully account for the added variability due to parameter uncertainty we consider the unconditional mean and variance of the sum total for an individual:

$$\begin{aligned} E[W] &= E_{\boldsymbol{\beta},\mathbf{\Sigma}|\mathbf{Y}}\left[E[W \mid \boldsymbol{\beta}, \mathbf{\Sigma}]\right] = E_{\boldsymbol{\beta},\mathbf{\Sigma}|\mathbf{Y}}\left[\mathbf{1}^T \boldsymbol{\beta}^T \mathbf{x}_i\right], \\ \operatorname{Var}[W] &= E_{\boldsymbol{\beta},\mathbf{\Sigma}|\mathbf{Y}}\left[\operatorname{Var}[W \mid \boldsymbol{\beta}, \mathbf{\Sigma}]\right] + \operatorname{Var}_{\boldsymbol{\beta},\mathbf{\Sigma}|\mathbf{Y}}\left[E[W \mid \boldsymbol{\beta}, \mathbf{\Sigma}]\right] \\ &= E_{\boldsymbol{\beta},\mathbf{\Sigma}|\mathbf{Y}}\left[\mathbf{1}^T \mathbf{\Sigma} \mathbf{1}\right] + E_{\boldsymbol{\beta},\mathbf{\Sigma}|\mathbf{Y}}\left[\left(\mathbf{1}^T \boldsymbol{\beta}^T \mathbf{x}_h\right)^2\right] \\ &\quad - \left(E_{\boldsymbol{\beta},\mathbf{\Sigma}|\mathbf{Y}}\left[\mathbf{1}^T \boldsymbol{\beta}^T \mathbf{x}_h\right]\right)^2. \end{aligned} \tag{34}$$

Here it is understood and the notation implies that the resulting expectations in (34) are in fact taken with respect to the posterior distributions of $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$, given $\mathbf{Y}$. Equation (34) can be calculated based upon the results from the MCMC procedures in the following manner:

$$\begin{aligned} \widehat{E}_h &= \frac{1}{N} \sum_{t=l}^{T} \mathbf{1}^T \boldsymbol{\beta}^{(t)^T} \mathbf{x}_h, \\ \widehat{V}_h &= \frac{1}{N} \sum_{t=l}^{T} \mathbf{1}^T \mathbf{\Sigma}^{(t)} \mathbf{1} + \frac{1}{N} \sum_{t=l}^{T} \left(\mathbf{1}^T \boldsymbol{\beta}^{(t)^T} \mathbf{x}_h\right)^2 - \widehat{E}_h^2, \end{aligned} \tag{35}$$

where T is the total number of iterations of the MCMC algorithm, l is the number of burn-in iterations defined in Section 3, and $N = T - l$.

Notice that by calculating the variance of the sum total score in this fashion we fully incorporate the obvious covariance structure amongst the response variables. Furthermore, in a Bayesian sense we capture the added variation due to parameter uncertainty.

As a particular example we investigated a hypothetical student. The student is a female of Hispanic or Spanish decent. Her GPA is 3.0 and she attends a private nonelite school in a rural community of the Pacific region of the United States. Thus, for this particular example we have the following characterization of the $(k \times 1)$ vector of explanatory variables $\mathbf{x}_h$. $x_{h1} = x_{h10} = x_{h12} = x_{h20} = x_{h21} = x_{h26} = 1$, $x_{h2} = 3$, and all the other remaining elements of $\mathbf{x}_h$ are equal to zero.

We obtained the following estimates of the Bayesian posterior means for the individual area exams, 50.9732, 51.2566, 51.5953, 52.1122, 50.0902, 53.7476, and 52.1576 for the VOCAB, READ, MATH 1, MATH 2, SCI, WRITE, and CIVIC exams, respectively. This results in an estimated sum total of 361.9328 with a standard error of 44.2357.

5. Conclusion

In conclusion, we have demonstrated how a flexible prior specification for the covariance structure of a multivariate multiple regression model can provide a richer class of distributions than the inverse Wishart family. We discussed how the likelihood function for the covariance structure can be approximated based upon Bellman's solution of a linear Volterra integral equation. We discussed the shrinkage properties of the posterior mean of the covariance structure. This highlighted the concept of the posterior means as a compromise between the prior information and the information contained in the data.

All posterior estimates were calculated based upon the numerical results of a MHWG procedure. The Metropolis-Hastings algorithm was employed to account for sampling from an approximate posterior distribution for the covariance structure.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] A. van der Merwe and J. V. Zidek, "Multivariate regression analysis and canonical variates," *The Canadian Journal of Statistics*, vol. 8, no. 1, pp. 27–39, 1980.

[2] G. Tiao and A. Zellner, "On the bayesian estimation of multivariate regression," *Journal of the Royal Statistical Society B*, vol. 26, pp. 277–285, 1964.

[3] C. F. Chen, "Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis," *Journal of the Royal Statistical Society, Series B: Methodological*, vol. 41, no. 2, pp. 235–248, 1979.

[4] J. Dickey, D. Lindley, and S. Press, "Bayesian estimation of the dispersion matrix of a multivariate normal distribution," *Communications in Statistics—Theory and Methods*, vol. 14, no. 5, pp. 1019–1034, 1985.

[5] I. J. Evans, "Bayesian estimation of parameters of a multivariate normal distribution," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 27, pp. 279–283, 1965.

[6] A. P. Dawid, "Some matrix-variate distribution theory: notational considerations and a Bayesian application," *Biometrika*, vol. 68, no. 1, pp. 265–274, 1981.

[7] T. Leonard and J. S. J. Hsu, "Bayesian inference for a covariance matrix," *Annals of Statistics*, vol. 20, no. 4, pp. 1669–1696, 1992.

[8] C.-W. Hsu, M. S. Sinay, and J. S. J. Hsu, "Bayesian estimation of a covariance matrix with flexible prior specification," *Annals of the Institute of Statistical Mathematics*, vol. 64, no. 2, pp. 319–342, 2012.

[9] M. S. Sinay, C. Hsu, and J. S. J. Hsu, "Bayesian estimation with flexible prior for the covariance structure of linear mixed effects models," *International Journal of Statistics and Probability*, vol. 2, pp. 29–41, 2013.

[10] R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, New York, NY, USA, 1960.

[11] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Chapman & Hall, New York, NY, USA, 2nd edition, 2004.

[12] National Center for Education Statistics, *High School and Beyond, 1980: A Longitudinal Survey of Students in the United States*, produced by National Opinion Research Center, Chicago, Ill, USA, 1980, distributed by Inter-University Consortium for Political and Social Research, Ann Arbor, Mich, USA, 2nd edition, 1986.

[13] M. S. Srivastava and C. G. Khatri, *An Introduction to Multivariate Statistics*, Elsevier North Holland, New York, NY, USA, 1979.

[14] T. Y. M. Chiu, T. Leonard, and K.-W. Tsui, "The matrix-logarithmic covariance model," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 198–210, 1996.

[15] R. Y. Rubinstein, *Simulation and the Monte Carlo Method*, John Wiley & Sons, New York, NY, USA, 1981.

[16] T. Leonard, J. S. J. Hsu, and K. Tsui, "Bayesian marginal inference," *Journal of the American Statistical Association*, vol. 84, pp. 1051–1058, 1989.

[17] N. M. Astone and S. S. McLanahan, "Family structure, parental practices and high school completion," *American Sociological Review*, vol. 56, no. 3, pp. 309–320, 1991.

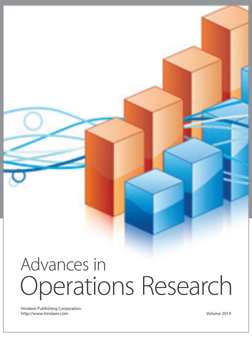
[18] J. Grogger, "School expenditures and post-schooling earnings: evidence from high school and beyond," *Review of Economics and Statistics*, vol. 78, no. 4, pp. 628–637, 1996.

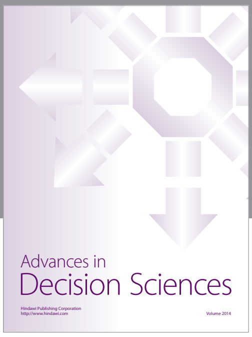
[19] E. P. St. John, "Price response in enrollment decisions: an analysis of the High School and Beyond Sophomore cohort," *Research in Higher Education*, vol. 31, no. 2, pp. 161–176, 1990.

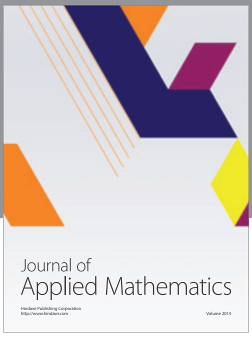
[20] R. Zwick and J. C. Sklar, "Predicting college grades and degree completion using high school grades and SAT scores: the role of student ethnicity and first language," *The American Educational Research Journal*, vol. 42, no. 3, pp. 439–464, 2005.

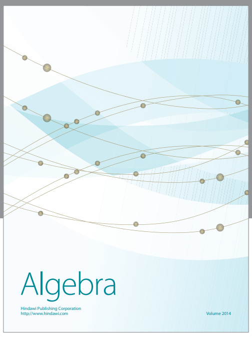
[21] S. Van Buuren and C. G. M. Oudshoorn, "Multivariate Imputation by Chained Equations: MICE V1.0 User's manual," Tech. Rep. PG/VGZ/00.038, TNO Prevention and Health, Leiden, The Netherlands, 2000.

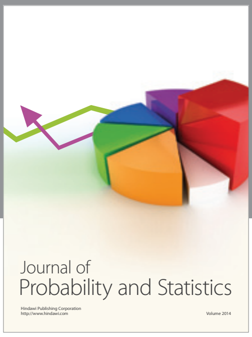
[22] J. J. Faraway, *Linear Models with R*, Chapman & Hall, New York, NY, USA, 2005.


Advances in Operations Research

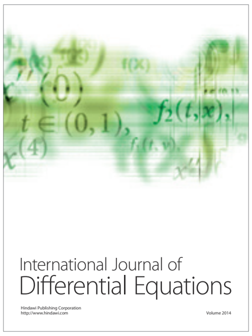

Advances in Decision Sciences


Journal of Applied Mathematics

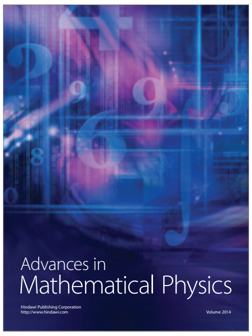

Algebra

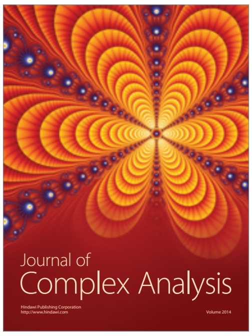

Journal of Probability and Statistics

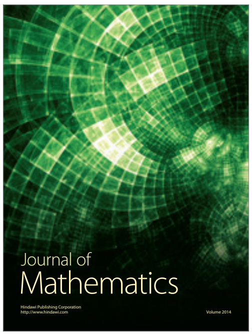
The Scientific World Journal

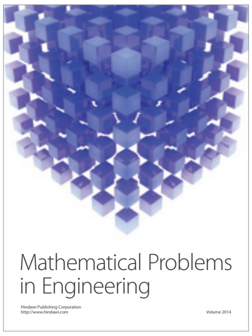

International Journal of Differential Equations

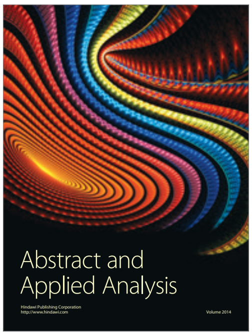
International Journal of Combinatorics

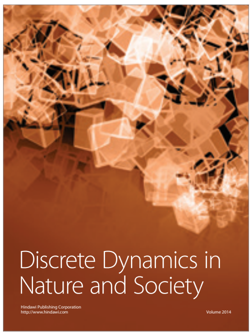

Advances in Mathematical Physics

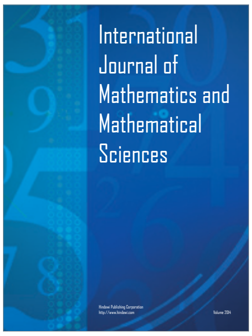

Journal of Complex Analysis


Journal of Mathematics


Mathematical Problems in Engineering

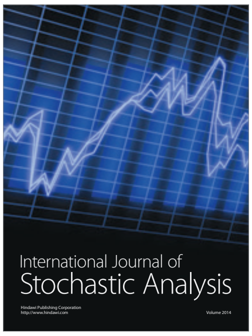

Abstract and Applied Analysis


Discrete Dynamics in Nature and Society


International Journal of Mathematics and Mathematical Sciences

Journal of Discrete Mathematics

Journal of Function Spaces


International Journal of Stochastic Analysis

Journal of Optimization