Ann Inst Stat Math (2012) 64:319–342
DOI 10.1007/s10463-010-0314-5

Bayesian estimation of a covariance matrix with flexible prior specification

Chih-Wen Hsu · Marick S. Sinay · John S. J. Hsu

Received: 14 January 2009 / Revised: 17 March 2010 / Published online: 24 September 2010

Abstract Bayesian analysis for a covariance structure has been in use for decades. The commonly adopted Bayesian setup involves the conjugate inverse Wishart prior specification for the covariance matrix. Here we depart from this approach and adopt a novel prior specification by considering a multivariate normal prior for the elements of the matrix logarithm of the covariance structure. This specification allows for a richer class of prior distributions for the covariance structure with respect to strength of beliefs in prior location hyperparameters and the added ability to model potential correlation amongst the covariance structure. We provide three computational methods for calculating the posterior moment of the covariance matrix. The moments of interest are calculated based upon computational results via Importance sampling, Laplacian approximation and Markov Chain Monte Carlo/Metropolis–Hastings techniques. As a particular application of the proposed technique we investigate educational test score data from the project talent data set.

C.-W. Hsu
Center for Teacher Education, National Taiwan Sport University,
250 Wen Hua 1st Road, Kweishan, Taoyuan 333, Taiwan, ROC
e-mail: cwhsu@mail.ntsu.edu.tw

M. S. Sinay (✉)
Corporate Treasury, Risk Capital and Portfolio Analysis, Bank of America,
315 Montgomery Street, 12th Floor, San Francisco, CA 94014, USA
e-mail: marick.sinay@bankofamerica.com

J. S. J. Hsu
Department of Statistics and Applied Probability, University of California Santa Barbara,
Santa Barbara, CA 93106, USA
e-mail: hsu@pstat.ucsb.edu

Keywords Gibbs sampling · Hierarchical analysis · Importance sampling · Laplacian approximation · Markov Chain Monte Carlo · Matrix logarithm transformation · Metropolis–Hastings algorithm · Volterra integral equation

1 Introduction

Multivariate analysis is of particular relevancy when the goal is inference for a covariance matrix. In this way, the correlation structure amongst observations can most appropriately be modeled. In addition, formal Bayesian analysis has long been used in multivariate analysis. However, in contrast to the common Bayesian method we will not make use of the inverse Wishart conjugate prior distributional specification for the covariance matrix for reasons stated below.

The Wishart distribution arises quite naturally in multivariate statistics. Suppose we have a random sample of p dimensional multivariate normal random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{iid}{\sim} \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, where N_p denotes the p dimensional multivariate normal distribution, $\boldsymbol{\theta}$ is a $(p \times 1)$ mean vector and $\boldsymbol{\Sigma}$ is a $(p \times p)$ symmetric positive definite covariance matrix. Define the $(p \times 1)$ sample mean vector as $\overline{\mathbf{Y}} = n^{-1} \sum_{i=1}^{n} \mathbf{Y}_i$. It follows that the $(p \times p)$ matrix $\sum_{i=1}^{n} \left(\mathbf{Y}_i - \overline{\mathbf{Y}}\right)\left(\mathbf{Y}_i - \overline{\mathbf{Y}}\right)^{\mathrm{T}}$ is distributed $\mathrm{W}_p(n-1, \boldsymbol{\Sigma})$ for $n > p$, where W_p denotes the p dimensional Wishart distribution, the degree of freedom parameter is equal to $(n-1)$ and the $(p \times p)$ matrix $\boldsymbol{\Sigma}$ is the scale matrix parameter.

If the $(p \times p)$ random matrix $\mathbf{M} \sim \mathrm{W}_p(\nu, \boldsymbol{\Psi})$, then $\mathbf{M}^{-1}$ exists almost surely and $\mathbf{M}^{-1} \sim \mathrm{IW}_p\left(\nu, \boldsymbol{\Psi}^{-1}\right)$ where IW_p denotes the p dimensional inverse Wishart distribution (Dawid 1981). Note that the inverse Wishart is fully parameterized by a single degree of freedom parameter ν and a scale matrix parameter $\boldsymbol{\Psi}$. Smaller values of the degree of freedom parameter imply an increasingly more diffuse distribution. On the other hand, larger values for the degree of freedom parameter yield a more highly concentrated distribution about the scale matrix parameter.

In Bayesian statistics, the inverse Wishart distribution is commonly used in multivariate analysis to provide a convenient conjugate prior distribution for the multivariate normal covariance matrix (Chen 1979; Dickey et al. 1985; Evans 1965). Since the inverse Wishart distribution is a conjugate prior, it is both analytically convenient and tractable. However, the inverse Wishart is limited in its flexibility to model prior information. There are two main shortcomings of the inverse Wishart when used as a prior distribution specification for a covariance matrix.

The first disadvantage is that the degree of freedom hyperparameter ν is the sole expression of the confidence level in *all* the elements of the prior hyperparameter matrix. That is, one value represents the strength of prior beliefs for the entire prior scale matrix. This is unappealing in settings where the strength of prior information about the covariance structure is not homogeneous. We may possibly have more or less certainty in our prior knowledge with respect to the location of the elements of the random matrix of interest. Unfortunately, the inverse Wishart prior distribution does not possess the means by which to model this asymmetric level of confidence.

The second shortcoming of the inverse Wishart prior distribution is that it does not allow for the ability to flexibly model any potential interdependency amongst the elements of the covariance matrix. That is, the inverse Wishart prior specification does not provide the ability to easily model the correlation within the covariance structure.

Leonard and Hsu (1992) presented an alternative approach that remedies both of these shortcomings and allows for greater flexibility in the prior specification. In a univariate normal model setting, the normal distribution has long been used as a prior distribution for the logarithm of the variance parameter (Berger 1985, p. 400). In this same vein, Leonard and Hsu consider the matrix logarithm transformation of the covariance matrix for the multivariate case. Making use of a result from Bellman (1970, p. 171), it can be demonstrated that the exponential terms of a multivariate normal likelihood function can be expressed in the form of a linear Volterra integral equation. An approximation of the likelihood function can then be obtained via Bellman's iterative solution to the linear Volterra integral equation. The resulting approximate likelihood function has a multivariate normal form with respect to the unique elements of the matrix logarithm of the covariance matrix. This allows a multivariate normal prior specification to act as a conjugate prior distribution, thereby yielding an approximate multivariate normal posterior distribution 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 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.

Leonard and Hsu computed posterior moments using importance sampling methods. Here we focus on two other competing techniques for computing posterior moments. We first demonstrate how a Laplacian approximation procedure can be utilized to calculate posterior moments. We then turn to Markov Chain Monte Carlo (MCMC) techniques for computing posterior moments. Specifically, we employ a Metropolis–Hastings algorithm within a Gibbs sampling routine (Gelman et al. 2005, p. 291).

The general outline of the article is as follows. We begin with a few brief comments concerning three various loss functions and their respective Bayesian estimators. We then introduce the likelihood function for the covariance matrix under the assumption that the mean vector is known. We report how an approximation can be made with respect to the unique elements of the matrix logarithm of the covariance structure. The resulting approximate likelihood function for the covariance structure will be a multivariate normal form. Then we proceed with the Bayesian analysis under the assumption of vague prior information for the covariance structure. We will transiently discuss the results from Leonard and Hsu (1992) wherein which they calculated the posterior moments via Importance sampling. In addition, we introduce a Laplacian approximation technique for calculating the posterior mean of the covariance

structure under vague prior information. From there we proceed with the hierarchical Bayesian analysis. Specifically, we assume a multivariate normal prior specification for the unique elements of the matrix logarithm of the covariance structure. Computationally, we calculate the posterior mean for the covariance matrix employing Importance sampling, a Laplacian approximation technique as well as MCMC procedures. With respect to the Laplacian approximation technique a novel methodology for approximating the likelihood function of the covariance structure is discussed. Finally, we conclude the article by analyzing the more general setup with an unknown mean vector and hierarchical prior specification. The numerical results in that section will be accomplished via MCMC techniques.

2 Loss functions and Bayes risk

With respect to a covariance matrix three common loss functions are utilized. Specifically, pseudoentropy loss with respect to $\boldsymbol{\Sigma}$, quadratic loss and pseudoentropy loss with respect to $\boldsymbol{\Sigma}^{-1}$, which are given by $L_{\boldsymbol{\Sigma}1}\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) = \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\right) - \log\left|\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\right| - p$, $L_{\boldsymbol{\Sigma}2}\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) = \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \mathbf{I}_p\right)^2$ and $L_{\boldsymbol{\Sigma}3}\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) = \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\right) - \log\left|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\right| - p$, respectively (Ni and Sun 2005). The three associated Bayes estimators that minimize the Bayes risk for these three loss functions are given by $\widehat{\boldsymbol{\Sigma}}_1 = E\left[\boldsymbol{\Sigma} \mid \mathbf{y}\right]$, $\operatorname{Vec}\left(\widehat{\boldsymbol{\Sigma}}_2\right) = \left[E\left[\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right) \mid \mathbf{y}\right]\right]^{-1} \operatorname{Vec}\left(E\left[\boldsymbol{\Sigma}^{-1} \mid \mathbf{y}\right]\right)$ and $\widehat{\boldsymbol{\Sigma}}_3 = E\left[\boldsymbol{\Sigma}^{-1} \mid \mathbf{y}\right]$, respectively, where $\operatorname{Vec}(\cdot)$ is the standard matrix operator that stacks the columns of its argument. Notice that under the pseudoentropy loss with respect to $\boldsymbol{\Sigma}$, denoted by $L_{\boldsymbol{\Sigma}1}$, the Bayes estimator is given by the posterior mean of the covariance matrix. We will in fact make use of this result for our analysis here. The practitioner may use different estimators when different loss functions are considered. For further discussion of Bayesian estimation of a covariance matrix from a decision theoretic viewpoint please refer to such references as Dey and Srinivasan (1985) and Yang and Berger (1994).

3 Likelihood function

Suppose we have a random sample of size n such that $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \mid \boldsymbol{\theta}, \boldsymbol{\Sigma} \overset{iid}{\sim} \mathrm{N}_p\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}\right)$ where N_p denotes the p dimensional multivariate normal distribution and $\boldsymbol{\Sigma}$ is a $(p \times p)$ positive definite symmetric covariance matrix. For this section, we will assume that $\boldsymbol{\theta}$ is known and without loss of generality that $\boldsymbol{\theta} = \mathbf{0}$. Later we will relax this assumption and treat the more general case with an unknown mean vector.

If we denote the realizations of the random sample as $\mathbf{y} = \left[\mathbf{y}_1, \ldots, \mathbf{y}_n\right]$ then the exact likelihood function for $\boldsymbol{\Sigma}$ can be expressed as

$$l\left(\boldsymbol{\Sigma} \mid \mathbf{y}\right) = (2\pi)^{-\frac{np}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \mathbf{y}_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{y}_i\right\} = (2\pi)^{-\frac{np}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}\operatorname{tr}\left[\mathbf{S}\boldsymbol{\Sigma}^{-1}\right]\right\} \tag{1}$$

where the maximum likelihood estimator of $\mathbf{\Sigma}$ is given by $\mathbf{S} = n^{-1} \sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^{\mathrm{T}}$ and tr $(\cdot)$ is the standard trace operator from matrix algebra.

Leonard and Hsu (1992) demonstrate that the likelihood function (1) can be approximated via Bellman's iterative solution to the linear Volterra integral equation. The chief benefit of such a procedure is that the approximate likelihood function can be expressed in terms of a multivariate normal distribution with respect to the unique elements of the matrix logarithm of the covariance matrix. As we will later see below this allows for a multivariate normal distribution to act as a conjugate prior distribution for the covariance structure.

In Bayesian analysis for a univariate normal model, the logarithm of the variance parameter has long been modeled by a univariate normal prior distribution (Berger 1985, p. 400). In a multivariate setting the matrix logarithm of a covariance matrix has also been investigated by Chiu et al. (1996). Along these same lines, we consider the matrix logarithm of $\mathbf{\Sigma}$ and $\mathbf{S}$, $\underset{(p\times p)}{\mathbf{A}} = \log(\mathbf{\Sigma}) = \mathbf{E}\left[\log(\mathbf{D})\right]\mathbf{E}^{\mathrm{T}}$ $\underset{(p\times p)}{\mathbf{\Lambda}} = \log(\mathbf{S}) = \mathbf{E}_0\left[\log(\mathbf{D}_0)\right]\mathbf{E}_0^{\mathrm{T}}$ 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 $\mathbf{\Sigma}$. $\mathbf{E}_0$ and $\mathbf{D}_0$ are defined analogously for $\mathbf{S}$. It is understood that the diagonal matrices log $(\mathbf{D})$ and log $(\mathbf{D}_0)$ are defined to be equal to the matrices whose diagonal elements are equal to the logarithm of the corresponding diagonal elements of the matrices $\mathbf{D}$ and $\mathbf{D}_0$, respectively, and all off diagonal elements of log $(\mathbf{D})$ and log $(\mathbf{D}_0)$ are equal to zero. Note that we have clearly define the matrix logarithm which makes use of the spectral decomposition of a matrix, rather than simply the element-wise logarithm of a matrix. Using the fact that $\mathbf{A} = \log(\mathbf{\Sigma})$ and noting that $|\mathbf{\Sigma}| = \exp\{\mathrm{tr}\,[\mathbf{A}]\}$ we can express the exact likelihood function (1) in the following equivalent fashion.

$$l(\mathbf{A}|\,\mathbf{y}) = (2\pi)^{-\frac{np}{2}} \exp\left\{-\frac{n}{2}\mathrm{tr}\left[\mathbf{A} + \mathbf{S}\exp\{-\mathbf{A}\}\right]\right\} \tag{2}$$

Note that by the invariance property for maximum likelihood estimators (2) is maximized when $\mathbf{A} = \mathbf{\Lambda}$.

We now define the following unconventional matrix operator Vec$^*(\cdot)$. Let a_{ij} be the (i, j)th element of the matrix $\mathbf{A}$, then $\underset{(q\times 1)}{\boldsymbol{\alpha}} = \mathrm{Vec}^*(\mathbf{A}) = \left[a_{11}, a_{22}, \ldots, a_{pp} \,\middle|\, a_{12}, a_{23}, \ldots, a_{p-1,p} \,\middle|\, \ldots \,\middle|\, a_{1,p-1}, a_{2p} \,\middle|\, a_{1p}\right]^{\mathrm{T}}$ where $\boldsymbol{\alpha} = \left[\alpha_1, \ldots, \alpha_q\right]^{\mathrm{T}}$ is a $(q \times 1)$ vector and $q = \frac{1}{2}p(p+1)$. The elements of $\boldsymbol{\alpha}$ are equal to the upper triangular elements of $\mathbf{A}$ starting with the first p main diagonal elements and then moving successively upward and to the right of the main diagonal. We analogously define the $(q \times 1)$ vector $\boldsymbol{\lambda} = \mathrm{Vec}^*(\mathbf{\Lambda})$.

The term $\exp\{-\mathbf{A}\}$ in the exponent of (2) can be expressed as a linear Volterra integral equation (Bellman 1970, p. 175). An approximation of the likelihood function for $\boldsymbol{\alpha}$ can then be obtained via Bellman's iterative solution to the linear Volterra integral equation. Leonard and Hsu (1992) demonstrate that the approximate likelihood function based upon a second-order expansion of said linear Volterra integral equation about $\mathbf{A} = \mathbf{\Lambda}$ is given by

$$l^*(\boldsymbol{\alpha} \mid \mathbf{y}) = (2\pi e)^{-\frac{np}{2}} |\mathbf{S}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\lambda})^{\mathrm{T}} \mathbf{Q}(\boldsymbol{\alpha} - \boldsymbol{\lambda})\right\} \tag{3}$$

where $\mathbf{Q}_{(q\times q)} = \frac{n}{2}\sum_{i=1}^{p} \mathbf{f}_{ii}\mathbf{f}_{ii}^{\mathrm{T}} + n\sum_{i<j}^{p}\sum^{p} \xi_{ij}\mathbf{f}_{ij}\mathbf{f}_{ij}^{\mathrm{T}}$ is the likelihood information matrix of $\boldsymbol{\alpha}$ such that: $\xi_{ij} = \frac{(d_i - d_j)^2}{d_i d_j[\log(d_i) - \log(d_j)]^2}$, $\mathbf{f}_{ij} = \mathbf{e}_i * \mathbf{e}_j$ denotes the $(q \times 1)$ vector that satisfies the condition $\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{e_i} * \mathbf{e_j}) = \mathbf{e_i}^{\mathrm{T}}\mathbf{A}\mathbf{e_j}$ and d_j and $\mathbf{e}_j$ are the jth normalized eigenvalue and eigenvector, respectively, of $\mathbf{S}$ for $j = 1, \ldots, p$.

We see that the approximate likelihood function (3) is a multivariate normal form with respect to $\boldsymbol{\alpha}$. Specifically, the approximate likelihood function for $\boldsymbol{\alpha}$ is a q dimensional multivariate normal distribution with mean vector equal to $\boldsymbol{\lambda}$ and covariance matrix equal to $\mathbf{Q}^{-1}$. This functional form of the approximate likelihood function in Eq. (3) will be the driving mechanism in the Bayesian analysis for $\boldsymbol{\alpha}$. For further details of the derivation of the approximate likelihood function please refer to Leonard and Hsu (1992).

Observe that by combining the approximate likelihood function as given in Eq. (3) with a multivariate normal prior specification for $\boldsymbol{\alpha}$, this will result in a multivariate normal posterior distribution for $\boldsymbol{\alpha}$. This lends itself quite nicely to analytical tractability. We first address the vague prior specification for $\boldsymbol{\alpha}$ and discuss a novel generalization of the finite sampling likelihood approximation technique that will be utilized in the Laplacian approximation. The hierarchical multivariate normal prior specification will be addressed below.

4 Vague prior specification

In this section we begin the formal Bayesian analysis. We follow Leonard and Hsu (1992) and assume a vague prior specification for $\boldsymbol{\alpha}$. Formally, a priori we assume that $\pi(\boldsymbol{\alpha}) \propto 1$. This is mainly done for ease of exposition and comparison purposes with the hierarchical prior specification we outline below.

Combining this vague prior distribution with the exact and approximate likelihood functions we observe that the exact and approximate posterior distributions for $\boldsymbol{\alpha}$ will in fact simply be proportional to Eqs. (2) and (3), respectively.

$$\pi(\boldsymbol{\alpha} \mid \mathbf{y}) \propto \overline{\pi}(\boldsymbol{\alpha} \mid \mathbf{y}) = (2\pi)^{-\frac{np}{2}} \exp\left\{-\frac{n}{2}\mathrm{tr}\left[\mathbf{A} + \mathbf{S}\exp\{-\mathbf{A}\}\right]\right\} \tag{4}$$

$$\pi^*(\boldsymbol{\alpha} \mid \mathbf{y}) \propto \overline{\pi}^*(\boldsymbol{\alpha} \mid \mathbf{y}) = (2\pi e)^{-\frac{np}{2}} |\mathbf{S}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\lambda})^{\mathrm{T}} \mathbf{Q}(\boldsymbol{\alpha} - \boldsymbol{\lambda})\right\} \tag{5}$$

Posterior moments can then be calculated via importance sampling. For an overview of importance sampling please refer to Rubinstein (1981), Leonard et al. (1994) and Robert and Casella (2004, p. 92).

4.1 Laplacian approximation under vague prior specification

As an alternative to the Importance sampling procedure consider calculating the posterior mean via a Laplacian approximation technique. The general Laplacian approximation procedure is widely used throughout applied mathematics and is first attributed to Laplace (Laplace 1986; Stigler 1986). In particular, the method can be used to approximate marginal posterior densities of any single parameter from a multivariate distribution. Leonard (1982), Leonard et al. (1989), Kass et al. (1989) and Leonard and Hsu (1999) discuss the Laplacian approximation technique as it relates to deriving posterior distributions.

Tierney and Kadane (1986) consider the posterior expected value of $g(\boldsymbol{\phi})$ where $g(\cdot)$ is a smooth, positive function on the parameter space and $\boldsymbol{\phi}$ represents a general parameter vector of interest. They note that the posterior mean of $g(\boldsymbol{\phi})$ can be written as

$$E\left[g(\boldsymbol{\phi})\,|\,\mathbf{y}\right] = \frac{\int_{\boldsymbol{\phi}} g(\boldsymbol{\phi})\, l(\boldsymbol{\phi}\,|\,\mathbf{y})\,\pi(\boldsymbol{\phi})\;\mathrm{d}\boldsymbol{\phi}}{\int_{\boldsymbol{\phi}} l(\boldsymbol{\phi}\,|\,\mathbf{y})\,\pi(\boldsymbol{\phi})\;\mathrm{d}\boldsymbol{\phi}} = \frac{\int_{\boldsymbol{\phi}} \exp\{nL^*(\boldsymbol{\phi})\}\;\mathrm{d}\boldsymbol{\phi}}{\int_{\boldsymbol{\phi}} \exp\{nL(\boldsymbol{\phi})\}\;\mathrm{d}\boldsymbol{\phi}} \tag{6}$$

where $L^*(\boldsymbol{\phi}) = \frac{1}{n}\log\left[g(\boldsymbol{\phi})\, l(\boldsymbol{\phi}\,|\,\mathbf{y})\,\pi(\boldsymbol{\phi})\right]$ and $L(\boldsymbol{\phi}) = \frac{1}{n}\log\left[l(\boldsymbol{\phi}\,|\,\mathbf{y})\,\pi(\boldsymbol{\phi})\right]$. Then applying Laplace's technique to both the numerator and denominator in Eq. (6), they approximate the posterior expected value of $g(\boldsymbol{\phi})$ by

$$E\left[g(\boldsymbol{\phi})\,|\,\mathbf{y}\right] \approx \left(\frac{|\boldsymbol{\Sigma}^*|}{|\boldsymbol{\Sigma}|}\right)^{\frac{1}{2}} \exp\left\{n\left[L^*\left(\widehat{\boldsymbol{\phi}}^*\right) - L\left(\widehat{\boldsymbol{\phi}}\right)\right]\right\} \tag{7}$$

where $\widehat{\boldsymbol{\phi}}^*$ and $\widehat{\boldsymbol{\phi}}$ maximize $L^*(\boldsymbol{\phi})$ and $L(\boldsymbol{\phi})$, respectively, and $\boldsymbol{\Sigma}^*$ and $\boldsymbol{\Sigma}$ are minus the inverse Hessians of $L^*(\boldsymbol{\phi})$ and $L(\boldsymbol{\phi})$ evaluated at $\widehat{\boldsymbol{\phi}}^*$ and $\widehat{\boldsymbol{\phi}}$, respectively.

However, if the function $g(\cdot)$ is not a positive function bounded away from zero then applying Laplace's approximation technique to the numerator of Eq. (6) may not be appropriate. Hence, the approximation method from Eq. (7) for calculating posterior moments is not directly applicable. Tierney et al. (1989) suggested that we may overcome this by considering the moment generating function $E\left[\exp\{g(\boldsymbol{\phi})\, t\}\right]$ where $|t| < h$ for some $h > 0$, since $\exp\{g(\boldsymbol{\phi})\, t\}$ is a nonnegative function. Therefore, for our case here we can calculate the posterior moment generating function of $g(\boldsymbol{\alpha})$ via

$$M_{g(\boldsymbol{\alpha})\,|\,\mathbf{y}}(t) = \frac{\int_{\boldsymbol{\alpha}} \exp\{g(\boldsymbol{\alpha})\, t\}\,\overline{\pi}(\boldsymbol{\alpha}\,|\,\mathbf{y})\;\mathrm{d}\boldsymbol{\alpha}}{\int_{\boldsymbol{\alpha}} \overline{\pi}(\boldsymbol{\alpha}\,|\,\mathbf{y})\;\mathrm{d}\boldsymbol{\alpha}} \tag{8}$$

where $\overline{\pi}(\boldsymbol{\alpha}\,|\,\mathbf{y})$ is defined in Eq. (4). Numerically we may calculate posterior moments of $g(\boldsymbol{\alpha})$ by making use of the definition of the derivative for the posterior moment

generating function of $g(\boldsymbol{\alpha})$. For a sufficiently small value of t we have

$$E\left[g(\boldsymbol{\alpha}) \mid \mathbf{y}\right] = \lim_{t \to 0} \left[\frac{M_{g(\boldsymbol{\alpha}) \mid \mathbf{y}}(t) - M_{g(\boldsymbol{\alpha}) \mid \mathbf{y}}(0)}{t} \right] \approx \frac{1}{t} \left[\frac{\int_{\boldsymbol{\alpha}} \exp\{g(\boldsymbol{\alpha})\, t\}\, \overline{\pi}(\boldsymbol{\alpha} \mid \mathbf{y})\ \mathrm{d}\boldsymbol{\alpha}}{\int_{\boldsymbol{\alpha}} \overline{\pi}(\boldsymbol{\alpha} \mid \mathbf{y})\ \mathrm{d}\boldsymbol{\alpha}} - 1 \right]. \tag{9}$$

However, the integrals involved in Eq. (9) are not tractable. For the integral in the denominator, an obvious choice is to consider the approximate posterior distribution as given in Eq. (5). The integral in the numerator will require a generalization of the finite sample likelihood approximation technique as was previously outlined above and will be discussed in greater detail below.

Recall that the approximate posterior distribution for $\boldsymbol{\alpha}$ as given in Eq. (5) is of a multivariate normal form. Thus, the integral in the denominator of (9) can be approximated by

$$\int_{\boldsymbol{\alpha}} \overline{\pi}(\boldsymbol{\alpha} \mid \mathbf{y})\ \mathrm{d}\boldsymbol{\alpha} \approx \int_{\boldsymbol{\alpha}} \overline{\pi}^*(\boldsymbol{\alpha} \mid \mathbf{y})\ \mathrm{d}\boldsymbol{\alpha} = e^{-\frac{np}{2}} |\mathbf{S}|^{-\frac{n}{2}} |\mathbf{Q}|^{-\frac{1}{2}} \tag{10}$$

where $\overline{\pi}(\boldsymbol{\alpha} \mid \mathbf{y})$ and $\overline{\pi}^*(\boldsymbol{\alpha} \mid \mathbf{y})$ are as in Eqs. (4) and (5), respectively, and $\mathbf{Q}$ is defined in Eq. (3). The integral in the numerator of Eq. (9)

$$\int_{\boldsymbol{\alpha}} \exp\{g(\boldsymbol{\alpha})\, t\}\, \overline{\pi}(\boldsymbol{\alpha} \mid \mathbf{y})\ \mathrm{d}\boldsymbol{\alpha} = \int_{\boldsymbol{\alpha}} (2\pi)^{-\frac{np}{2}} \exp\left\{ g(\boldsymbol{\alpha})\, t - \frac{n}{2} \mathrm{tr}\left[\mathbf{A} + \mathbf{S} \exp\{-\mathbf{A}\}\right] \right\}\ \mathrm{d}\boldsymbol{\alpha} \tag{11}$$

is more involved. We require a generalization of the finite sample likelihood function approximation technique as originally set forth by Leonard and Hsu (1992). To approximate the term $\exp\{-\mathbf{A}\}$ they consider $\exp\{-\mathbf{A}\omega\}$ where $0 < \omega < \infty$. Bellman (1970, p. 171) shows that $\exp\{-\mathbf{A}\omega\} = \mathbf{X}(\omega)$ where $\mathbf{X}(\omega)$ satisfies the linear Volterra integral equation

$$\mathbf{X}(\omega) = \mathbf{S}^{*-\omega} - \int_0^{\omega} \left(\mathbf{A} - \boldsymbol{\Lambda}^*\right) \mathbf{X}(v)\ \mathrm{d}v \quad 0 < \omega < \infty \tag{12}$$

and $\boldsymbol{\Lambda}^* = \log(\mathbf{S}^*)$ maximizes the integrand of Eq. (11) with respect to $\mathbf{A}$. A Taylor series expansion of $\mathbf{X}(\omega)$, about $\mathbf{A} = \boldsymbol{\Lambda}^*$, is now available by successively substituting the right side of Eq. (12) for the function $\mathbf{X}(\cdot)$ in the integrand. When $\omega = 1$, this

yields

$$\begin{aligned}\exp\{-\mathbf{A}\} = \mathbf{X}(1) = \mathbf{S}^{*-1} &- \int_0^1 \mathbf{S}^{*v-1}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{S}^{*-v}\,\mathrm{d}v \\ &+ \int_0^1\int_0^v \mathbf{S}^{*v-1}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{S}^{*u-v}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{S}^{*-u}\,\mathrm{d}u\,\mathrm{d}v \\ &+ \text{cubic and higher order terms.}\end{aligned}$$

Ignoring the cubic and higher order terms of this expansion for $\exp\{-\mathbf{A}\}$ we have

$$\begin{aligned}\mathrm{tr}\left[\mathbf{S}\exp\{-\mathbf{A}\}\right] \approx \mathrm{tr}\left[\mathbf{S}\mathbf{S}^{*-1}\right] &- \int_0^1 \mathrm{tr}\left[\mathbf{S}\mathbf{S}^{*v-1}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{S}^{*-v}\right]\mathrm{d}v \\ &+ \int_0^1\int_0^v \mathrm{tr}\left[\mathbf{S}\mathbf{S}^{*v-1}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{S}^{*u-v}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{S}^{*-u}\right]\mathrm{d}u\,\mathrm{d}v. \qquad (13)\end{aligned}$$

Define the spectral decomposition of $\mathbf{S}^* = \mathbf{E}^*\mathbf{D}^*\mathbf{E}^{*\mathrm{T}}$ where $\mathbf{E}^*$ is a matrix of the normalized eigenvectors of $\mathbf{S}^*$ and $\mathbf{D}^*$ is a diagonal matrix of the eigenvalues of $\mathbf{S}^*$. Furthermore, if we let $\mathbf{B} = \mathbf{E}^{*\mathrm{T}}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{E}^*$ and $\mathbf{C} = \mathbf{E}^{*\mathrm{T}}\mathbf{S}\mathbf{E}^*$ then the single integral in Eq. (13) can be expressed as

$$\begin{aligned}&\int_0^1 \mathrm{tr}\left[\mathbf{S}\mathbf{S}^{*v-1}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{S}^{*-v}\right]\mathrm{d}v \\ &= \int_0^1 \mathrm{tr}\left[\mathbf{D}^{*v-1}\mathbf{B}\mathbf{D}^{*-v}\mathbf{C}\right]\mathrm{d}v \\ &= \int_0^1 \sum_{i=1}^p\sum_{j=1}^p d_i^{*v-1}d_j^{*-v}b_{ij}c_{ji}\,\mathrm{d}v \\ &= \sum_{i=1}^p \frac{b_{ii}c_{ii}}{d_i^*} + \sum_{i,j:i\neq j}^p \frac{b_{ij}c_{ji}\left(d_i^*-d_j^*\right)}{d_i^*d_j^*\left[\log\left(d_i^*\right)-\log\left(d_j^*\right)\right]} \qquad (14)\end{aligned}$$

where d_i^* denotes the ith eigenvalue of $\mathbf{S}^*$ and b_{ij} and c_{ij} denote the $(i,j)\,th$ element of the matrices $\mathbf{B}$ and $\mathbf{C}$, respectively. Recall that $\mathbf{B} = \mathbf{E}^{*\mathrm{T}}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{E}^*$ so that $b_{ij} = \mathbf{e}_i^{*\mathrm{T}}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{e}_j^*$ where $\mathbf{e}_i^*$ is the ith normalized eigenvector of $\mathbf{S}^*$. Analogous to $\mathbf{f}_{ij}$ from Sect. 3.2, define the operator $\mathbf{f}_{ij}^* = \mathbf{e}_i^* * \mathbf{e}_j^*$ where $\mathbf{e}_i^* * \mathbf{e}_j^*$ is the $(q\times 1)$ vector that satisfies the condition $\boldsymbol{\alpha}^{\mathrm{T}}\left(\mathbf{e}_i^* * \mathbf{e}_j^*\right) = \mathbf{e}_i^{*\mathrm{T}}\mathbf{A}\mathbf{e}_j^*$. Therefore, making use of these definitions we can express the last line of Eq. (14) as

$$\int_0^1 \mathrm{tr}\left[\mathbf{S}\mathbf{S}^{*v-1}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{S}^{*-v}\right]\mathrm{d}v = \left(\boldsymbol{\alpha}-\boldsymbol{\lambda}^*\right)^{\mathrm{T}}\mathbf{L} \qquad (15)$$

where $\boldsymbol{\lambda}^* = \text{Vec}^*(\boldsymbol{\Lambda}^*)$ and $\mathbf{L} = \sum_{i=1}^{p} \frac{c_{ii}}{d_i^*}\mathbf{f}_{ii}^* + \sum_{i,j:i\neq j}^{p} \frac{c_{ji}}{d_j^*\log(d_i^*) - \log\left(d_j^*\right)}\mathbf{f}_{ij}^*$. The integrand in the double integral (13) can be simplified using the same notation as above $\text{tr}\left[\mathbf{S}\mathbf{S}^{*v-1}(\mathbf{A}-\boldsymbol{\Lambda}^*)\mathbf{S}^{*u-v}(\mathbf{A}-\boldsymbol{\Lambda}^*)\mathbf{S}^{*-u}\right] = \text{tr}\left[\mathbf{D}^{*u-v}\mathbf{B}\mathbf{D}^{*-u}\mathbf{C}\mathbf{D}^{*v-1}\mathbf{B}\right]$. Making use of the fact that $b_{ij} = \mathbf{e}_i^{*\mathrm{T}}(\mathbf{A}-\boldsymbol{\Lambda}^*)\mathbf{e}_j^*$ and the operator $\mathbf{f}_{ij}^*$ we can express the double integral as

$$\begin{aligned}&\int_0^1\int_0^v \text{tr}\left[\mathbf{S}\mathbf{S}^{*v-1}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{S}^{*u-v}\left(\mathbf{A}-\boldsymbol{\Lambda}^*\right)\mathbf{S}^{*-u}\right]\,\mathrm{d}u\,\mathrm{d}v\\&= \left(\boldsymbol{\alpha}-\boldsymbol{\lambda}^*\right)^{\mathrm{T}}\mathbf{Q}^*\left(\boldsymbol{\alpha}-\boldsymbol{\lambda}^*\right),\end{aligned} \tag{16}$$

where

$$\begin{aligned}\underset{(q\times q)}{\mathbf{Q}^*} &= \sum_{i=1}^{p}\xi_i^{(1)}\mathbf{f}_{ii}^*\mathbf{f}_{ii}^{*\mathrm{T}} + \sum_{i,l:i\neq l}^{p}\xi_{il}^{(2)}\mathbf{f}_{il}^*\mathbf{f}_{ii}^{*\mathrm{T}} + \sum_{i,j:i\neq j}^{p}\xi_{ij}^{(3)}\mathbf{f}_{ij}^*\mathbf{f}_{ji}^{*\mathrm{T}}\\&\quad + \sum_{i,j:i\neq j}\xi_{ij}^{(4)}\mathbf{f}_{ii}^*\mathbf{f}_{ji}^{*\mathrm{T}} + \sum_{i,j,l:i\neq j\neq l}^{p}\xi_{ijl}^{(5)}\mathbf{f}_{il}^*\mathbf{f}_{ji}^{*\mathrm{T}},\end{aligned} \tag{17}$$

and

$$\xi_i^{(1)} = \frac{c_{ii}}{2d_i^*},\ \xi_{il}^{(2)} = \frac{c_{li}}{d_i^*}\left[\frac{1}{\left[\log\left(\frac{d_i^*}{d_l^*}\right)\right]^2}\left(\frac{d_i^*}{d_l^*}-1\right) - \frac{1}{\log\left(\frac{d_i^*}{d_l^*}\right)}\right],$$

$$\xi_{ij}^{(3)} = \frac{c_{jj}}{d_j^*}\left[\frac{1}{\log\left(\frac{d_i^*}{d_j^*}\right)} + \frac{1}{\left[\log\left(\frac{d_i^*}{d_j^*}\right)\right]^2}\left(\frac{d_j^*}{d_i^*}-1\right)\right],$$

$$\xi_{ij}^{(4)} = \frac{c_{ij}}{d_j^*}\left[\frac{d_j^*}{d_i^*}\log\left(\frac{d_j^*}{d_i^*}\right) - \frac{1}{\left[\log\left(\frac{d_j^*}{d_i^*}\right)\right]^2}\left(\frac{d_j^*}{d_i^*}-1\right)\right],$$

$$\xi_{ijl}^{(5)} = \frac{c_{lj}}{d_j^*}\left[\frac{1}{\log\left(\frac{d_i^*}{d_l^*}\right)\log\left(\frac{d_j^*}{d_l^*}\right)}\left(\frac{d_j^*}{d_l^*}-1\right) - \frac{1}{\log\left(\frac{d_i^*}{d_l^*}\right)\log\left(\frac{d_j^*}{d_i^*}\right)}\left(\frac{d_j^*}{d_i^*}-1\right)\right].$$

However, $\mathbf{Q}^*$ is not necessarily symmetric. If $\mathbf{Q}^*$ is in fact not symmetric we can always make it symmetric by simply replacing $\mathbf{Q}^*$ with $\frac{1}{2}\left(\mathbf{Q}^* + \mathbf{Q}^{*\mathrm{T}}\right)$, which will

always be symmetric. Combining the results from Eqs. (13), (15) and (16) we have the following expression for the approximation to the integrand in Eq. (11)

$$\exp\{g(\boldsymbol{\alpha})\,t\}\,\overline{\pi}(\boldsymbol{\alpha}\mid\mathbf{y}) \approx (2\pi)^{-\frac{np}{2}} \exp\left\{g(\boldsymbol{\alpha})\,t+\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{U}_A-\frac{n}{2}\mathrm{tr}\left[\mathbf{S}\mathbf{S}^{*-1}\right]\right.$$
$$\left.+\frac{n}{2}\left(\boldsymbol{\alpha}-\boldsymbol{\lambda}^{*}\right)^{\mathrm{T}}\mathbf{L}-\frac{n}{2}\left(\boldsymbol{\alpha}-\boldsymbol{\lambda}^{*}\right)^{\mathrm{T}}\mathbf{Q}^{*}\left(\boldsymbol{\alpha}-\boldsymbol{\lambda}^{*}\right)\right\} \quad (18)$$

where $\mathbf{U}_A$ is a $(q\times 1)$ vector with the first p elements equal to $-\frac{n}{2}$ and the remaining elements zero. For details concerning the approximation procedure refer to Hsu (2001). The posterior mean of $g(\boldsymbol{\alpha})$ can be approximated via the technique from Eq. (9) by utilizing the integral of Eq. (18) with respect to $\boldsymbol{\alpha}$ and Eq. (10). The integral over Eq. (18) with respect to $\boldsymbol{\alpha}$ can be calculated analytically for specific functional forms of $g(\boldsymbol{\alpha})$. Without loss of generality, assume the parameter of interest is α_1. We can take $g(\boldsymbol{\alpha})=\alpha_1=\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{c}_1$ where $\mathbf{c}_1$ is a $(q\times 1)$ vector with one in the first element and the remaining elements zero. Then in this case the integral in the numerator of Eq. (9) can be approximated by

$$\begin{aligned}
&\int_{\boldsymbol{\alpha}} e^{\alpha_1 t}\overline{\pi}(\boldsymbol{\alpha}\mid\mathbf{y})\;\mathrm{d}\boldsymbol{\alpha}\\
&\approx\int_{\boldsymbol{\alpha}}(2\pi)^{-\frac{np}{2}}\exp\left\{t\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{c}_1+\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{U}_A-\frac{n}{2}\mathrm{tr}\left[\mathbf{S}\mathbf{S}^{*-1}\right]+\frac{n}{2}\left(\boldsymbol{\alpha}-\boldsymbol{\lambda}^{*}\right)^{\mathrm{T}}\mathbf{L}\right.\\
&\quad\left.-\frac{n}{2}\left(\boldsymbol{\alpha}-\boldsymbol{\lambda}^{*}\right)^{\mathrm{T}}\mathbf{Q}^{*}\left(\boldsymbol{\alpha}-\boldsymbol{\lambda}^{*}\right)\right\}\;\mathrm{d}\boldsymbol{\alpha}\\
&\approx\left|n\mathbf{Q}^{*}\right|^{-\frac{1}{2}}\exp\left\{-\frac{n}{2}\mathrm{tr}\left[\mathbf{S}\mathbf{S}^{*-1}\right]+\boldsymbol{\lambda}^{*\mathrm{T}}\left(\mathbf{U}_A+\mathbf{c}_1 t\right)\right.\\
&\quad\left.+\frac{1}{8}\mathbf{U}_1^{\mathrm{T}}\left(n\mathbf{Q}^{*}\right)^{-1}\mathbf{U}_1\right\},
\end{aligned} \quad (19)$$

where $\mathbf{U}_1=-2\left(\mathbf{U}_A+\mathbf{c}_1 t\right)-n\mathbf{L}$. Therefore, based upon the results from Eqs. (9), (10) and (19), we can approximate the posterior mean of α_1, $E\left[\alpha_1\mid\mathbf{y}\right]$, by

$$\frac{1}{t}\left[\frac{\left|n\mathbf{Q}^{*}\right|^{-\frac{1}{2}}}{e^{-\frac{np}{2}}\left|\mathbf{S}\right|^{-\frac{n}{2}}\left|\mathbf{Q}\right|^{-\frac{1}{2}}}\exp\left\{-\frac{n}{2}\mathrm{tr}\left[\mathbf{S}\mathbf{S}^{*-1}\right]+\boldsymbol{\lambda}^{*\mathrm{T}}\left(\mathbf{U}_A+\mathbf{c}_1 t\right)+\frac{1}{8}\mathbf{U}_1^{\mathrm{T}}\left(n\mathbf{Q}^{*}\right)^{-1}\mathbf{U}_1\right\}-1\right] \quad (20)$$

where t is chosen sufficiently small. We found $t=10^{-6}$ to be sufficient. In practice, we may successively employ smaller values of t until the estimates do not change within the third significant digit. For instance, $t=10^{-6}$ and $t=10^{-7}$ produced the same value for $E\left[\alpha_1\mid\mathbf{y}\right]$ within the third significant digit. The posterior variance can be calculated in a similar fashion.

It is important to point out that all posterior estimates of the covariance matrix are guaranteed to be positive definite. This is true because upon taking the matrix exponential transformation based upon the spectral decomposition the eigenvalues must all be positive since the exponential function is a strictly positive function. Furthermore,

Table 1 Posterior mean of **A** with vague prior via Laplacian approximation

	1	2	3	4	5	6	7	8
1	−5.287	0.693	0.378	0.475	0.395	0.298	0.019	0.405
2	0.693	−4.631	0.304	0.504	0.139	0.219	0.355	0.056
3	0.378	0.304	−5.631	0.390	0.047	−0.034	0.091	0.527
4	0.475	0.504	0.390	−4.055	0.453	0.189	0.306	0.441
5	0.395	0.139	0.047	0.453	−3.707	0.363	0.190	0.210
6	0.298	0.219	−0.034	0.189	0.363	−3.937	0.438	0.352
7	0.019	0.355	0.091	0.306	0.190	0.438	−3.593	0.316
8	0.405	0.056	0.527	0.441	0.210	0.352	0.316	−4.097

it should be noted that it is generally the case that Bayesian estimators approach maximum likelihood estimators as the sample size increases in nitely. In turn, maximum likelihood estimators are consistent estimators. Thus, not only are we guaranteed a positive definite estimator we are also assured a consistent estimator.

4.2 Project talent data set description

Throughout we will make use of the project talent data set (Flanagan and Tiedeman 1979) for application of the specific computational techniques. The project talent data has been extensively analyzed by Cooley and Lohnes (1971, p. 14) and a more detailed description of the data set can be found therein. The project talent survey was administered to high school students throughout the United States. A stratified random sample of the country's high schools was carefully obtained to ensure adequate representation of all types of schools (Cooley and Lohnes 1971, p. 14). Here we will utilize a subset of the data from the Project Talent study. In particular, we use the data generated from 18-year-old individuals who participated in the survey of which there were 70, 80 students.

Specifically, we will make use of eight standardized test scores in a number of fields including two general informational tests (parts I and II), an English test, a reading and comprehension test, creativity test, a mechanical reasoning test, an abstract reasoning test and a mathematics test. Please refer to Table 1 for the posterior mean of **A** calculated via Laplacian approximation under the vague prior specification.

In comparison to the importance sampling estimation procedure as described in Leonard and Hsu (1992), the Laplacian approximation technique produces quite similar numerical results as we would expect. The great advantage of the Laplacian approximation technique is its rapidity. Since no simulation routines are involved, posterior moments can be computed virtually instantaneously with modern statistical software. The disadvantage of this methodology, as its name implies, is that the technique is of course approximate. However, judging from the results of the two methods the approximation seems to be reasonable. In contrast, the Importance sampling routine is in fact an exact technique. However, its main disadvantage is the length of time to

convergence. Depending upon the particular application this can take quite a bit of computational time and resources. In our particular application we obtained convergence after two million simulations which took approximately three hours of runtime. We used the statistical and mathematical programming language R. The specifications of the hardware that the computations were made on are the following: Dual Quad-Core Intel Xeon with central processing unit E5450 at 3.0 GHz (8 cores) with 32 GB of memory, SAS hard drive and running MOSIX clustering software with ten nodes. The other nodes are comprised of Dual 2.2 GHz Intel Xeon processors with 2 GB of random access memory.

5 Hierarchical prior specification

We have already seen that the approximate likelihood function for $\boldsymbol{\alpha}$ possesses a multivariate normal form. Combining the multivariate normal approximate likelihood function with a multivariate normal prior distribution for $\boldsymbol{\alpha}$ will of course result in an multivariate normal approximate posterior distribution. In this way, the approximate likelihood function allows a multivariate normal prior specification to act as a conjugate prior for the covariance structure. In general, we will assume a priori that $\boldsymbol{\alpha} \mid \boldsymbol{\eta}, \boldsymbol{\Upsilon} \sim \mathrm{N}_q(\boldsymbol{\eta}, \boldsymbol{\Upsilon})$ where $\boldsymbol{\eta}$ is a $(q \times 1)$ prior mean location hyperparameter vector and $\boldsymbol{\Upsilon}$ is a $(q \times q)$ covariance hyperparameter matrix. In this way we can combine the approximate likelihood function (3) with a multivariate normal prior distribution to obtain a multivariate normal approximate posterior distribution for $\boldsymbol{\alpha}$. Thus, conjugacy is achieved, as will be demonstrated below, through use of the approximate likelihood function and a multivariate normal prior distribution.

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 while at the same time maintaining the tractability of conjugacy. 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 $\boldsymbol{\Upsilon}$. In contrast, the inverse Wishart prior is fully parameterized by a scale hyperparameter matrix and a single degree of freedom hyperparameter that only allows for one sole specification of the overall degree of confidence in all the elements of the hyperparameter scale matrix. In addition, with the multivariate normal prior we are able to model potential interdependency among the elements of $\boldsymbol{\alpha}$ because we can specify non-trivial 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 $\boldsymbol{\alpha}$. In comparison, the inverse Wishart provides no means by which to model interdependency within the covariance structure. In short, we are now able to craft a more complex and accurate prior specification for the covariance structure with the convenience and tractability of conjugacy.

5.1 Simulation study to investigate flexible prior specification

To further investigate the flexible prior specification we constructed the following simulation study. Suppose $\mathbf{Y}_1, \ldots, \mathbf{Y}_{100}$ constitutes a random sample of size $n = 100$ from a bivariate normal distribution with zero mean vector $\mathbf{0} = [0, 0]^{\mathrm{T}}$ and covariance matrix $\boldsymbol{\Sigma}$. Assume that the *true* prior distribution of $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3]^{\mathrm{T}} = [a, b, c]^{\mathrm{T}}$ is a multivariate normal distribution with mean vector equal to $\boldsymbol{\eta} = [1, 9, 2]^{\mathrm{T}}$ and covariance matrix $\boldsymbol{\Upsilon}$, where

$$\underset{(2\times 2)}{\mathbf{A}} = \log(\boldsymbol{\Sigma}) = \begin{bmatrix} a & c \\ c & b \end{bmatrix} \quad \text{and} \quad \underset{(3\times 3)}{\boldsymbol{\Upsilon}} = \begin{bmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 4 \end{bmatrix}.$$

Note that under this prior specification a and b are highly correlated and the variance of c is significantly larger than the variances of a and b. We make use of the pseudoentropy loss with respect to $\boldsymbol{\Sigma}$ which is given by $L_{\boldsymbol{\Sigma}1}\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) = \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\right) - \log\left|\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\right| - p$, where in this case $p = 2$ is the dimension of $\boldsymbol{\Sigma}$. As discussed above under this loss function the Bayes estimator that minimizes the Bayes risk is given by the posterior mean. Posterior means using the inverse Wishart priors with the mean equal to the true prior mean $E[\boldsymbol{\Sigma}]$ and a wide range of degree of freedom parameters ν are compared to the posterior mean when the true prior is used. Please note that the inverse Wishart mean is valid only when $\nu - p - 1 = \nu - 3 > 0$.

In our study, a random sample of 1,000 $\boldsymbol{\alpha}$'s were simulated from the true prior multivariate normal distribution given by $N_3(\boldsymbol{\eta}, \boldsymbol{\Upsilon})$. Then for each simulated $\boldsymbol{\alpha}$, or equivalently $\boldsymbol{\Sigma} = e^{\mathbf{A}}$, a random sample of 100 $\mathbf{y}$'s were simulated from a multivariate normal distribution given by $N_2(\mathbf{0}, \boldsymbol{\Sigma})$. Five Bayes estimators were considered for comparison: $\widehat{\boldsymbol{\Sigma}}_1$ is the posterior mean, when the true prior is used, $\widehat{\boldsymbol{\Sigma}}_2$, $\widehat{\boldsymbol{\Sigma}}_3$, $\widehat{\boldsymbol{\Sigma}}_4$ and $\widehat{\boldsymbol{\Sigma}}_5$ are the posterior means using the inverse Wishart priors, with the mean equal to the prior mean of $\boldsymbol{\Sigma}$, and $\nu = 3.1, 4, 10$ and 100 degrees of freedom, respectively. The average losses over 1,000 $\boldsymbol{\alpha}$'s were 0.02904, 0.91566, 2.66477, 4.47883 and 6.65841 for $\widehat{\boldsymbol{\Sigma}}_1, \ldots, \widehat{\boldsymbol{\Sigma}}_5$, respectively. The averaged losses using the inverse Wishart priors, which possess the same mean as the true mean for a wide range of ν's, are significantly higher than the one using the true prior. This indicates that the inverse Wishart family is restrictive and cannot fully describe the situation when we have different degrees of confidence for the covariance components or when the components are correlated.

5.2 Flexible prior specification

A subjective Bayesian may in fact wish to specify all $q + \frac{1}{2}q(q+1)$ 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 so called 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 $\boldsymbol{\alpha} \mid \boldsymbol{\mu}, \boldsymbol{\Delta} \sim \mathrm{N}_q (\mathbf{J}\boldsymbol{\mu}, \boldsymbol{\Delta})$ for the prior distribution. We have the following prior distributional form,

$$\pi (\boldsymbol{\alpha} \mid \boldsymbol{\mu}, \boldsymbol{\Delta}) \propto |\boldsymbol{\Delta}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\boldsymbol{\alpha} - \mathbf{J}\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Delta}^{-1} (\boldsymbol{\alpha} - \mathbf{J}\boldsymbol{\mu}) \right\} \tag{21}$$

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

$$\underset{(q \times 2)}{\mathbf{J}} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix}^{\mathrm{T}} \qquad \underset{(q \times q)}{\boldsymbol{\Delta}} = \begin{bmatrix} \sigma_1^2 \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \mathbf{I}_{q-p} \end{bmatrix}. \tag{22}$$

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. 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 location hyperparameters and two different levels of confidence in our choice of location hyperparameters for the variance components separately from the covariance components.

It should be noted that this particular prior specification relies upon an exchangeability condition for the random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. In our particular application, it is reasonable to assume that the standardized test scores for various subject area exams from the project talent (Flanagan and Tiedeman 1979) data set satisfy this requirement.

For the hyperparameters $\boldsymbol{\mu} = [\mu_1, \mu_2]^{\mathrm{T}}$ and $\boldsymbol{\Delta} = h\left(\sigma_1^2, \sigma_2^2\right)$ we will assume a vague prior distribution $\pi (\boldsymbol{\mu}, \boldsymbol{\Delta}) \propto 1$. Note here that the vague 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 non-trivial specifications quite easily.

Having stated all the prior distributional assumptions we turn to the posterior Bayesian analysis. We begin this by first examining the exact joint posterior distribution. The complexity of the functional form of the exact joint posterior will motivate our consideration of the computational procedures described in detail below.

5.3 Exact joint posterior distribution

The exact joint posterior distribution for all parameters and hyperparameters will be proportional to the product of Eq. (2) the exact likelihood function, Eq. (21) the prior distribution for $\boldsymbol{\alpha}$ and the vague prior distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\Delta}$. Note that here we will use $\boldsymbol{\alpha}$ interchangeably with $\mathbf{A}$.

$$\pi\left(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Delta} \mid \mathbf{y}\right) \propto \left|\boldsymbol{\Delta}\right|^{-\frac{1}{2}} \exp\left\{-\frac{n}{2} \operatorname{tr}\left[\mathbf{A} + \mathbf{S}\exp\left\{-\mathbf{A}\right\}\right] - \frac{1}{2}\left(\boldsymbol{\alpha} - \mathbf{J}\boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Delta}^{-1}\left(\boldsymbol{\alpha} - \mathbf{J}\boldsymbol{\mu}\right)\right\}.$$

We clearly see that the exact joint posterior distribution is in fact not analytically tractable. This is the driving motivation behind the implementation of the numerical techniques.

5.4 Approximate conditional posterior distribution for $\boldsymbol{\alpha}$ Given $\boldsymbol{\Delta}$

The prior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$ can be obtained by integrating over Eq. (21) with respect to $\boldsymbol{\mu}$. The resulting prior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$ is

$$\pi\left(\boldsymbol{\alpha}, \boldsymbol{\Delta}\right) \propto \left|\boldsymbol{\Delta}\right|^{-\frac{1}{2}} \left|\mathbf{J}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\mathbf{J}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{G}^{*}\boldsymbol{\alpha}\right\} \tag{23}$$

where $\underset{(q\times q)}{\mathbf{G}^{*}} = \left[\mathbf{I}_q - \mathbf{J}\left(\mathbf{J}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\right]^{\mathrm{T}} \boldsymbol{\Delta}^{-1}\left[\mathbf{I}_q - \mathbf{J}\left(\mathbf{J}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\right]$ and $\mathbf{I}_q$ is a $(q \times q)$ identity matrix. Note that with respect to $\boldsymbol{\alpha}$ Eq. (23) is a multivariate normal form. The approximate joint posterior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$ will be proportional to the product of the approximate likelihood function (3) and the joint prior distribution (23).

$$\begin{aligned}\pi^{*}\left(\boldsymbol{\alpha}, \boldsymbol{\Delta} \mid \mathbf{y}\right) \propto & \left(\sigma_1^2\right)^{-\frac{p-1}{2}} \left(\sigma_2^2\right)^{-\frac{q-p-1}{2}} \\ & \times \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\alpha} - \boldsymbol{\lambda}\right)^{\mathrm{T}}\mathbf{Q}\left(\boldsymbol{\alpha} - \boldsymbol{\lambda}\right) + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{G}^{*}\boldsymbol{\alpha}\right]\right\}\end{aligned} \tag{24}$$

Recall that $\boldsymbol{\lambda} = \mathrm{Vec}^{*}\left(\boldsymbol{\Lambda}\right)$ where $\boldsymbol{\Lambda}$ the matrix logarithm of the sample covariance matrix as defined above. We complete the square for the terms in the exponent of (24). Subsequent to that a proportionality is taken with respect to the terms that involve $\boldsymbol{\alpha}$ to yield the following approximate posterior distribution for $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\Delta}$.

$$\pi^{*}\left(\boldsymbol{\alpha} \mid \mathbf{y}, \boldsymbol{\Delta}\right) \propto \overline{\pi}^{*}\left(\boldsymbol{\alpha} \mid \mathbf{y}, \boldsymbol{\Delta}\right) = \exp\left\{-\frac{1}{2}\left(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{*}\right)^{\mathrm{T}}\left(\mathbf{Q} + \mathbf{G}^{*}\right)\left(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{*}\right)\right\} \tag{25}$$

where the $(q \times 1)$ vector $\boldsymbol{\alpha}^{*} = \left(\mathbf{Q} + \mathbf{G}^{*}\right)^{-1}\mathbf{Q}\boldsymbol{\lambda}$. Recall that $\mathbf{Q}$ and $\mathbf{G}^{*}$ are as defined in (3) and (23), respectively. Thus, we have the following approximate posterior distribution for $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\Delta}$.

$$\boldsymbol{\alpha} \mid \mathbf{y}, \boldsymbol{\Delta} \sim \mathrm{N}_q\left(\boldsymbol{\alpha}^{*}, \left(\mathbf{Q} + \mathbf{G}^{*}\right)^{-1}\right) \tag{26}$$

This demonstrates the conjugacy of utilizing the approximate likelihood function. Approximate posterior moments for $\boldsymbol{\alpha}$ can easily be calculated. In addition, numerical methods such as Importance sampling, Laplacian approximation and MCMC procedures can readily be implemented by making use of (26). In short, we have developed

a highly flexible while at the same time tractable Bayesian methodology for the covariance structure.

5.5 Exact conditional posterior distribution for $\boldsymbol{\alpha}$ given $\boldsymbol{\Delta}$

As previously mentioned above since our MCMC procedure makes use of an approximate posterior distribution for $\boldsymbol{\alpha}$ it is appropriate to include a Metropolis–Hastings accept reject algorithm with respect to simulated candidate values for $\boldsymbol{\alpha}$ (Gelman et al. 2005, p. 291). To implement the Metropolis–Hastings algorithm we need the exact posterior distribution for $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\Delta}$. This exact posterior distribution will in fact be proportional to the product of (2) the exact likelihood function multiplied by (23) the joint prior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$. Note that again here we will use $\boldsymbol{\alpha}$ interchangeably with $\mathbf{A}$.

$$\pi\left(\boldsymbol{\alpha} \mid \mathbf{y}, \boldsymbol{\Delta}\right) \propto \overline{\pi}\left(\boldsymbol{\alpha} \mid \mathbf{y}, \boldsymbol{\Delta}\right) = \exp\left\{-\frac{n}{2}\ \mathrm{tr}\ \left[\mathbf{A} + \mathbf{S}\exp\left\{-\mathbf{A}\right\}\right] - \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{G}^{*}\boldsymbol{\alpha}\right\} \tag{27}$$

In contrast to (25), we clearly see that (27) is not a multivariate normal form with respect to $\boldsymbol{\alpha}$. Analytically calculating posterior moments for $\boldsymbol{\alpha}$ based upon Eq. (27) is not feasible.

5.6 Importance sampling under hierarchical prior specification

To implement both the importance sampling and the Laplacian approximation procedures we require the posterior distribution for $\boldsymbol{\Delta}$. However, the exact posterior distribution for $\boldsymbol{\Delta}$ is analytically difficult to obtain. Therefore, we consider the approximate posterior distribution for $\boldsymbol{\Delta}$ which can be obtained by integrating over the joint approximate posterior distribution with respect to both $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$. The joint approximate posterior distribution will be proportional to the product of Eq. (3) the approximate likelihood function, Eq. (21) the prior distribution for $\boldsymbol{\alpha}$ given $\boldsymbol{\mu}$ and $\boldsymbol{\Delta}$ and the vague prior distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\Delta}$. Thus, we have

$$\begin{aligned}\pi^{*}\left(\boldsymbol{\Delta}|\,\mathbf{y}\right) &\propto \int_{\boldsymbol{\alpha}}\int_{\boldsymbol{\mu}} |\mathbf{Q}|^{\frac{1}{2}}\,|\boldsymbol{\Delta}|^{-\frac{1}{2}} \\ &\quad\times \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\alpha}-\boldsymbol{\lambda})^{\mathrm{T}}\,\mathbf{Q}\,(\boldsymbol{\alpha}-\boldsymbol{\lambda}) + (\boldsymbol{\alpha}-\mathbf{J}\boldsymbol{\mu})^{\mathrm{T}}\,\boldsymbol{\Delta}^{-1}\,(\boldsymbol{\alpha}-\mathbf{J}\boldsymbol{\mu})\right]\right\}\,\mathrm{d}\boldsymbol{\mu}\,\mathrm{d}\boldsymbol{\alpha} \\ &\propto |\boldsymbol{\Delta}|^{-\frac{1}{2}}\left|\mathbf{J}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\mathbf{J}\right|^{-\frac{1}{2}}\left|\mathbf{Q}+\mathbf{G}^{*}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Q}\left(\mathbf{Q}+\mathbf{G}^{*}\right)^{-1}\mathbf{G}^{*}\boldsymbol{\lambda}\right\}\end{aligned}$$

where $\mathbf{Q}$ and $\mathbf{G}^{*}$ have been previously defined in Eqs. (3) and (23), respectively. The constant of integration for the above posterior density can be obtained via numerical techniques. Please refer to Ogata (1989) for a discussion of numerical integration in higher dimensionality.

Now under the hierarchical prior specification the Importance sampling procedure will have to be performed via iterated expectations $E\left[g\left(\boldsymbol{\alpha}\right) \mid \mathbf{y}\right] = E_{\boldsymbol{\Delta} \mid \mathbf{y}}\left[E\left[g\left(\boldsymbol{\alpha}\right) \mid \mathbf{y}, \boldsymbol{\Delta}\right]\right]$. We first calculate the inner expectation for various values of σ_1^2 and σ_2^2 chosen over a grid on $\mathbb{R}^2$, then perform two dimensional numerical integration with respect to σ_1^2 and σ_2^2 for the outer expectation.

5.7 Laplacian approximation under hierarchical prior specification

In this section, we demonstrate how we can again use the Laplacian approximation technique to calculate posterior moments under a hierarchical prior specification. The analysis here follows closely to that of Sect. 4.1. The key difference here is the presence of the hyperparameter $\boldsymbol{\Delta}$. The hyperparameter $\boldsymbol{\mu}$ does not play a role since we have already shown that it can be integrated out. As with the importance sampling routine we need to make use of the conditional posterior distribution of $\boldsymbol{\alpha}$ given $\boldsymbol{\Delta}$.

Note that the functional form for the posterior mean of $g\left(\boldsymbol{\alpha},\right)$ conditional on a given value of $\boldsymbol{\Delta}$, will be exactly the same as in Eq. (9) except that here we replace $\overline{\pi}\left(\boldsymbol{\alpha} \mid \mathbf{y}\right)$ with $\overline{\pi}\left(\boldsymbol{\alpha} \mid \mathbf{y}, \boldsymbol{\Delta}\right)$ as defined in Eq. (27). As we have previously done before we need to find the two maximizers, which we denote here as $\boldsymbol{\Lambda}_N^* = \log\left(\mathbf{S}_N^*\right)$ and $\boldsymbol{\Lambda}_D^* = \log\left(\mathbf{S}_D^*\right)$, of the integrands in both the numerator and denominator of the expression for the posterior moment generating function, respectively. Furthermore, we let $\boldsymbol{\lambda}_N^* = \text{Vec}^*\left(\boldsymbol{\Lambda}_N^*\right)$ and $\boldsymbol{\lambda}_D^* = \text{Vec}^*\left(\boldsymbol{\Lambda}_D^*\right)$.

Assume without loss of generality that the primary parameter of interest is α_1. Then, for various values of σ_1^2 and σ_2^2 chosen over a grid on $\mathbb{R}^2$, the required integrals can be approximated by

$$\int_{\boldsymbol{\alpha}} \overline{\pi}\left(\boldsymbol{\alpha} \mid \mathbf{y}, \boldsymbol{\Delta}\right) \, \mathrm{d}\boldsymbol{\alpha} \approx \left|n\mathbf{Q}_D^* + \mathbf{G}^*\right|^{-\frac{1}{2}} \exp\left\{ -\frac{n}{2}\text{tr}\left[\mathbf{S}\mathbf{S}_D^{*\,-1}\right] + \boldsymbol{\lambda}_D^{*\,\mathrm{T}}\mathbf{U}_A - \frac{1}{2}\boldsymbol{\lambda}_D^{*\,\mathrm{T}}\mathbf{G}^*\boldsymbol{\lambda}_D^* + \frac{1}{8}\mathbf{U}_3^{\mathrm{T}}\left(n\mathbf{Q}_D^* + \mathbf{G}^*\right)^{-1}\mathbf{U}_3 \right\} \tag{28}$$

$$\int_{\boldsymbol{\alpha}} e^{\alpha_1 t}\overline{\pi}\left(\boldsymbol{\alpha} \mid \mathbf{y}, \boldsymbol{\Delta}\right) \, \mathrm{d}\boldsymbol{\alpha} \approx \left|n\mathbf{Q}_N^* + \mathbf{G}^*\right|^{-\frac{1}{2}} \exp\left\{ -\frac{n}{2}\text{tr}\left[\mathbf{S}\mathbf{S}_N^{*\,-1}\right] + \boldsymbol{\lambda}_N^{*\,\mathrm{T}}\left(\mathbf{U}_A + \mathbf{c}_1 t\right) - \frac{1}{2}\boldsymbol{\lambda}_N^{*\,\mathrm{T}}\mathbf{G}^*\boldsymbol{\lambda}_N^* + \frac{1}{8}\mathbf{U}_4^{\mathrm{T}}\left(n\mathbf{Q}_N^* + \mathbf{G}^*\right)^{-1}\mathbf{U}_4 \right\} \tag{29}$$

where analogous to Eq. (17), $\mathbf{Q}_N^*$ and $\mathbf{Q}_D^*$ are functions of the normalized eigenvalues of $\boldsymbol{\Lambda}_N^*$ and $\boldsymbol{\Lambda}_D^*$, respectively, $\mathbf{U}_3 = -2\mathbf{U}_A - n\mathbf{L}_D + 2\mathbf{G}^*\boldsymbol{\lambda}_D^*$, $\mathbf{U}_4 = -2\mathbf{U}_A - 2t\mathbf{c}_1 - n\mathbf{L}_N - 4\lambda_{N1}^* t\mathbf{c}_1$, $\mathbf{L}_N$ and $\mathbf{L}_D$ are defined analogously to Eq. (15), but based on $\mathbf{S}_N^*$ and $\mathbf{S}_D^*$, respectively, $\mathbf{U}_A$ was defined in Eq. (18) and c_1 was defined in Eq. (19) and λ_{N1}^* is the first element of the vector $\boldsymbol{\lambda}_N^*$. Therefore, the posterior mean of α_1 conditional on $\boldsymbol{\Delta}$, that is $E\left[\alpha_1 \mid \mathbf{y}, \boldsymbol{\Delta}\right]$, can be approximated by utilizing Eqs. (28) and (29) in a similar fashion as was done above in Eq. (20) under the vague prior specification. We can then numerically integrate over $\boldsymbol{\Delta}$ in order to obtain the unconditional posterior mean

Table 2 Posterior mean of **A** with hierarchical prior via Laplacian approximation

	1	2	3	4	5	6	7	8
1	−5.224	0.539	0.345	0.426	0.367	0.302	0.127	0.367
2	0.539	−4.623	0.291	0.426	0.200	0.247	0.334	0.154
3	0.345	0.291	−5.611	0.341	0.122	0.072	0.148	0.433
4	0.426	0.426	0.341	−4.052	0.395	0.237	0.308	0.390
5	0.367	0.200	0.122	0.395	−3.755	0.336	0.231	0.248
6	0.302	0.247	0.072	0.237	0.336	−3.971	0.384	0.331
7	0.127	0.334	0.148	0.308	0.231	0.384	−3.641	0.312
8	0.367	0.154	0.433	0.390	0.248	0.331	0.312	−4.113

$$E\left[\alpha_1 \,|\, \mathbf{y}\right] = E_{\boldsymbol{\Delta}\,|\,\mathbf{y}}\left[E\left[\alpha_1 \,|\, \mathbf{y}, \boldsymbol{\Delta}\right]\right] \approx \int_{\sigma_1^2}\int_{\sigma_2^2} E\left[\alpha_1 \,|\, \mathbf{y}, \boldsymbol{\Delta}\right] \pi^*\left(\boldsymbol{\Delta} \,|\, \mathbf{y}\right)\, \mathrm{d}\sigma_1^2\, \mathrm{d}\sigma_2^2.$$

Please refer to Table 2 for the posterior mean of **A** calculated via Laplacian approximation under the hierarchical prior specification. The estimate presented in Table 2 exhibits the intra-class form, up to some rounding error, as we expect. To observe this first note that the mean of the on diagonal elements of Table 1 is approximately equal to −4.367. In comparison, the associated diagonal elements of Table 2 shrink towards this mean. Similarly, the off diagonal elements of Table 2 are pulled towards the mean of the off diagonal elements of Table 1 which is approximately equal to 0.304. In this way, we can observe the shrinkage impact of the intra-class hierarchical prior specification on the posterior mean.

5.8 Exact conditional posterior distribution for $\boldsymbol{\Delta}$ given $\boldsymbol{\alpha}$

To implement the MCMC procedures we require the exact posterior distribution for $\boldsymbol{\Delta}$ conditional on $\boldsymbol{\alpha}$. This will in fact be proportional to (23) the joint prior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$. Note that the exact likelihood function (2) does not depend upon $\boldsymbol{\Delta}$ and thus can be omitted entirely.

$$\begin{aligned}\pi\left(\boldsymbol{\Delta} \,|\, \mathbf{y}, \boldsymbol{\alpha}\right) &\propto |\boldsymbol{\Delta}|^{-\frac{1}{2}} \left|\mathbf{J}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\mathbf{J}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{G}^*\boldsymbol{\alpha}\right\} \\ &\propto \left(\sigma_1^2\right)^{-\frac{p-1}{2}} \left(\sigma_2^2\right)^{-\frac{q-p-1}{2}} \exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{p}(\alpha_i - \overline{\alpha}_v)^2 \right. \\ &\quad \left. -\frac{1}{2\sigma_2^2}\sum_{i=p+1}^{q}(\alpha_i - \overline{\alpha}_c)^2\right\}\end{aligned}$$

where $\overline{\alpha}_v = p^{-1}\sum_{i=1}^{p}\alpha_i$ and $\overline{\alpha}_c = (q-p)^{-1}\sum_{i=p+1}^{q}\alpha_i$ 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 density functions.

$$\sigma_1^2 \Big| \mathbf{y}, \boldsymbol{\alpha} \sim \text{Inverse Gamma}\left(\frac{p-3}{2}, \frac{1}{2}\sum_{i=1}^{p} (\alpha_i - \overline{\alpha}_v)^2 \right) \tag{30}$$

$$\sigma_2^2 \Big| \mathbf{y}, \boldsymbol{\alpha} \sim \text{Inverse Gamma}\left(\frac{q-p-3}{2}, \frac{1}{2}\sum_{i=p+1}^{q} (\alpha_i - \overline{\alpha}_c)^2 \right) \tag{31}$$

This result is theoretically appealing in that the posterior distribution for σ_1^2, the variance hyperparameter for the location hyperparameter μ_1, depends only on the variance terms $\alpha_1, \ldots, \alpha_p$. Whereas, the posterior distribution for σ_2^2, the variance hyperparameter for the location hyperparameter μ_2, depends only on the covariance terms $\alpha_{p+1}, \ldots, \alpha_q$. This draws out the intra-class matrix form wherein which we model 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.

5.9 Markov Chain Monte Carlo under hierarchical prior specification

Based upon the theoretical results derived above in this section we outline the procedure for implementing the MCMC algorithm. From Eqs. (26), (27), (30) and (31), we have a formal setup for implementing a MCMC procedure with a Metropolis–Hastings accept reject algorithm (Gelman et al. 2005, p. 291). Below we delineate the specific steps involved.

1. Simulate $\sigma_1^{2(t)}$ and $\sigma_2^{2(t)}$ from Eqs. (30) and (31), respectively.
2. Simulate a candidate value $\widetilde{\boldsymbol{\alpha}}$ from Eq. (26) based upon $\sigma_1^{2(t)}$ and $\sigma_2^{2(t)}$ from step one. Then let

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

where $\rho = \dfrac{\pi(\widetilde{\boldsymbol{\alpha}} \mid \mathbf{y}, \boldsymbol{\Delta}^{(t)})}{\pi^*(\widetilde{\boldsymbol{\alpha}} \mid \mathbf{y}, \boldsymbol{\Delta}^{(t)})} \Big/ \dfrac{\pi(\boldsymbol{\alpha}^{(t)} \mid \mathbf{y}, \boldsymbol{\Delta}^{(t)})}{\pi^*(\boldsymbol{\alpha}^{(t)} \mid \mathbf{y}, \boldsymbol{\Delta}^{(t)})}$ and $\pi^*(\cdot \mid \cdot)$ and $\pi(\cdot \mid \cdot)$ are as defined in (25) and (27), respectively. Please refer to Tables 3 and 4 for the posterior mean of $\mathbf{A}$ and $\boldsymbol{\Sigma}$, respectively, calculated via MCMC procedures under the hierarchical prior specification.

The chief advantage of the MCMC sampling algorithm is the relative ease of implementation. Since the analysis is done solely with respect to the conditional posterior distributions the theoretical Bayesian analysis is somewhat simplified. Most modern statistical software programs greatly facilitate MCMC sampling techniques since they include many of the popular distributions for simulation purposes. Despite the use of the approximate posterior distribution for $\boldsymbol{\alpha}$ as given in Eq. (25), which is of course

Table 3 Posterior mean of **A** with hierarchical prior via MCMC/Metropolis–Hastings algorithm

	1	2	3	4	5	6	7	8
1	−5.235	0.539	0.344	0.424	0.366	0.301	0.125	0.368
2	0.539	−4.634	0.289	0.426	0.200	0.247	0.333	0.153
3	0.344	0.289	−5.626	0.344	0.120	0.069	0.149	0.431
4	0.424	0.426	0.344	−4.065	0.396	0.237	0.307	0.390
5	0.366	0.200	0.120	0.396	−3.768	0.335	0.230	0.246
6	0.301	0.247	0.069	0.237	0.335	−3.985	0.384	0.332
7	0.125	0.333	0.149	0.307	0.230	0.384	−3.653	0.312
8	0.368	0.153	0.431	0.390	0.246	0.332	0.312	−4.130

Table 4 Posterior mean of **Σ** with hierarchical prior via MCMC/Metropolis–Hastings algorithm

	1	2	3	4	5	6	7	8
1	0.01382	0.01142	0.00675	0.01450	0.01317	0.01158	0.01113	0.01239
2	0.01142	0.01888	0.00710	0.01571	0.01255	0.01209	0.01418	0.01165
3	0.00675	0.00710	0.00763	0.00989	0.00751	0.00663	0.00785	0.00938
4	0.01450	0.01571	0.00989	0.03238	0.01905	0.01576	0.01807	0.01802
5	0.01317	0.01255	0.00751	0.01905	0.03489	0.01638	0.01591	0.01522
6	0.01158	0.01209	0.00663	0.01576	0.01638	0.02946	0.01768	0.01529
7	0.01113	0.01418	0.00785	0.01807	0.01591	0.01768	0.03794	0.01643
8	0.01239	0.01165	0.00938	0.01802	0.01522	0.01529	0.01643	0.02796

appropriately accounted for by inclusion of the Metropolis–Hastings accept reject algorithm, the final numerical results are exact. In contrast to the importance sampling and Laplacian approximation procedures the MCMC sampling estimates are in fact the most accurate. The reason for this is twofold. The first reason is due to the numerical integration that was involved in both the Importance sampling and the Laplacian approximation. The numerical integration can potentially be an added source of diminished accuracy. Note however that this is entirely avoided under the MCMC sampling routine. In addition, recall that numerical integration was over the approximate posterior distribution for $\boldsymbol{\Delta}$. The use of this approximate posterior distribution in its own right can also lead to a lack of accuracy. Recall that the MCMC sampling algorithm used the exact posterior distribution for $\boldsymbol{\Delta}$ conditional on $\boldsymbol{\alpha}$ thus avoiding the use of any further approximate posterior distributions. The numerical results of the MCMC algorithm seem quite reasonable in comparison to the other methods. The disadvantage of MCMC methods is of course the relatively lengthly time spent until satisfactory convergence is attained.

6 Hierarchical Bayesian analysis with unknown mean vector

In this section, we relax the assumption of a known mean vector. Formally, we treat the case where we have a random sample such that $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \mid \boldsymbol{\theta}, \boldsymbol{\Sigma} \overset{iid}{\sim} \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Here

$\boldsymbol{\theta}$ is an unknown p dimensional mean vector and $\boldsymbol{\Sigma}$ is a $(p \times p)$ unknown positive definite symmetric covariance matrix. Thus, we have the following familiar multivariate Normal likelihood function for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$.

$$l\left(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \mathbf{y}\right) = (2\pi)^{-\frac{np}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i - \boldsymbol{\theta}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_i - \boldsymbol{\theta}\right)\right\}$$

In addition to the prior specifications for $\boldsymbol{\alpha}$ and $\boldsymbol{\Delta}$ as previously stated above we further assume $\boldsymbol{\theta} \mid \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^* \sim \mathrm{N}_p\left(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*\right)$. As a limiting case of this prior specification we consider $\pi\left(\boldsymbol{\theta}\right) \propto 1$. Analogous to Eq. (27) we have the following exact posterior distribution for $\boldsymbol{\alpha}$ given $\boldsymbol{\theta}$ and $\boldsymbol{\Delta}$

$$\pi\left(\boldsymbol{\alpha} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\Delta}\right) \propto \exp\left\{-\frac{n}{2}\ \mathrm{tr}\ \left[\mathbf{A} + \mathbf{W}\exp\left\{-\mathbf{A}\right\}\right] - \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{G}^*\boldsymbol{\alpha}\right\} \tag{32}$$

where $\underset{(p\times p)}{\mathbf{W}} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{y}_i - \boldsymbol{\theta}\right)\left(\mathbf{y}_i - \boldsymbol{\theta}\right)^{\mathrm{T}} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{y}_i - \overline{\mathbf{y}}\right)\left(\mathbf{y}_i - \overline{\mathbf{y}}\right)^{\mathrm{T}} + \left(\boldsymbol{\theta} - \overline{\mathbf{y}}\right)\left(\boldsymbol{\theta} - \overline{\mathbf{y}}\right)^{\mathrm{T}}$. Then under a vague prior specification for $\boldsymbol{\theta}$ the exact posterior distribution for $\boldsymbol{\theta}$ conditional on $\boldsymbol{\Sigma}$ will be proportional to the exact likelihood function.

$$\pi\left(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\Sigma}\right) \propto \exp\left\{-\frac{n}{2}\left(\boldsymbol{\theta} - \overline{\mathbf{y}}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta} - \overline{\mathbf{y}}\right)\right\}$$

Therefore, we see that the exact posterior distribution for $\boldsymbol{\theta}$ conditional on $\boldsymbol{\Sigma}$ is given by the following.

$$\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\Sigma} \sim \mathrm{N}_p\left(\overline{\mathbf{y}}, n^{-1}\boldsymbol{\Sigma}\right) \tag{33}$$

The posterior distributions for σ_1^2 and σ_2^2, the only two unknown hyperparameters involved in $\boldsymbol{\Delta}$, will be identical to the previously derived posterior distributions as stated in Eqs. (30) and (31). Thus, the MCMC will be the same as the algorithm that was previously outlined above except here will include one additional step which is to simulate $\boldsymbol{\theta}$ according to Eq. (33).

Note that based upon Eq. (33) the posterior mean of $\boldsymbol{\theta}$ is in fact equal to the sample mean vector $\overline{\mathbf{y}}$. In particular, we computed the following posterior means for the various exams 0.5138, 0.4958, 0.7840, 0.6773, 0.4551, 0.4461, 0.6017 and 0.4103, where all values are represented as proportions. Posterior moments for the other parameters of interest can be calculated based upon the results from the MCMC. Please refer to Table 5 for the posterior mean of $\mathbf{A}$ calculated via MCMC sampling under the hierarchical prior specification with an unknown mean vector.

7 Conclusion

We have estimated the covariance structure of a multivariate normal distribution from a Bayesian perspective. In contrast to the usual inverse Wishart conjugate prior specification we have made use of a highly flexible and tractable multivariate normal

Table 5 Posterior mean of **A** with an unknown mean under hierarchical prior via MCMC/Metropolis–Hastings algorithm

	1	2	3	4	5	6	7	8
1	−5.235	0.540	0.344	0.425	0.366	0.300	0.125	0.368
2	0.540	−4.635	0.289	0.427	0.200	0.248	0.333	0.153
3	0.344	0.289	−5.626	0.344	0.120	0.069	0.149	0.432
4	0.425	0.427	0.344	−4.066	0.396	0.237	0.308	0.391
5	0.366	0.200	0.120	0.396	−3.770	0.334	0.230	0.246
6	0.300	0.248	0.069	0.237	0.334	−3.984	0.384	0.332
7	0.125	0.333	0.149	0.308	0.230	0.384	−3.652	0.311
8	0.368	0.153	0.432	0.391	0.246	0.332	0.311	−4.130

prior specification for the unique elements of the matrix logarithm of the covariance matrix. In this way we have been able to model varying degrees of confidence in the prior location hyperparameters as well model potential interdependencies amongst the covariance structure. An approximation of the likelihood function for the covariance matrix based upon a second-order expansion of a linear Volterra integral equation was made. In this way the approximate likelihood function can be expressed in a multivariate normal form. Thus, we achieved approximate conjugacy. Under this Bayesian formulation we computed posterior moments via Importance sampling, Laplacian approximation and finally MCMC sampling. With respect to the Laplacian approximation we demonstrated how a generalized finite sample likelihood function approximation could be utilized to facilitate the required integration. In addition, a Metropolis–Hastings algorithm was employed in the MCMC sampling procedures to account for the approximation. We applied the estimation procedures to standardized test scores from the Project Talent educational data set.

Acknowledgments The authors are grateful to Tom Leonard for his valuable advice. We are also grateful to an Associate Editor and two referees for their helpful comments and suggestions that improved the paper.

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.

Chen, C.-F. (1979). Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological), 41*(2), 235–248.

Chiu, T. Y. M., Leonard, T., Tsui, K.-W. (1996). The matrix logarithmic covariance model. *Journal of the American Statistical Association, 91*(433), 198–210.

Cooley, W. W., Lohnes, P. R. (1971). *Multivariate data analysis*. New York: Wiley.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika, 68*(1), 265–274.

Dey, D. K., Srinivasan, C. (1985). Estimation of a covariance matrix under stein's loss. *The Annals of Statistics, 13*(4), 1581–1591.

Dickey, J., Lindley, D., Press, S. (1985). Bayesian estimation of the dispersion matrix of a multivariate normal distribution. *Communications in Statistics Theory and Methods, 14*(5), 1019–1034.

Evans, I. G. (1965). Bayesian estimation of parameters of a multivariate normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological), 27*(2), 279–283.
Flanagan, J. C., Tiedeman, D. V. (1979). *Project Talent public use file [computer file]*. Technical report, American Institutes for Research [producer], Palo Alto, California.
Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2005). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall.
Hsu, C. W. (2001). *Bayesian estimation of a covariance matrix and its application to mixed effects models*. PhD thesis, University of California Santa Barbara, Santa Barbara, California.
Kass, R. E., Tierney, L., Kadane, J. B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika, 76*(4), 663–674.
Laplace, P. S. (1986). Memoir on the probability of the causes of events. *Statistical Science, 1*(3), 364–378.
Leonard, T. (1982). A simple predicitive density function: Comment. *Journal of the American Statistical Association, 77*(379), 657–658.
Leonard, T., Hsu, J. S. J. (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics, 20*(4), 1669–1696.
Leonard, T., Hsu, J. S. J. (1999). *Bayesian methods*. New York: Cambridge University Press.
Leonard, T., Hsu, J. S. J., Tsui, K.-W. (1989). Bayesian marginal inference. *Journal of the American Statistical Association, 84*(408), 1051–1058.
Leonard, T., Hsu, J. S. J., Tsui, K.-W., Murray, J. F. (1994). Bayesian and likelihood inference from equally weighted mixtures. *Annals of the Institute of Statistical Mathematics, 46*(2), 203–220.
Ni, S., Sun, D. (2005). Bayesian estimates for vector autoregressive models. *Journal of Business and Economic Statistics, 23*(1), 105–117.
Ogata, Y. (1989). A Monte Carlo method for high dimensional integration. *Numerische Mathematik, 55*(2), 137–157.
Robert, C. P., Casella, G. (2004). *Monte Carlo statistical methods*. New York: Springer.
Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo method*. New York: Wiley.
Stigler, S. M. (1986). Laplace's 1774 memoir on inverse probability. *Statistical Science, 1*(3), 359–363.
Tierney, L., Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association, 81*(393), 82–86.
Tierney, L., Kass, R. E., Kadane, J. B. (1989). Fully exponential laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association, 84*(407), 710–716.
Yang, R., Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics, 22*(3), 1195–1211.