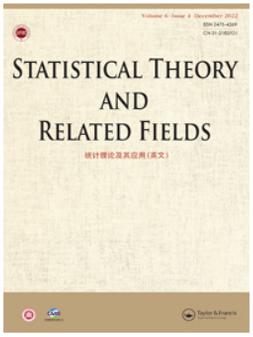

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Posterior propriety of an objective prior for generalized hierarchical normal linear models

Cong Lin, Dongchu Sun & Chengyuan Song

To cite this article: Cong Lin, Dongchu Sun & Chengyuan Song (2022) Posterior propriety of an objective prior for generalized hierarchical normal linear models, Statistical Theory and Related Fields, 6:4, 309-326, DOI: 10.1080/24754269.2021.1978206

To link to this article: https://doi.org/10.1080/24754269.2021.1978206

STATISTICAL THEORY AND RELATED FIELDS
2022, VOL. 6, NO. 4, 309–326
https://doi.org/10.1080/24754269.2021.1978206

Taylor & Francis
Taylor & Francis Group

Posterior propriety of an objective prior for generalized hierarchical normal linear models

Cong Lin[a], Dongchu Sun[a,b] and Chengyuan Song[c]

[a]School of Statistics, East China Normal University, Shanghai, People's Republic of China; [b]Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE, USA; [c]Boehringer Ingelheim (China), Shanghai, People's Republic of China

ABSTRACT

Bayesian Hierarchical models has been widely used in modern statistical application. To deal with the data having complex structures, we propose a generalized hierarchical normal linear (GHNL) model which accommodates arbitrarily many levels, usual design matrices and 'vanilla' covariance matrices. Objective hyperpriors can be employed for the GHNL model to express ignorance or match frequentist properties, yet the common objective Bayesian approaches are infeasible or fraught with danger in hierarchical modelling. To tackle this issue, [Berger, J., Sun, D., & Song, C. (2020b). An objective prior for hyperparameters in normal hierarchical models. *Journal of Multivariate Analysis*, 178, 104606. https://doi.org/10.1016/j.jmva.2020.104606] proposed a particular objective prior and investigated its properties comprehensively. Posterior propriety is important for the choice of priors to guarantee the convergence of MCMC samplers. James Berger conjectured that the resulting posterior is proper for a hierarchical normal model with arbitrarily many levels, a rigorous proof of which was not given, however. In this paper, we complete this story and provide an user-friendly guidance. One main contribution of this paper is to propose a new technique for deriving an elaborate upper bound on the integrated likelihood, but also one unified approach to checking the posterior propriety for linear models. An efficient Gibbs sampling method is also introduced and outperforms other sampling approaches considerably.

ARTICLE HISTORY
Received 23 January 2021
Revised 31 August 2021
Accepted 31 August 2021

KEYWORDS
Hierarchical linear model; linear mixed-effect model; objective Bayesian analysis; posterior propriety; Gibbs sampling

1. Introduction

Bayesian hierarchical models (or multilevel models) have been extensively used in the modern application, including education (Raudenbush & Bryk, 1986), psychology (Lindenberger & Pötter, 1998), clinical trials (Xia et al., 2011), economics (Shimotsu, 2010) and many other applied statistical fields. The fundamental idea of hierarchical modelling is to think of the lowest-level units (smallest and most numerous) as organized into a hierarchy of successively higher-level units. For example, students are in classes, classes are in schools, schools are in districts, and districts are in states. Accordingly, hierarchical models are naturally applicable to the survey, observational or experimental data involved with complicated nesting. However, the most commonly used and fully discussed hierarchical models are merely of two levels. Goldstein (2011) and Berger et al. (2020b) have ever defined 3-level hierarchical model and implemented statistical analysis on that. The hierarchical model with more levels was usually avoided by the researchers for the reason of analytical difficulty and intractable computation. To the best of authors' knowledge, a general hierarchical linear model with arbitrary levels seems to have never been defined or studied. In this paper, we will introduce the definition of a *generalized hierarchical normal linear* (GHNL) model and carry out an in-depth theoretical investigation of Bayesian inference for the GHNL models.

In order to implement fully Bayesian analysis, priors are supposed to be specified on the hyperparameters (parameters at higher levels of the hierarchical model). Improper (objective) priors are often used to express ignorance or to match frequentist properties (see the review article, Consonni et al., 2018). When using improper priors, an important issue whether the resulting posterior distributions are proper arises. As Hobert and Casella (1996) stated, without proper precaution, misuse of improper priors, sometimes unknowingly, will result in practical difficulties, such as the non-convergence of the Gibbs sampler. The enormous practical importance of posterior propriety motivates us to explore it in the framework of GHNL modelling afterwards. There is also a vast modern literature investigating the posterior propriety of improper priors applied to a large variety of models, such as, Sun et al. (2001), Speckman and Sun (2003), Berger et al. (2005) and Michalak and Morris (2016).

A great deal of efforts have been devoted to the development of objective hyperpriors in hierarchical

CONTACT Cong Lin conglin_stat@163.com

modelling, such as, Daniels and Kass (1999), Everson and Morris (2000), Gelman (2006), Gustafson et al. (2006), Berger et al. (2005) and Berger et al. (2020b). Formal objective Bayesian approaches, like the Jeffreys-rule prior or reference prior are only feasible for the simple hierarchical settings. For instance, the exact Jeffreys-rule prior for covariance matrices at higher level depends on the parameters from the lower level of the model, leading to plenty of difficulties in formulation and computation. Therefore, a common way is to use less formal approaches, such as applying formal objective priors from non-hierarchical models to hierarchical modelling. Unfortunately, the non-hierarchical Jefferys-rule prior and reference prior typically yield improper posteriors in the hierarchical settings (cf. Berger et al., 2005). Those who can recognize this problem often use constant priors instead for higher level variance components. However, the constant prior is so diffuse that it requires twice as many observations as logically needed to achieve posterior propriety (cf. Berger et al., 2005 and Berger et al., 2020b). In other words, the extra observations required are wasted on correcting the over-diffuse tail of the constant prior. The most powerful tool known for detecting over-diffuse hyperpriors is by looking at the frequentist notion of admissibility of resulting estimators (see Berger et al., 2005 for discussions and references). Sensible choices of objective hyperpriors are on the boundary of admissibility, being as diffuse as possible without leading to inadmissible estimators.

Berger et al. (2005) studied the propriety and admissibility of a number of hyperprios, but no overall conclusion was reached as to a specific prior to recommend. The reasons are as follows: (a) the admissibility of the leading candidate prior was unable to get proved; (b) the proposed computation methods were only efficient for relatively low-dimensional covariance matrices and remained quite challenging for the candidate priors; (c) the hierarchical model discussed was merely of two levels, and the results are not adaptive to a general hierarchical model with many levels. To address this issue, Berger et al. (2020b) recommended a particular objective prior for use in all normal hierarchical models. Consider the following canonical form of 2-level hierarchical normal model. Suppose that, independently, $\boldsymbol{y}_i \sim N_k(\boldsymbol{\theta}_i, \boldsymbol{I}_k)$ and $\boldsymbol{\theta}_i \sim N_k(\boldsymbol{\beta}, \boldsymbol{V})$ for $i = 1, \ldots, m$, where $N_k(\cdot, \cdot)$ denotes the k-dimensional normal distribution, $\boldsymbol{y}_i$ are $k \times 1$ observation vectors, $\boldsymbol{\theta}_i$ are the $k \times 1$ unobserved mean vectors, $\boldsymbol{\beta}$ is a $p \times 1$ 'hypermean' vector, and $\boldsymbol{V} \in \mathbb{R}^{k\times k}$ is an unknown 'hypercovariance' matrix. Berger et al. (2020b) proposed a particular combination of independent priors on hyperparameters $\boldsymbol{\beta}$ and $\boldsymbol{V}$ as

$$\pi(\boldsymbol{\beta}) \propto \frac{1}{(1 + \|\boldsymbol{\beta}\|^2)^{(k-1)/2}},$$
$$\pi(\boldsymbol{V}) \propto \frac{1}{|\boldsymbol{V}|^{1-1/(2k)} \prod_{1\le s<t\le k} (\nu_s - \nu_t)}, \qquad (1)$$

where $\nu_1 > \nu_2 > \cdots > \nu_k > 0$ are the ordered eigenvalues of $\boldsymbol{V}$. The recommendation (1) for hyperpriors was justified by Berger et al. (2020b) from the aspects of admissibility, ease of computation and performance. Most importantly, prior (1) is adapted to being used at any level in hierarchical modelling, which is not true for other proposed objective priors as previously mentioned.

Since it is hazardous to skip demonstrating propriety at the risk of making inference from improper posterior, Berger et al. (2020b) has shown the posterior propriety of a 3-level hierarchical model using prior (1), while assuming square design matrices for a technical reason. Berger et al. (2020b) also conjectured that the posterior is proper with the recommended prior being utilized at all levels of a hierarchical normal model with arbitrarily many levels, a rigorous proof of which was unable to be provided, however. In this paper, we complete this story and prove the posterior propriety for the GHNL models in general situations. Besides, as pointed out in Michalak and Morris (2016), researchers have been finding it daunting and time-consuming to inspect posterior propriety when using improper priors, except in the simplest models. For this reason, we supply a user-friendly guidance for checking posterior propriety to practitioners in different practical situations.

In Section 2, we give an explicit definition of the GHNL model which accommodates arbitrarily many levels and usual design matrices. It is important to note that we are considering the 'vanilla' covariance matrix problem herein. We are not assuming any special structures or sparsity for hypercovariance matrices. The association between the GHNL model and a linear mixed-effect model is also discussed. In Section 3, we demonstrate that recommended prior yields a proper posterior in the framework of GHNL modelling. In addition, we exhibit a guidance for checking posterior propriety. An efficient MCMC algorithm for sampling from the posterior is introduced in Section 4. Section 5 provides some concluding remarks and further generalizations.

2. Generalized hierarchical normal linear model

In this section, we will introduce the definition of a GHNL model with $(r + 1)$ levels, where $r \ge 1$. The association between the GHNL model and a linear mixed-effect model will be demonstrated as well, which brings an insight into the GHNL model. At last, the recommended prior on the hyperparameters of the GHNL model will be presented and discussed. Firstly, we introduce some notations to be used in the main body of this paper.

Notations Let $[k] = \{1, 2, \ldots, k\}$ for a positive integer k; $1_{\{\cdot\}}$ stands for the indicator function; $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

represents the k-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$; $\mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a k-dimensional normal random variable with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$; for a symmetric matrix $\boldsymbol{A}$, $\boldsymbol{A} > (<)0$ means that $\boldsymbol{A}$ is a positive (negative) definite matrix, and $\boldsymbol{A} \geq (\leq)0$ denotes that $\boldsymbol{A}$ is a non-negative (non-positive) definite matrix.

2.1. Model structure

Berger et al. (2020b) proposed a 3-level hierarchical model with the form:

$$\begin{cases} \text{Level 1}: \ \boldsymbol{y}_i = \boldsymbol{\theta}_i + \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{I}_k), & i \in [m]; \\ \text{Level 2}: \boldsymbol{\theta}_i = \boldsymbol{Z}_i\boldsymbol{\beta} + \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{V}), & \boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_s^\top); \\ \text{Level 3}: \ \boldsymbol{\beta}_j = \boldsymbol{\eta} + \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{W}), & j \in [s], \end{cases} \tag{2}$$

where $\boldsymbol{y}_i$ are $k \times 1$ observation vectors, $\boldsymbol{\theta}_i$ are the $k \times 1$ unobserved mean vectors, $\boldsymbol{\eta}$ is a $p \times 1$ 'hypermean' vector, $\boldsymbol{V} \in \mathbb{R}^{k\times k}$ and $\boldsymbol{W} \in \mathbb{R}^{p\times p}$ are unknown 'hyper-covariance' matrices, and $\boldsymbol{Z}_i$ are $k \times sp$ known matrices. At last, all the normal random variables in model (2) are mutually independent. Based on the 3-level hierarchical normal model, a more general hierarchical model with $(r+1)$ levels $(r \geq 1)$ can be constructed as

$$\begin{cases} \text{Level 1}: & \boldsymbol{y}_{i_0} = \boldsymbol{Z}_{0i_0}\boldsymbol{\theta}_1 + \mathcal{N}_{k_0}(\boldsymbol{0}, \boldsymbol{I}_{k_0}), & i_0 \in [m_0], \ \boldsymbol{\theta}_1^\top = (\boldsymbol{\theta}_{11}^\top, \ldots, \boldsymbol{\theta}_{1m_1}^\top); \\ \text{Level 2}: & \boldsymbol{\theta}_{1i_1} = \boldsymbol{Z}_{1i_1}\boldsymbol{\theta}_2 + \mathcal{N}_{k_1}(\boldsymbol{0}, \boldsymbol{V}_1), & i_1 \in [m_1], \ \boldsymbol{\theta}_2^\top = (\boldsymbol{\theta}_{21}^\top, \ldots, \boldsymbol{\theta}_{2m_2}^\top); \\ \vdots & \vdots & \vdots \\ \text{Level } r: & \boldsymbol{\theta}_{r-1,i_{r-1}} = \boldsymbol{Z}_{r-1,i_{r-1}}\boldsymbol{\theta}_r + \mathcal{N}_{k_{r-1}}(\boldsymbol{0}, \boldsymbol{V}_{r-1}), & i_{r-1} \in [m_{r-1}], \ \boldsymbol{\theta}_r^\top = (\boldsymbol{\theta}_{r1}^\top, \ldots, \boldsymbol{\theta}_{rm_r}^\top); \\ \text{Level } r+1: & \boldsymbol{\theta}_{ri_r} = \boldsymbol{Z}_{ri_r}\boldsymbol{\eta} + \mathcal{N}_{k_r}(\boldsymbol{0}, \boldsymbol{V}_r), & i_r \in [m_r]. \end{cases} \tag{3}$$

Firstly, all the normal random variables noted in the above model are mutually independent. Within model (3), the output of level $(j+1)$ consists of m_j units whose values are $k_j \times 1$ vectors for $j = 0, 1, \ldots, r$. By stacking the output units of level $(j+1)$ on top of one another, we can obtain the *outcome vector* of level $(j+1)$ as $\boldsymbol{\theta}_j$ for $j \in [r]$ and $\boldsymbol{y} = \left(\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_{m_0}^\top\right)^\top$ for level 1. Then $\boldsymbol{\theta}_j$'s are $(m_jk_j) \times 1$ vectors and $\boldsymbol{y}$ is a $(m_0k_0) \times 1$ vector. In fact, only the outcome of the lowest level can be observed, and the outcomes of higher levels are inaccessible and latent variables. Hence, the outcome variables of interest are always situated at the lowest level of the hierarchy. Different units in the same level share in common *input effects* (intercept can be included) which are exactly the outcome vectors from the upper level, except that the input effect of level $(r+1)$ is $\boldsymbol{\eta}$ which is a $d \times 1$ vector of fixed effects. In addition, the units from the same level have the same variance component. The variance component within level $(j+1)$ is denoted by $\boldsymbol{V}_j \in \mathbb{R}^{k_j\times k_j}$ for $j \in [r]$ and accounts for the magnitude of random variation within the corresponding level. The covariance matrices $\boldsymbol{V}_j$ are unobserved for $j \in [r]$. The matrices $\boldsymbol{Z}_{ji_j}$ are $k_j \times (m_{j+1}k_{j+1})$ matrices and denote the matrices of observed covariates for unit i_j in level j, where $j = 0, 1, \ldots, r$ and $i_j \in [m_j]$. It is natural to assume that there exist at least two units in each level and the dimensions of all units and $\boldsymbol{\eta}$ are no less than 1, mathematically, $m_j \geq 2$ and $k_j \geq 1$ for $j = 0, 1, \ldots, r$, and $d \geq 1$. Table 1 summarizes several important notations that will mainly affect the results for the posterior propriety in Section 3.

Table 1. Summary of certain important notations within model (3) and $j = 0, 1, \ldots, r$.

Notation	Meaning
$r+1$	Total number of levels
m_j	Number of the output units in level $(j+1)$
k_j	Dimension of each output unit in level $(j+1)$
d	Dimension of the fixed effect $\boldsymbol{\eta}$

The extensions from Berger et al. (2020b)'s model (2) to model (3) are two-fold, and model (3) accommodates arbitrarily many levels and usual covariate matrices. Further define $\boldsymbol{Z}_j = \left\{\boldsymbol{Z}_{j1}^\top, \ldots, \boldsymbol{Z}_{jm_j}^\top\right\}^\top$ for $j = 0, 1, \ldots, r$. Then $\boldsymbol{Z}_j$ are $(m_jk_j) \times (m_{j+1}k_{j+1})$ matrices for $j = 0, 1, \ldots, (r-1)$, $\boldsymbol{Z}_r$ is an $(m_rk_r) \times d$ matrix, and an alternative representation of the $(r+1)$-level hierarchical model (3) is thereby given by

$$\begin{cases} \text{Level 1}: & (\boldsymbol{y}|\boldsymbol{\theta}_1) & \sim & N_{m_0k_0}(\boldsymbol{Z}_0\boldsymbol{\theta}_1, \boldsymbol{I}_{m_0k_0}); \\ \text{Level 2}: & (\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{V}_1) & \sim & N_{m_1k_1}(\boldsymbol{Z}_1\boldsymbol{\theta}_2, \boldsymbol{I}_{m_1} \otimes \boldsymbol{V}_1); \\ \vdots & \vdots & \vdots & \\ \text{Level } r: & (\boldsymbol{\theta}_{r-1}|\boldsymbol{\theta}_r, \boldsymbol{V}_{r-1}) & \sim & N_{m_{r-1}k_{r-1}}(\boldsymbol{Z}_{r-1}\boldsymbol{\theta}_r, \boldsymbol{I}_{m_{r-1}} \otimes \boldsymbol{V}_{r-1}); \\ \text{Level } r+1: & (\boldsymbol{\theta}_r|\boldsymbol{\eta}, \boldsymbol{V}_r) & \sim & N_{m_rk_r}(\boldsymbol{Z}_r\boldsymbol{\eta}, \boldsymbol{I}_{m_r} \otimes \boldsymbol{V}_r). \end{cases} \tag{4}$$

Remark 2.1: If we assume that the covariance matrix for the units from level 1 in model (3) is a positive definite matrix $\boldsymbol{\Sigma}_0$ instead of the identity matrix, when $\boldsymbol{\Sigma}_0$ is known, the two assumptions are actually equivalent by taking reparameterization that $\boldsymbol{y}_{i_0}^* = \boldsymbol{\Sigma}_0^{-\frac{1}{2}}\boldsymbol{y}_{i_0}$ and $\boldsymbol{Z}_{i_0}^* = \boldsymbol{\Sigma}_0^{-\frac{1}{2}}\boldsymbol{Z}_{i_0}$. Furthermore, for a technical reason, $\boldsymbol{\Sigma}_0$ must be assumed as known throughout this paper, and this reason will be explained in Section 5.

2.2. Connection with the linear mixed-effect model (LMM)

The two-level hierarchical normal models are often referred to as LMMs in many places. As for the GHNL model (4), let $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r\}$ denote the set of unobserved outcome vectors and $\mathcal{V} = \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_r\}$ represent the set of unknown covariance matrices. If we take $\boldsymbol{\theta}_j$'s as intermediate variables, then marginalizing out over $\boldsymbol{\Theta}$ yields

$$(\boldsymbol{y}|\boldsymbol{\eta}, \mathcal{V}) \sim N_{m_0k_0}(\boldsymbol{X}_r\boldsymbol{\eta}, \boldsymbol{\Delta}), \tag{5}$$

where

$$\boldsymbol{\Delta} = \boldsymbol{I}_{m_0k_0} + \sum_{t=1}^{r} \boldsymbol{X}_{t-1}(\boldsymbol{I}_{m_t} \otimes \boldsymbol{V}_t)\boldsymbol{X}_{t-1}^\top, \quad \text{and}$$

$$\boldsymbol{X}_j = \prod_{s=0}^{j} \boldsymbol{Z}_s,\ j = 0, 1, \ldots, r. \tag{6}$$

$\boldsymbol{\Delta}$ is a $(m_0k_0) \times (m_0k_0)$ matrix and $\boldsymbol{X}_j$ are $(m_0k_0) \times (m_{j+1}k_{j+1})$ matrices for $j = 0, 1, \ldots, r$. Suppose that $\boldsymbol{Z}_j$ are of full column ranks. Then by Sylvester's rank inequality $\boldsymbol{X}_j$ are also of full column ranks, $j = 0, 1, \ldots, r$. In the rest of the paper, $\boldsymbol{Z}_j$ are assumed to be of full column ranks for $j = 0, 1, \ldots, r$.

If we consider a particular LMM as

$$\boldsymbol{y} = \boldsymbol{X}_r\boldsymbol{\eta} + \boldsymbol{X}_0\boldsymbol{\theta}_1^* + \cdots + \boldsymbol{X}_{r-1}\boldsymbol{\theta}_r^* + \boldsymbol{\epsilon}, \tag{7}$$

where $\boldsymbol{\eta}$ is the fixed effect, $\boldsymbol{\theta}_j$'s are random effects and independently distributed as $N_{m_jk_j}(\boldsymbol{0}, \boldsymbol{I}_{m_j} \otimes \boldsymbol{V}_j)$ for $j \in [r]$, and $\boldsymbol{\epsilon}$ denotes the vector of random errors and is distributed as $N_{m_0k_0}(\boldsymbol{0}, \boldsymbol{I}_{m_0k_0})$. By integrating out the random effects, the marginal distribution of $\boldsymbol{y}$ conditioning on $(\boldsymbol{\eta}, \mathcal{V})$ is identical to the distribution (5). In one word, the GHNL is equivalent to an LMM in the sense of the marginal distribution of observations after integrating out intermediate outcome vectors or random effects. The equivalence between the GHNL models and the LMMs can be illustrated by an example of mixed-effect ANOVA model as well.

Example 2.1 (Mixed-effect ANOVA model): Suppose we can observe the scores of p courses for student (ijk) as $\boldsymbol{y}_{ijk}$ for $i = 1, \ldots, s_1$, $j = 1, \ldots, s_2$ and $k = 1, \ldots, s_3$. The observed data are within a hierarchy of three levels: student (ijk) is nested within class (ij), and class (ij) is nested within school i. Thus, we have total s_1 schools, each school has s_2 classes and each class has s_3 students. Consider a mixed-effect ANOVA model as

$$\boldsymbol{y}_{ijk} = \boldsymbol{\eta} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_{ij} + \boldsymbol{\epsilon}_{ijk}, \tag{8}$$

where $\boldsymbol{y}_{ijk}$, $\boldsymbol{\eta}$, α_i, $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{\epsilon}_{ijk}$ are all $p \times 1$ vectors for $i = 1, \ldots, s_1$, $j = 1, \ldots, s_2$ and $k = 1, \ldots, s_3$, $\boldsymbol{\eta}$ denotes the overall mean and is fixed effect, $\boldsymbol{\alpha}_i \sim N_p(\boldsymbol{0}, \boldsymbol{V}_\alpha)$ is the effect of school, $\boldsymbol{\beta}_{ij}$ is distributed as $N_p(\boldsymbol{0}, \boldsymbol{V}_\beta)$ and represents the effect of class, the student-level independent random error is denoted by $\boldsymbol{\epsilon}_{ijk}$ and has distribution $N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$, and $\boldsymbol{\Sigma}_0$ is a known matrix. At the same time, $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{\epsilon}_{ijk}$ are independently distributed. Consequently, $\boldsymbol{V}_\alpha$, $\boldsymbol{V}_\beta$ and $\boldsymbol{\Sigma}_0$ are the variance components describing the school-level, class-level and student-level variations, respectively. Due to the hierarchical structure of the observations, we can naturally build a hierarchical model as

$$\boldsymbol{y}_{ijk} \sim N_p(\boldsymbol{\beta}_{ij}^*, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\beta}_{ij}^* \sim N_p(\boldsymbol{\alpha}_i^*, \boldsymbol{V}_\beta) \quad \text{and}$$
$$\boldsymbol{\alpha}_i^* \sim N_p(\boldsymbol{\eta}, \boldsymbol{V}_\alpha) \tag{9}$$

independently, for $i = 1, \ldots, s_1$, $j = 1, \ldots, s_2$ and $k = 1, \ldots, s_3$. Denote that

$$\boldsymbol{Y}_i = \begin{pmatrix} \boldsymbol{y}_{i11} & \cdots & \boldsymbol{y}_{is_21} \\ \vdots & \ddots & \vdots \\ \boldsymbol{y}_{i1s_3} & \cdots & \boldsymbol{y}_{is_2s_3} \end{pmatrix},$$

$$\boldsymbol{E}_i = \begin{pmatrix} \boldsymbol{\epsilon}_{i11} & \cdots & \boldsymbol{\epsilon}_{is_21} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\epsilon}_{i1s_3} & \cdots & \boldsymbol{\epsilon}_{is_2s_3} \end{pmatrix},$$

$$\boldsymbol{\beta}_i = \begin{pmatrix} \boldsymbol{\beta}_{i1} \\ \vdots \\ \boldsymbol{\beta}_{is_2} \end{pmatrix} \quad \text{and}$$

$$\boldsymbol{\beta}_i^* = \begin{pmatrix} \boldsymbol{\beta}_{i1}^* \\ \vdots \\ \boldsymbol{\beta}_{is_2}^* \end{pmatrix},$$

where $\boldsymbol{Y}_i$ and $\boldsymbol{E}_i$ are both $(s_3p) \times s_2$ matrices, and $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_i^*$ are both $(s_2p) \times 1$ vectors. Let $\boldsymbol{y}_i = \text{vec}(\boldsymbol{Y}_i)$ and $\boldsymbol{\epsilon}_i = \text{vec}(\boldsymbol{E}_i)$, where $\text{vec}(\boldsymbol{A})$ denotes the column vector obtained by stacking the columns of the matrix $\boldsymbol{A}$ on top of one another. Define that

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_{s_1} \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_{s_1} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_{s_1} \end{pmatrix},$$

$$\boldsymbol{\beta}^* = \begin{pmatrix} \boldsymbol{\beta}_1^* \\ \vdots \\ \boldsymbol{\beta}_{s_1}^* \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_{s_1} \end{pmatrix}, \quad \boldsymbol{\alpha}^* = \begin{pmatrix} \boldsymbol{\alpha}_1^* \\ \vdots \\ \boldsymbol{\alpha}_{s_1}^* \end{pmatrix}.$$

Thus, $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ are $(m_0p) \times 1$ vectors, and $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ are $(m_1p) \times 1$ vectors, and $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$ are $(m_2p) \times 1$ vectors, where $m_0 = s_1s_2s_3$, $m_1 = s_1s_2$ and $m_2 = s_1$. Then the hierarchical normal model (9) can be expressed as a GHNL model with the form

$$\begin{cases} \text{Level 1}: & (\boldsymbol{y}|\boldsymbol{\beta}^*, \boldsymbol{\Sigma}_0) & \sim & N_{m_0p}(\boldsymbol{Z}_0\boldsymbol{\beta}^*, \boldsymbol{I}_{m_0} \otimes \boldsymbol{\Sigma}_0); \\ \text{Level 2}: & (\boldsymbol{\beta}^*|\boldsymbol{\alpha}^*, \boldsymbol{V}_\beta) & \sim & N_{m_1p}(\boldsymbol{Z}_1\boldsymbol{\alpha}^*, \boldsymbol{I}_{m_1} \otimes \boldsymbol{V}_\beta); \\ \text{Level 3}: & (\boldsymbol{\alpha}^*|\boldsymbol{\eta}, \boldsymbol{V}_\alpha) & \sim & N_{m_2p}(\boldsymbol{Z}_2\boldsymbol{\eta}, \boldsymbol{I}_{m_2} \otimes \boldsymbol{V}_\alpha), \end{cases} \tag{10}$$

where

$$\boldsymbol{Z}_0 = \mathrm{diag}\{\underbrace{\mathbf{1}_{s_3} \otimes \boldsymbol{I}_p, \ldots, \mathbf{1}_{s_3} \otimes \boldsymbol{I}_p}_{s_1 s_2}\},$$

$$\boldsymbol{Z}_1 = \mathrm{diag}\{\underbrace{\mathbf{1}_{s_2} \otimes \boldsymbol{I}_p, \ldots, \mathbf{1}_{s_2} \otimes \boldsymbol{I}_p}_{s_1}\},\ \boldsymbol{Z}_2 = \mathbf{1}_{s_1} \otimes \boldsymbol{I}_p,$$

where $\mathbf{1}_q$ denotes the $q \times 1$ vector with all elements being one, and $\boldsymbol{Z}_0, \boldsymbol{Z}_1, \boldsymbol{Z}_2$ are $(m_0 p) \times (m_1 p)$, $(m_1 p) \times (m_2 p)$, $(m_2 p) \times p$ matrices, respectively. Denote that

$$\boldsymbol{X}_0 \triangleq \boldsymbol{Z}_0,$$

$$\boldsymbol{X}_1 \triangleq \mathrm{diag}\{\underbrace{\mathbf{1}_{s_2 s_3} \otimes \boldsymbol{I}_p, \ldots, \mathbf{1}_{s_2 s_3} \otimes \boldsymbol{I}_p}_{s_1}\} = \boldsymbol{Z}_0 \boldsymbol{Z}_1,$$

$$\boldsymbol{X}_2 \triangleq \mathbf{1}_{s_1 s_2 s_3} \otimes \boldsymbol{I}_p = \boldsymbol{Z}_0 \boldsymbol{Z}_1 \boldsymbol{Z}_2,$$

and $\boldsymbol{X}_0$, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ are $(m_0 p) \times (m_1 p)$, $(m_0 p) \times (m_2 p)$, $(m_0 p) \times p$ matrices, respectively. Thus, model (8) can be summarized as

$$\boldsymbol{y} = \boldsymbol{X}_2 \boldsymbol{\eta} + \boldsymbol{X}_1 \boldsymbol{\alpha} + \boldsymbol{X}_0 \boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{11}$$

where $\boldsymbol{\alpha} \sim N_{m_2 p}(\mathbf{0}, \boldsymbol{I}_{m_2} \otimes \boldsymbol{V}_\alpha)$, $\boldsymbol{\beta} \sim N_{m_1 p}(\mathbf{0}, \boldsymbol{I}_{m_1} \otimes \boldsymbol{V}_\beta)$ and $\boldsymbol{\epsilon} \sim N_{m_0 p}(\mathbf{0}, \boldsymbol{I}_{m_0} \otimes \boldsymbol{\Sigma}_0)$, independently. By integrating out $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$, the marginal distributions of $\boldsymbol{y}$ for model (11) and model (10) are identical and of the form:

$$(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{V}_\alpha, \boldsymbol{V}_\beta, \boldsymbol{\Sigma}_0) \sim N_{m_0 p}(\boldsymbol{X}_2 \boldsymbol{\eta}, \boldsymbol{\Omega}),$$

$$\boldsymbol{\Omega} = \boldsymbol{I}_{m_0} \otimes \boldsymbol{\Sigma}_0 + \boldsymbol{X}_0 (\boldsymbol{I}_{m_1} \otimes \boldsymbol{V}_\beta) \boldsymbol{X}_0^\top + \boldsymbol{X}_1 (\boldsymbol{I}_{m_2} \otimes \boldsymbol{V}_\alpha) \boldsymbol{X}_1^\top.$$

Example 2.1 provides a simple example to illustrate how the hierarchical model and the mixed-effect model can be constructed based on the nested data, and the equivalence between two models is also presented. In Appendix 1, we define a special LMM which is a special case of model (7) with $m_j = 1$ for all $j \in [r]$, and the theoretical investigation of this special LMM is distinct from that of the GHNLM. This special LMM could be common in application, and we just want to provide some theoretical results for those who have interests to refer to. We will still focus on the GHNL model in the following sections.

2.3. Priors on the hyperparameters

In order to implement fully Bayesian analysis, we should specify hyperpriors on the parameters $(\boldsymbol{\eta}, \mathcal{V})$. It follows the recommendation from Berger et al. (2020b) that we can assume priors on $(\boldsymbol{\eta}, \mathcal{V})$ as:

$$\pi(\boldsymbol{\eta}) \propto \frac{1}{(1 + \|\boldsymbol{\eta}\|^2)^{(d-1)/2}}, \quad \boldsymbol{\eta} \in \mathbb{R}^d, \tag{12}$$

$$\pi(\boldsymbol{V}_j) \propto \frac{1}{|\boldsymbol{V}_j|^{1-1/(2k_j)} \prod_{s<t} (\omega_{js} - \omega_{jt})},$$

$$\boldsymbol{V}_j > 0,\ j \in [r], \tag{13}$$

where $\omega_{j1} > \omega_{j2} > \cdots > \omega_{jk_j} > 0$ are the decreasingly ordered eigenvalues of $\boldsymbol{V}_j$, $j \in [r]$. Apart from prior (12), common choices of prior on $\boldsymbol{\eta}$ include the constant prior and conjugate prior. None of the three priors will result in improper priors or difficulties in computation. However, among the three priors, prior (12) is the most perfect for all k from the perspective of admissibility. Besides, it refers to Berger et al. (2005) that the prior (12) is a mixture-of-normal prior of the hierarchical structure as

$$(\boldsymbol{\eta}|\lambda) \sim N_d(\mathbf{0}, \lambda \boldsymbol{I}_d) \quad \text{and} \quad [\lambda] \propto \lambda^{-1/2} \exp\left(-\frac{1}{2\lambda}\right), \tag{14}$$

and those mixture-of-normal priors have shown great success in shrinkage estimation particularly (cf. Fourdrinier et al., 1998) and robust Bayesian estimation generally (cf. Berger, 1980). Therefore, prior (12) was actually recommended by Berger et al. (2005) for default use.

As for prior (13) on the unknown covariance matrices $\boldsymbol{V}_j$, $j \in [r]$, consider the transformation from $\boldsymbol{V}_j$ to $\boldsymbol{\Omega}_j = \mathrm{diag}\left(\omega_{j1}, \ldots, \omega_{jk_j}\right)$ and the orthogonal matrix $\boldsymbol{\Gamma}_j$ of corresponding eigenvectors, the Jacobian is

$$\left| \frac{\partial\, \boldsymbol{V}_j}{\partial\, (\boldsymbol{\Omega}_j, \boldsymbol{\Gamma}_j)} \right| = \prod_{s<t} (\omega_{js} - \omega_{jt}). \tag{15}$$

Consequently, the prior (13) on $\boldsymbol{V}_j$ becomes the prior density of $(\boldsymbol{\Omega}_j, \boldsymbol{\Gamma}_j)$ as

$$\pi(\boldsymbol{\Omega}_j, \boldsymbol{\Gamma}_j) \propto \frac{1}{|\boldsymbol{\Omega}_j|^{1-1/(2k_j)}} \tag{16}$$

with respect to Lebesgue measure on $\left(\omega_{j1}, \ldots, \omega_{jk_j}\right)$ and the invariant Haar measure over the space $\left\{\boldsymbol{\Gamma} : \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top = \boldsymbol{I}_{k_j}\right\}$. Note that the prior on $\boldsymbol{\Omega}_j$ is improper and, independently, the prior on $\boldsymbol{\Gamma}_j$ is constant. Use of a uniform prior for $\boldsymbol{\Gamma}_j$ ranging over a compact space is natural and non-controversial and has no influence on the eigenvalues. The term $\prod_{s<t} (\omega_{js} - \omega_{jt})$ is eliminated after changing variables for prior (13). In contrast, the commonly used priors on the covariance matrix, such as inverse Wishart, Jeffreys-rule and constant priors, contain the term $\prod_{s<t} (\omega_{js} - \omega_{jt})$ in the transformed space. This special term gives low mass to close eigenvalues and hence effectively force the eigenvalues apart. It is contrary to the common intuition which would suggest that one should choose a prior that pushes the eigenvalues closer together. As a result, prior (13) is essentially neutral as to expansion or shrinkage of the eigenvalues.

In the context of 2-level hierarchical normal model, Theorem 1 from Berger et al. (2020b) has demonstrated that the combination of priors (12) and (13) on $(\boldsymbol{\eta}, \mathcal{V})$

is on the boundary of admissibility, being as diffuse as possible without yielding inadmissible estimators. Furthermore, it is shown that the generalization that allows covariates at all levels of the hierarchical model will not affect the result of admissibility (cf. Berger et al., 2020b). Nonetheless, the admissibility of the recommended prior for the $(r+1)$-level hierarchical model with $r \geq 2$ is not clear. Generally speaking, this is a very difficult question to answer, and we mainly justify the recommendation of hyperpriors from the angles of posterior propriety and computation afterwards in the framework of the GHNL model.

3. Posterior propriety

Berger et al. (2020b) has shown that the resulting posterior of the recommended prior is proper for the 3-level hierarchical model (2), but under a narrow set of assumptions. They also conjectured the posterior propriety for a hierarchical model with any number of levels, a rigorous one of which was not given yet. In this section, we will comprehensively investigate the conditions for the posterior propriety of the GHNL model (4) using the recommended prior in more general situations. The dimension of $\boldsymbol{\eta}$ affects the investigation of posterior propriety considerably, and two cases $d \geq 2$ or $d = 1$ will be discussed separately.

Based on (5) and (14), by integrating out $\boldsymbol{\eta}$, we can obtain the marginal distribution of $\boldsymbol{y}$ conditioning on $(\mathcal{V}, \lambda)$ as

$$(\boldsymbol{y}|\mathcal{V}, \lambda) \sim N_{m_0 k_0}\left(\mathbf{0}, \boldsymbol{\Delta} + \lambda \boldsymbol{X}_r \boldsymbol{X}_r^\top\right). \tag{17}$$

The posterior propriety of the GHNL model (4) employing priors (13) and (14) is defined as

$$m(\boldsymbol{y}) = \int f(\boldsymbol{y}|\mathcal{V}, \lambda)\pi(\lambda) \prod_{s=1}^{r} \pi(\boldsymbol{V}_s)\, \mathrm{d}\lambda \prod_{t=1}^{r} \mathrm{d}\boldsymbol{V}_t < \infty, \tag{18}$$

where $m(\boldsymbol{y})$ denotes the marginal density of the observation vector. Next, we display some definitions and additional notations which are frequently used in this section.

More Notations Let $\mathrm{card}(A)$ denote the cardinality of the set A. For $0 < I_1, I_2 \leq \infty$, denote that $I_1 \simeq I_2$ if there exist constants $0 < C_1 \leq C_2$ such that $C_1 I_2 \leq I_1 \leq C_2 I_2$; For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\lambda_i(\boldsymbol{A})$ represents the i-th largest eigenvalue of $\boldsymbol{A}$, namely, $\lambda_1(\boldsymbol{A}) \geq \cdots \geq \lambda_n(\boldsymbol{A})$; Let $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ denote the maximum and minimum eigenvalues of an arbitrary symmetric matrix $\boldsymbol{A}$.

Definition: For convenience, let $\omega_{r+1,1} \triangleq \lambda$, $m_{r+1} \triangleq d$ and $k_{r+1} \triangleq 1$. Define a function of an arbitrary non-empty set E as $\mathcal{F}(E) \triangleq \{D \,|\, D \subset E,\ D \neq \emptyset\}$, such that $\mathcal{F}(E)$ denotes the set of all non-empty subsets of E. For any $R \in \mathcal{F}([r+1])$, let $H(R) \triangleq \left\{(j,l) \,|\, j \in R,\ l \in [k_j]\right\}$. Further define the composition of $\mathcal{F}$ and H as

$$\mathcal{S}(R) \triangleq (\mathcal{F} \circ H)(R) = \{D \,|\, D \subset H(R),\ D \neq \emptyset\}$$

for any $R \in \mathcal{F}([r+1])$. Define that

$$c_{jl,s} = 1_{\{m_j(l-1) < s \leq m_j l\}}, \tag{19}$$

for $j \in [r+1]$, $l \in [k_j]$ and $s \in [m_0 k_0]$.

3.1. Two key lemmas

Before we formally investigate the posterior propriety, we first introduce two important lemmas which dominate in the process of proving the main theorems in this paper.

Lemma 3.1: *Assume that $\boldsymbol{A}_j$ are $p_j \times p_j$ positive definite matrices, $j \in [r]$. Let $\boldsymbol{X}_j$ be $n \times p_j$ matrices of full column ranks, $j \in [r]$. Define that*

$$\boldsymbol{H} = \sum_{j=1}^{r} \boldsymbol{X}_j \boldsymbol{A}_j \boldsymbol{X}_j^\top. \tag{20}$$

Then

(a) $\lambda_{\max}(\boldsymbol{H}) \leq C_1 \sum_{j=1}^{r} \lambda_{\max}(\boldsymbol{A}_j)$, *where* $C_1 = \max_{j\in[r]} \lambda_{\max}(\boldsymbol{X}_j^\top \boldsymbol{X}_j)$.

(b) Also,

$$|\boldsymbol{H}| \geq \left(\frac{C_2}{r}\right)^n \left|\sum_{j=1}^{r} \boldsymbol{D}_j\right|, \tag{21}$$

where $C_2 = \min_{j\in[r]} \lambda_{\min}(\boldsymbol{X}_j^\top \boldsymbol{X}_j) > 0$. For any $j \in [r]$, $\boldsymbol{D}_j = \mathrm{diag}(a_{j1}, \ldots, a_{jn})$, where $a_{jk} = \lambda_k\left(\boldsymbol{A}_j\right)$ for $k \in [p_j]$ and $a_{jk} = 0$ for $p_j < k \leq n$.

Lemma 3.1 mainly demonstrates two inequalities with respect to the summation of a series of quadratic forms which have matrical inputs. Since $\boldsymbol{X}_i$'s have full column rank and $\boldsymbol{A}_j$'s are positive definite, then $n \geq p_j$ and rank $\left(\boldsymbol{X}_j \boldsymbol{A}_j \boldsymbol{X}_j^\top\right) = p_j$. It is worthwhile to note that the non-zero diagonal elements of $\boldsymbol{D}_j$ are the decreasingly ordered eigenvalues of $\boldsymbol{A}_j$ in the lower bound of $|\boldsymbol{H}|$, and this relation will deeply influence the last result when we derive the sufficient condition for posterior propriety afterwards. Besides, we can never find a constant $C^* > 0$ such that $|\boldsymbol{H}|_+ \leq C^* \left|\sum_{j=1}^{r} \boldsymbol{D}_j\right|_+$, where $|\boldsymbol{M}|_+$ denotes the product of all non-zero eigenvalues of $\boldsymbol{M}$. The proof of Lemma 3.1 can be found in Appendix A.2.

Lemma 3.2: *Assume a positive integer k. Suppose $\mathcal{M}$ is a subset of $\mathcal{F}([k])$ with cardinality n and let $C_1, \ldots, C_n$*

denote the sequence of all the elements within $\mathcal{M}$. Define an integral

$$I = \int_{\Omega} \frac{1}{\left[\prod_{j=1}^{k} \lambda_j^{a_j}\right]\left[\prod_{i=1}^{n}(1+\sum_{r\in C_i}\lambda_r)^{b_i}\right]} \mathrm{d}\boldsymbol{\lambda},$$

where a_j $j = 1, 2, \ldots, k$, and b_i, $i = 1, 2, \ldots, n$, are real constants, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)^\top \in \Omega \equiv [0, \infty]^k$. Then the integral I is finite if and only if (iff) the following two conditions are both satisfied.

(a) $a_j < 1, j \in [k]$;
(b) *inequalities*

$$\sum_{j\in D} a_j + \sum_{i\in \mathcal{G}_D} b_i > \mathrm{card}(D) \tag{22}$$

for all $D \in \mathcal{F}([k])$ hold, where $\mathcal{G}_D = \left\{i \mid D \bigcap C_i \neq \emptyset, i \in [n]\right\}$.

Here, Lemma 3.2 (b) may not be straightforward enough, so we will use the following example to elaborate on how Lemma 3.2 can be employed.

Example 3.1: Consider integral

$$I_0 = \int_{\Omega} \frac{1}{\left[\prod_{j=1}^{3} \lambda_j^{a_j}\right](1+\lambda_1)^{b_1}(1+\lambda_1+\lambda_2)^{b_2}(1+\lambda_2+\lambda_3)^{b_3}} \mathrm{d}\lambda_1\, \mathrm{d}\lambda_2\, \mathrm{d}\lambda_3,$$

where a_1, a_2, a_3, b_1, b_2, b_3 are all real constants. Similar to Lemma 3.2, we can define $\mathcal{M} = \{\{1\}, \{1, 2\}, \{2, 3\}\}$. Then $I_0 < \infty$ iff all the following inequalities hold: (a) $a_j < 1$, $j \in [3]$; (b)

$$\begin{aligned}
&\text{for } D = \{1\}, \quad a_1 + b_1 + b_2 > 1;\\
&\text{for } D = \{2\}, \quad a_2 + b_2 + b_3 > 1;\\
&\text{for } D = \{3\}, \quad a_3 + b_3 > 1;\\
&\text{for } D = \{1, 2\}, \quad a_1 + a_2 + b_1 + b_2 + b_3 > 2;\\
&\text{for } D = \{1, 3\}, \quad a_1 + a_3 + b_1 + b_2 + b_3 > 2;\\
&\text{for } D = \{2, 3\}, \quad a_2 + a_3 + b_2 + b_3 > 2;\\
&\text{for } D = \{1, 2, 3\}, \quad a_1 + a_2 + a_3 + b_1 + b_2 + b_3 > 3.
\end{aligned}$$

Note that, no matter how $\mathcal{M}$ is defined, we always need to check all the inequalities corresponding to all $D \in \mathcal{F}([k])$. Even though some inequalities could be trivial after being written down, for the sake of assurance, we would better take all non-empty subsets of $[k]$ into account in the early stage. Lemma 3.2 plays a crucial role in obtaining the follow-up theorems, and detailed proof of Lemma 3.2 sees Appendix A.2.

3.2. Conditions for the posterior to be proper when $d \geq 2$

In this subsection, the case $d \geq 2$ is mainly considered.

Theorem 3.1: *Consider the GHNL model* (4) *with priors* (12) *and* (13) *on $\boldsymbol{\eta}$ and $\boldsymbol{V}_i \in \mathcal{V}$, respectively. When $d \geq 2$, a sufficient condition for the posterior propriety is given by*

$$\sum_{s=1}^{m_0 k_0} 1_{\left\{\sum_{(j,l)\in D} c_{jl,s} > 0\right\}} > \sum_{(j,l)\in D} \frac{1}{k_j}, \tag{23}$$

for any $D \in \mathcal{S}([r+1])$.

Proof: It follows (17) that the integrated likelihood of $(\mathcal{V}, \lambda)$ after marginalizing out $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r, \eta)$ is given by

$$L(\mathcal{V}, \lambda; \boldsymbol{y}) \propto \frac{1}{|\boldsymbol{\Delta} + \lambda \boldsymbol{X}_r \boldsymbol{X}_r^\top|^{1/2}} \times \exp\left\{-\frac{1}{2}\boldsymbol{y}^\top \left(\boldsymbol{\Delta} + \lambda \boldsymbol{X}_r \boldsymbol{X}_r^\top\right)^{-1} \boldsymbol{y}\right\}. \tag{24}$$

By dropping the exponent term involving $\boldsymbol{y}$ (since it is less than one), we have

$$L(\mathcal{V}, \lambda; \boldsymbol{y}) < \frac{1}{|\boldsymbol{\Delta} + \lambda \boldsymbol{X}_r \boldsymbol{X}_r^\top|^{1/2}}.$$

By applying Lemma 3.1 (b), we can further bound the integrated likelihood as

$$L(\mathcal{V}, \lambda; \boldsymbol{y}) \leq \frac{C_1}{|\boldsymbol{M}_1|^{1/2}} \tag{25}$$

where C_1 is a positive constant that only depends on $\boldsymbol{X}_j$'s and

$$\boldsymbol{M}_1 = \boldsymbol{I}_{m_0 k_0} + \sum_{j=1}^{r+1} \boldsymbol{D}_j \quad \text{and}$$

$$\boldsymbol{D}_j = \begin{pmatrix} \boldsymbol{\Omega}_j \otimes \boldsymbol{I}_{m_j} & \\ & \boldsymbol{O}_{q_j} \end{pmatrix}_{(m_0 k_0)\times(m_0 k_0)}, \tag{26}$$

where $\boldsymbol{\Omega}_j$ are the diagonal matrices of the decreasingly ordered eigenvalues of $\boldsymbol{V}_j$ for $j \in [r]$, $\boldsymbol{\Omega}_{r+1} = (\omega_{r+1,1})$, $\boldsymbol{O}_{q_j}$ are $q_j \times q_j$ zero matrices and $q_j = m_0 k_0 - m_j k_j$ for $j \in [r+1]$. Since $\boldsymbol{\Omega}_{r+1} = (\omega_{r+1,1})$, then $\boldsymbol{\Omega}_{r+1}$ is a degenerate matrix as scalar λ and the prior on λ becomes

$$\pi(\boldsymbol{\Omega}_{r+1}) \propto \frac{1}{|\boldsymbol{\Omega}_{r+1}|^{1-1/(2k_{r+1})}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Omega}_{r+1}^{-1}\right)\right\}. \tag{27}$$

Combining (16), (18), (27) and (25), we have

$$\begin{aligned}
m(\boldsymbol{y}) &\leq \int \frac{C_1 \exp\left\{-\mathrm{tr}\left(\boldsymbol{\Omega}_{r+1}^{-1}\right)/2\right\} \prod_{j=1}^{r+1} 1_{\left\{\omega_{j1} > \cdots > \omega_{jk_j} > 0\right\}}}{|\boldsymbol{M}_1|^{1/2} \prod_{j=1}^{r+1} |\boldsymbol{\Omega}_j|^{1-\frac{1}{2k_j}}} \times \left[\prod_{j=1}^{r} \mathrm{d}\boldsymbol{\Omega}_j\, \mathrm{d}\boldsymbol{\Gamma}_j\right] \mathrm{d}\boldsymbol{\Omega}_{r+1}\\
&< \int \frac{C_1}{|\boldsymbol{M}_1|^{1/2} \prod_{j=1}^{r+1} |\boldsymbol{\Omega}_j|^{1-\frac{1}{2k_j}}} \prod_{j=1}^{r+1} \mathrm{d}\boldsymbol{\Omega}_j \triangleq I_0.
\end{aligned}$$

The definition of $c_{jl,s}$ in (19) yields

$$|\boldsymbol{M}_1| = \prod_{s=1}^{m_0k_0}\left(1+\sum_{j=1}^{r+1}\sum_{l=1}^{k_j} c_{jl,s}\omega_{jl}\right).$$

Therefore,

$$I_0 \simeq \int \frac{\prod_{j=1}^{r+1} \mathrm{d}\boldsymbol{\Omega}_j}{\prod_{j=1}^{r+1}|\boldsymbol{\Omega}_j|^{1-\frac{1}{2k_j}}\prod_{s=1}^{m_0k_0}\left(1+\sum_{j=1}^{r+1}\sum_{l=1}^{k_j} c_{jl,s}\omega_{jl}\right)^{\frac{1}{2}}},$$

which is finite iff

$$\sum_{(j,l)\in D}\left(1-\frac{1}{2k_j}\right)+\frac{1}{2}\sum_{s=1}^{m_0k_0} 1_{\left\{\sum_{(j,l)\in D} c_{jl,s}>0\right\}} > \mathrm{card}(D) \tag{28}$$

for any $D \in \mathcal{S}([r+1])$ by employing Lemma 3.2. It's obvious that the inequality (28) is equivalent to that in (23). ■

Applying Lemma 3.1 yields the upper bound on the integrated likelihood function of hyperparameters as $C_0|\boldsymbol{M}_1|^{-\frac{1}{2}}$, where $\boldsymbol{M}_1$ is an $(m_0k_0)\times(m_0k_0)$ matrix and defined in (26). Therefore, the special notation $c_{jl,s}$ can be understood as the indicator of whether eigenvalue ω_{jl} appears in the s-th diagonal element of $\boldsymbol{M}_1$ for $j \in [r+1]$, $l \in [k_j]$ and $s \in [m_0k_0]$. At the same time, the left-hand side of inequality (23) actually stands for the cardinality of set $\left\{s \mid \exists\,(j,l) \in D, \text{ such that } c_{jl,s} > 0\right\}$ for any $D \in \mathcal{S}([r+1])$.

The cardinality of $\mathcal{S}([r+1])$ is $2^{\sum_{j\in[r+1]}k_j}-1$, which means that the total number of the inequalities to be checked in (23) is exponential with r and the dimensions k_j. It has to be admitted that this will impose considerably heavy computational burden in common practice by applying Theorem 3.1 directly. Nevertheless, the researchers have no need to be anxious about the heavy computational burden, because most of the inequalities in Theorem 3.1 are trivial. To conclude this point, we have the following corollary.

Corollary 3.1: *Recursively define that*

$$\begin{aligned} R_1 &= \left\{j \mid m_j \le r+1, \quad j\in[r+1]\right\};\\ R_2 &= \left\{j \mid m_j \le \mathrm{card}(R_1), \quad j\in R_1\right\};\\ &\vdots\\ R_p &= \left\{j \mid m_j \le \mathrm{card}(R_{p-1}), \quad j\in R_{p-1}\right\},\end{aligned}$$

where p is the smallest positive integer i such that $\left\{j \mid m_j > card(R_i), j\in R_i\right\} = \emptyset$. We call the levels within R_p as kernel levels and denote R_p by R_{ker}. The inequality (23) holds for any $D \in \mathcal{S}([r+1])$ iff the inequality (23) holds for any $D \in \mathcal{S}(R_{\mathrm{ker}})$. Consequently, if $R_{\mathrm{ker}} = \emptyset$, then the posterior is always proper.

Proof: Let $R_1^c = [r+1] - R_1$, thus, $m_j > r+1$ for $j \in R_1^c$ by the definition of R_1. For any $D \in \mathcal{S}([r+1])$, if there exists $j^* \in R_1^c$ such that $(j^*, l) \in D$ for any $l \in [k_j^*]$, then

$$\begin{aligned}\sum_{s=1}^{m_0k_0} 1_{\left\{\sum_{(j,l)\in D} c_{jl,s}>0\right\}} &\ge \max_{(j,l)\in D} m_j > r+1\\ &= \max_{D\in\mathcal{S}([r+1])}\sum_{(j,l)\in D}\frac{1}{k_j},\end{aligned}$$

which is a trivial one. As a result, inequality (23) holds for any $D \in \mathcal{S}([r+1])$ iff inequality (23) holds for any $D \in \mathcal{S}(R_1)$. Since

$$\max_{D\in\mathcal{S}(R_i)}\sum_{(j,l)\in D}\frac{1}{k_j} = \mathrm{card}(R_i),$$

it can be recursively shown that inequality (23) holds for any $D \in \mathcal{S}([r+1])$ iff inequality (23) holds for any $D \in \mathcal{S}(R_i)$ and $i \in [p]$, where p is the smallest positive integer i such that $R_i - R_{i+1} = \emptyset$. ■

By using the technique of extracting kernel levels, we dramatically narrow down the checking region for posterior propriety. Since we should only check the inequalities for the levels within R_{ker}, substantially reducing the number of inequalities to be checked from $2^{\sum_{j\in[r+1]}k_j}-1$ to $2^{\sum_{j\in R_{\mathrm{ker}}}k_j}-1$. Moreover, Corollary 3.1 also indicates two interesting conclusions depicted as follows.

(a) First, it reveals the mechanism how three roles, number of levels, numbers of units in levels and dimensions of levels, affect the posterior propriety simultaneously. Roughly speaking, in the context of GHNL, the more levels with fewer units having lower dimensions, the less likely the posterior is to be proper. For example, if $m_{r-2} = m_{r-1} = m_r = 2$ and $k_{r-2} = k_{r-1} = k_r = 1$, for $D = \left\{(j,l) \mid j = r-2,\ r-1,\ r,\ l=1\right\}$, we have

$$2 = \sum_{s=1}^{m_0k_0} 1_{\left\{\sum_{(j,l)\in D} c_{jl,s}>0\right\}} < \sum_{(j,l)\in D}\frac{1}{k_j} = 3. \tag{29}$$

Therefore, the posterior propriety can hardly be guaranteed by Theorem 3.1. Conversely, if the units in each level are adequate enough such that the set of kernel levels is empty, i.e., $R_{\mathrm{ker}} = \emptyset$, then the posterior is always proper. As a consequence, more attention should be focused on the levels with small number of units, namely, the kernel levels.

(b) Second, the recommended prior for use at any level in hierarchical modelling is further justified from

the aspect of posterior propriety and ease of implementation. For instance, if we switch the prior on $\boldsymbol{V}_j, j \in [r]$ from (13) to

$$\pi(\boldsymbol{V}_j) \propto \frac{1}{|\boldsymbol{V}_j|^{1-a} \prod_{1 \le s < t \le k_j} (\omega_{js} - \omega_{jt})}, \quad \boldsymbol{V}_j > 0,\ j \in [r], \tag{30}$$

where $0 < a \le 1$ (a has to be larger than zero by Lemma (3.2)). Then the condition in Theorem 3.1 becomes

$$\sum_{s=1}^{m_0 k_0} 1_{\left\{\sum_{(j,l)\in D} c_{jl,s} > 0\right\}} > 2a \times \mathrm{card}(D), \quad D \in \mathcal{S}([r+1]). \tag{31}$$

On one hand, when a is greater than but close to zero (denoted by $a \gtrsim 0$), all the inequalities in (31) always hold. Thus, the posterior will be always proper. However, it is impractical to decide how small is for a and find such a fixed value that fits all levels. On the other hand, when $a \lesssim 1$, which denotes that a is less than and close to 1, then $2a * \mathrm{card}(D) > \sum_{(j,l)\in D} \frac{1}{k_j}$, concluding that inequality (31) is harder to get reached than inequality (23), especially for large dimensions k_j's. Therefore, the posterior using prior (30) is rather unlikely to be proper than using prior (13). Similar to Corollary 3.1, we can recursively define that

$$\begin{aligned} R_1^* &= \left\{j \mid m_j \le 2a \times \mathrm{card}\,(E_0), \quad j \in [r+1]\right\}; \\ R_2^* &= \left\{j \mid m_j \le 2a \times \mathrm{card}(E_1^*), \quad j \in R_1^*\right\}; \\ &\vdots \\ R_{p^*}^* &= \left\{j \mid m_j \le 2a \times \mathrm{card}(E_{p^*-1}^*), \quad j \in R_{p^*-1}^*\right\}, \end{aligned}$$

where $E_0 = H([r+1])$, $E_i^* = H(R_i^*)$ for $i \in [p^*]$ and p^* is the smallest positive integer l such that $\left\{j \mid m_j > 2a * \mathrm{card}(E_l^*),\ j \in R_l^*\right\} = \emptyset$. Then the posterior using prior (30) instead is proper if (31) holds for any $D \in \mathcal{S}\left(R_{p^*}^*\right)$. When $a \lesssim 1$, $\mathrm{card}\left(R_{p^*}^*\right)$ will be remarkably larger than $\mathrm{card}(R_{\mathrm{ker}})$ for large values of k_j's, imposing a dramatically heavier burden of checking inequalities than that for prior (13). Above all, one sensible choice for a is that let a be inversely proportional to k_j for level j, which takes both practical and theoretical considerations into account.

The upper bound on $\sum_{(j,l)\in D} \frac{1}{k_j}$ as card (R_i) for $D \in \mathcal{S}(R_i)$, leads to an effective way to extract kernel levels as presented in Corollary 3.1, but this upper bound is still too rough. Next, an elaborate upper bound on $\sum_{(j,l)\in D} \frac{1}{k_j}$ is demonstrated and a sufficient condition of clean form for posterior propriety is derived then.

Theorem 3.2: *Consider the GHNL model* (4) *with priors* (12) *and* (13) *on* $\boldsymbol{\eta}$ *and* $\boldsymbol{V}_i \in \mathcal{V}$, *respectively. Denote that* $m^* = \min_{j\in[r+1]} m_j = \min_{j\in R_{\mathrm{ker}}} m_j$. *When* $d \ge 2$, *the posterior is always proper if*

$$\sum_{j\in R_{\mathrm{ker}}} \frac{1}{k_j} < m^*. \tag{32}$$

Proof: For $D \in \mathcal{S}([r+1])$, define that

$$L(D) = \sum_{s=1}^{m_0 k_0} 1_{\left\{\sum_{(j,l)\in D} c_{jl,s} > 0\right\}}, \quad \text{and}$$
$$R(D) = \left\{j \mid (j,l) \in D\right\}. \tag{33}$$

It follows from Corollary 3.1 that we only need to prove

$$\sum_{(j,l)\in D} \frac{1}{k_j} < L(D), \quad \forall D \in \mathcal{S}\left(R_{\mathrm{ker}}\right).$$

Also, for any D belonging to $\mathcal{S}\left(R_{\mathrm{ker}}\right)$, we have

$$\begin{aligned} \sum_{(j,l)\in D} \frac{1}{k_j} &= \sum_{j\in R(D)} \frac{1}{k_j} \sum_{l=1}^{k_j} 1_{\{(j,l)\in D\}} \\ &\le \left(\sum_{j\in R(D)} \frac{1}{k_j}\right) \max_{j\in R(D)} \sum_{l=1}^{k_j} 1_{\{(j,l)\in D\}}. \end{aligned} \tag{34}$$

Distinct eigenvalues from the same level ($\le r$-th) never occur in the same row of matrix $\boldsymbol{M}_1$. Mathematically, $c_{jl_1,s} c_{jl_2,s} = 0$, for $1 \le l_1 < l_2 \le k_j$, $j \in [r]$, $s \in [m_0 k_0]$. Thus, for any $j \in [r]$,

$$\sum_{l=1}^{k_j} 1_{\{(j,l)\in D\}} \le \frac{1}{m_j} L(D), \tag{35}$$

which is also true for $j = r+1$ since $k_{r+1} = 1$. It is obvious that $\min_{j\in[r+1]} m_j = \min_{j\in R_{\mathrm{ker}}} m_j$ by the definition of R_{ker}, and that is denoted by m^*. Combining (35) with (34) yields

$$\begin{aligned} \sum_{(j,l)\in D} \frac{1}{k_j} &\le \left(\sum_{j\in R(D)} \frac{1}{k_j}\right) \frac{L(D)}{\min_{j\in R(D)} m_j} \\ &\le \frac{1}{m^*} \left(\sum_{j\in R_{\mathrm{ker}}} \frac{1}{k_j}\right) L(D) < L(D). \end{aligned} \tag{36}$$

■

According to the proof procedure of Theorem 3.2, it can be deduced that $\sum_{j\in[r+1]} \frac{1}{k_j} < m^*$ is also sufficient for posterior propriety. Obviously, the condition in Theorem 3.2 is easier to be satisfied. Theorem 3.2 reveals that for fixed m^*, the posterior is more likely to be proper for higher dimensions of the units in the kernel levels. Theorem 3.2 also provides the researchers

with a powerful tool to check the posterior propriety quickly.

Remark 3.1: Consider the model (4) with $r = 1$, namely, a two-level hierarchical model. When $d \geq 2$, we have $m^* = \min\{d, m_1\} \geq 2$. Then the posterior using the recommended prior is always proper for $k_1 \geq 2$ referring to Theorem 3.2.

Example 3.2 (Continue with Example 2.1): Consider the GHNL modelling of the mixed-effect ANOVA as (10), which is a 3-level hierarchical model with $r = 2$, $m_0 = s_1 s_2 s_3$, $m_1 = s_1 s_2$, $m_2 = s_1$, $m_3 = p$, $k_0 = k_1 = k_2 = p$ and $k_3 = 1$. It is natural to assume that we have at least two schools, each school has at least two classes and each class has at least two students, namely, $s_1 \geq 2$, $s_2 \geq 2$ and $s_3 \geq 2$. If (a) $p > 2$ and $s_1 \geq 2$ or (b) $p \geq 2$ and $s_1 > 2$ holds, it can be readily derived that the set of kernel levels is empty. Thus, the posterior is always proper according to Corollary 3.1. When $s_1 = p = 2$, the set of kernel levels is $R_2 = \{2, 3\}$, since $\frac{1}{k_2} + \frac{1}{k_3} < 2$, then the posterior is proper by applying Theorem 3.2. In conclusion, the posterior using the recommended prior is always proper when $p \geq 2$.

In Berger et al. (2020b)'s work, for a technical reason, they assumed $k = sp$ for 3-level hierarchical normal model (2) such that the design matrices for units within level 2 and 3 are square matrices. They eventually reached a conclusion that the posterior employing the recommended prior in model (2) is always proper for $k = sp$ and $p \geq 2$ (p is the dimension of hypermean in model (2)). However, the assumption that the design matrices for the units at high levels are square matrices appears to be unnatural and hard to be interpreted in practice. Nevertheless, we still generalize Berger et al. (2020b)'s result to the GHNL model in the following so as to draw a consistent conclusion with theirs.

Corollary 3.2: *Consider the GHNL model* (4) *with priors* (12) *and* (13) *on* $\boldsymbol{\eta}$ *and* $\boldsymbol{V}_i \in \mathcal{V}$, *respectively. Assume that* $d \geq 2$ *and* $k_j = m_{j+1} k_{j+1}$, $j \in [r]$. *Then the posterior is always proper.*

Proof: Since $m_j \geq 2$, $j \in [r+1]$, then $m^* \geq 2$. It remains to show that $\sum_{j \in R_{\ker}} \frac{1}{k_j} < 2$ by Theorem 3.2. By utilizing the condition that $k_j = m_{j+1} k_{j+1}$ for $j \in [r]$ and $k_{r+1} = 1$, we have $k_j = \prod_{s=j+1}^{r+1} m_s \geq 2^{r+1-j}$ for $j \in [r]$. Thus,

$$\sum_{j \in R_{\ker}} \frac{1}{k_j} \leq \sum_{j \in [r+1]} \frac{1}{k_j} \leq \sum_{j \in [r+1]} \frac{1}{2^{r+1-j}} = 2 - \frac{1}{2^r} < 2,$$

which completes the proof. ■

Above all, a general procedure for checking the posterior propriety of the GHNL models (4) employing the recommended prior for $d \geq 2$ can be summarized as follows.

Guidance for checking the posterior propriety when $d \geq 2$:

(a) If the design matrices for each unit in each level are square matrices, then the posterior is proper, otherwise, turn to (b).
(b) Derive the set of kernel levels, $R_{\ker}$. If $R_{\ker} = \emptyset$ or inequality (32) holds, the the posterior is proper. If neither, turn to (c).
(c) Check inequality (23) for all D belonging to $\mathcal{S}(R_{\ker})$. If that always holds, the the posterior is proper. If not, the posterior propriety can hardly be guaranteed.

3.3. Conditions for the posterior to be proper when $d = 1$

It is quite common that the dimension of the fixed effect $\boldsymbol{\eta}$ is only one in practice. However, when $d = 1$, note that

$$1 = m_{r+1} = \sum_{s=1}^{m_0 k_0} 1_{\left\{\sum_{(j,l) \in D} c_{jl,s} > 0\right\}} = \sum_{(j,l) \in D} \frac{1}{k_j} = \frac{1}{k_{r+1}} = 1$$

for $D = \{(r+1, 1)\}$, resulting in the failure of the sufficient condition in Theorem 3.1. Therefore, in this subsection, we mainly reinvestigate the conditions for the the posterior to be proper for $d = 1$.

Theorem 3.3: *Consider the GHNL model* (4) *with constant prior on* η *and prior* (13) *on* $\boldsymbol{V}_j \in \mathcal{V}$. *When* $d = 1$, *the posterior is proper if*

$$\sum_{s=1}^{m_0 k_0} 1_{\left\{\sum_{(j,l) \in D} c_{jl,s} > 0\right\}} > 1_{\{\exists j \in [r],\ (j,1) \in D\}} + \sum_{(j,l) \in D} \frac{1}{k_j} \tag{37}$$

holds for all $D \in \mathcal{S}([r])$.

Proof: When $d = 1$, vector $\boldsymbol{\eta}$ will degenerate into a scalar η. Hence the prior (12) on $\boldsymbol{\eta}$ becomes a constant prior on η. Based on (5), by integrating out over η, we can get the upper bound on the integrated likelihood of $\mathcal{V}$ after dropping the exponential term (less than one)

$$L(\mathcal{V}) < \frac{1}{|\boldsymbol{\Delta}|^{\frac{1}{2}} \left|\boldsymbol{X}_r^\top \boldsymbol{\Delta}^{-1} \boldsymbol{X}_r\right|^{\frac{1}{2}}}.$$

Since $\left|\boldsymbol{X}_r^\top \boldsymbol{\Delta}^{-1}\boldsymbol{X}_r\right| \geq \boldsymbol{X}_r^\top \boldsymbol{X}_r \lambda_{\min}\left(\boldsymbol{\Delta}^{-1}\right) = \boldsymbol{X}_r^\top \boldsymbol{X}_r \lambda_{\max}(\boldsymbol{\Delta})^{-1}$, using Lemma 3.1 (a) and (b), we have

$$L(\mathcal{V}) < C_0\,|\boldsymbol{\Delta}|^{-\frac{1}{2}}\left(1+\sum_{j=1}^{r}\omega_{j1}\right)^{\frac{1}{2}} \leq C_1\left(1+\sum_{j=1}^{r}\omega_{j1}\right)^{\frac{1}{2}}|\boldsymbol{M}_2|^{-\frac{1}{2}},$$

where C_0 and C_1 are constants that are independent of the $\boldsymbol{\Omega}_j$ for $j \in [r]$, $\boldsymbol{M}_2 = \boldsymbol{I}_{m_0k_0} + \sum_{j=1}^{r}\boldsymbol{D}_j$ and $\boldsymbol{D}_j$'s are defined in (26). Similar to the proof of Theorem 3.1, we can derive the upper bound on $m(\boldsymbol{y})$ as

$$m(\boldsymbol{y}) \leq \int \frac{C_1\left(1+\sum_{j=1}^{r}\omega_{j1}\right)^{\frac{1}{2}}}{|\boldsymbol{M}_2|^{1/2}\prod_{j=1}^{r}\left|\boldsymbol{\Omega}_j\right|^{1-\frac{1}{2k_j}}}\prod_{j=1}^{r}\mathrm{d}\boldsymbol{\Omega}_j \triangleq I_1.$$

It follows from the definition of $c_{jl,s}$ in (19) that

$$|\boldsymbol{M}_2| = \prod_{s=1}^{m_0k_0}\left(1+\sum_{j=1}^{r}\sum_{l=1}^{k_j}c_{jl,s}\omega_{jl}\right).$$

Thus,

$$I_0 \simeq \int \frac{\left(1+\sum_{j=1}^{r}\omega_{j1}\right)^{\frac{1}{2}}\prod_{j=1}^{r}\mathrm{d}\boldsymbol{\Omega}_j}{\prod_{j=1}^{r}\left|\boldsymbol{\Omega}_j\right|^{1-\frac{1}{2k_j}}\prod_{s=1}^{m_0k_0}\left(1+\sum_{j=1}^{r}\sum_{l=1}^{k_j}c_{jl,s}\omega_{jl}\right)^{\frac{1}{2}}},$$

which is finite iff

$$\sum_{(j,l)\in D}\left(1-\frac{1}{2k_j}\right)+\frac{1}{2}\sum_{s=1}^{m_0k_0}1_{\left\{\sum_{(j,l)\in D}c_{jl,s}>0\right\}} - \frac{1}{2}1_{\{\exists j\in[r],\,(j,1)\in D\}} > \mathrm{card}(D) \tag{38}$$

for any $D \in \mathcal{S}([r])$ by employing Lemma 3.2. It's obvious that inequality (38) is equivalent to that in (37). ■

Resembling the interpretation of Theorem 3.1, the left-hand side of inequality (37) actually denotes the cardinality of set $\left\{s \mid \exists\,(j,l) \in D, \text{ such that } c_{jl,s} > 0\right\}$ for any $D \in \mathcal{S}([r])$. To reduce the burden of checking inequalities, we have the following corollary.

Corollary 3.3: *Recursively define that*

$$\tilde{R}_1 = \left\{j \mid m_j \leq r+1,\ j \in [r]\right\};$$
$$\tilde{R}_2 = \left\{j \mid m_j \leq \mathrm{card}(\tilde{R}_1)+1,\ j \in \tilde{R}_1\right\};$$
$$\vdots$$
$$\tilde{R}_q = \left\{j \mid m_j \leq \mathrm{card}(\tilde{R}_{p-1})+1,\ j \in \tilde{R}_{p-1}\right\},$$

where q is the smallest positive integer i such that $\left\{j \mid m_j > card(\tilde{R}_i)+1,\ j \in \tilde{R}_i\right\} = \emptyset$. We call the levels within $\tilde{R}_q$ as kernel levels and denote $\tilde{R}_q$ by $\tilde{R}_{\mathrm{ker}}$. Inequality (37) *holds for any $D \in \mathcal{S}([r])$ iff inequality* (37) *holds for any $D \in \mathcal{S}\left(\tilde{R}_{\mathrm{ker}}\right)$. Consequently, if $\tilde{R}_{\mathrm{ker}} = \emptyset$, then the resulting posterior is always proper.*

In the process of extracting kernel levels, the thresholds of m_j to split up levels are increased by one in Corollary 3.3, when compared with that in Corollary 3.1, and this is because the upper bound on the right-hand side of inequality (37) is increased by one than that of inequality (23). Except for this point, the proof of Corollary 3.1 is same as that of Corollary 3.3. For good measure, a simple tool to check the posterior propriety is shown as follows, which is a counterpart of Theorem 3.2 for $d = 1$.

Theorem 3.4: *Consider the GHNL model* (4) *with constant prior on η and prior* (13) *on $\boldsymbol{V}_j \in \mathcal{V}$. When $d = 1$, the posterior is always proper if*

$$\sum_{j\in\tilde{R}_{\mathrm{ker}}}\frac{1}{k_j} < m^* - 1, \tag{39}$$

where $m^* = \min_{j\in[r]} m_j$ and $\tilde{R}_{\mathrm{ker}}$ is the derived set of kernel levels.

Proof: For any $D \in \mathcal{S}([r])$, define $L(D)$ and $R(D)$ in the same way as that in (33). According to Corollary (3.3), it suffices to show that

$$L(D) > \sum_{(j,l)\in D}\frac{1}{k_j}+1$$

for any $D \in \mathcal{S}\left(\tilde{R}_{\mathrm{ker}}\right)$. Similar to (36), we have

$$\sum_{(j,l)\in D}\frac{1}{k_j} \leq \frac{1}{m^*}\left(\sum_{j\in R(D)}\frac{1}{k_j}\right)L(D) \leq \frac{1}{m^*}\left(\sum_{j\in\tilde{R}_{\mathrm{ker}}}\frac{1}{k_j}\right)L(D) < L(D) - \frac{L(D)}{m^*},$$

for any $D \in \mathcal{S}\left(\tilde{R}_{\mathrm{ker}}\right)$. Since $L(D) \geq m^*$ always holds, then the proof is completed. ■

Remark 3.2: In model (4), when $r = 1$ and $d = 1$, suppose $m_1 \geq 2$ and $k_1 \geq 2$. The posterior using the recommended prior is always proper by applying Theorem 3.4.

Remark 3.3: If all k_j's are equal to one, the sufficient condition in Theorem 3.3 can be simplified as

$$\mathrm{card}(D) < \max_{j\in D} m_j - 1,$$

where $D \in \mathcal{F}([r])$. By employing the technique of extracting kernel levels, Theorem 3.4 is equivalent to Theorem 3.3, rather than a mere sufficient condition.

Example 3.3 (Continue with Example 2.1): Consider model (10) with assuming that $s_1 \geq 2$, $s_2 \geq 2$ and $s_3 \geq 2$. If $p = 1$, we have $k_0 = k_1 = k_2 = k_3 = m_3 = 1$. When $s_1 > 2$, we can easily derive the set of kernel levels as empty set. Thus, the posterior is always proper by Corollary 3.3. If $s_1 = 2$, inequality (37) fails, and the posterior propriety can hardly be guaranteed. Consequently, when $p = 1$, the posterior using the recommended prior is always proper for $s_1 > 2$.

Next, we generalize Berger et al. (2020b)'s result to the GHNL models for $d = 1$, assuming that the design matrices $\boldsymbol{Z}_{jij}$'s for units are square matrices.

Corollary 3.4: *When $d = 1$, consider the same model and prior as Theorem 3.3. Suppose $m_j \geq 3$ and $k_j = m_{j+1}k_{j+1}$, $j \in [r]$. Then the posterior is always proper.*

Proof: It follows from Theorem 3.4 that we should only present

$$\sum_{j\in[r]} \frac{1}{k_j} < m^* - 1.$$

According to the conditions that $k_j = m_{j+1}k_{j+1}, j \in [r]$ and $m_{r+1} = k_{r+1} = 1$, we have $k_j = \prod_{s=j+1}^{r} m_s, j \in [r]$. Thus,

$$\sum_{j\in[r]} \frac{1}{k_j} \leq \sum_{j\in[r]} \frac{1}{3^{r-j}} = \frac{3}{2}\left(1 - \frac{1}{3^r}\right) < \frac{3}{2} < 2 \leq m^* - 1.$$

■

Summing up the theoretical results above, a general procedure for checking the posterior propriety of the GHNL models (4) employing the recommended prior for $d = 1$ can be depicted as follows.

Guidance for checking the posterior propriety when $d = 1$:

(a) If the design matrices for each unit in each level are square matrices and $m_j \geq 3$, $j \in [r]$, then the posterior is proper, otherwise, turn to (b).
(b) Derive the set of kernel levels $\tilde{R}_{\text{ker}}$. If $\tilde{R}_{\text{ker}} = \emptyset$ or inequality (39) holds, then the posterior is proper. If neither, turn to (c).
(c) Check inequality (37) for all D belonging to $\mathcal{S}\left(\tilde{R}_{\text{ker}}\right)$. If that always holds, then the posterior is proper. If not, the posterior propriety can hardly be guaranteed.

4. Computation

In this section, we consider the MCMC sampling from the posterior arising from the model in Section 2. For the GHNL model (4) with prior (14) and (13) on $\boldsymbol{\eta}$ and $\mathcal{V}$, respectively, the joint posterior of $(\boldsymbol{\Theta}, \mathcal{V}, \boldsymbol{\eta}, \lambda)$ can be written as

$$\pi\left(\boldsymbol{\Theta}, \mathcal{V}, \boldsymbol{\eta}, \lambda|\boldsymbol{y}\right) \propto f(\boldsymbol{y}|\boldsymbol{\theta}_1) \prod_{j=1}^{r-1} f(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{j+1}, \boldsymbol{V}_j) f(\boldsymbol{\theta}_r|\boldsymbol{\eta}, \boldsymbol{V}_r) \times \prod_{s=1}^{r} \pi(\boldsymbol{V}_s)\pi(\boldsymbol{\eta}|\lambda)\pi(\lambda). \tag{40}$$

Sampling $(\boldsymbol{\Theta}, \mathcal{V}, \boldsymbol{\eta}, \lambda)$ from the posterior density (40) can be handled by Gibbs sampling method. The main difficulty of the computation is to sample the covariance matrices $\boldsymbol{V}_j$'s efficiently.

4.1. Gibbs sampling for input effects

The full conditionals of the input effects $(\boldsymbol{\Theta}, \boldsymbol{\eta})$ can be derived from the joint posterior (40) and are illustrated as follows.

(a) Conditioning on $\boldsymbol{\theta}_2$ and $\boldsymbol{V}_1$, the posterior distribution of $\boldsymbol{\theta}_1$ is

$$(\boldsymbol{\theta}_1 \,|\, \boldsymbol{\theta}_2, \boldsymbol{V}_1; \boldsymbol{y}) \sim N_{m_1k_1}\left(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{V}}_1\right), \tag{41}$$

where

$$\tilde{\boldsymbol{V}}_1 = \left(\boldsymbol{Z}_0^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_0 + \boldsymbol{I}_{m_1} \otimes \boldsymbol{V}_1^{-1}\right)^{-1},$$
$$\tilde{\boldsymbol{\theta}}_1 = \tilde{\boldsymbol{V}}_1\left[\boldsymbol{Z}_0^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{y} + \left(\boldsymbol{I}_{m_1} \otimes \boldsymbol{V}_1^{-1}\right)\boldsymbol{Z}_1\boldsymbol{\theta}_2\right].$$

(b) The full conditional posteriors of $\boldsymbol{\theta}_j$, $j = 2, \ldots, r$ have the forms:

$$(\boldsymbol{\theta}_j \,|\, \boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_{j+1}, \boldsymbol{V}_{j-1}, \boldsymbol{V}_j) \sim N_{m_jk_j}\left(\tilde{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{V}}_j\right), \tag{42}$$

where

$$\tilde{\boldsymbol{V}}_j = \left[\boldsymbol{Z}_{j-1}^\top\left(\boldsymbol{I}_{m_{j-1}} \otimes \boldsymbol{V}_{j-1}^{-1}\right)\boldsymbol{Z}_{j-1} + \boldsymbol{I}_{m_j} \otimes \boldsymbol{V}_j^{-1}\right]^{-1},$$
$$\tilde{\boldsymbol{\theta}}_j = \tilde{\boldsymbol{V}}_j\left[\boldsymbol{Z}_{j-1}^\top\left(\boldsymbol{I}_{m_{j-1}} \otimes \boldsymbol{V}_{j-1}^{-1}\right)\boldsymbol{\theta}_{j-1} + \left(\boldsymbol{I}_{m_j} \otimes \boldsymbol{V}_j^{-1}\right)\boldsymbol{Z}_j\boldsymbol{\theta}_{j+1}\right],$$

and $\boldsymbol{\theta}_{r+1} \triangleq \boldsymbol{\eta}$.

(c) By using (14), the full conditional of $\boldsymbol{\eta}$ can be derived as

$$(\boldsymbol{\eta} \,|\, \boldsymbol{\theta}_r, \lambda, \boldsymbol{V}_r) \sim N_d\left(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{V}}_{\boldsymbol{\eta}}\right), \tag{43}$$

where

$$\tilde{\boldsymbol{V}}_{\boldsymbol{\eta}} = \left[\boldsymbol{Z}_r^\top\left(\boldsymbol{I}_{m_r} \otimes \boldsymbol{V}_r^{-1}\right)\boldsymbol{Z}_r + \lambda^{-1}\boldsymbol{I}_d\right]^{-1},$$
$$\tilde{\boldsymbol{\eta}} = \tilde{\boldsymbol{V}}_{\boldsymbol{\eta}}\boldsymbol{Z}_r^\top\left(\boldsymbol{I}_{m_r} \otimes \boldsymbol{V}_r^{-1}\right)\boldsymbol{\theta}_r.$$

Input effects $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}$ and $\boldsymbol{\eta}$ can be readily sampled from their conditionals during the Gibbs sampling procedure, as their full conditional posterior distributions are all standard distributions.

4.2. Gibbs sampling for variance components

The variance components which include $\boldsymbol{V}_j$'s and λ can be updated from their full conditionals, and these conditionals have densities as follows.

(a) Given $\boldsymbol{\eta}$, the conditional posterior density of λ is

$$\pi(\lambda|\boldsymbol{\eta}) \propto \lambda^{-\frac{d+1}{2}} \exp\left\{-\frac{1+\|\boldsymbol{\eta}\|^2}{2\lambda}\right\}, \tag{44}$$

which is actually an inverse gamma distribution as $IG\left(\frac{d-1}{2}, \frac{1+\|\boldsymbol{\eta}\|^2}{2}\right)$.

(b) For $j \in [r]$, define that $t_j = \frac{m_j}{2} + 1 - \frac{1}{2k_j}$. The marginal posterior density of $\boldsymbol{V}_j$ given $(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{j+1})$ is

$$\pi\left(\boldsymbol{V}_j|\boldsymbol{\theta}_j, \boldsymbol{\theta}_{j+1}\right) \propto \frac{1}{\left|\boldsymbol{V}_j\right|^{t_j} \prod_{1\le s<t\le k_j}\left(\omega_{js}-\omega_{jt}\right)} \times \operatorname{etr}\left\{-\frac{1}{2}\boldsymbol{V}_j^{-1}\boldsymbol{H}_j\right\}, \tag{45}$$

where $\operatorname{etr}(\boldsymbol{A})$ denotes $\exp(\operatorname{tr}(\boldsymbol{A}))$ for a square matrix $\boldsymbol{A}$, and

$$\boldsymbol{H}_j \triangleq \boldsymbol{H}_j(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{j+1}) = \sum_{i=1}^{m_j}(\boldsymbol{\theta}_{ji} - \boldsymbol{Z}_{ji}\boldsymbol{\theta}_{j+1})(\boldsymbol{\theta}_{ji} - \boldsymbol{Z}_{ji}\boldsymbol{\theta}_{j+1})^\top.$$

The updating of λ can be simply carried out by sampling from an inverse gamma distribution. The full conditional posteriors of $\boldsymbol{V}_j$, (45) are actually distributed as a recently proposed class of prior distributions by Berger et al. (2020a) for the covariance matrix, which is called the Shrinkage Inverse Wishart (SIW) distributions. The new class $\mathrm{SIW}(a, \boldsymbol{H})$ for a $k \times k$ covariance matrix $\boldsymbol{W}$ has the density as

$$\pi^{\mathrm{SIW}}(\boldsymbol{W} \mid a, \boldsymbol{H}) \propto \frac{\operatorname{etr}\left(-\frac{1}{2}\boldsymbol{W}^{-1}\boldsymbol{H}\right)}{|\boldsymbol{W}|^a \prod_{i<j}\left(\nu_i - \nu_j\right)}, \tag{46}$$

where $\nu_1 > \nu_2 > \cdots > \nu_k > 0$ are the ordered eigenvalues of $\boldsymbol{W}$, a is a real constant and $\boldsymbol{H}$ is a $k \times k$ non-negative definite matrix. Thus, $\boldsymbol{V}_j$ are distributed as $\mathrm{SIW}\left(t_j, \boldsymbol{H}_j\right), j \in [r]$. To sample the covariance matrices from the full conditional posteriors, the previously suggested methods include the Metropolis-Hastings algorithm (cf. Berger et al., 2005) and Hit-and-run method (cf. Yang & Berger, 1994). The two methods both generate full candidate matrices by utilizing full-parameter proposal distributions, resulting in that they only work for moderate dimensions of the covariance matrices. To tackle this issue, Berger et al. (2020a) proposed a powerful Gibbs method for efficiently sampling the covariance matrices from their conditional densities and this new method works for higher dimensions k. The audience can refer to Berger et al. (2020a) or Appendix 3 for details of this Gibbs sampling method. According to the simulation results of Berger et al. (2020a), the new Gibbs method outperforms the Metropolis-Hastings and Hit-and-run methods for moderate dimensions and work for k up to 100, while the other two algorithms break down in much lower dimensions.

In the framework of 2-level HNLM, Berger et al. (2020b) compared the numerical performance, from the mean square error (MSE) perspective, of a dozen of objective hyperpriors, which are the product of three objective hyperpriors for the hypermean and four objective hyperpriors for the hypercovariance matrix. Priors on the hypermean include constant prior, conjugate prior and the recommended prior (12). Priors on the hypercovariance matrix include constant prior, hierarchical Jefferys prior, hierarchical reference prior and the recommended prior (13). Their simulation results have shown that the recommended combination of hyperpriors dominates all the others in terms of Bayes risk, and the constant prior on the hypercovariance performs the worst. However, neither of the two remaining choices for hypercovariance is computationally easy. Considering the 4-level HNLM, Song et al. (2020) performed numerical experiment to compare the recommended prior with constant prior for the hypercovariance matrices, and the other two priors were canceled due to intractable computation. Also, Song et al. (2020)'s result presented the domination of the recommended hyperpriors over other priors. In conclusion, both Berger et al. (2020b) and Song et al. (2020) have provided strong numerical evidence of the superiority of the recommended hyperpriors for use in GHNLM, since 2-level and 4-level HNLMs both are specific GHNLM.

5. Discussions

We have proposed a generalized hierarchical normal linear model applicable to the nested data with complex structures. The GHNL model proves to be equivalent to a LMM model, while the GHNL model is more natural for researchers to model nested data from scratch, especially when incorporating covariates at high levels. Like generalizations to the simple normal linear model, the GHNL can be generalized to the hierarchical model with generalized liner model in the first level, and thus discrete observations can be handled. Besides, the first level (or even higher levels) of the GHNL model can be also extended to the setting of semiparametric regression models, such as, single index model and partially linear model. The technique of modelling and investigation in this paper can be applied to the linear part of the models mentioned above. The statistical analysis could be complicated and such explorations are beyond the scope of this paper, however.

Berger et al. (2020b) put an end to the endless search for the appropriate hyperpriors in hierarchical

modelling and investigated properties comprehensively to justify the recommendation. Nonetheless, when it came to the propriety of the resulting posterior, they only suspected that that is true for use at any level for a general hierarchical normal model, the conditions for which were not given. To complete the story, we have studied the conditions for the posterior to be proper in more general situations than Berger et al. (2020b), when employing the recommended prior for the GHNL model. Theorems 3.1 and 3.3 demonstrate the main result, and Corollaries 3.1 and 3.3 reduce the computational burdens by defining kernel sets for $d \geq 2$ and $d = 1$, respectively. In addition, Theorems 3.2 and 3.4 provide powerful tools of simple forms for checking propriety of posterior for $d \geq 2$ and $d = 1$, separately. The user-friendly guidance for checking posterior propriety is eventually supplied. Note that our results only present sufficient conditions, and necessary conditions have never been discussed. The reason is because the derivation of the lower bound on the integrated likelihood of hyperparameters is intractable. Moreover, it is not worthwhile to investigate necessary conditions, as the derived upper bounds are tight enough such that the corresponding sufficient conditions are very modest, according to the remarks and examples in Section 3. At last, an efficient and powerful Gibbs sampling method for sampling from the posterior is introduced, overcoming the bottleneck of computation that the previously proposed sampling method only works for low dimensions or moderate dimensions inefficiently. The numerical evidence supporting the superiority of the recommended prior for hierarchical models was presented in Berger et al. (2020b) and Song et al. (2020).

Though we have made much progress in the hierarchical linear modelling, a major obstacle to applying our results is that the variance component for the first level is supposed to be known, which can hardly be satisfactory in practice. If we assume an unknown covariance matrix $\mathbf{\Sigma}_0$ for the first level and specify prior (13) on it, the exponential term within the likelihood can not be dropped simply any longer when deriving the upper bound, otherwise, the resulting integral will be always infinite. The upper bound on the exponential term with respect to the eigenvalues of covariance matrices is very tricky to be obtained, and the condition for the integrability of the resulting integral remains to be further studied. Thus, the GHNL modelling with unknown $\mathbf{\Sigma}_0$ can be taken as a sequential study of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research was supported by the National Natural Science Foundation of China [grant number 11671146].

References

Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Annals of Statistics*, *8*(4), 716–761. https://doi.org/10.1214/aos/1176345068

Berger, J., Strawderman, W., & Tang, D. (2005). Posterior propriety and admissibility of hyperpriors in normal hierarchical models. *Annals of Statistics*, *33*(2), 606–646. https://doi.org/10.1214/009053605000000075

Berger, J., Sun, D., & Song, C. (2020a). Bayesian analysis of the covariance matrix of a multivariate normal distribution with a new class of priors. *Annals of Statistics*, *48*(4), 2381–2403. https://doi.org/10.1214/19-AOS1891

Berger, J., Sun, D., & Song, C. (2020b). An objective prior for hyperparameters in normal hierarchical models. *Journal of Multivariate Analysis*, *178*(2020), 1–13. https://doi.org/10.1016/j.jmva.2020.104606

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, *13*(2), 627–679. https://doi.org/10.1214/18-BA1103

Daniels, M. J., & Kass, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association.*, *94*(448), 1254–1263. https://doi.org/10.1080/01621459.1999.10473878

Everson, P. J., & Morris, C. N. (2000). Inference for multivariate normal hierarchical models. *Journal of the Royal Statistical Society: Series B*, *62*(2), 399–412. https://doi.org/10.1111/rssb.2000.62.issue-2

Fourdrinier, D., Strawderman, W. E., & Wells, M. T. (1998). On the construction of Bayes minimax estimators. *Annals of Statistics*, *26*(2), 660–671. https://doi.org/10.1214/aos/1028144853

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). John Wiley & Sons.

Gustafson, P., Hossain, S., & Macnab, Y. C. (2006). Conservative prior distributions for variance parameters in hierarchical models. *Canadian Journal of Statistics*, *34*(3), 377–390. https://doi.org/10.1002/cjs.v34:3

Hobert, J. P., & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, *91*(436), 1461–1473. https://doi.org/10.1080/01621459.1996.10476714

Hoff, P. D. (2009b). Simulation of the matrix Bingham-Von Mises-Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, *18*(2), 438–456. https://doi.org/10.1198/jcgs.2009.07177

Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Lindenberger, U., & Pötter, U. (1998). The complex nature of unique and shared effects in hierarchical linear regression: Implications for developmental psychology. *Psychological Methods*, *3*(2), 218–230. https://doi.org/10.1037/1082-989X.3.2.218

Michalak, S. E., & Morris, C. N. (2016). Posterior propriety for hierarchical models with log-Likelihoods that have norm bounds. *Bayesian Analysis*, *11*(2), 545–571. https://doi.org/10.1214/15-BA962

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*(1), 1–17. https://doi.org/10.2307/2112482

Shimotsu, K. (2010). Exact local Whittle estimation of fractional integration with unknown mean and time trend. *Econometric Theory*, *26*(2), 501–540. https://doi.org/10.1017/S0266466609100075

Song, C., Sun, D., Fan, K., & Mu, R. (2020). Posterior propriety of an objective prior in a 4-Level normal hierarchical model. *Mathematical Problems in Engineering*, *2020*. https://doi.org/10.1155/2020/8236934

Speckman, P. L., & Sun, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, *90*(2), 289–302. https://doi.org/10.1093/biomet/90.2.289

Sun, D., Tsutakawa, R. K., & He, Z. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica*, *11*(1), 77–95. http://www.jstor.org/stable/24306811

Xia, A., Ma, H., & Carlin, B. P. (2011). Bayesian hierarchical modeling for detecting safety signals in clinical trials. *Journal of Biopharmaceutical Statistics*, *21*(5), 1006–1029. https://doi.org/10.1080/10543406.2010.520181

Yang, R., & Berger, J. (1994). Estimation of a covariance matrix using the reference prior. *Annals of Statistics*, *22*(3), 1195–1211. https://doi.org/10.1214/aos/1176325625

Appendices

Appendix 1. A special LMM

Consider a special LMM of the form:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \cdots + \boldsymbol{Z}_r\boldsymbol{u}_r + \boldsymbol{\epsilon}, \tag{A1}$$

where $\boldsymbol{y}$ denotes the observations and is an $n \times 1$ vector, $\boldsymbol{\beta}$ is the vector of fixed effects and is an $p \times 1$ vector. For $i \in [r]$, $\boldsymbol{u}_i$'s are $q_i \times 1$ vectors and represent the vectors of random effects, and $\boldsymbol{u}_i$'s are assumed to be independently distributed as $N_{q_i}(\boldsymbol{0}, \boldsymbol{W}_i)$, where $\boldsymbol{W}_i$'s are $q_i \times q_i$ positive definite matrices and unknown. $\boldsymbol{X}$ is an $n \times p$ matrix, $\boldsymbol{Z}_i$'s are $n \times q_i$ matrices, and $\boldsymbol{X}$ and $\boldsymbol{Z}_i$'s are known design matrices. $\boldsymbol{\epsilon}$ is the vector of random errors and distributed as $N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ is an $n \times n$ positive definite matrix and given.

It follows from Berger et al. (2020b) that we can assume independent priors on $(\boldsymbol{\beta}, \boldsymbol{W}_1, \ldots, \boldsymbol{W}_r)$ as

$$\begin{aligned} \pi(\boldsymbol{\beta}) &\propto \frac{1}{(1+\|\boldsymbol{\beta}\|^2)^{(p-1)/2}}, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \\ \pi(\boldsymbol{W}_j) &\propto \frac{1}{|\boldsymbol{W}_j|^{1-1/(2q_j)} \prod_{1\le s<t\le q_j}(\nu_{js}-\nu_{jt})}, \\ &\boldsymbol{W}_j > 0,\ j \in [r], \end{aligned} \tag{A2}$$

where $\nu_{j1} > \nu_{j2} > \cdots > \nu_{jq_j} > 0$ are the ordered eigenvalues of $\boldsymbol{W}_j$, $j \in [r]$. The prior on $\boldsymbol{\beta}$ has a hierarchical structure of the form

$$(\boldsymbol{\beta}|\tau) \sim N_d(\boldsymbol{0}, \tau\boldsymbol{I}_d) \quad \text{and} \quad [\tau] \propto \tau^{-1/2}\exp\left(-\frac{1}{2\tau}\right).$$

The posterior propriety results for the special LMM (A1) is displayed as follows. Firstly, let $\tau \triangleq \nu_{01}$ and $q_0 = 1$. Denote the index set of the variance scale or the eigenvalues of the covariance matrices by $F = \{(j,l) \mid j = 0, 1, \ldots, r,\ l \in [q_j]\}$, and $\mathcal{T} = \{D \mid D \subseteq F, D \neq \emptyset\}$ represents the set of the non-empty subsets of F. Define that $c_{jl,s} = 1_{\{l=s\}}$ for $j \in [r], l \in [q_j]$ and $s \in [n]$.

Theorem A.1: *Consider linear mixed effect model* (A1) *with prior* (A2) *on* $(\boldsymbol{\beta}, \boldsymbol{W}_1, \ldots, \boldsymbol{W}_r)$. *Assume* $p > 1$, *and then the posterior is proper if*

$$\sum_{s=1}^{n} 1_{\left\{1_{\{(0,1)\in D\}} 1_{\{s\le p\}} + \sum_{j\neq 0,(j,l)\in D} c_{jl,s} > 0\right\}} > \sum_{(j,l)\in D} \frac{1}{q_j} \tag{A3}$$

holds for any $D \in \mathcal{T}$.

The proof of Theorem A.1 is similar to that of Theorem 3.1 and is omitted here.

Fact A.1: When $p > 1$, (A3) holds for any $D \in \mathcal{T}$ iff

$$\sum_{j\in[r]} \frac{1}{q_j} < 1 \quad \text{and} \quad p > 1 + \sum_{j\in[r]} \frac{\min(p, q_j)}{q_j}. \tag{A4}$$

Proof: For any $D \in \mathcal{T}$ and $(0,1) \notin D$, (A3) is equivalent to

$$\sum_{s=1}^{n} 1_{\left\{\sum_{(j,l)\in D} c_{jl,s} > 0\right\}} > \sum_{(j,l)\in D} \frac{1}{q_j}. \tag{A5}$$

It can be deduced that inequality (A5) holds for any $D \in \mathcal{T}$ and $(0,1) \notin D$ iff

$$L > \sum_{j\in[r]} \frac{\min(L, q_j)}{q_j}, \quad \text{for } L \in [n], \tag{A6}$$

which is equivalent to $\sum_{j\in[r]} \frac{1}{q_j} < 1$ since $q_j \ge 1$ for $j \in [r]$.

Inequality (A3) holds for any $D \in \mathcal{T}$ with $(0,1) \in D$ iff

$$L > 1 + \sum_{j\in[r]} \frac{\min(L, q_j)}{q_j}, \quad \text{for } L = p, \ldots, n. \tag{A7}$$

Under the condition that $\sum_{j\in[r]} \frac{1}{q_j} < 1$, (A7) is equivalent to

$$p > 1 + \sum_{j\in[r]} \frac{\min(p, q_j)}{q_j}.$$

■

Corollary A.1: *Consider model* (A1) *with prior* (A2) *on parameters. The posterior is proper if one of the following condition holds,*

(a) $p > 1 + r$ *and* $\sum_{j\in[r]} \frac{1}{q_j} < 1$;
(b) $p > 1$ *and* $\sum_{j\in[r]} \frac{1}{q_j} < 1 - \frac{1}{p}$.

Proof: Since

$$\sum_{j\in[r]} \frac{\min(p, q_j)}{q_j} \le r \quad \text{and} \quad \sum_{j\in[r]} \frac{\min(p, q_j)}{q_j} \le p \sum_{j\in[r]} \frac{1}{q_j},$$

(a) and (b) follow from Fact A.1 directly. ■

Remark A.1: Consider model (A1) with $r = 1$. The posterior using prior (A2) is prior if either (a) $p \ge 2$, $q_1 \ge 3$ or (b) $p \ge 3$, $q_1 \ge 2$ holds. If $p = 1$, the posterior propriety can hardly be satisfied, the reason of which is two-fold. First, inequality (A3) fails for $D = \{(0,1)\}$. Second, if we follow the thread of deriving the condition in Theorem 3.3, a sufficient condition can be derived as

$$\sum_{s=1}^{n} 1_{\left\{\sum_{(j,l)\in D} c_{jl,s} > 0\right\}} > 1_{\{\exists j\in[r],\ (j,1)\in D\}} + \sum_{(j,l)\in D} \frac{1}{q_j}, \tag{A8}$$

for any $D \in \mathcal{T}$ with $(0,1) \notin D$. However, inequality (A8) does not hold for $D = \{(j,1)\}, j \in [r]$.

Appendix 2. Proof of lemmas in Section 3.1

Lemma A.1: ***min–max theorem, cf. Horn & Johnson, (2012):*** *For an $n \times n$ symmetric matrix $\boldsymbol{A}$ and a non-zero $n \times 1$ vector $\boldsymbol{x}$, the Rayleigh quotient for $\boldsymbol{A}$ and $\boldsymbol{x}$ can be defined as*

$$R(\boldsymbol{A}, \boldsymbol{x}) = \frac{\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle},$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. Then

$$\lambda_k(\boldsymbol{A}) = \max_U \left\{ \min_{\boldsymbol{x}} \{R(\boldsymbol{A}, \boldsymbol{x}) \mid \boldsymbol{x} \in U, \|\boldsymbol{x}\| \neq 0\} \;\middle|\; \dim(U) = k \right\}, \quad k \in [n] \tag{A9}$$

where U denotes the linear subspace of $\mathbb{R}^n$. Especially,

$$\lambda_{\max}(\boldsymbol{A}) = \max_{\boldsymbol{x}} \{R(\boldsymbol{A}, \boldsymbol{x}) \mid \|\boldsymbol{x}\| \neq 0\} \quad \text{and}$$
$$\lambda_{\min} = \min_{\boldsymbol{x}}(\boldsymbol{A}) \{R(\boldsymbol{A}, \boldsymbol{x}) \mid \|\boldsymbol{x}\| \neq 0\}. \tag{A10}$$

Lemma A.2: *For $n \times n$ symmetric matrices $\boldsymbol{A}_j$, $j \in [r]$, we have*

(a) $\lambda_{\max}\left(\sum_{j=1}^r \boldsymbol{A}_j\right) \leq \sum_{j=1}^r \lambda_{\max}\left(\boldsymbol{A}_j\right)$;
(b) *supposing $\boldsymbol{A}_j \geq 0$, $j \in [r]$, then*

$$\lambda_k\left(\sum_{j=1}^r \boldsymbol{A}_j\right) \geq \frac{1}{r}\sum_{j=1}^r \lambda_k\left(\boldsymbol{A}_j\right), \quad k \in [n].$$

Proof: For (a), given an $n \times 1$ non-zero vector $\boldsymbol{x}$, by (A10),

$$R\left(\sum_{j=1}^r \boldsymbol{A}_j, \boldsymbol{x}\right) = \sum_{j=1}^r R(\boldsymbol{A}_j, \boldsymbol{x}) \leq \sum_{j=1}^r \max_{\boldsymbol{x}_j \neq \boldsymbol{0}} R(\boldsymbol{A}_j, \boldsymbol{x}_j) = \sum_{j=1}^r \lambda_{\max}\left(\boldsymbol{A}_j\right). \tag{A11}$$

The proof for (a) is completed by using (A10) again.

For (b), it suffices to prove that for any $j \in [r]$

$$\lambda_k\left(\sum_{j=1}^r \boldsymbol{A}_j\right) \geq \lambda_k\left(\boldsymbol{A}_j\right), \quad i \in [n]. \tag{A12}$$

Since $\boldsymbol{A}_l \geq 0$, $l \in [r]$, for any $j \in [r]$, $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{x} \neq \boldsymbol{0}$,

$$R\left(\sum_{l=1}^r \boldsymbol{A}_l, \boldsymbol{x}\right) = \sum_{l=1}^r R(\boldsymbol{A}_l, \boldsymbol{x}) \geq R(\boldsymbol{A}_j, \boldsymbol{x}).$$

Minimize the both sides of the inequality above over $\{\boldsymbol{x} \mid \boldsymbol{x} \in U, \boldsymbol{x} \neq \boldsymbol{0}\}$ first. Then take the maximum over $\{U \mid U \subseteq \mathbb{R}^n, \dim(U) = k\}$. (A12) can be easily obtained by using Lemma A.1. ■

A.1 Proof of Lemma 3.1

For (a), by Lemma A.2 (a), it suffices to prove that $\lambda_{\max}(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top) \leq C_1\lambda_{\max}(\boldsymbol{A}_j)$, $j \in [r]$. For any $j \in [r]$, applying (A10), yields

$$0 < \lambda_{\max}(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top) = R\left(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top, \boldsymbol{x}^*\right) = \max_{\boldsymbol{x} \neq \boldsymbol{0}} R\left(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top, \boldsymbol{x}\right).$$

It is obvious that $\boldsymbol{X}_j^\top \boldsymbol{x}^* \neq \boldsymbol{0}$, otherwise, it will result in $R\left(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top, \boldsymbol{x}^*\right) = 0$, which contradicts. In addition, since $\left\langle \boldsymbol{X}_j^\top \boldsymbol{x}^*, \boldsymbol{X}_j^\top \boldsymbol{x}^* \right\rangle \leq \lambda_{\max}\left(\boldsymbol{X}_j^\top \boldsymbol{X}_j\right) \langle \boldsymbol{x}^*, \boldsymbol{x}^* \rangle \leq C_1 \langle \boldsymbol{x}^*, \boldsymbol{x}^* \rangle$, we have

$$\begin{aligned} R\left(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top, \boldsymbol{x}^*\right) &\leq C_1 R\left(\boldsymbol{A}_j, \boldsymbol{X}_j^\top \boldsymbol{x}^*\right) \\ &\leq C_1 \max\left\{R\left(\boldsymbol{A}_j, \boldsymbol{z}\right) \;\middle|\; \boldsymbol{z} \in \mathbb{R}^{p_j}, \boldsymbol{z} \neq \boldsymbol{0}\right\} \\ &= C_1\lambda_{\max}(\boldsymbol{A}_j). \end{aligned}$$

Therefore, we have proved part (a).

For (b), we only need to prove that

$$\lambda_k(\boldsymbol{H}) \geq \frac{C_2}{r}\sum_{j=1}^r a_{jk}, \quad k \in [n].$$

It follows from Lemma A.2 (b) that for any $k \in [n]$

$$\lambda_k(\boldsymbol{H}) \geq \frac{1}{r}\sum_{j=1}^r \lambda_k(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top).$$

Since $\operatorname{rank}(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top) = \operatorname{rank}(\boldsymbol{A}_j)$, then $\lambda_k(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top) = 0, p_j < k \leq n$. Thus, it remains to show that

$$\lambda_k(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top) \geq C_2\lambda_k(\boldsymbol{A}_j), \quad j \in [r], \; k \in [p_j]. \tag{A13}$$

Firstly, for any $j \in [r]$, we introduce a linear transformation $\mathcal{L}_j : \mathbb{R}^n \to \mathbb{R}^{p_j}$ defined as $\mathcal{L}_j(\boldsymbol{v}) = \boldsymbol{X}_j^\top \boldsymbol{v}$, $\boldsymbol{v} \in \mathbb{R}^n$. The *kernel* space of $\mathcal{L}_j$ is denoted by $\operatorname{Ker}(\mathcal{L}_j) = \left\{\boldsymbol{v} \in \mathbb{R}^n : \mathcal{L}_j(\boldsymbol{v}) = \boldsymbol{0}\right\}$. Since $\boldsymbol{X}_j$ is of full column rank $p_j \leq n$, then the dimension of the complementary space of $\operatorname{Ker}(\mathcal{L}_j)$ is p_j, i.e., $\dim\left(\operatorname{Ker}(\mathcal{L}_j)^\perp\right) = p_j$. Thus, the mapping $\mathcal{L}_j : \operatorname{Ker}(\mathcal{L}_j)^\perp \mapsto \mathbb{R}^{p_j}$ is a one-to-one mapping. For any $U \subseteq \mathbb{R}^{p_j}$ with $\dim(U) = k$, $k \in [p_j]$, define that

$$\mathcal{L}_j^*(U) = \left\{\boldsymbol{v} \in \operatorname{Ker}(\mathcal{L}_j)^\perp : \mathcal{L}_j(\boldsymbol{v}) = \boldsymbol{x}, \quad \boldsymbol{x} \in U\right\}.$$

It is obvious that $\mathcal{L}_j^*(U) \subseteq \operatorname{Ker}(\mathcal{L}_j)^\perp$ and $\dim\left(\mathcal{L}_j^*(U)\right) = k$. For any $U \subseteq \mathbb{R}^{p_j}$ with $\dim(U) = k$ and any $\boldsymbol{x} \in U$, there exists one and only one $\boldsymbol{v} \in \mathcal{L}_j^*(U)$ such that $\mathcal{L}_j(v) = \boldsymbol{x}$. Since $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = \boldsymbol{v}^\top(\boldsymbol{X}_j^\top \boldsymbol{X}_j)\boldsymbol{v} \geq C_2 \langle \boldsymbol{v}, \boldsymbol{v} \rangle$, we have

$$R\left(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top, \boldsymbol{v}\right) \geq C_2 R\left(\boldsymbol{A}_j, \boldsymbol{x}\right). \tag{A14}$$

It refers to Lemma A.1 that

$$\lambda_k(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top) = \max_V \left\{\min_{\boldsymbol{v}} \left\{R(\boldsymbol{X}_j\boldsymbol{A}_j\boldsymbol{X}_j^\top, \boldsymbol{v}) \mid \boldsymbol{v} \in V, \|\boldsymbol{v}\| \neq 0\right\} \;\middle|\; V \subseteq \operatorname{Ker}(\mathcal{L}_j)^\perp, \dim(V) = k\right\}, \tag{A15}$$

for $k \in [p_j]$. Minimize the both sides of inequality (A14) over $\{\boldsymbol{x} \mid \boldsymbol{x} \in U, \boldsymbol{x} \neq \boldsymbol{0}\}$ first. Then take the maximum over $\left\{U \mid U \subseteq \mathbb{R}^{p_j}, \dim(U) = k\right\}$, (A13) can be easily obtained by using Lemma A.1 and (A15).

A.2 Proof of Lemma 3.2

The Domain of the integral can be divided into

$$\begin{aligned} &\Omega_0 = \{\boldsymbol{\lambda} \mid 0 \leq \lambda_j \leq 1, \; j \in [k]\}, \\ &\Omega_D = \{\boldsymbol{\lambda} \mid \lambda_j > 1, \; j \in D \text{ and } 0 \leq \lambda_i \leq 1, \; i \in [k]/D\}, \\ &\quad D \in \mathcal{F}([k]) \end{aligned}$$

i.e., $\Omega = \left(\bigcup_{D \in \mathcal{F}([k])} \Omega_D\right) \bigcup \Omega_0$. Thus, the integral I is finite iff the integrals over Ω_0 and Ω_D for each $D \in \mathcal{F}([k])$ are finite.

Denote the integrand as $F(\boldsymbol{\lambda})$. Then

$$\int_{\Omega_0} F(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda} \simeq \int_{\Omega_0} \frac{1}{\prod_{j=1}^{k} \lambda_j^{a_j}}\,\mathrm{d}\boldsymbol{\lambda},$$

which is finite iff condition (a) is satisfied.

To verify condition (b), we only need to justify the following statement. Also, we assume condition (a) is always satisfied hereafter.

Fact A.2: For all $D \in \mathcal{F}([k])$ with $\mathrm{card}(D) = L$, $1 \le L \le k$, the integrals $\int_{\Omega_D} F(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda}$ are finite iff inequalities

$$\sum_{j\in E} a_j + \sum_{i\in \mathcal{G}_E} b_i > \mathrm{card}(E) \tag{A16}$$

hold for all $E \in \mathcal{F}([k])$ with $\mathrm{card}(E) \le L$ and $\mathcal{G}_E = \{i \mid E \bigcap C_i \neq \emptyset, i \in [n]\}$. Under the condition above,

$$\int_{\Theta_D(t)} G_D(\boldsymbol{\lambda}) \left(\prod_{r\in D} \mathrm{d}\lambda_r\right) \simeq \exp\left\{-\log t \left(\sum_{j\in D} a_j + \sum_{i\in \mathcal{G}_D} b_i - \mathrm{card}(D)\right)\right\} \tag{A17}$$

always holds, where

$$\Theta_D(t) = \left\{\boldsymbol{\lambda} \mid \lambda_j \ge t,\ j \in D \quad \text{and} \quad 0 \le \lambda_i \le 1,\ i \in [k]/D\right\},$$

$$G_D(\boldsymbol{\lambda}) = \frac{1}{\left[\prod_{j\in D} \lambda_j^{a_j}\right]\left[\prod_{i\in \mathcal{G}_D}(1 + \sum_{r\in C_i} \lambda_r)^{b_i}\right]}.$$

The reason why formula (A17) is required is that it plays an important role in verifying condition (A16).

Proof: We prove the result by the technique of mathematical induction. First, we assume that the statement in Fact A.2 is true for $L = l$, $1 \le l \le (k-1)$. With this assumption, we must show that the statement is true for its successor, $L = (l+1)$. Write an arbitrary set $D \in \mathcal{F}([k])$ with cardinality $(l+1)$ as $\{j_1, \ldots, j_{l+1}\}$, where $1 \le j_1 < \cdots < j_{l+1} \le k$. Denote that $D_{-j_i} = D/\left[\{j_i\}\right]$, $i = 1, \ldots, (l+1)$.

Step 1: We first prove that $\int_{\Omega_D} F(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda}$ is finite iff inequalities (A16) hold for $L = (l+1)$.

Region Ω_D can be divided into

$$\Sigma_1 = \{\boldsymbol{\lambda} | \lambda_j \ge \lambda_{j_1} > 1, j \in D_{-j_1} \text{ and } 0 \le \lambda_i \le 1, i \in [k]/D\}$$

$$\vdots$$

$$\Sigma_{l+1} = \{\boldsymbol{\lambda} | \lambda_j \ge \lambda_{j_{l+1}} > 1, j \in D_{-j_{l+1}} \text{ and } 0 \le \lambda_i \le 1, i \in [k]/D\}.$$

Therefore, integral $\int_{\Omega_D} F(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda}$ is finite iff $\int_{\Sigma_i} F(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda} < \infty$ for any $i = 1, \ldots, (l+1)$. For $i = 1, \ldots, (l+1)$, we have

$$\int_{\Sigma_i} F(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda} \simeq \int_{\Sigma_i} \frac{1}{\prod_{s\notin D} \lambda_s^{a_s}} \left(\frac{1}{\lambda_{j_i}}\right)^{a_{j_i} + \sum_{r\in \mathcal{H}_i} b_r} G_{D_{-j_i}}(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda},$$

where $\mathcal{H}_i = \{r \mid C_r \bigcap D_{j_i} = \emptyset \text{ and } j_i \in C_r, r \in [n]\}$, and it's easy to see that $\mathcal{G}_D = \mathcal{G}_{D_{-j_i}} \bigcup \mathcal{H}_i$ and $\mathcal{G}_{D_{-j_i}} \bigcap \mathcal{H}_i = \emptyset$. Since

$$\int_{\Sigma_i} G_{D_{-j_i}}(\boldsymbol{\lambda}) \left(\prod_{r\in D_{-j_i}} \mathrm{d}\lambda_r\right) \simeq \int_{\Theta_{D_{-j_i}}(\lambda_{j_i})} G_{D_{-j_i}}(\boldsymbol{\lambda}) \left(\prod_{r\in D_{-j_i}} \mathrm{d}\lambda_r\right), \tag{A18}$$

$$\int_{\Omega_{D_{-j_i}}} F(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda} \simeq \int_{\Omega_{D_{-j_i}}} \frac{1}{\prod_{s\notin D_{-j_i}} \lambda_s^{a_s}} G_{D_{-j_i}}(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda} \tag{A19}$$

and the RHS (Right-Hand Side) of (A19) and (A18) are finite simultaneously under condition (a). Hence, The LHS (Left-Hand Side) of (A18) is finite iff $\int_{\Omega_{D_{-j_i}}} F(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda}$ is finite.

Furthermore, by assumption and (A17), we have

$$\int_{\Sigma_i} G_{D_{-j_i}}(\boldsymbol{\lambda}) \left(\prod_{r\in D_{-j_i}} \mathrm{d}\lambda_r\right) \simeq \exp\left\{-\log \lambda_{j_i} \left(\sum_{j\in D_{-j_i}} a_j + \sum_{r\in \mathcal{G}_{D_{-j_i}}} b_r - l\right)\right\}.$$

Thus, under condition (a) and assumption, we have

$$\int_{\Sigma_i} F(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda} \simeq \int_1^{\infty} \left(\frac{1}{\lambda_{j_i}}\right)^{\sum_{j\in D} a_j + \sum_{r\in \mathcal{G}_D} b_r - l} \mathrm{d}\lambda_{j_i},$$

the RHS of which is finite iff $\sum_{j\in D} a_j + \sum_{r\in \mathcal{G}_D} b_r > 1 + l = \mathrm{card}(D)$.

In conclusion, $\int_{\Omega_D} F(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda}$ is finite iff $\sum_{j\in D} a_j + \sum_{r\in \mathcal{G}_D} b_r > \mathrm{card}(D)$ and $\int_{\Omega_{D_{-j_i}}} F(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda}$ is finite for any $i \in [l+1]$. Since D is arbitrary and $\mathrm{card}(D_{-j_i}) = l$, we have accomplished the goal of Step 1.

Step 2: Next, we prove that formula (A17) holds for D with cardinality $(l+1)$.

Region $\Theta_D(t)$ can be divided into

$$\Theta_D^{(1)}(t) = \{\boldsymbol{\lambda} | \lambda_j \ge \lambda_{j_1} \ge t, j \in D_{-j_1} \text{ and } 0 \le \lambda_i \le 1, i \in [k]/D\}$$

$$\vdots$$

$$\Theta_D^{(l+1)}(t) = \{\boldsymbol{\lambda} | \lambda_j \ge \lambda_{j_{l+1}} \ge t, j \in D_{-j_{l+1}} \text{ and } 0 \le \lambda_i \le 1, i \in [k]/D\}.$$

Similar to the proof of Step 1, we can prove that for $i = 1, \ldots, (l+1)$,

$$\int_{\Theta_D^{(i)}(t)} G_D(\boldsymbol{\lambda}) \left(\prod_{r\in D} \mathrm{d}\lambda_r\right) \simeq \int_t^{\infty} \left(\frac{1}{\lambda_{j_i}}\right)^{\sum_{j\in D} a_j + \sum_{r\in \mathcal{G}_D} b_r - l} \mathrm{d}\lambda_{j_i} = \exp\left\{-\log t \left(\sum_{j\in D} a_j + \sum_{r\in \mathcal{G}_D} b_r - \mathrm{card}(D)\right)\right\}.$$

Therefore, we get Step 2 proved.

Step 3: We need to present that the statement is true for $L = 1$ to complete the proof, on the basis of mathematical induction.

Denote that $D = \{r\}$, $r = 1, \ldots, k$. Then

$$\int_{\Omega_D} F(\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda} \simeq \int_1^{\infty} \left(\frac{1}{\lambda_r}\right)^{a_r + \sum_{i\in \mathcal{G}_D} b_i} \mathrm{d}\lambda_r,$$

which is finite iff $\sum_{j\in D} a_j + \sum_{i\in\mathcal{G}_D} b_i > 1 = \mathrm{card}(D)$, under which,

$$\int_{\Theta_D(t)} G_D(\boldsymbol{\lambda}) \left(\prod_{i\in D} \mathrm{d}\lambda_i\right)$$
$$\simeq \int_t^\infty \left(\frac{1}{\lambda_r}\right)^{a_r+\sum_{i\in\mathcal{G}_D} b_i} \mathrm{d}\lambda_r$$
$$= \exp\left\{-\log t\left(\sum_{j\in D} a_j + \sum_{i\in\mathcal{G}_D} b_i - \mathrm{card}(D)\right)\right\},$$

which accomplishes the proof of Fact A.2. ■

Appendix 3. Gibbs sampling from the SIW distributions

As for the SIW distribution (46), we first consider the change of variables from $\boldsymbol{W}$ to $\boldsymbol{\Xi} = \mathrm{diag}\,(\nu_1, \ldots, \nu_k)$ and the orthogonal matrix $\boldsymbol{O}$ of corresponding eigenvectors. The Jacobian is

$$\left|\frac{\partial\, \boldsymbol{W}}{\partial\,(\boldsymbol{\Xi}, \boldsymbol{O})}\right| = \prod_{i<j} (\nu_i - \nu_j)\,. \tag{A20}$$

According to (A20) and Lemma 4 in Berger et al. (2020a), (46) can be transformed to

$$\pi(\boldsymbol{\Xi}, \boldsymbol{O}) \propto \frac{1}{|\boldsymbol{\Xi}|^a}\,\mathrm{etr}\left(-\frac{1}{2}\boldsymbol{\Xi}^{-1}\boldsymbol{O}'\boldsymbol{H}\boldsymbol{O}\right). \tag{A21}$$

Gibbs sampling of $\boldsymbol{\Xi}$: We first sample $\boldsymbol{\Xi}$ given $(\boldsymbol{O}, \boldsymbol{H})$ from

$$\pi(\boldsymbol{\Xi} \mid \boldsymbol{O}, \boldsymbol{H}) \propto \frac{1}{\prod_{i=1}^k \nu_i^a}\,\mathrm{etr}\left(-\frac{1}{2}\boldsymbol{\Xi}^{-1}\boldsymbol{O}'\boldsymbol{H}\boldsymbol{O}\right)$$
$$= \prod_{i=1}^k \frac{1}{\nu_i^a}\,\mathrm{etr}\left(-\frac{c_i}{\nu_i}\right),$$

where c_i is the i-th diagonal element of $\boldsymbol{O}'\boldsymbol{H}\boldsymbol{O}/2$, $i \in [k]$. Therefore, we can sample ν_i independently from $IG\,(a-1, c_i)$.

Gibbs sampling of $\boldsymbol{O}$: Given $(\boldsymbol{\Xi}, \boldsymbol{H})$, the marginal density of $\boldsymbol{O}$ has the form:

$$\pi(\boldsymbol{O} \mid \boldsymbol{\Xi}, \boldsymbol{H}) \propto \mathrm{etr}\left(-\frac{1}{2}\boldsymbol{H}\boldsymbol{O}\boldsymbol{\Xi}^{-1}\boldsymbol{O}'\right).$$

Let $\boldsymbol{H} = \boldsymbol{L}\boldsymbol{U}\boldsymbol{L}^\top$, where $\boldsymbol{L}\boldsymbol{L}^\top = \boldsymbol{I}_k$ and $\boldsymbol{U} = \mathrm{diag}\,(u_1, \ldots, u_k)$ is the diagonal matrix of corresponding eigenvalues with $u_1 \geq \cdots \geq u_k$. Define $\boldsymbol{G} = \boldsymbol{L}^\top\boldsymbol{O}$. Since the invariant right Haar measure is invariant to the orthonormal transformation, the conditional density of $\boldsymbol{G}$ is

$$\pi(\boldsymbol{G} \mid \boldsymbol{\Xi}, \boldsymbol{H}) \propto \mathrm{etr}\left(-\frac{1}{2}\boldsymbol{U}\boldsymbol{G}\boldsymbol{\Xi}^{-1}\boldsymbol{G}'\right). \tag{A22}$$

The updating of $\boldsymbol{G}$ from (A22) can be implemented by applying a Gibbs update to two randomly selected columns (cf. Hoff, 2009b) or rows (cf. Berger et al., 2020a). The two ways are essentially equivalent when $\mathrm{rank}(\boldsymbol{H}) = k$, but Berger et al. (2020a)'s method is considerably faster if $\mathrm{rank}(\boldsymbol{H}) < k$. Without any loss, assume that the two randomly selected rows are the first and second rows. The updated value of $\boldsymbol{G}$ can be written as $\boldsymbol{G}^{\mathrm{new}} = \mathrm{diag}\left(\boldsymbol{\Phi}, \boldsymbol{I}_{k-2}\right)\begin{pmatrix}\boldsymbol{G}_{12}^{\mathrm{old}}\\ \boldsymbol{G}_{-12}^{\mathrm{old}}\end{pmatrix}$, where $\boldsymbol{G}_{12}^{\mathrm{old}}$ denotes the first two rows of the old value of $\boldsymbol{G}$ which is $\boldsymbol{G}^{\mathrm{old}}$, $\boldsymbol{G}_{-12}^{\mathrm{old}}$ is the remaining $k-2$ rows of $\boldsymbol{G}^{\mathrm{old}}$ and

$$\boldsymbol{\Phi} = \boldsymbol{D}_\epsilon\boldsymbol{\Phi}_0 = \begin{pmatrix}\epsilon_1 & 0\\ 0 & \epsilon_2\end{pmatrix}\begin{pmatrix}\cos\phi & -\sin\phi\\ \sin\phi & \cos\phi\end{pmatrix},$$

with $\phi \in (-\frac{\pi}{2}, \frac{\pi}{2}]$ and $\epsilon_i = \pm 1$ for $i = 1, 2$. Let $\boldsymbol{U}_1 = \mathrm{diag}(u_1, u_2)$. The full conditional density of ϕ has the form:

$$\pi(\phi \mid \boldsymbol{G}^{\mathrm{old}}, \boldsymbol{\Xi}, \boldsymbol{H}) \propto \mathrm{etr}\left\{-\frac{1}{2}\boldsymbol{U}_1\boldsymbol{\Phi}_0\boldsymbol{G}_{12}^{\mathrm{old}}\boldsymbol{\Xi}^{-1}\left(\boldsymbol{G}_{12}^{\mathrm{old}}\right)^\top\boldsymbol{\Phi}_0^\top\right\}.$$

Write

$$\boldsymbol{G}_{12}^{\mathrm{old}}\boldsymbol{\Xi}^{-1}\left(\boldsymbol{G}_{12}^{\mathrm{old}}\right)^\top = \begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}s_1 & 0\\ 0 & s_2\end{pmatrix}$$
$$\times\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}$$

where $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2}]$ and $s_1 > s_2$. Then the conditional density of ϕ can be rewritten as

$$\pi(\phi \mid \boldsymbol{G}^{\mathrm{old}}, \boldsymbol{\Xi}, \boldsymbol{H}) \propto \exp\left\{-c_0\cos^2(\phi+\theta)\right\},$$

where $c_0 = \frac{1}{2}(s_1 - s_2)(u_1 - u_2) \geq 0$. Define $\alpha = \cos^2(\phi + \theta)$. Then the full conditional density of α has the form:

$$\pi(\alpha \mid \boldsymbol{G}^{\mathrm{old}}, \boldsymbol{\Xi}, \boldsymbol{H}) \propto \exp\{-c_0\alpha\}\,\alpha^{-\frac{1}{2}}(1-\alpha)^{-\frac{1}{2}},$$
$$\alpha \in [0,\ 1].$$

Simulating $\alpha \in [0, 1]$ can proceed with a rejection sampler by setting the proposal distribution as Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$.