The Annals of Statistics
2017, Vol. 45, No. 1, 196–222
DOI: 10.1214/16-AOS1449

A BAYESIAN APPROACH FOR ENVELOPE MODELS

BY KSHITIJ KHARE[1], SUBHADIP PAL AND ZHIHUA SU[2]

University of Florida

The envelope model is a new paradigm to address estimation and prediction in multivariate analysis. Using sufficient dimension reduction techniques, it has the potential to achieve substantial efficiency gains compared to standard models. This model was first introduced by [Statist. Sinica **20** (2010) 927–960] for multivariate linear regression, and has since been adapted to many other contexts. However, a Bayesian approach for analyzing envelope models has not yet been investigated in the literature. In this paper, we develop a comprehensive Bayesian framework for estimation and model selection in envelope models in the context of multivariate linear regression. Our framework has the following attractive features. First, we use the matrix Bingham distribution to construct a prior on the orthogonal basis matrix of the envelope subspace. This prior respects the manifold structure of the envelope model, and can directly incorporate prior information about the envelope subspace through the specification of hyperparamaters. This feature has potential applications in the broader Bayesian sufficient dimension reduction area. Second, sampling from the resulting posterior distribution can be achieved by using a block Gibbs sampler with standard associated conditionals. This in turn facilitates computationally efficient estimation and model selection. Third, unlike the current frequentist approach, our approach can accommodate situations where the sample size is smaller than the number of responses. Lastly, the Bayesian approach inherently offers comprehensive uncertainty characterization through the posterior distribution. We illustrate the utility of our approach on simulated and real datasets.

1. Introduction.

1.1. *Background.* Consider the standard multivariate linear regression model, given by

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^r$ is the multivariate response vector, $\mathbf{X} \in \mathbb{R}^p$ is the vector of nonstochastic predictors, $\boldsymbol{\mu} \in \mathbb{R}^r$ and $\boldsymbol{\beta} \in \mathbb{R}^{r \times p}$ are unknown intercept and regression

Received July 2014; revised January 2016.
[1]Supported by NSF Grant DMS-11-06084.
[2]Supported by NSF Grant DMS-14-07460.
MSC2010 subject classifications. Primary 62F15, 62F30; secondary 60J20, 62H12.
Key words and phrases. Sufficient dimension reduction, envelope model, Gibbs sampling, Stiefel manifold, matrix Bingham distribution.

coefficients, and the errors $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ are assumed to have a multivariate normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}$. Our interest lies in the estimation of $\boldsymbol{\beta}$, as it depicts the relationship between the responses and predictors. The standard multivariate regression estimator for $\boldsymbol{\beta}$ does not take the stochastic relationships among the response variables into account. Exploiting these stochastic relationships can lead to improved efficiency in estimation of $\boldsymbol{\beta}$.

A substantial step forward along these lines is the envelope model introduced in [5]. An envelope estimator of $\boldsymbol{\beta}$ is more efficient than the standard estimator when there is immaterial information contained in $\mathbf{Y}$; in other words, the distribution of some elements in $\mathbf{Y}$ or some linear combinations of the elements in $\mathbf{Y}$ is invariant to the changes in $\mathbf{X}$. To put this more concretely, let $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$ be two matrices such that $[\boldsymbol{\Gamma} \ \ \boldsymbol{\Gamma}_0] \in \mathbb{R}^{r\times r}$ is an orthogonal matrix. We assume that (a) $\boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X} \sim \boldsymbol{\Gamma}_0^T\mathbf{Y}$, where $\sim$ means identically distributed and (b) $\boldsymbol{\Gamma}^T\mathbf{Y}$ is uncorrelated with $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ given $\mathbf{X}$. Hence, the matrix $[\boldsymbol{\Gamma} \ \ \boldsymbol{\Gamma}_0]$ divides $\mathbf{Y}$ into two parts, $\boldsymbol{\Gamma}^T\mathbf{Y}$ (the material part) and $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ (the immaterial part). Conditions (a) and (b) imply that $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ is uncorrelated with both $\boldsymbol{\Gamma}^T\mathbf{Y}$ and $\mathbf{X}$, therefore, it does not carry any information about $\boldsymbol{\beta}$ and it is immaterial to the regression.

Cook et al. [5] showed that (a) and (b) are equivalent to the following two conditions: (a′) $\mathcal{B} \subseteq \operatorname{span}(\boldsymbol{\Gamma})$, and (b′) $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 = \mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{P}_{\boldsymbol{\Gamma}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{Q}_{\boldsymbol{\Gamma}}$. Here, $\mathcal{B} = \operatorname{span}(\boldsymbol{\beta})$, $\mathbf{P}_{(\cdot)}$ denotes the linear operator that projects onto the subspace indicated by its argument and $\mathbf{Q}_{(\cdot)} = \mathbf{I} - \mathbf{P}_{(\cdot)}$. By [3], if $\boldsymbol{\Gamma}$ satisfies (b), $\operatorname{span}(\boldsymbol{\Gamma})$ is called a reducing subspace of $\boldsymbol{\Sigma}$. The $\boldsymbol{\Sigma}$-envelope of $\mathcal{B}$, denoted by $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, is defined as the smallest reducing subspace of $\boldsymbol{\Sigma}$ that contains $\mathcal{B}$ (see [5]). Consequently, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ can be used to decompose $\boldsymbol{\Sigma}$ into variation from the material and immaterial parts of $\mathbf{Y}$: $\boldsymbol{\Sigma}_1 = \operatorname{Var}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y})$ and $\boldsymbol{\Sigma}_2 = \operatorname{Var}(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{Y})$. We call (1) an *envelope model* when (a′) and (b′) are imposed. As $\boldsymbol{\beta}$ is related only with the material variation, this decomposition of $\boldsymbol{\Sigma}$ suggests that we can achieve efficiency gains by accounting for all the immaterial information when estimating $\boldsymbol{\beta}$. Particularly, massive efficiency gains can be obtained when $\|\boldsymbol{\Sigma}_2\| \gg \|\boldsymbol{\Sigma}_1\|$, where $\|\cdot\|$ denotes the spectral norm.

The coordinate form of an envelope model can be written as

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \tag{2}$$

where the coefficients $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ is an orthogonal basis of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and u denotes the dimension of the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. The positive definite matrices $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ carry the coordinates of $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$, and $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$ carries the coordinates of $\boldsymbol{\beta}$ with respect to $\boldsymbol{\Gamma}$. When $u = r$, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathbb{R}^r$, the envelope model degenerates to the standard model and no efficiency gains are offered.

Recently, more developments ([6, 15–17]) have taken place in the context of envelope models. In particular, some restrictions on the data structure have been removed, and new models have been formulated to achieve efficiency gains beyond

those offered by the original envelope model. Additionally, an improvement of the likelihood ratio test for the envelope model has been proposed in [14], and a connection between envelope models and partial least squares (PLS) has also been established in [4].

In this paper, we develop a Bayesian approach for analyzing envelope models in the context of multivariate linear regression. Although Bayesian methods have been proposed for the related field of sufficient dimension reduction (SDR) (see [10, 11, 18]), such an investigation has not been undertaken for envelope models. The attractive features of our approach are as follows. First, we introduce a class of priors that can incorporate prior information on subspaces. To be more specific, we put a joint matrix Bingham prior on the orthogonal basis matrix $\mathbf{\Gamma}$ (see Section 3). This method respects the fact that the elements of $\mathbf{\Gamma}$ jointly lie on an appropriate Stiefel manifold. Hence, prior information about the envelope subspace can be incorporated through specification of hyperparameters; see, for example, Remark 2. When prior information about envelope subspace is not available, this class of priors is also flexible to include objective priors (such as the uniform flat prior for all parameters). Second, a block Gibbs sampling algorithm with standard conditionals is derived to sample from the resulting posterior distribution. The ability to generate samples from the posterior distribution allows us compute posterior expectations and quantiles efficiently, and also perform model selection (selection of u) using the Deviance Information Criterion (DIC). One of the conditionals in the Gibbs sampler is a new generalized version of the matrix Bingham distribution, and we derive an efficient rejection sampler to sample from this distribution. Third, while the frequentist approach [5] cannot handle small sample size situations, especially when $n < r$, our Bayesian approach is stable and achieves efficiency gains under those situations; see Theorem 1 and the simulation study in Section 4.2. Lastly, the Bayesian approach offers a comprehensive uncertainty characterization through the posterior distribution. With finite sample size, while the frequentist envelope model would use bootstrap to get the standard error of $\hat{\boldsymbol{\beta}}$, the Bayesian envelope model addresses estimation uncertainty directly by construction of posterior credible intervals. Our prior construction in the context of envelope models also has useful and important implications for Bayesian SDR. Currently, all the existing Bayesian SDR models are based on putting an independent normal prior on each element of the basis of the central subspace (see, e.g., [10, 11, 18]). Under this prior, it is hard to incorporate prior information on central subspaces. However, under our formulation and the subsequent construction of priors, it should be very easy to incorporate prior information about the central subspace.

The paper is organized as follows. The rest of this section is devoted to a review of some relevant distributions and an introduction of a generalized Bingham distribution, which will be used for the construction of priors in the Bayesian envelope model. Section 2 provides an alternative parameterization of the envelope model, which is essential for model formulation. In Section 3, we introduce the class of

prior distributions and provide sufficient conditions for posterior propriety. In Section 3.3, we develop a block Gibbs sampling algorithm to sample from the posterior distributions corresponding to the priors introduced in Section 3. Section 4 illustrates the applicability of the Bayesian framework developed in the paper on simulated and real datasets. In Section 5, we include the proofs of the technical results. A few additional technical results as well as other details are provided in the supplementary material [9].

1.2. *Some relevant distributions.* We now establish notation for some probability distributions, which will be useful in constructing a class of prior distributions for the envelope model, and also in generating samples from the corresponding posterior distribution.

In preparation, we first introduce some notation. Let $M_{a,b}$ denote the space of $a \times b$ matrices, and $S_{a,b}$ denote the Stiefel manifold comprised of all $a \times b$ semi-orthogonal matrices, that is, if $\mathbf{\Gamma} \in S_{a,b}$ with $a \geq b$, then $\mathbf{\Gamma}^T \mathbf{\Gamma} = \mathbf{I}_b$. Let $S^+_{a,b}$ denote the collection of all $a \times b$ semi-orthogonal matrices such that the maximum entry (in absolute value) for each column of the matrix is positive, that is, $\max(c_i) > |\min(c_i)|$ where $(c_1, \ldots, c_p)^T$ is a column. Note that the compact unimodular group $S_{a,a}$ has a unique Haar measure, which through obvious mappings, gives rise to *induced Haar measures* on $S_{a,b}$ and $S^+_{a,b}$. Let $O_a \subset \mathbb{R}^a$ denote the set of vectors with positive entries arranged in decreasing order (the first entry is at least as large as the second entry, and so on). Let n, r, p be positive integers, and let u be a nonnegative integer satisfying $u \leq r$.

DEFINITION 1. An $a \times b$ random matrix $\mathbf{H}$ is defined to follow a matrix-variate normal distribution $[\mathrm{MN}_{a,b}(\mathbf{M}, \mathbf{A}_1, \mathbf{A}_2)]$ if its density function (on the space $M_{a,b}$) is given by

$$\frac{|\mathbf{A}_1|^{-b/2}|\mathbf{A}_2|^{-a/2}}{\sqrt{2\pi}^{ab}} e^{(-1/2)\,\mathrm{tr}(\mathbf{A}_1^{-1}(\mathbf{H}-\mathbf{M})\mathbf{A}_2^{-1}(\mathbf{H}-\mathbf{M})^T)}.$$

Here, $\mathbf{M} \in M_{a,b}$. Also, $\mathbf{A}_1 \in M_{a,a}$ and $\mathbf{A}_2 \in M_{b,b}$ are both positive definite matrices.

It is known that if $\mathbf{H} \sim \mathrm{MN}_{a,b}(\mathbf{M}, \mathbf{A}_1, \mathbf{A}_2)$, then $\mathbf{H}$ has rank equal to $\min(a, b)$ with probability 1. Note that $\mathbf{H} \sim \mathrm{MN}_{a,b}(\mathbf{M}, \mathbf{A}_1, \mathbf{A}_2)$ if and only if $\mathrm{vec}(\mathbf{H}) \sim N_{ab}(\mathrm{vec}(\mathbf{M}), \mathbf{A}_2 \otimes \mathbf{A}_1)$. Here, $\mathrm{vec}(\mathbf{H})$ stands for the vector obtained by stacking the columns of the matrix $\mathbf{H}$ on top of each other, and $\otimes$ stands for the Kronecker product. Hence, a draw from a matrix-variate normal distribution can be obtained by making a draw from the corresponding vectorized multivariate normal distribution.

DEFINITION 2. A random variable Z is defined to have a truncated Inverse-Gamma distribution with parameters $a, b, d > 0, c \geq 0$ [Inverse-Gamma(a, b, c, d)] if the probability density function of Z is proportional to $z^{-a-1}e^{-b/z}1_{z\in(c,d)}$.

Methods to sample from the truncated Inverse-Gamma distribution are discussed in the the supplementary material, Section E [9].

DEFINITION 3. Let $\mathbf{A}_1 \in M_{a,a}$ and $\mathbf{B}_1 \in M_{b,b}$ both be positive definite matrices. A random matrix $\mathbf{O}$ is defined to have a matrix Bingham distribution on $S_{a,b}$ with parameters $\mathbf{A}_1$ and $\mathbf{B}_1$ $[B_{a,b}(\mathbf{A}_1, \mathbf{B}_1)]$ if the probability density function of $\mathbf{O}$ (with respect to the induced Haar measure on $S_{a,b}$) is proportional to

$$e^{(-1/2)\,\mathrm{tr}(\mathbf{B}_1\mathbf{O}^T\mathbf{A}_1\mathbf{O})}. \tag{3}$$

The matrix Bingham distribution is in turn a matrix version of the Bingham distribution introduced in Bingham [2]. See [8] and the references therein for more details. Note that the density in (3) is invariant under arbitrary sign changes to the columns of $\mathbf{O}$. Hence, the $B_{a,b}(\mathbf{A}_1, \mathbf{B}_1)$ density on $S_{a,b}$ induces a density on $S^+_{a,b}$ which is given exactly by the expression in (3), up to proportionality. We will refer to this density as the $B_{a,b}(\mathbf{A}_1, \mathbf{B}_1)$ density on $S^+_{a,b}$.

We now introduce a generalized version of the matrix Bingham distribution on the Stiefel manifold $S_{2,2}$. This distribution will play a crucial role in the block Gibbs sampling algorithm in Section 3.3.

DEFINITION 4. A random matrix $\mathbf{Z} = [\mathbf{Z}_1 \quad \mathbf{Z}_2]$ is defined to have a generalized matrix Bingham distribution on $S_{2,2}$ with parameters $\mathbf{A}_1$ and $\mathbf{A}_2$ $[\mathrm{GB}_{2,2}(\mathbf{A}_1, \mathbf{A}_2)]$ if the probability density function of $\mathbf{Z}$ (with respect to the Haar measure on $S_{2,2}$) is proportional to

$$e^{-\mathbf{Z}_1^T\mathbf{A}_1\mathbf{Z}_1 - \mathbf{Z}_2^T\mathbf{A}_2\mathbf{Z}_2}. \tag{4}$$

The parameters $\mathbf{A}_1$ and $\mathbf{A}_2$ here are both semi-positive definite matrices. Hoff [8] provides a rejection sampling method to generate samples from the matrix Bingham distribution on $S_{2,2}$. A direct adaptation of this approach in the $\mathrm{GB}_{2,2}$ setting turns out to be inefficient. We provide a more nuanced rejection sampling method to draw exact samples from the $\mathrm{GB}_{2,2}(\mathbf{A}_1, \mathbf{A}_2)$ distribution in the supplementary material, Section A [9].

2. A reparameterization of the envelope model. The envelope model (2) in [5] is parameterized in terms of Grassmann manifolds. In particular, the envelope subspace $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is a point in an $r \times u$ Grassmann manifold. The specification of $\mathbf{\Gamma}$ is not unique, but $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathrm{span}(\mathbf{\Gamma})$ is unique. Since it is usually hard to specify a subspace, but it is much easier to specify a basis, we start by introducing an alternative parametrization of the envelope model, which is in terms of Stiefel manifolds. Under this parameterization, we use a unique orthogonal basis to represent an envelope subspace. This leads to a more transparent view of the structure of the envelope subspace, and also eases the incorporation of prior information about the envelope subspace into the specification of hyperparameters.

We consider the following parametrization of the envelope model:

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \tag{5}$$

where $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, $\boldsymbol{\Gamma} \in S_{r,u}^+$, and $\boldsymbol{\eta} \in M_{u,p}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are diagonal matrices with diagonal entries arranged in decreasing order, and $\boldsymbol{\Gamma}_0 \in S_{r,r-u}^+$ with $\boldsymbol{\Gamma}_0^T\boldsymbol{\Gamma} = 0$. We use $\boldsymbol{\omega} \in O_u$ and $\boldsymbol{\omega}_0 \in O_{r-u}$ to denote the diagonal vectors of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$, respectively. Note that if $\boldsymbol{\Gamma} \in S_{r,u}^+$, and $\boldsymbol{\Gamma}_0 \in S_{r,r-u}^+$, then $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in S_{r,r}^+$. Hence, the parameter that needs to be estimated is $(\boldsymbol{\mu}, \boldsymbol{\eta}, (\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0), \boldsymbol{\omega}, \boldsymbol{\omega}_0)$, and the parameter space is $M_{r,1} \times M_{u,p} \times S_{r,r}^+ \times O_u \times O_{r-u}$.

The construction of (2) from (5) can be worked out as follows. For the envelope model in (2), the parameters $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are full symmetric (not necessarily diagonal) matrices. Let $\boldsymbol{\Omega} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T$ and $\boldsymbol{\Omega}_0 = \mathbf{P}_0\boldsymbol{\Lambda}_0\mathbf{P}_0^T$ denote the spectral decompositions of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$, respectively, with the diagonal entries of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}_0$ arranged in decreasing order. Now, set $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}_0$ as the "new" $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. Also, set $\boldsymbol{\Gamma}\mathbf{P}$ and $\boldsymbol{\Gamma}_0\mathbf{P}_0$ (with changes of signs for columns whose maximum entry is not positive) as the "new" $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$, respectively. The resulting model is precisely given by (5).

Hence, although the parameterizations in (2) and (5) look similar, there are important differences. The $\boldsymbol{\Gamma}$ parameter in (5) is the unique orthogonal basis of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ that belongs to $S_{r,u}^+$, and diagonalizes the "old" $\boldsymbol{\Omega}$ [in (2)] with diagonal elements, or eigenvalues, in decreasing order. The restrictions in (5) on the diagonal entries in $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$, and the signs of the columns of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$, ensure parameter identifiability. This is crucial in facilitating an effective Bayesian analysis.

REMARK 1. If $u = r$, it can be easily shown that the envelope model (5) is equivalent to the standard multivariate regression model, by considering the transformation $(\boldsymbol{\Gamma}, \boldsymbol{\eta}, \boldsymbol{\Omega}) \to (\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}, \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T)$.

3. Prior distributions for the envelope model. Let ν_1 denote the Lebesgue measure on $M_{r,1}$, ν_2 denote the Lebesgue measure on $M_{u,p}$, ν_3 denote the projection of the Haar measure for $S_{r,r}$ on $S_{r,r}^+$, ν_4 denote the Lebesgue measure on O_u, and ν_5 denote the Lebesgue measure on O_{r-u}. We specify a (possibly improper) prior density for the parameter of interest (with respect to $\nu_1 \times \nu_2 \times \nu_3 \times \nu_4 \times \nu_5$) as follows.

• $\boldsymbol{\mu}$ is a priori independent of the other parameters and we put a flat improper prior on $\boldsymbol{\mu}$, that is, the (improper) prior density of $\boldsymbol{\mu}$ with respect to ν_1 is given by

$$\pi(\boldsymbol{\mu}) \propto 1. \tag{6}$$

• Fix a $p \times p$ positive semi-definite matrix $\mathbf{C}$, and $\mathbf{e} \in M_{r,p}$. The prior density of $\boldsymbol{\eta}$ (with respect to ν_2) conditioned on $((\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0), \boldsymbol{\omega}, \boldsymbol{\omega}_0)$ is proportional to

$$|\boldsymbol{\Omega}|^{-p/2} e^{(-1/2)\operatorname{tr}(\boldsymbol{\Omega}^{-1}(\boldsymbol{\eta}-\boldsymbol{\Gamma}^T\mathbf{e})\mathbf{C}(\boldsymbol{\eta}-\boldsymbol{\Gamma}^T\mathbf{e})^T)}. \tag{7}$$

If $\mathbf{C}$ is positive definite, then (7) the prior distribution of $\boldsymbol{\eta}$ conditioned on $((\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0), \boldsymbol{\omega}, \boldsymbol{\omega}_0)$ is matrix normal. In particular,

$$\boldsymbol{\eta} \mid (\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0), \boldsymbol{\omega}, \boldsymbol{\omega}_0 \sim \mathrm{MN}_{u,p}\big(\boldsymbol{\Gamma}^T \mathbf{e}, \boldsymbol{\Omega}, \mathbf{C}^{-1}\big).$$

Note that $\boldsymbol{\Gamma}^T \mathbf{e}$ can be interpreted as the conditional prior mean of $\boldsymbol{\eta}$. Also, let ω_i be the ith element in $\boldsymbol{\omega}$, then $\omega_i \mathbf{C}$ can be interpreted as the conditional covariance matrix of the ith row of $\boldsymbol{\eta}$. Note that we allow for impropriety by allowing $\mathbf{C}$ to be singular. For example, choosing $\mathbf{C} = 0$ would lead to the flat (improper) prior on $M_{u,p}$.

• Fix an $r \times r$ diagonal matrix $\mathbf{D}$ with positive diagonal entries, and an $r \times r$ positive semi-definite matrix $\mathbf{G}$. The marginal prior density of $\mathbf{O} = [\boldsymbol{\Gamma} \ \ \boldsymbol{\Gamma}_0]$ (with respect to ν_3) is proportional to

$$e^{(-1/2)\,\mathrm{tr}(\mathbf{D}^{-1}\mathbf{O}^T\mathbf{G}\mathbf{O})}. \tag{8}$$

It follows from (8) that the prior distribution of $\mathbf{O} = [\boldsymbol{\Gamma} \ \ \boldsymbol{\Gamma}_0]$ is a matrix Bingham distribution. In particular,

$$\mathbf{O} \sim B_{r,r}\big(\mathbf{G}, \mathbf{D}^{-1}\big).$$

The following lemma shows that the prior mode for $\mathbf{O}$ is an appropriately permuted version of the matrix of eigenvectors of $\mathbf{G}$, and thereby provides a concrete interpretation for the hyperparameter $\mathbf{G}$.

LEMMA 1. *Suppose that the eigenvalues of* $\mathbf{G}$ *are distinct. Then the mode for the prior density of* $\mathbf{O}$ *is unique and can be specified as follows. Let* $\mathbf{a}$ *denote the vector of diagonal entries of* $\mathbf{D}$. *For every* $i = 1, \ldots, r$, *let* t_i *denote the rank of the* a_i *among all entries in* $\mathbf{a}$ (*from the smallest to the largest*). *Then for every* $i = 1, \ldots, r$, *the* i*th column of the prior mode of* $\mathbf{O}$ *is given by the eigenvector* (*whose maximum entry is positive*) *of* $\mathbf{G}$ *corresponding to the* t_i*th largest eigenvalue of* $\mathbf{G}$.

For example, if $\mathbf{a} = (1, 5, 3, 10, 16)$, then $t_1 = 1$, $t_2 = 3$, $t_3 = 2$, $t_4 = 4$, $t_5 = 5$. The prior mode is reached when the columns of $\mathbf{O}$ are the eigenvectors of $\mathbf{G}$, corresponding to the $(1, 3, 2, 4, 5)$th largest eigenvalues of $\mathbf{G}$.

The relative magnitude of the diagonal elements of $\mathbf{D}$ represents the strength of the prior information on the envelope subspace. If $\mathbf{D} = \mathbf{I}_r$, then $\mathrm{tr}(\mathbf{D}^{-1}\mathbf{O}^T\mathbf{G}\mathbf{O}) = \mathrm{tr}(\mathbf{G})$, which leads to a flat noninformative prior density for $\mathbf{O}$. On the other hand, if the prior information on the envelope subspace is reliable, we can make the diagonal elements of $\mathbf{D}$ widely apart, such that their ranking is clear. For example, if $r = 5$, the vector of diagonal elements of $\mathbf{D}$ can be chosen as $(0.01, 0.1, 1, 10, 100)$. Also, if we are only confident about part of the information on the envelope subspace, we can make the corresponding diagonal elements in $\mathbf{D}$ to be widely apart, and the rest of the elements in $\mathbf{D}$ close to each other. For example, suppose we know that the leading eigenvector $\mathbf{v}_1$ of $\mathbf{G}$ is likely to be the first column in $\boldsymbol{\Gamma}$,

but we do not have any other information about $\mathbf{\Gamma}$ and $\mathbf{\Gamma}_0$. Then we make the first diagonal element of $\mathbf{D}$ to be quite larger than the others, and set all other elements to be the same. For example, if we choose the diagonal elements of $\mathbf{D}$ to be $(10, 1, 1, 1, 1)$, then $-\operatorname{tr}(\mathbf{D}^{-1}\mathbf{O}^T\mathbf{G}\mathbf{O}) = -\operatorname{tr}(\mathbf{G}) + 0.9\mathbf{O}_{\cdot 1}^T\mathbf{G}\mathbf{O}_{\cdot 1}$, which is maximized when $\mathbf{O}_{\cdot 1} = \mathbf{v}_1$. Another example is presented in Remark 2.

• $(\boldsymbol{\omega}, \boldsymbol{\omega}_0)$ and $\mathbf{O}$ are a priori independent. Given $\alpha, \alpha_0, \lambda, \lambda_0 > 0$, the entries of $\boldsymbol{\omega}$ are a priori distributed as order statistics of u i.i.d. observations from the Inverse-Gamma$(\alpha, \lambda, 0, \infty)$ distribution, and (independently) the entries of $\boldsymbol{\omega}_0$ are a priori distributed as order statistics of $r - u$ i.i.d. observations from the Inverse-Gamma$(\alpha_0, \lambda_0, 0, \infty)$ distribution. In particular, the joint prior density of $\boldsymbol{\omega}$ and $\boldsymbol{\omega}_0$ (with respect to $\nu_4 \times \nu_5$) is proportional to

$$\prod_{i=1}^{u} \omega_i^{-\alpha-1} e^{-\lambda/\omega_i} \prod_{i=1}^{r-u} \omega_{0,i}^{-\alpha_0-1} e^{-\lambda_0/\omega_{0,i}}. \tag{9}$$

Additionally, we allow for impropriety by allowing α to take values in $[-(p/2 + 1), 0]$, α_0 to take values in $[-1, 0]$, and λ, λ_0 to take zero values. For example, the choice $\alpha = -1$, $\alpha_0 = -1$ and the choice $\lambda = \lambda_0 = 0$ will lead to the flat (improper) prior on $O_u \times O_{r-u}$.

The hyperparameters that are needed to specify a prior from the above class are $\mathbf{e}$, $\mathbf{C}$, $\mathbf{G}$, $\mathbf{D}$, α, α_0 and λ, λ_0. Each of these has a natural interpretation in terms of the respective parameters. Hence, in case subjective prior knowledge exists, it can be incorporated in our model through these hyperparameters (see Section 3.1). Of course, if there is no subjective prior knowledge, as illustrated below, our class of prior distributions is flexible enough to allow for objective (often improper) priors.

3.1. *Two specific prior choices.* A natural choice of an objective prior is the following. If we choose $\mathbf{e} = 0$, $\mathbf{C} = 0$, $\mathbf{G} = 0$, $\alpha = -(p/2 + 1)$, $\alpha_0 = -1$ and $\lambda = \lambda_0 = 0$, it follows from the above discussion that the joint (improper) prior density for the parameters (on the space $M_{r,1} \times M_{u,p} \times S_{r,r}^+ \times O_u \times O_{r-u}$ with respect to measure $\nu_1 \times \nu_2 \times \nu_3 \times \nu_4 \times \nu_5$) is given by

$$\pi\big(\boldsymbol{\mu}, \boldsymbol{\eta}, (\mathbf{\Gamma}, \mathbf{\Gamma}_0), \boldsymbol{\omega}, \boldsymbol{\omega}_0\big) \propto 1, \tag{10}$$

for every $(\boldsymbol{\mu}, \boldsymbol{\eta}, (\mathbf{\Gamma}, \mathbf{\Gamma}_0), \boldsymbol{\omega}, \boldsymbol{\omega}_0) \in M_{r,1} \times M_{u,p} \times S_{r,r}^+ \times O_u \times O_{r-u}$. We will refer to the density in (10) as the uniform Haar improper prior density for the envelope model.

We also suggest another prior as follows. Su and Cook [16], Section 3.5, provide a detailed description and discussion of a procedure for generating a suitable initial value of the parameters for an iterative optimization procedure in the context of partial envelope models. An adaptation of this approach to the current setting is summarized in the supplementary material, Section B [9]. Let $(\boldsymbol{\eta}^*, (\mathbf{\Gamma}^*, \mathbf{\Gamma}_0^*), \boldsymbol{\omega}^*, \boldsymbol{\omega}_0^*)$ denote the parameter values obtained after performing this

procedure. We now make the following choice of hyperparameters. Set $\mathbf{e} = \boldsymbol{\Gamma}^* \boldsymbol{\eta}^*$. If $u \geq 2$, solve for α and λ from the equations

$$\frac{\alpha}{\lambda} = \frac{1}{u} \sum_{i=1}^{u} \frac{1}{\omega_i^*}, \quad \text{and} \quad \frac{\alpha^2}{\lambda^2} + \frac{\alpha}{\lambda^2} = \frac{1}{u} \sum_{i=1}^{u} \frac{1}{(\omega_i^*)^2}.$$

Here, we are essentially using the method of moments, and the fact that the entries of $\boldsymbol{\omega}$ are a priori distributed as order statistics of u i.i.d. observations from the Inverse-Gamma$(\alpha, \lambda, 0, \infty)$ distribution. If $u = 1$, we choose $\alpha = \frac{1}{\omega_1^*}$ and $\lambda = 1$.

In a similar fashion, if $u < r - 1$, we solve for α_0 and λ_0 from the equations

$$\frac{\alpha_0}{\lambda_0} = \frac{1}{r-u} \sum_{i=1}^{r-u} \frac{1}{\omega_{0,i}^*}, \quad \text{and} \quad \frac{\alpha_0^2}{\lambda_0^2} + \frac{\alpha_0}{\lambda_0^2} = \frac{1}{r-u} \sum_{i=1}^{r-u} \frac{1}{(\omega_{0,i}^*)^2}.$$

If $u = r - 1$, we choose $\alpha_0 = \frac{1}{\omega_{0,1}^*}$ and $\lambda_0 = 1$.

We assign the elements of $(\boldsymbol{\omega}^*, \boldsymbol{\omega}_0^*)$ as the diagonal elements of the hyperparameter $\mathbf{D}$. Alternatively, the diagonal elements of $\mathbf{D}$ can also be generated (i.i.d.) from a distribution supported on the positive part of the real line. Based on Lemma 1, we now provide a procedure to choose the hyperparameter $\mathbf{G}$ to incorporate the prior information about the envelope subspace as follows. We first generate the eigenvalues of $\mathbf{G}$, $\lambda_1, \lambda_2, \ldots, \lambda_r$ i.i.d. from a distribution supported on positive part of the real line. Let $\lambda_{(1)}, \ldots, \lambda_{(r)}$ be the ordered λ's (from the largest to the smallest). The relative magnitude of the eigenvalues can also represent our confidence in the prior information. The more confident we are about our prior information on $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$, the larger variance we put on the distribution used for drawing the $\lambda_{(i)}$'s. In the wheat protein data example in Section 4.4, we have used the chi-square distribution. This is because chi-square distribution has only one parameter that is directly associated with its variance. For every $i = 1, \ldots, r$, let t_i denote the rank of a_i^* among all the entries in $\mathbf{a}^*$. Following the example after Lemma 1, if $\mathbf{a}^* = (1, 5, 3, 10, 16)$, then $t_1 = 1$, $t_2 = 3$, $t_3 = 2$, $t_4 = 4$, $t_5 = 5$. We then construct a diagonal matrix $\boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda}_{ii} = \lambda_{(t_i)}$ for every $i = 1, \ldots, r$. Let $\mathbf{O}^* = [\boldsymbol{\Gamma}^* \ \ \boldsymbol{\Gamma}_0^*]$. Finally, we set $\mathbf{G} = \mathbf{O}^* \boldsymbol{\Lambda} (\mathbf{O}^*)^T$. By Lemma 1, we have ensured through the above procedure that $\mathbf{O}^*$ is the prior mode of $\mathbf{O}$. We set the remaining hyperparameter $\mathbf{C}$ to the zero matrix, thereby completing our hyperparameter choice. Henceforth, we will refer to this prior choice as the empirical prior. Both the prior choices introduced in this section will be illustrated in the wheat protein data application in Section 4.4.

REMARK 2. If only partial information is available for the envelope subspace, it can also be incorporated in a similar manner. For example, in the preceding example with $r = 5$, suppose that $u = 2$ and we know a unit vector $\mathbf{v}_1$ is likely to be the first column in $\boldsymbol{\Gamma}$ and unit vectors $\mathbf{v}_2$ and $\mathbf{v}_3$ are likely to be the first and third columns in $\boldsymbol{\Gamma}_0$. In this case, we know part of the envelope subspace,

and part of its orthogonal complement. We then generate $\lambda_1, \ldots, \lambda_5$ i.i.d. from a distribution supported on positive part of the real line. Let $(\mathbf{v}_4, \mathbf{v}_5) \in S^+_{5,2}$ be a basis of the orthogonal complement of $\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$. Then $\mathbf{G} = \mathbf{O}^* \boldsymbol{\Lambda} (\mathbf{O}^*)^T$, where $\mathbf{O}^* = (\mathbf{v}_1, \mathbf{v}_4, \mathbf{v}_2, \mathbf{v}_5, \mathbf{v}_3)$ and $\boldsymbol{\Lambda}_{ii} = \lambda_{(t_i)}$ for $i = 1, \ldots, 5$. Here, the diagonal elements of $\mathbf{D}$ can be set as $\mathbf{a}^* = (a_1, a_2, a_3, a_4, a_5)$, with $a_1 > a_2 > a_3 > a_4 = a_5$, reflecting our lack of prior knowledge on $\mathbf{v}_4$ and $\mathbf{v}_5$.

3.2. *Derivation of posterior density and conditions for posterior propriety.* Suppose we have n independent observation vectors $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ satisfying (5) with corresponding covariate vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. Then the likelihood function $L(\boldsymbol{\mu}, \boldsymbol{\eta}, (\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0), \boldsymbol{\omega}, \boldsymbol{\omega}_0)$ is given by

$$
\begin{aligned}
&(2\pi)^{-(nr)/2} |\boldsymbol{\Omega}|^{-n/2} |\boldsymbol{\Omega}_0|^{-n/2} \\
&\quad \times e^{(-1/2)\operatorname{tr}\{(\mathbb{Y} - \mathbf{1}_n \boldsymbol{\mu}^T - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0 \boldsymbol{\Gamma}_0^T)^{-1} (\mathbb{Y} - \mathbf{1}_n \boldsymbol{\mu}^T - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)^T\}},
\end{aligned} \tag{11}
$$

where $\mathbb{Y}$ is an $n \times r$ matrix with rows given by $\mathbf{Y}_1^T, \mathbf{Y}_2^T, \ldots, \mathbf{Y}_n^T$, $\mathbb{X}$ is an $n \times p$ matrix with rows given by $\mathbf{X}_1^T, \mathbf{X}_2^T, \ldots, \mathbf{X}_n^T$ and $\mathbf{1}_n \in \mathbb{R}^{n \times 1}$ is a column vector of 1's. Without loss of generality, we assume that the columns of the covariate matrix $\mathbb{X}$ are centered. If $n > p$, we will assume that $\mathbb{X}$ is of full column rank. Clearly, if $n \le p$, then such an assumption does not hold. It follows from (6), (7), (8), (9) and (11) that the posterior density of the parameter of interest (with respect to $\nu_1 \times \nu_2 \times \nu_3 \times \nu_4 \times \nu_5$) is given by

$$
\begin{aligned}
&\pi\big((\boldsymbol{\mu}, \boldsymbol{\eta}, (\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0), \boldsymbol{\omega}, \boldsymbol{\omega}_0) \mid \mathbb{Y}\big) \\
&\quad \propto (2\pi)^{-(nr)/2} |\boldsymbol{\Omega}|^{-n/2} |\boldsymbol{\Omega}_0|^{-n/2} \\
&\qquad \times e^{(-1/2)\operatorname{tr}\{(\mathbb{Y} - \mathbf{1}_n \boldsymbol{\mu}^T - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0 \boldsymbol{\Gamma}_0^T)^{-1} (\mathbb{Y} - \mathbf{1}_n \boldsymbol{\mu}^T - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)^T\}} \\
&\qquad \times e^{(-1/2)\operatorname{tr}(\boldsymbol{\Omega}^{-1}(\boldsymbol{\eta} - \boldsymbol{\Gamma}^T \mathbf{e})\mathbf{C}(\boldsymbol{\eta} - \boldsymbol{\Gamma}^T \mathbf{e})^T)} e^{(-1/2)\operatorname{tr}(\mathbf{D}^{-1}\mathbf{O}^T \mathbf{G} \mathbf{O})} \\
&\qquad \times \prod_{i=1}^{u} \omega_i^{-\alpha - p/2 - 1} e^{-\lambda/\omega_i} \prod_{i=1}^{r-u} \omega_{0,i}^{-\alpha_0 - 1} e^{-\lambda_0/\omega_{0,i}}.
\end{aligned} \tag{12}
$$

Since the prior density is allowed to be improper, it is important to establish conditions under which the posterior density is proper, that is, the expression on the right-hand side in (12) can be normalized to obtain a probability density function. The next theorem establishes sufficient conditions for posterior propriety (proof is provided in Section 5.2).

THEOREM 1. *The posterior density in* (12) *is proper under either of the following conditions*:

1. $n > \max(r, p + 3)$.
2. $n + 2\alpha > 1$, $\lambda, \lambda_0 > 0$ *and* $\mathbf{C}$ *is positive definite*.

3.3. *Sampling from the posterior distribution*. In order to perform Bayesian statistical inference, it is crucial to compute quantities related to the posterior distribution, for example, the posterior mean or posterior quantiles. The posterior density for our model is intractable in the sense that it is not possible to compute these quantities analytically or to generate i.i.d. samples from the posterior distribution. However, we show in the supplementary material, Section C [9], that various conditional posterior densities have a standard form, and we develop systematic scan and random scan Gibbs sampling algorithm to generate samples from the posterior distribution. It follows easily that the Markov chains corresponding to both the random scan and systematic scan Gibbs samplers have the density in (12) as a stationary density. The theorem below shows that both the random scan and systematic scan versions of the Gibbs sampler for the generalized matrix Bingham distribtuion are, in fact, Harris ergodic. This provides theoretical guarantees that the Gibbs sampling algorithms provide approximate samples from the density in (12).

THEOREM 2. *The systematic scan and random scan Gibbs samplers specified in the supplementary material*, *Section* C [9], *are Harris ergodic*.

A proof of this theorem is provided in Section 5.3 below. Note that the theorem above holds for an arbitrary initial value of the parameter vector, and allows us to use the Markov chain averages to approximate intractable posterior quantities of interest (such as posterior means, posterior standard deviations or posterior quantiles). We note here that while we will use posterior means to obtain point estimates of parameters of interest, the posterior mode of the parameter vector can be obtained by an iterative optimization approach specified in the supplementary material, Section D [9].

4. Simulation and data analysis. In this section, we provide one simulated data example and one real example to demonstrate the utility of the methodology developed in this paper. Note that, once u is fixed, methods for computing posterior quantities of interest have been developed in Section 3.3. We now discuss how to select the dimension of the envelope subspace u using the Deviance Information Criterion (DIC) in Section 4.1.

4.1. *Selection of* u. An important issue that needs to be dealt with for fitting an envelope model is the choice of u. Possible choices of u are $u = 0, 1, \ldots, r$. Let

$$\boldsymbol{\theta} = \left(\boldsymbol{\mu}^T, \boldsymbol{\eta}^T, \operatorname{vec}\left([\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0]\right)^T, \boldsymbol{\omega}^T, \boldsymbol{\omega}_0^T\right)^T$$

denote the parameter vector. Notice that the effective number of parameters changes with u, by Section 3.1 in [5], the effective number of parameters is $r + up + r(r+1)/2$. For each u, we fit a Bayesian envelope model with a given

choice of prior. Let $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{M}$ represent values sampled from the relevant posterior distribution after performing M steps of the Gibbs sampling Markov chain (after an appropriate burn-in). For a given parameter vector $\boldsymbol{\theta}$, we define the *deviance* as $D(\boldsymbol{\theta}) := -2\log L(\boldsymbol{\theta})$, where $L(\boldsymbol{\theta})$ is the likelihood function defined in (11). Let $\bar{D} := \sum_{i=1}^{M} D(\boldsymbol{\theta}^{(i)})/M$. The DIC score (see, e.g., [7], page 185) with the given choice of u is then computed as $DIC = \bar{D} + \sum_{i=1}^{M}(D(\boldsymbol{\theta}^{(i)}) - \bar{D})^2/(2(M-1))$. The value of u which corresponds to the minimum DIC score is chosen.

4.2. *Simulated data*. In this section, based on common practice in the literature, we perform a frequentist evaluation of the Bayesian procedures developed in this paper. This will be done by using replicated datasets which are generated from a known "true" model, and evaluating the average model selection and estimation performance of our Bayesian procedures in comparison with the true model.

We consider a setting with $r = 5$, $p = 2$ and $u = 2$. The subscript "true" is attached to a quantity if it is associated with the "true" model. Elements in $\boldsymbol{\eta}_{\text{true}} \in \mathbb{R}^{u\times p}$ were generated from i.i.d. Uniform[0, 5]. We obtained the $r \times r$ orthogonal matrix $\mathbf{O}_{\text{true}}$ from the left singular vectors of an $r \times r$ matrix with i.i.d. standard normal entries. The basis of the envelope subspace $\boldsymbol{\Gamma}_{\text{true}}$ was chosen to be the first two columns of $\mathbf{O}_{\text{true}}$, and $\boldsymbol{\Gamma}_{0,\text{true}}$ was chosen to be the remaining three columns. We set $\boldsymbol{\omega}_{\text{true}}$ to be $(1, 2)^T$ and $\boldsymbol{\omega}_{0,\text{true}}$ to be $(20, 16, 12)^T$; the error covariance matrix $\boldsymbol{\Sigma}_{\text{true}}$ is then $\boldsymbol{\Sigma}_{\text{true}} = \boldsymbol{\Gamma}_{\text{true}}\boldsymbol{\Omega}_{\text{true}}\boldsymbol{\Gamma}_{\text{true}}^T + \boldsymbol{\Gamma}_{0,\text{true}}\boldsymbol{\Omega}_{0,\text{true}}\boldsymbol{\Gamma}_{0,\text{true}}^T$. The intercept vector $\boldsymbol{\mu}_{\text{true}}$ was a vector of zeros. We varied the sample sizes from 30, 100 and 200. For each sample size, 200 datasets were generated. The elements in $\mathbf{X}$ were i.i.d. Uniform[0, 5] variables, and $\mathbf{Y}$ was computed following the envelope model in (5).

We first demonstrate the performance of DIC in dimension selection. To ensure a fair comparison among various choices of u, we use the uniform Haar prior [as described in (10)] for each choice of u. With every sample size, we counted the number of replications for which the minimum DIC score corresponds to each given value of u between 0 and 5. The results are provided in Table 1. The parameter vector of the true model has effective dimension 24. We notice that with sample size 30, DIC selects the correct value of u 98.5% of the time, and it selects the correct dimension 100% of the time for sample sizes 100 and 200. Hence, with small

TABLE 1
Number of replications (out of 200) for which a given value of u corresponds to the minimum DIC score

	$u=0$	$u=1$	$u=2$	$u=3$	$u=4$	$u=5$
$n=30$	0	3	197	0	0	0
$n=100$	0	0	200	0	0	0
$n=200$	0	0	200	0	0	0

sample size, it seems that DIC would occasionally underestimate the dimension, but this problem is quickly mitigated for moderate sample size.

Now we focus on comparing the efficiency and accuracy of the posterior mean estimators of the Bayesian envelope model versus the Bayesian standard multivariate linear regression model. We consider two versions: the Bayesian standard model, which refers to a Bayesian envelope model with $u = r$ (see Remark 1); and the Bayesian Wishart standard model, which specifies an inverse Wishart distribution for $\boldsymbol{\Sigma}$ and a matrix normal prior for $\boldsymbol{\beta} \mid \boldsymbol{\Sigma}$ (see [12], page 33). Now, for all datasets mentioned above, we fit a Bayesian envelope model with $u = 2$ using the uniform Haar prior, and also a Bayesian standard model again with a uniform Haar prior. For both models, posterior mean estimates for $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ were obtained for each of the datasets using the Gibbs sampling procedure outlined in Section 3.3 (and the supplementary material, Section C [9]). A burn-in of 1000 iterations was used, and the posterior estimates were computed using the next 2500 iterations.

For each sample size $n = 30, 100, 200$, we computed the estimation variance and average squared error (an estimate of MSE) of the posterior mean estimates for each element in $\boldsymbol{\beta}$ over 200 replications for both the Bayesian envelope model and the Bayesian standard model. More specifically, for each model, let $\hat{\beta}_{ij}^{(k)}$ be the estimate of the (i, j)th element from the kth replication, the estimation variance is given by $\sum_{k=1}^{200}(\hat{\beta}_{ij}^{(k)} - \bar{\beta}_{ij})^2/200$, where $\bar{\beta}_{ij} = \sum_{k=1}^{200}\hat{\beta}_{ij}^{(k)}/200$, and the average squared error is given by $\sum_{k=1}^{200}(\hat{\beta}_{ij}^{(k)} - \beta_{ij,\text{true}})^2/200$. The estimation variances and average squared errors for estimating β_{11} from both models are plotted in Figure 1. The other elements in $\boldsymbol{\beta}$ all have similar pattern. From Figure 1, we first notice that the red lines (estimation variance) and the blue lines (average squared error) are overlapping with each other for both methods, indicating that estimation variance is the major contributor of average squared error. For each sample size, the estimator from the Bayesian envelope model has much smaller estimation variance and average squared error than the Bayesian standard model. To quantify

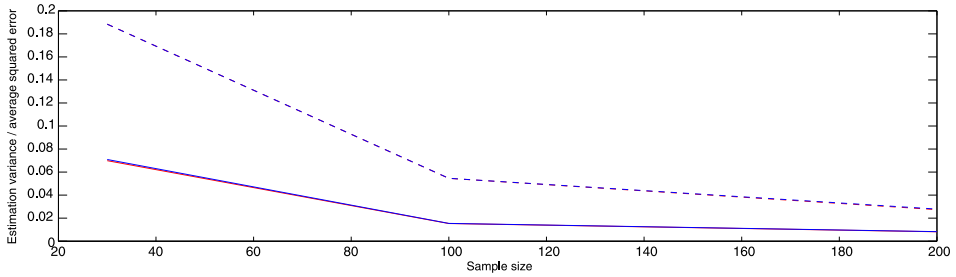

FIG. 1. *Comparison of the Bayesian standard model versus the Bayesian envelope model. The solid lines mark the Bayesian envelope model and the dashed lines mark the Bayesian standard model. The red lines mark estimation variances and the blue lines mark the average squared errors. The red lines are overlapped with the blue lines in this plot because estimation variance is the main contributor of the average square errors.*

TABLE 2

Ratio of estimation variance and average squared error in estimation of $\boldsymbol{\beta}$. $Ratio_{\mathrm{Var}}$ denotes the ratio of estimation variance of the Bayesian standard model versus the Bayesian envelope model. $Ratio_{\mathrm{MSE}}$ denotes the ratio of average squared error of the Bayesian standard model versus the Bayesian envelope model

	$n = 30$		$n = 100$		$n = 200$	
	$Ratio_{\mathrm{Var}}$	$Ratio_{\mathrm{MSE}}$	$Ratio_{\mathrm{Var}}$	$Ratio_{\mathrm{MSE}}$	$Ratio_{\mathrm{Var}}$	$Ratio_{\mathrm{MSE}}$
$\beta_{1,1}$	2.867	2.828	3.447	3.462	2.868	3.187
$\beta_{1,2}$	1.986	1.985	1.740	1.749	2.249	2.286
$\beta_{1,3}$	2.772	2.765	2.498	2.499	3.573	3.572
$\beta_{1,4}$	3.454	3.444	2.910	2.929	2.827	2.906
$\beta_{1,5}$	2.863	2.841	3.480	3.494	4.475	4.637
$\beta_{2,1}$	2.161	2.141	2.058	2.029	1.870	1.908
$\beta_{2,2}$	1.867	1.859	1.254	1.253	1.557	1.558
$\beta_{2,3}$	2.244	2.242	1.965	1.953	2.383	2.382
$\beta_{2,4}$	2.376	2.367	2.181	2.184	1.796	1.783
$\beta_{2,5}$	1.872	1.868	2.169	2.172	2.127	2.216

the efficiency gains by Bayesian envelope model, we computed the ratios of the estimation variance and average squared error of the Bayesian standard estimator versus the Bayesian envelope estimator for each sample size. From Table 2, we find that the ratios stays about the same across different sample sizes, which suggests that the relative performance of the Bayesian envelope model versus the Bayesian standard model is quite stable with the change of sample size. For sample size 200, the estimation variance ratio ranges from 1.56 to 4.48, which is a substantial efficiency gain in many applications. We also compared the Bayesian envelope model with the Bayesian Wishart standard model, and the results are similar to those in Table 2. The details are contained in the supplementary material, Section G [9].

As pointed out in [5], the efficiency gains for the envelope model are substantial especially when the immaterial part is more variant than the material part, that is, the largest entry in $\boldsymbol{\omega}_{0,\mathrm{true}}$ is larger than the largest entry in $\boldsymbol{\omega}_{\mathrm{true}}$. When the immaterial part is less variant than the material part, the envelope model is still at least as efficient as the standard model, but the gains are not as substantial as the ones presented in Table 2. The Bayesian envelope model inherits these properties. To demonstrate this, and for balance, we added simulation results with $\boldsymbol{\omega}_{\mathrm{true}} = (20, 16)^T$, $\boldsymbol{\omega}_{0,\mathrm{true}} = (0.5, 1, 2)^T$ and $\boldsymbol{\omega}_{\mathrm{true}} = (1, 1)^T$, $\boldsymbol{\omega}_{0,\mathrm{true}} = (1, 1, 1)^T$ in the supplementary material, Section H [9]. Also, since our estimator is derived from the normal likelihood, we also investigated its robustness under deviation of normality. Based on our numerical experiments (not shown here), we find that a moderate departure from normality does not have a notable affect on the performance of the Bayesian envelope estimator.

The Bayesian envelope model also works well with small sample size. We set $r = 50$ and $n = 30$. We let $p = 10$, $\boldsymbol{\omega}_{\mathrm{true}} = (2, 1)$, $\boldsymbol{\omega}_{0,\mathrm{true}} = (100, 99, \ldots, 53)$, and

TABLE 3

Ratio of estimation variance and average squared error for ten randomly selected elements in $\boldsymbol{\beta}$, with $n = 30$ and $r = 50$. $Ratio_{\mathrm{Var}}$ denotes the ratio of estimation variance of the Bayesian standard model versus the Bayesian envelope model. $Ratio_{\mathrm{MSE}}$ denotes the ratio of average squared error of the Bayesian standard model versus the Bayesian envelope model

	$\beta_{6,2}$	$\beta_{9,10}$	$\beta_{19,8}$	$\beta_{15,6}$	$\beta_{10,9}$	$\beta_{48,9}$	$\beta_{3,5}$	$\beta_{15,2}$	$\beta_{17,9}$	$\beta_{34,2}$
$Ratio_{\mathrm{MSE}}$	4.390	5.423	5.101	7.917	8.870	7.653	5.412	2.712	8.543	3.919
$Ratio_{\mathrm{Var}}$	4.390	5.383	5.099	9.914	8.837	7.971	5.357	3.130	8.627	3.936

generated the elements in $\boldsymbol{\eta}_{\mathrm{true}}$ independently from $N(0, 10^2)$. The other parameters were generated the same way as in the settings that produced Table 2. We used the empirical Bayes prior discussed in Section 3.1 to set the hyperparameters. By Theorem 1, when $n < r$, the hyperparameter $\mathbf{C}$ needs to be positive definite for posterior propriety. Hence, we set $\mathbf{C}$ to be a diagonal matrix with diagonal elements being independent χ^2_1 variates. We find that the Bayesian envelope model achieves substantial efficiency gains compared to the Bayesian standard model. Table 3 provides the ratio of estimation variance and average squared error for ten randomly selected elements of $\boldsymbol{\beta}$. For all of the elements in Table 3, the ratio of estimation variance is greater than 1 with some of them much larger than 1, indicating the Bayesian envelope model is more efficient. We also compare the Bayesian envelope model with the Bayesian Wishart standard model using a uniform improper prior for $\boldsymbol{\beta} \mid \boldsymbol{\Sigma}$, and an Inverse-Wishart$(2r + 1, \frac{\mathrm{tr}(\mathbb{Y}_c^T(\mathbf{I}-\mathbf{P}_{\mathbb{X}})\mathbb{Y}_c)}{n}\mathbf{I}_r)$ prior for $\boldsymbol{\Sigma}$. The results are included in the supplementary material, Section I [9].

4.3. *Comparison of the Bayesian envelope estimator and non-Bayesian envelope estimator.* We used the same data that produced Table 2 (200 simulated data sets each for $n = 30, 100, 200$), and we fit both the Bayesian envelope model and the non-Bayesian envelope model to the data. The estimation variance and average squared errors are computed, and their ratios are displayed in Table 4. We notice that the ratios are around 1, indicating that their performance is similar in this setting. We would like to remind the reader that the Bayesian framework developed in this paper offers the following advantages compared to the non-Bayesian framework.

- Ability to address uncertainty by constructing posterior credible intervals (using the same Markov chain used for computing posterior expectations). To illustrate this, Table 5 provides 95% posterior credible intervals for each element of $\boldsymbol{\beta}$ for *a sample data set* (out of the 200 simulated datasets) for sample size $n = 100, 200$.
- Ability to incorporate prior information. We illustrate this by implementing the scenario described in Remark 2. We use the same settings that were used for generating Table 2. Suppose we know a unit vector $\mathbf{v}_1$ is very likely to be the first

TABLE 4

Ratio of estimation variance and average squared error in estimation of $\boldsymbol{\beta}$. $\text{Ratio}_{\text{Var}}$ denotes the ratio of estimation variance of the non-Bayesian envelope model versus the Bayesian envelope model. $\text{Ratio}_{\text{MSE}}$ denotes the ratio of average squared error of the non-Bayesian envelope model versus the Bayesian envelope model

	$n = 30$		$n = 100$		$n = 200$	
	$\text{Ratio}_{\text{Var}}$	$\text{Ratio}_{\text{MSE}}$	$\text{Ratio}_{\text{Var}}$	$\text{Ratio}_{\text{MSE}}$	$\text{Ratio}_{\text{Var}}$	$\text{Ratio}_{\text{MSE}}$
$\beta_{1,1}$	1.031	1.016	1.033	1.017	0.994	1.002
$\beta_{1,2}$	1.019	1.023	1.019	1.024	0.994	0.991
$\beta_{1,3}$	0.995	0.995	0.995	0.994	0.998	0.997
$\beta_{1,4}$	1.000	1.005	1.001	1.005	1.008	1.008
$\beta_{1,5}$	1.014	0.998	1.012	0.996	1.000	0.997
$\beta_{2,1}$	1.011	1.006	1.039	1.035	0.997	0.999
$\beta_{2,2}$	1.017	1.005	1.019	1.009	1.008	1.016
$\beta_{2,3}$	1.020	1.012	1.039	1.029	0.989	0.994
$\beta_{2,4}$	1.028	1.012	1.025	1.011	1.006	1.002
$\beta_{2,5}$	1.019	1.020	1.064	1.065	1.003	1.001

column in $\boldsymbol{\Gamma}$, and unit vectors $\mathbf{v}_2$ and $\mathbf{v}_3$ are very likely to be the first and third columns of $\boldsymbol{\Gamma}_0$. In principle, we can complete $\mathbf{O}^* = (\mathbf{v}_1, \mathbf{v}_4, \mathbf{v}_2, \mathbf{v}_5, \mathbf{v}_3)$ by picking $(\mathbf{v}_4, \mathbf{v}_5) \in S^+_{5,2}$ to be any orthogonal basis of $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)^\perp$. We, however, make a careful choice of $(\mathbf{v}_4, \mathbf{v}_5)$ using the procedure in the supplementary material, Section B [9] as follows. Obtain $\widetilde{\boldsymbol{\Gamma}}$ (as described in the supplementary material, Section B [9]) and denote its two columns by $\mathbf{g}_1$ and $\mathbf{g}_2$. Project $\mathbf{g}_1$ and $\mathbf{g}_2$ on $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)^\perp$, and denote the projection as $\tilde{\mathbf{g}}_1$ and $\tilde{\mathbf{g}}_2$. Without loss

TABLE 5

95% *posterior credible intervals for each element of $\boldsymbol{\beta}$ for a sample data set with $n = 100$ and with $n = 200$*

	$n = 100$		$n = 200$	
	Lower	**Upper**	**Lower**	**Upper**
$\beta_{1,1}$	3.52	3.74	3.60	3.73
$\beta_{1,2}$	−4.60	−4.43	−4.55	−4.40
$\beta_{1,3}$	0.78	0.97	0.89	1.00
$\beta_{1,4}$	0.94	1.10	1.02	1.15
$\beta_{1,5}$	−4.53	−4.37	−4.51	−4.41
$\beta_{2,1}$	−3.06	−2.81	−2.90	−2.71
$\beta_{2,2}$	5.98	6.16	5.84	5.98
$\beta_{2,3}$	1.18	1.45	1.37	1.53
$\beta_{2,4}$	−3.06	−2.80	−3.62	−3.46
$\beta_{2,5}$	3.40	3.61	3.75	3.88

TABLE 6

Ratio of estimation variance and average squared error in estimation of $\boldsymbol{\beta}$. $\text{Ratio}_{\text{Var}}$ denotes the ratio of estimation variance of the non-Bayesian envelope model versus the Bayesian envelope model with informative prior. $\text{Ratio}_{\text{MSE}}$ denotes the ratio of average squared error of the non-Bayesian envelope model versus the Bayesian envelope model

	$n = 30$		$n = 100$		$n = 200$	
	$\text{Ratio}_{\text{Var}}$	$\text{Ratio}_{\text{MSE}}$	$\text{Ratio}_{\text{Var}}$	$\text{Ratio}_{\text{MSE}}$	$\text{Ratio}_{\text{Var}}$	$\text{Ratio}_{\text{MSE}}$
$\beta_{1,1}$	1.362	1.361	1.170	1.167	1.144	1.147
$\beta_{1,2}$	1.440	1.473	1.225	1.203	1.343	1.354
$\beta_{1,3}$	1.249	1.245	1.125	1.120	1.068	1.068
$\beta_{1,4}$	1.135	1.144	1.048	1.046	1.040	1.041
$\beta_{1,5}$	1.198	1.197	1.065	1.072	1.129	1.131
$\beta_{2,1}$	2.491	2.607	2.187	2.223	2.050	2.047
$\beta_{2,2}$	2.923	3.033	2.500	2.527	3.466	3.490
$\beta_{2,3}$	1.773	1.797	1.411	1.403	1.675	1.668
$\beta_{2,4}$	2.197	2.194	1.435	1.426	1.440	1.450
$\beta_{2,5}$	2.012	2.054	1.156	1.156	1.708	1.721

of generality, suppose the norm of $\tilde{\mathbf{g}}_1$ is greater than or equal to $\tilde{\mathbf{g}}_2$ (otherwise switch them). Take $\mathbf{v}_4 = \tilde{\mathbf{g}}_1/\|\tilde{\mathbf{g}}_1\|$, and $\mathbf{v}_5$ as a basis of $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4)^{\perp}$. With $\mathbf{O}^*$ in hand, we now choose the hyperaparameter $\mathbf{G}$ as specified in Remark 2. The generation of other hyperparameters follows the empirical approach specified in the supplementary material, Section B [9] and Section 3.1. Results from a comparison between the Bayesian envelope with this informative prior, and the non-Bayesian envelope model are summarized in Table 6. Compared to Table 4, where we used a noninformative prior, we obtain extra gains by using the informative prior, especially when the sample size is smaller. If we have full information about the full orientation of the envelope subspace and its orthogonal complement, essentially the same procedure as above [without the complication of choosing $(\mathbf{v}_4, \mathbf{v}_5)$] can be used to construct an informative prior.

- Ability to analyze data with $n < r$ (see Table 3).

4.4. *Example*: *Analysis of wheat data*. We now illustrate the Bayesian envelope model using the wheat protein data [5]. The data consists of $r = 6$ responses, which measure the log infrared reflectance at six different wavelengths for 50 ground wheat samples. The predictor is a binary indicator, taking 0 or 1 if a sample has high or low protein content. There are 26 samples with high protein content, and 24 samples with low protein content. We used both the prior choices developed in Section 3.1 for the Bayesian envelope model. To select the dimension of the envelope subspace, we fit a Bayesian envelope model with each u $(u = 0, 1, \ldots, 6)$ and computed the DIC scores. The corresponding DIC scores, obtained using the procedure outlined in Section 4.1, are provided in Table 7. For both prior choices,

TABLE 7
DIC values for all possible dimensions of the Bayesian envelope model with uniform Haar prior and empirical prior

	$u=0$	$u=1$	$u=2$	$u=3$	$u=4$	$u=5$	$u=6$
Uniform prior	1257.7	1201.9	1206.3	1208.0	1209.3	1211.1	1216.0
Empirical prior	1254.5	1197.5	1374.3	1245.8	1266.2	1333.0	1668.0

the model corresponding to $u = 1$ has the lowest DIC score. It is worth mentioning that for the empirical prior case, the gap between the lowest and the second lowest DIC score is much larger as compared to the uniform Haar prior case.

We now turn to estimation of the regression coefficients $\boldsymbol{\beta}$. In this example, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_6)^T$ is a 6×1 vector. For each prior choice, we fit the Bayesian envelope model with $u = 1$ as suggested by DIC, and then compute the posterior mean and posterior standard deviation for each β_i, $i = 1, \ldots, 6$. Based on Remark 1, we use the Bayesian envelope model with $u = 6$ to get the posterior mean and standard deviation of β_i's for the Bayesian standard model. Results are provided in Table 8 and Table 9. It is clear from both these tables that using the envelope model can lead to a substantial reduction in variability of the regression coefficients. The ratios of the posterior standard deviation of the Bayesian standard model versus the Bayesian envelope model range from 6.3 to 65.2 with the uniform Haar prior, and from 6.3 to 64.4 for the empirical prior. Hence, the Bayesian envelope model has much smaller posterior standard deviation, and should produce more reliable estimators.

5. Technical proofs.

5.1. *Proof of Lemma* 1. Without loss of generality, we will assume that the entries of $\mathbf{a}$ are arranged in decreasing order. Let $\mathbf{G} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T$ denote the spectral

TABLE 8
Posterior means and standard deviations of $\boldsymbol{\beta}$ for wheat protein data, using a uniform Haar prior

	Bayesian envelope model		**Bayesian standard model**	
Coefficient	**Post. mean**	**Post. SD**	**Post. mean**	**Post. SD**
β_1	−1.039	0.378	2.934	10.479
β_2	4.406	0.498	7.745	8.630
β_3	3.630	0.417	7.219	9.273
β_4	−5.880	0.644	−2.395	10.157
β_5	0.594	0.224	2.799	14.601
β_6	−1.610	0.904	0.410	5.759

TABLE 9
Posterior means and standard deviation of $\boldsymbol{\beta}$ for wheat protein data, using an empirical prior

	Bayesian envelope model		**Bayesian standard model**	
Coefficient	**Post. mean**	**Post. SD**	**Post. mean**	**Post. SD**
β_1	−1.054	0.372	3.308	9.216
β_2	4.394	0.480	8.064	7.606
β_3	3.619	0.405	7.567	8.170
β_4	−5.868	0.620	−2.046	8.948
β_5	0.588	0.212	3.259	12.815
β_6	−1.555	0.888	0.656	5.115

decomposition of $\mathbf{G}$, where $\mathbf{P} \in S^+_{r,r}$ and the diagonal entries of $\mathbf{\Lambda}$ are arranged in increasing order. Let $\mathbf{A}$ denote the diagonal matrix with the entries of $\mathbf{a}$ as its diagonal elements.

To prove Lemma 1, it enough to prove that $\operatorname{tr}(\mathbf{A}\mathbf{O}^T\mathbf{G}\mathbf{O})$ is uniquely minimized over $S^+_{r,r}$ at $\mathbf{O} = \mathbf{P}$. Let $b_r = a_r$, and $b_i = a_i - a_{i+1}$ for every $i = 1, \ldots, r-1$. Let ℓ_i denote the ith column of $\mathbf{O}$ for every $i = 1, \ldots, r$. It follows that

$$\begin{aligned}\operatorname{tr}(\mathbf{A}\mathbf{O}^T\mathbf{G}\mathbf{O}) &= \sum_{i=1}^{r} a_i \ell_i^T \mathbf{G}\ell_i = \sum_{i=1}^{r}\left(\sum_{j=i}^{r} b_j\right)\ell_i^T\mathbf{G}\ell_i = \sum_{j=1}^{r} b_j\left(\sum_{i=1}^{j}\ell_i^T\mathbf{G}\ell_i\right)\\ &= \sum_{j=1}^{r} b_j \operatorname{tr}(\mathbf{O}_j^T\mathbf{G}\mathbf{O}_j),\end{aligned}$$

where $\mathbf{O}_j$ denotes the submatrix of the first j columns of $\mathbf{O}$. Now, by [13], Theorem 1.2, it follows that for every $j = 1, \ldots, r$, $\operatorname{tr}(\mathbf{O}_j^T\mathbf{G}\mathbf{O}_j)$ is uniquely minimized over $S^+_{r,r}$ when $\mathbf{O}_j$ corresponds to the submatrix of the first j columns of $\mathbf{P}$. Lemma 1 now follows by the arguments above.

5.2. *Proof of Theorem 1.* We start with a lemma which will be a crucial ingredient in investigating propriety of the posterior density. Let $\mathbb{Y}_c \in \mathbb{R}^{n\times r}$ be the centered data matrix of $\mathbf{Y}$ whose ith row is $(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$.

LEMMA 2. *If $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ are independent observations satisfying (5), with corresponding covariate vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ and errors $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_n$, respectively, then with probability 1, $\mathbb{Y}_c^T\mathbb{Y}_c$ is positive definite. If $n > \max(p, r)$, then with probability 1, $\mathbb{Y}_c^T(\mathbf{I} - \mathbf{P}_{\mathbb{X}})\mathbb{Y}_c$ is a positive definite matrix.*

PROOF. Since $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_n$ are i.i.d. multivariate normal, it follows that the $n \times r$ matrix $\mathbf{E}$, whose ith row is given by $\boldsymbol{\varepsilon}_i^T$, follows a $\mathrm{MN}_{n,r}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma})$ distribution. Hence, $\mathbb{Y}_c$ follows a $\mathrm{MN}_{n,r}(\mathbb{X}\boldsymbol{\eta}^T\mathbf{\Gamma}^T, \mathbf{I}_n, \mathbf{\Sigma})$ distribution. Hence, $\mathbb{Y}_c$ has rank

r with probability 1 (since $n > r$). It follows that $\mathbb{Y}_c^T \mathbb{Y}_c$ is positive definite with probability 1.

Since $(\mathbf{I} - \mathbf{P}_{\mathbb{X}})\mathbb{X} = 0$, it follows that $\mathbb{Y}_c^T (\mathbf{I} - \mathbf{P}_{\mathbb{X}})\mathbb{Y}_c = \mathbf{E}^T (\mathbf{I} - \mathbf{P}_{\mathbb{X}})\mathbf{E}$. If $\mathbf{E}^T (\mathbf{I} - \mathbf{P}_{\mathbb{X}})\mathbf{E} = \mathbf{E}^T (\mathbf{I} - \mathbf{P}_{\mathbb{X}})^2 \mathbf{E}$ is not positive definite, then there exists $\mathbf{z} \in \mathbb{R}^r$ such that $(\mathbf{I} - \mathbf{P}_{\mathbb{X}})\mathbf{E}\mathbf{z} = 0$, that is, $\mathbf{E}\mathbf{z}$ lies in the column space of $\mathbb{X}$. Note that $\mathbf{E}\mathbf{z}$ has a multivariate normal distribution on $\mathbb{R}^n$, whereas $\mathbb{X}$ has rank less than or equal to p. Hence, the probability that $\mathbf{E}\mathbf{z}$ lies in the column space of $\mathbb{X}$ is zero. It follows that $\mathbb{Y}_c^T (\mathbf{I} - \mathbf{P}_{\mathbb{X}})\mathbb{Y}_c$ is positive definite with probability 1. □

By Lemma 2, if $n > \max(p, r)$, it is safe to assume that the specific observed values of $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ (and the corresponding $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$) satisfy the property that $\mathbb{Y}_c^T (\mathbf{I} - \mathbf{P}_{\mathbb{X}})\mathbb{Y}_c$ is positive definite. With the above lemma in hand, we now commence with the main proof.

Let A denote the integral of the unnormalized density on the right-hand side of (12) (over $M_{r,1} \times M_{u,p} \times S_{r,r}^+ \times O_u \times O_{r-u}$) with respect to $\nu_1 \times \nu_2 \times \nu_3 \times \nu_4 \times \nu_5$. We will prove that A is finite under the conditions in Theorem 1.

Note that

$$
\begin{aligned}
&\operatorname{tr}\{(\mathbb{Y} - \mathbf{1}_n \boldsymbol{\mu}^T - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0 \boldsymbol{\Gamma}_0^T)^{-1} (\mathbb{Y} - \mathbf{1}_n \boldsymbol{\mu}^T - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)^T\} \\
&\quad = \operatorname{tr}\{(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0 \boldsymbol{\Gamma}_0^T)^{-1} (\mathbb{Y} - \mathbf{1}_n \boldsymbol{\mu}^T - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)^T \\
&\qquad \times (\mathbb{Y} - \mathbf{1}_n \boldsymbol{\mu}^T - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)\} \\
&\quad = \operatorname{tr}\{(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0 \boldsymbol{\Gamma}_0^T)^{-1} (\mathbb{Y} - \mathbf{1}_n \boldsymbol{\mu}^T - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)^T \\
&\qquad \times (\mathbf{P}_{\mathbf{1}_n} + \mathbf{Q}_{\mathbf{1}_n})(\mathbb{Y} - \mathbf{1}_n \boldsymbol{\mu}^T - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)\} \\
&\quad = \operatorname{tr}\{(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0 \boldsymbol{\Gamma}_0^T)^{-1} \\
&\qquad \times [n(\bar{\mathbf{Y}} - \boldsymbol{\mu})(\bar{\mathbf{Y}} - \boldsymbol{\mu})^T + (\mathbb{Y} - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)^T \mathbf{Q}_{\mathbf{1}_n} (\mathbb{Y} - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)]\} \qquad (13) \\
&\quad = \operatorname{tr}\{(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0 \boldsymbol{\Gamma}_0^T)^{-1} \\
&\qquad \times [n(\boldsymbol{\mu} - \bar{\mathbf{Y}})(\boldsymbol{\mu} - \bar{\mathbf{Y}})^T + (\mathbb{Y}_c - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)^T (\mathbb{Y}_c - \mathbb{X}\boldsymbol{\eta}^T \boldsymbol{\Gamma}^T)]\} \\
&\quad = \operatorname{tr}\{n(\boldsymbol{\mu} - \bar{\mathbf{Y}})^T (\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0 \boldsymbol{\Gamma}_0^T)^{-1} (\boldsymbol{\mu} - \bar{\mathbf{Y}})\} \\
&\qquad + \operatorname{tr}\{(\mathbb{Y}_c \boldsymbol{\Gamma} - \mathbb{X}\boldsymbol{\eta}^T) \boldsymbol{\Omega}^{-1} (\mathbb{Y}_c \boldsymbol{\Gamma} - \mathbb{X}\boldsymbol{\eta}^T)^T\} + \operatorname{tr}(\mathbb{Y}_c \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0^{-1} \boldsymbol{\Gamma}_0^T \mathbb{Y}_c^T).
\end{aligned}
$$

Note that under the conditions in Theorem 1, $\mathbb{X}^T \mathbb{X} + \mathbf{C}$ is a positive definite matrix. Hence,

$$
\begin{aligned}
&\operatorname{tr}((\mathbb{Y}_c \boldsymbol{\Gamma} - \mathbb{X}\boldsymbol{\eta}^T) \boldsymbol{\Omega}^{-1} (\mathbb{Y}_c \boldsymbol{\Gamma} - \mathbb{X}\boldsymbol{\eta}^T)^T) + \operatorname{tr}(\boldsymbol{\Omega}^{-1} (\boldsymbol{\eta} - \boldsymbol{\Gamma}^T \mathbf{e}) \mathbf{C} (\boldsymbol{\eta} - \boldsymbol{\Gamma}^T \mathbf{e})^T) \\
&\quad = \operatorname{tr}(\boldsymbol{\Omega}^{-1} (\mathbb{Y}_c \boldsymbol{\Gamma} - \mathbb{X}\boldsymbol{\eta}^T)^T (\mathbb{Y}_c \boldsymbol{\Gamma} - \mathbb{X}\boldsymbol{\eta}^T)) \\
&\qquad + \operatorname{tr}(\boldsymbol{\Omega}^{-1} (\boldsymbol{\eta} - \boldsymbol{\Gamma}^T \mathbf{e}) \mathbf{C} (\boldsymbol{\eta} - \boldsymbol{\Gamma}^T \mathbf{e})^T) \qquad (14)
\end{aligned}
$$

$$
\begin{aligned}
&= \operatorname{tr}\big(\boldsymbol{\Omega}^{-1}\big(\boldsymbol{\eta}\mathbb{X}^T\mathbb{X}\boldsymbol{\eta}^T - 2\boldsymbol{\eta}\mathbb{X}^T\mathbb{Y}_c\boldsymbol{\Gamma} + \boldsymbol{\Gamma}^T\mathbb{Y}_c^T\mathbb{Y}_c\boldsymbol{\Gamma}\big)\big)\\
&\quad + \operatorname{tr}\big(\boldsymbol{\Omega}^{-1}\big(\boldsymbol{\eta}\mathbf{C}\boldsymbol{\eta}^T - 2\boldsymbol{\eta}\mathbf{C}\mathbf{e}^T\boldsymbol{\Gamma} + \boldsymbol{\Gamma}^T\mathbf{e}\mathbf{C}\mathbf{e}^T\boldsymbol{\Gamma}\big)\big)\\
&= \operatorname{tr}\big(\boldsymbol{\Omega}^{-1}\big(\boldsymbol{\eta} - \boldsymbol{\Gamma}^T\tilde{\mathbf{e}}\big)\big(\mathbb{X}^T\mathbb{X} + \mathbf{C}\big)\big(\boldsymbol{\eta} - \boldsymbol{\Gamma}^T\tilde{\mathbf{e}}\big)^T\big) + \operatorname{tr}\big(\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\tilde{\mathbf{G}}\boldsymbol{\Gamma}\big),
\end{aligned}
$$

where

$$
\tilde{\mathbf{e}} = \big(\mathbb{Y}_c^T\mathbb{X} + \mathbf{e}\mathbf{C}\big)\big(\mathbb{X}^T\mathbb{X} + \mathbf{C}\big)^{-1}
$$

and

$$
\tilde{\mathbf{G}} = \mathbb{Y}_c^T\mathbb{Y}_c + \mathbf{e}\mathbf{C}\mathbf{e}^T - \tilde{\mathbf{e}}\big(\mathbb{X}^T\mathbb{X} + \mathbf{C}\big)\tilde{\mathbf{e}}^T.
$$

Since

$$
\begin{aligned}
&\int_{M_{r,1}} e^{(-1/2)\operatorname{tr}\{n(\boldsymbol{\mu}-\bar{\mathbf{Y}})^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)^{-1}(\boldsymbol{\mu}-\bar{\mathbf{Y}})\}}\nu_1(d\boldsymbol{\mu})\\
&\qquad = (2\pi)^{r/2}n^{r/2}|\boldsymbol{\Omega}|^{1/2}|\boldsymbol{\Omega}_0|^{1/2},
\end{aligned}
$$

it follows from (12) and (14) that

$$
\begin{aligned}
A = \int_{S_{r,r}^+\times O_u\times O_{r-u}} &\Bigg(\int_{M_{u,p}} e^{(-1/2)\operatorname{tr}\{\boldsymbol{\Omega}^{-1}(\boldsymbol{\eta}-\boldsymbol{\Gamma}^T\tilde{\mathbf{e}})(\mathbb{X}^T\mathbb{X}+\mathbf{C})(\boldsymbol{\eta}-\boldsymbol{\Gamma}^T\tilde{\mathbf{e}})^T\}}\nu_2(d\boldsymbol{\eta})\Bigg)\\
&\times (2\pi)^{-(n-1)r/2}n^{r/2}|\boldsymbol{\Omega}|^{-(n+p-1)/2}|\boldsymbol{\Omega}_0|^{-(n-1)/2}\\
&\times e^{(-1/2)\operatorname{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\tilde{\mathbf{G}}\boldsymbol{\Gamma})}e^{(-1/2)\operatorname{tr}(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T(\mathbb{Y}_c^T\mathbb{Y}_c)\boldsymbol{\Gamma}_0)}\\
&\times \prod_{i=1}^{u}\omega_i^{-\alpha-1}e^{-\lambda/\omega_i}\prod_{i=1}^{r-u}\omega_{0,i}^{-\alpha_0-1}\\
&\times e^{-\lambda_0/\omega_{0,i}}e^{(-1/2)\operatorname{tr}(\mathbf{D}^{-1}\mathbf{O}^T\mathbf{G}\mathbf{O})}\nu_3\big(d(\boldsymbol{\Gamma},\boldsymbol{\Gamma}_0)\big)\nu_4(d\boldsymbol{\omega})\nu_5(d\boldsymbol{\omega}_0) \qquad (15)\\
= \int_{S_{r,r}^+\times O_u\times O_{r-u}} &\frac{n^{r/2}|\boldsymbol{\Omega}|^{-(n-1)/2}|\boldsymbol{\Omega}_0|^{-(n-1)/2}|(\mathbb{X}^T\mathbb{X}+\mathbf{C})|^{-u/2}}{(\sqrt{2\pi})^{r(n-1)-up}}\\
&\times e^{(-1/2)\operatorname{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\tilde{\mathbf{G}}\boldsymbol{\Gamma})}e^{(-1/2)\operatorname{tr}(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T(\mathbb{Y}_c^T\mathbb{Y}_c)\boldsymbol{\Gamma}_0)}\\
&\times \prod_{i=1}^{u}\omega_i^{-\alpha-1}e^{-\lambda/\omega_i}\prod_{i=1}^{r-u}\omega_{0,i}^{-\alpha_0-1}\\
&\times e^{-\lambda_0/\omega_{0,i}}e^{(-1/2)\operatorname{tr}(\mathbf{D}^{-1}\mathbf{O}^T\mathbf{G}\mathbf{O})}\nu_3\big(d(\boldsymbol{\Gamma},\boldsymbol{\Gamma}_0)\big)\nu_4(d\boldsymbol{\omega})\nu_5(d\boldsymbol{\omega}_0).
\end{aligned}
$$

Since $O_k \subset \mathbb{R}_+^k$ for every $k \in \mathbb{N}$, we get from (15) that

$$
\begin{aligned}
A \le \int_{S_{r,r}^+\times\mathbb{R}_u^+\times\mathbb{R}_{r-u}^+} &\frac{n^{r/2}|\boldsymbol{\Omega}|^{-(n-1)/2}|\boldsymbol{\Omega}_0|^{-(n-1)/2}|(\mathbb{X}^T\mathbb{X}+\mathbf{C})|^{-u/2}}{(\sqrt{2\pi})^{r(n-1)-up}}\\
&\times e^{(-1/2)\operatorname{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\tilde{\mathbf{G}}\boldsymbol{\Gamma})}e^{(-1/2)\operatorname{tr}(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T(\mathbb{Y}_c^T\mathbb{Y}_c)\boldsymbol{\Gamma}_0)} \qquad (16)
\end{aligned}
$$

$$\times \prod_{i=1}^{u} \omega_i^{-\alpha-1} e^{-\lambda/\omega_i} \prod_{i=1}^{r-u} \omega_{0,i}^{-\alpha_0-1}$$
$$\times e^{-\lambda_0/\omega_{0,i}} e^{(-1/2)\operatorname{tr}(\mathbf{D}^{-1}\mathbf{O}^T\mathbf{G}\mathbf{O})} v_3\big(d(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)\big)\, d\boldsymbol{\omega}\, d\boldsymbol{\omega}_0.$$

It can be easily checked that the matrices

$$M_1 = \begin{bmatrix} \mathbb{Y}_c^T\mathbb{Y}_c & \mathbb{Y}_c^T\mathbb{X} \\ \mathbb{X}^T\mathbb{Y}_c & \mathbb{X}^T\mathbb{X} \end{bmatrix} \quad \text{and} \quad M_2 = \begin{bmatrix} \mathbf{e}\mathbf{C}\mathbf{e}^T & \mathbf{e}\mathbf{C} \\ \mathbf{C}\mathbf{e}^T & \mathbf{C} \end{bmatrix}$$

are nonnegative definite. If $n > \max(r, p)$, then

$$\begin{vmatrix} \mathbb{Y}_c^T\mathbb{Y}_c & \mathbb{Y}_c^T\mathbb{X} \\ \mathbb{X}^T\mathbb{Y}_c & \mathbb{X}^T\mathbb{X} \end{vmatrix} = |\mathbb{X}^T\mathbb{X}|\big|\mathbb{Y}_c^T(\mathbf{I} - \mathbf{P}_{\mathbb{X}})\mathbb{Y}_c\big| > 0.$$

Hence, the matrix M_1 is positive definite if $n > p$. It follows that the matrix

$$M_1 + M_2 = \begin{bmatrix} \mathbb{Y}_c^T\mathbb{Y}_c + \mathbf{e}\mathbf{C}\mathbf{e}^T & \mathbb{Y}_c^T\mathbb{X} + \mathbf{e}\mathbf{C} \\ \mathbb{X}^T\mathbb{Y}_c + \mathbf{C}\mathbf{e}^T & \mathbb{X}^T\mathbb{X} + \mathbf{C} \end{bmatrix}$$

is nonnegative definite [and positive definite if $n > \max(r, p)$]. Since $\mathbb{X}^T\mathbb{X} + \mathbf{C}$ is a positive definite matrix under the conditions in Theorem 1, it follows by the properties of block partitioned nonnegative definite matrices that the matrix

$$\begin{aligned} &\mathbb{Y}_c^T\mathbb{Y}_c + \mathbf{e}\mathbf{C}\mathbf{e}^T - (\mathbb{Y}_c^T\mathbb{X} + \mathbf{e}\mathbf{C})(\mathbb{X}^T\mathbb{X} + \mathbf{C})^{-1}(\mathbb{X}^T\mathbb{Y}_c + \mathbf{C}\mathbf{e}^T) \\ &\quad = \mathbb{Y}_c^T\mathbb{Y}_c + \mathbf{e}\mathbf{C}\mathbf{e}^T - \tilde{\mathbf{e}}(\mathbb{X}^T\mathbb{X} + \mathbf{C})\tilde{\mathbf{e}}^T \\ &\quad = \tilde{\mathbf{G}} \end{aligned}$$

is nonnegative definite [and positive definite if $n > \max(r, p)$]. It follows that under the conditions in Theorem 1, the matrices $\tilde{\mathbf{G}} + 2\lambda\mathbf{I}_r$ and $\mathbb{Y}_c^T\mathbb{Y}_c + 2\lambda_0\mathbf{I}_r$ are positive definite matrices.

Note that

$$|\boldsymbol{\Omega}| = \prod_{i=1}^{u} \omega_i, \qquad |\boldsymbol{\Omega}_0| = \prod_{i=1}^{r-u} \omega_{0,i}, \tag{17}$$

$$\operatorname{tr}\big(\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\tilde{\mathbf{G}}\boldsymbol{\Gamma}\big) = \sum_{i=1}^{u} \frac{(\boldsymbol{\Gamma}^T\tilde{\mathbf{G}}\boldsymbol{\Gamma})_{ii}}{\omega_i}, \tag{18}$$

$$\operatorname{tr}\big(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T(\mathbb{Y}_c^T\mathbb{Y}_c)\boldsymbol{\Gamma}_0\big) = \sum_{i=1}^{r-u} \frac{(\boldsymbol{\Gamma}_0^T(\mathbb{Y}_c^T\mathbb{Y}_c)\boldsymbol{\Gamma}_0)_{ii}}{\omega_{0,i}}. \tag{19}$$

Note that for $a, b > 0$,

$$\int_0^\infty x^{-a-1} e^{-b/2x}\, dx = b^{-a} K_a,$$

where K_a is a finite constant which depends only on a. It follows that

$$\begin{aligned}&\int_{\mathbb{R}_u^+} |\boldsymbol{\Omega}|^{-(n-1)/2} e^{(-1/2)\operatorname{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\tilde{\mathbf{G}}\boldsymbol{\Gamma})} \left(\prod_{i=1}^{u} \omega_i^{-\alpha-1} e^{-\lambda/\omega_i}\right) d\boldsymbol{\omega} \\ &\quad = (K_{(n+2\alpha-1)/2})^u \prod_{i=1}^{u} \left(\boldsymbol{\Gamma}^T\tilde{\mathbf{G}}\boldsymbol{\Gamma} + 2\lambda \mathbf{I}_u\right)_{ii}^{-(n+2\alpha-1)/2},\end{aligned} \tag{20}$$

and

$$\begin{aligned}&\int_{\mathbb{R}_{r-u}^+} |\boldsymbol{\Omega}_0|^{-(n-1)/2} e^{(-1/2)\operatorname{tr}(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T(\mathbb{Y}_c^T\mathbb{Y}_c)\boldsymbol{\Gamma}_0)} \\ &\qquad \times \left(\prod_{i=1}^{r-u} \omega_{0,i}^{-\alpha_0-1} e^{-\lambda_0/\omega_{0,i}}\right) d\boldsymbol{\omega}_0 \\ &\quad = (K_{(n+2\alpha_0-1)/2})^{r-u} \prod_{i=1}^{r-u} \left(\boldsymbol{\Gamma}_0^T(\mathbb{Y}_c^T\mathbb{Y}_c)\boldsymbol{\Gamma}_0 + 2\lambda_0 \mathbf{I}_{r-u}\right)_{ii}^{-(n+2\alpha_0-1)/2}.\end{aligned} \tag{21}$$

Since $\mathbf{G}$ is positive semi-definite, and $\mathbf{D}$ is a diagonal matrix with positive diagonal entries, we get that

$$e^{(-1/2)\operatorname{tr}(\mathbf{D}^{-1}\mathbf{O}^T\mathbf{G}\mathbf{O})} \le 1$$

for every $\mathbf{O} \in S_{r,r}^+$. It follows from (16), (20) and (21) that there exists a finite constant A_0 such that

$$\begin{aligned}A \le A_0 &\int_{S_{r,r}^+} \prod_{i=1}^{u} \left(\boldsymbol{\Gamma}^T\tilde{\mathbf{G}}\boldsymbol{\Gamma} + 2\lambda \mathbf{I}_u\right)_{ii}^{-(n+2\alpha-1)/2} \\ &\times \prod_{i=1}^{r-u} \left(\boldsymbol{\Gamma}_0^T(\mathbb{Y}_c^T\mathbb{Y}_c + \mathbf{G})\boldsymbol{\Gamma}_0 + 2\lambda_0 \mathbf{I}_{r-u}\right)_{ii}^{-(n+2\alpha_0-1)/2} \nu_3\big(d(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)\big).\end{aligned}$$

Let $\mathbf{e}_i \in \mathbb{R}^u$ be the vector with ith entry equal to 1, and all other entries equal to 0. Note that $\mathbf{e}_i^T\mathbf{e}_i = 1$. Then for every $\boldsymbol{\Gamma} \in S_{r,u}^+$, it follows that $(\boldsymbol{\Gamma}\mathbf{e}_i)^T(\boldsymbol{\Gamma}\mathbf{e}_i) = \mathbf{e}_i^T\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}\mathbf{e}_i = \mathbf{e}_i^T\mathbf{e}_i = 1$. Hence,

$$\begin{aligned}\left(\boldsymbol{\Gamma}^T\tilde{\mathbf{G}}\boldsymbol{\Gamma}\right)_{ii} &= \mathbf{e}_i^T\boldsymbol{\Gamma}^T\tilde{\mathbf{G}}\boldsymbol{\Gamma}\mathbf{e}_i \\ &= (\boldsymbol{\Gamma}\mathbf{e}_i)^T\tilde{\mathbf{G}}(\boldsymbol{\Gamma}\mathbf{e}_i) \\ &\ge \lambda_{\min}(\tilde{\mathbf{G}}),\end{aligned}$$

where $\lambda_{\min}(\tilde{\mathbf{G}})$ denotes the smallest eigenvalue of $\tilde{\mathbf{G}}$. Similarly, for every $\boldsymbol{\Gamma}_0 \in S_{r,r-u}^+$,

$$\left(\boldsymbol{\Gamma}_0^T(\mathbb{Y}_c^T\mathbb{Y}_c)\boldsymbol{\Gamma}_0\right)_{ii} \ge \lambda_{\min}\left(\mathbb{Y}_c^T\mathbb{Y}_c\right).$$

Since ν_3 is a probability measure on $S_{r,r}^+$, it follows that

$$\begin{aligned} A &\le A_0\big(\lambda_{\min}(\tilde{\mathbf{G}}) + 2\lambda\big)^{-u(n+2\alpha-1)/2}\big(\lambda_{\min}\big(\mathbb{Y}_c^T\mathbb{Y}_c\big) + 2\lambda_0\big)^{-(r-u)(n+2\alpha_0-1)/2} \\ &\quad \times \int_{S_{r,r}^+} \nu_3\big(d(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)\big) \\ &\le A_0\big(\lambda_{\min}(\tilde{\mathbf{G}}) + 2\lambda\big)^{-u(n+2\alpha-1)/2}\big(\lambda_{\min}\big(\mathbb{Y}_c^T\mathbb{Y}_c\big) + 2\lambda_0\big)^{-(r-u)(n+2\alpha_0-1)/2} \\ &< \infty. \end{aligned}$$

5.3. *Proof of Theorem 2.* In order to prove Theorem 2, we first prove a general mathematical result about orthogonal matrices.

LEMMA 3. *Let*

$$\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_r) \in \mathbb{R}^{r\times r}$$

be an orthogonal matrix, and let

$$\mathbf{I} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_r) \in \mathbb{R}^{r\times r}$$

be an identity matrix. Let $\mathbf{R}_{i,j,\theta} \in \mathbb{R}^{r\times r}$ *be an identity matrix except its elements* (i,i), (i,j), (j,i) *and* (j,j) *are* $\sin(\theta)$, $-\cos(\theta)$, $\cos(\theta)$ *and* $\sin(\theta)$. *Define the set* $\mathcal{R} = \{\mathbf{R}_{i,j,\theta} : 0 \le \theta < 2\pi, 1 \le i < j \le n\}$. *Then there exists a matrix* $\mathbf{M}$ *such that* $\mathbf{OM} = \mathbf{I}$ *and* $\mathbf{M}$ *is a multiplication of* r *members in* $\mathcal{R}$.

PROOF. We will use mathematical induction for the proof of the above statement. When $r = 3$, we perform the following procedure:

1. Rotate $(\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3)$ to $(\mathbf{o}_1^{(1)}, \mathbf{o}_2^{(1)}, \mathbf{o}_3^{(1)})$ such that $\mathbf{o}_3^{(1)} = \mathbf{o}_3$ and $\mathbf{o}_2^{(1)} \in \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$.
2. Rotate $(\mathbf{o}_1^{(1)}, \mathbf{o}_2^{(1)}, \mathbf{o}_3^{(1)})$ to $(\mathbf{o}_1^{(2)}, \mathbf{o}_2^{(2)}, \mathbf{o}_3^{(2)})$ such that $\mathbf{o}_2^{(2)} = \mathbf{o}_2^{(1)}$ and $\mathbf{o}_3^{(2)} \in \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$.
3. After Step 1 and Step 2, both $\mathbf{o}_2^{(2)}$ and $\mathbf{o}_3^{(2)}$ are in $\operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$, then we must have $\mathbf{o}_1^{(2)} = \mathbf{e}_1$. Now we rotate $(\mathbf{o}_2^{(2)}, \mathbf{o}_3^{(2)})$ to align with $(\mathbf{e}_2, \mathbf{e}_3)$.

From the preceding procedure, notice that we always rotate two vectors at a time. Now we prove the validity of Step 1, the validity of Step 2 and Step 3 follows similarly. In Step 1, there exists a unit length vector $\mathbf{a}$ such that $\mathbf{a} \in \operatorname{span}(\mathbf{o}_1, \mathbf{o}_2) \cap \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$. If not, then we must have $\operatorname{span}(\mathbf{o}_1, \mathbf{o}_2) \perp \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$, as $\dim(\operatorname{span}(\mathbf{o}_1, \mathbf{o}_2)) = 2$ and $\dim(\operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)) = 2$, it cannot happen when $r = 3$. Let $\mathbf{o}_2^{(1)} = \mathbf{a}$, then $\mathbf{o}_2^{(1)} \in \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$. As we also have $\mathbf{o}_2^{(1)} \in \operatorname{span}(\mathbf{o}_1, \mathbf{o}_2)$, there exists a rotation presented by $\mathbf{R}_{1,2,\theta_1}$ such that $\mathbf{o}_2\mathbf{R}_{1,2,\theta_1} = \mathbf{a}$. Notice that $\mathbf{o}_3\mathbf{R}_{1,2,\theta_1} = \mathbf{o}_3$. Let $\mathbf{o}_1^{(1)} = \mathbf{o}_1\mathbf{R}_{1,2,\theta_1}$. Then we have $(\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3)\mathbf{R}_{1,2,\theta_1} = (\mathbf{o}_1^{(1)}, \mathbf{o}_2^{(1)}, \mathbf{o}_3^{(1)})$. In Step 2, the rotation is presented as right multiplying $(\mathbf{o}_1^{(1)}, \mathbf{o}_2^{(1)}, \mathbf{o}_3^{(1)})$ by $\mathbf{R}_{1,3,\theta_2}$,

and similarly in Step 3, $\mathbf{R}_{2,3,\theta_3}$ is right multiplied to $(\mathbf{o}_1^{(2)}, \mathbf{o}_2^{(2)}, \mathbf{o}_3^{(2)})$. Therefore, the statement holds for $r = 3$.

Suppose the statement holds for $r = k-1$, $k \geq 4$. When $r = k$, we perform the following procedure:

1. Rotate $(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_k)$ to $(\mathbf{o}_1^{(1)}, \mathbf{o}_2^{(1)}, \ldots, \mathbf{o}_k^{(1)})$ such that $\mathbf{o}_i^{(1)} = \mathbf{o}_i$, $i \geq 3$ and $\mathbf{o}_2^{(1)} \in \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_k)$.
2. Rotate $(\mathbf{o}_1^{(1)}, \mathbf{o}_2^{(1)}, \ldots, \mathbf{o}_k^{(1)})$ to $(\mathbf{o}_1^{(2)}, \mathbf{o}_2^{(2)}, \ldots, \mathbf{o}_k^{(2)})$ such that $\mathbf{o}_i^{(2)} = \mathbf{o}_i^{(1)}$ for $i = 2, 4, 5, \ldots, k$, and $\mathbf{o}_3^{(2)} \in \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_k)$.
3. ....

We follow the pattern in Step 1 and Step 2 for the other steps. At Step i, $i = 3, \ldots, k-1$, we keep all the $\mathbf{o}_j^{(i-1)}$'s (j runs from 1 to k except 1 and $i+1$) fixed and only rotate $\mathbf{o}_1^{(i-1)}$ and $\mathbf{o}_{i+1}^{(i-1)}$ such that $\mathbf{o}_{i+1}^{(i)} \in \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_k)$.

Each of the preceding step is valid. We take Step 1 as an example. Because $\dim(\operatorname{span}(\mathbf{o}_1, \mathbf{o}_2)) = 2$, $\dim(\operatorname{span}(\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_k)) = k-1$, and the space is k dimensional, there exists a unit length vector $\mathbf{a}$ such that $\mathbf{a} \in \operatorname{span}(\mathbf{o}_1, \mathbf{o}_2) \cap \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_k)$. Let $\mathbf{o}_2^{(1)} = \mathbf{a}$, then $\mathbf{o}_2^{(1)} \in \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_k)$. As $\mathbf{o}_2^{(1)} \in \operatorname{span}(\mathbf{o}_1, \mathbf{o}_2)$, there exists a rotation presented by $\mathbf{R}_{1,2,\theta_1}$ such that $\mathbf{o}_2 \mathbf{R}_{1,2,\theta_1} = \mathbf{a}$. Notice that $\mathbf{o}_i \mathbf{R}_{1,2,\theta_1} = \mathbf{o}_i$ for $i \geq 3$. Let $\mathbf{o}_1^{(1)} = \mathbf{o}_1 \mathbf{R}_{1,2,\theta_1}$. Then we have $(\mathbf{o}_1^{(1)}, \mathbf{o}_2^{(1)}, \ldots, \mathbf{o}_k^{(1)}) = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_k)\mathbf{R}_{1,2,\theta_1}$. The validity for the rest of the steps results from the same argument.

After all the preceding steps, we have $\mathbf{o}_j^{(k-1)} \in \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_k)$ for $j = 2, \ldots, k$. Then we must have $\mathbf{o}_1^{(k-1)} = \mathbf{e}_1$. As $\operatorname{span}(\mathbf{o}_2^{(k-1)}, \mathbf{o}_3^{(k-1)}, \ldots, \mathbf{o}_k^{(k-1)}) = \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_k)$, this is a $k-1$ dimensional subspace and the first element of all $\mathbf{o}_j^{(k-1)}$, $j = 2, \ldots, k$, is 0. Using the result for $r = k-1$, the statement also holds for $r = k$. By mathematical induction, the statement holds. □

We now make some observations.

(a) Note that sampling from the conditional distribution of $[\mathbf{O}_{\cdot i} : \mathbf{O}_{\cdot j}]$ specified in (C.5), (C.8) or (C.9) corresponds to multiplying by a matrix of the form $\mathbf{R}_{i,j,\theta}$ considered in Lemma 3, and that for every $1 \leq i < j \leq r$, $\mathbf{R}_{i,j,0} = \mathbf{I}_r$.

(b) For every $1 \leq i \leq u$ and $1 \leq j \leq r-u$, the conditional density of ω_i in (C.3) and $\omega_{0,j}$ in (C.4) is strictly positive on $(\omega_{i-1}, \omega_{i+1})$ and $(\omega_{0,j-1}, \omega_{0,j+1})$, respectively.

(c) Let $\boldsymbol{\omega}, \widetilde{\boldsymbol{\omega}} \in O_u$ be arbitrarily chosen. Let $c = \min(\omega_1, \widetilde{\omega}_1)$, and $\boldsymbol{\omega}^* = (c/(u-i+2))_{i=1}^u \in O_u$. We now illustrate the following method of moving from $\boldsymbol{\omega}$ to $\widetilde{\boldsymbol{\omega}}$. Note that $0 < \boldsymbol{\omega}_1^* < \boldsymbol{\omega}_2$, and $\boldsymbol{\omega}_{i-1}^* < \boldsymbol{\omega}_i^* < \boldsymbol{\omega}_{i+1}$. Hence, by (b), one can move from $\boldsymbol{\omega}$ to $\boldsymbol{\omega}^*$ as part of one step of the systematic or random scan Gibbs sampling chain. Now consider keeping all the entries of $\boldsymbol{\omega}^*$ the same, except changing the last entry to $\widetilde{\omega}_u$. Since $\omega_{u-1}^* < \widetilde{\omega}_u$, it follows by (b) that this move

can be achieved as part of one step of both the Gibbs sampling algorithms. Since $\omega^*_{u-2} < \widetilde{\omega}_{u-1} < \widetilde{\omega}_u$, it again follows by observation (b) that we can make a move which keeps all entries of the current vector same, except changing the $(u-1)$th entry to $\widetilde{\omega}_{u-1}$, as part of one step of both Gibbs sampling chain. Continuing this process, and combining the above arguments, it can thus be shown that it is possible to move from $\boldsymbol{\omega}$ to $\widetilde{\boldsymbol{\omega}}$ as part of $u+1$ steps of both Gibbs sampling Markov chains.

(d) Let $\boldsymbol{\omega}_0, \widetilde{\boldsymbol{\omega}}_0 \in O_{r-u}$ be arbitrarily chosen. By exactly the same arguments as those used in (c), it can be shown that it is possible to move from $\boldsymbol{\omega}_0$ to $\widetilde{\boldsymbol{\omega}}_0$ as part of $r-u+1$ steps of both Gibbs sampling Markov chains.

It follows by observations (a) and (b) above, and the continuity of all the conditional densities involved (in the interior of their respective supports) in the Gibbs sampling algorithms that the 1-step Markov transition density of staying at the current value is strictly positive for both the systematic scan and random scan Gibbs samplers considered in Theorem 2. Also, by Lemma 3, and observations (a), (b), (c) and (d) above, that the $r+1$-step Markov transition densities of both the systematic scan and random scan Gibbs samplers considered in Lemma 2 are strictly positive everywhere on $O_u \times O_{r-u} \times S^+_{r,r}$. This implies that both Markov chains are irreducible and aperiodic. As noted earlier, both Markov chains have the density in (C.2) as a stationary density. It follows that both the systematic scan and random scan Gibbs samplers in Theorem 2 are Harris ergodic (see [1]).

SUPPLEMENTARY MATERIAL

Supplement to "A Bayesian approach for envelope models" (DOI: 10.1214/16-AOS1449SUPP; .pdf). The supplement [9] provides additional details and proofs for many of the results in the authors' paper.

REFERENCES

[1] ASMUSSEN, S. and GLYNN, P. W. (2011). A new proof of convergence of MCMC via the ergodic theorem. *Statist. Probab. Lett.* **81** 1482–1485. MR2818658

[2] BINGHAM, C. (1974). An antipodally symmetric distribution on the sphere. *Ann. Statist.* **2** 1201–1225. MR0397988

[3] CONWAY, J. B. (1990). *A Course in Functional Analysis*, 2nd ed. *Graduate Texts in Mathematics* **96**. Springer, New York. MR1070713

[4] COOK, R. D., HELLAND, I. S. and SU, Z. (2013). Envelopes and partial least squares regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 851–877. MR3124794

[5] COOK, R. D., LI, B. and CHIAROMONTE, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica* **20** 927–960. MR2729839

[6] COOK, R. D. and SU, Z. (2013). Scaled envelopes: Scale-invariant and efficient estimation in multivariate linear regression. *Biometrika* **100** 939–954. MR3142342

[7] GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2027492

[8] HOFF, P. D. (2009). Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *J. Comput. Graph. Statist.* **18** 438–456. MR2749840

[9] KHARE, K., PAL, S. and SU, Z. (2016). Supplement to "A Bayesian approach for envelope models." DOI:10.1214/16-AOS1449SUPP.

[10] MAO, K., LIANG, F. and MUKHERJEE, S. (2010). Supervised dimension reduction using Bayesian mixture modeling. In *International Conference on Artificial Intelligence and Statistics* 501–508.

[11] REICH, B. J., BONDELL, H. D. and LI, L. (2011). Sufficient dimension reduction via Bayesian mixture modeling. *Biometrics* **67** 886–895. MR2829263

[12] ROSSI, P. E., ALLENBY, G. M. and MCCULLOCH, R. (2005). *Bayesian Statistics and Marketing*. Wiley, Chichester. MR2193403

[13] SAMEH, A. H. and WISNIEWSKI, J. A. (1982). A trace minimization algorithm for the generalized eigenvalue problem. *SIAM J. Numer. Anal.* **19** 1243–1259. MR0679663

[14] SCHOTT, J. R. (2013). On the likelihood ratio test for envelope models in multivariate linear regression. *Biometrika* **100** 531–537. MR3068454

[15] SU, Z. and COOK, D. (2012). Inner envelopes: Efficient estimation in multivariate linear regression. *Biometrika* **99** 687–702. MR2966778

[16] SU, Z. and COOK, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika* **98** 133–146. MR2804215

[17] SU, Z. and COOK, R. D. (2013). Estimation of multivariate means with heteroscedastic errors using envelope models. *Statist. Sinica* **23** 213–230. MR3076165

[18] TOKDAR, S. T., ZHU, Y. M. and GHOSH, J. K. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Anal.* **5** 319–344. MR2719655

DEPARTMENT OF STATISTICS
102 GRIFFIN-FLOYD HALL
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32605
USA
E-MAIL: zhihuasu@stat.ufl.edu