A Comprehensive Bayesian Framework for Envelope Models

Saptarshi Chakraborty
and
Zhihua Su
Department of Biostatistics, University at Buffalo
and
Department of Statistics, University of Florida

August 9, 2023

Abstract

The envelope model aims to increase efficiency in multivariate analysis by utilizing dimension reduction techniques. It has been used in many contexts including linear regression, generalized linear models, matrix/tensor variate regression, reduced rank regression, and quantile regression, and has shown the potential to provide substantial efficiency gains. Virtually all of these advances, however, have been made from a frequentist perspective, and the literature addressing envelope models from a Bayesian point of view is sparse. The objective of this paper is to propose a Bayesian framework that is applicable across various envelope model contexts. The proposed framework aids straightforward interpretation of model parameters and allows easy incorporation of prior information. We provide a simple block Metropolis-within-Gibbs MCMC sampler for practical implementations of our method. Simulations and data examples are included for illustration.

Keywords: Envelope model, sufficient dimension reduction, Metropolis-within-Gibbs MCMC sampler, Bayesian partial least squares, Harris ergodicity

1 Introduction

Enveloping (Cook, 2018) is a dimension reduction technique that aims to gain estimation efficiency in multivariate analysis. In the era of high-throughput technology, many contemporary datasets are high-dimensional and contain information that is immaterial to established analyses. The envelope model seeks to identify and remove such immaterial information, making the subsequent analysis more efficient, sometimes equivalent to taking thousands of additional observations. Because of its proven ability to gain efficiency, envelope models have been an active area of research over the past decade. Initially derived for multivariate linear regression model (Cook et al., 2010), envelope models have since been developed in many different contexts including generalized linear models (Cook and Zhang, 2015), partial least squares (PLS) (Cook et al., 2013), matrix or tensor variate regression (Ding and Cook, 2018; Li and Zhang, 2017), quantile regression (Ding et al., 2021), spatial regression (Rekabdarkolaee et al., 2019) and variable selection (Su et al., 2016).

Virtually all of these advances, however, have been made from a frequentist perspective, and the literature addressing envelope models from a Bayesian perspective is still sparse. This is due to the fact that a key parameter in envelope models resides on a Grassmann manifold. Prior elicitation on such a restricted topological space is extremely difficult. The only Bayesian approach proposed so far is due to Khare et al. (2017). There the authors first reparameterize the model to define the key parameter on a Stiefel manifold and then put a matrix Bingham prior (Bingham, 1974) on the parameter. An appealing feature of this approach is that prior information on the key parameter can be incorporated through the specification of the hyperparameters in the matrix Bingham distribution. However, the approach crucially depends on the specific form of the response envelope model to achieve conjugacy, which makes extension to other contexts difficult. Moreover, the Gibbs sam-

pler proposed therein for implementation of the model requires sampling from generalized matrix Bingham distributions and truncated inverse gamma distributions which make the algorithm computationally expensive, and thus slow in sizable problems.

The purpose of this paper is to formulate a Bayesian framework that is easy to implement and is applicable to envelope models arising in diverse contexts. The proposed approach is completely different from the approach of Khare et al. (2017) in that it is free of any manifold structure. Consequently, prior constructions and practical implementation are more straightforward. We demonstrate our approach through the response envelope model, predictor envelope model, and probit model as examples, each of which has its own distinct modeling flavor. It is of note that the proposed Bayesian predictor envelope model leads to a Bayesian development of partial least squares (PLS), owing to a connection between the predictor envelope model and PLS (Cook et al., 2013). For implementation we propose a simple block Metropolis-within-Gibbs MCMC samplers to draw samples from the target posteriors. The proposed samplers are shown to be Harris ergodic, which provides theoretical guarantees on the quality of the MCMC samples with sufficient number of iterations. The general techniques used in this paper could potentially be utilized in other problems with parameters defined on manifolds, including envelope models in other contexts and Bayesian sufficient dimension reduction.

2 Review of the Envelope Model

This section reviews the envelope model from a frequentist perspective. The envelope model was first developed for the multivariate linear regression model (Cook et al., 2010)

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

where $\boldsymbol{Y} \in \mathbb{R}^r$ is the multivariate response vector, $\boldsymbol{X} \in \mathbb{R}^p$ is the vector of predictors, $\boldsymbol{\mu} \in \mathbb{R}^r$ and $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ are unknown intercept and regression coefficients, and $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ are errors with zero mean and a positive definite covariance matrix $\boldsymbol{\Sigma}$.

The goal of the envelope model is to improve the efficiency in the estimation of $\boldsymbol{\beta}$ by utilizing the relationships among the response variables. It is based on the assumption that some linear combinations of the response variables do not depend on $\boldsymbol{X}$ and are immaterial to the estimation of $\boldsymbol{\beta}$. Specifically, let $\mathbb{P}_{\mathcal{E}}$ denote the projection onto a subspace $\mathcal{E} \subseteq \mathbb{R}^r$ and $\mathbb{Q}_{\mathcal{E}} = \boldsymbol{I}_r - \mathbb{P}_{\mathcal{E}}$, where $\boldsymbol{I}_r$ is the r-dimensional identity matrix. Then we can decompose the response vector $\boldsymbol{Y}$ into two parts, $\mathbb{P}_{\mathcal{E}}\boldsymbol{Y}$ and $\mathbb{Q}_{\mathcal{E}}\boldsymbol{Y}$. Suppose that these two parts satisfy the following two conditions: (a) $\mathbb{Q}_{\mathcal{E}}\boldsymbol{Y} \mid \boldsymbol{X} \sim \mathbb{Q}_{\mathcal{E}}\boldsymbol{Y}$, where $\sim$ means equal in distribution; (b) $\mathrm{cor}(\mathbb{P}_{\mathcal{E}}\boldsymbol{Y}, \mathbb{Q}_{\mathcal{E}}\boldsymbol{Y} \mid \boldsymbol{X}) = 0$. Then the distribution of $\mathbb{Q}_{\mathcal{E}}\boldsymbol{Y}$ is not affected directly by $\boldsymbol{X}$ or indirectly via its correlation with $\mathbb{P}_{\mathcal{E}}\boldsymbol{Y}$. We call $\mathbb{P}_{\mathcal{E}}\boldsymbol{Y}$ the *material part* of $\boldsymbol{Y}$ and $\mathbb{Q}_{\mathcal{E}}\boldsymbol{Y}$ the *immaterial part* of $\boldsymbol{Y}$. Conditions (a) and (b) are equivalent to the following two conditions on the parameters in (1): (I) $\mathrm{span}(\boldsymbol{\beta}) \subseteq \mathcal{E}$, and (II) $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 = \mathbb{P}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbb{P}_{\mathcal{E}} + \mathbb{Q}_{\mathcal{E}}\boldsymbol{\Sigma}\mathbb{Q}_{\mathcal{E}}$, i.e., $\mathcal{E}$ is a reducing subspace of $\boldsymbol{\Sigma}$ (Conway, 1990). The $\boldsymbol{\Sigma}$-envelope of $\boldsymbol{\beta}$, denoted by $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$, is defined formally as the smallest reducing subspace of $\boldsymbol{\Sigma}$ that contains $\mathrm{span}(\boldsymbol{\beta})$. We use u $(0 \le u \le r)$ to denote the dimension of the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$. Let $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ and $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)}$ be orthonormal bases for $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ and $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})^{\perp}$ respectively, where $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})^{\perp}$ denotes the orthogonal complement of $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$. Then by (I), $\boldsymbol{\beta}$ can be written as $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$ carries the coordinates of $\boldsymbol{\beta}$ with respect to $\boldsymbol{\Gamma}$. By (II), $\boldsymbol{\Sigma}$ has the structure: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, where $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ are positive definite matrices carrying the coordinates of $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$ respectively. The matrix $\boldsymbol{\Sigma}_1$ carries the variation of the material part and $\boldsymbol{\Sigma}_2$ carries the variation of the

immaterial part. Thus, the coordinate form of the envelope model can be written as

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

When $u = r$, the envelope model reduces to the standard linear regression model (1).

Cook et al. (2010) estimates the parameters by optimizing a normal likelihood. Note that $\boldsymbol{\Gamma}$ is not identifiable, and only $\text{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ is identifiable. So the estimation entails optimization on an $r \times u$ Grassmann manifold, which is defined as the set of all u-dimensional subspaces in an r-dimensional space. Once we have an estimator of the envelope subspace $\widehat{\mathcal{E}} = \widehat{\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})}$, the envelope estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \mathbb{P}_{\widehat{\mathcal{E}}}\,\widehat{\boldsymbol{\beta}}_{\text{OLS}}$, i.e., the projection of the OLS estimator $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ onto the estimated envelope subspace $\widehat{\mathcal{E}}$.

The formulation of the envelope model (2) is based on dimension reduction of the response vector $\boldsymbol{Y}$. Therefore we call (2) the *response envelope model* henceforth. The construction of envelope subspaces is flexible and can be based on the dimension reduction of predictor vector $\boldsymbol{X}$ or other objects as well. It can also be extended beyond linear regression. As examples, we will discuss the predictor envelope model in Section 4, and the envelope model in generalized linear regression in Section 5.

3 A New Bayesian Response Envelope Model

3.1 Formulation

To derive the new Bayesian response envelope model, we first consider a reparameterization of (2) that identifies the envelope subspace via a parameter defined on a Euclidean space (Ma and Zhu, 2013; Cook et al., 2016). Let $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ be an arbitrary orthonormal basis of $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$. Let $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ denote the matrices formed with the top u and the bottom $r-u$ rows

of $\boldsymbol{\Gamma}$, respectively. Without loss of generality, we assume that $\boldsymbol{\Gamma}_1$ is non-singular; otherwise we can permute the order of the elements in $\boldsymbol{Y}$ to achieve that. Then

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_1 \\ \boldsymbol{\Gamma}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_u \\ \boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_1^{-1} \end{pmatrix} \boldsymbol{\Gamma}_1 \equiv \begin{pmatrix} \boldsymbol{I}_u \\ \boldsymbol{A} \end{pmatrix} \boldsymbol{\Gamma}_1. \tag{3}$$

It can be shown that $\boldsymbol{A} \in \mathbb{R}^{(r-u)\times u}$ depends on $\boldsymbol{\Gamma}$ only through $\text{span}(\boldsymbol{\Gamma})$, and there is a one-to-one correspondence between $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ and $\boldsymbol{A}$ (Su et al., 2016).

Remark 3.1. The permutation of the elements of $\boldsymbol{Y}$ can be determined from any consistent estimator of the basis $\boldsymbol{\Gamma}$, for example, the maximum likelihood estimator of $\boldsymbol{\Gamma}$ or any of the four staring values discussed in Cook et al. (2016). Given a consistent estimator, say $\widehat{\boldsymbol{\Gamma}}$, we can apply the Gaussian elimination with partial pivoting to find the u rows in $\widehat{\boldsymbol{\Gamma}}$ that constitute a non-singular matrix, say rows $i_1, \ldots, i_u$. The elements in $\boldsymbol{Y}$ are then permuted such that rows $i_1, \ldots, i_u$ become new rows $1, \ldots, u$.

The key advantage of the above reparameterization lies in the identification of $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ and $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})^{\perp}$ with the Euclidean parameter $\boldsymbol{A}$ described as follows. Let

$$\boldsymbol{C}_{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{I}_u \\ \boldsymbol{A} \end{pmatrix} \in \mathbb{R}^{r\times u}, \quad \text{and} \quad \boldsymbol{D}_{\boldsymbol{A}} = \begin{pmatrix} -\boldsymbol{A}^T \\ \boldsymbol{I}_{r-u} \end{pmatrix} \in \mathbb{R}^{r\times(r-u)}. \tag{4}$$

Chen et al. (2020) show that if $\boldsymbol{C}_{\boldsymbol{A}}$ is a basis of $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ then $\boldsymbol{D}_{\boldsymbol{A}}$ is a basis of $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})^{\perp}$. Consequently $\boldsymbol{\Gamma}(\boldsymbol{A}) = \boldsymbol{C}_{\boldsymbol{A}}(\boldsymbol{C}_{\boldsymbol{A}}^T\boldsymbol{C}_{\boldsymbol{A}})^{-1/2}$ and $\boldsymbol{\Gamma}_0(\boldsymbol{A}) = \boldsymbol{D}_{\boldsymbol{A}}(\boldsymbol{D}_{\boldsymbol{A}}^T\boldsymbol{D}_{\boldsymbol{A}})^{-1/2}$ form orthonormal bases of $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ and $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})^{\perp}$ respectively. With this parameterization and letting $\boldsymbol{\beta} = \boldsymbol{\Gamma}(\boldsymbol{A})\boldsymbol{\eta}$, we can reformulate the response envelope model (2) as

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}(\boldsymbol{A})\boldsymbol{\eta}\boldsymbol{X} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 = \boldsymbol{\Gamma}(\boldsymbol{A})\boldsymbol{\Omega}\boldsymbol{\Gamma}^T(\boldsymbol{A}) + \boldsymbol{\Gamma}_0(\boldsymbol{A})\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T(\boldsymbol{A}). \tag{5}$$

Here, the uniqueness of the basis matrix $\boldsymbol{C}_{\boldsymbol{A}}$ ensures identifiability of $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$.

A justification of the model identifiability is discussed in Section C.1 of the Supplement and the difference between the reparameterization in (4) and the one used in Khare et al. (2017) is included in Section C.2 of the Supplement.

Remark 3.2. When $u = 0$, $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ is the trivial space $\{0\}$, $\boldsymbol{Y}$ is uncorrelated to $\boldsymbol{X}$, $\boldsymbol{\Gamma}_0(\boldsymbol{A}) = \boldsymbol{I}_r$ and $\boldsymbol{\Sigma} = \boldsymbol{\Omega}_0$. On the other hand, when $u = r$, then $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) = \mathbb{R}^r$, $\boldsymbol{\Gamma}(\boldsymbol{A}) = \boldsymbol{I}_r, \boldsymbol{\Sigma} = \boldsymbol{\Omega}$, and model (5) degenerates to a standard linear regression model.

Let $\mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the $d-$variate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. A parametric representation of model (5) is given by

$$\boldsymbol{Y} \mid \boldsymbol{X} \sim \mathrm{N}_r \left(\boldsymbol{\mu} + \boldsymbol{\Gamma}(\boldsymbol{A})\boldsymbol{\eta}\boldsymbol{X}, \boldsymbol{\Gamma}(\boldsymbol{A})\boldsymbol{\Omega}\boldsymbol{\Gamma}^T(\boldsymbol{A}) + \boldsymbol{\Gamma}_0(\boldsymbol{A})\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T(\boldsymbol{A}) \right), \tag{6}$$

Here $\boldsymbol{X}$ can be either deterministic, or stochastic with a distribution free of parameters $\boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0$ and $\boldsymbol{A}$. Suppose that we have n independent observations $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i), i = 1, \ldots, n\}$ from model (6). For notational convenience, we define $\mathbb{Y} \in \mathbb{R}^{n \times r}$ and $\mathbb{X} \in \mathbb{R}^{n \times p}$ as $\mathbb{Y}^T = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ and $\mathbb{X}^T = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, and $\mathbf{1}_n$ to denote the n-dimensional vector of 1's. Further, we let $\overline{\boldsymbol{Y}} = \mathbf{1}_n^T\mathbb{Y}/n$, $\overline{\boldsymbol{X}} = \mathbf{1}_n^T\mathbb{X}/n$, $\mathbb{X}_c = \mathbb{X} - \mathbf{1}_n\overline{\boldsymbol{X}}^T$ and $\mathbb{Y}_c = \mathbb{Y} - \mathbf{1}_n\overline{\boldsymbol{Y}}^T$.

3.2 Prior and Posterior Distributions

To assign prior distributions for the model parameters $\boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0$, and $\boldsymbol{A}$, we consider standard conjugate priors for analytical and computational tractability. Such priors involve the inverse Wishart and matrix normal distributions, which are briefly reviewed in Section A of the Supplementary document. We use $\mathbb{S}_+^{m \times m}$ to denote the set of all $m \times m$ symmetric positive definite matrices. The proposed prior density is assumed to be of the form $\pi(\boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{A}) = \pi(\boldsymbol{\mu})\pi(\boldsymbol{\eta} \mid \boldsymbol{A}, \boldsymbol{\Omega})\pi(\boldsymbol{\Omega})\pi(\boldsymbol{\Omega}_0)\pi(\boldsymbol{A})$, where

(i) $\pi(\boldsymbol{\mu}) \propto 1$ is an improper flat density (with respect to the Lebesgue measure on $\mathbb{R}^r$).

(ii) $\pi(\boldsymbol{\Omega})$ and $\pi(\boldsymbol{\Omega}_0)$ are the inverse Wishart $\mathrm{IW}_u(\boldsymbol{\Psi}, \nu)$ and $\mathrm{IW}_{r-u}(\boldsymbol{\Psi}_0, \nu_0)$ densities, respectively, where $\boldsymbol{\Psi} \in \mathbb{S}_+^{u\times u}$ and $\boldsymbol{\Psi}_0 \in \mathbb{S}_+^{(r-u)\times(r-u)}$ are fixed positive definite matrices, $\nu > u-1$ and $\nu_0 > r-u-1$.

(iii) Conditional on $\boldsymbol{A}$ and $\boldsymbol{\Omega}$, $\pi(\boldsymbol{\eta} \mid \boldsymbol{A}, \boldsymbol{\Omega})$ is the matrix normal $\mathrm{MN}_{u,p}\left(\boldsymbol{\Gamma}(\boldsymbol{A})^T\boldsymbol{e},\ \boldsymbol{\Omega},\ \boldsymbol{M}^{-1}\right)$ density, where $\boldsymbol{M} \in \mathbb{S}_+^{p\times p}$ and $\boldsymbol{e} \in \mathbb{R}^{r\times p}$ are fixed hyper-parameters.

(iv) $\pi(\boldsymbol{A})$ is the matrix normal $\mathrm{MN}_{r-u,u}\left(\boldsymbol{A}_0, \boldsymbol{K}, \boldsymbol{L}\right)$ density, where $\boldsymbol{A}_0 \in \mathbb{R}^{(r-u)\times u}$ and positive definite matrices $\boldsymbol{K} \in \mathbb{S}_+^{(r-u)\times(r-u)}$ and $\boldsymbol{L} \in \mathbb{S}_+^{u\times u}$ are fixed hyper-parameters.

If we have prior information on the most likely envelope subspace (i.e., the prior mode) $\widehat{\mathcal{E}}_{\text{prior}}$, we can find the corresponding prior mode of $\boldsymbol{A}$, say $\widehat{\boldsymbol{A}}_{\text{prior}}$, via the process in (3), and set $\boldsymbol{A}_0 = \hat{\boldsymbol{A}}_{\text{prior}}$. If only partial prior information is available on $\widehat{\mathcal{E}}_{\text{prior}}$, it can still be incorporated in the current framework. For example, suppose $u = 4$ and we know that the unit vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are contained in $\widehat{\mathcal{E}}_{\text{prior}}$, and that $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are linearly independent. We then generate two random vectors $\boldsymbol{v}_3$ and $\boldsymbol{v}_4$ from $\text{span}(\boldsymbol{v}_1, \boldsymbol{v}_2)^\perp$ as follows. Let $\boldsymbol{G}_0 \in \mathbb{R}^{r\times(r-2)}$ be an orthonormal basis of $\text{span}(\boldsymbol{v}_1, \boldsymbol{v}_2)^\perp$, and let $\boldsymbol{C}$ be an $(r-2)\times 2$ matrix with each element independently generated from the standard normal distribution. Let $(\boldsymbol{v}_3, \boldsymbol{v}_4) = \boldsymbol{G}_0\boldsymbol{C}(\boldsymbol{C}^T\boldsymbol{C})^{-1/2}$, and construct $\widehat{\boldsymbol{\Gamma}}_{\text{prior}} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4)$. Then we can obtain $\widehat{\boldsymbol{A}}_{\text{prior}}$ via the process in (3) from $\widehat{\boldsymbol{\Gamma}}_{\text{prior}}$. The matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ in the prior distribution of $\boldsymbol{A}$ represent our confidence about the prior: the variance matrix of column i of $\boldsymbol{A}$ is $l_{ii}\boldsymbol{K}$, where l_{ii} is the ith diagonal element of $\boldsymbol{L}$ and the variance matrix of row j of $\boldsymbol{A}$ is $k_{jj}\boldsymbol{L}$, where k_{jj} is the jth diagonal element of $\boldsymbol{K}$. If we are confident about some specific row(s) or column(s) of $\boldsymbol{A}_{\text{prior}}$, we can make the corresponding k_{jj}'s and l_{ii}'s small.

Theorem 3.1 establishes the propriety of the posterior density (explicit form given in (S4) of the Supplement) associated with the above prior. A proof is provided in Supplement C.4.

Theorem 3.1. *The posterior density for* $(\boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{A})$ *is proper with respect to Lebesgue measure on* $\mathbb{R}^r \times \mathbb{R}^{u\times p} \times \mathbb{S}_+^{u\times u} \times \mathbb{S}_+^{(r-u)\times(r-u)} \times \mathbb{R}^{(r-u)\times u}$.

3.3 Sampling from the Posterior Density

Direct generation of independent random samples from the intractable posterior distribution is infeasible. As an alternative, we provide a Metropolis-within-Gibbs sampler to generate MCMC draws from the posterior. Starting from some initial value, Algorithm 3.1 produces a set of MCMC samples, which can then used to approximate the posterior. Derivations of the conditional distributions used in Algorithm 3.1 are in Supplement C.6.

Algorithm 3.1. One iteration of the Metropolis-within-Gibbs MCMC sampler for updating $(\boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{A})$ with envelope dimension $u \in \{1, \ldots, r-1\}$.

S.1 Generate a Metropolis-Hastings realization for $\boldsymbol{A}$ from the target conditional posterior density $\pi(\boldsymbol{A} \mid \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \mathbb{X}, \mathbb{Y})$ specified in (S13). Subsequently obtain $\boldsymbol{\Gamma}(\boldsymbol{A})$ and $\boldsymbol{\Gamma}_0(\boldsymbol{A})$ through (4). A detailed note on the Metropolis-Hastings scheme is in Supplement C.7.

S.2 Generate $\boldsymbol{\Omega}$ from $\mathrm{IW}_u\left(\boldsymbol{\Psi} + \boldsymbol{\Gamma}(\boldsymbol{A})^T\widetilde{\boldsymbol{G}}\,\boldsymbol{\Gamma}(\boldsymbol{A}),\ n-1+\nu\right)$, where $\widetilde{\boldsymbol{G}} = \mathbb{Y}_c^T\mathbb{Y}_c + \boldsymbol{e}\boldsymbol{M}\boldsymbol{e}^T - \check{\boldsymbol{e}}\left(\mathbb{X}_c^T\mathbb{X}_c + \boldsymbol{M}\right)\check{\boldsymbol{e}}^T$ with $\check{\boldsymbol{e}} = \left(\mathbb{Y}_c^T\mathbb{X}_c + \boldsymbol{e}\boldsymbol{M}\right)\left(\mathbb{X}_c^T\mathbb{X}_c + \boldsymbol{M}\right)^{-1}$.

S.3 Generate $\boldsymbol{\Omega}_0$ from $\mathrm{IW}_{r-u}\left(\boldsymbol{\Psi}_0 + \boldsymbol{\Gamma}(\boldsymbol{A})^T\mathbb{Y}_c^T\mathbb{Y}_c\boldsymbol{\Gamma}(\boldsymbol{A}),\ n-1+\nu_0\right)$.

S.4 Generate $\boldsymbol{\eta}$ from $\mathrm{MN}_{u,p}\left(\boldsymbol{\Gamma}(\boldsymbol{A})^T\check{\boldsymbol{e}},\ \boldsymbol{\Omega},\ \left(\mathbb{X}_c^T\mathbb{X}_c + \boldsymbol{M}\right)^{-1}\right)$, with $\check{\boldsymbol{e}}$ defined in Step S.2.

S.5 Generate $\boldsymbol{\mu}$ from $\mathrm{N}_r\left(\overline{\boldsymbol{Y}} + \boldsymbol{\Gamma}(\boldsymbol{A})\boldsymbol{\eta}\overline{\boldsymbol{X}}, \frac{1}{n}\left(\boldsymbol{\Gamma}(\boldsymbol{A})\boldsymbol{\Omega}\boldsymbol{\Gamma}(\boldsymbol{A})^T + \boldsymbol{\Gamma}_0(\boldsymbol{A})\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0(\boldsymbol{A})^T\right)\right)$.

Remark 3.3. Algorithm 3.1 can be easily modified to account for the cases $u = 0$ and $u = r$, by discarding the steps that involve generations of parameters that are not present in the model. In particular, $\boldsymbol{A}$ is not present in the model when $u = 0$ or $u = r$; hence the Metropolis step S.1 is not needed in such cases. This means that Algorithm 3.1 turns into a Gibbs sampler when $u = 0$ or $u = r$. In addition, when $u = 0$, then $\boldsymbol{\eta} = 0, \boldsymbol{\Gamma}_0(\boldsymbol{A}) = \boldsymbol{I}_r$ and $\boldsymbol{\Sigma} = \boldsymbol{\Omega}_0$, then steps S.4 and S.2 are not needed. On the other hand, when $u = r$, $\boldsymbol{\Gamma}(\boldsymbol{A}) = \boldsymbol{I}_r$ and $\boldsymbol{\Sigma} = \boldsymbol{\Omega}$, and step S.3 is to be skipped.

Theorem 3.2 establishes the Harris ergodicity of Algorithm 3.1. This provides theoretical guarantees against the existence of *pathological* initial values from which the chain may fail to converge. While such pathological initial points have collective measure zero for a Metropolis-within-Gibbs chain, in practice they may arise naturally, as demonstrated in (Roberts and Rosenthal, 2006). Theorem 3.2 ensures that the resulting MCMC samples are asymptotically (i.e., when run long enough) from the correct target posterior density for *all* starting points (not just *almost* all). Thus the MCMC samples will provide strongly consistent estimators of various posterior quantities, such as the posterior mean, posterior variance and posterior quantiles (see, e.g., Tierney, 1994; Chan and Geyer, 1994). A proof of Theorem 3.2 is provided in Supplement C.5.

Theorem 3.2. *The MCMC sampler in Algorithm 3.1 and its extension to the cases $u = 0$ and $u = r$ described in Remark 3.3 is Harris ergodic, i.e., (a) ϕ-irreducible for some measure ϕ, (b) aperiodic and (c) Harris recurrent.*

Remark 3.4. If interest lies in a computationally fast point estimator of the model parameters, one may consider the maximum a posteriori (MAP) estimator. We provide an algorithm for MAP estimation along with some numerical results in Supplement C.11.

4 Bayesian Predictor Envelope / Partial Least Squares

In this section, we first review the predictor envelope model and its connection with partial least square (PLS) from the frequentist perspective. Then we demonstrate how the proposed Bayesian framework (Section 3) applies to the predictor envelope model.

4.1 Review of the Predictor Envelope Model

The predictor envelope model aims to achieve efficiency gains in the estimation of $\boldsymbol{\beta}$ by performing dimension reduction on the predictors. We slightly reparameterize (1) to

$$\boldsymbol{Y} = \boldsymbol{\mu}_{\boldsymbol{Y}} + \boldsymbol{\beta}^T(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}}) + \boldsymbol{\varepsilon}, \tag{7}$$

because we need to utilize the marginal distribution of $\boldsymbol{X}$ to facilitate the discussion. The predictor $\boldsymbol{X} \in \mathbb{R}^p$ is assumed to be stochastic having mean $\boldsymbol{\mu}_{\boldsymbol{X}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$. The response $\boldsymbol{Y} \in \mathbb{R}^r$ can be univariate ($r = 1$) or multivariate ($r > 1$). For the sake of notational clarity we denote the covariance matrix of $\boldsymbol{\varepsilon}$ by $\boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}}$ instead of $\boldsymbol{\Sigma}$ hereafter.

The predictor envelope model imposes the envelope structure on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ by constructing the $\boldsymbol{\Sigma}_{\boldsymbol{X}}$-envelope of span($\boldsymbol{\beta}$), denoted by $\mathcal{E}_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}(\boldsymbol{\beta})$. It can be shown that $(\mathbb{P}_{\mathcal{E}}\boldsymbol{X}, \boldsymbol{Y})$ is uncorrelated with $\mathbb{Q}_{\mathcal{E}}\boldsymbol{X}$ (Cook, 2018), so the immaterial part $\mathbb{Q}_{\mathcal{E}}\boldsymbol{X}$ does not carry any information on $\boldsymbol{Y}$ directly or indirectly via its correlation with $\mathbb{P}_{\mathcal{E}}\boldsymbol{X}$. All the information on $\boldsymbol{\beta}$ is encapsulated in $\mathbb{P}_{\mathcal{E}}\boldsymbol{X}$. Hence the predictor envelope model is formulated as

$$\boldsymbol{Y} = \boldsymbol{\mu}_{\boldsymbol{Y}} + \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}}) + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \tag{8}$$

where $\boldsymbol{\beta}^T = \boldsymbol{\Gamma}\boldsymbol{\eta}$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times m}$ is an orthonormal basis for $\mathcal{E}_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}(\boldsymbol{\beta})$ and m ($0 \le m \le p$) is the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}(\boldsymbol{\beta})$. The matrix $\boldsymbol{\eta}$ carries the coordinates of $\boldsymbol{\beta}^T$ with respect to $\boldsymbol{\Gamma}$. The matrices $\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T = \text{var}(\mathbb{P}_{\mathcal{E}}\boldsymbol{X})$ and $\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T = \text{var}(\mathbb{Q}_{\mathcal{E}}\boldsymbol{X})$ respectively quantify the variations

in the material and immaterial parts of $\boldsymbol{X}$, where $\boldsymbol{\Gamma}_0$ is an orthonormal basis of $\mathcal{E}_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}(\boldsymbol{\beta})^{\perp}$, and $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are positive definite matrices.

PLS originated in econometrics (Wold, 1966), and is widely used in various applied scientific disciplines, such as chemometrics, behavioral science, genetics, and social sciences as a method to improve the prediction performance of OLS. Despite its popularity, its Bayesian development has been limited, primarily because PLS has been historically defined in terms of iterative algorithms and not as a model-based approach. Similar to the predictor envelope model, PLS also aims a dimension reduction of $\boldsymbol{X}$ in the linear regression model (7). It operates by reducing $\boldsymbol{X}$ to a few linear combinations $\boldsymbol{G}^T\boldsymbol{X}$, where $\boldsymbol{G} \in \mathbb{R}^{p\times m}$ has full column rank, and $m(\leq p)$ is called the number of components. There exist multiple PLS algorithms to estimate $\boldsymbol{G}$; a popular algorithm is SIMPLS (De Jong, 1993) which uses a sequential moment-based procedure to obtain $\widehat{\boldsymbol{G}}_{\text{PLS}}$.

Cook et al. (2013) showed that $\text{span}(\widehat{\boldsymbol{G}}_{\text{PLS}})$ is a $\sqrt{n}$-consistent estimator of the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}(\boldsymbol{\beta})$. This implies that the predictor envelope estimator of $\boldsymbol{\beta}$ and the SIMPLS estimator of $\boldsymbol{\beta}$ both estimate the same dimension reduction subspace $\mathcal{E}_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}(\boldsymbol{\beta})$, and the SIMPLS estimator can be studied through the predictor envelope model.

4.2 Formulation of the Bayesian Predictor Envelope Model

To construct a Bayesian framework for the predictor envelope model, we follow the parametrization in (3) and construct $\boldsymbol{C}_{\boldsymbol{A}}$ and $\boldsymbol{D}_{\boldsymbol{A}}$ as:

$$\boldsymbol{C}_{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{I}_m \\ \boldsymbol{A} \end{pmatrix} \in \mathbb{R}^{p\times m}, \quad \text{and } \boldsymbol{D}_{\boldsymbol{A}} = \begin{pmatrix} -\boldsymbol{A}^T \\ \boldsymbol{I}_{p-m} \end{pmatrix} \in \mathbb{R}^{p\times(p-m)}, \tag{9}$$

where $\boldsymbol{A} \in \mathbb{R}^{(p-m)\times m}$ is an unconstrained matrix. Then we can express $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$ as explicit functions of $\boldsymbol{A}$: $\boldsymbol{\Gamma}(\boldsymbol{A}) = \boldsymbol{C}_{\boldsymbol{A}}\left(\boldsymbol{C}_{\boldsymbol{A}}^T\boldsymbol{C}_{\boldsymbol{A}}\right)^{-1/2}$ and $\boldsymbol{\Gamma}_0(\boldsymbol{A}) = \boldsymbol{D}_{\boldsymbol{A}}\left(\boldsymbol{D}_{\boldsymbol{A}}^T\boldsymbol{D}_{\boldsymbol{A}}\right)^{-1/2}$.

A parametric representation of the predictor envelope model (8) is formulated as

$$\begin{aligned} \boldsymbol{Y} \mid \boldsymbol{X} &\sim \mathrm{N}_r \left(\boldsymbol{\mu}_{\boldsymbol{Y}} + \boldsymbol{\eta}^T \boldsymbol{\Gamma}^T(\boldsymbol{A})(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}}), \boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}} \right) \\ \boldsymbol{X} &\sim \mathrm{N}_p \left(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Gamma}(\boldsymbol{A}) \boldsymbol{\Omega} \boldsymbol{\Gamma}^T(\boldsymbol{A}) + \boldsymbol{\Gamma}_0(\boldsymbol{A}) \boldsymbol{\Omega}_0 \boldsymbol{\Gamma}_0^T(\boldsymbol{A}) \right). \end{aligned} \tag{10}$$

Notice that the marginal distribution of $\boldsymbol{X}$ is included in (10) because $\boldsymbol{\beta}$ is linked with $\boldsymbol{\Sigma}_{\boldsymbol{X}}$. More precisely, the marginal distribution of $\boldsymbol{X}$ aids identification of the material and immaterial parts in (8). An expression for the log-likelihood with n data points $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i) : i = 1, \ldots, n\}$ from model (10) is given in (S16) in the Supplement.

4.3 Prior and Posterior distributions

The joint prior density for the model parameters in (10), viz., $\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\mu}_{\boldsymbol{Y}}, \boldsymbol{\eta}, \boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0$ and $\boldsymbol{A}$ is assumed to be of the form $\pi(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\mu}_{\boldsymbol{Y}}, \boldsymbol{\eta}, \boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{A}) = \pi(\boldsymbol{\mu}_{\boldsymbol{X}})\pi(\boldsymbol{\mu}_{\boldsymbol{Y}})\pi(\boldsymbol{\Omega})\pi(\boldsymbol{\Omega}_0)$ $\pi(\boldsymbol{A})\pi(\boldsymbol{\eta} \mid \boldsymbol{A}, \boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}})\pi(\boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}})$. Here

(i) $\pi(\boldsymbol{\mu}_{\boldsymbol{X}}) \propto 1$ and $\pi(\boldsymbol{\mu}_{\boldsymbol{Y}}) \propto 1$ are improper flat densities (with respect to Lebesgue measures on $\mathbb{R}^p$ and $\mathbb{R}^r$ respectively).

(ii) $\pi(\boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}})$, $\pi(\boldsymbol{\Omega})$ and $\pi(\boldsymbol{\Omega}_0)$ are inverse Wishart $\mathrm{IW}_r(\boldsymbol{\Psi}_{\boldsymbol{Y}}, \nu_{\boldsymbol{Y}})$, $\mathrm{IW}_m(\boldsymbol{\Psi}_{\boldsymbol{X}}, \nu_{\boldsymbol{X}})$ and $\mathrm{IW}_{p-m}(\boldsymbol{\Psi}_{0,\boldsymbol{X}}, \nu_{0,\boldsymbol{X}})$ densities respectively, where $\boldsymbol{\Psi}_{\boldsymbol{Y}} \in \mathbb{S}_+^{r\times r}$, $\boldsymbol{\Psi}_{\boldsymbol{X}} \in \mathbb{S}_+^{m\times m}$, $\boldsymbol{\Psi}_{0,\boldsymbol{X}} \in \mathbb{S}_+^{(p-m)\times(p-m)}$, $\nu_{\boldsymbol{Y}} > r-1$, $\nu_{\boldsymbol{X}} > m-1$ and $\nu_{0,\boldsymbol{X}} > p-m-1$ are fixed hyper-parameters.

(iii) $\pi(\boldsymbol{A})$ is the matrix normal $\mathrm{MN}_{p-m,m}(\boldsymbol{A}_0, \boldsymbol{K}, \boldsymbol{L})$ density, where $\boldsymbol{K} \in \mathbb{S}_+^{(p-m)\times(p-m)}$, $\boldsymbol{L} \in \mathbb{S}_+^{m\times m}$, and $\boldsymbol{A}_0 \in \mathbb{R}^{(p-m)\times m}$ are fixed hyper-parameters.

(iv) Conditional on $\boldsymbol{A}$ and $\boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}}$, $\pi(\boldsymbol{\eta})$ is the matrix normal $\mathrm{MN}_{m,r}\left(\boldsymbol{M}^{-1}\boldsymbol{\Gamma}^T(\boldsymbol{A})\boldsymbol{e}, \boldsymbol{M}^{-1}, \boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}}\right)$ density, where $\boldsymbol{M} \in \mathbb{S}_+^{m\times m}$, and $\boldsymbol{e} \in \mathbb{R}^{p\times r}$ are fixed hyper-parameters.

Prior information on $\mathcal{E}_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}(\boldsymbol{\beta})$ can be incorporated in a similar way as discussed in Section 3.2. The form of the resulting posterior density is given in (S17). Theorem 4.1 establishes propriety of the posterior distribution. A proof is provided in the Supplement D.2.

Theorem 4.1. *The posterior density for* $(\boldsymbol{\mu}_{\boldsymbol{Y}}, \boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{A})$ *is proper with respect to the Lebesgue measure on* $\mathbb{R}^r \times \mathbb{R}^p \times \mathbb{S}_+^{r\times r} \times \mathbb{R}^{m\times r} \times \mathbb{S}_+^{m\times m} \times \mathbb{S}_+^{(p-m)\times(p-m)} \times \mathbb{R}^{(p-m)\times m}$.

4.4 Sampling from the Posterior Density

Similar to the response envelope model, generating i.i.d. samples directly from the posterior distribution for the predictor envelope model is not feasible. A Metropolis-within-Gibbs MCMC sampler for generating approximate samples from the target posterior density is provided in Algorithm D.1 in Supplement D.1. The proposed MCMC sampler for the predictor envelope model is similar to that for the response envelope model except that here we have a couple of extra parameters to sample. The following theorem establishes Harris ergodicity of the sampler. A proof is provided in Section D.3 of the Supplement. A note on the MAP estimation under a predictor envelope model is provided in Supplement D.4.

Theorem 4.2. *The Metropolis-within-Gibbs sampler in Algorithm D.1 (Supplement), and its extension to the cases with* $m = 0$ *or* $m = p$*, is Harris ergodic, i.e., (a)* ϕ*-irreducible for some measure* ϕ*, (b) aperiodic and (c) Harris recurrent.*

Because of the connection between the predictor envelope model and PLS (more specifically, SIMPLS), the proposed approach also serves as a Bayesian framework for PLS. Note that all the parameters in model (10) are well defined and identifiable, which avoids the identification issue present in the Bayesian PLS model of Vidaurre et al. (2013). The current framework also makes it easy to incorporate prior information on the dimension

reduction subspace. If we know a priori the most likely dimension reduction subspace, we can find a matrix $\widehat{\boldsymbol{A}}_{\text{prior}}$ such that the corresponding $\boldsymbol{C}_{\boldsymbol{A}}$ is a basis of this subspace. Then we can set the mode of the prior $\boldsymbol{A}_0$, to be $\widehat{\boldsymbol{A}}_{\text{prior}}$. The Bayesian approach for other PLS variants, such as NIPALS, can also be developed using this framework.

5 Bayesian Envelope Model for Generalized Linear Regression

This section demonstrates how our Bayesian envelope approach can be applied to a generalized linear regression model. We consider the probit model as an example for illustration purposes; other generalized linear regression models can be studied using similar techniques.

5.1 Formulation

The probit model is formulated as $P(Y = 1 \mid \boldsymbol{X}) = \Phi(\mu_Y + \boldsymbol{\beta}^T(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}}))$, where Y is a binary response, $\boldsymbol{X} \in \mathbb{R}^p$ is the stochastic predictor vector, $\mu_Y \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\Phi(\cdot)$ denotes the cumulative distribution function for the standard normal distribution.

We impose the envelope structure on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ as follows. Consider the $\boldsymbol{\Sigma}_{\boldsymbol{X}}$-envelope of $\text{span}(\boldsymbol{\beta})$, i.e., $\mathcal{E}_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}(\boldsymbol{\beta})$, then we have $\text{span}(\boldsymbol{\beta}) \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}(\boldsymbol{\beta})$, and that $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ can be decomposed into the sum of the variation of the material part $\text{var}(\mathbb{P}_{\mathcal{E}}\boldsymbol{X})$, and the variation of the immaterial part $\text{var}(\mathbb{Q}_{\mathcal{E}}\boldsymbol{X})$. Let m $(0 \leq m \leq p)$ be the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}(\boldsymbol{\beta})$, and $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times m}$ be an orthonormal basis of $\mathcal{E}_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}(\boldsymbol{\beta})$. Then the envelope probit model is formulated as

$$P(Y = 1 \mid \boldsymbol{X}) = \Phi(\mu_Y + \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})), \qquad \boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T,$$

where $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p\times(p-m)}$ is an orthonormal basis of $\mathcal{E}_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}(\boldsymbol{\beta})^{\perp}$, $\boldsymbol{\eta} \in \mathbb{R}^m$ contains the coordinates

of $\boldsymbol{\beta}$ with respect to $\boldsymbol{\Gamma}$, and $\boldsymbol{\Omega} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(p-m)\times(p-m)}$ are positive definite matrices that contain the coordinates of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ with respect to $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$.

Again we use the parameterization in (3), so that $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$ can be expressed as functions of $\boldsymbol{A}$, i.e. $\boldsymbol{\Gamma}(\boldsymbol{A}) = \boldsymbol{C}_{\boldsymbol{A}}(\boldsymbol{C}_{\boldsymbol{A}}^T\boldsymbol{C}_{\boldsymbol{A}})^{-1/2}$ and $\boldsymbol{\Gamma}_0(\boldsymbol{A}) = \boldsymbol{D}_{\boldsymbol{A}}(\boldsymbol{D}_{\boldsymbol{A}}^T\boldsymbol{D}_{\boldsymbol{A}})^{-1/2}$, where $\boldsymbol{C}_{\boldsymbol{A}}$ and $\boldsymbol{D}_{\boldsymbol{A}}$ are as defined in (9). Under this parameterization, the envelope probit model is formulated as

$$\begin{aligned} P(Y = 1 \mid \boldsymbol{X}) &= \Phi\left(\mu_Y + \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T(\boldsymbol{A})(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})\right) \\ \boldsymbol{X} &\sim \mathrm{N}_p\left(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Gamma}(\boldsymbol{A})\boldsymbol{\Omega}\boldsymbol{\Gamma}^T(\boldsymbol{A}) + \boldsymbol{\Gamma}_0(\boldsymbol{A})\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T(\boldsymbol{A})\right). \end{aligned} \tag{11}$$

Note that the marginal distribution of $\boldsymbol{X}$ also appears in the envelope probit model as it is utilized in the estimation and inference of $\boldsymbol{\beta}$. The interpretations of the parameters $\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0$ and $\boldsymbol{A}$ are analogous to the interpretations of the corresponding parameters in the predictor envelope model (Section 4). However, the intercept parameter μ_Y now represents the unconditional mean of Y on a *probit scale*.

5.2 Prior and Posterior Distributions

The joint prior density of the parameters $\mu_Y, \boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0$, and $\boldsymbol{A}$ in model (11) is assumed to be of the form $\pi(\mu_Y, \boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{A}) = \pi(\mu_Y)\pi(\boldsymbol{\mu}_{\boldsymbol{X}})\pi(\boldsymbol{\eta} \mid \boldsymbol{A})\pi(\boldsymbol{\Omega})\pi(\boldsymbol{\Omega}_0)\pi(\boldsymbol{A})$, where

(i) $\pi(\mu_Y)$ and $\pi(\boldsymbol{\mu}_{\boldsymbol{X}})$ are the univariate normal $\mathrm{N}(\alpha_Y, \Xi_Y)$ and p-variate normal $\mathrm{N}_p(\boldsymbol{\alpha}_{\boldsymbol{X}}, \boldsymbol{\Xi}_{\boldsymbol{X}})$ densities respectively, where $\alpha_Y \in \mathbb{R}$, $\Xi_Y > 0$, $\boldsymbol{\alpha}_{\boldsymbol{X}} \in \mathbb{R}^p$ and $\boldsymbol{\Xi}_{\boldsymbol{X}} \in \mathbb{S}_+^{p\times p}$ are fixed hyper-parameters.

(ii) $\pi(\boldsymbol{\Omega})$ and $\pi(\boldsymbol{\Omega}_0)$ are inverse Wishart $\mathrm{IW}_m(\boldsymbol{\Psi}_{\boldsymbol{X}}, \nu_{\boldsymbol{X}})$ and $\mathrm{IW}_{p-m}(\boldsymbol{\Psi}_{0,\boldsymbol{X}}, \nu_{0,\boldsymbol{X}})$ densities respectively, where $\boldsymbol{\Psi}_{\boldsymbol{X}} \in \mathbb{S}_+^{m\times m}$, $\boldsymbol{\Psi}_{0,\boldsymbol{X}} \in \mathbb{S}_+^{(p-m)\times(p-m)}$, $\nu_{\boldsymbol{X}} > m-1$, and $\nu_{0,\boldsymbol{X}} > p-m-1$ are fixed hyper-parameters.

(iii) $\pi(\boldsymbol{A})$ is the matrix normal $\mathrm{MN}_{p-m,m}(\boldsymbol{A}_0, \boldsymbol{K}, \boldsymbol{L})$ density, where $\boldsymbol{A}_0 \in \mathbb{R}^{(p-m)\times m}$, $\boldsymbol{K} \in \mathbb{S}_+^{(p-m)\times(p-m)}$ and $\boldsymbol{L} \in \mathbb{S}_+^{m\times m}$ are fixed hyper-parameters.

(iv) Conditional on $\boldsymbol{A}$, $\pi(\boldsymbol{\eta} \mid \boldsymbol{A})$ is the m-variate normal $\mathrm{N}_m\left(\boldsymbol{M}^{-1}\boldsymbol{\Gamma}^T(\boldsymbol{A})\boldsymbol{e}, \boldsymbol{M}^{-1}\right)$ density, where $\boldsymbol{M} \in \mathbb{S}_+^{m\times m}$ and $\boldsymbol{e} \in \mathbb{R}^p$ are fixed hyper-parameters.

The explicit form of the posterior density is given in (S27) of the Supplement.

Remark 5.1. It is known that improper flat priors on the parameters in a standard Bayesian probit regression model can lead to an improper posterior distribution. Chen and Shao (2001) provide sufficient conditions on the non-stochastic design matrix to ensure posterior propriety under improper flat prior. In the current settings, however, posterior propriety is guaranteed almost surely as the joint prior distribution is proper.

5.3 Data Augmentation MCMC Sampler

The posterior density (S27) is more complicated than (S4) or (S17) in the supplement due to the $\log \Phi(\cdot)$ term in the full conditional posterior of $\mu_Y, \boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\eta}, \boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. This precludes direct formulation of Metropolis-within-Gibbs algorithms similar to Algorithm 3.1 or Algorithm D.1 (supplement). Fortunately, the data augmentation technique for Bayesian probit regression models (Albert and Chib, 1993) can be adopted here. For each Y_i, one introduces a latent Gaussian random variable U_i with $E(U_i \mid \boldsymbol{X}_i) = \mu_Y + \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T(\boldsymbol{A})(\boldsymbol{X}_i - \boldsymbol{\mu}_{\boldsymbol{X}})$, $\mathrm{var}(U_i \mid \boldsymbol{X}_i) = 1$ and $Y_i = \mathbb{1}\{U_i \geq 0\}$, where $\mathbb{1}(\cdot)$ is the indicator function.

Straightforward derivations reveal that the full conditional distribution of U_i given $Y_i, \boldsymbol{X}_i$ and the parameters $\{\mu_Y, \boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{A}\}$ is $\mathrm{TN}(\mu_Y + \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T(\boldsymbol{A})(\boldsymbol{X}_i - \boldsymbol{\mu}_{\boldsymbol{X}}), 1, Y_i)$, where $\mathrm{TN}(\mu, \sigma^2, \omega)$ denotes the truncated normal distribution, which is a normal $\mathrm{N}(\mu, \sigma^2)$ distribution truncated to be non-negative or negative depending on $\omega = 1$ or 0. There exist

multiple efficient algorithms for random generation from truncated normal distribution (Robert, 1995; Marsaglia, 1964), implemented in various softwares including `R`. The advantage of defining the latent variable U_i lies in the simplification of the posterior distribution of the model parameters. In particular, given U_i, the information on Y_i is superfluous, and the data $(\boldsymbol{X}_i, U_i)$, $i = 1, \ldots, n$ follow a predictor envelope model considered in Section 4, with fixed variance $\text{var}(U_i \mid \boldsymbol{X}_i) = 1$. Let $\mathbb{U} = (U_1, \ldots, U_n)^T$, then the posterior density $\pi(\mu_Y, \boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{A} \mid \mathbb{U}, \mathbb{X})$ has the same form as the posterior density for predictor envelope model (S17), except for the following differences: r is restricted to be equal to 1, $\boldsymbol{\mu}_{\boldsymbol{X}}$ and μ_Y now have proper normal priors, $\mathbb{Y}$ is replaced by $\mathbb{U}$, and the $\boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}}$ term is not present ($\text{var}(U_i \mid \boldsymbol{X}_i)$ is 1). A data augmentation Metropolis-within-Gibbs sampler (Algorithm E.1 in Supplement E.1) is derived to draw samples from the posterior density (S27). Theorem 5.1 establishes Harris ergodicity of the Markov chain generated using Algorithm E.1. The proof is provided in Supplement E.3.

Theorem 5.1. *The Metropolis-within-Gibbs sampler in Algorithm E.1 and its extension to the cases $m = 0$ and $m = p$ described in Remark E.1 is Harris ergodic.*

An expectation conditional maximization (ECM) algorithm (Meng and Rubin, 1993) for the MAP estimation of the parameters is provided in Section E.2 of the Supplement.

We have demonstrated the incorporation of the Probit model in the proposed envelope framework via the Albert and Chib (1993) data-augmentation approach. The framework is quite flexible and can incorporate other generalized linear regression models. For example, with the logistic regression or multinomial regression, the data-augmentation scheme in Polson et al. (2013) and Holmes and Held (2006) can be applied in a similar way as in the Probit model.

6 Bayesian Inference on the Envelope Dimension

The envelope dimension ("u" in the response envelope; "m" in the predictor and probit envelope models) is an unknown parameter that requires specification. Optimal selection of this parameter can be viewed as a model selection problem as considered in Khare et al. (2017); however, the Bayesian paradigm permits a coherent approach to infer this parameter via posterior probabilities obtained after eliciting a prior distribution. For expository purposes, below we consider the response envelope model; analogous methods can be obtained for predictor and probit envelope models.

Given a prior distribution $\pi(u)$ on u elicited independently of the other model parameters, our interest lies in the posterior $\Pr(u \mid \text{data})$. However, computation of this posterior requires evaluation of the marginal likelihood, which is an extremely challenging problem for a complicated model like ours. In Supplement B we discuss some potential approaches for this computation along with their challenges. Here we consider a simple BIC-based approximation of $\Pr(u \mid \text{data})$ from Kass and Raftery (1995):

$$\Pr(u = k \mid \text{data}) \approx \frac{\exp(-\text{BIC}(k)/2)\ \pi(u = k)}{\sum_{k'=0}^{r} \exp(-\text{BIC}(k')/2)\ \pi(u = k')};\ k = 0, 1, \ldots, r \tag{12}$$

where $\text{BIC}(u) = -2 \log \widetilde{L}(u) + \rho(u) \log n$, with $\widetilde{L}(u)$ denoting the maximized value of the likelihood function, and $\rho(u) = r(r+1)/2 + r + pu$ being the effective number of parameters in the response envelope model. Given $\Pr(u \mid \text{data})$, one may consider, e.g., the posterior mode $\hat{u} = \arg\max_{k=0,\ldots,r} \Pr(u = k \mid \text{data})$ as a point estimate of u, and then estimate the other parameters $\{\boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{A}\}$ conditional on $u = \hat{u}$. A more coherent *Bayesian model averaging* (BMA) approach acknowledges the estimation uncertainty in u and considers the BMA posterior for the model parameters:

$$\pi(\mu, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}, \boldsymbol{A} \mid \text{data}) = \sum_{k=0}^{r} \pi(\mu, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}, \boldsymbol{A} \mid u = k, \text{data})\ \Pr(u = k \mid \text{data}). \tag{13}$$

The associated marginal posterior mean for $\boldsymbol{\beta}$ is

$$E(\boldsymbol{\beta} \mid \text{data}) = \sum_{k=0}^{r} E(\boldsymbol{\beta} \mid u = k, \text{data}) \Pr(u = k \mid \text{data}) \tag{14}$$

which is a weighted average of posterior means of $\boldsymbol{\beta}$ obtained from $u = k$ specific models. Supplement B contains more details on estimation and inference. The form of the point estimator (14) resembles the frequentist weighted envelope estimator (Eck and Cook, 2017), which is a weighted average of the estimator of $\boldsymbol{\beta}$ under envelope models with various u's. We note here that (13) permits coherent Bayesian inference on $\boldsymbol{\beta}$. Furthermore, $\Pr(u = k \mid \text{data})$ can incorporate prior information on u, unlike the frequentist approach.

7 Illustration

This section demonstrates efficiency gains achieved by the Bayesian envelope models via simulations and data examples. Vague priors are considered for all model parameters. In particular, for any univariate, multivariate or matrix normal prior, the mean is set to be zero, and the covariance matrix is set to be 10^6 times the identity matrix. In the inverse Wishart prior $\text{IW}_d(\boldsymbol{\Psi}, \nu)$, $\boldsymbol{\Psi}$ and ν are taken to be $10^{-6}\boldsymbol{I}_d$ and d respectively. We estimate the model parameters using the posterior means. The tuning parameters in the Metropolis steps, if present, are tuned during burn-in. They are adaptively increased/decreased at every fifth iteration to ensure an acceptance rate of 30–50%. For inference on the envelope dimension, we use the BIC-approximated posterior probabilities.

7.1 Simulation Study

We consider the response envelope model (6) for illustration. Analogous simulation results on predictor envelope model and envelope probit model are provided in Supplement D.6

and E.4. We fixed $r = 20$, $p = 7$ and $u_{\text{true}} = 2$, where the subscript "true" denotes the value of a parameter in the data generation process. The elements of $\boldsymbol{\mu}_{\text{true}}, \boldsymbol{\eta}_{\text{true}}$ and $\boldsymbol{A}_{\text{true}}$ were independently generated from Uniform$(0, 10)$, Uniform$(0, 10)$ and Uniform$(-1, 1)$ respectively. The matrices $\boldsymbol{\Omega}_{\text{true}}$ and $\boldsymbol{\Omega}_{0,\text{true}}$ were diagonal, with diagonal elements being independent Uniform$(0, 1)$ and Uniform$(5, 10)$ variates. The sample size n was varied from $50, 100, 200, 500$ and 1000. For each sample size, 200 replicated datasets were generated. On each dataset, we ran Algorithm 3.1 to generate 13,333 MCMC draws (after a burn-in of 6,667 draws) for each u in $\{0, 1, \ldots, r\}$. A uniform prior is consider for u. The median running times for the MCMC algorithms were 7.55, 5.43, 7.53, 6.02, and 7.94 minutes for $n = 50$, 100, 200, 500, and 1000 respectively under $u = u_{\text{true}}$. Computations were done on SLURM HPC parallel computing clusters, with computing nodes having 16 GB of allocated memory each, and clock speeds ranging between 2.10 – 2.40 GHz.

n	$0 \leq u \leq 1$	$u = 2$	$u = 3$	$u = 4$	$5 \leq u \leq 7$
50	0.000	0.239	0.427	0.280	0.003
100	0.000	0.701	0.275	0.024	0.000
200	0.000	0.929	0.071	0.000	0.000
500	0.000	0.988	0.012	0.000	0.000
1000	0.000	1.000	0.000	0.000	0.000

Table 1: Posterior probabilities of envelope dimension u.

Table 1 contains the posterior probabilities $\Pr(u \mid \text{data})$ averaged across replicates. As n grows, we notice that $\Pr(u \mid \text{data})$ concentrates more at $u_{\text{true}} = 2$. With smaller n's $\Pr(u \mid \text{data})$ tends to put nontrivial masses on $\{u > u_{\text{true}}\}$ but not on $\{u < u_{\text{true}}\}$, effectuating

overestimation of u rather than underestimation. In general, the cost of overestimation of u is lower than underestimation, since underestimation induces bias while overestimation usually does not induce bias but loses some efficiency gains.

We next focused on the estimation of $\boldsymbol{\beta}$. On each replicate, we considered four estimators of $\boldsymbol{\beta}$ from (a) the envelope model with $\hat{u} = \arg\max_{0\leq u\leq r} \Pr(u \mid \text{data})$ (model $\mathcal{M}_{\hat{u}}$), (b) the envelope model with $u_{\text{true}} = 2$ (model $\mathcal{M}_{u_{\text{true}}}$), (c) envelope model with BMA approach using (14) (model $\mathcal{M}_{\text{BMA}}$), and (d) the standard Bayesian regression model (model $\mathcal{M}_{\text{std}}$). We adopt a "frequentist evaluation" of the estimators using mean squared errors (MSE) and estimation variances. (Bayesian evaluations through posterior standard deviations are provided in Supplement C.10.) For each model $\mathcal{M} \in \{\mathcal{M}_{\hat{u}}, \mathcal{M}_{u_{\text{true}}}, \mathcal{M}_{\text{BMA}}, \mathcal{M}_{\text{std}}\}$ and element (i, j) in $\boldsymbol{\beta}$ (denoted by $\beta_{i,j}$), we calculated the MSE $M_{i,j,\mathcal{M}} = \sum_{k=1}^{200} \left(\hat{\beta}^k_{i,j,\mathcal{M}} - \beta_{i,j,\text{true}}\right)^2 /200$ and the estimation variance $V_{i,j,\mathcal{M}} = \sum_{k=1}^{200} \left(\hat{\beta}^k_{i,j,\mathcal{M}} - \overline{\beta}_{i,j,\mathcal{M}}\right)^2 /200$. Here $\hat{\beta}^k_{i,j,\mathcal{M}}$ denotes the estimator of $\beta_{i,j}$ obtained from model $\mathcal{M}$ in the k-th replicate, and $\overline{\beta}_{i,j,\mathcal{M}} = \sum_{k=1}^{200} \hat{\beta}^k_{i,j,\mathcal{M}}/200$. The ratios of these MSEs and variances for the envelope estimators to the standard Bayesian

n	$M_{i,j,\mathcal{M}_{\text{std}}}/M_{i,j,\mathcal{M}_{\hat{u}}}$	$M_{i,j,\mathcal{M}_{\text{std}}}/M_{i,j,\mathcal{M}_{\text{BMA}}}$	$M_{i,j,\mathcal{M}_{\text{std}}}/M_{i,j,\mathcal{M}_{u_{\text{true}}}}$
50	4.89 (2.72, 8.69)	5.03 (2.74, 9.29)	7.68 (3.54, 17.29)
100	6.48 (3.73, 17.09)	6.63 (3.75, 17.29)	7.58 (3.98, 20.98)
200	7.12 (3.47, 17.03)	7.20 (3.49, 16.99)	7.56 (3.58, 18.69)
500	7.62 (3.66, 21.41)	7.66 (3.66, 21.72)	7.81 (3.66, 22.68)
1000	7.65 (3.73, 23.39)	7.65 (3.73, 23.39)	7.65 (3.73, 23.39)

Table 2: Medians (ranges) of the component-wise MSE ratios.

regression estimators are summarized Table 2 and Supplementary Table S11 respectively.

These ratios are all well above 1, which demonstrates the envelope model is able to achieve efficiency gains in the estimation of every element in $\boldsymbol{\beta}$. When $n = 1000$, the ratios for all three envelope estimators become nearly identical as $\Pr(u \mid \text{data})$ concentrates at u_{true}. For smaller n, the estimator from $\mathcal{M}_{\hat{u}}$ loses some efficiency due to overestimated u in some cases. The estimator from $\mathcal{M}_{\text{BMA}}$ is more efficient as it still incorporates information from $\mathcal{M}_{u_{\text{true}}}$ with positive probability. However, the efficiency losses in $\mathcal{M}_{\hat{u}}$ and $\mathcal{M}_{\text{BMA}}$ are mild, and the resulting estimators are still considerably more efficient than the estimator of $\mathcal{M}_{\text{std}}$. Figure 1 takes a closer look at $\beta_{1,1}$ for visual assessments of the competing estimators. The figure shows the MSEs and estimation variance of the estimators under various sample sizes. Throughout, the $\mathcal{M}_{u_{\text{true}}}$ estimator enjoys the most efficiency gain, followed by $\mathcal{M}_{\text{BMA}}$ and $\mathcal{M}_{\hat{u}}$. The $\mathcal{M}_{\text{std}}$ estimator has the largest MSE, mainly due to its estimation variance. When $n \geq 100$, the estimators from $\mathcal{M}_{\text{BMA}}$ and $\mathcal{M}_{\hat{u}}$ are very close to the $\mathcal{M}_{u_{\text{true}}}$ estimator.

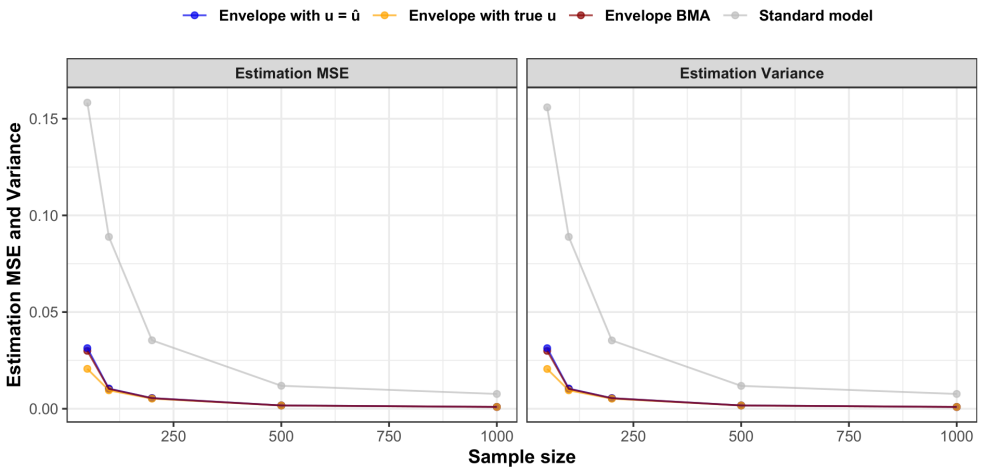

Figure 1: MSE and estimation variances of competing estimators of $\beta_{1,1}$ at different n.

In addition, we performed a range of simulations investigating the performance of the proposed approaches under various scenarios. First, under the above simulation settings,

we compared the MSE and prediction errors of the proposed approaches with a few more competing methods including (a) manifold Bayesian envelope model (Khare et al., 2017) (b) frequentist envelope model (Cook et al., 2010), (c) frequentist weighted envelope estimator (Eck and Cook, 2017), (d) approximate mean-field variational Bayes envelope estimator, (e) frequentist reduced rank regression, (f) Bayesian reduced rank regression, and (g) remMap estimator (Peng et al., 2010). (Supplement C.13). Second, we varied the signal-to-noise ratios, and compared estimation and prediction performances of the proposed approaches with the competitors (Supplement C.15). Third, we assessed the proposed MCMC algorithm under larger sample sizes and dimensions (Supplement C.14). We also discovered that although the variational Bayes implementation is much faster to compute, it induces bias in estimation and produces unreliable uncertainty quantification. Finally, we compared posterior standard deviations of the proposed Bayesian estimators with bootstrap standard errors of the frequentist envelope estimators in Supplement C.10.

7.2 Real Data Analyses

7.2.1 Response Envelope Model

We apply the Bayesian response envelope model to the *Arabidopsis thaliana* dataset from the genetic association study in Wille et al. (2004). This dataset was analyzed in Mukherjee et al. (2015) in the context of multivariate linear regression to understand how the gene expression levels of downstream pathways, carotenoid, and phytosterol are affected by the gene expression levels of two isoprenoid biosynthesis pathways, mevalonate and non-mevalonate. The predictors correspond to 39 genes from mevalonate and non-mevalonate, and the responses correspond to 36 genes from carotenoid and phytosterol. The dataset

consists of a total of 118 samples collected from GeneChip microarray experiments. We ran Algorithm 3.1 to generate 10,000 MCMC samples after discarding a burn-in of 5,000 iterations with $u = 0, 1, \ldots, 36$ and subsequently computed the marginal posterior probabilities for u. Up to three decimal places, the posterior probabilities were non-zero for only two u's viz., $\Pr(u = 1 \mid \text{data}) = 0.998$ and $\Pr(u = 2 \mid \text{data}) = 0.002$. With these posterior probabilities, the BMA model would produce a posterior essentially identical to the posterior of an envelope model with $u = 1$. Hence for subsequent analysis, we focused on the envelope model with $u = 1$. We compared the posterior mean and standard deviation of $\boldsymbol{\beta}$ from the Bayesian envelope model and the Bayesian standard model. The results are in the left panel of Figure 2. The horizontal axis indicates the coordinates of $\text{vec}(\boldsymbol{\beta})$, where $\text{vec}(\cdot)$ is the vector operation that stacks a matrix to a vector columnwise. As depicted through the posterior standard deviations, the Bayesian envelope estimators enjoy substantially less estimation uncertainty than the Bayesian standard linear regression estimator.

We now compare the prediction performance of the proposed Bayesian envelope estimator with the Bayesian standard estimator, frequentist envelope estimator, reduced rank regression (RRR) estimator, and remMap estimator. We performed cross-validations with 100 random $80\% - 20\%$ training-test splits. The predictor performance is evaluated by the unit average prediction error (defined in Supplement C.13.2). For the two Bayesian approaches, we also used the test-set log posterior predictive density (LPPD) as a measure of Bayesian predictive accuracy (Gelman et al., 2013, Section 7.1), with a higher LPDD indicating a better fit. The prediction errors and LPPD are plotted in panels B and C of Figure 2. In panel B, the prediction performance of the Bayesian envelope model appears to be quite similar to those of the frequentist envelope and RRR models. RemMap turns out to have a poorer prediction performance, comparable to the Bayesian standard model.

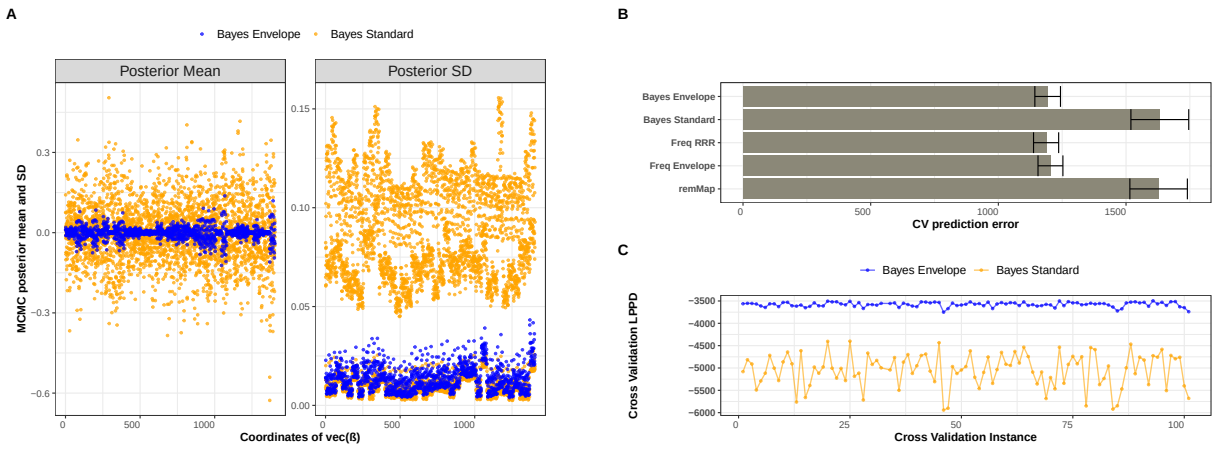

Figure 2: Panel A: Posterior means and standard deviations for all elements in $\boldsymbol{\beta}$. Panel B: Cross validation prediction errors (replication mean $\pm$ SD) for five competing models. Panel C: Cross-validation LPPDs for the Bayesian envelope and standard models.

In panel C, the LPPDs show uniform superiority of the Bayesian envelope model over the Bayesian standard model.

Finally, we looked into the posterior distributions of $\boldsymbol{\beta}$ to investigate the associations between isoprenoid biosynthesis pathway genes (predictors) and downstream pathway genes (responses). For this purpose, we computed a 95% posterior credible interval (based on 2.5th and 97.5th percentile) for each element in $\boldsymbol{\beta}$ under both the envelope and standard model. We labeled an element as "significant" if the corresponding credible interval excluded zero, and "non-significant" otherwise. The results are displayed in Figure 3 as heatmaps. The figure depicts a noticeably more regular pattern for the significant associations obtained from the envelope model compared to the standard model. This is a consequence of the reduced dimensionality (recall that $\Pr(u = 1 \mid \text{data}) \approx 0.998$), which aids noise reduction. The envelope analysis suggests that only a handful of isoprenoid

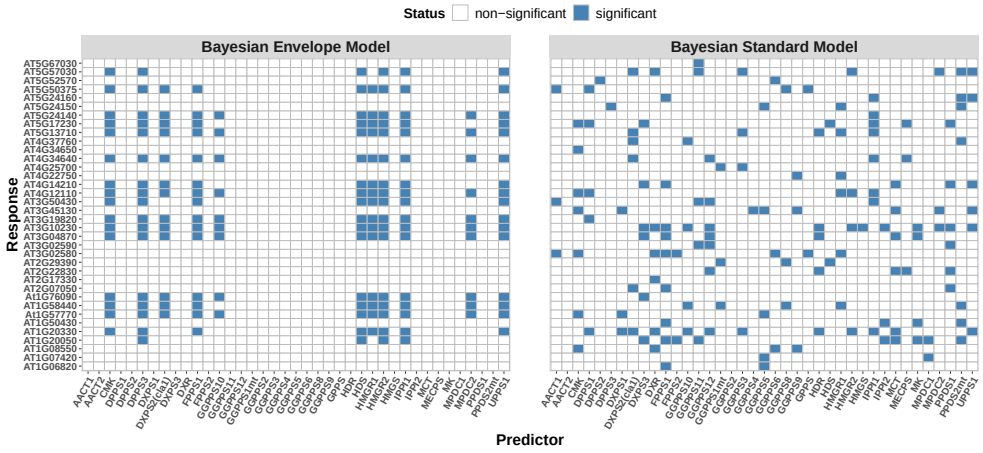

Figure 3: Significance of regression coefficients under Bayesian envelope model and Bayesian standard model.

biosynthesis pathway genes, such as CMK, DXPS2, FPPS1, and HDS, are primarily associated with the downstream pathway genes. Our findings are consistent with existing scientific knowledge. In particular, the association of DXPS2 and HDS with downstream pathway carotenoid, and the association of FPPS1 and downstream pathway phytosterol has been demonstrated in Wille et al. (2004). In contrast, the signals detected in the standard model are noticeably noisier and more sporadic, due to the presence of immaterial variability in the estimates. This is also consistent with the larger posterior standard deviations presented in panel A of Figure 2. By appropriately identifying and discarding the immaterial variability, the envelope model enhances signal detection. The signal detection performance of the frequentist envelope and RRR models is included in Supplement C.17.

7.2.2 Predictor Envelope Model

The yeast cell cycle dataset was analyzed in Chun and Keleş (2010) under the context of sparse PLS. It contains the mRNA levels measured at 18 evenly spaced time points and the binding information of 106 transcription factors for 542 genes. It is known that transcription factors control the rate at which DNA is transcribed into mRNA. So we took the measurements of mRNA levels as responses and the binding information of transcription factors as predictors. We ran Algorithm D.1 to generate 10,000 MCMC samples after discarding a burn-in of 5,000 iterations for every $m = 0, 1, \ldots, 106$. Up to 3 decimal places, the marginal posterior of m had $\Pr(m = 2 \mid \text{data}) = 1$. We subsequently focused on the predictor envelope model with $m = 2$, which is equivalent to the envelope BMA model. We compared the posterior mean and standard deviation of the Bayesian predictor envelope model and the Bayesian standard model for each element in $\boldsymbol{\beta}$, and the results are in the left panel Figure 4. The posterior standard deviations of the standard model are noticeably larger than those of the envelope model, which confirms the efficiency gains obtained by the envelope estimator.

We also investigate the prediction performance via LPPD values. We followed the same procedure as in Section 7.2.1 and obtained the test-set LPPD values for Bayesian envelope and standard models based on 100 random $80 - 20\%$ partitions of the data. The right panel of Figure 4 shows that for each random partition, the LPPD for the Bayesian predictor envelope model is higher than that of the standard model, thus confirming a better prediction performance for the envelope model.

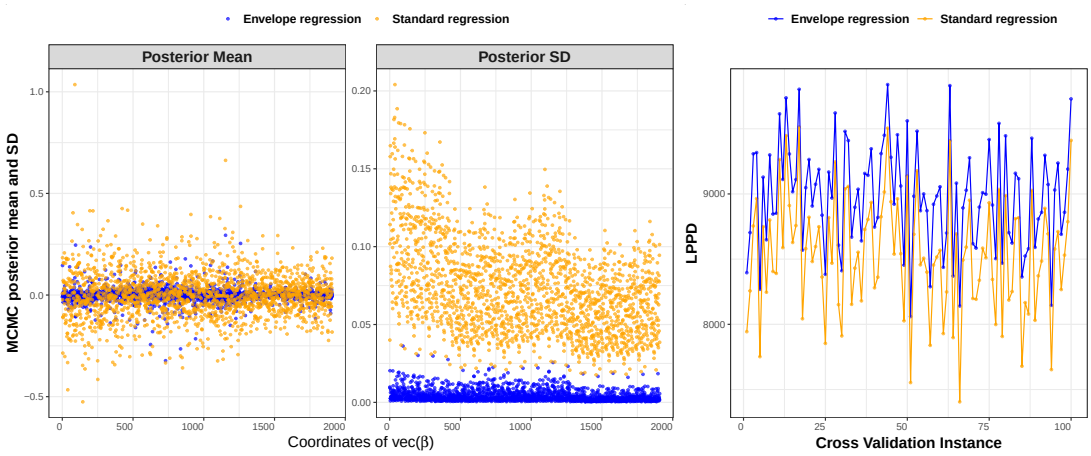

Figure 4: Left panel: Comparison on posterior means and standard deviations for all elements in $\boldsymbol{\beta}$. Right panel: Comparison on LPPD.

7.2.3 Envelope Probit Model

The wine dataset (Aeberhard et al., 1992) contains measurements on 13 characteristics of wines from two different cultivars in the same region of Italy. There are 48 and 71 samples from the first and second cultivars respectively. A binary response takes value 1 if the wine is from the second cultivar and 0 otherwise. For demonstration purposes, we only keep $p = 8$ predictors, which are malic acid, ash, alkalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols and color intensity. This is because having all characteristics in the model creates a perfect linear separation in the data, making estimation of $\boldsymbol{\beta}$ an ill-posed problem. We ran Algorithm E.1 (supplement) to generate 20,000 MCMC samples after discarding a burn-in of 80,000 iterations. The posterior probabilities $\Pr(m \mid \text{data})$ were obtained to be 0.053, 0.568, 0.302, 0.04, 0.018, and 0.019 for $m = 3$, 4, 5, 6, 7, and 8 respectively. If we take a point estimator, then $\hat{m} = \arg\max_k \Pr(m = k \mid \text{data}) = 4$.

Table 3 compares the MCMC posterior mean and standard deviation for each element in $\boldsymbol{\beta}$ from the Bayesian envelope probit model with $\hat{m} = 4$, envelope BMA, and the Bayesian standard probit model. We note that the posterior standard deviations of the standard probit estimator are about twice as large as their envelope model counterparts, again exhibiting the efficiency gains achieved by the envelope model.

Model	Measure	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Envelope $\hat{m}$	mean	-0.86	-0.16	-0.09	-0.04	2.51	-0.22	2.89	-1.72
	sd	0.45	0.13	0.21	0.04	1.05	0.11	1.15	0.47
Envelope BMA	mean	-1.21	-0.24	0.11	-0.05	1.38	0.21	1.26	-1.89
	sd	0.59	1.34	0.23	0.04	1.77	2.95	1.48	0.74
Standard	mean	-1.56	-8.00	0.17	0.06	7.87	17.73	3.51	-4.84
	sd	0.78	5.40	0.32	0.08	4.05	12.26	1.97	2.22

Table 3: Posterior means and standard deviations of each element in $\boldsymbol{\beta}$ from the envelope probit model with $\hat{m} = 4$, envelope BMA and the Bayesian standard probit model.

8 Discussion

This paper proposes a Bayesian framework for envelope models aiding a unified approach for multiple different contexts. Our framework can potentially be further extended to derive novel Bayesian envelope methodologies for several other contexts, such as the quantile/expectile envelope model, envelope models in matrix/tensor variate regression, and envelope models in reduced rank regression, to name a few. Variable selection, either on predictor variables or response variables, can be accommodated in this framework through

sparsity-inducing priors such as the spike and slab priors (Mitchell and Beauchamp, 1988), global-local priors (Polson and Scott, 2010) or beta-prime priors (Bai and Ghosh, 2021). The Bayesian sparse predictor envelope model may give rise to a Bayesian sparse PLS; such a model is currently under investigation.

9 Acknowledgement

This work was partly supported by grant DMS-1407460 from the US National Science Foundation, grant 632688 from Simons Foundation, and the Graduate School Fellowship at the University of Florida. The authors sincerely thank Prof. Hani Doss for his helpful comments. Computing support is provided by the Center for Computational Research at the University at Buffalo. We thank the Associate Editor and two reviewers for the constructive comments and helpful suggestions that have substantially improved the paper.

References

Aeberhard, S., D. Coomans, and O. de Vel (1992). The classification performance of rda. Technical report, Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*(422), 669–679.

Bai, R. and M. Ghosh (2021). On the beta prime prior for scale parameters in high-dimensional bayesian regression models. *Statistica Sinica 31*, 1–23.

Bingham, C. (1974). An antipodally symmetric distribution on the sphere. *The Annals of Statistics 2*(6), 1201–1225.

Chan, K. S. and C. J. Geyer (1994). Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics 22*(4), 1747–1758.

Chen, M.-H. and Q.-M. Shao (2001). Propriety of posterior distribution for dichotomous quantal response models. *Proceedings of the American Mathematical Society 129*(1), 293–302.

Chen, T., Z. Su, Y. Yang, and S. Ding (2020). Efficient estimation in expectile regression using envelope models. *Electronic Journal of Statistics 14*(1), 143–173.

Chun, H. and S. Keleş (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society, Ser. B 72*(1), 3–25.

Conway, J. (1990). *A Course in Functional Analysis*. Graduate Texts in Mathematics. New York: Springer.

Cook, R., I. Helland, and Z. Su (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society, Ser. B 75*(5), 851–877.

Cook, R. D. (2018). *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics*. John Wiley & Sons.

Cook, R. D., L. Forzani, and Z. Su (2016). A note on fast envelope estimation. *Journal of Multivariate Analysis 150*, 42–54.

Cook, R. D., B. Li, and F. Chiaromonte (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica 20*, 927–960.

Cook, R. D. and X. Zhang (2015). Foundations for envelope models and methods. *Journal of the American Statistical Association 110*(510), 599–611.

De Jong, S. (1993). Simpls: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems 18*(3), 251–263.

Ding, S. and R. D. Cook (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society, Ser. B 80*(2), 387–408.

Ding, S., Z. Su, G. Zhu, and L. Wang (2021). Envelope quantile regression. *Statistica Sinica 31*, 79–106.

Eck, D. J. and R. D. Cook (2017). Weighted envelope estimation to handle variability in model selection. *Biometrika 104*(3), 743–749.

Gelman, A., H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Holmes, C. C. and L. Held (2006, 03). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis 1*(1), 145–168.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the american statistical association 90*(430), 773–795.

Khare, K., S. Pal, and Z. Su (2017). A Bayesian approach for envelope models. *The Annals of Statistics 45*(1), 196–222.

Li, L. and X. Zhang (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association 112*(519), 1131–1146.

Ma, Y. and L. Zhu (2013). Efficiency loss and the linearity condition in dimension reduction. *Biometrika 100*(2), 371–383.

Marsaglia, G. (1964). Generating a variable from the tail of the normal distribution. *Technometrics 6*(1), 101–102.

Meng, X.-L. and D. B. Rubin (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika 80*(2), 267–278.

Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*(404), 1023–1032.

Mukherjee, A., K. Chen, N. Wang, and J. Zhu (2015). On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika 102*(2), 457–477.

Peng, J., J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack, and P. Wang (2010, March). Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer. *The annals of applied statistics 4*(1), 53–77.

Polson, N. G. and J. G. Scott (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics 9*, 501–538.

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using pólya-gamma latent variables. *Journal of the American Statistical Association 108*(504), 1339–1349.

Rekabdarkolaee, H. M., Q. Wang, Z. Naji, and M. Fuentes (2019). New parsimonious multivariate spatial model: Spatial envelope. *Statistica Sinica To appear*.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing 5*(2), 121–125.

Roberts, G. O. and J. S. Rosenthal (2006). Harris recurrence of metropolis-within-gibbs and trans-dimensional markov chains. *The Annals of Applied Probability 16*(4), 2123–2139.

Su, Z., G. Zhu, X. Chen, and Y. Yang (2016). Sparse envelope model: efficient estimation and response variable selection in multivariate linear regression. *Biometrika 103*(3), 579–593.

Tierney, L. (1994, 12). Markov chains for exploring posterior distributions. *The Annals of Statistics 22*(4), 1701–1728.

Vidaurre, D., M. A. van Gerven, C. Bielza, P. Larrañaga, and T. Heskes (2013). Bayesian sparse partial least squares. *Neural Computation 25*(12), 3318–3339.

Wille, A., P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelić, P. von Rohr, L. Thiele, et al. (2004). Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biology 5*(11), R92.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. Krishnaiah (Ed.), *In Multivariate Analysis*, Volume 59, pp. 391–420. Academic Press, NY.