High-dimensional multivariate posterior consistency under global-local shrinkage priors

Ray Bai[a,*], Malay Ghosh[a]

[a]*Department of Statistics, University of Florida, Gainesville, FL 32611, USA*

Abstract

We consider sparse Bayesian estimation in the classical multivariate linear regression model with p regressors and q response variables. In univariate Bayesian linear regression with a single response y, shrinkage priors which can be expressed as scale mixtures of normal densities are popular for obtaining sparse estimates of the coefficients. In this paper, we extend the use of these priors to the multivariate case to estimate a $p \times q$ coefficients matrix $\mathbf{B}$. We derive sufficient conditions for posterior consistency under the Bayesian multivariate linear regression framework and prove that our method achieves posterior consistency even when $p > n$ and even when p grows at nearly exponential rate with the sample size. We derive an efficient Gibbs sampling algorithm and provide the implementation in a comprehensive R package called `MBSP`. Finally, we demonstrate through simulations and data analysis that our model has excellent finite sample performance.

Keywords: Heavy tail, High-dimensional data, Posterior consistency, Shrinkage estimation, Sparsity, Variable selection

1. Introduction

1.1. Background

We consider the classical multivariate normal linear regression model, viz.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{1}$$

where $\mathbf{Y} = (y_1, \ldots, y_q)$ is an $n \times q$ response matrix of n samples and q continuous response variables, $\mathbf{X}$ is an $n \times p$ matrix of n samples and p covariates, $\mathbf{B} \in \mathbb{R}^{p \times q}$ is the coefficient matrix, and $\mathbf{E} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is an $n \times q$ noise matrix. Under normality, we assume that $\varepsilon_1, \ldots, \varepsilon_n$ are iid $\mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma})$. In other words, each row of $\mathbf{E}$ is identically distributed with mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}$. Throughout this paper, we also assume that $\mathbf{Y}$ and $\mathbf{X}$ are centered so there is no intercept term in $\mathbf{B}$.

Our focus is on sparse Bayesian estimation and variable selection on the coefficients matrix $\mathbf{B}$ in (1). In practical settings, particularly in high-dimensional settings when $p > n$, it is important not only to provide robust estimates of $\mathbf{B}$, but to choose a subset of regressor variables from the p rows of $\mathbf{B}$ which are good for prediction on the q responses. Although p may be large, the number of predictors that are actually associated with the responses is generally quite small. A parsimonious model also tends to give far better estimation and prediction performance than a dense model, which further motivates the need for sparse estimates of $\mathbf{B}$.

In frequentist approaches to univariate regression, the most commonly used method for inducing sparsity is through imposing regularization penalties on the coefficients of interest. Popular choices of penalty functions include the lasso [40] and its many variants, e.g., [38, 47, 50, 51]. Many of these penalized regression methods include

*Corresponding author.
Email address: `raybai07@ufl.edu` (Ray Bai)

either an ℓ_1 penalty function or a combination of an ℓ_1 and ℓ_2 penalty to shrink irrelevant predictors or groups of predictors to exactly zero.

These ℓ_1 and ℓ_2 regularization methods have been naturally extended to the multivariate regression setup where sparsity in the coefficients matrix is desired. For example, Rothman et al. [34] used an ℓ_1 penalty on each individual coefficient of $\mathbf{B}$ in (1), in addition to an ℓ_1 penalty on the off-diagonal entries of the covariance matrix to perform joint sparse estimation of $\mathbf{B}$ and $\mathbf{\Sigma}$. Li et al. [26] proposed the multivariate sparse group lasso, which utilizes a combination of a group ℓ_2 penalty on rows of $\mathbf{B}$ and an ℓ_1 penalty on the individual coefficients b_{ij} to perform sparse estimation and variable selection at both the group and within-group levels. Wilms and Croux [45] also consider a model which imposes an ℓ_2 penalty on the rows of $\mathbf{B}$ to shrink entire rows to zero, while simultaneously estimating the covariance matrix $\mathbf{\Sigma}$. Much of the frequentist literature on sparse estimation of (1) has focused on producing robust point estimates of $\mathbf{B}$, rather than on characterizing uncertainty of the estimates. In contrast, Bayesian methods naturally provide a vehicle for uncertainty quantification through the posterior density.

In the Bayesian univariate regression model, spike-and-slab priors, introduced by Mitchell and Beauchamp [30], have been a popular choice for inducing sparsity in the coefficients for regression problems. These priors are a mixture density with a point mass at zero used to force some coefficients to zero (the "spike") and a continuous density (the "slab") to model the nonzero coefficients. Since then, many variants of spike-and-slab have been developed. George and McCulloch [20] proposed a stochastic search variable selection (SSVS) method, which places a mixture prior of two normal densities with different variances (one small and one large) on each of the coefficients and which facilitates efficient Gibbs sampling. Recently, Ishwaran and Rao [25] and Narisetty and He [31] also used the mixture prior of normals but used rescaling of the variances (dependent upon the sample size n) in order to better control the amount of shrinkage for each individual coefficient. In order to perform group estimation and group variable selection, Xu and Ghosh [46] also introduced the Bayesian group lasso with spike-and-slab priors (BGL-SS), which is a mixture density with a point mass at a vector $\mathbf{0}_{m_g} \in \mathbb{R}^{m_g}$, where m_g denotes the size of group g and a normal distribution to model the "slab."

This two-components mixture approach has been extended to the multivariate framework by several authors [9, 28, 29]. In particular, Brown et al. [9] and Liquet et al. [29] first facilitate variable selection by associating each of the p rows $\mathbf{b}_1, \ldots, \mathbf{b}_p$ of $\mathbf{B}$ with a p-dimensional binary vector $\gamma = (\gamma_1, \ldots, \gamma_p)$, where each entry in γ follows a Bernoulli distribution. The selected $\mathbf{b}_i$s are then estimated by placing a multivariate Zellner g-prior (see Zellner [48]) on the sub-matrix of the selected covariates. Liquet et al. [28] extend the work of Xu and Ghosh [46] to the multivariate case with a method called Multivariate Group Selection with Spike and Slab Prior (MBGL-SS). Under MBGL-SS, rows of $\mathbf{B}$ are grouped together and modeled with a prior mixture density with a point mass at $\mathbf{0} \in \mathbb{R}^{m_g q}$ having positive probability (where m_g denotes the size of the gth group and q is the number of responses). Liquet et al. [28] use the posterior median $\widehat{\mathbf{B}} = (\widehat{b_{ij}})_{p\times q}$ as the estimate for $\mathbf{B}$, so that entire rows are estimated to be exactly zero.

Finally, both frequentist and Bayesian reduced rank regression (RRR) approaches have been developed to tackle the problem of sparse estimation of $\mathbf{B}$ in (1). RRR constrains the coefficient matrix $\mathbf{B}$ to be rank-deficient. Chen and Huang [14] proposed a rank-constrained adaptive group lasso approach to recover a low-rank matrix with some rows of $\mathbf{B}$ estimated to be exactly zero. Bunea et al. [10] also proposed a joint sparse and low-rank estimation approach and derived its non-asymptotic oracle bounds. The RRR approach was recently adapted to the Bayesian framework by Goh et al. [23] and Zhu et al. [49]. In the Bayesian framework, rank-reducing priors are used to shrink most of the rows and columns in $\mathbf{B}$ towards $\mathbf{0}_p \in \mathbb{R}^p$ or $\mathbf{0}_q^\top \in \mathbb{R}^q$.

1.2. Global-local shrinkage priors

When p is large, (point mass) spike-and-slab priors can face computational problems since they require either searching over 2^p possible models. This has led to the creation of a wide number of absolutely continuous shrinkage priors which behave similarly to spike-and-slab priors but which require significantly less computational effort. In univariate regression, these priors can be placed on each of the individual coefficients $\beta_1, \ldots, \beta_p$ and are represented as scale-mixtures of Gaussian distributions, viz.

$$\beta_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \tau\xi_i), \quad \xi_i \sim \pi(\xi_i), \tag{2}$$

where $\pi(\xi_i)$ typically follows a heavy-tailed density. These types of priors are known as global-local (GL) shrinkage priors. For GL priors, τ represents a global parameter that shrinks all coefficients to zero, while ξ_i is a tuning parameter

that controls the degree of shrinkage for each individual β_i. These priors contain significant probability around zero so that most coefficients are shrunk to zero. However, they retain heavy enough tails in order to correctly identify and prevent overshrinkage of the true signals (or non-zero coefficients). This combination of heavy mass around zero and tail robustness makes global-local shrinkage priors especially appealing when inducing sparsity.

Examples of GL shrinkage priors include the popular horseshoe prior [12] and the Bayesian lasso [32]. Priors of the type (2) have also been considered by numerous authors; see, e.g., [3, 4, 6, 24, 33, 37]. Armagan et al. [2] noted that a number of these priors utilize a beta prime density as the prior for $\pi(\xi_i)$. This family of generalized beta priors was first studied by Libby and Novick [27], and Armagan et al. [2] referred to this general class of shrinkage priors as the "three parameter beta normal" (TPBN) mixture family. The TPBN family in particular includes the horseshoe, the Strawderman–Berger [4, 37], and the normal-exponential-gamma (NEG) [24] priors. Polson and Scott [33] also generalized the beta prime density to the family of hypergeometric inverted beta (HIB) priors. Finally, Armagan et al. [3] introduced another general class of priors called the generalized double Pareto (GDP) family.

These priors have been studied extensively and have been shown to have a number of good theoretical properties. For example, Armagan et al. [1] gave sufficient conditions for posterior consistency in univariate linear regression when several well-known shrinkage priors are placed on the coefficients. Ghosh and Chakrabarti [21] and van der Pas et al. [41, 42] showed that when these priors are used to estimate sparse normal means, the posterior distributions concentrate around the true means at the minimax rate under mild conditions. van der Pas et al. [44] also obtained minimax-optimal posterior contraction rates for the horseshoe under both empirical Bayes and hierarchical Bayesian choices for the global shrinkage parameter τ in (2). For the normal means model, the theoretical properties of model selection (including the variable selection method applied in this article) and uncertainty quantification under scale-mixture priors were also recently investigated by Salomond [36] and van der Pas et al. [43]. Finally, in the context of multiple hypothesis testing, Bhadra et al. [6], Datta and Ghosh [16], Ghosh and Chakrabarti [21], and Ghosh et al. [22] showed that multiple testing rules induced by these shrinkage priors can achieve optimal Bayes risk in terms of 0–1 symmetric loss (or expected number of misclassified signals).

Ghosh et al. [22] observed that for a large number of global-local shrinkage priors of the form (2), the local parameter ξ_i has a hyperprior distribution $\pi(\xi_i)$ that can be written as

$$\pi(\xi_i) = K\xi_i^{-a-1}L(\xi_i), \tag{3}$$

where $K > 0$ is the constant of proportionality, a is positive real number, and L is a positive measurable, non-constant, slowly varying function over $(0, \infty)$.

Definition 1. *A positive measurable function L defined over (A, ∞), for some $A \geq 0$, is said to be slowly varying (in Karamata's sense) if for every fixed $\alpha > 0$, $L(\alpha x)/L(x) \to 1$ as $x \to \infty$.*

A thorough treatment of functions of this type can be found in the classical text by Bingham et al. [8]. Table 1 provides a list of several well-known global-local shrinkage priors that fall in the class of priors of the form (2), the corresponding density $\pi(\xi_i)$ for ξ_i, and the slowly-varying component $L(\xi_i)$ in (3). Following Tang et al. [39], we refer to these scale-mixture priors as polynomial-tailed priors.

Although polynomial-tailed priors have been studied extensively in univariate regression, their potential utility for multivariate analysis seems to have been largely overlooked. In this paper, we introduce a new Bayesian approach

Table 1: Polynomial-tailed priors, their respective prior densities for $\pi(\xi_i)$ up to normalizing constant C, and the slowly-varying component $L(\xi_i)$.

Prior	$\pi(\xi_i)/C$	$L(\xi_i)$
Student's t	$\xi_i^{-a-1}\exp(-a/\xi_i)$	$\exp(-a/\xi_i)$
Horseshoe	$\xi_i^{-1/2}(1+\xi_i)^{-1}$	$\xi_i^{a+1/2}/(1+\xi_i)$
Horseshoe+	$\xi_i^{-1/2}(\xi_i-1)^{-1}\ln(\xi_i)$	$\xi_i^{a+1/2}(\xi_i-1)^{-1}\ln(\xi_i)$
NEG	$(1+\xi_i)^{-1-a}$	$\{\xi_i/(1+\xi_i)\}^{a+1}$
TPBN	$\xi_i^{u-1}(1+\xi_i)^{-a-u}$	$\{\xi_i/(1+\xi_i)\}^{a+u}$
GDP	$\int_0^\infty (\lambda^2/2)\exp(-\lambda^2\xi_i/2)\lambda^{2a-1}\exp(-\eta\lambda)d\lambda$	$\int_0^\infty t^a \exp(-t-\eta\sqrt{2t/\xi_i})dt$
HIB	$\xi_i^{u-1}(1+\xi_i)^{-(a+u)}\exp\left(-\frac{s}{1+\xi_i}\right)\times\left(\phi^2+\frac{1-\phi^2}{1+\xi_i}\right)^{-1}$	$\{\xi_i/(1+\xi_i)\}^{a+u}\exp\left(-\frac{s}{1+\xi_i}\right)\left(\phi^2+\frac{1-\phi^2}{1+\xi_i}\right)^{-1}$

for estimating the unknown $p \times q$ coefficient matrix $\mathbf{B}$ in (1) using polynomial-tailed priors. We call our method the Multivariate Bayesian model with Shrinkage Priors (MBSP).

While there have been many methodological developments for Bayesian multivariate linear regression, theoretical results in this domain have not kept pace with applications. There appears to be very little theoretical justification for adopting Bayesian methodology in multivariate regression. In this article, we take a step towards resolving this gap by providing sufficient conditions under which Bayesian multivariate linear regression models can obtain posterior consistency. To our knowledge, our paper is the first one to give general conditions for posterior consistency under model (1) when $p > n$ and when p grows at nearly exponential rate with sample size n. We further illustrate that our method based on polynomial-tailed priors achieves strong posterior consistency in both low-dimensional and ultra-high-dimensional settings.

The rest of our paper is organized as follows. In Section 2, we introduce the MBSP model and provide some insight into how it facilitates sparse estimation and variable selection. In Section 3, we present sufficient conditions for our model to achieve posterior consistency in both the cases where p grows slower than n and the case when p grows at nearly exponential rate with n. In Section 4, we show how to implement MBSP using the TPBN family of priors as a special case and how to utilize our method for variable selection. Efficient Gibbs sampling and computational complexity considerations are also discussed. In Section 5, we illustrate our method's finite-sample performance through simulations and analysis of a real data set. Finally, in Section 6, we discuss some directions for future research.

2. Multivariate Bayesian model with shrinkage priors (MBSP)

2.1. Preliminary notation and definitions

We first introduce the following notation and definitions.

Definition 2. *A random matrix* $\mathbf{Y}$ *is said to have the matrix-normal density if* $\mathbf{Y}$ *has the density function (on the space* $\mathbb{R}^{a\times b}$*):*

$$f(\mathbf{Y}) = \frac{|\mathbf{U}|^{-b/2}|\mathbf{V}|^{-a/2}}{(2\pi)^{ab/2}} e^{-\frac{1}{2}\mathrm{tr}\{\mathbf{U}^{-1}(\mathbf{Y}-\mathbf{M})\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{M})^\top\}}, \tag{4}$$

where $\mathbf{M} \in \mathbb{R}^{a\times b}$*, and* $\mathbf{U}$ *and* $\mathbf{V}$ *are positive definite matrices of dimension* $a \times a$ *and* $b \times b$ *respectively. If* $\mathbf{Y}$ *is distributed as a matrix-normal distribution with pdf given in* (4)*, we write* $\mathbf{Y} \sim \mathcal{MN}_{a\times b}(\boldsymbol{M, U, V})$*.*

Definition 3. *The matrix* $\mathbf{O} \in \mathbb{R}^{a\times b}$ *denotes the* $a \times b$ *matrix with all zero entries.*

2.2. MBSP model

Our multivariate Bayesian model formulation for model (1) with shrinkage priors (henceforth referred to as MBSP) is as follows:

$$\mathbf{Y}|\mathbf{X}, \mathbf{B}, \mathbf{\Sigma} \sim \mathcal{MN}_{n\times q}(\mathbf{XB}, \mathbf{I}_n, \mathbf{\Sigma}), \quad \mathbf{B}|\xi_1, \ldots, \xi_p, \mathbf{\Sigma} \sim \mathcal{MN}_{p\times q}[\mathbf{O}, \tau\,\mathrm{diag}(\xi_1, \ldots, \xi_p), \mathbf{\Sigma}], \quad \xi_1, \ldots, \xi_p \overset{\text{ind}}{\sim} \pi, \tag{5}$$

where π is a polynomial-tailed prior density of the form (3).

2.3. Handling sparsity

In this section, we illustrate how the MBSP model induces sparsity. First note that in (5), an alternative way of writing the density $\mathbf{Y}|\mathbf{X}, \mathbf{B}, \mathbf{\Sigma}$ is

$$\mathbf{Y}|\mathbf{X}, \mathbf{B}, \mathbf{\Sigma} \propto |\mathbf{\Sigma}|^{-nq/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i - \sum_{j=1}^{p} x_{ij}\mathbf{b}_j\right)^\top \mathbf{\Sigma}^{-1}\left(\mathbf{y}_i - \sum_{j=1}^{p} x_{ij}\mathbf{b}_j\right)\right\}, \tag{6}$$

where $\mathbf{b}_j$ denotes the jth row of $\mathbf{B}$.

Following from (6), we see that under (5) and known $\mathbf{\Sigma}$, the joint prior density $\pi(\mathbf{B}, \xi_1, \ldots, \xi_p)$ is

$$\pi(\mathbf{B}, \xi_1, \ldots, \xi_p) \propto \prod_{j=1}^{p} \xi_j^{-q/2} e^{-\frac{1}{2\xi_j}\|\mathbf{b}_j(\tau\mathbf{\Sigma})^{-1/2}\|_2^2} \pi(\xi_j), \tag{7}$$

where $\|\cdot\|_2$ denotes the ℓ_2 vector norm. Since the p rows of $\mathbf{B}$ are independent, we see from (7) that this choice of prior induces sparsity on the rows of $\mathbf{B}$, while also accounting for the covariance structure of the q responses. This ultimately facilitates sparse estimation of $\mathbf{B}$ as a whole and variable selection from the p regressors.

For example, if $\pi(\xi_j) \overset{\text{ind}}{\sim} \mathcal{IG}(\alpha_j, \gamma_j/2)$ (where $\mathcal{IG}$ denotes the inverse-gamma density), then the marginal density for $\mathbf{B}$ (after integrating out the ξ_js) is proportional to

$$\prod_{j=1}^{p} \left\{\|\mathbf{b}_j(\tau\mathbf{\Sigma})^{-1/2}\|_2^2 + \gamma_j\right\}^{-(\alpha_j+q/2)}, \tag{8}$$

which corresponds to a multivariate Student's t density.

On the other hand, if $\pi(\xi_j) \propto \xi_j^{q/2-1}(1+\xi_j)^{-1}$, then the joint density in is proportional to

$$\prod_{j=1}^{p} \xi_j^{-1}(1+\xi_j)^{-1} e^{-\frac{1}{2\xi_j}\|\mathbf{b}_j(\tau\mathbf{\Sigma})^{-1/2}\|_2^2}, \tag{9}$$

and integrating out the ξ_js gives a multivariate horseshoe density function.

As examples (8) and (9) demonstrate, our model allows us to obtain sparse estimates of $\mathbf{B}$ by inducing row-wise sparsity in $\mathbf{B}$ with a matrix-normal scale mixture using global-local shrinkage priors. This row-wise sparsity also facilitates variable selection from the p variables.

3. Posterior consistency of MBSP

3.1. Notation

We first introduce some notation that will be used throughout the paper. For any two sequences of positive real numbers a_n and b_n with $b_n \neq 0$, we write $a_n = O(b_n)$ if $|a_n/b_n| \leq M$ for all $n \in \mathbb{N}$, for some positive real number M independent of n, and $a_n = o(b_n)$ to denote $\lim_{n\to\infty} a_n/b_n = 0$. Therefore, $a_n = o(1)$ if $\lim_{n\to\infty} a_n = 0$.

For a vector $v \in \mathbb{R}^n$, $\|v\|_2 = (v_1^2 + \cdots + v_n^2)^{1/2}$ denote the ℓ_2 norm. For a matrix $\mathbf{A} \in \mathbb{R}^{a\times b}$ with entries a_{ij}, $\|\mathbf{A}\|_F = \{\text{tr}(\mathbf{A}^\top\mathbf{A})\}^{1/2} = \{\sum_{i=1}^{a}\sum_{j=1}^{b} a_{ij}^2\}^{1/2}$ denotes the Frobenius norm of $\mathbf{A}$. For a symmetric matrix $\mathbf{A}$, we denote its minimum and maximum eigenvalues by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ respectively. Finally, for an arbitrary set $\mathcal{A}$, we denote its cardinality by $|\mathcal{A}|$.

3.2. Definition of posterior consistency

For this section, we denote the number of predictors by p_n to emphasize that p depends on n and is allowed to grow with n. Suppose that the true model is

$$\mathbf{Y}_n = \mathbf{X}_n\mathbf{B}_{0n} + \mathbf{E}_n, \tag{10}$$

where $\mathbf{Y}_n = (\mathbf{Y}_{n,1}, \ldots, \mathbf{Y}_{n,q})$ and $\mathbf{E}_n \sim \mathcal{MN}_{n\times q}(\mathbf{O}, \mathbf{I}_n, \mathbf{\Sigma})$. For convenience, we denote $\mathbf{B}_{0n}$ as $\mathbf{B}_0$ going forward, noting $\mathbf{B}_0$ depends on p_n (and therefore on n).

Let $\{\mathbf{B}_0\}_{n\geq 1}$ be the sequence of true coefficient matrices, and let $\mathbb{P}_0$ denote the distribution of $\{\mathbf{Y}_n\}_{n\geq 1}$ under (10). Let $\{\pi_n(\mathbf{B}_n)\}_{n\geq 1}$ and $\{\pi_n(\mathbf{B}_n|\mathbf{Y}_n)\}_{n\geq 1}$ denote the sequences of prior and posterior densities for coefficients matrix $\mathbf{B}_n$. Analogously, let $\{\Pi_n(\mathbf{B}_n)\}_{n\geq 1}$ and $\{\Pi_n(\mathbf{B}_n|\mathbf{Y}_n)\}_{n\geq 1}$ denote the sequences of prior and posterior distributions. In order to achieve consistent estimation of $\mathbf{B}_0(\equiv \mathbf{B}_{0n})$, the posterior probability that $\mathbf{B}_n$ lies in a ε-neighborhood of $\mathbf{B}_0$ should converge to 1 almost surely with respect to $\mathbb{P}_0$ measure as $n \to \infty$. We therefore define strong posterior consistency as follows.

Definition 4. ***(Posterior consistency)*** *Let* $\mathcal{B}_n = \{\mathbf{B}_n : \|\mathbf{B}_n - \mathbf{B}_0\|_F > \varepsilon\}$*, where* $\varepsilon > 0$*. The sequence of posterior distributions of* $\mathbf{B}_n$ *under prior* $\pi_n(\mathbf{B}_n)$ *is said to be strongly consistent under* (10) *if, for any* $\varepsilon > 0$*,*

$$\Pi_n(\mathcal{B}_n|Y_n) = \Pi_n(\|\mathbf{B}_n - \mathbf{B}_0\|_F > \varepsilon|Y_n) \to 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \to \infty.$$

Using Definition 4, we now state two general theorems and a corollary that provide general conditions under which priors on $\mathbf{B}$ (not just the MBSP model) may achieve strong posterior consistency in both low-dimensional and ultra-high-dimensional settings.

3.3. Sufficient conditions for posterior consistency

For our theoretical investigation, we assume $\mathbf{\Sigma}$ to be fixed and known and dimension of the response variables q to be fixed. In practice, $\mathbf{\Sigma}$ is typically unknown, and one can estimate it from the data. In Section 4, we present a fully Bayesian implementation of MBSP by placing an appropriate inverse-Wishart prior on $\mathbf{\Sigma}$.

Theorem 1 applies to the case where the number of predictors p_n diverges to ∞ at a rate slower than n as $n \to \infty$, while Theorem 2 applies to the case where p_n grows to ∞ at a faster rate than n as $n \to \infty$. To handle these two cases, we require different sets of regularity assumptions. Proofs for both theorems are shown in Section 1 of the Online Supplement (see Appendix A).

3.3.1. Low-dimensional case

We first impose the following regularity conditions which are all standard ones used in the literature and relatively mild; see, e.g., Armagan et al. [1]. In particular, Assumption (A2) ensures that the design matrix $\mathbf{X}_n^\top \mathbf{X}_n$ is positive definite for all $n \in \mathbb{N}$ and that $\mathbf{B}_0$ is estimable.

Regularity conditions

(A1) $p_n = o(n)$ and $p_n \leq n$ for all $n \in \mathbb{N}$.

(A2) There exist constants c_1, c_2 so that

$$0 < c_1 < \liminf_{n\to\infty} \lambda_{\min}(\mathbf{X}_n^\top \mathbf{X}_n/n) \leq \limsup_{n\to\infty} \lambda_{\max}(\mathbf{X}_n^\top \mathbf{X}_n/n) < c_2 < \infty.$$

(A3) There exist constants d_1 and d_2 so that $0 < d_1 < \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) < d_2 < \infty$.

Using these conditions, we are able to attain a very simple sufficient condition for strong posterior consistency under (10), as defined in Definition 4, which we state in the next theorem.

Theorem 1. *Assume that conditions (A1)–(A3) hold. Then the posterior of* $\mathbf{B}_n$ *under any prior* $\pi_n(\mathbf{B}_n)$ *is strongly consistent under* (10)*, i.e., for* $\mathcal{B}_n = \{\mathbf{B}_n : \|\mathbf{B}_n - \mathbf{B}_0\|_F > \varepsilon\}$ *and any arbitrary* $\varepsilon > 0$*,*

$$\Pi_n(\mathcal{B}_n|\mathbf{Y}_n) \to 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \to \infty$$

if

$$\Pi_n\left(\mathbf{B}_n : \|\mathbf{B}_n - \mathbf{B}_0\|_F < \frac{\Delta}{n^{\rho/2}}\right) > \exp(-kn) \tag{11}$$

for all $0 < \Delta < \varepsilon^2 c_1 d_1^{1/2}/(48 c_2^{1/2} d_2)$ *and* $0 < k < \varepsilon^2 c_1/(32 d_2) - 3\Delta c_2^{1/2}/(2 d_1^{1/2})$*, where* $\rho > 0$*.*

Condition (11) in Theorem 1 states that as long as the prior distribution for $\mathbf{B}_n$ captures $\mathbf{B}_0$ inside a ball of radius $\Delta/n^{\rho/2}$ with sufficiently high probability for large n, the posterior of $\mathbf{B}_n$ will be strongly consistent.

3.3.2. Ultra-high-dimensional case

To achieve posterior consistency when $p_n \gg n$ and $p_n \geq O(n)$, we require additional restrictions on the eigenstructure of $\mathbf{X}_n$ and an additional assumption on the size of the true model. Working under the assumption of sparsity, we assume that the true model (10) contains only a few nonzero predictors. That is, most of the rows of $\mathbf{B}_0$ should contain only zero entries.We denote $S^* \subset \{1, \ldots, p_n\}$ as the set of indices of the rows of $\mathbf{B}_0$ with at least one nonzero entry and let $s^* = |S^*|$ be the size of S^*. We need the following regularity conditions.

Regularity conditions

(B1) $p_n > n$ for all $n \in \mathbb{N}$, and $\ln(p_n) = O(n^d)$ for some $d \in (0, 1)$.

(B2) The rank of $\mathbf{X}_n$ is n.

(B3) Let $\mathcal{J}$ denote a set of indices, where $\mathcal{J} \subset \{1, \ldots, p_n\}$ such that $|\mathcal{J}| \leq n$. Let $\mathbf{X}_{\mathcal{J}}$ denote the submatrix of $\mathbf{X}_n$ that contains the columns with indices in $\mathcal{J}$. For any such set $\mathcal{J}$, there exists a finite constant $\widetilde{c}_1 > 0$ so that

$$\liminf_{n\to\infty} \lambda_{\min}(\mathbf{X}_{\mathcal{J}}^\top \mathbf{X}_{\mathcal{J}}/n) \geq \widetilde{c}_1.$$

(B4) There is finite constant $\widetilde{c}_2 > 0$ so that

$$\limsup_{n\to\infty} \lambda_{\max}(\mathbf{X}_n^\top \mathbf{X}_n/n) < \widetilde{c}_2.$$

(B5) There exist constants d_1 and d_2 so that $0 < d_1 < \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) < d_2 < \infty$.

(B6) S^* is nonempty for all $n \in \mathbb{N}$, and $s^* = o\{n/\ln(p_n)\}$.

Condition (B1) allows the number of predictors p_n to grow at nearly exponential rate. In particular, p_n may grow at a rate of e^{n^d}, where $d \in (0, 1)$. In the high-dimensional literature, it is a standard assumption that $\ln(p_n) = o(n)$. Condition (B3) assumes that for any submatrix of $\mathbf{X}_n$ that is full rank, its minimum singular value is bounded below by $n\widetilde{c}_1$. This condition is needed to overcome potential identifiability issues, since trivially, the smallest singular value of $\mathbf{X}_n$ is zero. (B4) imposes a supremum on the maximum singular value of $\mathbf{X}_n$, which poses no issue. Finally, Condition (B6) allows the true model size to grow with n but at a rate slower than $n/\ln(p_n)$. Condition (B6) is a standard requirement that has been used to establish estimation consistency when p_n grows at nearly exponential rate with n for frequentist point estimators, such as the Dantzig estimator [11], the scaled lasso [38], and the lasso [40]. In ultra-high-dimensional problems, it is generally agreed that s^* must be small relative to both p and n in order to attain estimation consistency and minimax convergence rates, and hence, this restriction on the growth rate of s^*.

Under these regularity conditions, we are able to attain a simple sufficient condition for posterior consistency under (10) even when p_n grows faster than n. Theorem 2 gives the sufficient condition for strong consistency.

Theorem 2. *Assume that conditions (B1)–(B6) hold. Then the posterior of* $\mathbf{B}_n$ *under any prior* $\pi_n(\mathbf{B}_n)$ *is strongly consistent under* (10)*, i.e., for* $\mathcal{B}_n = \{\mathbf{B}_n : \|\mathbf{B}_n - \mathbf{B}_0\|_F > \varepsilon\}$ *and any arbitrary* $\varepsilon > 0$*,*

$$\Pi_n(\mathcal{B}_n|\mathbf{Y}_n) \to 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \to \infty$$

if

$$\Pi_n\left(\mathbf{B}_n : \|\mathbf{B}_n - \mathbf{B}_0\|_F < \frac{\widetilde{\Delta}}{n^{\rho/2}}\right) > \exp(-kn) \tag{12}$$

for all $0 < \widetilde{\Delta} < \varepsilon^2\widetilde{c}_1 d_1^{1/2}/(48\widetilde{c}_2^{1/2} d_2)$ *and* $0 < k < \varepsilon^2\widetilde{c}_1/(32d_2) - 3\widetilde{\Delta}\widetilde{c}_2^{1/2}/(2d_1^{1/2})$*, where* $\rho > 0$*.*

Similar to (11) in Theorem 1, Condition (12) in Theorem 2 states that as long as the prior distribution for $\mathbf{B}_n$ captures $\mathbf{B}_0$ inside a ball of radius $\widetilde{\Delta}/n^{\rho/2}$ with sufficiently high probability for large n, the posterior of $\mathbf{B}_n$ will be strongly consistent. To our knowledge, our paper is the first one in the literature to address the issue of ultra-high-dimensional consistency in Bayesian multivariate linear regression. There has been very little theoretical investigation done in the framework of Bayesian multivariate regression, and our paper takes a step towards narrowing this gap.

Now that we have provided simple sufficient conditions for posterior consistency in Theorems 1 and 2, we are ready to state our main theorems which demonstrate the power of the MBSP model (5) under polynomial-tailed hyperpriors (3).

3.4. Sufficient conditions for posterior consistency of MBSP

We now establish posterior consistency under the MBSP model (5), assuming that $\mathbf{\Sigma}$ is fixed and known, q is fixed, and that $\tau = \tau_n$ is a tuning parameter that depends on n.

As in Section 3.3, we assume that most of the rows of $\mathbf{B}_0$ are zero, i.e., that the true model $S \subset \{1, \ldots, p_n\}$ is small relative to the total number of predictors. As before, we consider the cases where $p_n = o(n)$ and $p_n \geq O(n)$ separately. We also require the following regularity assumptions which turn out to be sufficient for both the low-dimensional and ultra-high-dimensional cases. Here, b^0_{jk} denotes an entry in $\mathbf{B}_0$.

Regularity conditions

(C1) For the slowly varying function $L(t)$ in the priors for $\xi_1, \ldots, \xi_p$, in (3), $\lim_{t\to\infty} L(t) \in (0, \infty)$. That is, there exists $c_0 > 0$ such that $L(t) \geq c_0$ for all $t \geq t_0$, for some t_0 which depends on both L and c_0.

(C2) There exists $M > 0$ so that $\sup_{j,k} |b^0_{jk}| \leq M < \infty$ for all $n \in \mathbb{N}$, i.e., the maximum entry in $\mathbf{B}_0$ is uniformly bounded above in absolute value.

(C3) $\tau_n \in (0, 1)$ for all $n \in \mathbb{N}$, and $\tau_n = o(p_n^{-1} n^{-\rho})$ for $\rho > 0$.

Remark 1. Condition (C1) is a very mild condition which ensures that L is slow-varying. Ghosh et al. [22] established that (C1) holds for L in the TPBN priors ($L(\xi_i) = (1 + \xi_i)^{-(\alpha+\beta)}$) and the GDP priors, viz.

$$L(\xi_i) = 2^{-\alpha/2-1} \int_0^\infty e^{-\beta\sqrt{2u/\xi_i}} e^{-u} u^{(\alpha/2+1)-1} du.$$

The TPBN family in particular includes many well-known one-group shrinkage priors, such as the horseshoe prior ($\alpha = 0.5, \beta = 0.5$), the Strawderman–Berger prior ($\alpha = 1$, $\beta = 0.5$), and the normal-exponential-gamma prior ($\alpha = 1, \beta > 0$). As remarked by Ghosh and Chakrabarti [21], one easily verifies that Assumption (C1) also holds for the inverse-gamma priors ($\pi(\xi_i) \propto \xi_i^{-\alpha-1} e^{-b/\xi_i}$) and the half-$t$ priors ($\pi(\xi_i) \propto (1 + \xi/\nu)^{-(\nu+1)/2}$).

Remark 2. Condition (C2) is a mild condition that bounds the entries of $\mathbf{B}_0$ in absolute value for all $n \in \mathbb{N}$, while (C3) specifies an appropriate rate of decay for τ_n. It is possible that the upper bound on the rate for τ_n can be loosened for individual GL priors. However, since we wish to encompass all possible priors of the form (3), we provide a general rate that works for all the polynomial-tailed priors considered in this paper.

We are now ready to state our main theorem for posterior consistency of the MBSP model. The proof of Theorem 3 can be found in Section 2 of the Online Supplement (see Appendix A).

Theorem 3 (**Low-dimensional case**). *Suppose that we have the MBSP model* (5) *with hyperpriors* (3)*. Provided that Assumptions (A1)–(A3) and (C1)–(C3) hold, our model achieves strong posterior consistency. That is, for any* $\varepsilon > 0$*,*

$$\Pi_n(\mathbf{B}_n : \|\mathbf{B}_n - \mathbf{B}_0\|_F > \varepsilon | \mathbf{Y}_n) \to 0 \quad a.s.\ \mathbb{P}_0 \text{ as } n \to \infty.$$

Theorem 3 establishes posterior consistency for the MBSP model only when $p_n = o(n)$. We also note that in the low-dimensional setting where $p_n = o(n)$, we place *no* restrictions on the growth on the number of nonzero predictors in the true model relative to sample size n. This contrasts with a previous result by Armagan et al. [1], who required that the number of true nonzero covariates grow slower than $n/\ln(n)$.

In the ultra-high-dimensional case where $p_n \geq O(n)$, we can still achieve posterior consistency under the MBSP model, with additional mild restrictions on the design matrix $\mathbf{X}_n$ and on the size of the true model. Theorem 4 deals with the ultra-high-dimensional scenario. The proof for Theorem 4 can be found in Section 2 of the Online Supplement (see Appendix A).

Theorem 4 (**Ultra high-dimensional case**). *Suppose that we have the MBSP model* (5) *with hyperpriors* (3)*. Provided that Assumptions (B1)–(B6) and (C1)–(C3) hold, our model achieves strong posterior consistency. That is, for any* $\varepsilon > 0$*,*

$$\Pi_n(\mathbf{B}_n : \|\mathbf{B}_n - \mathbf{B}_0\|_F > \varepsilon | \mathbf{Y}_n) \to 0 \quad a.s.\ \mathbb{P}_0 \text{ as } n \to \infty.$$

Interestingly enough, to ensure posterior consistency in the ultra-high-dimensional setting, the only thing that needs to be controlled is the tuning parameter τ_n, provided that our hyperpriors in (5) have the form (3). However, in the high-dimensional regime, p_n is allowed to grow at nearly exponential rate, and therefore, the rate of decay for τ_n from Condition (C3) necessarily needs to be much faster. Intuitively, this makes sense because we must sum over $p_n q$ terms in order to compute the Frobenius normed difference in Theorem 4.

Taken together, Theorems 3 and 4 both provide theoretical justification for the use of global-local shrinkage priors for multivariate linear regression. Even when we allow the number of predictors to grow at nearly exponential rate, the posterior distribution under MBSP (5) is able to consistently estimate $\mathbf{B}_0$ in (10). Our result is also very general in that a wide class of shrinkage priors, as indicated in Table 1, can be used for the hyperpriors ξ_is in (5).

4. Implementation of the MBSP model

In this section, we demonstrate how to implement the MBSP model using the three parameter beta normal (TPBN) mixture family [2, 27]. We choose the TPBN family because it is rich enough to generalize several well-known polynomial-tailed priors. Although we focus on the TPBN family, our model can easily be implemented for other global-local shrinkage priors (such as the Student's t prior or the generalized double Pareto prior) using similar techniques as the ones we describe below.

4.1. TPBN family

A random variable y said to follow the three parameter beta density, denoted as $\mathcal{TPB}(u, a, \tau)$, if

$$\pi(y) = \frac{\Gamma(u+a)}{\Gamma(u)\Gamma(a)}\,\tau^a y^{a-1}(1-y)^{u-1}\left\{1-(1-\tau)y\right\}^{-(u+a)}.$$

In univariate regression, a global-local shrinkage prior of the form given, for all $i \in \{1, \ldots, p\}$, by

$$\beta_i|\tau, \xi_i \sim \mathcal{N}(0, \tau\xi_i), \quad \pi(\xi_i) = \frac{\Gamma(u+a)}{\Gamma(u)\Gamma(a)}\,\xi_i^{u-1}(1+\xi_i)^{-(u+a)}, \tag{13}$$

may therefore be represented alternatively as

$$\beta_i|\nu_i \sim \mathcal{N}(0, \nu_i^{-1} - 1), \quad \nu_i \sim \mathcal{TPB}(u, a, \tau). \tag{14}$$

After integrating out ν_i in (14), the marginal prior for β_i is said to belong to the TPBN family. Special cases of (14) include the horseshoe prior ($u = 0.5, a = 0.5$), the Strawderman–Berger prior ($u = 1, a = 0.5$), and the normal-exponential-gamma (NEG) prior ($u = 1, a > 0$). By Proposition 1 of Armagan et al. [2], (13) and (14) can also be written as a hierarchical mixture of two Gamma distributions,

$$\beta_i|\psi_i \sim \mathcal{N}(0, \psi_i), \quad \psi_i|\zeta_i \sim \mathcal{G}(u, \zeta_i), \quad \zeta_i \sim \mathcal{G}(a, \tau), \tag{15}$$

where $\psi_i = \xi_i\tau$.

4.2. The MBSP-TPBN model

Taking our MBSP model (5) with the TPBN family as our chosen prior and placing an inverse-Wishart conjugate prior on $\mathbf{\Sigma}$, we can construct a specific variant of the MBSP model which we term the MBSP-TPBN model. For our theoretical study of MBSP, we assumed $\mathbf{\Sigma}$ to be known and the dimension of the responses q to be fixed (and thus, $q < n$ for large n). However, in order for our model to be implemented in finite samples, q can be of any size (including $q \gg n$), provided that the posterior distribution is proper. The use of an inverse-Wishart prior ensures posterior propriety.

Reparametrizing the variance terms $\tau\xi_1, \ldots, \tau\xi_p$ in terms of the ψ_is from (15), the MBSP-TPBN model is as follows:

$$\begin{aligned}
\mathbf{Y}|\mathbf{X}, \mathbf{B}, \mathbf{\Sigma} &\sim \mathcal{MN}_{n\times q}(\mathbf{XB}, \mathbf{I}_n, \mathbf{\Sigma}),\\
\mathbf{B}|\psi_1, \ldots, \psi_p, \mathbf{\Sigma} &\sim \mathcal{MN}_{p\times q}[\mathbf{O}, \mathrm{diag}(\psi_1, \ldots, \psi_p), \mathbf{\Sigma}],\\
\psi_i|\zeta_i &\overset{\text{ind}}{\sim} \mathcal{G}(u, \zeta_i), \quad i \in \{1, \ldots, p\},\\
\zeta_i &\overset{\text{iid}}{\sim} \mathcal{G}(a, \tau), \quad i \in \{1, \ldots, p\},\\
\mathbf{\Sigma} &\sim \mathcal{IW}(d, k\mathbf{I}_q),
\end{aligned} \tag{16}$$

where u, a, d, k, and τ are appropriately chosen hyperparameters. The MBSP-TPBN model can be implemented using the R package `MBSP`, which is available on the Comprehensive R Archive Network (CRAN).

4.2.1. Computational details

The full conditional densities under model (16) are available in closed form, and hence, can be implemented straightforwardly using Gibbs sampling. Moreover, by suitably modifying an algorithm introduced by Bhattacharya et al. [7] for drawing from the matrix-normal density (4), we can significantly reduce the computational complexity of sampling from the full conditional density for $\mathbf{B}$ from $O(p^3)$ to $O(n^2 p)$ when $p \gg n$. We provide technical details for our Gibbs sampling algorithm and our algorithm for sampling efficiently from the conditional density for $\mathbf{B}$ in Section 3 of the supplemental materials (see Appendix A).

In our experience, with good initial estimates for $\mathbf{B}$ and $\mathbf{\Sigma}$, $(\mathbf{B}^{(\text{init})}, \mathbf{\Sigma}^{(\text{init})})$, the Gibbs sampler converges quite quickly, usually within 5000 iterations. In Section 3 of the Online Supplement (see Appendix A), we describe how to initialize $(\mathbf{B}^{(\text{init})}, \mathbf{\Sigma}^{(\text{init})})$. In the Online Supplement, we also provide history plots of the draws from the Gibbs sampler for individual coefficients of $\mathbf{B}$ from experiment 5 ($n = 100$, $p = 500$, $q = 3$) and experiment 6 ($n = 150$, $p = 1000$, $q = 4$) of our simulation studies in Section 5.1, which illustrate rapid convergence.

Although our algorithm is efficient, Gibbs sampling can still be prohibitive if p is extremely large (say, on the order of millions). In this case, we recommend first screening the p covariates based on the magnitude of their marginal correlations with the responses $(y_1, \ldots, y_q)$ and then implementing the MBSP model on the reduced subset of covariates. This marginal screening technique for dimension reduction has long been advocated for ultra-high-dimensional problems, even for non-Bayesian approaches (e.g., [17, 18]). Faster alternatives to MCMC to handle extremely large p are also worth exploring in the future.

4.2.2. Specification of hyperparameters τ, d, and k

Just as in (5), the τ in (16) continues to act as a global shrinkage parameter. A natural question is how to specify an appropriate value for τ. Armagan et al. [2] recommend setting τ to the expected level of sparsity. Given our theoretical results in Theorems 3 and 4, we set $\tau \equiv \tau_n = 1/(p\sqrt{n \ln n})$. This choice of τ satisfies the sufficient conditions for posterior consistency in both the low-dimensional and the high-dimensional settings when $\mathbf{\Sigma}$ is fixed and known.

In order to specify the hyperparameters d and k in the $\mathcal{IW}(d, k\mathbf{I}_q)$ prior for $\mathbf{\Sigma}$, we appeal to the arguments made by Brown et al. [9]. As noted by Brown et al. [9], if we set $d = 3$, then $\mathbf{\Sigma}$ has a finite first moment, with $\mathbb{E}(\mathbf{\Sigma}) = k/(d-2)\mathbf{I}_q = k\mathbf{I}_q$. Additionally, as argued in Bhadra and Mallick [5] and Brown et al. [9], k should *a priori* be comparable in size with the likely variances of $\mathbf{Y}$ given $\mathbf{X}$. Accordingly, we take our initial estimate of $\mathbf{B}$ from the Gibbs sampler, $\mathbf{B}^{(\text{init})}$ (specified in Section 4.2.1), and take k as the variance of the residuals, $\mathbf{Y} - \mathbf{X}\mathbf{B}^{(\text{init})}$.

4.3. Variable selection

Although the MBSP model (5) and the MBSP-TPBN model (16) produce robust estimates for $\mathbf{B}$, they do not produce exact zeros. In order to use model (16) for variable selection, we recommend looking at the 95% credible intervals for each entry b_{ij} in row i and column j. If the credible intervals for every single entry in row $i \in \{1, \ldots, p\}$, contain zero, then we classify predictor i as an irrelevant predictor. If at least one credible interval in row $i \in \{1, \ldots, p\}$ does not contain zero, then we classify i as an active predictor. The empirical performance of this variable selection method seems to work well, as shown in Section 5.

5. Simulations and data analysis

5.1. Simulation studies

For our simulation studies, we implement model (16) using our R package `MBSP`. We specify $u = 0.5$, $a = 0.5$ so that the polynomial-tailed prior that we utilize is the horseshoe prior. The horseshoe is known to perform well in simulations [12, 41]. We set $\tau = 1/(p\sqrt{n \ln n})$, $d = 3$, and k comparable to the size of likely variance of $\mathbf{Y}$ given $\mathbf{X}$.

In all of our simulations, we generate data from the multivariate linear regression model (1) as follows. The rows of the design matrix $\mathbf{X}$ are independently generated from $N_p(\mathbf{0}, \mathbf{\Gamma})$, where $\mathbf{\Gamma} = (\Gamma_{ij})_{p \times p}$ with $\Gamma_{ij} = 0.5^{|i-j|}$. The sparse $p \times q$ matrix $\mathbf{B}$ is generated by first randomly selecting an active set of predictors, $\mathcal{A} \subset \{1, \ldots, p\}$. For rows with indices in the set $\mathcal{A}$, we independently draw every row element from $\mathcal{U}([-5, -0.5] \cup [0.5, 5])$. All the other rows in $\mathbf{B}$,

i.e., $\mathcal{A}^C$, are then set equal to zero. Finally, the rows of the noise matrix $\mathbf{E}$ are independently generated from $\mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (\Sigma_{ij})_{q\times q}$ with $\Sigma_{ij} = \sigma^2(0.5)^{|i-j|}$, $\sigma^2 = 2$. We consider six different simulation settings with varying levels of sparsity.

a) Experiment 1 ($p < n$): $n = 60$,$p = 30$, $q = 3$, 5 active predictors (sparse model).

b) Experiment 2 ($p < n$): $n = 80$, $p = 60$, $q = 6$, 40 active predictors (dense model).

c) Experiment 3 ($p > n$): $n = 50$, $p = 200$, $q = 5$, 20 active predictors (sparse model).

d) Experiment 4 ($p > n$): $n = 60$, $p = 100$, $q = 6$, 40 active predictors (dense model).

e) Experiment 5 ($p \gg n$): $n = 100$, $p = 500$, $q = 3$, 10 active predictors (ultra-sparse model).

f) Experiment 6 ($p \gg n$): $n = 150$, $p = 1000$, $q = 4$, 50 active predictors (sparse model).

The Gibbs sampler described in Section 4.2.1 is efficient in handling the two $p \gg n$ setups in experiments 5 and 6. Running on an Intel Xeon E5-2698 v3 processor, the Gibbs sampler runs about 761 iterations per minute for Experiment 5 and about 134 iterations per minute for Experiment 6. In all our experiments, we run Gibbs sampling for 15,000 iterations, discarding the first 5000 iterations as burn-in.

As our point estimate for $\mathbf{B}$, we take the posterior median $\widehat{\mathbf{B}} = (\widehat{b_{ij}})_{p\times q}$. To perform variable selection, we inspect the 95% individual credible interval for every entry and classify predictors as irrelevant if all of the q intervals in that row contain 0, as described in Section 4.3. We compute mean squared errors (MSEs) rescaled by a factor of 100, as well as the false discovery rate (FDR), false negative rate (FNR), and overall misclassification probability (MP) as follows:

$$\text{MSE}_{\text{est}} = 100 \times \|\widehat{\mathbf{B}} - \mathbf{B}\|_F^2/(pq), \quad \text{MSE}_{\text{pred}} = 100 \times \|\mathbf{X}\widehat{\mathbf{B}} - \mathbf{X}\mathbf{B}\|_F^2/(nq),$$

$$\text{FDR} = \text{FP} / (\text{TP} + \text{FP}), \quad \text{FNR} = \text{FN} / (\text{TN} + \text{FN}), \quad \text{MP} = (\text{FP} + \text{FN})/(pq),$$

where FP, TP, FN, and TN denote the number of false positives, true positives, false negatives, and true negatives, respectively.

We compare the performance of the MBSP-TPBN estimator with that of four other row-sparse estimators of $\mathbf{B}$. An alternative Bayesian approach based on the spike-and-slab formulation is studied. Namely, we consider the multivariate Bayesian group lasso posterior median estimator with a spike-and-slab prior (MBGL-SS), introduced by Liquet et al. [28], which applies a spike-and-slab prior with a point mass $\mathbf{0}^{m_g q}$ for the gth group of covariates, which corresponds to m_g rows of $\mathbf{B}$. When the grouping structure of the covariates is not available, we can still utilize the MBGL-SS method by applying the spike-and-slab prior to each individual row of $\mathbf{B}$. In our study, we consider each predictor as its own "group" (i.e., $m_1 = \cdots = m_p = 1$) so that individual rows are shrunk to $\mathbf{0}_q^\top$. This method can be implemented in R using the `MBSGS` package.

In addition, we compare the performance of MBSP-TPBN to three different frequentist point estimators obtained through regularization penalties on the rows of $\mathbf{B}$. In the R package `glmnet` [19], there is an option to fit the following model to multivariate data, which we call the multivariate lasso (MLASSO) method:

$$\widehat{\mathbf{B}}^{\text{MLASSO}} = \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times q}} \left(\|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda \sum_{j=1}^{p} \|\mathbf{b}_j\|_2 \right).$$

The MLASSO model applies an ℓ_1 penalty to each of the rows of $\mathbf{B}$ to shrink entire row estimates to be $\mathbf{0}_q^\top$. We also compare the MBSP-TPBN estimator to the row-sparse reduced-rank regression (SRRR) estimator, introduced by Chen and Huang [14], which uses an adaptive group lasso penalty on the rows of $\mathbf{B}$, but which further constrains the solution to be rank-deficient. Finally, we compare our method to the sparse partial least squares estimator (SPLS), introduced by Chun and Keleş [15]. SPLS combines partial least squares (PLS) regression with a regularization penalty on the rows of $\mathbf{B}$ in order to obtain a row-sparse PLS estimate of $\mathbf{B}$. The SRRR and SPLS methods are available in the R packages `rrpack` and `spls`.

Table 2 shows the results averaged across 100 replications for the MBSP-TPBN model (16), compared with MBGL-SS, LSGL, and SRRR. As the results illustrate, the Bayesian methods tend to outperform the frequentist ones

in the low-dimensional case where $p < n$. In the two low-dimensional experiments (experiments 1 and 2), the MBGL-SS estimator performs the best across all of our performance metrics, with the MBSP-TPBN model following closely behind.

However, in all the high-dimensional ($p > n$) settings, MBSP-TPBN significantly outperforms all of its competitors. Table 2 shows that the MBSP-TPBN model has a lower MSE_{est} than the other four methods in experiments 3 through 6. In experiments 5 and 6 (the $p \gg n$ scenarios), the MSE_{est} and MSE_{pred} are both much lower for the MBSP-TPBN model than for the other methods.

Additionally, using the 95% credible interval technique in Section 4.3 to perform variable selection, the FDR and the overall MP are also consistently low for the MBSP-TPBN model. Even when the true underlying model is not sparse, as in experiments 2 and 4, MBSP performs very well and correctly identifies most of the signals. In both the ultra-high-dimensional settings we considered in experiments 5 and 6, the other four methods all seem to report high FDR, while the MBSP's FDR remains very small.

In short, our experimental results show that the MBSP model (2) has excellent finite sample performance for both estimation and selection, is robust to non-sparse situations, and scales very well to large p compared to the other methods. In addition to its strong empirical performance, the MBSP model (as well as the MBGL-SS model) provides a vehicle for uncertainty quantification through the posterior credible intervals.

5.2. Yeast cell cycle data analysis

We illustrate the MBSP methodology on a yeast cell cycle data set. This data set was first analyzed by Chun and Keleş [15] and is available in the `spls` package in R. Transcription factors (TFs) are sequence-specific DNA binding proteins which regulate the transcription of genes from DNA to mRNA by binding specific DNA sequences. In order to understand their role as a regulatory mechanism, one often wishes to study the relationship between TFs and their target genes at different time points. In this yeast cell cycle data set, mRNA levels are measured at 18 time points seven minutes apart (every 7 minutes for a duration of 119 minutes). The 542×18 response matrix $\mathbf{Y}$ consists of 542 cell-cycle-regulated genes from an α factor arrested method, with columns corresponding to the mRNA levels at the 18 distinct time points. The 542×106 design matrix $\mathbf{X}$ consists of the binding information of a total of 106 TFs.

In practice, many of the TFs are not actually related to the genes, so our aim is to recover a parsimonious model with only a tiny number of the truly statistically significant TFs. To perform variable selection, we fit the MBSP-TPBN model (16) and then use the 95% credible interval method described in Section 4.3. Beyond identifying significant TFs, we assess the predictive performance of the MBSP-TPBN model (16) by performing five-fold cross validation, using 80% of the data as our training set to obtain an estimate of $\mathbf{B}$, $\widehat{\mathbf{B}}^{\text{train}}$. We take the posterior median as $\widehat{\mathbf{B}}^{\text{train}} = (\widehat{b}_{ij})^{\text{train}}$ and use it to compute the mean squared error of the residuals on the remaining 20% of the left-out data. We repeat this five times, using different training and test sets each time, and take the average MSE as our mean squared predictor error (MSPE). To make our analysis more clear, we scale the MSPE by a factor of 100.

Table 3 shows our results compared with the MBGL-SS, MLASSO, SRRR, and SPLS methods. MBSP-TPBN selects 12 of the 106 TFs as significant, so we do recover a parsimonious model. All five methods selected the TFs, ACE2, SWI5, and SWI6. The two Bayesian methods seem to recover a much more sparse model than the frequentist methods. In particular, the MLASSO method has lowest MSPE, but it selects 78 of the 106 TFs as significant, suggesting that there may be overfitting in spite of the regularization penalty on the rows of $\mathbf{B}$. Our results suggest that the frequentist methods may have good predictive performance on this particular data set, but at the expense of parsimony. In practice, sparse models are preferred for the sake of interpretability, and our numerical results illustrate that the MBSP model recovers a sparse model with competitive predictive performance.

Finally, Figure 1 illustrates the posterior median estimates and the 95% credible bands for four of the 10 TFs that were selected as significant by the MBSP-TPBN model. These plots illustrate that the standard errors under the MBSP-TPBN model are not too large. One of the potential drawbacks of using credible intervals for selection is that these intervals may be too conservative, but we see that it is not the case here. This plot, combined with our earlier simulation results and our data analysis results, provide empirical evidence for using the MBSP model for estimation and variable selection. However, further theoretical investigation is warranted in order to justify the use of marginal credible intervals for variable selection. In particular, van der Pas et al. [43] showed that marginal credible intervals may provide overconfident uncertainty statements for certain large signal values when applied to estimating normal mean vectors, and the same issue could be present here.

Table 2: Simulation results for MBSP-TPBN, compared with MBGL-SS, MLASSO, SRRR, and SPLS, averaged across 100 replications.

Experiment 1: $n = 60$, $p = 30$, $q = 3$. 5 active predictors (sparse model)

Method	**MSE$_{\text{est}}$**	**MSE$_{\text{pred}}$**	**FDR**	**FNR**	**MP**
MBSP	1.146	24.842	0.015	0	0.003
MBGL-SS	**0.718**	**17.074**	**0.005**	0	**0.001**
MLASSO	2.181	41.424	0.6412	0	0.335
SRRR	1.646	29.256	0.3270	0	0.128
SPLS	2.428	43.879	0.1093	0.0019	0.028

Experiment 2: $n = 80$, $p = 60$, $q = 6$, 40 active predictors (dense model)

Method	**MSE$_{\text{est}}$**	**MSE$_{\text{pred}}$**	**FDR**	**FNR**	**MP**
MBSP	5.617	104.88	0.0034	0	0.0023
MBGL-SS	**5.202**	**101.40**	**0.0007**	0	**0.0005**
MLASSO	10.478	130.90	0.3307	0	0.330
SRRR	5.695	104.67	0.0491	0	0.038
SPLS	244.136	3633.77	0.2071	0	0.223

Experiment 3: $n = 50$, $p = 200$, $q = 5$, 20 active predictors (sparse model)

Method	**MSE$_{\text{est}}$**	**MSE$_{\text{pred}}$**	**FDR**	**FNR**	**MP**
MBSP	**1.357**	**117.52**	**0.0117**	0	**0.0013**
MBGL-SS	57.25	694.81	0.858	0.02	0.619
MLASSO	8.400	169.026	0.7758	0	0.349
SRRR	17.46	161.70	0.698	0	0.307
SPLS	48.551	2006.03	0.422	0.033	0.103

Experiment 4: $n = 60$, $p = 100$, $q = 6$, 40 active predictors (dense model)

Method	**MSE$_{\text{est}}$**	**MSE$_{\text{pred}}$**	**FDR**	**FNR**	**MP**
MBSP	**11.030**	**172.89**	**0.0266**	0	**0.0114**
MBGL-SS	204.33	318.80	0.505	0.1265	0.415
LSGL	44.635	188.81	0.544	0	0.479
SRRR	242.67	193.64	0.594	0	0.587
SPLS	213.19	3909.07	0.135	0.0005	0.005

Experiment 5: $n = 100$, $p = 500$, $q = 3$, 10 active predictors (ultra-sparse model)

Method	**MSE$_{\text{est}}$**	**MSE$_{\text{pred}}$**	**FDR**	**FNR**	**MP**
MBSP	**0.0374**	**12.888**	**0.064**	0	**0.0015**
MBGL-SS	1.327	155.51	0.483	0.0005	0.092
MLASSO	0.2357	75.961	0.837	0	0.115
SRRR	0.9841	49.428	0.688	0	0.104
SPLS	0.3886	138.62	0.1355	0.0005	0.005

Experiment 6: $n = 150$, $p = 1000$, $q = 4$, 50 active predictors (sparse model)

Method	**MSE$_{\text{est}}$**	**MSE$_{\text{pred}}$**	**FDR**	**FNR**	**MP**
MBSP	**0.0155**	**8.934**	**0.0025**	0.00003	**0.00016**
MBGL-SS	1.327	155.51	0.483	0.0005	0.092
MLASSO	1.982	181.95	0.810	0	0.214
SRRR	0.9841	49.428	0.688	0	0.104
SPLS	25.560	8631.92	0.420	0.021	0.051

Table 3: Results for analysis of the yeast cell cycle data set. The MSPE has been scaled by a factor of 100. In particular, all fives models selected the three TFs, ACE2, SWI5, and SWI6 as significant.

Method	Number of Proteins Selected	MSPE
MBSP	12	18.673
MBGL-SS	7	20.093
MLASSO	78	17.912
SRRR	44	18.204
SPLS	44	18.904

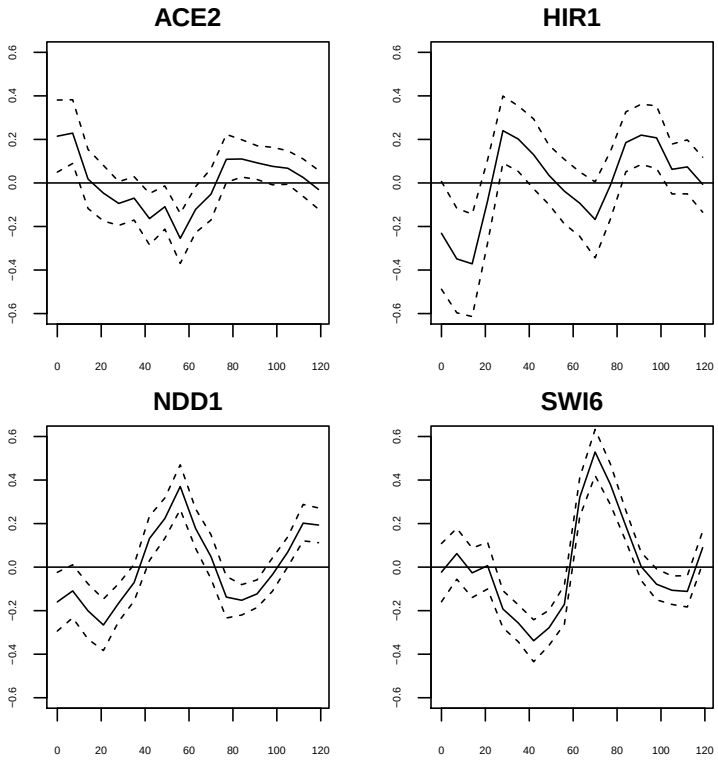

Figure 1: Plots of the estimates and 95% credible bands for four of the 10 TFs that were deemed as significant by the MBSP-TPBN model. The x-axis indicates time (minutes) and the y-axis indicates the estimated coefficients.

6. Conclusion and future work

In this paper, we have introduced a method for sparse multivariate Bayesian estimation with shrinkage priors (MBSP). Previously, global-local shrinkage priors have mainly been used in univariate regression or in the estimation of normal mean vectors. Our paper extends their use to the multivariate linear regression framework.

Our paper makes several important contributions to methodology and theory. First, our model may be used for sparse multivariate estimation for p, n, and q of any size. To motivate the MBSP model, we have shown that the posterior distribution can consistently estimate $\mathbf{B}$ in (1) in both the low-dimensional and ultra-high-dimensional settings where p is allowed to grow nearly exponentially with n (with the response dimension q fixed). This appears to be the first paper to provide sufficient conditions for ultra-high-dimensional posterior consistency under model (1) in the statistical literature. Moreover, our method is general enough to encompass a large family of heavy-tailed priors, including the Student's t prior, the horseshoe prior, the generalized double Pareto prior, and others.

The MBSP model (5) can be implemented using straightforward Gibbs sampling. We implemented a fully Bayesian version of it with an appropriate prior on $\mathbf{\Sigma}$ and with polynomial-tailed priors belonging to the TPBN

family, using the horseshoe prior as a special case. By examining the 95% posterior credible intervals for every element in each row of the posterior conditional distribution of $\mathbf{B}$, we also showed how one could use the MBSP model for variable selection. Through simulations and data analysis on a real data set, we have illustrated that our model has excellent performance in finite samples for both estimation and variable selection.

6.1. Future work

Although our paper addresses a long-standing gap between theory and application for Bayesian multivariate linear regression, much still remains unknown. In this paper, we demonstrated that the MBSP model (5) could achieve posterior consistency in both low-dimensional ($p = o(n)$) and ultra-high-dimensional ($\ln p = o(n)$) settings. The next step is to quantify the posterior contraction rate. In the present context of multivariate linear regression, we say that the posterior distribution contracts at the rate r_n if

$$\Pi_n(\|\mathbf{B}_n - \mathbf{B}_0\|_F > M_n r_n | \mathbf{Y}_n) \to 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \to \infty,$$

for every $M_n \to \infty$ as $n \to \infty$. In the context of high-dimensional *univariate* regression, several authors (e.g., [13], [35]) have attained optimal posterior contraction rates of $O\{\sqrt{s \ln(p)/n}\}$ with respect to the ℓ_1 and ℓ_2 norms (where s denotes the number of active predictors). It is worth noting that $\sqrt{s \ln(p)/n}$ is the familiar minimax rate of convergence under squared error loss for a number of frequentist point estimators, including the Dantzig selector [11], the scaled lasso [38], and the lasso [40]. We conjecture that under suitable regularity conditions and compatibility conditions on the design matrix, the MBSP model can attain a similarly optimal posterior rate of contraction.

Additionally, we could investigate if posterior consistency and optimal posterior contraction rates can be achieved if we allow the number of response variables q to diverge to infinity in the MBSP model. From an implementation standpoint, q can be of any size, but for our theoretical investigation of the MBSP model, we assumed q to be fixed. If q is allowed to grow as sample size grows, then some sort of sparsity assumption for the response variables may need to be imposed. We surmise that novel techniques would also be needed to prove posterior consistency in this scenario, since the distributional theory we used to prove our consistency results may not apply if q is no longer fixed.

Extension of our posterior consistency results to the case where $\mathbf{\Sigma}$ is unknown and endowed with a prior also remains an open problem. In this case, we need to integrate out $\mathbf{\Sigma}$ in order to work with the marginal density of the prior on $\mathbf{B}$. If we assume the standard inverse-Wishart prior on $\mathbf{\Sigma}$, this gives rise to a matrix-variate t distribution. Handling this density is very nontrivial and would require significantly different techniques than the ones we used to establish posterior consistency in Section 3.4. Nevertheless, this warrants future investigation.

For variable selection with the MBSP model, we relied on the post hoc method of examining the 95% credible intervals for each entry of the estimated coefficients matrix for $\mathbf{B}$. Further theoretical justification for this selection method is needed. Other possible thresholding rules should also be investigated. Because scale-mixture shrinkage priors place zero probability at exactly zero, we must necessarily use thresholding to perform variable selection. How to optimally choose this threshold (or thresholds) in high-dimensional settings remains an active area of research.

All the aforementioned are very important open problems in Bayesian multivariate linear regression, and we hope that the methodology and theory introduced here can serve as the foundation for further developments in this area.

Acknowledgments

The authors are grateful to the Editor-in-Chief, the Associate Editor and two referees for their helpful comments on an earlier version of this paper.

Appendix A. Supplementary Material

Supplementary material related to this article can be found online.

References

[1] A. Armagan, D.B. Dunson, J. Lee, W.U. Bajwa, N. Strawn, Posterior consistency in linear models under shrinkage priors, Biometrika 100 (2013) 1011–1018.

[2] A. Armagan, M. Clyde, D.B. Dunson, Generalized beta mixtures of Gaussians, In: J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, K.Q. Weinberger, Eds., Advances in Neural Information Processing Systems 24, pp. 523–531, 2011.

[3] A. Armagan, D.B. Dunson, J. Lee, Generalized double Pareto shrinkage, Statist. Sinica 23 (2013) 119–143.

[4] J.O. Berger, A robust generalized Bayes estimator and confidence region for a multivariate normal mean, Ann. Statist. 8 (1980) 716–761.

[5] A. Bhadra, B.K. Mallick, Joint high-dimensional Bayesian variable and covariance selection with an application to EQTL analysis, Biometrics 69 (2013) 447–457.

[6] A. Bhadra, J. Datta, N.G. Polson, B. Willard, The horseshoe+ estimator of ultra-sparse signals, Bayesian Anal. 12 (2017) 1105–1131.

[7] A. Bhattacharya, A. Chakraborty, B.K. Mallick, Fast sampling with Gaussian scale mixture priors in high-dimensional regression, Biometrika 103 (2016) 985–991.

[8] N.H. Bingham, C.M. Goldie, J.L. Teugels, Regular variation, In: Encyclopedia of Mathematics and its Applications, Cambridge University Press, 1987.

[9] P.J. Brown, M. Vannucci, T. Fearn, Multivariate Bayesian variable selection and prediction, J. R. Stat. Soc. Ser. B 60 (1998) 627–641.

[10] F. Bunea, Y. She, M.H. Wegkamp, Joint variable and rank selection for parsimonious estimation of high-dimensional matrices, Ann. Statist. 40 (2012) 2359–2388.

[11] E. Candès, T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n, Ann. Statist. 35 (2007) 2313–2351.

[12] C.M. Carvalho, N.G. Polson, J.G. Scott, The horseshoe estimator for sparse signals, Biometrika 97 (2010) 465–480.

[13] I. Castillo, J. Schmidt-Hieber, A.W. van der Vaart, Bayesian linear regression with sparse priors, Ann. Statist. 43 (2015) 1986–2018.

[14] L. Chen, J.Z. Huang, Sparse reduced-rank regression for simultaneous dimension reduction and variable selection, J. Amer. Statist. Assoc. 107 (2012) 1533–1545.

[15] H. Chun, S. Keleş, Sparse partial least squares regression for simultaneous dimension reduction and variable selection, J. R. Stat. Soc. Ser. B 72 (2010) 3–25.

[16] J. Datta, J.K. Ghosh, Asymptotic properties of BayesBayes risk for the horseshoe prior, Bayesian Anal. 8 (2013) 111–132.

[17] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, J. R. Stat. Soc. Ser. B 70 (2008) 849–911.

[18] J. Fan, R. Song, Sure independence screening in generalized linear models with np-dimensionality, Ann. Statist. 38 (2010) 3567–3604.

[19] J. Friedman, T. Hastie, R.J. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2010) 1–22.

[20] E.I. George, R.E. McCulloch, Variable selection via gibbs sampling, J. Amer. Statist. Assoc. 88 (1993) 881–889.

[21] P. Ghosh, A. Chakrabarti, Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems, Bayesian Anal. 12 (2017) 1133–1161.

[22] P. Ghosh, X. Tang, M. Ghosh, A. Chakrabarti, Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity, Bayesian Anal. 11 (2016) 753–796.

[23] G. Goh, D.K. Dey, K. Chen, Bayesian sparse reduced rank multivariate regression, J. Multivariate Anal. 157 (2017) 14 – 28.

[24] J.E. Griffin, P.J. Brown, Some priors for sparse regression modelling, Bayesian Anal. 8 (2013) 691–702.

[25] H. Ishwaran, J.S. Rao, Spike and slab variable selection: Frequentist and Bayesian strategies, Ann. Statist. 33 (2005) 730–773.

[26] Y. Li, B. Nan, J. Zhu, Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure, Biometrics 71 (2015) 354–363.

[27] D.L. Libby, M.R. Novick, Multivariate generalized beta distributions with applications to utility assessment, J. Educ. Stat. 7 (1982) 271–294.

[28] B. Liquet, K. Mengersen, A.N. Pettitt, M. Sutton, Bayesian variable selection regression of multivariate responses for group data, Bayesian Anal. 12 (2017) 1039–1067.

[29] B. Liquet, L. Bottolo, G. Campanella, S. Richardson, M. Chadeau-Hyam, `R2GUESS`: A graphics processing unit-based R package for Bayesian variable selection regression of multivariate responses. J. Stat. Softw. 69 (2016) 1–32.

[30] T.J. Mitchell, J.J. Beauchamp, Bayesian variable selection in linear regression, J. Amer. Statist. Assoc. 83 (1988) 1023–1032.

[31] N.N. Narisetty, X. He, Bayesian variable selection with shrinking and diffusing priors, Ann. Statist. 42 (2014) 789–817.

[32] T. Park, G. Casella, The Bayesian lasso, J. Amer. Statist. Assoc. 103 (2008) 681–686.

[33] N.G. Polson, J.G. Scott, On the half-Cauchy prior for a global scale parameter, Bayesian Anal. 7 (2012) 887–902.

[34] A.J. Rothman, E. Levina, J. Zhu, Sparse multivariate regression with covariance estimation, J. Comput. Graph. Stat. 19 (2010) 947–962.

[35] V. Ročková, E.I. George, The spike-and-slab lasso, J. Amer. Statist. Assoc. 0 (2018) 0–0.

[36] J.-B. Salomond, Risk quantification for the thresholding rule for multiple testing using Gaussian scale mixtures, ArXiv e-prints (2017).

[37] W.E. Strawderman, Proper Bayes minimax estimators of the multivariate normal mean, Ann. Math. Statist. 42 (1971) 385–388.

[38] T. Sun, C.-H. Zhang, Scaled sparse linear regression, Biometrika 99 (2012) 879–898.

[39] X. Tang, X. Xu, M. Ghosh, P. Ghosh, Bayesian variable selection and estimation based on global-local shrinkage priors, Sankhyā A (2017).

[40] R.J. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.

[41] S.L. van der Pas, B.J.K. Kleijn, A.W. van der Vaart, The horseshoe estimator: Posterior concentration around nearly black vectors, Electron. J. Statist. 8 (2014) 2585–2618.

[42] S.L. van der Pas, J.-B. Salomond, J. Schmidt-Hieber, Conditions for posterior contraction in the sparse normal means problem, Electron. J. Statist. 10 (2016) 976–1000.

[43] S.L. van der Pas, B. Szabó, A.W. van der Vaart, Uncertainty quantification for the horseshoe (with discussion), Bayesian Anal. 12 (2017) 1221–1274.

[44] S.L. van der Pas, B. Szabó, A.W. van der Vaart, Adaptive posterior contraction rates for the horseshoe, Electron. J. Statist. 11 (2017) 3196–3225.

[45] I. Wilms, C. Croux, An algorithm for the multivariate group lasso with covariance estimation, J. Appl. Stat. 45 (2018) 668–681.

[46] X. Xu, M. Ghosh, Bayesian variable selection and estimation for group lasso, Bayesian Anal. 10 (2015) 909–936.
[47] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc. Ser. B 68 (2006) 49–67.
[48] A. Zellner, On assessing prior distributions and Bayesian regression analysis with *g* prior distributions, In: P.K. Goel, A. Zellner, Eds., Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, Studies in Bayesian Econometrics, pp. 233–243, 1986.
[49] H. Zhu, Z. Khondker, Z. Lu, J.G. Ibrahim, Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers, J. Amer. Statist. Assoc. 109 (2014) 997–990.
[50] H. Zou, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418–1429.
[51] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B 67 (2005) 301–320.