An approach of Bayesian variable selection for ultrahigh-dimensional multivariate regression

Xiaotian Dai | Guifang Fu | Randall Reese | Shaofei Zhao | Zuofeng Shang

1Department of Mathematics and Statistics, SUNY Binghamton University, Binghamton, NY, USA
2Idaho National Laboratory, Idaho Falls, ID, USA
3Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ, USA

Correspondence
Guifang Fu, Department of Mathematics and Statistics SUNY Binghamton University, Binghamton, NY 13902, USA.
Email: gfu@binghamton.edu

Funding information
National Science Foundation

In many practices, scientists are particularly interested in detecting which of the predictors are truly associated with a multivariate response. It is more accurate to model multiple responses as one vector rather than separating each component one by one. This is particularly true for complex traits having multiple correlated components. A Bayesian multivariate variable selection (BMVS) approach is proposed to select important predictors influencing the multivariate response from a candidate pool with ultrahigh dimension. By applying the sample-size-dependent spike and slab priors, the BMVS approach satisfies the strong selection consistency property under certain conditions, which represents the advantages of BMVS over other existing Bayesian multivariate regression-based approaches. The proposed approach considers the covariance structure of multiple responses without assuming independence and integrates the estimation of covariance-related parameters together with all regression parameters into one framework through a fast-updating Markov chain Monte Carlo (MCMC) procedure. It is demonstrated through simulations that the BMVS approach outperforms some other relevant frequentist and Bayesian approaches. The proposed BMVS approach possesses a flexibility of wide applications, including genome-wide association studies with multiple correlated phenotypes and a large scale of genetic variants and/or environmental variables, as demonstrated in the real data analyses section. The computer code and test data of the proposed method are available as an R package.

KEYWORDS
Bayesian analysis, gene selection, high dimension modelling, multivariate genome-wide association studies

1 INTRODUCTION

Modelling the association between a multivariate response and a set of predictors has been attracting considerable attention in various disciplines, including ecology, geology, psychology, genetics, among others. For example, De’Ath (2002) modelled the relationship between multispecies and environmental factors. Tsitsika et al. (2009) selected important predictors associated with multiple internet addiction characteristics for adolescents. Influential single-nucleotide polymorphisms (SNPs) affecting multiple biological traits were detected through genome-wide association studies (GWAS) (Bottolo et al., 2013; Liquet et al., 2016; O’Reilly et al., 2012; Yang et al., 2010; Zhou & Stephens, 2014). When the number of predictors is much larger than the number of observations \( p > n \) (O’Reilly et al., 2012; Scott et al., 2012), overfitting and low prediction accuracy may become problematic. Therefore, variable selection is crucial for high-dimensional data analyses.

As a well-known frequentist method, penalized regression has been widely used for variable selection purposes in high-dimensional data analysis (Li et al., 2011; Rothman et al., 2010). These include LASSO (Tibshirani, 1996), adaptive LASSO (Zou, 2006) and elastic net (Zou &...
Hastie, 2005), to name a few. Despite LASSO's popularity in shrinking unimportant predictors to zero, its theoretical properties are controversial. Some research claims that the selection consistency of LASSO cannot be established and its performance for high-dimensional predictors is not well explored (Fan & Li, 2001; Narisetty & He, 2014; Zou, 2006). Breiman and Friedman (1997) proposed the Curds and Whey algorithm to enhance the prediction accuracy of multivariate regression by taking advantage of the correlations between response variables. However, their algorithm fitted a full model using $\ell_2$ penalty, which cannot shrink unimportant predictors to produce a sparse model (Fan & Li, 2001). She and Chen (2017) proposed using reduced-rank regression to handle high-dimensional multivariate data analysis and set a penalty constraint not only on the regression coefficients but also on the rank of the coefficient matrix. The disadvantages of the reduced-rank approach include its sensitivity on the rank constraint and outliers of the data for low-rank coefficient matrix estimation (She & Chen, 2017). In addition, the theoretical properties of reduced-rank regression investigated by She and Chen (2017) focused on non-asymptotic robust analysis and prediction accuracy improvement, instead of the asymptotic selection consistency.

Bayesian variable selection approaches have been well-established in univariate regression settings where there is only one response variable. Consider a univariate regression model here: $Y_{i,1} = X_{i,1} \beta_{p,1} + \epsilon_{i,1}$. Bayesian methods usually introduce a latent variable $Z_i$ for each predictor, in which $Z_i = 0$ or 1, $i = 1, ..., p$, indicates whether or not the $i$ th predictor should be included into the final model. Hence a binary vector $Z$ of length $p$ can indicate a particular model. There have been many efforts devoted to develop “objective” priors for selecting the optimal model, such as the Jeffreys’ prior, Zellner’s g-prior and their related model selection literature (Bayarri & García-Donato, 2008; Berger & Pericchi, 1996; Bové et al., 2015; Cano & Salmerón, 2013; Cui & George, 2008; De Santis & Spezzaferri, 1999; Fouskakis & Ntzoufras, 2016; Johnson & Rossell, 2012; Kass & Wasserman, 1995; Laud & Ibrahim, 1995; Liang et al., 2008; Maruyama & George, 2010; Maruyama & Strawderman, 2010; Moreno et al., 1998; Pérez & Berger, 2002; Zellner & Siow, 1980, 1984). The word “objective” indicates that these priors were not “subjective” and applicable to various model selection problems (Bayarri et al., 2012).

As the advent of new data collection technologies, Bayesian regularized regression, such as the Bayesian lasso (Li et al., 2011; Park & Casella, 2008) and the spike and slab priors, was investigated to cope with the challenges of high dimensionality. The connections between regularized regressions and the spike and slab priors have been studied in Narisetty and He (2014). Mitchell and Beauchamp (1988) proposed the use of spike and slab priors on regression coefficients $\beta_s$. When $Z_i = 0$, the $i$ th predictor is considered inactive and $\beta_i$ will have a prior distribution with concentrated probability mass around zero, which is referred to as the spike prior. When $Z_i = 1$, the $i$ th predictor is considered active in the model, and $\beta_i$ will have a prior distribution with a diffusing probability density, which is referred to as the slab prior. George and McCulloch (1993) proposed a stochastic search variable selection (SSVS) approach to select “promising” subsets of predictors. Their framework used a normal distribution with zero mean and a small fixed variance as the spike prior and another normal distribution with zero mean and a large fixed variance as the slab prior. Most recently, Narisetty and He (2014) questioned the theoretical selection consistency property for the approaches using fixed hyperparameters for spike and slab priors and showed that SSVS did not guarantee selection consistency. As the first work changing the fixed variances of the spike and slab priors into sample-size-dependent shrinking and diffusing priors, Narisetty and He (2014) proved that the criteria for strong selection consistency were satisfied in univariate regression (Narisetty & He, 2014).

In terms of how many variables should be selected, traditional Bayesian literature employed a latent vector $Z_{p,1}$ to indicate the inclusions and exclusions of all the $p$ predictors, and then the optimal model was determined by finding the latent vector yielding the highest model posterior probability (Berger & Pericchi, 1996; Cui & George, 2008; Fouskakis & Ntzoufras, 2016; Laud & Ibrahim, 1995; Liang et al., 2008; Maruyama & George, 2010; Maruyama & Strawderman, 2010; Moreno et al., 1998; Pérez & Berger, 2002; Shang & Clayton, 2011; Zellner & Siow, 1980). Barbieri and Berger (2004) suggested that an optimal model in the Bayesian variable selection was often the “median probability model,” which was defined to be the model consisting of the predictors whose overall posterior inclusion probability is at least 1/2. The overall posterior inclusion probability is the sum of the posterior probabilities of the models which contain that predictor. In a recent study of univariate Bayesian variable selection, Narisetty and He (2014) proposed using the marginal posterior probability of $Z_i = 1$ to select the optimal model and showed that the selection consistency can be guaranteed using sample-size-dependent shrinking and diffusing priors. This is a major improvement because the use of marginal posterior probability is definitely computationally advantageous compared with computing the model posterior probabilities among the entire model space.

Unlike the aforementioned univariate variable selection approaches, variable selection approaches for multivariate regressions are undeveloped. Brown et al. (1998) developed a Bayesian method for multivariate variable selection, but their approach inherited aforementioned limitation of using model posterior probabilities. Therefore, it was not computationally feasible when $p$ is greater than 20 because it involves $2^p$ choices (Brown et al., 1998). A more recent development on Bayesian multivariate regression-based variable selection approach was the GUESS algorithm proposed by Bottolo et al. (2013) and the R package implementing the GUESS algorithm (Liquet et al., 2016). Compared with the aforementioned Bayesian univariate regression-based approaches, GUESS was able to model multiple correlated responses and their covariance structure. However, similar to SSVS, the GUESS algorithm used the fixed hyperparameters for spike and slab priors and hence did not have theoretical guarantees about selection consistency. Moreover, GUESS also employed model posterior probabilities to search the optimal model in the entire model space. To overcome the computational cost of model posterior probabilities, Bottolo et al. (2013) adopted graphical processor unit (GPU) technologies, specifically using NVIDIA’s Complete Unified Device Architecture (CUDA). Most recently, Ning et al. (2020) proposed a novel
Bayesian method to estimate regression coefficients in high-dimensional multivariate regression using a product of independent spike and slab priors. These priors accommodate sparsity at the group level, as some predictors can be naturally clustered into groups.

In this article, we propose a novel Bayesian multivariate variable selection (BMVS) approach to detect truly influential predictors that are associated with a multivariate response from an ultrahigh-dimensional candidate pool (i.e., \( p \gg n \) or \( p = e^{O(n)} \)). We consider sample-size-dependent shrinking and diffusing priors instead of fixed variances of the spike and slab priors. After modelling a complex covariance matrix without assuming independence of multiple response variables, we are still able to establish the appealing strong selection consistency for the BMVS approach under certain conditions. All unknown parameters, including the covariance-related and regression-related coefficients, are estimated through a fast-updating Markov chain Monte Carlo (MCMC) algorithm. We demonstrate that the proposed BMVS method performs well in multiple empirical simulation studies. We compare the BMVS model to a few other approaches that have been widely used in multivariate regression literature, including not only the popular frequentist methods such as canonical correlation analysis (CCA) and multivariate LASSO/elastic net but also the Bayesian multivariate regression-based GUESS algorithm (Bottolo et al., 2013). The BMVS method outperforms those relevant approaches by achieving very promising false discovery rates (FDRs) and power in all simulation designs. To further demonstrate its accuracy in variable selection, we apply the proposed BMVS approach to two real data analyses in GWAS. The R package will be published on the Comprehensive R Archive Network (CRAN).

## 2 | METHODOLOGY

### 2.1 | Prior distributions and hyperparameters

Consider the multivariate regression model:

\[
Y = X\beta + E,  
\]  

(1)

where \( Y \in \mathbb{R}^{p \times q} \) is the response matrix, \( X \in \mathbb{R}^{p \times n} \) is the predictor matrix, \( \beta \in \mathbb{R}^{p \times q} \) is the parameter coefficient matrix and \( E \in \mathbb{R}^{p \times q} \) is the error matrix with each row following an independent and identical distribution \( N_q(0, \Sigma_Y) \). Here, \( p \) denotes the number of predictors, \( q \) denotes the number of response variables and \( n \) denotes the number of observations. Let \( Y_m \) and \( X_m \) denote the \( m \) th row vector in \( Y \) and \( X \), respectively, where \( m = 1, \ldots, n \). The \( i \) th row vector of \( \beta \), \( \beta_i \), having a \( 1 \times q \) dimension, corresponds to the coefficient vector of the \( i \) th predictor, where \( i = 1, \ldots, p \). For selection purpose, we assume that the true model is sparse (i.e., active/non-active ratio is small). Both the regression coefficients \( \beta \) and the covariance matrix \( \Sigma_Y \) are unknowns that need to be estimated.

For each of the predictors, we introduce a latent binary variable \( Z_i \) to decide whether or not it should be included in the model. An active predictor is indicated by \( Z_i = 1 \) and a non-active predictor by \( Z_i = 0 \), with distribution \( P(Z_i = 1) = 1 - P(Z_i = 0) = \phi \). Intuitively, the priors of \( \phi \) should be given by the influential/non-influential ratio of the candidate predictor pool, which nevertheless is unknown in real datasets. We will empirically verify through simulations that the choice of priors for \( \phi \) has trivial impacts on the final result.

The priors of the regression coefficients \( \beta \) can be given as

\[
\begin{align*}
Y_i|X_i, \beta, \Sigma_Y &\sim N_q(X_i\beta, \Sigma_Y), \\
\beta|\ldots &\sim N_q(0, \sigma^2 \Sigma_I) \\
\text{and } \sigma^2 &\sim IG(a_1, a_2)
\end{align*}
\]  

(2)

where \( I_q \) is a \( q \) by \( q \) identity matrix, \( \sigma^2 \) is a scalar parameter following an inverse gamma distribution with a shape parameter \( a_1 \) and a scale parameter \( a_2 \), as the inverse gamma distribution is conditionally conjugate and the priors are essentially non-informative when \( a_1 \) and \( a_2 \) are reasonably small (Gelman, 2006).

The sample-size-dependent parameters \( u_0^2 \) and \( u_1^2 \) control the priors of \( \beta_i \) to make it either a slab prior or a spike prior. If \( Z_i = 0 \), \( \beta_i \) will have a spike prior distribution with concentrated probability mass around zero. If \( Z_i = 1 \), \( \beta_i \) will have a slab prior distribution with a diffusing probability density. Inspired by Narisetti and He (2014), we design the prior hyperparameters of \( u_0^2 \) and \( u_1^2 \) as

\[
\begin{align*}
u_0^2 &= \frac{\sigma^2}{10n}, \\
u_1^2 &= \sigma^2 \max \left( \frac{pq}{100n \log n} \right)
\end{align*}
\]  

(3)
where \( \hat{\sigma}_i^2 \) is the average sample variances of multiple response variables \( \bar{\sigma}_i^2 = \frac{\sum_{k=1}^{q} \hat{\sigma}_i^2}{q} \). These choices of \( u_i^2 \) and \( u_i^2 \) asymptotically satisfy the conditions that \( u_i^2 \to 0 \) as \( n \to \).

A natural choice for the prior distribution of covariance matrix \( \Sigma_Y \) would be a probability distribution defined on positive-definite matrices with the conjugate property. According to Dawid (1981), directly modelling the matrix variate has the advantage of preserving the matrix structure without breaking the matrices down into multiple row or column vectors that dramatically increase complexity and computation costs. As such, we choose the inverse Wishart distribution:

\[
\Sigma_Y \sim IW(\nu, \Lambda),
\]

where \( \nu \) is the degree of freedom and \( \Lambda \) is a positive-definite scale matrix. The default values for these prior hyperparameters are \( \nu = q + 1 \) and \( \Lambda = I_k \) (Leonard & Hsu, 1992). Due to the conjugacy of the inverse Wishart distribution, \( \Sigma_Y \) can be conveniently estimated together with all the other unknown parameters.

### 2.2 Posterior distributions

Assuming that \( Z, s \) and \( \sigma^2 \) are conditionally independent of \( \Sigma_Y \), the full conditional posterior distributions of unknown parameters are derived below. The full details of the derivation can be found in the Supporting Information.

The full conditional posterior distribution of \( \beta_i \), the coefficients of the \( i \) th predictor, is

\[
f(\beta_i | Z_i, \sigma^2, \Sigma_Y, Y) \propto N_q(\mu_i, \Sigma_i),
\]

where \( \mu_i = X_{mi} - X_{i} \mu \). We also have

\[
\begin{align*}
\mu_i &= \left( \sigma^2 u_i^2 I_k \right)^{-1} + \sum_{m=1}^{n} X_{mi} \Sigma_Y^{-1} \sum_{m=1}^{n} \left( Y_m - X_{mi} \mu \right) X_{mi}^T, \\
\Sigma_i &= \left( \sigma^2 u_i^2 I_k \right)^{-1} + \sum_{m=1}^{n} X_{mi} \Sigma_Y^{-1} X_{mi}^T,
\end{align*}
\]

where \( X_{mi} \) is the \( m \)th observation of the \( i \)th predictor.

The conditional posterior distribution of \( \sigma^2 \) is

\[
f(\sigma^2 | Z, Y) \propto IG\left( \alpha_1 + pq, \alpha_2 + \sum_{i=1}^{p} \beta_i (u_i^2 I_k)^{-1} \beta_i^T \right).
\]

The conditional posterior probability of \( Z_i \) is

\[
P(Z_i = 1 | \beta, \sigma^2) = \frac{\phi_{\theta_q}(\beta_i; 0, \sigma^2 u_i^2 I_k)}{\phi_{\theta_q}(\beta_i; 0, \sigma^2 u_i^2 I_k) + (1 - \phi_q)\theta_q(\beta_i; 0, \sigma^2 u_i^2 I_k)},
\]

where \( \theta_q(\beta_i; .) \) is the value of a \( q \)-dimensional multivariate normal distribution evaluated at \( \beta_i \).

Finally, the conditional posterior distribution of \( \Sigma_Y \) is

\[
f(\Sigma_Y | \beta, Y) \propto IW\left( n + \nu + \sum_{m=1}^{n} (Y_m - X_{mi} \beta) (Y_m - X_{mi} \beta)^T \right).
\]

### 2.3 Estimation

Because all conditional posterior distributions have standard forms, we use Gibbs samplers to simulate and estimate them. The selection of the final model is based on the estimated marginal posterior probabilities of \( Z_i \), a higher of which indicates stronger associations between the corresponding predictor and the multivariate response. Therefore, we rank predictors by their estimated posterior probabilities \( P(Z_i = 1 | \beta, \sigma^2) \). To
determine how many predictors to keep and to identify the optimal model, we use the multivariate corrected Akaike information criterion (AICc) proposed by Bedrick and Tsai (1994):

\[ \text{AICc} = n\log|\Sigma_Y| + dq(n + p), \]

where \( d = n/(n - (p + q + 1)) \). The multivariate AICc considers the correlation between multiple response variables and puts a higher penalty on the model size than the conventional AIC and hence is more suitable for high-dimensional data with sparse structures; see Bedrick and Tsai (1994). In addition, it works better than many other univariate-based model selection criteria that treat multiple responses separately (Chen & Huang, 2012).

### 3 | SIMULATION

We examine the performance of the BMVS method using several simulation designs and also compare BMVS’s performance relative to many other relevant multivariate variable selection methods. We vary both the number of observations and the number of predictors. All the simulation results are calculated based on 100 replications. In each replicate, we perform 1000 burn-in iterations followed by 5000 update iterations to estimate unknown parameters.

We perform three different simulation settings based on the standard statistical multivariate regression model described in Equation (1). The predictors in these settings are continuous variables.

- **Setting 1**: Generate each predictor independently from a standard normal distribution: \( X_{ij} \approx \mathcal{N}(0, 1) \). Set the first ten predictors to be influential and generate their corresponding coefficients from a uniform distribution \( \beta_k \sim \text{Uniform}(1, 3) \) (yielding moderate effects), where \( 1 \leq k \leq 10 \) and \( 1 \leq j \leq q \). Set the coefficients of all other non-causal predictors to be zero. Let the \( kj \) th component of \( \Sigma_Y \) be \( \sigma_{ij} = 0.5^{k-1} \), where \( 1 \leq k \leq q \), and \( 1 \leq j \leq q \). Finally, connect \( Y \) with the truly influential predictors using model (1). Set the number of response variables to be \( q = 5 \).
- **Setting 2**: Maintain all other settings from Setting 1 but only increase the difficulty level by setting the number of response variables to be \( q = 30 \).
- **Setting 3**: Increase the difficulty level of Setting 1 from two changes: (1) Introduce correlations among predictors (instead of the independence outlined in Setting 1). The predictors are generated from a multivariate normal distribution: \( X_{ij} \sim \mathcal{N}_p(0, \Sigma_X) \). The \( kj \) th component of \( \Sigma_X \) is \( \sigma_{ij} = 0.5^{k-1} \), where \( 1 \leq k \leq p \) and \( 1 \leq j \leq p \). (2) Generate the coefficients of influential predictors from a uniform distribution \( \beta_k \sim \text{Uniform}(0.5, 0.8) \), which results in very weak signals that are not easily differentiated from the non-influential candidates. Maintain all other settings from Setting 1.
- **Setting 4**: We increase the number of influential predictors to twenty (i.e., the signal-to-noise ratio is also increased accordingly), which doubles the amount used in previous three settings. In addition, a random covariance matrix is employed. Let the \( kj \) th component of \( \Sigma_Y \) be \( \sigma_{ij} = \rho^{k-1} \), where \( \rho \sim \text{Uniform}(0.2, 0.8) \). Maintain all other settings from Setting 1.

We compare the BMVS method with three other variable selection methods: (1) CCA coupled with Wilk’s lambda test to assess the significance of each predictor, (2) multivariate LASSO (M-LASSO), (3) multivariate elastic net (M-EN) and (4) the GUESS algorithm. We implement M-LASSO and M-EN using the R package glmnet (Friedman et al., 2010), CCA using the R package CCA (González et al., 2008) and the GUESS algorithm using the R package R2GUESS (Liquet et al., 2016). To make the comparisons fair, we use the multivariate AICc to determine the model size for M-LASSO, M-EN and BMVS. The R2GUESS package comes with its own tuning procedure to choose the optimal model by ranking the model posterior probabilities (marginal likelihood \( \times \) prior probabilities) of all choices across the model space (Bottolo et al., 2013).

The results of the four simulation settings are summarized in Tables 1–4, respectively. The true model is denoted by \( t \), and the selected model is denoted by \( \hat{t} \). We used five criteria to assess the performance of these approaches: the average of posterior probabilities of all influential and non-influential predictors (mpp\(_1\) and mpp\(_0\), respectively), the power that the exact true model is selected (P(\( \hat{t} = t \))), the power that the true model is contained in the selected model (P(\( \hat{t} \supset t \))) and the FDR.

Despite the fact that all approaches have essentially equal high power in achieving \( \{ \hat{t} \supset t \} \), the advantage of the BMVS methods is that it maintains high power in selecting the exact true model \( \{ \hat{t} = t \} \) while also attaining low FDRs. In particular, when the M-EN and M-LASSO approaches fail to pick the exact model in any of the simulation replicate (indicated by \( P(\hat{t} = t) = 0 \)), the BMVS method achieves astonishingly high power of nearly 100% (see Tables 1 and 2). Note that it is much harder to achieve \( \{ \hat{t} = t \} \) than \( \{ \hat{t} \supset t \} \), and the difference between \( \{ \hat{t} = t \} \) and \( \{ \hat{t} \supset t \} \) is equivalent to false discoveries. It seems that the M-EN and M-LASSO approaches aggressively expand their model sizes while overfitting, scoring FDRs above 95%, to guarantee that the true model is contained in their selected sets. The BMVS approach, on the other hand, attains low FDRs (see Tables 1–4). The CCA approach also has a low FDR, but its power in \( P(\hat{t} = t) \) is only about one third of that of the BMVS approach. The average running times of the five methods compared (BMVS, CCA, M-LASSO, M-EN and R2GUESS) in Setting 4 are 223, 0.32,
### TABLE 1  Simulation results for Setting 1

|        | mpp$_1$ | mpp$_0$ | P($t = t$) | P($t \succ t$) | FDR  |
|--------|---------|---------|------------|----------------|------|
| BMVS   | 0.999   | 5.234 x 10^{-14} | 0.990      | 0.990          | 0.000 |
| CCA    | 0.390   | 0.960   | 0.085      |
| M-LASSO| 0.000   | 1.000   | 0.949      |
| M-EN   | 0.000   | 1.000   | 0.948      |
| R2GUESS| 0.990   | 1.000   | 0.001      |

- **n = 200, p = 500, q = 5, |t| = 10**

### TABLE 2  Simulation results for Setting 2

|        | mpp$_1$ | mpp$_0$ | P($t = t$) | P($t \succ t$) | FDR  |
|--------|---------|---------|------------|----------------|------|
| BMVS   | 1.000   | 3.009 x 10^{-105} | 1.000      | 1.000          | 0.000 |
| CCA    | 0.280   | 1.000   | 0.088      |
| M-LASSO| 0.000   | 1.000   | 0.945      |
| M-EN   | 0.000   | 1.000   | 0.941      |
| R2GUESS| 1.000   | 1.000   | 0.000      |

- **n = 200, p = 1000, q = 30, |t| = 10**

### TABLE 3  Simulation results for Setting 3

|        | mpp$_1$ | mpp$_0$ | P($t = t$) | P($t \succ t$) | FDR  |
|--------|---------|---------|------------|----------------|------|
| BMVS   | 0.995   | 0.001   | 0.750      | 0.950          | 0.001 |
| DC-SIS + BMVS | 0.880 | 0.950 | 0.006 |
| CCA    | 0.300   | 0.990   | 0.116      |
| M-LASSO| 0.000   | 1.000   | 0.949      |
| M-EN   | 0.000   | 1.000   | 0.948      |
| R2GUESS| 0.990   | 1.000   | 0.001      |

- **n = 200, p = 1000, q = 5, |t| = 10**
2.67, 0.25 and 35 s, respectively. The running times are evaluated on a Mac with 2.2 GHz Intel Core i7 CPU and 16 GB DDR4. As mentioned before, GUESS is a full likelihood algorithm, while the BMVS calculates the marginal likelihood, which requires less computational cost. Therefore, if the BMVS algorithm is able to be implemented in a more advanced platform such as the GPU like the GUESS does, it will also gain considerable improvement in the speed.

The results of BMVS are quite robust when the number of response variables increases from \( q = 5 \) to \( q = 30 \) in Simulation Setting 2 or when the number of influential predictors doubles from 10 to 20 in Simulation Setting 4. At a minimum, this simulation study demonstrates the potential of the BMVS approach in modelling a 30-dimensional multivariate response vector. This provides response dimensionality large enough for many real life applications. Setting 3 represents a more difficult case because its influential signal is very weak and the correlations between predictors add further confounding factors. As a result, it is to be expected that the results of Setting 3 are a little worse than those of the other settings. We demonstrate that an initial feature screening, distance correlation-based sure independence screening before applying BMVS (i.e., DC-SIS + BMVS), is able to improve the results (Li et al., 2012). Fan and Lv (2008) proposed the sure independence screening (SIS) procedure and proved that the Pearson correlation ranking procedure possesses a sure screening property. As an extension, Li et al. (2012) proposed the distance correlation-based SIS procedure, which can be applied to preselect important predictors based on their association strength with the multivariate response.

The performance of the GUESS is comparable with that of the BMVS approach for the majority cases and even performs better than BMVS in Simulation Setting 3. This is likely to be caused by the conservativeness of AICc in terms of selecting an optimal model size. As a result, the model selected by BMVS tends to be smaller than the true model in Simulation Setting 3, causing its power to be slightly lower than the other settings, but it does not imply that the posterior probabilities of the truly influential predictors generated by BMVS are also small. When a good sample size is given, BMVS significantly outperforms GUESS and accurately selects the exact true model 100% times in Simulation Setting 4.

Scott and Berger (2010) discussed the multiplicity correction issues in Bayesian variable selection and claimed that no fixed choice of \( \phi = P(Z_i = 1) \) that is independent of the sample size \( n \) can adjust for multiplicity. We recommend choosing \( \phi \) such that \( P(\sum_{i=1}^{p} Z_i > K) = 0.1 \), where \( K = \log n \), and the \( \phi \) value can be easily derived as a binomial distribution parameter. After varying the signal-to-noise ratios in the simulation studies, we find that the recommended prior probability is robust to different settings.

### Real Data Analyses

In this section, we apply our BMVS method to two real data examples in GWAS, each containing multiple response variables.

#### 4.1 Rice shape data

In this example, we analyse a dataset related to the shape of rice \((Oryza sativa)\). There are 3254 SNPs genotyped for each of the 179 rice accessions, with missing genotypes imputed by Iwata et al. (2015). The SNPs do not contain heterozygous genotypes due to the inbreeding nature of \( O. sativa \) (Zhao et al., 2011). Thus, the SNPs here only have genotype AA or aa. As a standard routine, SNPs with minor allele frequencies of less
than 5% are excluded in our analyses. Though using the same data, the previous work of Iwata et al. (2015) mainly focused on shape prediction, which is different from our emphasis for identifying the most influential genetic variants.

We retain component principal analysis on the shape descriptor, EFD coefficients, reported by Iwata et al. (2015). The first six PCs are retained as the response of our model \( (q = 6) \), which explain about 99% of the total variation in the shape descriptor coefficients (the first six PCs explain 94.66%, 1.85%, 1.00%, 0.57%, 0.52% and 0.26% of the total variation, respectively). We then perform a fivefold cross-validation on the entire observed data and calculate the following numerical assessments: (1) the average size of selected models \( \langle |t| \rangle \), (2) the Frobenius matrix norm of the difference between predicted phenotypes and observed phenotypes \( \| Y - \hat{Y} \|_F \) and (3) the multivariate AICc of selected models. The BMVS method is again compared with GUESS (R2GUESS), M-LASSO (glmnet) and M-EN (glmnet). As indicated in Section 2 and theoretical results (see the Supporting Information), this work focuses on variable selection in high-dimensional multivariate regressions. In this real data analysis section, we use the prediction accuracy as a way to assess the performance of variable selection because the ground truth is unknown in real datasets. Theoretically speaking, a model containing all the truth (and also minimum false positives) should have the best prediction power.

Table 5 shows the results of the fivefold cross-validation obtained using these four methods. The BMVS method produces the lowest values of AICc (724.6854) among all of the four methods. The M-LASSO approach selects an average model size of 139, which is around 12 times larger than that of BMVS, dramatically inflating both its prediction error and its AICc. Although M-EN yields the lowest prediction error, both its AICs and model size are large. The GUESS algorithm selects a slightly smaller model size than BMVS, but at the cost of increasing both AICc and prediction error. In addition to comparing BMVS with standard multivariate variable selection approaches, we also compare it with its univariate counterpart to assess whether a multivariate approach performs better than a univariate approach when handling multiple responses. Here, “univariate counterpart” refers to the process of individually fitting each component of the multivariate response using univariate approaches and repeating the process six times. To make comparison fair and also minimize all other possible confounding factors, we used the Bayesian variable selection method proposed by Narisetty and He (2014) from which the BMVS is extended. This “univariate counterpart” method yields a predictor error \( \| Y - \hat{Y} \|_F \) of 7.0076, which is about three times larger than that yielded by BMVS.

### 4.2 | Flowering time (FT) data

In this example, we consider the GWAS for three FT-related response of 272 rice O. sativa accessions: FT at Aberdeen, FT ratio of Arkansas/Aberdeen and FT ratio of Faridpur/Aberdeen. These three phenotypes of the same rice accession are correlated, sharing genetic basis in some way. The predictor data set includes 36,901 SNPs that were genotyped by Zhao et al. (2011) and imputed by Iwata et al. (2015). For this ultrahigh-dimensional setting, we perform the DC-SIS + BMVS process. Specifically, we keep 200 candidate SNPs from the initial feature screening and then perform BMVS on this subset to further detect the truly influential SNPs.

Zhao et al. (2011) conducted GWAS for the same dataset but used the linear mixed model to fit each individual response separately. They reported significant SNPs (\( p \text{ value} \leq 10^{-4} \)) for each individual phenotype. Here, we compare three sets of results as follows: (1) the SNPs selected by the BMVS method, denoted by \( \hat{t}_{BMVS} \); (2) the SNPs that are simultaneously significant for all three phenotypes as reported by Zhao et al. (2011), denoted by \( \hat{t}_1 \) (the intersection set); and (3) the SNPs that are significant for at least one phenotype as reported by Zhao et al. (2011), denoted by \( \hat{t}_2 \) (the union set).

The BMVS method selects nine SNPs, the intersection set \( \hat{t}_1 \) contains 70 SNPs and the union set \( \hat{t}_2 \) contains 215 SNPs. To perform a fair comparison, we apply a multivariate linear regression model to perform fivefold cross-validation for these three sets of results (see Table 6). BMVS outperforms the other methods by using the smallest model size to achieve the best results. Specifically, it achieves both the smallest prediction error (16 versus 64 or 997) and the smallest AICc (−179 versus 227 or 5294). The negative sign of the AICc obtained by BMVS may cause some confusion, but the AICc rules indicate smaller results are always more desirable than larger ones. Compared with \( \hat{t}_{BMVS} \), \( \hat{t}_2 \) is astonishingly large in both the selected model size and the predictor error. In addition, we further compare the significance of the three sets by directly fitting a multivariate regression model that uses all observations (without doing cross-validation). Eight out of nine selected SNPs are significant for \( \hat{t}_{BMVS} \), while only 10 out of 70 selected SNPs are significant for \( \hat{t}_1 \) and only 52 out of 215 are significant for \( \hat{t}_2 \). All these analyses are made at the 0.05 level without adjustment for multiple comparisons.

**TABLE 5** Results of real data analyses on rice shape data using fivefold cross-validation

|     | \( |t| \) | \( \| Y - \hat{Y} \|_F \) | AICc   |
|-----|-------|-----------------|-------|
| BMVS | 11.80 | 2.8626          | 724.6854 |
| R2GUESS | 8.40 | 6.0624          | 883.5024 |
| M-LASSO | 139.60 | 4.4521          | 9915.9830 |
| M-EN  | 101.60 | 1.7921          | 3716.6670 |
To visualize the associations between the three-dimensional response vector and each of the nine influential SNPs selected by BMVS (i.e., set $t_{BMVS}$), we use the multidimensional scaling (MDS) approach to demonstrate the Euclidean distances of 272 observations in a two-dimensional space (see Figure 1) (Kruskal & Wish, 1978). The MDS approach calculates the pairwise Euclidean distance matrix of all the observations and extract eigenvectors associated with the top two eigenvalues. This ensures that the pairwise Euclidean distances can be approximated.

### Table 6

Results of real data analyses on flowering time data using fivefold cross-validation

| $t_{BMVS}$ | $|Y - \hat{Y}|_F$ | AICc  |
|------------|------------------|-------|
| 9          | 16.6450          | -179.2434 |
| 70         | 64.7921          | 227.8176  |
| 215        | 997.0443         | 5294.179 |

**Figure 1** The multidimensional scaling plot demonstrating variation of the multivariate response vector under different genotype groups in a two-dimensional space for each of the nine influential SNPs selected by BMVS approach, with black dots representing genotype $aa$ and red dots representing genotype $AA$. 

To visualize the associations between the three-dimensional response vector and each of the nine influential SNPs selected by BMVS (i.e., set $t_{BMVS}$), we use the multidimensional scaling (MDS) approach to demonstrate the Euclidean distances of 272 observations in a two-dimensional space (see Figure 1) (Kruskal & Wish, 1978). The MDS approach calculates the pairwise Euclidean distance matrix of all the observations and extract eigenvectors associated with the top two eigenvalues. This ensures that the pairwise Euclidean distances can be approximated.
and visualized in a lower dimensional space (see Figure 1). In each subplot of Figure 1, the colour coding represents different genotypes for each of the nine influential SNPs, with black dots for genotype aa and red dots for genotype AA. Apparently, the variation of the response vector shows quite different patterns of distribution under two different genotypes of each of the nine influential SNPs, which may indicate that the associations are not trivial.

5 | DISCUSSION

One appealing characteristic of the BMVS method is that it skips the mundane process of tuning parameter selection, as required by penalized regression approaches. All simulation and real data examples use the same value for hyperparameters with no additional tunings required. Another appealing property is that the theoretical selection consistency (see the Supporting Information) holds for ultrahigh-dimensional cases, including situations where the number of predictors is exponentially larger than the sample size (i.e., \( p = e^{(n)} \)). This work primarily focuses on variable selection. We evaluated the variable selection performance of the BMVS method in Tables 1–4 through four different simulation settings. However, unlike simulated data with given truth, the true variables are not observable in real data, and hence, it is not feasible to evaluate variable selection performance. Therefore, prediction performance measures are used to indirectly compare how efficient the selected small number of variables obtained from the different variable selection methods in achieving good predictions. If prediction is the main interest for some future work, one can adjust the BMVS approach accordingly by adopting other criteria such as cross-validation or bootstrap prediction error in place of the multivariate AICc to further improve the prediction accuracy.

In general, Bayesian methods are more computationally expensive than frequentist approaches due to their large number of hyperparameters and long iteratively updating process. Bhattacharya et al. (2016) and Narisetty et al. (2018) proposed fast sampling methods for high-dimensional multivariate Gaussian distributions, which may be exploited to improve the speed for BMVS. For the ultrahigh-dimensional data such as GWAS that we analysed in this article, we suggest using a fast initial feature screening method before applying the BMVS method. As one can see from the results of our Simulation 3 and FT data analyses, the combination of DC-SIS and BMVS can further improve the accuracy.

Many studies modelled a multivariate response by separately fitting each component of the response vector using multiple univariate models (Chitwood et al., 2014; Fu et al., 2017; Narisetty & He, 2014; Zhang et al., 2009). Compared with the BMVS approach, which handles the multivariate response as one unit, the separate univariate regression approach has three notable disadvantages: (1) They fail to take advantage of the correlations between multiple response variables (Breiman & Friedman, 1997); but correlated variables indeed share information in a way that is important for discovering additional signals (O’Reilly et al., 2012). (2) When a series of models are fitted, deriving a generalized interpretation for multiple response variables is difficult (Chitwood et al., 2014; Fu et al., 2017). (3) It is more appropriate to group multiple correlated response variables together for some practical applications, and hence, separating them could produce incomplete or misleading results. Breiman and Friedman (1997) provided theoretical proofs demonstrating that prediction accuracy can be improved when modelling multiple correlated response variables by using one multivariate regression model compared with using multiple separate univariate regression models. This claim is empirically confirmed in two of our real data analyses.

For the future work, similar skills that are proposed for univariate selection can be directly adopted to determine the optimal model size for our multivariate model, that is, the threshold to determine how many predictors to keep.

CONFLICT OF INTEREST

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The authors confirm that all data underlying the findings are fully available without restriction. The phenotypic data and the SNP information of *Oryza sativa* can be found at https://www.nature.com/articles/ncomms1467. The EFD coefficients of O. sativa grain can be found at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4380318/.

REFERENCES

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32(3), 870–897.
Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3), 1550–1577.
Bayarri, M. J., & García-Donato, G. (2008). Generalization of Jeffreys divergence-based priors for Bayesian hypothesis testing. *Journal of the Royal Statistical Society: Series B*, 70(5), 981–1003.
Bedrick, E. J., & Tsai, C.-L. (1994). Model selection for multivariate regression in small samples. *Biometrics*, 50, 226–231.
Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433), 109–122.
How to cite this article: Dai, X., Fu, G., Reese, R., Zhao, S., & Shang, Z. (2022). An approach of Bayesian variable selection for ultrahigh-dimensional multivariate regression. Stat., 11(1), e476. https://doi.org/10.1002/sta4.476