Published in final edited form as:
J Multivar Anal. 2017 May ; 157: 14–28. doi:10.1016/j.jmva.2017.02.007.

Bayesian sparse reduced rank multivariate regression

Gyuhyeong Goh[a], **Dipak K. Dey**[b], and **Kun Chen**[b]

[a]Department of Statistics, Kansas State University, Manhattan, KS 66506, United States

[b]Department of Statistics, University of Connecticut, Storrs, CT 06269, United States

Abstract

Many modern statistical problems can be cast in the framework of multivariate regression, where the main task is to make statistical inference for a possibly sparse and low-rank coefficient matrix. The low-rank structure in the coefficient matrix is of intrinsic multivariate nature, which, when combined with sparsity, can further lift dimension reduction, conduct variable selection, and facilitate model interpretation. Using a Bayesian approach, we develop a unified sparse and low-rank multivariate regression method to both estimate the coefficient matrix and obtain its credible region for making inference. The newly developed sparse and low-rank prior for the coefficient matrix enables rank reduction, predictor selection and response selection simultaneously. We utilize the marginal likelihood to determine the regularization hyperparameter, so our method maximizes its posterior probability given the data. For theoretical aspect, the posterior consistency is established to discuss an asymptotic behavior of the proposed method. The efficacy of the proposed approach is demonstrated via simulation studies and a real application on yeast cell cycle data.

Keywords

Bayesian; Low rank; Posterior consistency; Rank reduction; Sparsity

1. Introduction

In various fields of scientific research such as genomics, economics, image processing, astronomy, etc., massive amount of data are routinely collected, and many associated statistical problems can be cast in the framework of multivariate regression, where both the number of response variables and the number of predictors are possibly of high dimensionality For example, in genomics study, it is critical to explore the relationship between genetic markers and gene expression profiles in order to understand the gene regulatory network; in a study of human lung disease mechanism, the detailed CT-scanned lung imaging data enable us to examine the systematic variations in airway tree measurements across various lung disease status and pulmonary function test results. To

Correspondence to: Gyuhyeong Goh.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

formulate, suppose we have n independent observations of the response vector $\mathbf{y}_i \in \mathbb{R}^q$ and the predictor vector $\mathbf{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$. Consider the multivariate linear regression model

$$\boldsymbol{Y}=\mathbf{X}\mathbf{C}+\mathbf{E}, \qquad (1)$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\top \in \mathbb{R}^{n\times q}$ is the response matrix, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n\times p}$ is the predictor matrix, $\mathbf{C} \in \mathbb{R}^{n\times q}$ is the unknown regression coefficient matrix, and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)^\top \in \mathbb{R}^{n\times q}$ is the error matrix with $\mathbf{e}_i$'s being independently and identically distributed (i.i.d.) with mean zero and covariance matrix $\mathbf{\Sigma}_e$, a $q \times q$ positive definite matrix. Following Bunea et al. [8, 9], Chen et al. [11] and Mukherjee et al. [28], we assume $\mathbf{\Sigma}_e = \sigma^2\mathbf{I}_q$. We further assume the response variables and the predictors are all centered, and there is no intercept term. In what follows, we use $\mathbf{a}_j^\top$ to denote the j^{th} row of a generic matrix $\mathbf{A}$ and $\tilde{\mathbf{a}}_\ell$ the ℓ^{th} column of $\mathbf{A}$, e.g., $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_p)^\top = (\tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_q)$. A fundamental goal of multivariate regression is thus to estimate and make inference about the coefficient matrix $\mathbf{C}$ so that meaningful dependence structure between the responses and predictors can be revealed.

When the predictor dimension p and the response dimension q are large relative to the sample size n, classical estimation methods such as ordinary least squares (OLS) may fail miserably. The curse of dimensionality can be mitigated by assuming that $\mathbf{C}$ admits certain low-dimensional structures, and regularization/penalization approaches are then commonly deployed to conduct dimension reduction and model estimation. The celebrated reduced rank regression (RRR) [2, 24, 32] achieved dimension reduction through constraining the coefficient matrix $\mathbf{C}$ to be rank deficient, building upon the belief that the response variables are related to the predictors through only a few latent directions, i.e., some linear combinations of the original predictors. As such, low-rank structure induces and models dependency among responses, which is the essence of conducting multivariate analysis. Bunea et al. [8] generalized the classical RRR to high dimensional settings, casting reduced-rank estimation as a penalized least squares problem with the penalty being proportional to the rank of $\mathbf{C}$. Yuan et al. [37] utilized the nuclear norm penalty, defined as the ℓ_1 norm of the singular values. See also, Chen et al. [11], Mukherjee and Zhu [29], Negahban and Wainwright [30], and Rohde and Tsybakov [33].

It is worth noting that low-rankness in $\mathbf{C}$ is of intrinsic multivariate nature; when combined with row and/or column-wise sparsity, it can further lift dimension reduction and facilitate model interpretation. For example, in the aforementioned genomics study, it is plausible that the gene expression profiles (responses) and the genetic markers (predictors) are associated through only a few latent pathways (linear combinations of possibly highly-correlated genetic markers), and moreover, very likely such linear associations only involve a small subset of genetic markers and/or gene profiles. Therefore, recovering a low-rank and also sparse coefficient matrix $\mathbf{C}$ in model (1) hold the key to reveal such interesting connections between the responses and predictors. Chen et al. [10] proposed a regularized sparse singular value decomposition (SVD) approach with known rank, in which each latent variable is constructed from only a subset of the predictors and is associated with only a subset of the responses. Chen and Huang [12] proposed a rank-constrained adaptive group Lasso approach to recover a low-rank coefficient matrix C with sparse rows; for each zero row in **C**, the corresponding predictor is then completely eliminated from the model. Bunea et al. [9] also proposed a joint sparse and low-rank estimation approach and derived its nonasymptotic oracle error bounds. Both methods required to solve the nonconvex rank-constrained problem by fitting models of various ranks. Recently, Ma et al. [27] proposed a subspace assisted regression with row sparsity method which was shown to achieve near optimal nonasymptotic minimax rates in estimation.

While all the aforementioned regularized regression techniques produce attractive point estimators of the coefficient matrix **C**, it remains a difficult problem to assess the uncertainty of the obtained estimators. To overcome this limitation, there has already been a rich literature on Bayesian approaches of the reduced rank regression. From a Bayesian perspective, the unknown parameter is considered as a random variable, and thus the statistical inference can be made by the posterior distribution. The first attempt to develop the Bayesian reduced rank regression was made by Geweke [20]. The coefficient matrix is assumed to be $\mathbf{C} = \mathbf{A}\mathbf{B}^\top$ with $\mathbf{A} \in \mathbb{R}^{p\times r}$ and $\mathbf{B} \in \mathbb{R}^{q\times r}$, where $r < \min(p, q)$ is assumed to be known. Then, by assigning Gaussian prior for each elements of **A** and **B**, the induced posterior achieves the low-rank structure of the prespecified rank. As an alternative, Lim and Teh [26] proposed to start from the largest possible rank $r = \min(p, q)$, assign a column-wise shrinkage Gaussian prior on each columns of **A** and **B**. The posterior for redundant columns of **A** and **B** is forced to be concentrated around zero, so the (approximate) rank reduction can be accomplished. The main challenge of this Bayesian approach is the choice of the hyperparameters of the Gaussian priors in order to control the amount of shrinkage. There have been several attempts to overcome this challenge by assigning priors on the hyperparameters, so that they can be determined in the estimation procedure. For instance, Salakhutdinov and Mnih [34] proposed to utilize the Wishart distribution as the hyperprior. Similar hierarchical Bayesian methods were also proposed in the context of matrix completion, matrix completion deals with missing values, but we do not [4, 40]. However, none of the aforementioned studies dealt with the sparsity of the coefficient matrix C. Recently, Zhu et al. [41] introduced a Bayesian low-rank regression model with high-dimensional responses and covariates. To enable sparse estimation under low rank constraint with a prefixed rank, they utilized a sparse singular value decomposition (SVD) structure [10] with Gaussian-mixtures of gamma priors on all the elements of the decomposed matrices. Then, the sparsity of C was achieved using Bayesian thresholding method. For a survey on Bayesian reduce rank models, see Alquier [1] and the references therein.

We develop in this article a novel Bayesian simultaneous dimension reduction and variable selection approach. Our method aims to tackle several challenges regarding both the estimation and inference in the sparse and low-rank regression problems. First, the proposed method enables us to simultaneously estimate the unknown rank and remove irrelevant predictors, in contrast to several existing methods in which rank selection has to be resolved by comparing fitted models of various ranks or by some ad hoc approach such as scree plot. In addition, we also seek potential column sparsity of the coefficient matrix, so that it is applicable to problems with high-dimensional responses where response selection is highly desirable (to be elaborated below). Second, by careful construction of the prior distribution, our method alleviates the many difficulties brought by the use of nonsmooth and non-convex penalty functions and by the tuning parameter selection procedure in penalized regression analysis. From a Bayesian perspective, the penalty function can be viewed as a negative logarithm of the prior density function [25, 31, 36]. We develop a general prior for **C** mimicking the rank penalty and the group ℓ_0 row/column penalty, to achieve simultaneous rank reduction and variable selection through the induced posterior distribution, yet the computation is kept tractable and efficient, where the group ℓ_0 penalty directly restricts the number of nonzero rows and columns. Since the tuning parameters are considered as random variables in our Bayesian formulation, the optimal ones are selected to achieve the highest posterior probability given the data. Furthermore, using our Bayesian approach, the credibility intervals for the regression coefficients and their functions can be easily constructed using the Markov Chain Monte Carlo (MCMC) technique. In contrast, there has been little work on quantifying the estimation uncertainty in regularized regression approaches.

We now formally state our assumptions or prior beliefs about the coefficient matrix **C** in model (1).

A1. (Reduced rank) $r^* \leq r$, where $r^* = \mathrm{rank}(\mathbf{C})$ indicates the rank of **C** and $r = \min(p, q)$.

A2. (Row-wise sparsity) $p^* \leq p$, where $p^* = \mathrm{card}(\{j : \mathbf{c}_j^\top \mathbf{c}_j \neq 0\})$ and $\mathbf{c}_j^\top$ denotes the j^{th} row of **C**, where card(·) denotes the cardinality of a set.

A3. (Column-wise sparsity) $q^* \leq q$, where $q^* = \mathrm{card}(\{\ell : \tilde{c}_l^\top \tilde{c}_\ell \neq 0\})$ and $\tilde{\mathbf{c}}_\ell$ denotes the ℓ^{th} column of **C**.

A1 states that **C** is possibly of low rank. In A2, excluding the j^{th} predictor from model (1) is equivalent to setting all entries of the j^{th} row of **C** as zero. Therefore, the first two assumptions concern rank reduction and predictor selection. The third assumption is about “response selection”, i.e., if the ℓ^{th} column of **C** is zero, the ℓ^{th} response is modeled as a noise variable. While such structural assumption can be treated as optional depending on the specific application, we stress that there are many circumstances where response selection is highly desirable [10]. For example, in many applications the dimension of the responses can be very high, and there may exist noise variables that are not related to any predictors in the model. In addition, eliminating irrelevant predictors dramatically reduces the number of free parameters of the model and thus it improves the accuracy of parameter estimation [5]. In addition, allowing possible response selection provides more flexibility and generality in the multivariate linear regression framework, since the case of selecting all responses can be viewed as a special case.

The remainder of the paper is organized as follows. In Section 2, we briefly introduce a general penalized regression approach for conducting sparse and low-rank estimation. In Section 3, we develop our new Bayesian approach, and explore the connections between the penalized least squares and our Bayes estimators. The full conditionals are obtained in Section 4, and we describe the posterior optimization algorithm and posterior sampling technique. In Section 5, we study the posterior consistency of the proposed method.

Simulation studies and a real application on yeast cycle data are presented in Section 6 and Section 7. Some concluding remarks are given in Section 8.

2. Penalized regression approach

In the regularized estimation framework, the unknown coefficient matrix **C** in model (1) can be estimated by the following penalized least squares (PLS) method,

$$\hat{\mathbf{C}}_{\text{pls}} = \arg\min_{\mathbf{C}} \left\{\frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}\mathbf{C}\|_F^2 + \mathscr{P}_\lambda(\mathbf{C})\right\}, \tag{2}$$

where $\|\mathbf{C}\|_F = \sqrt{\text{trace}(\mathbf{C}^\top\mathbf{C})} = \sqrt{\text{trace}(\mathbf{C}\mathbf{C}^\top)}$ denotes the Frobenius norm and $\mathscr{P}_{\boldsymbol{\lambda}}(\mathbf{C})$ is a penalty function with non-negative tuning parameter $\boldsymbol{\lambda}$ controlling the amount of regularization. It is natural to construct a penalty function of an additive form,

$$\mathscr{P}_\lambda(\mathbf{C}) = \mathscr{P}_{\lambda_1}^{\text{RR}}(\mathbf{C}) + \mathscr{P}_{\lambda_2}^{\text{RS}}(\mathbf{C}) + \mathscr{P}_{\lambda_3}^{\text{cs}}(\mathbf{C}),$$

where $\mathscr{P}_{\lambda_1}^{\text{RR}}(\mathbf{C})$, $\mathscr{P}_{\lambda_2}^{\text{RS}}(\mathbf{C})$ and $\mathscr{P}_{\lambda_3}^{\text{CS}}(\mathbf{C})$ induce the low-rankness, row-wise sparsity and column-wise sparsity in **C**, with tuning parameters λ_1, λ_2 and λ_3, respectively. There are numerous choices of the penalty functions. Note that the rank of matrix **C** is same as the number of non-zero singular values, i.e., rank(**C**) = card $(\{k: s_k(\mathbf{C}) > 0\}) = r^*$, where $s_k(\mathbf{C})$ denotes the k^{th} singular value of **C**. Hence, rank reduction can be achieved by penalizing the singular values of **C**, i.e.,

$$\mathscr{P}_{\lambda_1}^{\text{RR}}(\mathbf{C}) = \lambda_1 \sum_{k=1}^{r} \rho_1\{s_k(\mathbf{C})\}, \tag{3}$$

where ρ_1 is a sparsity-inducing penalty function. In particular, choosing $\rho_1(|a|) = \mathbf{1}\,(|a| \neq 0)$ corresponds to directly penalizing/restraining the rank of **C**, and that $\rho_1(|a|) = |a|^{\beta_1}$ gives the Schatten-β quasi-norm penalty when $0 < \beta_1 < 1$ and the convex nuclear norm penalty $\lambda_1\|\mathbf{C}\|_*$ when $\beta_1 = 1$, where $\|\mathbf{C}\|_* = \sum_{k=1}^{r} s_k(\mathbf{C})$ denotes the nuclear norm. For promoting rowwise/column-wise sparsity, selecting or eliminating parameters by groups is needed, which can be achieved by penalizing the row/column ℓ_2 norms of **C**,

$$\mathscr{P}_{\lambda_2}^{\text{RS}}(\mathbf{C}) = \frac{1}{2}\lambda_2 \sum_{j=1}^{p} \rho_2(\|\mathbf{c}_j\|_2), \tag{4}$$

$$\mathscr{P}^{\mathrm{CS}}_{\lambda_3}(\mathbf{C}) = \frac{1}{2}\lambda_3 \sum_{\ell=1}^{q} \rho_3 \|\tilde{\mathbf{c}}_\ell\|_2), \tag{5}$$

where $\|\mathbf{c}\|_2 = \sqrt{\mathbf{c}^\top \mathbf{c}}$ denotes the ℓ_2 norm. Choosing $\rho_2(|a|) = \mathbf{1}\ (|a| \neq 0)$ corresponds to directly counting and penalizing the number of nonzero rows, and $\rho_2(|a|) = |a|$ corresponds to the convex group Lasso penalty [38]. Other methods include group SCAD [16] and group MCP [7, 39]; see Fan and Lv [18] and Huang et al. [22] for comprehensive reviews. In principal, rank reduction and variable selection can be accomplished by solving the PLS problem (2) with any sparsity-inducing penalties ρ_1, ρ_2 and ρ_3.

The pros and cons of using convex penalties in model selection are well understood. In low-rank estimation, Bunea et al. [8] showed that while the convex nuclear norm penalized estimator has similar estimation properties to those of the nonconvex rank penalized estimator, the former requires stronger conditions and is in general not as parsimonious as the latter in rank selection. For sparse group selection, it is known that the convex group Lasso criterion often leads to over-selection and substantial estimation bias, and adopting nonconvex penalties may lead to superior properties in both model estimation and variable selection under milder conditions [23, 27]. Unfortunately, the nonconvexity of a penalized regression criterion also imposes great challenges in both understanding its theoretical properties and solving the optimization problem in computation. Therefore, trading off computation efficiency and statistical properties is critical in formulating penalized estimation criterion, and it is particularly relevant when dealing with large data applications. The problem of tuning parameter selection can also be troublesome, especially so for the problem of interest here as it requires multiple tuning parameters. Furthermore, it is still a largely unsolved problem on how to make statistical inference and attach error measures to any penalized estimator. All these concerns motivate us to tackle the sparse and low-rank estimation problem in a Bayesian fashion, to achieve a computationally efficient implementation and be able to make valid inference about the composite low-dimensional structure of **C**.

3. Bayesian sparse and low-rank regression

From a Bayesian perspective, the PLS estimator in (2) can be viewed as the maximum a posteriori (MAP) estimator from the following posterior density function,

$$\pi(\mathbf{C}|\boldsymbol{Y}, \lambda,) \propto f(\boldsymbol{Y}|\mathbf{C})\pi(\mathbf{C}|\lambda) \propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{Y} - \mathbf{X}\mathbf{C}\|_F^2\right) \exp\left\{-\frac{1}{\sigma^2}\mathscr{P}_\lambda(\mathbf{C})\right\},$$

where $f(\mathbf{Y} \mid \mathbf{C})$ denotes the likelihood function and $\pi(\mathbf{C} \mid \boldsymbol{\lambda})$ denotes the prior density function of **C** given the tuning parameter $\boldsymbol{\lambda}$. Since the MAP estimator of **C** is free of $\boldsymbol{\sigma}^2$, without loss of generality, throughout this paper we assume that $\boldsymbol{\sigma}^2 = 1$. Motivated by the connections between PLS and MAP and by the penalty function defined in (3)-(5), it is natural to consider the following prior,

$$\pi(\mathbf{C}|\lambda) \propto \exp\left\{-\tfrac{1}{2}\lambda_1 \sum_{k=1}^{r} \mathbf{1}(|s_k(\mathbf{C})|>\varepsilon)\right\} \\ \times\exp\left\{-\tfrac{1}{2}\lambda_2 \sum_{j=1}^{p} \mathbf{1}(\|\mathbf{c}_j\|_2>\varepsilon)\right\} \\ \times\exp\left\{-\tfrac{1}{2}\lambda_3 \sum_{\ell=1}^{q} \mathbf{1}(\|\tilde{c}_\ell\|_2>\varepsilon)\right\}, \qquad (6)$$

where $\varepsilon > 0$ is a suitably small value and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$. Note that as $\varepsilon \to 0$, this prior converges to the following penalty function under the MAP estimation:

$$\mathscr{P}_{\boldsymbol{\lambda}}(\mathbf{c}) = \frac{1}{2}\lambda_1\sum_{k=1}^{r}\ell_0(|s_k(\mathbf{C})|) + \frac{1}{2}\lambda_2\sum_{j=1}^{p}\ell_0(\|\mathbf{c}_j\|_2) + \frac{1}{2}\lambda_3\sum_{\ell=1}^{q}\ell_0(\|\tilde{\mathbf{c}}_\ell\|_2).$$

In penalized regression, this ℓ_0 penalty corresponds to directly penalize the rank, the number of nonzero rows, and the number of nonzero columns of **C**, which leads to an intractable combinatory problem. Similarly, in Bayesian framework, even though this setup directly targets on the desired structure of **C**, there are several difficulties in using such a prior distribution. Since the prior density function in (6) involves the singular values of **C**, it induces an improper posterior distribution, as it is generally difficult to be considered as a probability density function of **C**. Moreover, the non-differentiability of indicator function induces a discontinuous posterior density function.

To overcome the first difficulty, i.e., avoiding direct use of the singular values, we propose an indirect modeling method through decomposing the matrix **C**. We write $\mathbf{C} = \mathbf{A}\mathbf{B}^\top$, where **A** is a $p \times r$ matrix, **B** is a $q \times r$ matrix, and r is an upper bound of the true rank r^* of **C**, e.g., a trivial one is $r = \min(p, q)$. Apparently such a decomposition is not unique, as with any nonsingular $r \times r$ matrix **Q**, $\mathbf{C} = \mathbf{A}\mathbf{B}^\top = \mathbf{A}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{B}^\top = \tilde{\mathbf{A}}\tilde{\mathbf{B}}^\top$ where $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{Q}$ and $\tilde{\mathbf{B}} = \mathbf{B}(\mathbf{Q}^{-1})^\top$. Interestingly, the following lemma reveals that the low-rankness and the row/column sparsity of **C** can all be represented as certain row/column sparsity of **A** and **B**, and more importantly, the representations are invariant to any nonsingular transformation.

Lemma 1. Let $\mathbf{C} \in \mathbb{R}^{p\times q}$, and suppose $\mathbf{C} = \mathbf{A}\mathbf{B}^\top$ for some $\mathbf{A} \in \mathbb{R}^{p\times r}$, $\mathbf{B} \in \mathbb{R}^{q\times r}$ with $r = \min(p, q)$. Let $\tilde{\mathbf{a}}_k$ and $\tilde{\mathbf{b}}_k$ denote the k^{th} column of **A** and **B**, respectively. Let $\mathbf{a}_j^\top$ and $\mathbf{b}_\ell^\top$ denote the j^{th} row of **A** and the ℓ^{th} *row of* **B**, respectively. Then

1. $$\operatorname{rank}(\mathbf{C}) \le \min(p, q) - \textstyle\sum_{k=1}^{r}\mathbf{1}\left(\|\tilde{\mathbf{a}}_k\|_2 + \|\tilde{\mathbf{b}}_k\|_2 = 0\right).$$
2. $\{j: \|\mathbf{c}_j\|_2 \neq 0\} \subset \{j: \|\mathbf{a}_j\|_2 \neq 0\}$.
3. $\{\ell: \|\tilde{\mathbf{c}}_\ell\|_2 \neq 0\} \subset \{\ell: \|\mathbf{b}_\ell\|_2 \neq 0\}$.

The first statement in the above lemma suggests the following rank-reducing prior,

$$\pi^{\mathrm{RR}}(\mathbf{A},\mathbf{B}|\lambda_1) \propto \exp\left\{-\frac{1}{2}\lambda_1\sum_{k=1}^{r}\mathbf{1}\left(\|\tilde{\mathbf{a}}_k\|_2+\|\tilde{\mathbf{b}}_k\|_2>\varepsilon\right)\right\}, \quad (7)$$

where $\varepsilon > 0$ and $\lambda_1 > 0$. For a sufficiently small ε, this prior induces sparsity on columns of **A** and **B** simultaneously and thus reduces the rank of **C**. Similarly, Lemma 1 suggests the following row-wise and column-wise sparsity-inducing priors,

$$\pi^{\mathrm{RS}}(\mathbf{A}|\lambda_2) \propto \exp\left\{-\frac{1}{2}\lambda_2\sum_{j=1}^{p}\mathbf{1}(\|\mathbf{a}_j\|_2>\varepsilon)\right\}, \quad (8)$$

$$\pi^{\mathrm{cs}}(\mathbf{B}|\lambda_3) \propto \exp\left\{-\frac{1}{2}\lambda_3\sum_{\ell=1}^{q}\mathbf{1}(\|\mathbf{b}_\ell\|_2>\varepsilon)\right\}, \quad (9)$$

where $\varepsilon > 0$, $\lambda_2 > 0$ and $\lambda_3 > 0$. Combining (7), (8) and (9) leads us to the following prior,

$$\pi(\mathbf{A},\mathbf{B}|\boldsymbol{\lambda}) \propto \pi^{\mathrm{RR}}(\mathbf{A},\mathbf{B}|\lambda_1)\pi^{\mathrm{RS}}(\mathbf{A}|\lambda_2)\pi^{\mathrm{CS}}(\mathbf{B}|\lambda_3), \quad (10)$$

where $\lambda = (\lambda_1, \lambda_2, \lambda_3)$.

The discontinuity problem still presents in (10) due to the indicator functions. We address this problem by approximating ℓ_0-norm by a well-behaved smooth function. Let **D** be a $m \times n$ matrix. Define

$$\mathscr{P}_{\lambda,\omega}(\mathbf{D})=\frac{1}{2}\lambda\sum_{i=1}^{m}\frac{\mathbf{d}_i^\top\mathbf{d}_i}{(\omega+\mathbf{d}_i^\top\mathbf{d}_i)^{1-\frac{\beta}{2}}}, \quad (11)$$

where $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_\mathrm{m})^\top$, $0 \le \beta \le 1$ and $\omega > 0$. We have

$$\frac{\mathbf{d}_i^\top\mathbf{d}_i}{(\omega+\mathbf{d}_i^\top\mathbf{d}_i)^{1-\frac{\beta}{2}}} \to (\|\mathbf{d}_i\|_2)^\beta,$$

as $\omega \to 0$, where we define $(0)^0 = 0$. This implies that the proposed penalty approximates the group ℓ_β-norm penalty when ω is sufficiently small. In particular, when $\beta = 0$ and ω is chosen to be a small positive constant ω_0, (11) gives an approximate group ℓ_0-norm penalty and produces approximately sparse solutions, while it is continuous as well as differentiable with respect to **D**. In all our numerical studies, we set $\omega_0 = 10^{-10}$ and utilize tolerance level 10^{-5} to determine zero estimates. Fig. 1 shows the plots of $f(x) = x^2/(x^2 + \omega_0)$ for varying $\omega_0 = 10^{-k}$, k = 4, 6, 8,10. Indeed, when $w_0 = 10^{-10}$, the function closely mimics the ℓ_0 penalty.

We now propose the following prior distribution for our Bayesian Sparse and Reduced-rank Regression (BSRR) method,

$$\begin{aligned}\pi^{\mathrm{BSRR}}(\mathbf{A},\mathbf{B}|\boldsymbol{\lambda}) \propto \exp\left(-\tfrac{1}{2}\lambda_1\sum_{k=1}^{r}\frac{\tilde{\mathbf{a}}_k^\top\tilde{\mathbf{a}}_k+\tilde{\mathbf{b}}_k^\top\tilde{\mathbf{b}}_k}{\omega_0+\tilde{\mathbf{a}}_k^\top\tilde{\mathbf{a}}_k+\tilde{\mathbf{b}}_k^\top\tilde{\mathbf{b}}_k}\right)\\ \times\exp\left(-\tfrac{1}{2}\lambda_2\sum_{j=1}^{p}\frac{\mathbf{a}_j^\top\mathbf{a}_j}{\omega_0+\mathbf{a}_j^\top\mathbf{a}_j}\right)\\ \times\exp\left(-\tfrac{1}{2}\lambda_3\sum_{\ell=1}^{q}\frac{\mathbf{b}_\ell^\top\mathbf{b}_\ell}{\omega_0+\mathbf{b}_\ell^\top\mathbf{b}_\ell}\right),\end{aligned} \qquad (12)$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ and $\lambda_i, \geq 0$ (i = 1, 2, 3). The BSRR posterior is then given as

$$\pi^{\mathrm{BSRR}}(\mathbf{A},\mathbf{B}|\boldsymbol{Y},\boldsymbol{\lambda}) \propto f(\boldsymbol{Y}|\mathbf{C}{=}\mathbf{A}\mathbf{B}^\top)\pi^{\mathrm{BSRR}}(\mathbf{A},\mathbf{B}|\boldsymbol{\lambda}), \qquad (13)$$

and the MAP estimate $(\hat{\mathbf{A}}_{\mathrm{map}}, \hat{\mathbf{B}}_{\mathrm{map}})$ is defined as

$$(\hat{\mathbf{A}}_{\mathrm{map}},\hat{\mathbf{B}}_{\mathrm{map}}) = \arg\max_{\mathbf{A},\mathbf{B}}\{\pi(\mathbf{A},\mathbf{B}|\boldsymbol{Y},\boldsymbol{\lambda})\},$$

Since $\mathbf{C} = \mathbf{A}\mathbf{B}^\top$, the MAP estimator for **C**, named BSRR estimator, is given by $\hat{\mathbf{C}}_{\mathrm{BSRR}} = \hat{\mathbf{A}}_{\mathrm{map}}(\hat{\mathbf{B}}_{\mathrm{map}})^\top$.

The following lemma tells us the BSRR model can be expressed as a hierarchical Bayesian model by introducing auxiliary variables.

Lemma 2. Define $\mathbf{d} = (d_{1,1}, \ldots, d_{1,r}, d_{2,1},\ldots, d_{2,p}, d_{3,1},\ldots, d_{3,q})$. Let $\boldsymbol{\pi}(\mathbf{A}, \mathbf{B}, \mathbf{d} \mid \boldsymbol{\lambda})$ be a density function of (**A**, **B**, **d**) such that

$$\begin{aligned}\pi(\mathbf{A},\mathbf{B},\mathbf{d}|\boldsymbol{\lambda}) \propto \pi(\mathbf{A},\mathbf{B}|\mathbf{d})\pi(\mathbf{d}|\boldsymbol{\lambda})\\ \propto \exp\left[-\tfrac{1}{2}\left\{\sum_{k=1}^{r}d_{1,k}(\tilde{\mathbf{a}}_k^\top\tilde{\mathbf{a}}_k+\tilde{\mathbf{b}}_k^\top\tilde{\mathbf{b}}_k)\right\}\right]\\ \times\exp\left[-\tfrac{1}{2}\left\{\sum_{j=1}^{p}d_{2,j}(\mathbf{a}_j^\top\mathbf{a}_j)+\sum_{\ell=1}^{q}d_{3,\ell}(\mathbf{b}_\ell^\top\mathbf{b}_\ell)\right\}\right]\\ \times\prod_{i=1}^{3}\left[\prod_{j=1}^{m_i}\left\{(d_{i,j})^{\frac{\lambda_i}{2}}\exp\left(-\tfrac{\omega_0}{2}d_{i,j}\right)\right\}\right],\end{aligned} \qquad (14)$$

where $m_1 = r$, $m_2 = p$ and $m_3 = q$. Let $\boldsymbol{\pi}^{\mathrm{BSRR}}(\mathbf{A}, \mathbf{B} \mid \boldsymbol{\lambda})$ denote the *BSRR* prior defined in (12). Then, for any positive $\boldsymbol{\lambda}$ and ω_0, we have that

$$\pi^{\mathrm{BSRR}}(\mathbf{A},\mathbf{B}|\boldsymbol{\lambda}) \propto \max_{\mathrm{d}}\{\pi(\mathbf{A},\mathbf{B},\mathbf{d}|\boldsymbol{\lambda})\}.$$

The proof of Lemma 2 can be shown by differentiating Eq. (14) with respect to $d_{i,j}$'s, letting them to be zero, and finding the solutions of the equations. Based on Lemma 2, we introduce the following hierarchical Bayesian representation of the sparse reduced-rank regression model (HBSRR),

$$f(\boldsymbol{Y}|\mathbf{A},\mathbf{B}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{Y}-\mathbf{XAB}^{\top}\|_F^2\right), \quad (15)$$

$$\pi(\mathbf{A},\mathbf{B}|\mathbf{d}) \propto \exp\left\{-\frac{1}{2}\left(\|\mathbf{AD}_1^{1/2}\|_F^2+\|\mathbf{D}_2^{1/2}\mathbf{A}\|_F^2\right)\right\}\times\exp\left\{-\frac{1}{2}\left(\|\mathbf{BD}_1^{1/2}\|_F^2+\|\mathbf{D}_3^{1/2}\mathbf{B}\|_F^2\right)\right\}, \quad (16)$$

$$\begin{gathered}\mathbf{D}_i^{1/2}=\mathrm{diag}(\sqrt{d_{i,1}},\ldots,\sqrt{d_{i,m_i}}), i=1,2,3,\\ \pi(\mathbf{d}|\boldsymbol{\lambda})\alpha\prod_{i=1}^{3}\left[\prod_{j=1}^{m_i}\left\{(d_{i,j})^{\frac{\lambda_i}{2}}\exp\left(-\frac{\omega_0}{2}d_{i,j}\right)\right\}\right],\end{gathered} \quad (17)$$

where $m_1 = r$, $m_2 = p$, $m_3 = q$, $\mathbf{d} = \{\mathrm{diag}(\mathbf{D}_1), \mathrm{diag}(\mathbf{D}_2), \mathrm{diag}(\mathbf{D}_3)\}$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$, and $d_{i,m_i} > 0$ ($i = 1, 2, 3$). Let $(\hat{\mathbf{A}}_{\mathrm{mode}}, \hat{\mathbf{B}}_{\mathrm{mode}}, \hat{\mathbf{d}}_{\mathrm{mode}})$ be the mode of the induced posterior $\pi(\mathbf{A}, \mathbf{B}, \mathbf{d} \mid \mathbf{Y}, \boldsymbol{\lambda})$ from the above hierarchical model. Recall that $\hat{\mathbf{C}}_{\mathrm{BSRR}}$indicates the BSRR estimator. Then, using Lemma 2, it is straightforward to show that $\hat{\mathbf{C}}_{\mathrm{BSRR}} = \hat{\mathbf{A}}_{\mathrm{mode}} (\hat{\mathbf{B}}_{\mathrm{mode}})^{\top}$, almost surely. This enables us to easily find $\hat{\mathbf{C}}_{\mathrm{BSRR}}$ using the HBSRR. Since all full conditional distributions of the HBSRR are well-known distributions such as Gaussian and gamma, the estimation procedure can be conducted by standard Bayesian estimation algorithms.

4. Bayesian analysis

Since our posterior distribution is complex, the Bayesian inference procedure requires the implementation of iterated conditional modes (ICM) algorithm [6] or Markov chain Monte Carlo (MCMC) sampling techniques, to obtain Bayes estimators such as posterior mode, posterior mean, or credible set. We derive full conditional distributions from the joint posterior of HBSRR. Then, we describe the implementation of ICM and MCMC, and discuss the determination of the tuning parameter from a Bayesian perspective.

4.1. Full conditionals

To derive the full conditionals of the HBSRR, we write

$$\begin{aligned}\|\boldsymbol{Y}-\mathbf{XAB}^\top\|_F^2=&\mathrm{trace}\left\{\left(\boldsymbol{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top\right)\left(\boldsymbol{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top\right)^\top\right\}\\ &+\mathrm{trace}(\tilde{\mathbf{x}}_j\mathbf{a}_j^\top\mathbf{B}^\top\mathbf{B}\mathbf{a}_j\tilde{\mathbf{x}}_j^\top)-2\mathrm{trace}\left\{\tilde{\mathbf{x}}_j\mathbf{a}_j^\top\mathbf{B}^\top\left(\boldsymbol{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top\right)^\top\right\}\\ =&\mathrm{trace}\left\{\left(\boldsymbol{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top\right)\left(\boldsymbol{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top\right)^\top\right\}\\ &+\mathbf{a}_j^\top\mathbf{B}^\top\mathbf{B}\mathbf{a}_j\tilde{\mathbf{x}}_j^\top\tilde{\mathbf{x}}_j-2\mathbf{a}_j^\top\mathbf{B}^\top\left(\boldsymbol{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top\right)^\top\tilde{\mathbf{x}}_j.\end{aligned} \qquad (18)$$

Here we use the notation $\mathbf{C}_{(j)}$ to denote the submatrix of a generic matrix $\mathbf{C}$ by deleting its j^{th} row, and $\mathbf{C}_{(\tilde{j})}$ by deleting its j^{th} column. Using (15), (16) and (18), the full conditional distribution of $\mathbf{a}_j$ (j = 1,…, p) is determined to be

$$\mathbf{a}_j|\mathrm{Others}\overset{\mathrm{ind}}{\sim}\mathcal{N}_r(\boldsymbol{\mu}_j^{\mathbf{A}},\textstyle\sum_j^{\mathbf{A}}), \qquad (19)$$

where

$$\begin{aligned}\boldsymbol{\mu}_j^{\mathbf{A}}=&\textstyle\sum_j^A\mathbf{B}^\top\left(\boldsymbol{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top\right)^\top\tilde{\mathbf{x}}_j,\\ \textstyle\sum_j^A=&(\mathbf{B}^\top\mathbf{B}\tilde{\mathbf{x}}_j^\top\tilde{\mathbf{x}}_j+\mathbf{D}_1+d_{2,j}\mathbf{I}_r)^{-1},\end{aligned}$$

with $\mathbf{I}_r$ denoting the $r\times r$ identity matrix. Similar to (18), we have

$$\|\boldsymbol{Y}-\mathbf{XAB}^\top\|_F^2=\mathrm{trace}\left[\left\{\boldsymbol{Y}_{(\tilde{\ell})}-\mathbf{XA}(\mathbf{B}^\top)_{(\tilde{\ell})}\right\}^\top\left\{\boldsymbol{Y}_{(\tilde{\ell})}-\mathbf{XA}(\mathbf{B}^\top)_{(\tilde{\ell})}\right\}\right]+\mathbf{b}_\ell^\top(\mathbf{XA})^\top\mathbf{XA}\mathbf{b}_\ell-2\mathbf{b}_\ell^\top(\mathbf{XA})^\top\tilde{\mathbf{y}}_\ell+\tilde{\mathbf{y}}_\ell^\top\tilde{\mathbf{y}}_\ell.$$

The full conditional distribution of $\mathbf{b}_\ell$ for ℓ= 1,…, q, is given by

$$\mathbf{b}_\ell|\mathrm{Others}\overset{\mathrm{ind}}{\sim}\mathcal{N}_r(\boldsymbol{\mu}_\ell^{\mathbf{B}},\textstyle\sum_\ell^{\mathbf{B}}), \qquad (20)$$

where

$$\begin{aligned}\boldsymbol{\mu}_\ell^B=&\textstyle\sum_\ell^{\mathbf{b}}(\mathbf{XA})^\top\tilde{\mathbf{y}}_\ell,\\ \textstyle\sum_\ell^B=&\{(\mathbf{XA})^\top\mathbf{XA}+\mathbf{D}_1+d_{3,\ell}\mathbf{I}_r\}^{-1}.\end{aligned}$$

From (16) and (17), it is straightforward to show that the full conditionals for elements of **D** are written as

$$d_{1,k}|\text{Others} \overset{\text{ind}}{\sim} \mathscr{G}\left(\frac{\lambda_1}{2}, \frac{\omega_0+\tilde{\mathbf{a}}_k^T\tilde{\mathbf{a}}_k+\widetilde{\mathbf{b}}_k^T\widetilde{\mathbf{b}}_k}{2}\right),$$
$$d_{2,j}|\text{Others} \overset{\text{ind}}{\sim} \mathscr{G}\left(\frac{\lambda_2}{2}, \frac{\omega_0+\mathbf{a}_j^T\mathbf{a}_j}{2}\right),$$
$$d_{3,\ell}|\text{Others} \overset{\text{ind}}{\sim} \mathscr{G}\left(\frac{\lambda_3}{2}, \frac{\omega_0+\mathbf{b}_\ell^T b_\ell}{2}\right),$$

where $k = 1,\ldots, r$, $j = 1,\ldots, p$, and $\ell = 1, \ldots, q$.

4.2. Iterated conditional modes

All full conditionals in Section 4.1 are well-known distributions (normal or gamma distribution), and the modes are thus well-known. Consequently, using the full conditionals, we construct the following ICM algorithm to find the BSRR estimate $\hat{\mathrm{C}}_{\mathrm{BSRR}}$:

```
Algorithm 1 ICM algorithm for $\hat{\mathrm{C}}_{\mathrm{BSRR}}$

Step 1) Set initial values $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{d}}) = (\mathbf{A}(^{0}), \mathbf{B}(^{0}), \mathbf{d}^{(0)})$.
Step 2) Update $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{d}}) = (\mathbf{A}^{(t+1)}, \mathbf{B}^{(t+1)}, \mathbf{d}^{(t+1)})$ by
    $\mathbf{a}_j^{(t+1)} \leftarrow \{(\mathbf{B}^{(t)})^\top\mathbf{B}^{(t)}\tilde{\mathbf{x}}_j^\top\tilde{\mathbf{x}}_j+\mathbf{D}_1^{(t)}+d_{2,j}^{(t)}\mathbf{I}_r\}^{-1}$
        $\times(\mathbf{B}^{(t)})^\top\{\boldsymbol{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}^{(t)}(\mathbf{B}^{(t)})^\top\}^\top\tilde{\mathbf{x}}_j,$
    $\mathbf{b}_\ell^{(t+1)} \leftarrow \left\{(\mathbf{XA}^{(t+1)})^\top\mathbf{XA}^{(t+1)}+\mathbf{D}_1^{(t)}+d_{3,\ell}^{(t)}\mathbf{I}_r\right\}^{-1}(\mathbf{XA}^{(t+1)})^\top\tilde{\mathbf{y}}_\ell,$
    $d_{1,k}^{(t+1)} \leftarrow \lambda_1\left\{\omega_0+(\tilde{\mathbf{a}}_k^{(t+1)})^\top\tilde{\mathbf{a}}_k^{(t+1)}+(\widetilde{\mathbf{b}}_k^{(t+1)})^\top\widetilde{\mathbf{b}}_k^{(t+1)}\right\}^{-1},$
    $d_{2,j}^{(t+1)} \leftarrow \lambda_2\left\{\omega_0+(\mathbf{a}_j^{(t+1)})^\top\mathbf{a}_j^{(t+1)}\right\}^{-1},$
    $d_{3,\ell}^{(t+1)} \leftarrow \lambda_3\left\{\omega_0+(\mathbf{b}_\ell^{(t+1)})^\top\mathbf{b}_\ell^{(t+1)}\right\}^{-1}$
    for $k = 1, \ldots, r$; $j = 1, \ldots, p$; $\ell = 1, \ldots, q$.
Step 3) Repeat step 2 until convergence.
Step 4) Return $\hat{\mathrm{C}}_{\mathrm{BSRR}} = \hat{\mathbf{A}}\hat{\mathbf{B}}^\top$.
```

To set an initial value $(\mathbf{A}^{(0)}, \mathbf{B}^{(0)}, \mathbf{d}^{(0)})$, we propose to utilize the OLS estimate $\hat{\mathrm{C}}_{\mathrm{ols}} = (\mathbf{X}^\top\mathbf{X})^{-}\mathbf{X}^\top\mathbf{Y}$. Let $\hat{\mathrm{C}}_{\mathrm{ols}} = \mathbf{USV}^\top$ be the sigular value decomposition of $\hat{\mathrm{C}}_{\mathrm{ols}}$. Then, $(\mathbf{A}^{(0)}, \mathbf{B}^{(0)}, \mathbf{d}^{(0)})$ can be defined as

$$\begin{gathered}
\mathbf{A}^{(0)}=\mathbf{U}\mathbf{S}^{1/2}, \mathbf{B}^{(0)}=\mathbf{V}\mathbf{S}^{1/2},\\
d_{1,k}^{(0)}=\lambda_1\left\{\omega_0+(\tilde{\mathbf{a}}_k^{(0)})^\top\tilde{\mathbf{a}}_k^{(0)}+(\tilde{\mathbf{b}}_k^{(0)})^\top\tilde{\mathbf{b}}_k^{(0)}\right\}^{-1},\\
d_{2,j}^{(0)}=\lambda_2\left\{\omega_0+(\mathbf{a}_j^{(0)})^\top\mathbf{a}_j^{(0)}\right\}^{-1}\\
d_{3,\ell}^{(0)}=\lambda_3\left\{\omega_0+(\mathbf{b}_\ell^{(0)})^\top\mathbf{b}_\ell^{(0)}\right\}^{-1},
\end{gathered}$$

for $k = 1, \ldots, r$, $j = 1, \ldots, p$ and $\ell = 1, \ldots, q$.

4.3. Posterior sampling

Using the HBSRR, we introduce indirect sampling method to obtain the posterior samples for **C**, so that they can be used to construct the credible set for the BSRR estimate $\hat{\mathbf{C}}_{\mathrm{BSRR}}$. Recall that $(\hat{\mathbf{A}}_{\mathrm{mode}}, \hat{\mathbf{B}}_{\mathrm{mode}}, \hat{\mathbf{d}}_{\mathrm{mode}})$ denotes the posterior mode of the HBSRR. Then,

$$\begin{aligned}
\hat{\mathbf{C}}_{\mathrm{BSRR}}&=\arg\max_{\mathbf{C}=\mathbf{A}\mathbf{B}^\top}\left\{\max_{\mathbf{d}}\pi(\mathbf{A},\mathbf{B},\mathbf{d}|\boldsymbol{Y},\lambda)\right\}\\
&=\arg\max_{\mathbf{C}=\mathbf{A}\mathbf{B}^\top}\left\{f(\boldsymbol{Y}|\mathbf{A},\mathbf{B})\pi(\mathbf{A},\mathbf{B}|\hat{\mathbf{d}}_{\mathrm{mode}})\right\}.
\end{aligned}$$

Consequently, we can obtain the posterior sample of **C** from the following posterior

$$\pi(\mathbf{A},\mathbf{B}|\boldsymbol{Y},\hat{\mathbf{d}}_{\mathrm{mode}})\propto f(\boldsymbol{Y}|\mathbf{A},\mathbf{B})\pi(\mathbf{A},\mathbf{B}|\hat{\mathbf{d}}_{\mathrm{mode}}).$$

Note that $\hat{\mathbf{d}}_{\mathrm{mode}}$ can be obtained by the proposed ICM algorithm in the previous section. First, to generate MCMC samples from the above posterior distribution of (**A**, **B**), we consider a Gibbs sampler that iterates through the following steps:

1. update $\mathbf{a}_j$ for $j = 1, \ldots, p$;
2. update $\mathbf{b}_\ell$ for $\ell = 1, \ldots, q$.

The explicit forms of full conditionals of $\mathbf{a}_j$ and $\mathbf{b}_\ell$ respectively, are given in (19) and (20). In each Gibbs step, we update $\mathbf{a}_j$ and $\mathbf{b}_\ell$ by generating samples from

$$\begin{gathered}
\mathbf{a}_j|\boldsymbol{Y},\mathbf{A}_{(j)},\mathbf{B},\hat{\mathbf{d}}_{\mathrm{mode}}\sim\mathscr{N}_r\left(\boldsymbol{\mu}_{\mathbf{A},j},\sum\nolimits_{\mathbf{A},j}\right)\\
\boldsymbol{\mu}_{\mathbf{A},j}=\sum\nolimits_{\mathbf{A},j}\mathbf{B}^\top\left(\boldsymbol{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top\right)^\top\widetilde{\mathbf{x}}_j,\\
\sum\nolimits_{\mathbf{A},j}=\left(\mathbf{B}^\top\mathbf{B}\widetilde{\mathbf{x}}_j^\top\widetilde{\mathbf{x}}_j+\hat{\mathbf{D}}_1+\hat{d}_{2,j}\mathbf{I}_r\right)^{-1}\text{for} j=1,\ldots,p,\\
\mathbf{b}_\ell|\boldsymbol{Y},\mathbf{A},\hat{\mathbf{d}}_{\mathrm{mode}}\sim\mathscr{N}_r(\boldsymbol{\mu}_{\mathbf{B},\ell},\sum\nolimits_{\mathbf{B},\ell}),\\
\boldsymbol{\mu}_{\mathbf{B},\ell}=\sum\nolimits_{\mathbf{B},\ell}(\mathbf{X}\mathbf{A})^\top\widetilde{\mathbf{y}}_\ell,\\
\sum\nolimits_{\mathbf{B},\ell}=\left\{(\boldsymbol{X}\boldsymbol{A})^\top\boldsymbol{X}\boldsymbol{A}+\hat{\mathbf{D}}_1+\hat{d}_{3,j}\mathbf{I}_r\right\}^{-1}\text{for}\ell=1,\ldots,q,
\end{gathered}$$

where $\hat{\mathbf{d}}_{\text{mode}} = \{\text{diag}(\hat{\mathbf{D}}_1), \text{diag}(\hat{\mathbf{D}}_2), \text{diag}(\hat{\mathbf{D}}_3)\}$. Let $\{\mathbf{A}^i, \mathbf{B}^i\}_{i=1}^N$ be a set of obtained MCMC samples from the above sampling procedure. Then a set of posterior samples for **C** can be obtained by $\mathscr{S} = \left\{\mathbf{C}^i : \mathbf{C}^i = \mathbf{A}^i (\mathbf{B}^i)^\top\right\}_{i=1}^N$.

4.4. Tuning parameter selection

In practice, we are usually interested in selecting a tuning parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ (or hyperparameter) from the set of candidates $\mathscr{L} = \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K\}$. We assume that there is no preferred model, i.e., $\pi(\boldsymbol{\lambda}_k) = 1/K$ for $k = 1, \ldots, K$. In general, the tuning parameter is determined via a grid search strategy from a lower bound $\boldsymbol{\lambda}_L = (0^+, 0^+, 0^+)$ to a given upper bound $\boldsymbol{\lambda}_U$ which is the smallest value to induce the marginal null model (i.e., all estimates are zero). Hence, the set $\mathscr{L}$ is well-defined. Let $m(\mathbf{Y} \mid \boldsymbol{\lambda}_k) = \int f(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}) \Pi(d\mathbf{A}, d\mathbf{B} \mid \boldsymbol{\lambda}_k)$ be a marginal likelihood for a given $\boldsymbol{\lambda}_k$. Then, we can show that $m(\mathbf{Y} \mid \boldsymbol{\lambda}_k)$ is proportional to the posterior probability of $\boldsymbol{\lambda}_k$ given **Y**, that is

$$\prod(\boldsymbol{\lambda}_k|\boldsymbol{Y}) = \frac{m(\boldsymbol{Y}|\boldsymbol{\lambda}_k)\pi(\boldsymbol{\lambda}_k)}{\sum_{k=1}^K m(\boldsymbol{Y}|\boldsymbol{\lambda}_k)\pi(\boldsymbol{\lambda}_k)} \propto m(\boldsymbol{Y}|\boldsymbol{\lambda}_k). \tag{21}$$

From the above viewpoint, we define the optimal $\boldsymbol{\lambda}_{k^*}$ such that

$$\boldsymbol{\lambda}_{k^*} = \arg\max_{\boldsymbol{\lambda}_k \varepsilon \mathscr{L}} \{m(\boldsymbol{Y}|\boldsymbol{\lambda}_k)\}.$$

Let **C** be the $p \times q$ coefficient matrix with rank(**C**) = r^*. Without loss of generality, suppose that the first p^* rows and q^* columns of **C** are non-zero and the remaining rows and columns of **C** are zero. Then, the matrix **C** can be decomposed as

$$\mathbf{C} = \begin{pmatrix} \mathbf{I}_{r^*} \\ \hline \mathbf{C}_A \\ \hline \text{O}_A \end{pmatrix} (\mathbf{C}_B | \mathbf{O}_B),$$

where $\mathbf{I}_{r^*}$. denotes the identity matrix of order r^*, $\mathbf{C}_A$ is a $(p^* - r^*) \times r^*$ nonzero matrix, $\mathbf{C}_B$ is a $r^* \times q^*$ nonzero matrix, and $\mathbf{O}_A$ is the $(p - p^*) \times$ r* zero matrix, and $\mathbf{O}_B$ is the $r^* \times (q - q^*)$ zero matrix. The key to above parameterization of **C** is that the matrix $\mathbf{C}_A$ and $\mathbf{C}_B$ are uniquely determined. It can be seen that for given (r^*, p^*, q^*), the number of free parameters in **C** is $\dim(\mathbf{C}_\text{A}) + \dim(\mathbf{C}_B) = r^*(p^* + q^* - r^*)$. Suppose that a given tuning parameter $\boldsymbol{\lambda} = \boldsymbol{\lambda}_k$ results in an estimator with (r_k, p_k, q_k). Define $\boldsymbol{\theta} = \{vec(\mathbf{C}_A)^\top, vec(\mathbf{C}_B)^\top\}^\top$ such that $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{r_k(p_k + q_k - r_k)}$, where vec(·) denotes the vectorization of a matrix. Then, the marginal likelihood can be rewritten as

$$\begin{aligned} m(\boldsymbol{Y}|\boldsymbol{\lambda}_k) &= \int_\Theta f(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_k)d\boldsymbol{\theta} \\ &= \int_\Theta \exp\{(nq_k)g_n(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\lambda}_k)\}d\boldsymbol{\theta}, \end{aligned} \quad (22)$$

where $g_n(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\lambda}) = (nq_k)^{-1} \{\ln f(\mathbf{Y} \mid \boldsymbol{\theta}) + \ln \boldsymbol{\pi} (\boldsymbol{\theta} \mid \boldsymbol{\lambda}_k)\}$. Let $\hat{\boldsymbol{\theta}}$ be the mode of $g_n(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\lambda})$. By the Laplace approximation, the marginal likelihood in (22) can be expressed as

$$m(\boldsymbol{Y}|\boldsymbol{\lambda}) = \frac{(2\pi/nq_k)^{\frac{r_k(p_k+q_k-r_k)}{2}}}{\left|-G_n(\hat{\boldsymbol{\theta}})\right|^{1/2}} \exp\{(nq_k)g_n(\hat{\boldsymbol{\theta}}|\boldsymbol{Y},\boldsymbol{\lambda}_k)\}\{1+O_p(n^{-1})\}, \quad (23)$$

where

$$G_n(\hat{\boldsymbol{\theta}}) = \frac{\partial^2 g_n(\theta|\boldsymbol{Y},\boldsymbol{\lambda})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

By taking the logarithm of the formula (23) and ignoring the term of $O(1)$ and higher order terms, we have the following approximation of log marginal likelihood

$$\ln\{m(\boldsymbol{Y}|\boldsymbol{\lambda}_k)\} \approx \ln f(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}) - \frac{r_k(p_k+q_k-r_k)}{2}\ln(nq_k). \quad (24)$$

Let $\hat{\mathbf{C}}_k$ be the BRRR estimate for given $\boldsymbol{\lambda}_k$. Then, by substituting it in (24) and multiplying by −2, (24) reduces to the BRRR version of Bayesian information criterion [35],

$$\mathrm{BIC}(\boldsymbol{\lambda}_k) = -2\ln f(\boldsymbol{Y}|\hat{\mathbf{C}}_k) + r_k(p_k+q_k-r_k)\ln(nq_k). \quad (25)$$

According (21), we know that minimizing the BIC corresponds to maximizing the posterior probability of $\boldsymbol{\lambda}_k$ given $\mathbf{Y}$. Hence, we regard the tuning parameter $\boldsymbol{\lambda}_*$ as the optimum if $\boldsymbol{\lambda}_* = \mathrm{argmin}_{\boldsymbol{\lambda}_k \in \mathcal{L}}\ \mathrm{BIC}(\boldsymbol{\lambda}_k)$.

5. Posterior consistency

In Bayesian analysis, the posterior consistency assures that the posterior converges to point mass at the true parameter as more data are collected [13, 15, 21]. Here, we discuss the posterior consistency for the proposed BSRR method, following Armagan et al. [3]. We allow the number of predictors p to grow with sample size n, and the number of true non-zero coefficients p^* is assumed to be finite. Henceforth, we denote p as p_n. Similarly the response matrix $\mathbf{Y}$ and predictor matrix $\mathbf{X}$ are denoted by $\mathbf{Y}_n$ and $\mathbf{X}_n$, respectively. Unlike p_n, the number of response variables q is assumed to be fixed in our analysis.

Suppose that, given $\mathbf{X}_n$ and $\mathbf{C}^*$, $\mathbf{Y}_n$ is generated from

$$\boldsymbol{Y}_n = \mathbf{X}_n \mathbf{C}^* + E,$$

where $\mathbf{e}_i \overset{\text{i.i.d.}}{\sim} \mathscr{N}_q(\mathbf{0}, \sum)$ with a positive definite matrix $\boldsymbol{\Sigma}$ (assumed to be known) and $\mathbf{C}^*$ is a ($p_n \times q$) matrix such that card $(\{j : \mathbf{c}^{*\top}_j \mathbf{c}^*{}_j \neq 0\}) = p^*$, card $(\{\ell : \tilde{\mathbf{c}}_\ell^{*\top} \tilde{\mathbf{c}}_\ell^* \neq 0\}) = q^*$ and rank($\mathbf{C}^*$) = r^*. Further, we make the following assumptions.

- **I.** $p_n = o(n)$, but $p^* < \infty$ and $q^* \le q < \infty$.
- **II.** $0 < S_{\min} < \liminf_{n \to \infty} S_{n,\min}/\sqrt{n} \le \limsup_{n \to \infty} S_{n,\max}/\sqrt{n} < S_{\max} < \infty$, where $S_{n,\min}$ and $S_{n,\max}$ denote the smallest and the largest singular values of $\mathbf{X}$, respectively.
- **III.** $\mathrm{Sup}_{(j,\ell)}(c^*_{j\ell}) < \infty$, where $c^*_{j\ell}$ indicates the $(j, \ell)^{th}$ element of $\mathbf{C}^*$.

Our main results are presented in Theorems 3 and 4 below.

Theorem 3. Under assumptions I and II, if the prior $\Pi(\mathbf{A}, \mathbf{B})$ satisfies the following condition:

$$\prod \left\{ (\mathbf{A}, \mathbf{B}) : \|\mathbf{A}\mathbf{B}^\top - \mathbf{C}^*\|_F < \frac{\Delta}{n^{\frac{\rho}{2}}} \right\} > \exp(-dn),$$

for all $0 < \Delta < \varepsilon^2 S_{\min}^2 / (48 S_{\max}^2)$ and $0 < d < \varepsilon S_{\min}^2 / (32 \tau_{\max}) - 3 \Delta S_{\max} / (2 \tau_{\min})$ and some $\rho > 0$, where $\tau_{\min}$ and $\tau_{\max}$ denote, respectively, the smallest and the largest eigenvalue of Σ, then the posterior of ($\mathbf{A}$, $\mathbf{B}$) induced by the prior $\Pi(\mathbf{A}, \mathbf{B})$ is strongly consistent, i.e., for any $\varepsilon > 0$,

$$\prod \{ (\mathbf{A}, \mathbf{B}) : \|\mathbf{C} - \mathbf{C}^*\|_F > \varepsilon, \mathbf{C} = \mathbf{A}\mathbf{B}^\top | \boldsymbol{Y}_n \} \to 0 \text{almost surely},$$

as $n \to \infty$.

Theorem 4. Under assumptions I, II and III, the prior defined in (12) yields a strongly consistent posterior if $\lambda_i = \delta_i n^{\rho/2} \sqrt{p_n} \ln n$ for finite $\delta_i > 0$, $i = 1, 2, 3$.

In Theorem 3, we establish a sufficient condition on a prior distribution in order to achieve posterior consistency. Theorem 4 then shows that our BSRR prior in (12) satisfies the sufficient condition in Theorem 3, and consequently, our BSRR method possesses the desirable posterior consistency property. The proofs of both theorems are shown in the supplementary materials.

6. Simulation studies

To examine the performance of our BSRR method, we conduct Monte Carlo experiments under several possible scenarios. For purposes of comparison, we also consider the following two reduced priors:

$$\pi^{\mathrm{RR}}(\mathbf{A},\mathbf{B}|\lambda_1,\lambda_2):=\pi^{\mathrm{BSRR}}(\mathbf{A},\mathbf{B}|\boldsymbol{\lambda},\lambda_3=0), \quad (26)$$

$$\pi^{\mathrm{RC}}(\mathbf{A},\mathbf{B}|\lambda_2,\lambda_3):=\pi^{\mathrm{BSRR}}(\mathbf{A},\mathbf{B}|\boldsymbol{\lambda},\lambda_1=0). \quad (27)$$

We denote the Bayesian methods using (26) and (27) as RR (Row-wise-sparse and Reduced-rank) method and RC (Row-and-Column-wise sparse) method, respectively. Our BSRR method aims to recover all the low-dimensional structures in A1–A3, but RR and RC methods, respectively, do not consider the column-wise sparsity of **C** in A3 and the reduced rank structure of **C** in A1. Therefore, the RR method is analogous to the joint rank and predictor selection methods proposed by Chen and Huang [12] and Bunea et al. [9]. The RR and RC methods can be derived from BSRR method with setting $\lambda_3 = 0$ and $\lambda_1 = 0$ in (13), respectively. Hence, the BSRR estimate $\hat{\mathrm{C}}_{\mathrm{BSRR}}$ as well as RR and RC estimates, $\hat{\mathrm{C}}_{\mathrm{RR}}$ and $\hat{\mathrm{C}}_{\mathrm{RC}}$, are obtained by the proposed algorithm in Section 4.2. Similarly, the unknown tuning parameter λ for each model is estimated by the proposed BIC in (25).

We generate data from the multivariate regression model $\mathbf{Y} = \mathbf{XC} + \mathbf{E}$. For the $n \times p$ design matrix $\mathbf{X}$, its n rows are independently generated from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_{ij})_{p\times p}$ with $\Gamma_{ij} = (0.5)^{|i-j|}$. The $p \times q$ coefficient matrix **C** is defined as $\mathbf{C}=\sum_{k=1}^{r^*} s_k\mathbf{C}_k$, where $s_k = 5 + (k - 1)$ȳ$15/r^*$ɉ; the entries of $\boldsymbol{C}_k$ are all zero expect in its upper left $p^* \times q^*$ submatrix, which is generated by $\mathbf{z}_1\mathbf{z}_2^\top/(\|\mathbf{z}_1\|_2\|\mathbf{z}_2\|_2)$, where $\mathbf{z}_1 \in \mathbb{R}^{p^*}$, $\mathbf{z}_2 \in \mathbb{R}^{q^*}$, and all their entries are i.i.d samples from uniform($[-1, -0.3] \cup [0.3,1]$). The rows of the noise matrix **E** are independently generated from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_e)$, where $\boldsymbol{\Sigma}_e = (\Sigma_{ij})_{q\times q}$ with $\Sigma_{ij} = \boldsymbol{\sigma}^2(0.5)^{|i-j|}$ and $\boldsymbol{\sigma}^2$ is chosen according to the signal to noise ratio(SNR) defined by $s_{r^*}(\mathbf{XC})/s_1(\mathbf{P}_X\mathbf{E})$ with $\mathbf{P}_X = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top$.

In the first scenario, we generate models of moderate dimensions (i.e., p, $q < n$) in three setups:

(a1) $p = q = 25$, $n = 50$, $r^* = 3$, $p^* = 10$, $q^* = 10$. This setup favors our BSRR method.

(a2) $p = q = 25$, $n = 50$, $r^* = 3$, $p^* = 10$, $q^* = 25$. As all the responses are revelent in the model, this setup favors the RR method.

(a3) $p = q = 25$, $n = 50$, $r^* = 10$, $p^* = 10$, $q^* = 10$. This favors the RC method, which does not enforce rank reduction.

In the second scenario, we generate high-dimensional data (i.e., p, $q > n$) using similar settings as above, (b1) $p = 200$, $q = 170$, $n = 50$, $r^* = 3$, $p^* = 10$, $q^* = 10$.

(b2) $p = 200$, $q = 170$, $n = 50$, $r^* = 3$, $p^* = 10$, $q^* = 170$.

(b3) $p = 200$, $q = 170$, $n = 50$, $r^* = 10$, $p^* = 10$, $q^* = 10$.

The estimation accuracy is measured by the following three mean squared errors (MSEs):

$$\begin{aligned}\mathrm{MSE}_{\mathrm{est}} &= 100\|\hat{\mathbf{C}} - \mathbf{C}\|_F^2/(\mathscr{P}q),\\ \mathrm{MSE}_{\mathrm{pred}} &= 100\|\mathbf{X}\hat{\mathbf{C}} - \mathbf{X}\mathbf{C}\|_F^2/(nq),\\ \mathrm{MSE}_{\mathrm{dim}} &= 100\|\mathrm{s}(\hat{\mathbf{C}}) - \mathbf{s}(\mathbf{C})\|^2/\min(p,q),\end{aligned}$$

where **s**(**C**) denotes the vector of singular values for a matrix **C**. To assess the variable selection performance, we use false positive rate (FPR) and false negative rate (FNR) such that FPR% = 100FP/(TN + FP) and FNR% = 100FN/(TP + FN), where TP, FP, TN and FN denote the numbers of true nonzeros, false nonzeros, true zeros and false zeros, respectively. The rank selection performance is evaluated by the percentage of correct rank identification (CRI%). All measurements are estimated by the Monte Carlo method with 500 replications.

Tables 1 and 2 summarize the simulation results. As expected, in the cases (a1) and (b1), where rank reduction, predictor selection and response selection are all preferable, the BSRR method performs much better than the other two reduced methods. In cases (a2) and (b2), rank reduction and predictor selection are preferable while response selection is not necessary. The performance of the BSRR method is very similar to the RR method which assumes the correct model structure. In the cases (a3) and (b3), rank reduction becomes unnecessary when response and predictor selections are performed. While the BSRR method slightly underestimates the true rank ($r^* = 10$), its variable selection performance (FPR and FNR) is comparable to that of the RC method which assumes the correct model structure. The results are consistent for different SNR levels. Therefore, our BSRR approach provides a flexible and unified way for simultaneously exploring rank reduction, predictor selection and response selection.

7. Yeast cell cycle data

Transcription factors (TFs), also called sequence-specific DNA binding proteins, regulate the transcription of genes from DNA to mRNA by binding specific DNA sequences. In order to understand the regulatory mechanism, it is important to reveal the network structure between TFs and their target genes. The network structure can be formulated using the multivariate regression model in (1), where the row and column of the response matrix, respectively, correspond to genes and samples (arrays, tissue types, time points), and the design matrix includes the binding information representing the strength of interaction between TFs and the target genes. The regression coefficient matrix then describes actual transcription factor activities of TFs for genes. In practice, many TFs are not actually related to the genes and there exists dependency among the samples due to the design of experiment.

Here, we analyze an *Yeast cell cycle* data [14] using BSRR. The dataset is available in the *spls* package in R. The response matrix **Y** consists of 542 cell-cycle-regulated genes from an α factor arrest method, where mRNA levels are measured at every 7 minutes during 119 minutes, i.e., $n = 542$ and $q = 18$. The 542 × 106 predictor matrix **X** contains the binding information of the target genes for a total of 106 TFs, where Chromatin immunoprecipitation (ChIP) for the 542 genes was performed on each of these 106 TFs. In our analyses, **Y** and **X** are centered.

We apply the BSRR method to the dataset. We use the proposed BIC to choose the tuning parameter and obtain $\hat{\lambda}_1 = 5$, $\hat{\lambda}_2 = 1.5$ and $\hat{\lambda}_3 = 1.5$. As a result, 26 TFs are identified at 17 time points (105 min is eliminated) with the estimated rank $\hat{r} = 4$. Fig. 2 displays the obtained parameter estimates and 95% credible bands for randomly selected 4 TFs among the 26 TFs; see figures in the supplementary materials for all TFs. The same data set was also analyzed by the adaptive SRRR method of Chen and Huang [12]. In the adaptive SRRR, 32 TFs were identified at 18 time points with the optimal rank $\hat{r} = 4$ determined by a cross validation method. Among their selected 32 TFs, 21 TFs were also identified by our BSRR method. To compare variable selection performance between two methods, we define the following two models:

$$M^1 : \boldsymbol{Y} = \mathbf{X}_1\mathbf{C}_1 + \mathbf{E};$$
$$M^2 : \boldsymbol{Y} = \mathbf{X}_2\mathbf{C}_2 + \mathbf{E};$$

where $\mathbf{X}_1$ contains the information of the 542 genes for the 32 TFs identified by the adaptive SRRR, $\mathbf{X}_2$ contains the information of the 542 genes for the 26 TFs identified by BSRR, $\mathbf{C}_1$ is the 32 × 18 matrix with rank($\mathbf{C}_1$) = 4, and $\mathbf{C}_2$ is the 26 × 18 matrix with rank($\mathbf{C}_2$) = 4. To conduct a fair comparison, we consider the following reduced rank priors in Geweke [20] for the models M^1 and M^2, respectively:

$$\pi^1(\mathbf{A}_1, \mathbf{B}_1) \propto \exp\left\{-\frac{\tau}{2}(\|\mathbf{A}_1\|_F^2 + \|\mathbf{B}_1\|_F^2)\right\} \text{s.t.} \mathbf{C}_1 = \mathbf{A}_1\mathbf{B}_1^\top, \tag{28}$$

$$\pi^2(\mathbf{A}_2, \mathbf{B}_2) \propto \exp\left\{-\frac{\tau}{2}(\|\mathbf{A}_2\|_F^2 + \|\mathbf{B}_2\|_F^2)\right\} \text{s.t.} \mathbf{C}_2 = \mathbf{A}_2\mathbf{B}_2^\top, \tag{29}$$

where $\mathbf{A}_1$ is a 32 × 4 matrix, $\mathbf{B}_1$ is an 18 × 4 matrix, $\mathbf{A}_2$ is an 26 × 4 matrix, $\mathbf{B}_2$ is a 18 × 4 matrix and we set $\boldsymbol{\tau} = 0.0001$ to be a non-informative (flat) prior, so that the parameter estimates are determined nearly by the observations ($\mathbf{Y}_1$, $\mathbf{X}_1$) and ($\mathbf{Y}_2$, $\mathbf{X}_2$). As the Bayesian model selection criterion, using the priors in (28) and (29), we utilize the deviance information criteria (DIC) defined by

$$\begin{aligned}\mathrm{DIC}_1= & -4E_{\mathbf{A}_1,\mathbf{B}_1|\boldsymbol{Y},\mathbf{X}_1}[\ln\{f(\boldsymbol{Y}|\mathbf{X}_1,\mathbf{C}_1{=}\mathrm{A}_1\mathbf{b}_1^\top)\}]\\ & +2\ln\left\{f\left(\boldsymbol{Y}|\mathbf{X}_1,\mathbf{C}_1{=}\overline{\mathbf{A}_1\mathbf{B}_1^\top}\right)\right\},\\ \mathrm{DIC}_2= & -4E_{\mathbf{A}_2,\mathbf{B}_2|\boldsymbol{Y},\mathbf{X}_2}[\ln\{f(\boldsymbol{Y}|\mathbf{X}_2,\mathbf{C}_2{=}\mathbf{A}_2\mathbf{B}_2^\top)\}]\\ & +2\ln\left\{f\left(\boldsymbol{Y}|\mathbf{X}_2,\mathbf{C}_2{=}\overline{\mathbf{A}_2\mathbf{B}_2^\top}\right)\right\},\end{aligned}$$

where $\overline{\mathbf{AB}^\top}$ denotes the posterior mean. If DIC$_1$ > DIC$_2$, then it implies that the model $\mathcal{M}^2$ is more strongly supported by the given data than the model $\mathcal{M}^1$. Let $\{\mathbf{A}_1^i,\mathbf{B}_1^i\}_{i=1}^N$ and $\{\mathbf{A}_2^i,\mathbf{B}_2^i\}_{i=1}^N$ be MCMC samples from the posteriors $\pi^1(\mathbf{A}_1, \mathbf{B}_1 \mid \mathbf{Y}, \mathbf{X}_1)$ and $\pi^2(\mathbf{A}_2, \mathbf{B}_2 \mid \mathbf{Y}, \mathbf{X}_2)$, respectively. Note that the MCMC samples can be easily generated from multivariate normal distributions by using the Gibbs sampler. Define $\{\mathbf{C}_m^i{:}\mathbf{C}_m^i{=}\mathbf{A}_m^i(\mathbf{B}_m^i)^\top\}_{i=1}^N$, for m = 1, 2. Then the DIC can be estimated by the following Monte Carlo estimator:

$$\begin{aligned}\widehat{\mathrm{DIC}}_1{=}-4\left\{\frac{1}{N}\sum_{i=1}^N\ln f(\boldsymbol{Y}|\mathbf{X}_1,\mathbf{C}_1^i)\right\}+2\ln f\left(\boldsymbol{Y}|\mathbf{X}_1,\frac{1}{N}\sum_{i=1}^N\mathbf{C}_1^i\right),\\ \widehat{\mathrm{DIC}}_2{=}-4\left\{\frac{1}{N}\sum_{i=1}^N\ln f(\boldsymbol{Y}|\mathbf{X}_2,\mathbf{C}_2^i)\right\}+2\ln f\left(\boldsymbol{Y}|\mathbf{X}_2,\frac{1}{N}\sum_{i=1}^N\mathbf{C}_2^i\right).\end{aligned}$$

Based on 1,000 MCMC samples (after 1,000 burn-in iterations) with 100 replication, we obtain $\widehat{\mathrm{DIC}}_1{=}19824.46$ and $\widehat{\mathrm{DIC}}_2{=}19784.03$ with Monte Carlo errors 1.29 and 1.14, respectively. Since $\widehat{\mathrm{DIC}}_1{>}\widehat{\mathrm{DIC}}_2$, this result supports the model $\mathcal{M}^2$. Consequently, this implies that our BSRR method has better variable selection performance than the adaptive SRRR for Yeast cell cycle data. Recall that the response at 105 min was eliminated in the BSRR method. Table 3 displays the parameter estimates and 95% credible intervals (CIs) at 105 min from the model $\mathcal{M}^2$. Since all CIs include zero, this demonstrates that the response elimination at 105 min in the BSRR is valid. In other words, none of TFs activates at 105 min.

8. Discussion

We have developed a Bayesian sparse and low rank regression method, which achieves simultaneous rank reduction and predictor/response selection. There are many directions for future research. We have mainly focused on the ℓ_0 type sparsity-inducing penalties to construct prior distribution. The method can be extended to use other forms of penalties for inducing diverse lower-dimensional structures. The low-rank structure induces dependency among the response variables, and hence the error correlation structure is not explicitly considered in the current work. Incorporating the variance component into our model might improve the efficiency of the coefficient estimation. In a Bayesian framework, this can be accomplished by assigning an appropriate prior on the variance component; the choice of the prior should be carefully treated due to the lack of unimodality of the posterior [31]. In practice, the response variables could be binary or counts. It is thus pressing to utilize a general likelihood function with the proposed BSRR prior. The proposed ICM algorithm converges relatively fast, and in each iteration the main cost is to inverse a matrix of dimension min(n, p) owning to Woodbury matrix identity. However, this approach would still be inefficient when both p and n are extremely large. One way is to conduct some pre-screening procedure [17, 19] before implementation of the proposed method. It would also be interesting to study on-line learning and the divide-and-conquer strategies of the proposed model. We have established the posterior consistency of the proposed sparse and low-rank estimation method under a high-dimensional asymptotic regime, which characterizes the behavior of the posterior distribution when the number of predictors p_n increases with the sample size n. The theoretical analysis of the Bayesian (point) estimator itself could be of interest rather than the entire posterior distribution [1].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to the Associate Editor and two referees for their helpful comments on an earlier version of this paper. Chen's research was partially supported by U.S. National Institutes of Health grant U01-HL114494 and U.S. National Science Foundation grant DMS-1613295.

References

1. Alquier, P. Bayesian Methods for Low-Rank Matrix Estimation: Short Survey and Theoretical Study. Springer; Berlin Heidelberg: 2013. p. 309-323.
2. Anderson TW. Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions. The Annals of Mathematical Statistics. 1951; 22:327–351.
3. Armagan A, Dunson DB, Lee J, Bajwa WU, Strawn N. Posterior consistency in linear models under shrinkage priors. Biometrika. 2013; 100:1011–1018.
4. Babacan SD, Luessi M, Molina R, Katsaggelos AK. Low-rank matrix completion by variational sparse Bayesian learning. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2011:2188–2191.
5. Bahadori MT, Zheng Z, Liu Y, Lv J. Scalable Interpretable Multi-Response Regression via SEED. arXiv:1608.03686. 2016
6. Besag J. On the Statistical Analysis of Dirty Pictures. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1986; 48:259–302.
7. Breheny P, Huang J. Penalized methods for bi-level variable selection. Statistics and Its Interface. 2009; 2:369–380. [PubMed: 20640242]
8. Bunea F, She Y, Wegkamp MH. Optimal selection of reduced rank estimators of high-dimensional matrices. The Annals of Statistics. 2011; 39:1282–1309.
9. Bunea F, She Y, Wegkamp MH. Joint variable and rank selection for parsimonious estimation of high dimensional matrices. The Annals of Statistics. 2012; 40:2359–2388.
10. Chen K, Chan KS, Stenseth NC. Reduced rank stochastic regression with a sparse singular value decomposition. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2012; 74:203–221.
11. Chen K, Dong H, Chan KS. Reduced rank regression via adaptive nuclear norm penalization. Biometrika. 2013; 100:901–920. [PubMed: 25045172]
12. Chen L, Huang JZ. Sparse Reduced-Rank Regression for Simultaneous Dimension Reduction and Variable Selection. Journal of the American Statistical Association. 2012; 107:1533–1545.
13. Choi T, Ramamoorthi RV. Remarks on consistency of posterior distributions. Institute of Mathematical Statistics Collections. 2008; 3:170–186.

14. Chun H, Keles S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010; 72:3–25. [PubMed: 20107611]

15. Diaconis P, Freedman D. On the Consistency of Bayes Estimates. The Annals of Statistics. 1986; 14:1–26.

16. Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

17. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008; 70:849–911. [PubMed: 19603084]

18. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Statistica Sinica. 2010; 20:101–148. [PubMed: 21572976]

19. Fan J, Song R. Sure independence screening in generalized linear models with NP–dimensionality. The Annals of Statistics. 2010; 38:3567–3604.

20. Geweke J. Bayesian reduced rank regression in econometrics. Journal of Econometrics. 1996; 75:121–146.

21. Ghosh, JK., Delampady, M., Samanta, T. An Introduction to Bayesian Analysis: Theory and Methods. Springer-Verlag; New York: 2006.

22. Huang J, Breheny P, Ma S. A Selective Review of Group Selection in High Dimensional Models. Statistical Science. 2012; 27(4):481–499.

23. Huang J, Horowitz JL, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. The Annals of Statistics. 2008; 36:587–613.

24. Izenman AJ. Reduced-rank regression for the multivariate linear model. Journal of Multivariate Analysis. 1975; 5:248–264.

25. Kyung M, Gilly J, Ghosh M, Casella G. Penalized regression, standard errors, and Bayesian lassos. Bayesian Analysis. 2010; 5:369–412.

26. Lim YJ, Teh YW. Variational Bayesian approach to movie rating prediction. Proceedings of KDD Cup and Workshop. 2007

27. Ma Z, Ma Z, Sun T. Adaptive Estimation in Two-way Sparse Reduced-rank Regression. arXiv: 1403.1922. 2014

28. Mukherjee A, Chen K, Wang N, Zhu J. On the degrees of freedom of reduced-rank estimators in multivariate regression. Biometrika. 2015; 102:457–477. [PubMed: 26702155]

29. Mukherjee A, Zhu J. Reduced rank ridge regression and its kernel extensions. Statistical Analysis and Data Mining. 2011; 4:612–622. [PubMed: 22993641]

30. Negahban S, Wainwright MJ. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. The Annals of Statistics. 2011; 39:1069–1097.

31. Park T, Casella G. The Bayesian Lasso. Journal of the American Statistical Association. 2008; 103:681–686.

32. Reinsel, GC., Velu, PP. Multivariate reduced-rank regression: Theory and Applications. Springer-Verlag; New York: 1998.

33. Rohde A, Tsybakov AB. Estimation of high-dimensional low-rank matrices. The Annals of Statistics. 2011; 39:887–930.

34. Salakhutdinov, R., Mnih, A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In: Cohen, WW.Mccallum, A., Roweis, ST., editors. Proceedings of the 25th International Conference on Machine Learning (ICML-08). 2008. p. 880-887.

35. Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics. 1978; 6:461–464.

36. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1996; 58:267–288.

37. Yuan M, Ekici A, Lu Z, Monteiro R. Dimension reduction and coefficient estimation in multivariate linear regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2007; 69:329–346.

38. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006; 68:49–67.

39. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics. 2010; 38:894–942.

40. Zhou M, Wang C, Chen M, Paisley J, Dunson D, Carin L. Nonpara-metric Bayesian matrix completion. 2010 IEEE Sensor Array and Multichannel Signal Processing Workshop. 2010:213–216.

41. Zhu H, Khondker Z, Lu Z, Ibrahim JG. Bayesian Generalized Low Rank Regression Models for Neuroimaging Phenotypes and Genetic Markers. Journal of the American Statistical Association. 2014; 109:997–990. [PubMed: 25349462]

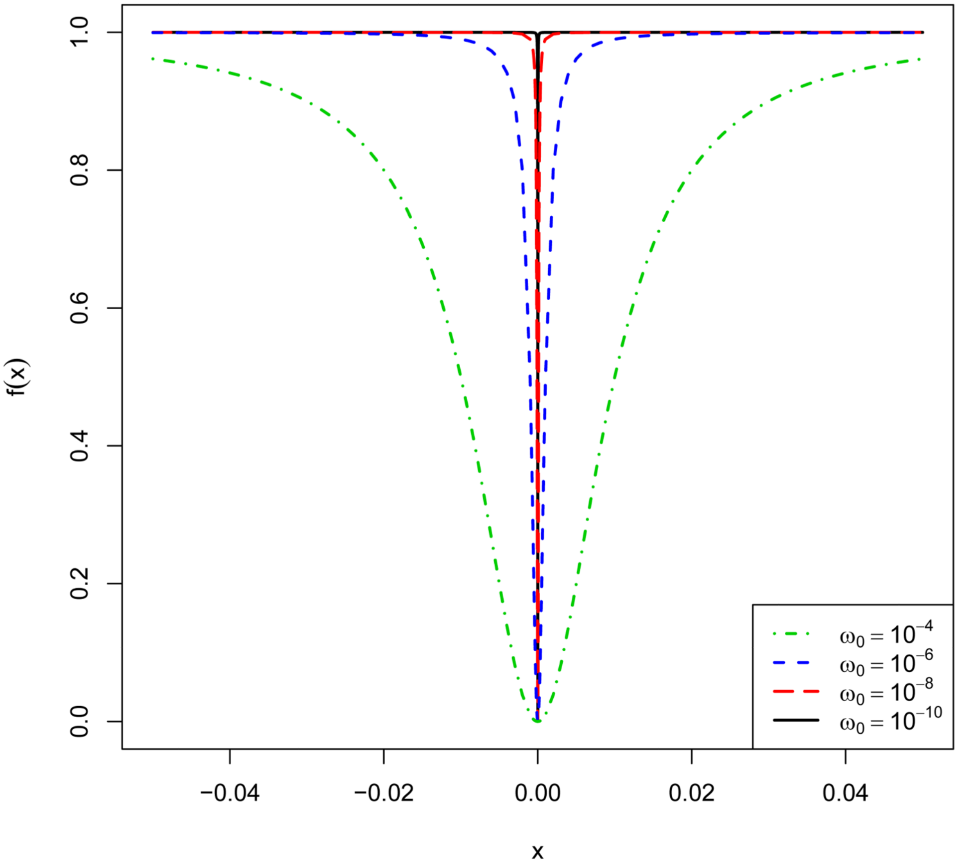
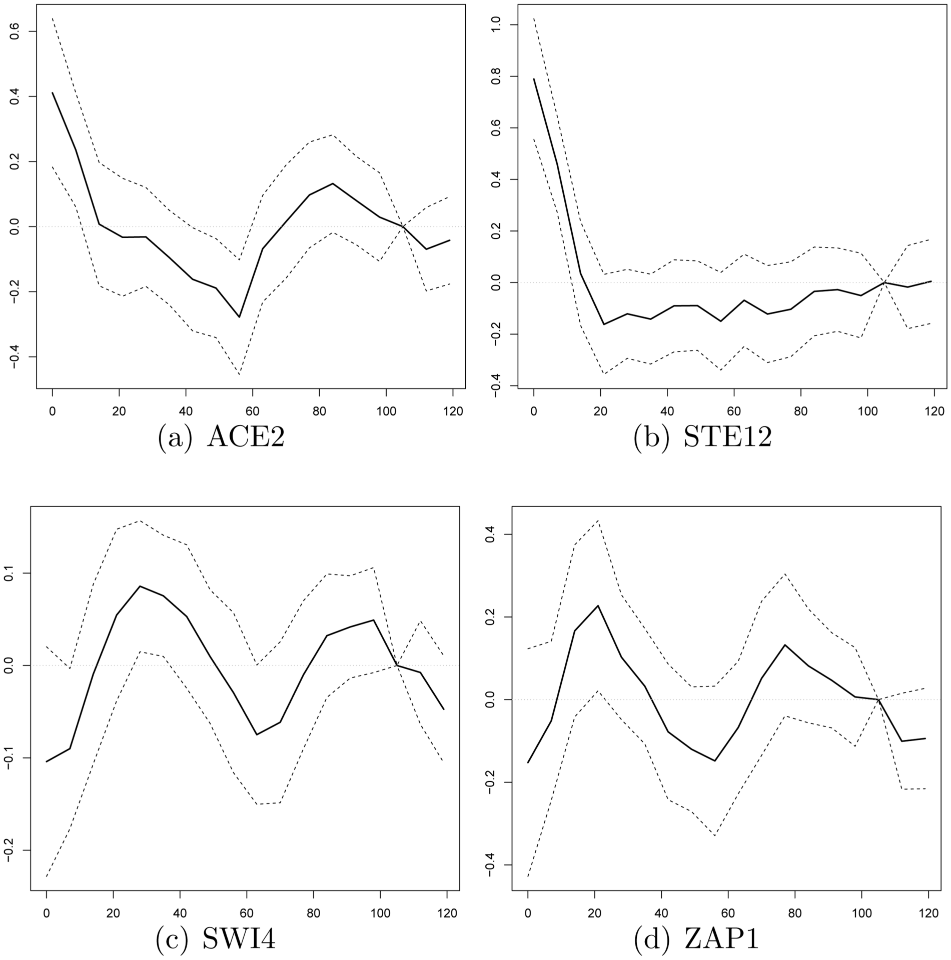


Figure 1.

Plot of $f(x) = x^2/(x^2 + \omega_0)$ for $\omega_0 = 10^{-k}$ ($k = 4, 6, 8, 10$).



(a) ACE2

(b) STE12

(c) SWI4

(d) ZAP1

Figure 2.

The parameter estimates and 95% credible bands for randomly selected 4 TFs from the BSRR, where *x*-axis indicates time (min) and *y*-axis indicates estimated coefficients.

Table 1

Summary of the simulation results for examples (a1)–(a3).

Case	SNR	Method	$\mathrm{MSE}_{\mathrm{est}}$	$\mathrm{MSE}_{\mathrm{pred}}$	$\mathrm{MSE}_{\mathrm{dim}}$	FPR%	FPR%	CRI%	$\hat{r}$
(a1)	0.50	**BSRR**	7.42	109.63	17.59	0.41	0.24	100.00	3.00
		RR	14.74	220.78	49.13	28.91	0.26	98.80	3.01
		RC	16.26	212.42	198.35	3.53	0.02	0.00	9.73
	0.75	**BSRR**	2.73	40.11	5.40	0.02	0.02	100.00	3.00
		RR	5.14	81.80	10.06	28.55	0.06	100.00	3.00
		RC	5.50	71.34	64.30	0.43	0.00	0.00	9.30
	1.00	**BSRR**	1.48	21.69	2.85	0.00	0.02	100.00	3.00
		RR	2.60	42.38	3.93	28.57	0.00	100.00	3.00
		RC	2.89	37.46	33.57	0.06	0.00	0.00	9.07
(a2)	0.50	BSRR	17.00	255.77	50.98	0.01	0.68	100.00	3.00
		RR	17.48	259.70	56.60	0.01	0.44	100.00	3.00
		RC	38.16	470.79	535.99	0.03	0.01	0.00	10.00
	0.75	BSRR	6.01	96.44	9.77	0.00	0.24	100.00	3.00
		RR	6.15	97.63	11.59	0.00	0.08	100.00	3.00
		RC	16.43	204.44	232.50	0.00	0.02	0.00	10.00
	1.00	BSRR	3.06	50.14	3.70	0.00	0.10	100.00	3.00
		RR	3.12	50.62	4.32	0.00	0.04	100.00	3.00
		RC	8.93	112.27	128.20	0.00	0.00	0.00	10.00
(a3)	0.50	BSRR	0.03	0.38	0.15	0.36	0.00	18.20	9.10
		RR	0.07	0.87	0.12	28.69	0.00	23.00	9.16
		RC	0.03	0.36	0.07	1.64	0.00	98.00	9.98
	0.75	BSRR	0.02	0.19	0.10	0.80	0.00	19.40	9.12
		RR	0.03	0.40	0.09	28.82	0.00	21.40	9.14
		RC	0.01	0.15	0.03	3.50	0.00	98.20	9.98
	1.00	BSRR	0.01	0.12	0.09	1.73	0.00	20.00	9.12
		RR	0.02	0.24	0.08	29.16	0.00	21.20	9.13
		RC	0.01	0.09	0.02	4.38	0.00	98.40	9.98

Table 2

Summary of the simulation results for examples (b1)–(b3).

Case	SNR	Method	MSE$_{est}$	MSE$_{pred}$	MSE$_{dim}$	FPR%	FNR%	CRI%	$\hat{r}$
(b1)	0.50	**BSRR**	0.03	3.12	0.47	0.00	0.06	98.00	3.00
		RR	0.35	44.87	21.49	4.72	0.08	30.40	3.69
		RC	0.06	6.67	5.68	0.04	0.02	0.00	9.14
	0.75	**BSRR**	0.01	1.37	0.23	0.00	0.04	99.20	3.00
		RR	0.15	19.28	8.68	4.72	0.02	27.20	3.73
		RC	0.02	2.29	1.90	0.00	0.02	0.00	8.66
	1.00	**BSRR**	0.01	0.79	0.17	0.00	0.00	99.80	3.00
		RR	0.07	9.97	3.54	4.72	0.02	51.20	3.49
		RC	0.01	1.22	0.94	0.00	0.00	0.00	8.20
(b2)	0.50	BSRR	0.30	44.79	4.78	0.00	0.79	99.60	3.00
		RR	0.31	45.36	6.04	0.00	0.16	99.40	3.00
		RC	1.29	128.08	168.35	0.00	0.03	0.00	10.00
	0.75	BSRR	0.12	18.79	0.82	0.00	0.38	99.20	3.00
		RR	0.12	18.77	1.09	0.00	0.00	99.20	3.00
		RC	0.56	55.61	72.10	0.00	0.07	0.00	10.00
	1.00	BSRR	0.07	10.44	0.27	0.00	0.26	99.60	3.00
		RR	0.07	10.36	0.36	0.00	0.00	99.40	3.00
		RC	0.30	30.43	39.46	0.00	0.06	0.00	10.00
(b3)	0.50	BSRR	0.00	0.02	0.01	0.02	0.00	19.40	9.13
		RR	0.00	0.21	0.02	4.68	0.00	27.20	9.22
		RC	0.00	0.01	0.00	0.00	0.00	98.20	9.98
	0.75	BSRR	0.00	0.01	0.01	0.03	0.00	19.20	9.13
		RR	0.00	0.10	0.01	4.67	0.00	24.60	9.19
		RC	0.00	0.01	0.00	0.00	0.00	98.40	9.98
	1.00	BSRR	0.00	0.01	0.01	0.02	0.00	19.20	9.13
		RR	0.00	0.06	0.01	4.66	0.00	22.40	9.16
		RC	0.00	0.00	0.00	0.00	0.00	98.60	9.99

Table 3

Parameter estimates (Est) with 95% credible intervals (CIs) at 105 min from the model M^2.

TF	Est	CIs	TF	Est	CIs
ACE2	−0.03	(−0.13, 0.10)	RME1	0.05	(−0.08, 0.17
ARG81	−0.05	(−0.19, 0.09)	RTG3	0.01	(−0.09, 0.11)
FKH2	0.06	(−0.01, 0.13)	SFP1	−0.05	(−0.15, 0.05)
HIR1	0.10	(−0.06, 0.26)	SOK2	−0.03	(−0.07, 0.02)
HIR2	0.07	(−0.06, 0.21)	STB1	0.01	(−0.05, 0.06)
IME4	−0.02	(−0.13, 0.08)	STE12	−0.04	(−0.21, 0.12)
MBP1	−0.04	(−0.13, 0.05)	SWI4	0.02	(−0.03, 0.08)
MCM1	−0.03	(−0.11, 0.04)	SWI5	−0.06	(−0.21, 0.09)
MET4	0.06	(−0.02, 0.13)	SWI6	−0.01	(−0.08, 0.07)
NDD1	0.05	(−0.08, 0.17)	YAP7	0.03	(−0.04, 0.10)
NRG1	0.02	(−0.06, 0.11)	YFL044C	0.04	(−0.06, 0.15)
PHD1	−0.02	(−0.09, 0.04)	YJL206C	0.08	(−0.05, 0.22)
REB1	0.03	(−0.03, 0.09)	ZAP1	−0.03	(−0.15, 0.09)