# Sparse Reduced-Rank Regression With Adaptive Selection of Groups of Predictors

Quan Wei*, Yujia Zhang†, and Ziping Zhao*
*School of Information Science and Technology, ShanghaiTech University, Shanghai, China
†School of Mathematical Sciences, East China Normal University, Shanghai, China
Email: weiquan@shanghaitech.edu.cn, yujiazhang@stu.ecnu.edu.cn, zhaoziping@shanghaitech.edu.cn

***Abstract*—Sparse reduced-rank regression (SRRR) model has been widely used for dimension reduction and variable selection with wide applications in machine learning, signal processing, econometrics, etc. In this paper, the SRRR estimation problem is investigated which aims at adaptively selecting the whole group of predictors. The problem is formulated to minimize the least squares loss with a sparsity-inducing penalty. For effective variable selection, a novel sparse estimation scheme, called Sorted L-One Penalized Estimation (SLOPE), is considered, where a group-SLOPE type penalty is adopted. An algorithm leveraging the block majorization-minimization framework is developed for efficient problem resolution. Simulations on synthetic data illustrate the effectiveness of the proposed algorithm. The competitive performance of the proposed model in prediction accuracy is numerically demonstrated on real data with comparison to the state-of-the-art methods.**

***Index Terms*—Multivariate regression, low-rank, group sparsity, variable selection, SLOPE.**

## I. Introduction

The reduced-rank regression (RRR) model is a multivariate linear regression model, where the coefficient matrix has a low-rank property [1]. The name of "reduced-rank regression" was first brought up in [2]. Denoting the response (or dependent) variables by $\mathbf{y}_i \in \mathbb{R}^q$ and predictor (or independent) variables by $\mathbf{x}_i \in \mathbb{R}^p$, a RRR model is given as follows:

$$\begin{aligned} \mathbf{y}_i &= \boldsymbol{\mu} + \mathbf{C}\mathbf{x}_i + \boldsymbol{\epsilon}_i \\ &= \boldsymbol{\mu} + \mathbf{A}\mathbf{B}^T\mathbf{x}_i + \boldsymbol{\epsilon}_i, \end{aligned} \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^q$ is the constant intercept, $\mathbf{C} \in \mathbb{R}^{q\times p}$ is the low-rank coefficient matrix with $\mathrm{rank}(\mathbf{C}) = r \leq \min(p, q)$, and $\boldsymbol{\epsilon}_i$ is the innovation term with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$. The low-rankness of $\mathbf{C}$ offers effective dimension reduction and improves the model interpretability, in which case it can be written as $\mathbf{C} = \mathbf{A}\mathbf{B}^T$ with $\mathbf{A} \in \mathbb{R}^{q\times r}$ and $\mathbf{B} \in \mathbb{R}^{p\times r}$. Matrix $\mathbf{A}$ is often called the exposure matrix and $\mathbf{B}$ is called the factor matrix with the linear combinations $\mathbf{B}^T\mathbf{x}_i$ being the latent factors. RRR is widely used in many areas related to data analytics like financial econometrics [3], [4], portfolio design [5], [6], computer vision [7], environmental engineering [8], wireless systems [9], etc.

Variable selection is important in data analytics since it can help with model interpretability and can improve estimation and forecasting accuracy. For instance, the LASSO-type [10] methods were proposed for variable selection for linear regression models. Like the low-rank structure for factor extraction, row-wise group sparsity on matrix $\mathbf{B}$ can be further considered for selecting the predictor variables, which leads to the sparse RRR (SRRR) model. Since $\mathbf{B}^T\mathbf{x}_i$'s are interpreted as the linear factors linking the response variables and the predictors, the SRRR can generate factors only with a subset of all the predictors. To realize simultaneous dimension reduction and variable selection, the group LASSO (gLASSO) [11] penalty was first considered in [12] for the SRRR estimation problem. In [12], an double-loop algorithm based on the alternating minimization method was proposed. However, such an algorithm can be slow in practice due to its double-loop nature where lots of inner-loop iterations may be necessary to get an accurate enough solution. To tackle this issue, an efficient single-loop algorithm is proposed in [13] to reduce the computational complexity.

In practice, variable selection procedure such as LASSO may lead to results with many irrelevant variables especially when the data is highly correlated, yielding poor predictive accuracy. This requires to seek a procedure that controls the expected proportion of irrelevant variables among the selected, i.e., the false discovery rate (FDR). Inspired by the popular Benjamini–Hochberg rule [14], a generalization of LASSO, named Sorted L-One Penalized Estimation (SLOPE) [15], is proposed to control FDR at a user specified level to adapt to the unknown sparsity in variable selection, which has demonstrated its superiority over LASSO in ubiquitous engineering applications. Taking the genetic or the financial data for an example, they can be explained by several groups of factors where there exist possibly high within-group correlation but low between-group correlation. In such a situation, it is natural to select entire groups of predictors rather than the individual significant ones, in which case the group SLOPE (gSLOPE) [16] are developed in the spirit of extending LASSO to gLASSO.

In this paper, the SRRR estimation problem is studied which is formulated to minimize the ordinary least squares (OLS) loss penalized by a gSLOPE penalty. To pursue cheap updating steps for the problem resolution, an efficient and globally convergent algorithm based on the block majorization-minimization (BMM) framework [17], [18] is proposed, which makes the variables updated in closed-forms. Simulations on

This work was supported in part by the National Nature Science Foundation of China (NSFC) under Grant 62001295 and in part by the Shanghai Sailing Program under Grant 20YF1430800.

synthetic and real data showcase the efficiency of the proposed algorithm and the effectiveness of our proposed model in terms of prediction accuracy and variable selection.

## II. Problem Formulation

Given a sample path of $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N (N \geq \max(p,q))$ from model (1), the SRRR estimation problem can be generally formulated as follows:

$$\begin{aligned} \underset{\mathbf{A},\mathbf{B}}{\text{minimize}} \quad & F(\mathbf{A},\mathbf{B}) = L(\mathbf{A},\mathbf{B}) + \sigma R(\mathbf{B}) \\ \text{subject to} \quad & \mathbf{A}^T\mathbf{A} = \mathbf{I}, \end{aligned} \tag{2}$$

where $L(\mathbf{A},\mathbf{B})$ is sample loss function and $R(\mathbf{B})$ is the sparsity-inducing penalty. The constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, a Stiefel manifold, is added for identification purpose to deal with the orthogonal invariance of the parameters [12]; i.e., the solution to problem (2) is unique up to an $r \times r$ orthogonal matrix. The $L(\mathbf{A},\mathbf{B})$ is chosen as the OLS loss which is given by[1]

$$L(\mathbf{A},\mathbf{B}) = \tfrac{1}{2}\sum_{i=1}^N \left\|\mathbf{y}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i\right\|_2^2 = \tfrac{1}{2}\left\|\mathbf{Y} - \mathbf{A}\mathbf{B}^T\mathbf{X}\right\|_F^2, \tag{3}$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$. Next we will characterize the regularizer $R(\mathbf{B})$. We first define the set $\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_m\}$ ($m$ denotes the number of groups of interest) which forms a partition of the set $\{1, \ldots, p\}$, that is, $\mathcal{I}_i$'s are nonempty sets, $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset$ for $i \neq j$, and $\cup \mathcal{I}_i = \{1, \ldots, p\}$, and then we define

$$\|\mathbf{B}\|_{\mathcal{I}} = \left[\|\mathbf{B}_1\|_F, \ldots, \|\mathbf{B}_m\|_F\right]^T,$$

where $\mathbf{B}_i \in \mathbb{R}^{|\mathcal{I}_i| \times r}$ is a submatrix of $\mathbf{B}$ composed of rows indexed by the set $\mathcal{I}_i$ with $|\mathcal{I}_i|$ denoting the number of elements in set $\mathcal{I}_i$. The classical gLASSO penalty [11], [13] can be written as

$$R_{\text{gL}}(\mathbf{B}) = \lambda_{\text{gL}} \left\|\mathbf{w} \odot \|\mathbf{B}\|_{\mathcal{I}}\right\|_1 = \lambda_{\text{gL}} \sum_{i=1}^m w_i \|\mathbf{B}_i\|_F, \tag{4}$$

where $\lambda_{\text{gL}} \geq 0$ is a penalty parameter, $\mathbf{w} = [w_1, \ldots, w_m]^T$ is a positive weight vector, and $\odot$ denotes the Hadamard product. The gLASSO penalty is easy for optimization and has been shown to favor sparser solutions, but its performance degenerates when there are groups of highly correlated predictors. Instead of the $\ell_1$-norm (as used in LASSO), the method of SLOPE is induced based on the FDR control properties of the sorted $\ell_1$-norm which is defined for and $\mathbf{x} \in \mathbb{R}^p$ as $J_{\boldsymbol{\lambda}}(\mathbf{x}) = \sum_{i=1}^p \lambda_i |x|_{(i)}$, where $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_p]^T$ with its elements satisfying $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ and $|x|_{(1)} \geq \cdots \geq |x|_{(p)}$ denotes the sequence of the absolute values in $\mathbf{x}$ sorted in nonincreasing order. Clearly, $J_{\boldsymbol{\lambda}}(\mathbf{x})$ reduces to the LASSO penalty when $\lambda_1 = \cdots = \lambda_p$. To identify the relevant groups of predictors, the gSLOPE penalty considered in this paper is given as follows:[2]

$$R(\mathbf{B}) = J_{\boldsymbol{\lambda}}\left(\mathbf{w} \odot \|\mathbf{B}\|_{\mathcal{I},\mathbf{K}}\right), \tag{5}$$

[1]In this paper, the intercept term has been omitted without loss of generality as in [12], since it can always be removed by assuming that the response and predictor variables have zero mean.

[2]In [19], the SLOPE penalty is referred to as ordered weighted $\ell_1$ norm.

with $\|\mathbf{B}\|_{\mathcal{I},\mathbf{K}} = \left[\|\mathbf{B}_1^T\mathbf{K}_1\|_F, \ldots, \|\mathbf{B}_m^T\mathbf{K}_m\|_F\right]^T$, where $\mathbf{K}_i$ is the submatrix with $|\mathcal{I}_i|$ rows composed of rows indexed by the set $\mathcal{I}_i$ for a general matrix $\mathbf{K}$ with $p$ rows. In practice, $\mathbf{K} = \mathbf{X}$ or $\mathbf{K} = \mathbf{I}$ are common used [20]. It can be proved the gSLOPE penalty $R(\mathbf{B})$ is convex in $\mathbf{B}$.

Finally, based on $L(\mathbf{A},\mathbf{B})$ and $R(\mathbf{B})$, the problem in (2) becomes a nonconvex nonsmooth optimization problem.

## III. Problem Solving via BMM

### A. The BMM Method

To solve problem (2), we adopt the block majorization-minimization (BMM) method [17], [18], which is briefly stated below. Consider the following optimization problem:

$$\underset{\mathbf{x}}{\text{minimize}} \quad f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{x} \in \mathcal{X},$$

where the optimization variable $\mathbf{x}$ can be partitioned into $n$ blocks as $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with $\mathbf{x}_i \in \mathcal{X}_i$ and $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$, and $f : \mathcal{X} \to \mathbb{R}$ is a continuous function. At the $k$-th iteration, the $i$-th block $\mathbf{x}_i$ is updated according to the following rules:

$$\begin{cases} \mathbf{x}_i^{(k)} \in \arg\min\limits_{\mathbf{x}_i \in \mathcal{X}_i} \overline{f}_i\big(\mathbf{x}_i, \mathbf{x}_{-i}^{(k-1)}\big) \\ \mathbf{x}_{-i}^{(k)} = \mathbf{x}_{-i}^{(k-1)} \text{ with } \mathbf{x}_{-i} = (\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n), \end{cases}$$

where $\overline{f}_i$ is a majorizing function of $f(\mathbf{x})$ w.r.t. $\mathbf{x}_i$ satisfying

A1) $\overline{f}_i\big(\mathbf{x}_i^{(k)}; \mathbf{x}^{(k)}\big) = f\big(\mathbf{x}^{(k)}\big)$,

A2) $\overline{f}_i\big(\mathbf{x}_i; \mathbf{x}^{(k)}\big) \geq f\big(\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n^{(k-1)}\big)$,

A3) $\nabla_{\mathbf{x}_i} \overline{f}_i\big(\mathbf{x}_i^{(k)}; \mathbf{x}^{(k)}, \mathbf{d}_i\big) = \nabla f\big(\mathbf{x}^{(k)}; \mathbf{d}\big)$, $\forall \mathbf{d}$ s.t. $\mathbf{x}_i^{(k)} + \mathbf{d}_i \in \mathcal{X}_i$ and $\mathbf{d} = (\mathbf{0}, \ldots, \mathbf{d}_i, \ldots, \mathbf{0})$.

In summary, BMM suffices to carry out a sequential block coordinate inexact update [21]. The majorizing functions in BMM can be chosen in a flexible way [18] while a properly chosen one can make the updates easy and lead to a fast convergence over iterations. In practice, the surrogate subproblems are applaudable if they are convex (thus, efficiently solvable) or have closed-form solutions. In the following, we will derive a BMM-based algorithm for problem (2).

### B. Solving The $\mathbf{A}$-Subproblem

When $\mathbf{B}$ is fixed at $\mathbf{B}^{(k)}$, the problem w.r.t. $\mathbf{A}$ becomes

$$\begin{aligned} \underset{\mathbf{A}}{\text{minimize}} \quad & \tfrac{1}{2}\left\|\mathbf{Y} - \mathbf{A}\mathbf{B}^{(k)T}\mathbf{X}\right\|_F^2 \\ \text{subject to} \quad & \mathbf{A}^T\mathbf{A} = \mathbf{I}. \end{aligned} \tag{6}$$

Problem (6) is an orthogonal Procrustes problem [22]. The optimal update $\mathbf{A}^\star$ is given by

$$\mathbf{A}^\star = \mathbf{U}\mathbf{V}^T, \tag{7}$$

where $\mathbf{U} \in \mathbb{R}^{q \times r}$ and $\mathbf{V} \in \mathbb{R}^{r \times r}$ are obtained based on the singular value decomposition $\mathbf{Y}\mathbf{X}^T\mathbf{B}^{(k)} = \mathbf{U}\mathbf{D}\mathbf{V}^T$.

### C. Solving The $\mathbf{B}$-Subproblem

Now we consider the optimization over $\mathbf{B}$ for fixed $\mathbf{A}^{(k)}$. Since $\mathbf{A}$ has orthonormal columns, there is always a matrix $\mathbf{A}^{\perp}$ with orthonormal columns lying in the orthogonal complement of $\mathbf{A}$ such that $[\mathbf{A}, \mathbf{A}^{\perp}]$ is orthogonal. Then we have

$$\begin{aligned}\left\|\mathbf{Y}-\mathbf{A}\mathbf{B}^T\mathbf{X}\right\|_F^2 &= \left\|[\mathbf{A},\mathbf{A}^{\perp}]^T[\mathbf{Y}-\mathbf{A}\mathbf{B}^T\mathbf{X}]\right\|_F^2 \\ &= \left\|\mathbf{A}^T\mathbf{Y}-\mathbf{B}^T\mathbf{X}\right\|_F^2 + \left\|(\mathbf{A}^{\perp})^T\mathbf{Y}\right\|_F^2.\end{aligned}$$

Then the subproblem w.r.t. $\mathbf{B}$ becomes

$$\underset{\mathbf{B}}{\text{minimize}} \;\; \tfrac{1}{2}\left\|\mathbf{A}^{(k)T}\mathbf{Y}-\mathbf{B}^T\mathbf{X}\right\|_F^2 + \sigma J_{\boldsymbol{\lambda}}\left(\mathbf{w}\odot\|\mathbf{B}\|_{\mathcal{I},\mathbf{K}}\right) \quad (8)$$

which is a convex problem and is readily solved by the off-the-shelf solvers or calling scripting languages like CVX [23].

To invoke a cheap update, we will find a closed-form solution for the $\mathbf{B}$-subproblem.

**Lemma 1.** *$\mathbf{K}_i$ can be factorized as $\mathbf{K}_i = \mathbf{L}_i\mathbf{Q}_i$ by LQ decomposition[3], where $\mathbf{Q}_i$ is a matrix with $l_i$ orthonormal rows, whose span coincides with the subspace spanned by the rows of $\mathbf{K}_i$, and $\mathbf{L}_i \in \mathbb{R}^{|\mathcal{I}_i|\times l_i}$ is a matrix of full column rank.*

Note that, in general $\mathbf{K}_i$ may not be of full rank, i.e., $\operatorname{rank}(\mathbf{K}_i) = l_i \le |\mathcal{I}_i|$. Based on Lemma 1, there is always a matrix $\mathbf{L}_i^{\perp} \in \mathbb{R}^{|\mathcal{I}_i|\times(|\mathcal{I}_i|-l_i)}$ such that $\left[\mathbf{L}_i, \mathbf{L}_i^{\perp}\right] \in \mathbb{R}^{|\mathcal{I}_i|\times|\mathcal{I}_i|}$. For $i = 1, \ldots, m$, we have

$$\mathbf{B}_i^T\left[\mathbf{L}_i, \mathbf{L}_i^{\perp}\right] = \left[\mathbf{B}_i^T\mathbf{L}_i, \mathbf{B}_i^T\mathbf{L}_i^{\perp}\right] = \left[\hat{\mathbf{B}}_i^T, \check{\mathbf{B}}_i^T\right],$$

where $\hat{\mathbf{B}}_i \in \mathbb{R}^{l_i\times r}$ and $\check{\mathbf{B}}_i \in \mathbb{R}^{(|\mathcal{I}_i|-l_i)\times r}$ and

$$\left[\mathbf{L}_i, \mathbf{L}_i^{\perp}\right]^{-1}\mathbf{X}_i = \begin{bmatrix}\hat{\mathbf{X}}_i \\ \check{\mathbf{X}}_i\end{bmatrix},$$

where $\hat{\mathbf{X}}_i \in \mathbb{R}^{l_i\times N}$ and $\check{\mathbf{X}}_i \in \mathbb{R}^{(|\mathcal{I}_i|-l_i)\times N}$. Then we obtain

$$\begin{aligned}\mathbf{B}^T\mathbf{X} &= \sum_{i=1}^m \mathbf{B}_i^T\mathbf{X}_i = \sum_{i=1}^m\left(\mathbf{B}_i^T\left[\mathbf{L}_i, \mathbf{L}_i^{\perp}\right]\times\left[\mathbf{L}_i, \mathbf{L}_i^{\perp}\right]^{-1}\mathbf{X}_i\right) \\ &= \sum_{i=1}^m\left(\hat{\mathbf{B}}_i^T\hat{\mathbf{X}}_i + \check{\mathbf{B}}_i^T\check{\mathbf{X}}_i\right) \\ &= \hat{\mathbf{B}}^T\hat{\mathbf{X}} + \check{\mathbf{B}}^T\check{\mathbf{X}},\end{aligned}$$

where $\hat{\mathbf{B}} \in \mathbb{R}^{\hat{p}\times r}$, $\check{\mathbf{B}} \in \mathbb{R}^{\check{p}\times r}$, $\hat{\mathbf{X}} \in \mathbb{R}^{\hat{p}\times n}$, and $\check{\mathbf{X}} \in \mathbb{R}^{\check{p}\times n}$ with $\hat{p} = l_1 + \ldots + l_m$ and $\check{p} = p - \hat{p}$. If we define the partition, $\hat{\mathcal{I}} = \left\{\hat{\mathcal{I}}_1, \ldots, \hat{\mathcal{I}}_m\right\}$, of the set $\left\{1, \ldots, \hat{p}\right\}$ with

$$\begin{cases}\hat{\mathcal{I}}_1 = \{1, \ldots, l_1\}, \\ \hat{\mathcal{I}}_2 = \{l_1+1, \ldots, l_1+l_2\}, \\ \vdots \\ \hat{\mathcal{I}}_m = \left\{\sum_{j=1}^{m-1} l_j + 1, \ldots, \sum_{j=1}^{m} l_j\right\}.\end{cases}$$

and define the following partition, $\check{\mathcal{I}} = \left\{\check{\mathcal{I}}_1, \ldots, \check{\mathcal{I}}_m\right\}$, of the set $\left\{1, \ldots, \check{p}\right\}$ with[4]

$$\begin{cases}\check{\mathcal{I}}_1 = \{1, \ldots, |\mathcal{I}_1| - l_1\}, \\ \check{\mathcal{I}}_2 = \{|\mathcal{I}_1| - l_1 + 1, \ldots, |\mathcal{I}_1| - l_1 + l_2\}, \\ \vdots \\ \check{\mathcal{I}}_m = \left\{\sum_{j=1}^{m-1}(|\mathcal{I}_j| - l_j) + 1, \ldots, \sum_{j=1}^{m}(|\mathcal{I}_j| - l_j)\right\}.\end{cases}$$

We have $\hat{\mathbf{B}}_i$ (resp. $\check{\mathbf{B}}_i$) is the submatrix of $\hat{\mathbf{B}}$ (resp. $\check{\mathbf{B}}$) composed of rows indexed by the set $\hat{\mathcal{I}}_i$ (resp. $\check{\mathcal{I}}_i$) and similar results hold for $\hat{\mathbf{X}}_i$ and $\check{\mathbf{X}}_i$. Also, we have[5]

$$\begin{aligned}\left[\|\mathbf{B}\|_{\mathcal{I},\mathbf{K}}\right]_i &= \|\mathbf{B}_i^T\mathbf{K}_i\|_F = \|\mathbf{B}_i^T\mathbf{L}_i\mathbf{Q}_i\|_F \\ &= \|\mathbf{B}_i^T\mathbf{L}_i\|_F = \|\hat{\mathbf{B}}_i\|_F.\end{aligned}$$

Finally, the objective in (8) can be equivalently recast as

$$\tfrac{1}{2}\left\|\mathbf{A}^T\mathbf{Y}-\hat{\mathbf{B}}^T\hat{\mathbf{X}}-\check{\mathbf{B}}^T\check{\mathbf{X}}\right\|_F^2 + \sigma J_{\boldsymbol{\lambda}}\big(\mathbf{w}\odot\|\hat{\mathbf{B}}\|_{\hat{\mathcal{I}}}\big), \quad (9)$$

where $\|\hat{\mathbf{B}}\|_{\hat{I}} = \left[\|\hat{\mathbf{B}}_1\|_F, \ldots, \|\hat{\mathbf{B}}_m\|_F\right]^T$. Based on this objective, problem (9) is minimized when $\check{\mathbf{B}}$ is chosen to be

$$\check{\mathbf{B}}^{\star} = \left(\check{\mathbf{X}}\check{\mathbf{X}}^T\right)^{-1}\check{\mathbf{X}}\left(\mathbf{A}^T\mathbf{Y}-\hat{\mathbf{B}}^T\hat{\mathbf{X}}\right)^T.$$

Substituting for $\check{\mathbf{B}}$ in (9), the $\mathbf{B}$ subproblem becomes

$$\underset{\hat{\mathbf{B}}}{\text{minimize}} \;\; \tfrac{1}{2}\left\|\mathbf{A}^{(k)T}\tilde{\mathbf{Y}}-\hat{\mathbf{B}}^T\tilde{\mathbf{X}}\right\|_F^2 + \sigma J_{\boldsymbol{\lambda}}\big(\mathbf{w}\odot\|\hat{\mathbf{B}}\|_{\hat{\mathcal{I}}}\big), \quad (10)$$

where $\mathbf{P} = \check{\mathbf{X}}^T\left(\check{\mathbf{X}}\check{\mathbf{X}}^T\right)^{-1}\check{\mathbf{X}}$, $\tilde{\mathbf{Y}} = \mathbf{Y}(\mathbf{I}-\mathbf{P})$, and $\tilde{\mathbf{X}} = \hat{\mathbf{X}}(\mathbf{I}-\mathbf{P})$. To obtain a simple update rule while guaranteeing convergence of the overall algorithm. We propose to update $\hat{\mathbf{B}}$ by solving a majorized problem for problem (10). We first give the following useful result.

**Lemma 2** (Quadratic Majorization [18])**.** *Let $\mathbf{M}, \mathbf{N} \in \mathbb{S}$ and $\mathbf{M} \preceq \mathbf{N}$, at any point $\mathbf{X}^{(k)}$, it follows that*

$$\begin{aligned}\operatorname{tr}\big(\mathbf{X}^T\mathbf{M}\mathbf{X}\big) \le \operatorname{tr}\big(\mathbf{X}^T\mathbf{N}\mathbf{X}\big) &- 2\operatorname{tr}\big(\mathbf{X}^T(\mathbf{N}-\mathbf{M})\mathbf{X}^{(k)}\big) \\ &+ \operatorname{tr}\big(\mathbf{X}^{(k)T}(\mathbf{N}-\mathbf{M})\mathbf{X}^{(k)}\big),\end{aligned}$$

*where the equality is achieved at $\mathbf{X} = \mathbf{X}^{(k)}$.*

Based on Lemma 1, the first quadratic part in the objective of problem (10) can be majorized at $\left(\mathbf{A}^{(k)}, \hat{\mathbf{B}}^{(k)}\right)$ as follows:

$$\tfrac{1}{2}\left\|\mathbf{A}^{(k)T}\tilde{\mathbf{Y}}-\hat{\mathbf{B}}^T\tilde{\mathbf{X}}\right\|_F^2 \le \tfrac{1}{2}t\left\|\hat{\mathbf{B}}-\mathbf{G}^{(k)}\right\|_F^2 + const., \quad (11)$$

where $t \ge \lambda_{\max}\big(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\big)$,

$$\mathbf{G}^{(k)} = \big(\mathbf{I}-t^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\big)\hat{\mathbf{B}}^{(k)} + t^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T\mathbf{A}^{(k)}, \quad (12)$$

and $const.$ denotes a variable-irrelevant constant.

Finally, the $\mathbf{B}$-subproblem after majorization is given in the following form:

$$\underset{\hat{\mathbf{B}}}{\text{minimize}} \;\; \tfrac{1}{2}\left\|\hat{\mathbf{B}}-\mathbf{G}^{(k)}\right\|_F^2 + \sigma J_{\tilde{\boldsymbol{\lambda}}}\big(\mathbf{w}\odot\|\hat{\mathbf{B}}\|_{\hat{\mathcal{I}}}\big), \quad (13)$$

[3]If $\mathbf{A} = \mathbf{Q}\mathbf{R}$ is the QR decomposition of $\mathbf{A}$ then $\mathbf{A}^T = \mathbf{R}^T\mathbf{Q}^T$ is the LQ decomposition of $\mathbf{A}^T$.

[4]Note that $\check{\mathcal{I}} = \emptyset$ and $\hat{\mathcal{I}} = \mathcal{I}$ when all $\mathbf{K}_i$'s are of full rank.

[5]The $[\mathbf{x}]_i$ is used to denote the $i$th element in $\mathbf{x}$.

**Algorithm 1:** BMM Algorithm for SRRR Estimation

**Input:** data $\mathbf{Y}, \mathbf{X}$ and parameter $\sigma, \boldsymbol{\lambda}, \mathbf{w}$
**Initialize** $k = 1$, $\mathbf{A}^{(0)}$ and $\mathbf{B}^{(0)}$;
**repeat**
  compute SVD of $\mathbf{Y}\mathbf{X}^T\mathbf{B}^{(k-1)}$;
  update $\mathbf{A}^{(k)}$ in closed-form solution Eq. (7);
  compute $\mathbf{G}^{(k)}$ in Eq. (12);
  update $\mathbf{B}^{(k)}$ in closed-form solution Eq. (14);
  $k \leftarrow k + 1$;
**until** *convergence*;
**Output:** $\mathbf{A}, \mathbf{B}$

where $\tilde{\boldsymbol{\lambda}} = \sigma\frac{\boldsymbol{\lambda}}{t}$. Problem (13) can be efficiently solved in closed-form via the proximal algorithm [24] as follows:

$$\begin{cases} \hat{\mathbf{B}}_i^\star = \left[\text{prox}_{J_{\tilde{\boldsymbol{\lambda}}}}\left(\mathbf{G}^{(k)}\right)\right]_i = c_i^\star \left(w_i \left\|\mathbf{G}_i^{(k)}\right\|_F\right)^{-1} \mathbf{G}_i^{(k)}, \\ \qquad\qquad \text{for } i = 1, \ldots, m \\ \text{with } \mathbf{c}^\star = \arg\min\limits_{\mathbf{c}\in\mathbb{R}^m} \left\{ \frac{1}{2}\sum_{i=1}^m \left(\left\|\mathbf{G}_i^{(k)}\right\|_F - w_i^{-1}c_i\right)^2 + J_{\tilde{\boldsymbol{\lambda}}}(\mathbf{c}) \right\} \end{cases} \tag{14}$$

Consequently, calculating $\text{prox}_{J_{\tilde{\boldsymbol{\lambda}}}}\left(\mathbf{G}^{(k)}\right)$ reduces to identifying $\mathbf{c}^\star$, which can be efficiently obtained via the fast proximal step in closed-from for regular SLOPE as provided in [15].

### D. The Overall Algorithm

Based on the BMM algorithm, to solve the original SRRR estimation problem (2), we just need to update two variable blocks with closed-form solutions alternately until convergence. The overall algorithm is summarized in Algorithm 1.

**Theorem 3.** *The sequence* $\{\mathbf{A}^{(k)}, \mathbf{B}^{(k)}\}$ *generated by Algorithm 1 converges globally to a stationary point of Problem (2).*

*Proof:* This result can be established by checking the satisfaction of the constraint qualification condition called linear independence constraint qualification [21] for $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ and then following the convergence proof in [17]. ■

## IV. Numerical Simulations

### A. Simulations on Synthetic Data

To validate the the performance of the proposed SRRR model and the algorithm, simulations on synthetic data are first conducted. An SRRR $(p = 20, q = 12, r = 4)$ with underlying group sparse structure for $\mathbf{B}$ is specified firstly.

We first examine the efficiency of the proposed BMM algorithm. We compare our BMM algorithm with the one where the $\mathbf{B}$-subproblem in (8) is solved via CVX. The convergence result of the objective function value with sample size $N = 1000$ is shown in Fig. 1. It is easy to see that our proposed algorithm can achieve a faster convergence in terms of CPU time.

We also examine the estimation accuracy of the proposed formulation and algorithm. It is evaluated by computing the mean squared error (MSE), defined as

$$\text{MSE} = \left\|(\mathbf{A}^\star\mathbf{B}^{\star T} - \mathbf{A}\mathbf{B}^T)\mathbf{X}\right\|_F^2 / Nq. \tag{15}$$

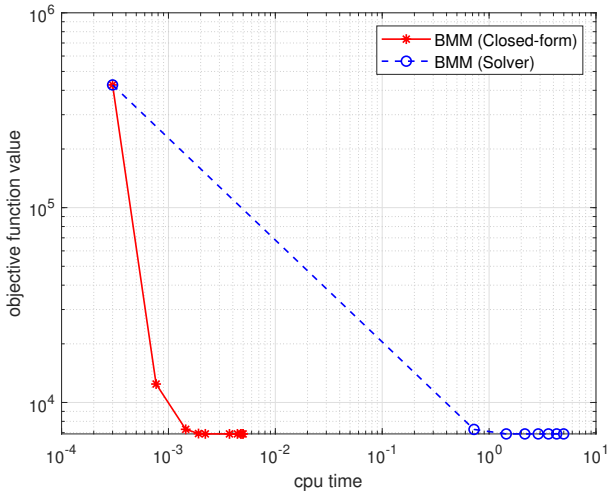


Fig. 1. Comparison of convergence speed for objective function value.

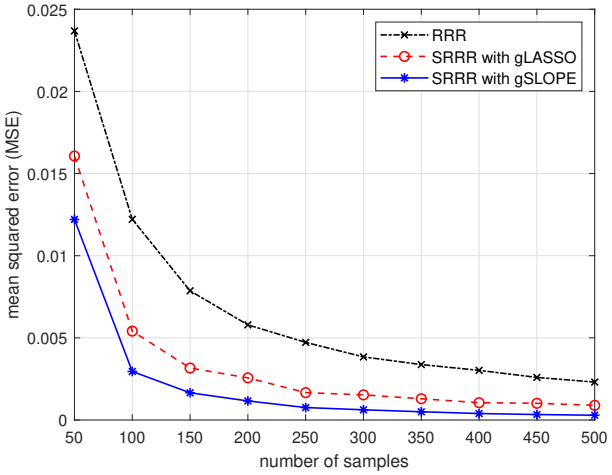


Fig. 2. Comparison of predictive accuracy.

We compared three cases which are RRR estimation (without sparsity), SRRR estimation with gLASSO, and SRRR estimation with gSLOPE. The result is averaged over 50 Monte Carlo simulations. Fig. 2 shows that the both SRRR problem formulations can exploit the group sparsity structures in $\mathbf{B}$ and the method of gSLOPE attains a better performance in comparison to the gLASSO method.

### B. Simulations on Real Data

TABLE I
Performance Comparisons

| | SRRR (gLASSO) | SRRR (gSLOPE) |
|---|---|---|
| MSE | 0.095 | 0.094 |
| Number of selected TFs $(s)$ | 70 | 68 |
| Number of confirmed TFs $(k)$ | 16 | 19 |
| Prob$(K \geq k)$[6] | 0.117 | 0.003 |

[6]Prob$(K \geq k)$ denotes the probability of observing at least $k$ confirmed variables out of 85 unconfirmed and 21 confirmed variables in a random draw of $s$ variables.

Real data analysis is concerned with identifying transition factors (TFs) that regulate the RNA transcript levels of yeast genes within the eukaryotic cell cycle. We utilize a yeast cell cycle gene expression data set from [25]. The genes whose RNA levels varied periodically were identified as cell-cycle-regulated genes. The experiments measure messenger ribonucleic acid levels every 7 min. for 119 min. with a total of 18 measurements covering two cell cycle periods ($q = 18$). The chromatin immunoprecipitation data contain binding information of a total of 106 TFs ($p = 106$). After excluding genes with missing values in either of the experiments, 542 cell-cycle-related genes are retained ($N = 542$). We apply the SRRR model with gSLOPE and gLASSO regularizer to these data. Five-fold cross-validation (CV) was used for selecting tuning parameters, including the number of factors $r$ for SRRR. In the experiment, we set $r = 4$ due to the minimum CV error.

For comparison of prediction accuracy, we randomly split the data into two halves, one as the training set and another as the test set. The training set is used to identify the best penalty parameters and then the specified model is used to make prediction on the test data. Table I summarizes the MSE on the testing set of the two methods.

For the purpose of variable selection comparison, we fit the entire dataset using the optimal penalty parameters selected by the five-fold CV and kept track on whether each of 106 TFs is chosen in the fitted model. Table I shows how many TFs are selected and also reports how many of these consistently selected TFs are among the 21 experimentally confirmed TFs that are related to the cell cycle regulation [26]. We can see that gSLOPE identifies a total of 68 TFs that are related to yeast cell cycle processes, including 19 of the 21 confirmed TFs, while gLASSO identifies 16 confirmed TFs of 70 selected TFs. It clearly demonstrates that gSLOPE can select more relevant variables than gLASSO. We also report a hypergeometric probability calculation quantifying chance occurrences of the number of confirmed TFs among the variables selected by each method. A comparison of these probabilities indicates that method of gSLOPE has more evidence that selection of a large number of confirmed TFs is not due to chance.

## V. Conclusions

In this paper, the SRRR model estimation problem has been considered. It has been formulated to minimize the least squares loss with the SLOPE sparsity-inducing penalty by considering an orthogonality constraint. An efficient and convergent algorithm has been proposed which exploits the problem structure and has a low computational complexity. Numerical simulations demonstrate the efficiency of the proposed algorithm and the model over the existing ones in literature.

## References


[1] G. C. Reinsel and R. P. Velu, *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer Science & Business Media, 1998.

[2] A. J. Izenman, "Reduced-rank regression for the multivariate linear model," *Journal of Multivariate Analysis*, vol. 5, no. 2, pp. 248–264, 1975.

[3] E. Bernardini and G. Cubadda, "Macroeconomic forecasting and structural analysis through regularized reduced-rank regression," *International Journal of Forecasting*, vol. 31, no. 3, pp. 682–691, 2015.

[4] Z. Zhao and D. P. Palomar, "Robust maximum likelihood estimation of sparse vector error correction model," in *Proc. the 2017 5th IEEE Global Conference on Signal and Information Processing*, Montreal, QC, Canada, Nov. 2017, pp. 913–917.

[5] ——, "Mean-reverting portfolio with budget constraint," *IEEE Transactions on Signal Processing*, vol. 66, no. 9, pp. 2342–2357, 2018.

[6] Z. Zhao, R. Zhou, and D. P. Palomar, "Optimal mean-reverting portfolio with leverage constraint for statistical arbitrage in finance," *IEEE Transactions on Signal Processing*, vol. 67, no. 7, pp. 1681–1695, 2019.

[7] D. Huang and F. De la Torre, "Bilinear kernel reduced rank regression for facial expression synthesis," in *European Conference on Computer Vision*, 2010, pp. 364–377.

[8] C. A. Glasbey, "A reduced rank regression model for local variation in solar radiation," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 41, no. 2, pp. 381–387, 1992.

[9] M. Nicoli and U. Spagnolini, "Reduced-rank channel estimation for time-slotted mobile communication systems," *IEEE Transactions on Signal Processing*, vol. 53, no. 3, pp. 926–944, 2005.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[11] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[12] L. Chen and J. Z. Huang, "Sparse reduced-rank regression for simultaneous dimension reduction and variable selection," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1533–1545, 2012.

[13] Z. Zhao and D. P. Palomar, "Sparse reduced rank regression with nonconvex regularization," in *2018 IEEE Statistical Signal Processing Workshop (SSP)*, 2018, pp. 811–815.

[14] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[15] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès, "SLOPE—adaptive variable selection via convex optimization," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1103–1140, 2015.

[16] D. Brzyski, A. Gossmann, W. Su, and M. Bogdan, "Group SLOPE—adaptive selection of groups of predictors," *Journal of the American Statistical Association*, vol. 114, no. 525, pp. 419–433, 2019.

[17] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.

[18] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017.

[19] M. Figueiredo and R. Nowak, "Ordered weighted L1 regularized regression with strongly correlated covariates: Theoretical aspects," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016, pp. 930–938.

[20] N. Simon and R. Tibshirani, "Standardization and the group lasso penalty," *Statistica Sinica*, vol. 22, pp. 983–1001, 2012.

[21] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[22] J. C. Gower and G. B. Dijksterhuis, *Procrustes Problems*. Oxford University Press, 2004.

[23] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[24] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[25] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[26] L. Wang, G. Chen, and H. Li, "Group SCAD regression analysis for microarray time course gene expression data," *Bioinformatics*, vol. 23, no. 12, pp. 1486–1494, 2007.