Envelope-based sparse reduced-rank regression for multivariate linear model

Wenxing Guo[a,*], N. Balakrishnan[b], Mu He[c]

[a]*Department of Mathematical Sciences, University of Essex, Colchester, United Kingdom.*
[b]*Department of Mathematics and Statistics, McMaster University, Hamilton, Canada.*
[c]*Department of Foundational Mathematics, Xi'an Jiaotong-Liverpool University, Suzhou, China.*

Abstract

Envelope models were first proposed by Cook et al. [10] as a method to reduce estimative and predictive variations in multivariate regression. Sparse reduced-rank regression, introduced by Chen and Huang [4], is a widely used technique that performs dimension reduction and variable selection simultaneously in multivariate regression. In this work, we combine envelope models and sparse reduced-rank regression method to propose an envelope-based sparse reduced-rank regression estimator, and then establish its consistency, asymptotic normality and oracle property in high-dimensional data. We carry out some Monte Carlo simulation studies and also analyze two datasets to demonstrate that the proposed envelope-based sparse reduced-rank regression method displays good variable selection and prediction performance.

Keywords: Dimension reduction, Envelope model, High dimension, Reduced-rank regression, Variable selection
2020 MSC: Primary 62F12, 62H12, Secondary 62J99

1. Introduction

In this work, we consider the following multivariate linear regression model:

$$Y_i = \boldsymbol{\beta} X_i + \boldsymbol{\varepsilon}_i, \quad i \in \{1, \ldots, n\}, \tag{1}$$

where $Y_i \in \mathbb{R}^r$ denotes a multivariate response vector, $X_i \in \mathbb{R}^p$ denotes a non-stochastic vector of predictors, $\boldsymbol{\varepsilon}_i$ is an error vector having mean 0, covariance matrix $\boldsymbol{\Sigma}$ and is independent of X_i, and $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ is the regression coefficient matrix in which we are primarily interested in. If X_i is a vector of random quantities during sampling, then the model is conditional on the observed values of X_i. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$, respectively. Without loss of generality, let us assume that the data are centered, so that the intercept can be excluded from the regression model. Then, model (1) can be re-expressed as

$$\boldsymbol{Y} = \boldsymbol{\beta} \boldsymbol{X} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\varepsilon}$ is $(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)$.

1.1. Notation and definitions

For positive integers r and p, $\mathbb{R}^{r\times p}$ represents the class of all real matrices of dimension $r \times p$, and $\mathbb{S}^{r\times r}$ denotes the class of all symmetric $r \times r$ matrices. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(\boldsymbol{A})$ stands for the trace of $\boldsymbol{A}$. For $\mathbf{B} \in \mathbb{R}^{r\times p}$, span $(\mathbf{B})$ stands for the subspace of $\mathbb{R}^r$ spanned by the columns of $\mathbf{B}$, the Frobenious norm of $\boldsymbol{B}$ is denoted by $\|\boldsymbol{B}\|_F = \sqrt{\mathrm{tr}\,(\boldsymbol{B}^T\boldsymbol{B})}$. $\boldsymbol{B}^+$ denotes the Moore–Penrose inverse of $\boldsymbol{B}$. For a column vector X, the Euclidean norm is represented as $\|X\|_2 = \sqrt{X^T X}$. A basis matrix for a subspace $\mathcal{S}$ is any matrix whose columns form a basis for $\mathcal{S}$. A subspace $\mathcal{R}$ of $\mathbb{R}^r$ is a reducing subspace of $\mathbf{M} \in \mathbb{R}^{r\times r}$ if $\mathbf{M}\mathcal{R} \subseteq \mathcal{R}$ and $\mathbf{M}\mathcal{R}^\perp \subseteq \mathcal{R}^\perp$.

*Corresponding author. Email address: wg22745@essex.ac.uk

1.2. Review of envelope models

Envelope method was developed for the multivariate linear model by Cook et al. [10]. It is built on a key assumption that the linear combination of some response variables is irrelevant as the predictors vary. The goal of this method is then to reduce the dimension of variables and improve efficiency. More specifically, let $\boldsymbol{P}_{\xi}\boldsymbol{Y}$ denote the projection of $\boldsymbol{Y}$ onto a subspace $\xi \subseteq \mathbb{R}^r$ with the following two properties: (i) The distribution of $\boldsymbol{Q}_{\xi}\boldsymbol{Y}|\boldsymbol{X}$ does not depend on $\boldsymbol{X}$, where $\boldsymbol{Q}_{\xi} = \boldsymbol{I}_r - \boldsymbol{P}_{\xi}$, and (ii) $\boldsymbol{P}_{\xi}\boldsymbol{Y}$ is independent of $\boldsymbol{Q}_{\xi}\boldsymbol{Y}$, given $\boldsymbol{X}$. The two conditions, when combined, imply that the distribution of $\boldsymbol{Q}_{\xi}\boldsymbol{Y}$ is not affected marginally by $\boldsymbol{X}$ or through an association with $\boldsymbol{P}_{\xi}\boldsymbol{Y}$. As a result, changes in $\boldsymbol{X}$ influence this distribution only through $\boldsymbol{P}_{\xi}\boldsymbol{Y}$. Furthermore, conditions (i) and (ii) hold if and only if (a) $\mathcal{B} \triangleq \text{span}(\boldsymbol{\beta})$ is the subspace of ξ and (b) ξ is a reducing subspace of $\boldsymbol{\Sigma}$. The $\boldsymbol{\Sigma}$-envelope of $\mathcal{B}$, denoted by $\xi_{\boldsymbol{\Sigma}}(\mathcal{B})$, is defined formally as the intersection of all reducing subspaces of $\boldsymbol{\Sigma}$ that contain $\mathcal{B}$. Let u=dim$\{\xi_{\boldsymbol{\Sigma}}(\mathcal{B})\}$ and $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{r\times r}$ be an orthogonal matrix with $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ being a column orthogonal matrix and span($\boldsymbol{\Gamma}$)=$\xi_{\boldsymbol{\Sigma}}(\mathcal{B})$. This then leads directly to the following envelope version of model (1):

$$Y_i = \boldsymbol{\beta}X_i + \varepsilon_i, \;\; \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_{\boldsymbol{0}}\boldsymbol{\Omega}_{\boldsymbol{0}}\boldsymbol{\Gamma}_{\boldsymbol{0}}^T, \;\; i \in \{1, \ldots, n\}, \tag{3}$$

where $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\gamma}$ with $\boldsymbol{\gamma} \in \mathbb{R}^{u\times p}$ representing the coordinates of $\boldsymbol{\beta}$ corresponding to the basis $\boldsymbol{\Gamma}$, $\boldsymbol{\Omega} \in \mathbb{S}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{S}^{(r-u)\times(r-u)}$ are both positive definite matrices, $\boldsymbol{\gamma}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ depend on the basis $\boldsymbol{\Gamma}$. It should be mentioned that the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ depend only on $\xi_{\boldsymbol{\Sigma}}(\mathcal{B})$ rather than on the basis. The estimators of the parameters in (3) can be achieved by maximum likelihood estimation, and dimension u of the envelope can be determined based on likelihood ratio test, information criteria, and cross-validation. The envelope estimator $\hat{\boldsymbol{\beta}}_{en}$ of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_{en} = \boldsymbol{P}_{\hat{\xi}}\hat{\boldsymbol{\beta}}_{OLS}$, is just the projection of the ordinary least-squares estimator $\hat{\boldsymbol{\beta}}_{OLS}$ of $\boldsymbol{\beta}$ onto the estimated envelope. A detailed review of envelope models can be found in Cook et al. [8] and Cook [7].

1.3. Review of reduced-rank regression

From model (2), the ordinary least-squares (OLS) estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}. \tag{4}$$

It is clear that the OLS estimator of multiple responses is equivalent to performing separate OLS estimation for each response variable, and so the estimator does not make use of the likely correlation existing between the multiple responses. It will, of course, be useful to consider the correlation between response variables. One way of incorporating possible interrelationships between response variables is to consider reduced-rank regression (RRR) model (Reinsel and Velu [20]). The reduced-rank regression would allow the rank of $\boldsymbol{\beta}$ to be less than $\min(p, r)$, and so the model parametrization can be expressed as $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{B}$, where $\boldsymbol{A} \in \mathbb{R}^{r\times d}$, $\boldsymbol{B} \in \mathbb{R}^{d\times p}$, and rank($\boldsymbol{A}$)=rank($\boldsymbol{B}$)=$d$. The decomposition $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{B}$ is non-unique since for any orthogonal matrix $\boldsymbol{O} \in \mathbb{R}^{d\times d}$, $\boldsymbol{A}^* = \boldsymbol{A}\boldsymbol{O}$ and $\boldsymbol{B}^* = \boldsymbol{O}^T\boldsymbol{B}$ will result in other valid decompositions satisfying $\boldsymbol{\beta} = \boldsymbol{A}^*\boldsymbol{B}^* = \boldsymbol{A}\boldsymbol{B}$. Nevertheless, the parameter $\boldsymbol{\beta}$ of interest is identifiable, as well as span($\boldsymbol{A}$)=span($\boldsymbol{\beta}$) and span($\boldsymbol{B}^T$)=span($\boldsymbol{\beta}^T$) (Cook et al. [9]). Under some constraints on $\boldsymbol{A}$ and $\boldsymbol{B}$, such as $\boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{I}_d$ or $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_d$, Anderson [1] and Reinsel and Velu [20] derived the maximum likelihood estimators of the RRR parameters. As there are some linear constraints on the regression coefficients, the number of effective parameters gets reduced and the prediction accuracy may therefore get improved. In high-dimensional data, a large number of predictor variables will be typically available, but some of them may not be useful for predictive purpose. For this reason, Chen and Huang [4] proposed sparse reduced-rank regression for simultaneous dimension reduction and variable selection in multivariate regression with fixed dimension of parameters in terms of penalty functions. Lian and Kim [17] provided sufficient conditions to guarantee the oracle estimator to be a local minimizer, and stronger conditions to guarantee that it is a global minimizer in an ultra-high dimensional setting for a class of nonconvex penalties. Chen et al. [2] made use of sparse singular value decomposition (SVD) of the coefficient matrix $\boldsymbol{\beta}$ to propose a regularized reduced-rank regression approach improving predictive accuracy and also facilitating good interpretations. Chen et al. [3] proposed an adaptive nuclear norm penalization approach for low-rank matrix approximation, and then used it to develop a new reduced-rank estimation method for high-dimensional multivariate regression. Cook et al. [9] incorporated the idea of envelopes into reduced-rank regression by proposing a reduced-rank envelope model, which has a total number of parameters to be no more than either of the reduced-rank regression or the envelope regression. The reduced-rank envelope estimator is at least as efficient as the two estimators mentioned above, but it is not sparse.

In many regression problems, we are often interested in finding important predictor variables for predicting the response variable, where each predictor variable may be represented by a group of derived input variables. For this reason, Yuan and Lin [26] proposed model selection and estimation in a general regression problem with grouped variables in terms of LASSO penalty. Nardi and Rinaldo [18] established asymptotic properties of the group LASSO estimator for general linear models. Zhao and Yu [27] studied model selection consistency of LASSO in the classical fixed p setting as well as in the setting when p grows with sample size n. For the classical linear regression model, Zou and Zhang [29] studied the model selection and estimation when the number of parameters diverges with sample size, in terms of the adaptive elastic-net penalty function. Guo et al. [13] established the oracle property of the group SCAD estimator in linear regression model under high-dimensional setting when the number of groups grows at a certain polynomial rate. Su et al. [24] proposed a sparse envelope model that performs response variable selection efficiently under the envelope model. In their model, it is assumed that the number of predictors p is fixed and is smaller than the sample size n, but r can be greater than n.

In the present work, we propose a sparse reduced-rank regression method based on the envelope model with adaptive group LASSO for multivariate linear model, which performs the tasks of dimension reduction of response and predictor variables, as well as group variable selection simultaneously. The proposed method is suitable for all n and p. Moreover, the cases when r and p are fixed, and r and p grow simultaneously with n, are also considered. We then establish the consistency, asymptotic normality and oracle property of the envelope-based sparse reduced-rank regression estimation developed here. Finally, with the use of Monte Carlo simulation studies as well as two datasets, we demonstrate that the method developed here displays good variable selection and prediction performance as compared to some well-known existing methods.

2. Envelope-based sparse reduced-rank regression estimator and its properties

From model (3), with $\boldsymbol{\beta}$ having a low rank structure, we have

$$\boldsymbol{\Gamma}^T\boldsymbol{Y} = \boldsymbol{\eta}\boldsymbol{B}\boldsymbol{X} + \boldsymbol{\Gamma}^T\boldsymbol{\varepsilon}, \tag{5}$$

where $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{B}$, $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\gamma}$ and $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{B}$ represent the reduced-rank method, the standard envelope method and the reduced-rank envelope method, respectively. Also, $\boldsymbol{\eta} \in \mathbb{R}^{u\times d}$, $u \geq d$, denotes the coordinates of $\boldsymbol{A}$ with respect to $\boldsymbol{\Gamma}$. If $\boldsymbol{\Gamma}$ is unknown, we can obtain an estimator $\hat{\boldsymbol{\Gamma}}$ of $\boldsymbol{\Gamma}$ by using the method described in Section 3. Then, the standard envelope estimator is obtained as

$$\hat{\boldsymbol{\beta}}_{en} = \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}. \tag{6}$$

Using singular value decomposition, we have

$$\boldsymbol{\eta}\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T, \tag{7}$$

where $\boldsymbol{U} \in \mathbb{R}^{u\times d}$ and $\boldsymbol{V} \in \mathbb{R}^{p\times d}$ are both rank-d matrices with orthogonal columns, and $\boldsymbol{D}$ is a $d \times d$ nonnegative diagonal matrix. Then, (5) can be re-expressed as

$$\boldsymbol{\Gamma}^T\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{F}\boldsymbol{X} + \boldsymbol{\epsilon}, \tag{8}$$

where $\boldsymbol{F} = \boldsymbol{D}\boldsymbol{V}^T$ and $\boldsymbol{\epsilon} = \boldsymbol{\Gamma}^T\boldsymbol{\varepsilon}$. Next, we consider the optimization over $\boldsymbol{F}$. Because $\boldsymbol{U}$ has orthogonal columns, let $\boldsymbol{U}^\perp$ be any column orthogonal matrix such that $(\boldsymbol{U}\vdots\boldsymbol{U}^\perp)$ is an orthogonal matrix. We then have

$$\|\boldsymbol{\Gamma}^T\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{F}\boldsymbol{X}\|_F^2 = \|(\boldsymbol{U}\vdots\boldsymbol{U}^\perp)^T(\boldsymbol{\Gamma}^T\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{F}\boldsymbol{X})\|_F^2 = \|\boldsymbol{U}^T\boldsymbol{\Gamma}^T\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{X}\|_F^2 + \|(\boldsymbol{U}^\perp)^T\boldsymbol{\Gamma}^T\boldsymbol{Y}\|_F^2. \tag{9}$$

Note that the second term on the right side of (9) does not include $\boldsymbol{F}$ and $\boldsymbol{U}^\perp$ only exists in the second term, and so the choice of $\boldsymbol{U}^\perp$ does not matter. If $\boldsymbol{\Gamma}$ and $\boldsymbol{U}$ are given, we can then achieve an estimator of $\boldsymbol{F}$, by minimizing the objective function

$$\begin{aligned} Q(\boldsymbol{F}) &= \frac{1}{2n}\|\boldsymbol{U}^T\boldsymbol{\Gamma}^T\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{X}\|_F^2 + \lambda_n\sum_{j=1}^{p}\omega_j(\|F_j\|_2) \\ &= \frac{1}{2n}\mathrm{tr}((\boldsymbol{U}^T\boldsymbol{\Gamma}^T\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{X})^T(\boldsymbol{U}^T\boldsymbol{\Gamma}^T\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{X})) + \lambda_n\sum_{j=1}^{p}\omega_j(\|F_j\|_2), \end{aligned} \tag{10}$$

where F_j denotes the jth column of $\boldsymbol{F}$, and $\|\cdot\|_2$ denotes the standard Euclidean norm. Also, λ_n denotes the parameter of penalty function and ω_j is the adaptive weight. Their choices are discussed later in Section 5.

Combining (5) and (8), we have the coefficient matrix to be $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{U}\boldsymbol{F}$. By minimizing the objective function in (10), a sparse estimator $\hat{\boldsymbol{F}}$ is achieved. We then have the following proposition.

Proposition 1. *If $\boldsymbol{\Gamma}$ and $\boldsymbol{U}$ are given, the proposed estimator $\hat{\boldsymbol{\beta}} = \boldsymbol{\Gamma}\boldsymbol{U}\hat{\boldsymbol{F}}$ incorporates the envelope method, the reduced-rank method and the adaptive group LASSO penalty technology.*

Remark 1: $\lambda_n \sum_{j=1}^{p} \omega_j(\|F_j\|_2)$ is the adaptive group LASSO penalty function. By using the penalty function, we may obtain a sparse estimator $\hat{\boldsymbol{F}}$ in which some column vectors of $\hat{\boldsymbol{F}}$ are exactly zero vectors. As the coefficient matrix is $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{U}\boldsymbol{F}$, it follows that some column vectors of the estimator $\hat{\boldsymbol{\beta}}$ (corresponding to those column vectors of $\hat{\boldsymbol{F}}$) are also zero vectors. In this way, we achieve a sparse estimator $\hat{\boldsymbol{\beta}}$. Moreover, if $\boldsymbol{\Gamma}$ and $\boldsymbol{U}$ are given, the estimator $\hat{\boldsymbol{\beta}}$ can have good properties that are consistent with the estimator $\hat{\boldsymbol{F}}$, which is established in Theorems 1 and 3.

Assume that the penalty function $\lambda_n \sum_{j=1}^{p} \omega_j(\|F_j\|_2)$ is equal to 0, as well as $\boldsymbol{U}$ and $\boldsymbol{\Gamma}$ are known. Then, by minimizing the function

$$Q(\boldsymbol{F}) = \mathrm{tr}((\boldsymbol{U}^T\boldsymbol{\Gamma}^T\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{X})^T(\boldsymbol{U}^T\boldsymbol{\Gamma}^T\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{X})), \tag{11}$$

we obtain $\hat{\boldsymbol{F}} = \boldsymbol{U}^T\boldsymbol{\Gamma}^T\boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$. Hence, $\hat{\boldsymbol{\beta}} = \boldsymbol{\Gamma}\boldsymbol{U}\hat{\boldsymbol{F}} = \boldsymbol{\Gamma}\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\Gamma}^T\boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$, which degenerates to the reduced-rank envelope estimator. Furthermore, if $r > u = d$, then $\hat{\boldsymbol{\beta}} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$, which is the standard envelope estimator. If $r = u > d$, then the estimator degenerates to the reduced-rank regression estimator. If $r = u = d$, then the estimator is an ordinary least-squares estimator, which is the case when there is no immaterial information to be reduced.

For $\min_u \|\boldsymbol{\Gamma}^T\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{F}\boldsymbol{X}\|_F^2$ with $\boldsymbol{U}^T\boldsymbol{U} = I_u$, if $\boldsymbol{F}$ is fixed, the optimization of $\boldsymbol{U}$ is an orthogonal Procrustes problem (Chen and Huang [4], Gower and Dijksterhuis [12]). The solution is $\hat{\boldsymbol{U}} = \boldsymbol{U}_*\boldsymbol{V}_*^T$, where $\boldsymbol{U}_*$ and $\boldsymbol{V}_*$ are achieved from the singular value decomposition $\boldsymbol{\Gamma}^T\boldsymbol{Y}\boldsymbol{X}^T\boldsymbol{F}^T = \boldsymbol{U}_*\boldsymbol{D}_*\boldsymbol{V}_*^T$. For fixed $\boldsymbol{U}$, motivated by the works of Yuan and Lin [26] and Chen and Huang [4], we can make use of the subgradient method (Friedman et al. [11]) to derive the optimal solution of (10) as

$$F_j = \frac{1}{X^jX^{jT}}\left(1 - \frac{n\lambda_n\omega_j}{\|\boldsymbol{L}_jX^{jT}\|_2}\right)_+ \boldsymbol{L}_jX^{jT}, \tag{12}$$

where $\boldsymbol{L}_j = \boldsymbol{U}^T\boldsymbol{\Gamma}^T\boldsymbol{Y} - \sum_{i\neq j}^{p} F_iX^i$, X^j and X^i denote the jth and ith rows of $\boldsymbol{X}$, respectively, and the subscript "+" denotes the positive part of a real number.

We now propose the following estimation algorithm by using the subgradient method.

Algorithm:

(a) Get $\hat{\boldsymbol{\Gamma}}$ by the method in Section 3;
(b) Give an initial value for $\boldsymbol{F}$;
(c) By SVD $\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y}\boldsymbol{X}^T\hat{\boldsymbol{F}}^T = \boldsymbol{U}_*\boldsymbol{D}_*\boldsymbol{V}_*^T$, update $\hat{\boldsymbol{U}} = \boldsymbol{U}_*\boldsymbol{V}_*^T$;
(d) Given $\boldsymbol{U} = \hat{\boldsymbol{U}}$, obtain $\hat{\boldsymbol{F}}$ from (12);
(e) Repeat steps (c) and (d) until $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{U}}\hat{\boldsymbol{F}}$ converges, i.e., $\|\hat{\boldsymbol{\beta}}_{new} - \hat{\boldsymbol{\beta}}_{old}\|/\|\hat{\boldsymbol{\beta}}_{old}\| < \nu$, where $\hat{\boldsymbol{\beta}}_{new}$ and $\hat{\boldsymbol{\beta}}_{old}$ denote the newly estimated and previously estimated values, respectively, and ν is the level of tolerance, say, $\nu = 10^{-6}$.

Let $\mathcal{A} = \left\{j : \|F_j\|_2 \neq 0, j \in \{1,\ldots,p\}\right\}$, $\mathcal{A}^c = \left\{j : \|F_j\|_2 = 0, j \in \{1,\ldots,p\}\right\}$, $\widehat{\mathcal{A}} = \left\{j : \|\widehat{F}_j\|_2 \neq 0, j \in \{1,\ldots,p\}\right\}$, $q = |\mathcal{A}|$ and $p - q = |\mathcal{A}^c|$. Without loss of generality, let $\boldsymbol{F} = (\boldsymbol{F}_{\mathcal{A}}, \boldsymbol{F}_{\mathcal{A}^c})$. Then, we have $\boldsymbol{\beta}_{\mathcal{A}^c} = \boldsymbol{\Gamma}\boldsymbol{U}\boldsymbol{F}_{\mathcal{A}^c} = \boldsymbol{0}$.

2.1. Case when r and p are fixed

In this section, we show that the proposed estimator has the oracle property when the sample size n increases, when r and p remain fixed. The following regularity conditions are assumed for establishing the asymptotic properties:

(A1) There exists a positive definite matrix $\boldsymbol{M}$ such that $\boldsymbol{X}\boldsymbol{X}^T/n \to \boldsymbol{M}$, as $n \to \infty$;
(A2) There exists a positive constant C_1 such that $\omega_j \leq C_1$, for all $j \in \mathcal{A}$.

Lemma 1. *Under regularity condition (A1), the model selection in (10) is consistent, that is, $P(\widehat{\mathcal{A}} = \mathcal{A}) \to 1$, as $n \to \infty$, only if for any $j \in \mathcal{A}^c$*

$$\sqrt{n}\lambda_n\omega_j \to \infty.$$

The proof is similar to that of Proposition 3.1 of Nardi and Rinaldo [18], and so we do not present it here.

Theorem 1. *Under regularity conditions (A1) and (A2), suppose $\hat{\mathbf{\Gamma}}$ and $\hat{\boldsymbol{U}}$ are $\sqrt{n}$-consistent estimators of $\mathbf{\Gamma}$ and $\boldsymbol{U}$, respectively, and that the errors are normally distributed. If $\sqrt{n}\lambda_n \to 0$ as $n \to \infty$, then there exists a local minimizer $\hat{\boldsymbol{F}}$ of $Q(\boldsymbol{F})$ such that $\hat{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$, that is, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F = O_p\left(n^{-1/2}\right)$, and that this $\hat{\boldsymbol{\beta}}$ must satisfy*

(a) Sparsity: $P(\hat{\boldsymbol{\beta}}_{\mathcal{A}^c} = 0) \to 1$,

(b) Asymptotic normality: $\sqrt{n}(vec(\hat{\boldsymbol{\beta}}_{\mathcal{A}}) - vec(\boldsymbol{\beta}_{\mathcal{A}})) \xrightarrow{D} N(0, \mathbf{\Sigma}_{\beta_{\mathcal{A}}})$,

where $\mathbf{\Sigma}_{\beta_{\mathcal{A}}}$ is the upper-left $pq \times pq$ block of $\mathbf{\Sigma}_{RE}$, which is the asymptotic covariance matrix of the reduced-rank envelope estimator [Proposition 8 of Cook et al. [9]].

2.2. Case when p and r grow with n

In this section, we show the consistency of model selection when p, r and q increase with the sample size n. For this reason, in the following, we use p_n, r_n and q_n instead of p, r and q to indicate that they can grow with n. Assume the following regularity conditions:

(B1) There exists a positive constant C_2 such that $\frac{1}{n}X^j X^{jT} \leq C_2$, for all $j \in \{1, \ldots, p_n\}$, and all n;

(B2) There exists a positive constant C_3 such that $\alpha^T \boldsymbol{R}_{11}\alpha \geq C_3$, for all $\|\alpha\|_2^2 = 1$, where $\boldsymbol{R}_{11} = \frac{\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^T}{n}$;

(B3) $q_n = O_p(n^{c_1})$ for some $0 < c_1 < 1$;

(B4) There exist positive constants c_2 and C_4 such that $c_1 < c_2 \leq 1$ and $n^{(1-c_2)/2} \min_{j \in \mathcal{A}} \|\beta_j\|_2 \geq C_4$.

The regularity conditions (B1)–(B4) stated above were first used by Zhao and Yu [27] for establishing the model selection consistency of the LASSO estimator, and by Kim et al. [14] for showing the oracle property of the SCAD estimator. The same conditions were also used by Guo et al. [13] for discussing the oracle property of the group SCAD estimator under the high-dimensional setting where the number of groups can grow at a certain polynomial rate.

To show the consistency of model selection when p and r increase with sample size n, we first introduce the following theorem.

Theorem 2. *Suppose $E(\|\boldsymbol{\varepsilon}\|_F)^{2k} < \infty$ for an integer $k \geq 1$ and $p_n(\sqrt{n}\lambda_n\omega)^{-2k} \to 0$, where $\omega = \min_{j \in \mathcal{A}^c} \omega_j$. Let $\boldsymbol{F} = (\boldsymbol{F}_{\mathcal{A}}, \boldsymbol{0})$, and define*

$$\hat{\boldsymbol{F}}_{\mathcal{A}} = \arg\min_{\boldsymbol{F}} \left\{ \frac{1}{2n}\|\hat{\boldsymbol{U}}^T\hat{\Gamma}^T\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{X}_{\mathcal{A}}\|_F^2 + \lambda_n \sum_{j \in \mathcal{A}} \omega_j(\|F_j\|_2) \right\}. \tag{13}$$

Then, with probability tending to 1, $(\hat{\boldsymbol{F}}_{\mathcal{A}}, \boldsymbol{0})$ is the solution of (10).

Theorem 3. *Under the regularity conditions (B1)–(B4), provided $\lambda_n = o(n^{-(1-c_2+c_1)/2})$ and $r_n = o(n^{c_2/2})$, then $P(\{j : \|\hat{\beta}_j\|_2 \neq 0\} = \mathcal{A}) \to 1$.*

Remark 2: When $\boldsymbol{\varepsilon}$ has all moments, p_n is allowed to grow much faster than n (up to n^θ, for any $\theta > 0$). Moreover, r_n can grow slower than $n^{c_2/2}$.

3. Estimation of the envelope

To achieve the estimation of the envelope $\xi_{\mathbf{\Sigma}}(\mathcal{B})$, Cook et al. [8] and Su et al. [24] developed an iterative algorithm which is fast and effective. Let

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_1 \\ \mathbf{\Gamma}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_u \\ \boldsymbol{N} \end{pmatrix} \mathbf{\Gamma}_1 \triangleq \boldsymbol{Q}_N \mathbf{\Gamma}_1,$$

where $\mathbf{\Gamma}_1$ consists of the first u rows of $\mathbf{\Gamma}$, and suppose it is nonsingular, and $\boldsymbol{N}$ represents $\mathbf{\Gamma}_2\mathbf{\Gamma}_1^{-1}$ which depends on $\mathbf{\Gamma}$ only through the space formed by the column vectors of $\mathbf{\Gamma}$. This is so because, for any orthogonal matrix $\boldsymbol{P} \in \mathbb{R}^{u \times u}$, if $\mathbf{\Gamma}^* = \mathbf{\Gamma}\boldsymbol{P}$, then $\mathbf{\Gamma}_1^* = \mathbf{\Gamma}_1\boldsymbol{P}$, $\mathbf{\Gamma}_2^* = \mathbf{\Gamma}_2\boldsymbol{P}$, and $\boldsymbol{N}^* = \mathbf{\Gamma}_2\boldsymbol{P}\boldsymbol{P}^{-1}\mathbf{\Gamma}_1^{-1} = \boldsymbol{N}$. The optimization problem estimating $\xi_{\mathbf{\Sigma}}(\mathcal{B})$ is then

$$\hat{\boldsymbol{N}} = \arg\min_{\boldsymbol{N} \in \mathbb{R}^{(r-u) \times u}} -2\log|\boldsymbol{Q}_N^T\boldsymbol{Q}_N| + \log|\boldsymbol{Q}_N(\boldsymbol{Y}\boldsymbol{Q}_X\boldsymbol{Y}^T/n)\boldsymbol{Q}_N| + \log|\boldsymbol{Q}_N(\boldsymbol{Y}\boldsymbol{Y}^T/n)^{-1}\boldsymbol{Q}_N|, \tag{14}$$

where $\boldsymbol{Q}_X = \boldsymbol{I}_n - \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}$. For the convenience of the following statement, let $\hat{\boldsymbol{\Sigma}}_Y$ and $\hat{\boldsymbol{\Sigma}}_{\text{res}}$ denote $\boldsymbol{Y}\boldsymbol{Y}^T/n$ and $\boldsymbol{Y}\boldsymbol{Q}_X\boldsymbol{Y}^T/n$, respectively.

If $n > r + p$, it follows that $\text{rank}(\hat{\boldsymbol{\Sigma}}_{\text{res}}) = \text{rank}(\hat{\boldsymbol{\Sigma}}_Y) = r$ with probability 1. Therefore, $\boldsymbol{Q}_N\hat{\boldsymbol{\Sigma}}_{\text{res}}\boldsymbol{Q}_N$ and $\boldsymbol{Q}_N\hat{\boldsymbol{\Sigma}}_Y^{-1}\boldsymbol{Q}_N$ are nonsingular. But, if $p > n$, then $\hat{\boldsymbol{\Sigma}}_{\text{res}}$ is singular. If $r > n$, then both $\hat{\boldsymbol{\Sigma}}_{\text{res}}$ and $\hat{\boldsymbol{\Sigma}}_Y$ are singular. In both these cases, optimization in (14) is not solvable as it depends on the inverse of $\hat{\boldsymbol{\Sigma}}_Y$. At the same time, the optimization algorithm for solving (14) needs the inverse of $\hat{\boldsymbol{\Sigma}}_{\text{res}}$. But, $\boldsymbol{\Sigma}_{\text{res}}^{-1}$ and $\boldsymbol{\Sigma}_Y^{-1}$ can be directly estimated by using methods such as positive definite estimators of large covariance matrices (Rothman [21]) and sparse permutation invariant covariance estimation (Rothman et al. [22]). Yet another suitable method is to use the ridge-type covariance estimators proposed by Ledoit and Wolf [16] for $\boldsymbol{\Sigma}_{\text{res}}$ and $\boldsymbol{\Sigma}_Y$. In this work, we use positive definite estimators of large covariance matrices and sparse permutation invariant covariance estimation for estimating $\hat{\boldsymbol{\Sigma}}_{\text{res}}^{-1}$ and $\hat{\boldsymbol{\Sigma}}_Y^{-1}$, respectively. Once we obtain $\hat{\boldsymbol{N}}$, then we have $\xi_{\boldsymbol{\Sigma}}(\mathcal{B}) = \text{span}(\hat{\boldsymbol{Q}}_N)$.

4. Selection of u

In the above discussion, we have assumed that u, the dimension of the envelope, is known. In practice, however, u will be unknown. There are a few ways to choose u such as cross-validation (CV), likelihood-ratio test (LRT) and information criterion such as AIC or BIC. Cook [7] has provided an elaborate discussion on all these methods. The AIC tends to select a model that contains the true model, and so it tends to overestimate u. The BIC tends to select the correct u with probability getting close to 1 as n goes to ∞, but, it can be slow to respond in case of small samples. The LRT method performs the best in case of small samples, but asymptotically the error probability is equal to the significance level. The cross-validation method tends to balance bias and variance when selecting u, which may lead to choices that are different from those provided by LRT and information criteria. Here, we use the cross-validation method for selecting u.

5. Tuning

The rank d can be selected by cross-validation (CV). The parameter of adaptive LASSO penalty function is denoted by λ_n. We set $\omega_j = 1/\|\beta^j\|_2^\delta$ as the adaptive weight. Let $\tilde{\omega}_j = 1/\|\tilde{\beta}^j\|_2^\delta$ be an estimator of ω, where $\tilde{\beta}^j$ is a consistently estimated value of β^j. When $n \geq p$, the reduced-rank envelope method can be used to estimate β^j. When $p > n$, by setting ω_j's all equal to ω_n, a reasonable estimator can be the solution of (10) with single penalty parameter $\lambda_n\omega_n$. In this paper, we use fivefold CV procedure to estimate λ_n. The fivefold CV procedure is as follows: Let $\mathcal{D}$ denote the full dataset, as well as $\mathcal{D} - \mathcal{D}^\tau$ and $\mathcal{D}^\tau$ denote training and test set, respectively, $\tau = 1, \ldots, 5$. For each λ_n and τ, we derive the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ using the training set $\mathcal{D} - \mathcal{D}^\tau$. The fivefold CV criterion is defined as

$$\text{CV}(\lambda_n) = \sum_{\tau=1}^{5} \sum_{(\boldsymbol{Y}_t, \boldsymbol{X}_t) \in \mathcal{D}^\tau} \left\| \boldsymbol{Y}_t - \hat{\boldsymbol{\beta}}^{(\tau)}(\lambda_n)\boldsymbol{X}_t \right\|_F^2.$$

We obtain a $\hat{\lambda}_n$ by minimizing $\text{CV}(\lambda_n)$.

6. Simulation study

6.1. Simulation Setups and Methods

Scenario I. We generated data with p and r being smaller than n, taking $\boldsymbol{\Omega} = \boldsymbol{I}_u$ and $\boldsymbol{\Omega}_0 = 10\boldsymbol{I}_{r-u}$. We assumed that elements of the first s columns in $\boldsymbol{\eta}$ were independent uniform (0, 10) variables, and the remaining elements of $p - s$ columns were all zeros. Then, $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, X_i follows multivariate normal distribution with mean 0 and covariance matrix $\boldsymbol{I}_p$, and $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ was obtained by standardizing an $r \times r$ matrix of independent uniform (0, 1) variables. The error covariance matrix was generated from $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$. The prediction mean squared error (PMSE) is then defined as

$$\text{PMSE} = E\|\hat{\boldsymbol{\beta}}\boldsymbol{X} - \boldsymbol{Y}\|^2/nr. \tag{15}$$

Table 1: Prediction comparisons of these methods based on PMSE using 100 simulation runs with p and r being smaller than n. PMSE denotes prediction mean squared error. OLS, ENV and ENRRR denote the ordinary least-squares estimator, standard envelope estimator and envelope-based reduced rank regression estimator, respectively. Further, SPLS denotes the sparse partial least-squares estimator, SRRR and aSRRR denote the sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively; ENSRRR and aENSRRR denote the envelope-based sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively.

	n=200, u = 10, s = 20, d = 5			
PMSE	p=30, r=20	p=50, r=30	p=100, r=50	p=50, r=100
OLS	1.661	2.331	4.142	3.074
ENV	1.474	1.843	2.331	2.365
ENRRR	1.440	1.780	2.163	2.341
SPLS	1.841	2.148	2.245	2.476
SRRR	1.465	1.821	2.238	2.402
aSRRR	1.459	1.798	2.129	2.367
ENSRRR	1.438	1.775	2.168	2.347
aENSRRR	1.436	1.771	2.127	2.335

Table 2: Variable selection comparisons of these methods based on ACR using100 simulation runs with p and r being smaller than n. ACR denotes the average correct ratio between the number of correct selection and the total number of relevant variables. SPLS denotes the sparse partial least-squares estimator, SRRR and aSRRR denote the sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively; ENSRRR and aENSRRR denote the envelope-based sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively.

	n=200, u = 10, s = 20, d = 5			
ACR	p=30, r=20	p=50, r=30	p=100, r=50	p=50, r=100
SPLS	0.79	0.65	0.94	0.82
SRRR	0.67	0.60	0.20	0.40
aSRRR	1.00	0.99	0.96	0.98
ENSRRR	0.69	0.65	0.30	0.55
aENSRRR	1.00	0.99	0.97	0.99

Table 3: Prediction comparisons of these methods based on PMSE using 100 simulation runs with p and r being greater than n. PMSE denotes prediction mean squared error. SPLS denotes the sparse partial least-squares estimator, SRRR and aSRRR denote the sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively; ENSRRR and aENSRRR denote the envelope-based sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively.

	$n = 60, p = 70, r = 70, d = 5$		$n = 100, p = 150, r = 150, d = 5$	
PMSE	$u = 10, s = 20$	$u = 20, s = 40$	$u = 10, s = 20$	$u = 20, s = 40$
SPLS	3.324	5.435	2.472	3.090
SRRR	2.687	4.915	2.525	3.173
aSRRR	2.658	4.055	2.506	3.044
ENSRRR	2.524	4.764	2.392	2.950
aENSRRR	2.446	3.882	2.381	2.930

Table 4: Variable selection comparisons of these methods based on ACR based on 100 simulation runs with p and r being greater than n. ACR denotes the average correct ratio between the number of correct selection and the total number of relevant variables. SPLS denotes the sparse partial least-squares estimator, SRRR and aSRRR denote the sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively; ENSRRR and aENSRRR denote the envelope-based sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively.

	$n = 60, p = 70, r = 70$		$n = 100, p = 150, r = 150$	
ACR	$u = 10, s = 20$	$u = 20, s = 40$	$u = 10, s = 20$	$u = 20, s = 40$
SPLS	0.83	0.68	0.96	0.76
SRRR	0.77	0.76	0.93	0.68
aSRRR	0.97	0.97	0.97	0.86
ENSRRR	0.82	0.80	0.95	0.70
aENSRRR	0.98	0.98	0.97	0.88

We compared prediction accuracy of all the methods in terms of PMSE. We also compared the accuracy of variable selection of these methods in terms of average correct ratio (ACR) between the number of correct selection and the total number of relevant variables, which measures the ability of selecting relevant variables.

Scenario II. We generated data with p and r being greater than n, taking $\mathbf{\Omega} = \boldsymbol{I}_u$ and $\mathbf{\Omega}_0 = 10\boldsymbol{I}_{r-u}$. We assumed that elements of the first s columns in $\boldsymbol{\eta}$ were independent uniform (0, 10) variables, and the remaining elements of $p - s$ colomns were all zeros. Then, $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$, X_i follows multivariate normal distribution with mean being 0 and covariance matrix $0.1\boldsymbol{I}_p$, and $(\mathbf{\Gamma},\ \mathbf{\Gamma}_0)$ was obtained by standardizing an $r \times r$ matrix of independent uniform (0, 1) variables. The error covariance matrix was generated by $\mathbf{\Sigma} = \mathbf{\Gamma\Omega\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$.

6.2. Simulation Results

When p and r are smaller than n, the results in Table 1 show that all methods outperform least-squares estimator in terms of prediction mean squared error. But, the PMSEs of aENSRRR are the smallest compared to all other estimators in all the cases considered. From the variable selection viewpoint, OLS, ENV and ENRRR have no variable selection provision. Table 2 shows that aENSRRR and aSRRR achieve the best performance in terms of the average correct ratio among all the methods considered, and the two methods are more stable and accurate when parameters change. Also, the technique with adaptive group LASSO can identify almost all correct zero groups.

When p and r are greater than n, since the OLS, ENV and ENRRR methods do not exist, we compare other methods in terms of PMSE and ACR. Table 3 shows that the PMSEs of aENSRRR are still the smallest and are also the most stable ones in all the cases considered when the parameters become larger. Similarly, Table 4 shows that the performance of aENSRRR based on ACR is still the best among all the methods considered. These results demonstrate the proposed method possesses good stability, good variable selection and prediction performance compared to some existing methods.

Table 5: Prediction comparisons of these methods based on PMSE using data split at random 100 times. PMSE denotes prediction mean squared error. OLS, ENV and ENRRR denote the ordinary least-squares estimator, standard envelope estimator and envelope-based reduced rank regression estimator, respectively. Further, SPLS denotes the sparse partial least-squares estimator, SRRR and aSRRR denote the sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively; ENSRRR and aENSRRR denote the envelope-based sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively.

	OLS	ENV	ENRRR	SPLS	SRRR	aSRRR	ENSRRR	aENSRRR
PMSE	0.534	0.512	0.490	0.416	0.466	0.415	0.401	0.399

7. Real-life Examples

7.1. Example 1: Yeast cell cycle data

A yeast cell cycle data set was first used by Spellman et al. [23], which is available in the R package spls. The response matrix $\boldsymbol{Y}$ consists of 542 cell-cycle-regulated genes. The cell cycle was measured by taking RNA levels on genes at 18 time points using the α-factor arrest method. The 542×106 predictor matrix $\boldsymbol{X}$ contains the binding information of the target genes for a total of 106 transition factors (TFs). This data set has been analyzed by some other authors including Chun and Keleş [6], Chen and Huang [4], Kong et al. [15] and Zhu and Su [28] in the context of reduced-rank regression. Our main goal here is to identify the TFs that contribute to the variation of the RNA transcript levels in cell cycles. We utilize approximately 2/3 of the data as training set and the remaining as testing set, and also repeat such splitting at random 100 times. In this case, we have $n = 360$, $r = 18$ and $p = 106$ in the training dataset. We centered and scaled both the predictor matrix $\boldsymbol{X}$ and response matrix $\boldsymbol{Y}$. By using fivefold cross-validation, we selected the number of factors $d = 4$ for SRRR, aSRRR, ENSRRR and aENSRRR. Similarly, the dimension of the envelope, u, was selected to be 6 by fivefold CV. For SPLS, we selected $K = 8$ as the number of hidden components.

To compare prediction accuracy of the methods, we use the training dataset to build models, and then use the testing dataset to assess the models. Table 5 shows average prediction errors from 100 random splits. From Table 5, we can see that OLS performs poorly in this case, and the proposed aENSRRR method has the lowest prediction error among all the methods considered. To compare the stability of selection of variables, we calculated numbers of selected predictors in the 100 splits and medians for all the methods. As Table 6 shows, the numbers of selected predictors of the aENSRRR method range from 48 to 81, and the fluctuation difference is 33, which is similar to those of SPLS and aSRRR methods. The performance of SPLS, aSRRR, and aENSRRR methods are all similar, and better than those of the other two methods in terms of stability of variable selection in this case.

Table 6: Variable selection comparisons of these methods based on MNSP and RNSP using data split at random 100 times. MNSP and RNSP denote median and range of the numbers of selected predictors in the 100 splits, respectively. SPLS denotes the sparse partial least-squares estimator, SRRR and aSRRR denote the sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively; ENSRRR and aENSRRR denote the envelope-based sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively.

	SPLS	SRRR	aSRRR	ENSRRR	aENSRRR
MNSP	30	77	64	76	64
RNSP	[19, 53]	[45, 89]	[46, 80]	[47, 88]	[48, 81]

7.2. Example 2: Breast cancer data

In this section, we consider a breast cancer dataset from Chin et al. [5], which consists of gene expression and DNA copy number measurements with 89 samples. The dataset is available in the R package PMA. This dataset has been earlier used by Witten et al. [25] and Chen et al. [3]. Peng et al. [19] showed that some types of cancer have the characteristics of abnormal alterations of DNA copy number. Our goal here is to identify the relationship between DNA copy numbers and the RNA expression levels. It is meaningful to regress gene expression profile on copy number changes, because amplification or deletion of DNA part corresponding to a given gene may lead to

corresponding increase or decrease of gene expression (Chen et al. [3]). In this case, we analyze chromosome 21 in which we have $r = 227$, $p = 44$ and $n = 89$. We centered and scaled both $\boldsymbol{X}$ and $\boldsymbol{Y}$. For comparison of prediction accuracy and variable selection performance, the data were randomly split into a training set of size 70 and a test set of size 19. As in Example 1, by using five-fold CV, we selected the parameters as $d = 3$, $u = 4$ and $K = 1$. From Table 7, we observe that the proposed aENSRRR method still performs the best as compared to all others in terms of prediction accuracy in this case. Moreover, as Table 8 reveals, SPLS and aENSRRR methods have similar performance in terms of stability of variable selection, being better than all other methods in this case.

Table 7: Prediction comparisons of these methods based on PMSE using data split at random 100 times. PMSE denotes prediction mean squared error. OLS, ENV and ENRRR denote the ordinary least-squares estimator, standard envelope estimator and envelope-based reduced rank regression estimator, respectively. Further, SPLS denotes the sparse partial least-squares estimator, SRRR and aSRRR denote the sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively; ENSRRR and aENSRRR denote the envelope-based sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively.

	OLS	ENV	ENRRR	SPLS	SRRR	aSRRR	ENSRRR	aENSRRR
PMSE	0.751	0.278	0.271	0.258	0.268	0.262	0.257	0.249

Table 8: Variable selection comparisons of these methods based on MNSP and RNSP using data split at random 100 times. MNSP and RNSP denote median and range of the numbers of selected predictors in the 100 splits, respectively. SPLS denotes the sparse partial least-squares estimator, SRRR and aSRRR denote the sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively; ENSRRR and aENSRRR denote the envelope-based sparse reduced-rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively.

	SPLS	SRRR	aSRRR	ENSRRR	aENSRRR
MNSP	28	26	6	27	8
RNSP	[19, 32]	[16, 37]	[3, 23]	[18, 36]	[5, 17]

Appendix

Proof of Theorem 1: Let $\alpha_n = n^{-\frac{1}{2}} + a_n$. It is then sufficient to show that, for any given υ, there exists a large constant C such that

$$P\left\{\inf_{\|\boldsymbol{W}\|_F=C} Q(\boldsymbol{F}+\alpha_n\boldsymbol{W}) \geq Q(\boldsymbol{F})\right\} \geq 1-\upsilon, \tag{16}$$

where $\boldsymbol{W}$ is a $r \times p$ constant matrix. This implies, with probability at least $1-\upsilon$, that there exists a local minimum in the ball $\{\boldsymbol{F}+\alpha_n\boldsymbol{W} : \|\boldsymbol{W}\|_F \leq C\}$. Hence, there exists a local minimizer such that $\|\hat{\boldsymbol{F}}-\boldsymbol{F}\|_F = O_p(\alpha_n)$. Let

$$D_n(\boldsymbol{W}) = Q(\boldsymbol{F}+\alpha_n\boldsymbol{W}) - Q(\boldsymbol{F}). \tag{17}$$

We then have

$$\begin{aligned} D_n(\boldsymbol{W}) &= \frac{1}{2n}\mathrm{tr}((\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y} - (\boldsymbol{F}+\alpha_n\boldsymbol{W})\boldsymbol{X})^T(\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y} - (\boldsymbol{F}+\alpha_n\boldsymbol{W})\boldsymbol{X})) \\ &\quad -\frac{1}{2n}\mathrm{tr}((\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{X})^T(\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{X})) + \lambda_n\sum_{j=1}^{p}\omega_j(\|F_j+\alpha_n W_j\|_2 - \|F_j\|_2), \end{aligned} \tag{18}$$

where W_j denotes the jth column of $\boldsymbol{W}$. By simple calculation, we obtain

$$\frac{1}{2n}\text{tr}((\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y}-(\boldsymbol{F}+\alpha_n\boldsymbol{W})\boldsymbol{X})^T(\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y}-(\boldsymbol{F}+\alpha_n\boldsymbol{W})\boldsymbol{X}))-\frac{1}{2n}\text{tr}((\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y}-\boldsymbol{F}\boldsymbol{X})^T(\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y}-\boldsymbol{F}\boldsymbol{X}))$$
$$=\frac{\alpha_n^2}{2n}\text{tr}(\boldsymbol{X}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X})-\frac{\alpha_n}{n}\text{tr}(\boldsymbol{X}^T\boldsymbol{W}^T(\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y}-\boldsymbol{F}\boldsymbol{X})). \tag{19}$$

Under regularity condition (A1), we know that $\boldsymbol{X}\boldsymbol{X}^T/n = O(1)$, and so

$$\frac{\alpha_n^2}{2n}\text{tr}(\boldsymbol{X}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}) = O(\alpha_n^2C^2) = O(n^{-1/2}\alpha_nC^2). \tag{20}$$

As we assume $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{U}}$ to be $\sqrt{n}$-consistent estimators of $\boldsymbol{\Gamma}$ and $\boldsymbol{U}$, respectively, and $\boldsymbol{\varepsilon}\boldsymbol{X}^T/\sqrt{n} = O_p(1)$, it follows that $(\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y}-\boldsymbol{F}\boldsymbol{X})\boldsymbol{X}^T/n = O_p(n^{-1/2})$, and so

$$\frac{\alpha_n}{n}\text{tr}(\boldsymbol{X}^T\boldsymbol{W}^T(\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y}-\boldsymbol{F}\boldsymbol{X})) = O_p(n^{-1/2}\alpha_nC). \tag{21}$$

The first term dominates the second term on the RHS of (19) by choosing a sufficiently large C. Next, let $D_3 = \lambda_n\sum_{j=1}^p\omega_j(\|F_j+\alpha_nW_j\|_2-\|F_j\|_2)$. Now, upon using Cauchy-Schwarz inequality and regularity condition (A2), we obtain

$$\begin{aligned}
D_3 &= \lambda_n\sum_{j=1}^p\omega_j(\|F_j+\alpha_nW_j\|_2-\|F_j\|_2)\geq\lambda_n\sum_{j\in\mathcal{A}}\omega_j(\|F_j+\alpha_nW_j\|_2-\|F_j\|_2)\\
&\geq -\lambda_n\sum_{j\in\mathcal{A}}\omega_j(\|F_j+\alpha_nW_j-F_j\|_2)\geq-\lambda_n\left(\max_{j\in\mathcal{A}}\omega_j\right)\alpha_n\sqrt{q}\|\boldsymbol{W}\|_F\\
&= -\sqrt{n}\lambda_n\left(\max_{j\in\mathcal{A}}\omega_j\right)\frac{\alpha_n}{\sqrt{n}}\sqrt{q}\|\boldsymbol{W}\|_F\geq-O_p(n^{-1/2}\alpha_nC).
\end{aligned}\tag{22}$$

It then follows that D_3 is also dominated by $\frac{\alpha_n^2}{2n}\text{tr}(\boldsymbol{X}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X})$ for a sufficiently large C. Upon combining (18)–(22), we get

$$D_n(\boldsymbol{W})\geq O(n^{-1/2}\alpha_nC^2)-O_p(n^{-1/2}\alpha_nC)-O_p(n^{-1/2}\alpha_nC). \tag{23}$$

Thus, by choosing a sufficiently large C, (16) holds true; that is, $\hat{\boldsymbol{F}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{F}$. Moreover, $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{U}}$ are $\sqrt{n}$-consistent estimators of $\boldsymbol{\Gamma}$ and $\boldsymbol{U}$, respectively. As $\hat{\boldsymbol{\beta}}$=$\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{U}}\hat{\boldsymbol{F}}$, then $\hat{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$. Next, we will establish that this $\hat{\boldsymbol{\beta}}$ has sparsity and asymptotic normality properties stated in (a) and (b). In fact, if (a) is true, by the oracle property of adaptive LASSO penalty function, the asymptotic normality of $\hat{\boldsymbol{\beta}}$ can be directly deduced from the asymptotic normality property of the reduced-rank envelope estimator (Cook et al. [9]). Therefore, we only need to prove that $\hat{\boldsymbol{\beta}}$ has sparsity, the property in (a). In the following, we assume that $\|\hat{F}_j\|_2\neq0$, for some $j\in\mathcal{A}^c$. Then, we have

$$\frac{1}{\sqrt{n}}(\boldsymbol{U}^T\boldsymbol{T}^T\boldsymbol{Y}-\hat{\boldsymbol{F}}\boldsymbol{X})X^{jT} = \sqrt{n}\lambda_n\omega_j\frac{\hat{F}_j}{\|\hat{F}_j\|_2}. \tag{24}$$

Further,

$$\|\frac{1}{\sqrt{n}}(\boldsymbol{U}^T\boldsymbol{T}^T\boldsymbol{Y}-\hat{\boldsymbol{F}}\boldsymbol{X})X^{jT}\|_2 = \sqrt{n}\lambda_n\omega_j. \tag{25}$$

The LHS of (25) is equal to $O_p(1)$, which implies that $\sqrt{n}\lambda_n\omega_j = O_p(1)$, which is in contradiction with $\sqrt{n}\lambda_n\omega_j\to\infty$ in Lemma 1. Thus, $P(\|\hat{F}_j\|_2=0)\to1$, for all $j\in\mathcal{A}^c$. It then follows that $\hat{\boldsymbol{\beta}}_{\mathcal{A}^c}=\boldsymbol{0}$. This completes the proof of the theorem. □

Proof of Theorem 2: To simplify the proof, let us use the notations $\hat{\boldsymbol{H}}_{\mathcal{A}} = \hat{\boldsymbol{F}}_{\mathcal{A}} - \boldsymbol{F}_{\mathcal{A}}$, $\boldsymbol{K}(1) = \frac{\boldsymbol{\varepsilon}\boldsymbol{X}_{\mathcal{A}}^T}{\sqrt{n}}$, $K^j(2) = \frac{\boldsymbol{\varepsilon}X^{jT}}{\sqrt{n}}$, $R_{12}^j = \frac{\boldsymbol{X}_{\mathcal{A}}X^{jT}}{n}$, $j \in \mathcal{A}^c$, and $\boldsymbol{S} = \left[\omega_1 \frac{\hat{F}_1}{\|\hat{F}_1\|_2}, \ldots, \omega_{q_n} \frac{\hat{F}_{q_n}}{\|\hat{F}_{q_n}\|_2}\right]$.
By the definition of $\hat{\boldsymbol{F}}_{\mathcal{A}}$, it is sufficient to show that

$$P\left(\forall j \in \mathcal{A}^c | \, \|\frac{1}{n}(\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y} - \hat{\boldsymbol{F}}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}})X^{jT}\|_2 \leq \lambda_n\omega_j\right) \to 1,$$

which is equivalent to showing that

$$P\left(\exists j \in \mathcal{A}^c | \, \|\frac{1}{n}(\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y} - \hat{\boldsymbol{F}}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}})X^{jT}\|_2 > \lambda_n\omega_j\right) \to 0. \tag{26}$$

Because $\boldsymbol{F} = (\boldsymbol{F}_{\mathcal{A}}, \boldsymbol{0})$, $\hat{\boldsymbol{H}}_{\mathcal{A}} = \hat{\boldsymbol{F}}_{\mathcal{A}} - \boldsymbol{F}_{\mathcal{A}}$, $K^j(2) = \frac{\boldsymbol{\varepsilon}X^{jT}}{\sqrt{n}}$, $R_{12}^j = \frac{\boldsymbol{X}_{\mathcal{A}}X^{jT}}{n}$, $j \in \mathcal{A}^c$, (26) can be re-expressed as

$$P\left(\exists j \in \mathcal{A}^c | \, \| \sqrt{n}\hat{\boldsymbol{H}}_{\mathcal{A}}R_{12}^j - \hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T K^j(2)\|_2 > \sqrt{n}\lambda_n\omega_j\right) \to 0. \tag{27}$$

Note that, by the definition of $\hat{\boldsymbol{F}}_{\mathcal{A}}$, we have

$$-\frac{1}{n}(\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{Y} - \hat{\boldsymbol{F}}\boldsymbol{X}_{\mathcal{A}})X^{jT} + \lambda_n\omega_j\frac{\hat{F}_j}{\|\hat{F}_j\|_2} = \boldsymbol{0} \Leftrightarrow \frac{1}{n}(\hat{\boldsymbol{F}}\boldsymbol{X}_{\mathcal{A}} - \boldsymbol{F}\boldsymbol{X}_{\mathcal{A}} - \hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{\varepsilon})X^{jT} + \lambda_n\omega_j\frac{\hat{F}_j}{\|\hat{F}_j\|_2} = \boldsymbol{0}, \ j \in \mathcal{A}.$$

It then follows that

$$\sqrt{n}\hat{\boldsymbol{H}}_{\mathcal{A}}\boldsymbol{R}_{11} - \hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{K}(1) + \sqrt{n}\lambda_n\boldsymbol{S} = \boldsymbol{0}. \tag{28}$$

Using (28), $\| \sqrt{n}\hat{\boldsymbol{H}}_{\mathcal{A}}R_{12}^j - \hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T K^j(2)\|_2 > \sqrt{n}\lambda_n\omega_j$, $j \in \mathcal{A}^c$, is implied by

$$\|\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{K}(1)\boldsymbol{R}_{11}^{-1}R_{12}^j - \hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T K^j(2)\|_2 \leq \sqrt{n}\lambda_n(\omega_j - \|\boldsymbol{S}\boldsymbol{R}_{11}^{-1}R_{12}^j\|_2), \ j \in \mathcal{A}^c.$$

Let $\zeta^j = \hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\boldsymbol{K}(1)\boldsymbol{R}_{11}^{-1}R_{12}^j - \hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T K^j(2)$. By the regularity conditions (B1) and (B2) and the fact that $E(\|\boldsymbol{\varepsilon}\|_F)^{2k} < \infty$, we obtain $E(\|\zeta^j\|_2)^{2k} < \infty$, $j \in \mathcal{A}^c$. Moreover, for any $t > 0$ and $j \in \mathcal{A}$, we have $P(\|\zeta^j\|_2 > t) = O(t^{-2k})$ by Markov inequality. Consequently, we have

$$P(\exists j \in \mathcal{A}^c | \, \| \sqrt{n}\hat{\boldsymbol{H}}_{\mathcal{A}}R_{12}^j - \hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T K^j(2)\|_2 > \sqrt{n}\lambda_n\omega_j) \leq \sum_{j\in\mathcal{A}^c} P(\| \sqrt{n}\hat{\boldsymbol{H}}_{\mathcal{A}}R_{12}^j - \hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T K^j(2)\|_2 > \sqrt{n}\lambda_n\omega_j)$$

$$\leq \sum_{j\in\mathcal{A}^c} P(\|\zeta^j\|_2 > \sqrt{n}\lambda_n\omega_j) \leq (p_n - q_n)O(\frac{1}{(\sqrt{n}\lambda_n\omega)^{2k}}) = O(\frac{p_n}{(\sqrt{n}\lambda_n\omega)^{2k}}) \to 0,$$

which completes the proof of the theorem. □

Proof of Theorem 3: In Theorem 2, we have proved that the proposed estimator is equal to $(\hat{\boldsymbol{F}}_{\mathcal{A}}, \boldsymbol{0})$ with probability tending to 1. Therefore, to establish consistency of the model selection, it suffices to show that $P(\min_{j\in\mathcal{A}} \|\hat{F}_j\|_2 > 0) \to 1$. Note that $\|\hat{F}_j\|_2 \geq \|F_j\|_2 - \|\hat{F}_j - F_j\|_2$. According to the regularity condition (B4), we have $\min_{j\in\mathcal{A}} \|F_j\|_2 = O(n^{-(1-c_2)/2})$. As $\lambda_n = o(n^{-(1-c_2+c_1)/2})$, it is sufficient to show that $\max_{j\in\mathcal{A}} \|\hat{F}_j - F_j\|_2 \leq o_p(n^{-(1-c_2)/2})$, that is, $\| \sqrt{n}\hat{H}_j\|_2 \leq o_p(n^{c_2/2})$.
From (28), we obtain

$$\| \sqrt{n}\hat{H}_j\|_2 \leq \|\hat{\boldsymbol{U}}^T\hat{\boldsymbol{\Gamma}}^T\|_F\|\boldsymbol{\varepsilon}X^{jT}/\sqrt{n}\|_2 + \sqrt{n}\lambda_n\omega_j \leq r_n\|\boldsymbol{\varepsilon}X^{jT}/\sqrt{n}\|_2 + \sqrt{n}\lambda_n\omega_j, \ j \in \mathcal{A}.$$

Upon combining the facts that $r_n = o(n^{c_2/2})$, $\|\boldsymbol{\varepsilon}X^{jT}/\sqrt{n}\|_2 = o(1)$, ω_j's are bounded and $\sqrt{n}\lambda_n = o(n^{(c_2-c_1)/2})$, we obtain $\max_{j\in\mathcal{A}} \| \sqrt{n}\hat{H}_j\|_2 = o(n^{c_2/2}) + o(n^{(c_2-c_1)/2}) = o_p(n^{c_2/2})$. Therefore, $P(\min_{j\in\mathcal{A}} \|\hat{F}_j\|_2 > 0) \to 1$. Moveover, by using the fact that $\|\beta_j\|_2 = \|\boldsymbol{\Gamma}\boldsymbol{U}F_j\|_2 = \|F_j\|_2$ for any j, we have $\min_{j\in\mathcal{A}} \|\hat{\beta}_j\|_2 = \min_{j\in\mathcal{A}} \|\hat{F}_j\|_2$. Thus, $P(\min_{j\in\mathcal{A}} \|\hat{\beta}_j\|_2 > 0) = P(\min_{j\in\mathcal{A}} \|\hat{F}_j\|_2 > 0) \to 1$, which completes the proof of the theorem. □

Acknowledgments

We express our sincere thanks to the anonymous reviewers and the Editor for their incisive comments on an earlier version of this manuscript which led to this much improved version.

References

[1] T. W. Anderson, Asymptotic distribution of the reduced rank regression estimator under general conditions, The Annals of Statistics 27 (1999) 1141–1154.
[2] K. Chen, K.-S. Chan, N. C. Stenseth, Reduced rank stochastic regression with a sparse singular value decomposition, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74 (2012) 203–221.
[3] K. Chen, H. Dong, K.-S. Chan, Reduced rank regression via adaptive nuclear norm penalization, Biometrika 100 (2013) 901–920.
[4] L. Chen, J. Z. Huang, Sparse reduced-rank regression for simultaneous dimension reduction and variable selection, Journal of the American Statistical Association 107 (2012) 1533–1545.
[5] K. Chin, S. DeVries, J. Fridlyand, P. T. Spellman, R. Roydasgupta, W.-L. Kuo, A. Lapuk, R. M. Neve, Z. Qian, T. Ryder, et al., Genomic and transcriptional aberrations linked to breast cancer pathophysiologies, Cancer cell 10 (2006) 529–541.
[6] H. Chun, S. Keleş, Sparse partial least squares regression for simultaneous dimension reduction and variable selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (2010) 3–25.
[7] R. D. Cook, An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics, in: Wiley Series in Probability and Statistics, Wiley, New York, 2018.
[8] R. D. Cook, L. Forzani, Z. Su, A note on fast envelope estimation, Journal of Multivariate Analysis 150 (2016) 42–54.
[9] R. D. Cook, L. Forzani, X. Zhang, Envelopes and reduced-rank regression, Biometrika 102 (2015) 439–456.
[10] R. D. Cook, B. Li, F. Chiaromonte, Envelope models for parsimonious and efficient multivariate linear regression (with discussion), Statistica Sinica 20 (2010) 927–1010.
[11] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, Pathwise coordinate optimization, The Annals of Applied Statistics 1 (2007) 302–332.
[12] J. C. Gower, G. B. Dijksterhuis, Procrustes Problems, Oxford University Press, London, 2004.
[13] X. Guo, H. Zhang, Y. Wang, J.-L. Wu, Model selection and estimation in high dimensional regression models with group scad, Statistics & Probability Letters 103 (2015) 86–92.
[14] Y. Kim, H. Choi, H.-S. Oh, Smoothly clipped absolute deviation on high dimensions, Journal of the American Statistical Association 103 (2008) 1665–1673.
[15] Y. Kong, D. Li, Y. Fan, J. Lv, Interaction pursuit in high-dimensional multi-response regression via distance correlation, The Annals of Statistics 45 (2017) 897–922.
[16] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, Journal of Multivariate Analysis 88 (2004) 365–411.
[17] H. Lian, Y. Kim, Nonconvex penalized reduced rank regression and its oracle properties in high dimensions, Journal of Multivariate Analysis 143 (2016) 383–393.
[18] Y. Nardi, A. Rinaldo, On the asymptotic properties of the group lasso estimator for linear models, Electronic Journal of Statistics 2 (2008) 605–633.
[19] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack, P. Wang, Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer, The Annals of Applied Statistics 4 (2010) 53–57.
[20] G. C. Reinsel, R. P. Velu, Multivariate Reduced-Rank Regression: Theory and Applications, Springer, New York, 1998.
[21] A. J. Rothman, Positive definite estimators of large covariance matrices, Biometrika 99 (2012) 733–740.
[22] A. J. Rothman, P. J. Bickel, E. Levina, J. Zhu, Sparse permutation invariant covariance estimation, Electronic Journal of Statistics 2 (2008) 494–515.
[23] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, Molecular Biology of the Cell 9 (1998) 3273–3297.
[24] Z. Su, G. Zhu, X. Chen, Y. Yang, Sparse envelope model: efficient estimation and response variable selection in multivariate linear regression, Biometrika 103 (2016) 579–593.
[25] D. M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (2009) 515–534.
[26] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (2006) 49–67.
[27] P. Zhao, B. Yu, On model selection consistency of lasso, The Journal of Machine Learning Research 7 (2006) 2541–2563.
[28] G. Zhu, Z. Su, Envelope-based sparse partial least squares, The Annals of Statistics 48 (2020) 161–182.
[29] H. Zou, H. H. Zhang, On the adaptive elastic-net with a diverging number of parameters, The Annals of Statistics 37 (2009) 1733–1751.