Sparse Reduced-Rank Regression for Simultaneous Dimension Reduction and Variable Selection in Multivariate Regression

Lisha Chen

Department of Statistics

Yale University, New Haven, CT 06511

email: lisha.chen@yale.edu

Jianhua Z. Huang

Department of Statistics

Texas A&M University, College Station, TX 77843

email: jianhua@stat.tamu.edu

March 27, 2012

Author's Footnote:

Lisha Chen (Email: lisha.chen@yale.edu) is Assistant Professor, Department of Statistics, Yale University, 24 Hillhouse Ave, New Haven, CT 06511. Jianhua Z. Huang (Email: jianhua@stat.tamu.edu) is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143. Huang's work was partly supported by grants from NCI (CA57030), NSF (DMS-0907170, DMS-1007618), and Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

Abstract

The reduced-rank regression is an effective method to predict multiple response variables from the same set of predictor variables, because it can reduce the number of model parameters as well as take advantage of interrelations between the response variables and therefore improve predictive accuracy. We propose to add a new feature to the reduced-rank regression that allows selection of relevant variables using sparsity inducing penalties. By treating each row of the matrix of the regression coefficients as a group, we propose a group-lasso type penalty and show that this penalty satisfies certain desirable invariance property. We develop two numerical algorithms to solve the penalized regression problem and establish the asymptotic consistency of the proposed method. In particular, the manifold structure of the reduced-rank regression coefficient matrix is respected and carefully studied in our theoretical analysis. In a simulation study and real data analysis, the new method is compared with several existing variable selection methods for multivariate regression and exhibits competitive performance in prediction and variable selection.

KEYWORDS: group lasso, reduced-rank regression, Stiefel manifold, variable selection

1. INTRODUCTION

We have seen increasing applications where several response variables are predicted or explained by a common set of predictors. For example, researchers in autism study are interested in predicting multiple clinical characterization variables using patients' attentional pattern summarized in multidimensional eye-tracking data. In genetic study, it is interesting to model gene expression level at multiple time points using multiple transcription factors. One might also model the returns of multiple stocks together using a set of econometric variables.

For such multiple-response problems, one can naively perform separate linear regression on each response by ignoring the possible interrelations between response variables. However, we expect that this naive solution can be much improved. In this paper we ask the question of how to improve the interpretability and predictability of the linear model by selecting important predictors and taking the advantage of the correlations between response variables. Our question has two aspects: One is dimension reduction, that is, to combine the predictor variables into fewer features which can be explained as latent factors that drive the variation in the multiple response variables. The other is variable selection, that is, to identify the relevant predictor variables and discard irrelevant variables when deriving those latent factors.

We develop a sparse reduced-rank regression (SRRR) method for multivariate regression by addressing the two aspects of the question. The dimension reduction aspect of multivariate regression is taken care of by the so-called reduced-rank regression (RRR) (Izenman 1975; Reinsel & Velu 1998). RRR makes a restriction on the rank of the regression coefficient matrix. The rank constraint implies that the effective number of parameters to be estimated is reduced and the efficiency of estimation is thus improved. It also implies that the coefficient matrix can be expressed as the product of two lower rank matrices. The predictors multiplied by one of the lower rank matrices produces the lower dimensional factors that drive the variation in the multiple responses. The variable selection aspect is addressed by adding a penalty to the least squares fitting criterion to enforce the sparsity of the reduced-rank coefficient matrix. We focus in this paper on sparsity inducing penalty for the purpose of variable selection. Note that the Ky-Fan norm penalty has been used for factor selection and shrinkage in multivariate regression (Yuan, Ekici, Lu & Monteiro 2007) but it is not applicable for variable selection.

Different from the single response regression that is extensively studied in the literature (Tibshirani 1996; Yuan & Lin 2006; Wang & Leng 2008), the variable selection for multivariate regression considered here is not only supervised jointly by all predictors but also is integrated with the dimension reduction procedure. We use some invariance consideration to decide on a reasonable sparsity-inducing penalty in a penalized regression formulation (Section 2). The resulting penalty is a group-lasso type penalty that treats each row of the regression coefficient matrix as a group. Two numerical algorithms are developed for computation and tunining parameter selection methods are proposed (Section 3). One important difference between our setting and the well-studied single response setting is that the range of the regression coefficient matrix is not a linear space and has certain manifold structure. We address this issue by making good use of the Stiefel manifold representation in our theoretical analysis (Section 4). In an asymptotic analysis, we obtain the consistency result in terms of parameter estimation and variable selection (Section 5). It is well known that RRR contains many classical multivariate regression models as special cases, including principal component and factor analysis, canonical correlation analysis, linear discriminant analysis, and correspondence analysis; see Chapter 6 of Izenman (2008). Therefore, our sparse RRR method provides a unified treatment of variable selection for these methods.

Variable selection for multivariate regression has started to attract attention recently. Several methods have been proposed in the literature, including the L_2SVS (Simila & Tikka 2007), L_∞SVS (Turlach, Venables & Wright 2005), RemMap (Peng, Zhu, Bergamaschi, Han, Noh, Pollack & Wang 2010), and SPLS (Chun & Keles 2010). The first three of these methods also use sparsity inducing penalty but none of them considers the reduced rank structure. Our SRRR method is compared with these methods and the separate penalized regression method in a simulation study and show competitive performance (Section 6). The SRRR method is illustrated using a yeast cell cycle dataset in Section 7.

2. SPARSE REDUCED-RANK REGRESSION

2.1 Multivariate linear regression and the reduced-rank model

Suppose we have multiple response variables $Y_1, Y_2, \ldots, Y_q$ and multiple predictor variables $X_1, X_2, \ldots, X_p$. The linear model assumes a linear relationship between each response variable and the

predictors, that is,

$$Y_j = \sum_{k=1}^{p} X_k c_{kj} + \epsilon_j, \qquad j = 1, 2, \ldots, q.$$

In the model above we have omitted the intercept term without loss of generality, since the intercept can be removed by assuming the response variables and predictor variables have mean zero. We also assume that the q error terms ϵ_j's are random variables with mean zero. With n observations we can write the model in the matrix notation as

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

where $\mathbf{Y}$ is the $n \times q$ response matrix, $\mathbf{X}$ is the $n \times p$ predictor matrix, $\mathbf{C}$ is the $p \times q$ matrix of regression coefficients and $\mathbf{E}$ is the $n \times q$ error matrix. Each row of $\mathbf{X}$ and $\mathbf{Y}$ corresponds to an observation. The generalization of the least squares criterion to the multiple response case is

$$\begin{aligned} \mathrm{RSS}(\mathbf{C}) &= \sum_{j=1}^{q} \sum_{i=1}^{n} \left(y_{ij} - \sum_{k=1}^{p} x_{ik} c_{kj} \right)^2 \\ &= tr[(\mathbf{Y} - \mathbf{XC})^T (\mathbf{Y} - \mathbf{XC})] \\ &= \|\mathbf{Y} - \mathbf{XC}\|^2 \end{aligned} \tag{2}$$

where $\| \cdot \|$ denotes the Frobenius norm for a matrix or the Euclidean norm for a vector. The ordinary least squares (OLS) estimate of $\mathbf{C}$ is

$$\hat{\mathbf{C}}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \tag{3}$$

Note that the OLS estimate (3) for multiple responses is equivalent to performing separate OLS estimation for each response variable, and it does not make use of the fact that the multiple responses are likely correlated. In practice, however, it is often the case that the multiple regressions which are considered together have correlated response variables. One way of taking advantage of possible interrelationships between response variables is to impose a constraint on the rank of $\mathbf{C}$, that is,

$$rank(\mathbf{C}) = r, \qquad r \leq \min(p, q), \tag{4}$$

resulting in the reduced-rank regression (RRR) model (Reinsel & Velu 1998). An immediate implication of the reduced-rank restriction is that there is a number of linear constraints on regression

coefficients, and hence the number of effective number of parameters is reduced and the estimation efficiency is improved. It also follows from the rank constraint that $\mathbf{C}$ can be expressed as a product of two rank r matrices as follows

$$\mathbf{C} = \mathbf{B}\mathbf{A}^T$$

where $\mathbf{B}$ is of dimension $p \times r$ and $\mathbf{A}$ is of dimension $q \times r$. The model (1) therefore can be rewritten as

$$\mathbf{Y} = (\mathbf{X}\mathbf{B})\,\mathbf{A}^T + \mathbf{E} \tag{5}$$

where $\mathbf{XB}$ is of reduced dimension with only r components. These r linear combinations of the predictor variables can be interpreted as unobservable latent factors that drive the variation in the responses. We expect that the correlations between the q responses are taken into account in the model as they are modeled by r ($r \leq q$) common latent factors. We therefore achieve the dimensionality reduction of the predictor variables.

For fixed r, we estimate the rank-constrained $\mathbf{C}$ by solving the optimization problem

$$\min_{\mathbf{C}:\, rank(\mathbf{C})=r} \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|^2, \tag{6}$$

or equivalently,

$$\min_{\mathbf{A},\mathbf{B}:\, rank(\mathbf{A})=r, rank(\mathbf{B})=r} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{A}^T\|^2. \tag{7}$$

Denote $\mathbf{S}_{xx} = (1/n)\mathbf{X}^T\mathbf{X}$, $\mathbf{S}_{xy} = (1/n)\mathbf{X}^T\mathbf{Y}$, and $\mathbf{S}_{yx} = (1/n)\mathbf{Y}^T\mathbf{X}$. The matrices $\mathbf{A}$ and $\mathbf{B}$ that solve (7) are determined only up to nonsingular transformations. A set of solution is provided by

$$\hat{\mathbf{A}}^{(r)} = \mathbf{V}, \qquad \hat{\mathbf{B}}^{(r)} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{V},$$

where $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$ and $\mathbf{v}_j$ is the eigenvector of $\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ corresponding to the jth largest eigenvalue λ_j. This solution satisfies the identifiability conditions that $\mathbf{A}^T\mathbf{A} = \mathbf{I}_r$ and $\mathbf{B}^T\mathbf{S}_{xx}\mathbf{B}$ being diagonal.

2.2 Sparse reduced-rank regression through penalized least squares

We now develop our method of variable selection for RRR using penalized regression. RRR allows the responses to borrow strength from each other through a set of common latent factors to improve prediction accuracy. However, each latent factor is a linear combination of all predictors. When a

large number of predictor variables are available, some of them might not be useful for prediction. Thus we would like to perform variable selection, or exclude the redundant predictors when forming the latent factors. Note that excluding a predictor corresponds to setting as zero an entire row of the matrix $\mathbf{B}$. Inspired by the penalized regression with a grouped lasso penalty (Yuan & Lin 2006), we consider the following optimization problem

$$\min_{\mathbf{A},\mathbf{B}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{A}^T\|^2 + \sum_{i=1}^{p} \lambda_i \|\mathbf{B}^i\| \quad \text{s.t.} \quad \mathbf{A}^T\mathbf{A} = \mathbf{I}, \tag{8}$$

where the superscript denotes a row of the matrix so that $\mathbf{B}^i$ is a *row* vector, the constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ is introduced for identifiability purpose, and $\lambda_i > 0$ are penalty parameters whose choice will be discussed in Section 3.5. In (8), each row of $\mathbf{B}$ is treated as a group and $\|\mathbf{B}^i\| = 0$ is equivalent to setting the i-th row of $\mathbf{B}$ as zeros. Thus the group lasso penalty encourages row-wise sparsity on the $\mathbf{B}$ matrix. An alternative way of introducing sparsity is to directly use the lasso penalty (Tibshirani 1996) on the entire matrix $\mathbf{B}$ that encourages element-wise sparsity. We shall show below that the group lasso penalty has certain invariance property that the lasso penalty does not have. Note that the collection of $\mathbf{B}\mathbf{A}^T$ does not form a linear space, and its manifold structure needs to be respected in optimization and in the theoretical analysis of the property of the corresponding estimator (see Sections 4 and 5 for more details). This is the key difference of (8) to the original group lasso problem (Yuan & Lin 2006).

The solution to the optimization problem (8) is not unique. Suppose $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ is a solution. For a $r \times r$ orthogonal matrix $\mathbf{Q}$, let $\widetilde{\mathbf{A}} = \widehat{\mathbf{A}}\mathbf{Q}$ and $\widetilde{\mathbf{B}} = \widehat{\mathbf{B}}\mathbf{Q}$. It follows from the fact $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$ that $\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}^T = \widehat{\mathbf{B}}\widehat{\mathbf{A}}^T$ and $\|\widetilde{\mathbf{B}}^i\| = \|\widehat{\mathbf{B}}^i\|$. Thus $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ is also a solution of (8). On the other hand, suppose $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ and $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ are two solutions of (8). Since both $\widehat{\mathbf{A}}$ and $\widetilde{\mathbf{A}}$ are of full column rank r, there is non-singular $r \times r$ matrix $\mathbf{Q}$ such that $\widetilde{\mathbf{A}} = \widehat{\mathbf{A}}\mathbf{Q}$. The orthogonality constraint implies that

$$\mathbf{I}_r = \widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}} = \mathbf{Q}^T\widehat{\mathbf{A}}^T\widehat{\mathbf{A}}\mathbf{Q} = \mathbf{Q}^T\mathbf{Q},$$

which in turn implies that $\mathbf{Q}$ is an orthogonal matrix. Using $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ and $\|\widetilde{\mathbf{B}}^i\| = \|\widetilde{\mathbf{B}}^i\mathbf{Q}^T\|$, we obtain that $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ also minimizes $\|\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{Q}^T(\mathbf{A}\mathbf{Q}^T)^T\|^2 + \sum_{i=1}^p \lambda_i \|\mathbf{B}^i\mathbf{Q}^T\|$. Since $\widehat{\mathbf{A}} = \widetilde{\mathbf{A}}\mathbf{Q}^T$, it follows that $(\widehat{\mathbf{A}}, \widetilde{\mathbf{B}})$ minimize $\|\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{Q}^T\mathbf{A}^T\|^2 + \sum_{i=1}^p \lambda_i \|\mathbf{B}^i\mathbf{Q}^T\|$. Thus, fixing $\mathbf{A}$ at $\widehat{\mathbf{A}}$, both $\widehat{\mathbf{B}}$ and $\widetilde{\mathbf{B}}\mathbf{Q}^T$ minimize the convex function $\|\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{A}^T\|^2 + \sum_{i=1}^p \lambda_i \|\mathbf{B}^i\|$, and therefore $\widehat{\mathbf{B}} = \widetilde{\mathbf{B}}\mathbf{Q}^T$,

or $\widetilde{\mathbf{B}} = \widehat{\mathbf{B}}\mathbf{Q}$. We have obtained the following result regarding the property of the optimization problem.

Lemma 1 *The solution to the optimization problem* (8) *is unique up to a* $r \times r$ *orthogonal matrix. More precisely, suppose* $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ *is a solution of* (8). $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ *is also a solution of* (8) *if and only if there is an orthogonal matrix* $\mathbf{Q}$ *such that* $\widetilde{\mathbf{A}} = \widehat{\mathbf{A}}\mathbf{Q}$ *and* $\widetilde{\mathbf{B}} = \widehat{\mathbf{B}}\mathbf{Q}$.

According to this lemma, $\mathbf{B}$ is determined only up to an orthogonal transformation, and thus setting to zero a single element of $\mathbf{B}$ does not have a clear meaning. We now show that our formulation ensures meaningful variable selection in that different solutions of the optimization problem correspond to selection of the same set of predictors. To formalize the idea, we need the following definition. Note that each row of $\mathbf{B}$ corresponds to a column of $\mathbf{X}$.

Definition 1 *If the entire row* j *of* $\mathbf{B}$ *is zero, then the predictor variable* X_j *is called a nonactive variable, otherwise it is called an active variable.*

Suppose $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ and $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ are two solutions of (8). By switching the order of variables we can write $\widehat{\mathbf{B}} = (\widehat{\mathbf{B}}_1^T, \mathbf{0})^T$ such that $\widehat{\mathbf{B}}_1$ does not include rows with all zeros. According to Lemma 1, there exists an orthogonal matrix $\mathbf{Q}$ such that $\widetilde{\mathbf{B}} = (\widehat{\mathbf{B}}_1^T, \mathbf{0})^T\mathbf{Q} = ((\widehat{\mathbf{B}}_1\mathbf{Q})^T, \mathbf{0})^T$. Note that none of the rows of $\widehat{\mathbf{B}}_1\mathbf{Q}$ is entirely zero; for any row $\widehat{\mathbf{B}}_1^i$ in the matrix $\widehat{\mathbf{B}}_1$, $\widehat{\mathbf{B}}_1^i\mathbf{Q} = \mathbf{0}$ if and only if $\widehat{\mathbf{B}}^i = \mathbf{0}$. Therefore, the active variables determined by $\widetilde{\mathbf{B}}$ are the same as those by $\widehat{\mathbf{B}}$. We summarize the result below as a lemma.

Lemma 2 *The set of active variables obtained by solving the optimization problem* (8) *is uniquely determined.*

This lemma provides some support to our use of group lasso penalty since it guarantees the identifiability of variable selection. It is easy to see that element-wise sparsity of $\mathbf{B}$ does change when $\mathbf{B}$ is multiplied by a rotation matrix.

3. NUMERICAL SOLUTION AND TUNING

This section presents two algorithms for solving the optimization problem (8), both iteratively optimizing with respect to $\mathbf{A}$ and $\mathbf{B}$. This section also discusses methods for specifying the tuning parameters.

3.1 Overview of the iterative optimization

For fixed $\mathbf{B}$, the optimization problem reduces to

$$\min_{\mathbf{A}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{A}^T\| \quad \text{s.t.} \quad \mathbf{A}^T\mathbf{A} = \mathbf{I}, \tag{9}$$

which is an orthogonal Procrustes problem (Gower & Dijksterhuis 2004). The solution of $\mathbf{A}$ is $\widehat{\mathbf{A}} = \mathbf{U}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are obtained from the singular value decomposition $\mathbf{Y}^T\mathbf{X}\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U}$ is $q \times r$, and $\mathbf{V}$ is $r \times r$.

Now we consider optimization over $\mathbf{B}$ for fixed $\mathbf{A}$. Since $\mathbf{A}$ has orthonormal columns, there is a matrix $\mathbf{A}^\perp$ with orthonormal columns such that $(\mathbf{A}, \mathbf{A}^\perp)$ is an orthogonal matrix. Then we have

$$\|\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{A}^T\|^2 = \|(\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{A}^T)(\mathbf{A}, \mathbf{A}^\perp)\|^2 = \|\mathbf{Y}\mathbf{A} - \mathbf{X}\mathbf{B}\|^2 + \|\mathbf{Y}\mathbf{A}^\perp\|^2. \tag{10}$$

The second term does not involve $\mathbf{B}$. Therefore for fixed A, the optimization problem (8) reduces to

$$\min_{\mathbf{B}} \|\mathbf{Y}\mathbf{A} - \mathbf{X}\mathbf{B}\|^2 + \sum_{i=1}^{p} \lambda_i \|\mathbf{B}^i\|. \tag{11}$$

Below, we present two methods for solving this problem.

3.2 Subgradient method

We use the subgradient method to solve (11) (Friedman, Hastie, Hofling, & Tibshirani 2007). The subgradient equations w.r.t. $\mathbf{B}^l$, the lth row of $\mathbf{B}$, are

$$2\mathbf{X}_l^T(\mathbf{X}\mathbf{B} - \mathbf{Y}\mathbf{A}) + \lambda_l s_l = 0, \qquad l = 1, 2, \ldots, p,$$

where $s_l = \mathbf{B}^l/\|\mathbf{B}^l\|$ if $\|\mathbf{B}^l\| \neq 0$, and s_l is a r-vector satisfying $\|s_l\| < 1$ if $\|\mathbf{B}^l\| = 0$. If $\|\mathbf{B}^l\| = 0$, the subgradient equations for $\mathbf{B}^l$ become

$$2\mathbf{X}_l^T\left(\sum_{k\neq l}^{p} \mathbf{X}_k\mathbf{B}^k - \mathbf{Y}\mathbf{A}\right) + \lambda_l s_l = 0.$$

Solving for s_l, one gets

$$s_l = -\frac{2}{\lambda_l}\mathbf{X}_l^T\left(\sum_{k\neq l}^{p} \mathbf{X}_k\mathbf{B}^k - \mathbf{Y}\mathbf{A}\right) = -\frac{2}{\lambda_l}\mathbf{X}_l^T\mathbf{R}_l,$$

where $\mathbf{R}_l = \mathbf{YA} - \sum_{k\neq l}^{p} \mathbf{X}_k\mathbf{B}^k$, which in turn is used to check whether $\|s_l\| < 1$. If $\|s_l\| < 1$ holds, set $\|\mathbf{B}^l\| = 0$. Otherwise, $\mathbf{B}^l$ can be solved from the first order conditions. We can transform the first order condition w.r.t. $\mathbf{B}^l$ ($\mathbf{B}^l \neq 0$) to

$$-2\mathbf{X}_l^T(\mathbf{R}_l - \mathbf{X}_l\mathbf{B}^l) + \lambda_l \frac{\mathbf{B}^l}{\|\mathbf{B}^l\|} = 0,$$

which has the solution

$$\mathbf{B}^l = \left(\mathbf{X}_l^T\mathbf{X}_l + \frac{\lambda_l}{2\,\|\mathbf{B}^l\|}\right)^{-1} \mathbf{X}_l^T\mathbf{R}_l. \tag{12}$$

Note that the right hand side involves $\|\mathbf{B}^l\|$ and therefore we need to obtain $\|\mathbf{B}^l\|$ in order to get $\mathbf{B}^l$. Denote $c = \|\mathbf{B}^l\|$ and we can obtain an equation of c using (12). It is easy to see this equation has the solution $c = (\|\mathbf{X}_l^T\mathbf{R}_l\| - \lambda_l/2)/\|\mathbf{X}_l\|^2$. Plug this solution into the right-hand side of (12) to get

$$\mathbf{B}^l = \frac{1}{\|\mathbf{X}_l\|^2}\left(1 - \frac{\lambda_l}{2\|\mathbf{X}_l^T\mathbf{R}_l\|}\right)\mathbf{X}_l^T\mathbf{R}_l.$$

Combining this result with that for $\mathbf{B}^l = 0$, we obtain that the optimal solution of (11) is

$$\mathbf{B}^l = \frac{1}{\mathbf{X}_l^T\mathbf{X}_l}\left(1 - \frac{\lambda_l}{2\|\mathbf{X}_l^T\mathbf{R}_l\|}\right)_+ \mathbf{X}_l^T\mathbf{R}_l. \tag{13}$$

This is a vector version of the soft-thresholding rule.

The discussion above leads to the following algorithm:

Algorithm 1: numerical algorithm using subgradient method

Input: $\mathbf{X}$,$\mathbf{Y}$,λ

Output: $\mathbf{A}$,$\mathbf{B}$

while value of objective function (8) not converged **do**
- For fixed $\mathbf{B}$, solve $\mathbf{A}$ by SVD as in Procrustes problem indicated by (9).
- **while** $\mathbf{B}$ not converged **do**
  - **for** each l **do**
    - solve $\mathbf{B}^l$ by (13)
  - check whether $\mathbf{B}$ has converged
- check whether objective function has converged

3.3 Variational method

An alternative method to optimize (11) over $\mathbf{B}$ for fixed $\mathbf{A}$ is the variational method that makes use of the following result

$$\min_c \frac{1}{2}\left(cx^2 + \frac{1}{c}\right) = |x|.$$

According to this result, optimizing the objective function in (11) over $\mathbf{B}$ is equivalent to optimizing the following objective function jointly over $\mathbf{B}$ and μ_i, $i = 1, ..., p$,

$$f = \|\mathbf{YA} - \mathbf{XB}\|^2 + \sum_{i=1}^{p} \frac{\lambda_i}{2}\left(\mu_i \|\mathbf{B}^i\|^2 + \frac{1}{\mu_i}\right), \tag{14}$$

which can be done iteratively. For fixed $\mathbf{B}$, the optimal μ_i is

$$\mu_i = 1/\|\mathbf{B}^i\|, \qquad i = 1, \ldots, p. \tag{15}$$

For fixed μ_i, the first order condition for $\mathbf{B}^i$ is

$$\frac{\partial f}{\partial \mathbf{B}^i} = -2\mathbf{X}_i^T(\mathbf{YA} - \mathbf{XB}) + \lambda_i \mu_i \mathbf{B}^i = 0, \qquad i = 1, 2, \ldots, p.$$

Collecting all these conditions and express them in the matrix form to obtain

$$-\mathbf{X}^T(\mathbf{YA} - \mathbf{XB}) + \frac{1}{2}\mathrm{diag}(\lambda_1\mu_1, ..., \lambda_p\mu_p)\,\mathbf{B} = 0.$$

Solving for $\mathbf{B}$ yields

$$\mathbf{B} = \{\mathbf{X}^T\mathbf{X} + \frac{1}{2}\mathrm{diag}(\lambda_1\mu_1, ..., \lambda_p\mu_p)\}^{-1}\mathbf{X}^T\mathbf{YA}. \tag{16}$$

One iterates between (15) and (16) until convergence. The variational method seldom produces $\|\mathbf{B}^i\|$ that is exactly zero, and so one needs to set a threshold for $\|\mathbf{B}^i\|$ to obtain the sparsity.

3.4 Comparison of the two algorithms

We compare the computational complexity of the two algorithms. Updating $\mathbf{A}$ is done in the same fashion in both methods and so we consider only the steps for updating B. In the following calculation, we ignore the one-time computation such as for $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{Y}$ and focus on the terms which appear in the iterative steps. Moreover, we assume that the rank number r is a fixed small constant. For the variational method, (15) requires $2pr$ operations and (16) requires $O(p^3)$(for inversion) + pqr(for $(\mathbf{X}^T\mathbf{Y})\mathbf{A}$) + p^2r operations. Since r is small, the computational

complexity for each iteration of the variational method is $O(pq + p^3)$. For the subgradient method, calculation of $\mathbf{R}_l$ requires $O(npr + nqr)$ operations, subsequent calculation of $\mathbf{X}_l^T \mathbf{R}_l$ requires $O(nr)$ operations, and thus calculation of $\mathbf{B}_l$ using (13) requires $O(npr+nqr)$ operations. Since there are p of $\mathbf{B}_l$'s, the computational complexity for one iteration of the subgradient method is $O(nrpq+nrp^2)$, or $O(npq+np^2)$ when r is a small number. It is clear that the variational method is computationally less expensive when $n \gg p$. If q is smaller or not much bigger than p, the subgradient method is computationally less costly when $p \gg n$.

3.5 Tuning

The cross-validation (CV) or K-fold CV can be used to select the rank number (or the number of factors) r and the penalty parameters. The parameters that give a smaller CV error are preferred. From the simulation study to be reported in Section 6, we observed that the 5-fold CV works very well in selecting the correct rank number when the data is generated from a reduced rank model; see Table 3. In practice, the reduced rank model is usually an approximation, and the CV error may continue to decline as a function of the rank number. Borrowing an idea from the principal components analysis, we can use the scree plot of the CV error to select an appropriate rank number; see Section 7 for an illustration using a real-world data set.

Since selection of p penalty parameters requires computationally intensive optimization, we propose two strategies to reduce the number of tuning parameters: One uses a single penalty parameter by setting λ_i's all equal; the other uses the idea of adaptive lasso (Zou 2006). The adaptive lasso penalty requires a pilot estimator $\widetilde{\mathbf{C}}$ and sets $\lambda_i = \lambda \|\tilde{\mathbf{C}}^i\|^{-\gamma}$, where $\|\tilde{\mathbf{C}}^i\|^{-\gamma}$'s are referred to as the adaptive weights and λ, $\gamma > 0$ are two tuning parameters that can be selected using CV. Our experience suggests that it is sufficient to choose γ from a small set of candidates such as $\{0.5, 1, 2, 4, 8\}$. When $n > p$, the solution of the unpenalized version of (8) provides a pilot estimator. When $n \leq p$, the solutions of unpenalized reduced-rank regression or ordinary multivariate regression are not well defined. For this case, a reasonable pilot estimator can be the solution of (8) with a single penalty parameter, which in turn is chosen by the CV.

4. GEOMETRY OF THE REDUCED-RANK MODEL

This section discusses some geometry of the reduced-rank model, which will play an important role in obtaining the asymptotic results in the next section. Let $\mathbf{C}_*$ be the rank-r coefficient matrix used to generate the data according to model (1) and $\mathbf{C}_* = \mathbf{U}_*\mathbf{D}_*{\mathbf{V}_*}^T$ be its reduced singular value decomposition, where $\mathbf{U}_*$ and $\mathbf{V}_*$ are respectively $p \times r$ and $q \times r$ rank-r matrices with orthonormal columns and $\mathbf{D}_*$ is a $r \times r$ nonnegative diagonal matrix. Although this reparameterization is not globally identifiable (considering, for example, $\widetilde{\mathbf{U}}_* = -\mathbf{U}_*$ and $\widetilde{\mathbf{V}}_* = -\mathbf{V}_*$), it is identifiable up to the sign switch of each column of $\mathbf{U}_*$ and $\mathbf{V}_*$. Let $\mathbf{C}$ be a $p \times q$ matrix in the neighborhood of $\mathbf{C}_*$ on the manifold of rank-r matrices. We now derive a local reparametrization of $\mathbf{C}$.

Let $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be the reduced singular value decomposition, where $\mathbf{U}$ and $\mathbf{V}$ are respectively $p \times r$ and $q \times r$ rank-r orthonormal matrices and $\mathbf{D}$ is a $r \times r$ nonnegative diagonal matrix. According to the Stiefel manifold representation of orthonormal matrices (Edelman, Arias & Smith 1998), the manifold of $p \times r$ orthonormal matrices can be represented as

$$\{\mathbf{U} = \mathbf{exp}(1, \mathbf{U}_*\mathbf{G}_u + \mathbf{H}_u) : \mathbf{G}_u \text{ is } r \times r, \mathbf{G}_u = -\mathbf{G}_u^T, \mathbf{H}_u \text{ is } p \times r, \mathbf{U}_*^T\mathbf{H}_u = \mathbf{0}\},$$

where $\exp(t, \mathbf{J}_u)$ is the exponential map that defines a geodesic emanating from $\mathbf{U}_*$ in the tangent direction $\mathbf{J}_u$. The exponential map can be expressed as $\mathbf{exp}(t, \mathbf{J}_u) = \mathbf{U}_*\mathbf{M}(t, \mathbf{J}_u) + \mathbf{Q}\mathbf{N}(t, \mathbf{J}_u)$, where $\mathbf{Q}\mathbf{R} = (\mathbf{I} - \mathbf{U}_*\mathbf{U}_*^T)\mathbf{J}_u$ is the compact QR decomposition ($\mathbf{Q}$ is $p \times r$, $\mathbf{R}$ is $r \times r$), and $\mathbf{M}(t, \mathbf{J}_u)$ and $\mathbf{N}(t, \mathbf{J}_u)$ are $r \times r$ matrices given by the matrix exponential

$$\begin{bmatrix} \mathbf{M}(t, \mathbf{J}_u) \\ \mathbf{N}(t, \mathbf{J}_u) \end{bmatrix} = \exp\left\{ t \begin{bmatrix} \mathbf{U}_*^T\mathbf{J}_u & -\mathbf{R}^T \\ \mathbf{R} & \mathbf{0} \end{bmatrix} \right\} \begin{bmatrix} \mathbf{I}_r \\ \mathbf{0} \end{bmatrix}.$$

Similarly, the manifold of $q \times r$ orthonormal matrices can be represented as

$$\{\mathbf{V} = \mathbf{exp}(1, \mathbf{V}_*\mathbf{G}_v + \mathbf{H}_v) : \mathbf{G}_v \text{ is } r \times r, \mathbf{G}_v = -\mathbf{G}_v^T, \mathbf{H}_v \text{ is } q \times r, \mathbf{V}_*^T\mathbf{H}_v = \mathbf{0}\},$$

where $\exp(t, \mathbf{J}_v)$ is the exponential map that defines a geodesic emanating from $\mathbf{V}_*$ in the tangent direction $\mathbf{J}_v$. The non-negative diagonal matrix $\mathbf{D}$ can be parametrized as $\mathbf{D} = \mathbf{D}_* \exp(\mathbf{K})$, where $\mathbf{K}$ is a $r \times r$ diagonal matrix whose diagonal elements are not constrained.

Using the the definition of the matrix exponential and the Taylor series expansion, we obtain

the following first order approximations for perturbations along the manifolds:

$$\mathbf{U} - \mathbf{U}_* = \mathbf{exp}(1, \mathbf{U}_*\mathbf{G}_u + \mathbf{H}_u) - \mathbf{U}_* = \mathbf{U}_*\mathbf{G}_u + \mathbf{H}_u + O(\|\mathbf{G}_u\|^2 + \|\mathbf{H}_u\|^2), \tag{17}$$

$$\mathbf{V} - \mathbf{V}_* = \mathbf{exp}(1, \mathbf{V}_*\mathbf{G}_v + \mathbf{H}_v) - \mathbf{V}_* = \mathbf{V}_*\mathbf{G}_v + \mathbf{H}_v + O(\|\mathbf{G}_v\|^2 + \|\mathbf{H}_v\|^2), \tag{18}$$

$$\mathbf{D} - \mathbf{D}_* = \mathbf{D}_* \exp(\mathbf{K}) - \mathbf{D}_* = \mathbf{D}_*\mathbf{K} + O(\|\mathbf{K}\|^2). \tag{19}$$

We say that two matrix norms are equivalent if their ratio is bounded away from zero and infinity. Note that $\|\mathbf{U}_*\mathbf{G}_u + \mathbf{H}_u\|^2 = \|\mathbf{G}_u\|^2 + \|\mathbf{H}_u\|^2$ and $\|\mathbf{V}_*\mathbf{G}_v + \mathbf{H}_v\|^2 = \|\mathbf{G}_v\|^2 + \|\mathbf{H}_v\|^2$. Thus (17)–(19) imply the following result.

Lemma 3 *$\|\mathbf{U} - \mathbf{U}_*\|$ and $\|\mathbf{V} - \mathbf{V}_*\|$ are locally equivalent to $\{\|\mathbf{G}_u\|^2 + \|\mathbf{H}_u\|^2\}^{1/2}$ and $\{\|\mathbf{G}_v\|^2 + \|\mathbf{H}_v\|^2\}^{1/2}$, respectively. If the diagonal elements of $\mathbf{D}_*$ are bounded away from zero and infinity, $\|\mathbf{D} - \mathbf{D}_*\|$ is locally equivalent to $\|\mathbf{K}\|$.*

The next result gives the first order approximation of $\mathbf{C} = \mathbf{UDV}^T$ in a neighborhood of $\mathbf{C}_* = \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T$. The proof relies on (17)–(19) and is given in the Appendix.

Lemma 4 *The following holds:*

$$\begin{aligned}
&\mathbf{UDV}^T - \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T \\
&\quad = \mathbf{U}_*(\mathbf{G}_u\mathbf{D}_* - \mathbf{D}_*\mathbf{G}_v)\mathbf{V}_*^T + \mathbf{U}_*\mathbf{D}_*\mathbf{H}_v^T + \mathbf{H}_u\mathbf{D}_*\mathbf{V}_*^T + \mathbf{U}_*\mathbf{D}_*\mathbf{K}\mathbf{V}_*^T \\
&\qquad + O(\|\mathbf{G}_u\|^2 + \|\mathbf{H}_u\|^2 + \|\mathbf{G}_v\|^2 + \|\mathbf{H}_v\|^2 + \|\mathbf{K}\|^2).
\end{aligned} \tag{20}$$

Note that the four leading terms in (20) are orthogonal to each other. The orthogonality of the terms involving $\mathbf{H}_u$ and $\mathbf{H}_v$ follows from $\mathbf{U}_*^T\mathbf{H}_u = \mathbf{0}$ and $\mathbf{V}_*^T\mathbf{H}_v = \mathbf{0}$. To show the orthogonality of the two terms not involving $\mathbf{H}_u$ and $\mathbf{H}_v$, note that

$$\langle \mathbf{U}_*(\mathbf{G}_u\mathbf{D}_* - \mathbf{D}_*\mathbf{G}_v)\mathbf{V}_*^T, \mathbf{U}_*\mathbf{D}_*\mathbf{K}\mathbf{V}_*^T \rangle = \langle \mathbf{G}_u\mathbf{D}_* - \mathbf{D}_*\mathbf{G}_v, \mathbf{D}_*\mathbf{K} \rangle.$$

Because of the anti-symmetry, the diagonal elements of $\mathbf{G}_u$ and $\mathbf{G}_v$ are all zero. Since $\mathbf{D}_*$ and $\mathbf{K}$ are diagonal matrices, the above inner product is zero. The orthogonality also implies that

$$\begin{aligned}
&\|\mathbf{U}_*(\mathbf{G}_u\mathbf{D}_* - \mathbf{D}_*\mathbf{G}_v)\mathbf{V}_*^T + \mathbf{U}_*\mathbf{D}_*\mathbf{H}_v^T + \mathbf{H}_u\mathbf{D}_*\mathbf{V}_*^T + \mathbf{U}_*\mathbf{D}_*\mathbf{K}\mathbf{V}_*^T\|^2 \\
&\quad = \mathrm{tr}\{(\mathbf{G}_u\mathbf{D}_* - \mathbf{D}_*\mathbf{G}_v)^T(\mathbf{G}_u\mathbf{D}_* - \mathbf{D}_*\mathbf{G}_v)\} \\
&\qquad + \mathrm{tr}(\mathbf{D}_*\mathbf{H}_u^T\mathbf{H}_u\mathbf{D}_*) + \mathrm{tr}(\mathbf{D}_*\mathbf{H}_v^T\mathbf{H}_v\mathbf{D}_*) + \mathrm{tr}(\mathbf{D}_*\mathbf{K}^2\mathbf{D}_*).
\end{aligned} \tag{21}$$

Let d_i denote the i-th diagonal element of $\mathbf{D}_*$. The next result is proved in the Appendix.

Lemma 5 *We have that*

$$\frac{1}{2}\{\min_{i<j}(d_i - d_j)^2\}(\|\mathbf{G}_u\|^2 + \|\mathbf{G}_v\|^2) \leq \|\mathbf{G}_u\mathbf{D}_* - \mathbf{D}_*\mathbf{G}_v\|^2 \leq 2\,(\max_i d_i^2)(\|\mathbf{G}_u\|^2 + \|\mathbf{G}_v\|^2).$$

To summarize the development so far, the manifold of rank-r matrices is locally reparametrized as $\boldsymbol{\Theta} = (\mathbf{G}_u, \mathbf{H}_u, \mathbf{G}_v, \mathbf{H}_v, \mathbf{K})$, where $\mathbf{G}_u$, $\mathbf{G}_v$, $\mathbf{H}_u$, $\mathbf{H}_v$, and $\mathbf{K}$ are specified as above, and $\boldsymbol{\Theta}_* = \mathbf{0}$ is the parameter value corresponding to the true coefficient matrix $\mathbf{C}_*$. Define $\|\boldsymbol{\Theta}\|^2 = \|\mathbf{G}_u\|^2 + \|\mathbf{H}_u\|^2 + \|\mathbf{G}_v\|^2 + \|\mathbf{H}_v\|^2 + \|\mathbf{K}\|^2$. Lemma 4, (21), and Lemma 5 together imply the following result:

Lemma 6 *If the diagonal elements of* $\mathbf{D}_*$ *are all distinct, positive and bounded, then the distance* $\|\mathbf{C} - \mathbf{C}_*\|$ *is locally equivalent to the norm* $\{\|\mathbf{G}_u\|^2 + \|\mathbf{H}_u\|^2 + \|\mathbf{G}_v\|^2 + \|\mathbf{H}_v\|^2 + \|\mathbf{K}\|^2\}^{1/2}$.

5. ASYMPTOTIC ANALYSIS

In this section, we study the asymptotic behavior of the estimator obtained as the solution of the problem (8) when the sample size n goes to infinity and p and q are fixed constants. Asymptotic analysis that allows p and/or q to grow with n is left for future research; see, however, Zou & Zhang (2009) for some relevant results for the univariate regression case. The following assumptions are made for the asymptotic results.

C1. There is a positive definite matrix $\boldsymbol{\Sigma}$ such that $\mathbf{X}\mathbf{X}^T/n \to \boldsymbol{\Sigma}$ as $n \to \infty$.

C2. The first p_0 variable are important and the rest are irrelevant; that is, $\|\mathbf{C}_*^i\| > 0$ if $i \leq p_0$ and $\|\mathbf{C}_*^i\| = 0$ if $i > p_0$, where $\mathbf{C}_*^i$ is the ith row of $\mathbf{C}_*$.

Theorem 1 *(Consistency of parameter estimation). Suppose that* $\lambda_i/\sqrt{n} = \lambda_{n,i}/\sqrt{n} \to 0$ *for all* $i \leq p_0$. *Then, (i) there is a local minimizer* $\widehat{\mathbf{C}}$ *of* (22) *that is* $\sqrt{n}$*-consistent in estimating* $\mathbf{C}_*$*; (ii) letting* $\widehat{\mathbf{C}} = \widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^T$ *be the singular value decomposition,* $\widehat{\mathbf{U}}$, $\widehat{\mathbf{D}}$, *and* $\widehat{\mathbf{V}}$ *are* $\sqrt{n}$*-consistent in estimating* $\mathbf{U}_*$, $\mathbf{D}_*$, *and* $\mathbf{V}_*$, *respectively.*

The proof of this theorem makes use of the geometric results in the previous section. Let $\mathbf{C} = \mathbf{B}\mathbf{A}^T$. If $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, then $\|\mathbf{C}^i\| = \|\mathbf{B}^i\|$, where $\mathbf{C}^i$ is the i-th row of $\mathbf{C}$. Thus, the optimization

problem (8) is equivalent to the minimization of the criterion function

$$Q(\mathbf{C}) = \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|^2 + \sum_{i=1}^{p} \lambda_{n,i} \|\mathbf{C}^i\| \tag{22}$$

subject to the constraint $\text{rank}(\mathbf{C}) = r$. With a slight abuse of notation, we use $Q(\mathbf{\Theta})$ to denote the optimizing objective function (22) in terms of the new parameterization given in the previous section. If we can show that for any positive number ε, there is a sufficiently large α such that

$$\liminf_n \Pr\left\{ \inf_{\|\mathbf{\Theta}\| = n^{-1/2}\alpha} Q(\mathbf{\Theta}) > Q(\mathbf{\Theta}_*) \right\} > 1 - \varepsilon, \tag{23}$$

then with probability tending to one, there exists a local minimizer $\widehat{\mathbf{\Theta}}$ of $Q(\mathbf{\Theta})$ located in the interior of the ball $\{\mathbf{\Theta} : \|\mathbf{\Theta}\| \leq n^{-1/2}\alpha\}$, and thus the corresponding $\widehat{\mathbf{C}}$ is $\sqrt{n}$-consistent, in light of Lemma 6. Details of the proof of Theorem 1 is given in the Appendix.

Theorem 2 *(Consistency of variable selection) If $\lambda_{n,i}/\sqrt{n} \to 0$ for $i \leq p_0$ and $\lambda_{n,i}/\sqrt{n} \to \infty$ for $i > p_0$, then*

$$P(\widehat{\mathbf{C}}^i = \mathbf{0}) = \Pr(\widehat{\mathbf{U}}^i = \mathbf{0}) \to 1, \qquad i > p_0.$$

The proof of this theorem is given in the Appendix. Theorem 2 in particular implies that an adaptive lasso penalty similar to that in Zou (2006) will yield consistency of variable selection. Specifically, letting $\lambda_{n,i} = \lambda \|\widetilde{\mathbf{C}}^i\|^{-\gamma}$, $\lambda, \gamma > 0$, if $\widetilde{\mathbf{C}}$ is a consistent estimate of $\mathbf{C}_*$, then the conditions on $\lambda_{n,i}$ are satisfied and so the consistency of variable selection holds. The unpenalized least squares reduced-rank estimator can be used as the pilot estimator $\widetilde{\mathbf{C}}$, since it is consistent and asymptotically normally distributed (Anderson 1999).

6. SIMULATION STUDY

In this section we use simulated data to illustrate the proposed SRRR method and compare it with five related methods that were proposed in the literature for variable selection in multivariate regression.

6.1 Related methods

The first three methods all solve certain penalized least squares problem but with different penalty functions. The ***L*$_2$SVS** (Simila & Tikka 2007) method solves the following optimization problem:

$$\min_{\mathbf{C}} \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|^2 + \lambda \sum_{i=1}^{p} \|\mathbf{C}^i\|,$$

where $\mathbf{C}^i$ is the ith row of $\mathbf{C}$. The L_2SVS is most closely related to our method SRRR but it has no low rank constraint on $\mathbf{C}$. The $\boldsymbol{L_\infty}$**SVS** (Turlach et al. 2005) method solves a similar optimization problem as L_2SVS but uses a different penalty. The optimization problem is

$$\min_{\mathbf{C}} \|\mathbf{Y} - \mathbf{XC}\|^2 + \lambda \sum_{i=1}^{p} \|\mathbf{C}^i\|_\infty$$

where $\|\mathbf{C}^i\|_\infty = \max(|\mathbf{C}_{i1}|, ..., |\mathbf{C}_{ip}|)$ is the L_∞-norm of the ith row of $\mathbf{C}$. Because of the L_∞-penalty, a predictor variable would be selected if it is useful for one response. In contrast to L_2SVS, $\|\mathbf{C}^i\|_\infty$ rather than $\|\mathbf{C}^i\|$ is used to measure the effect of the ith variable on the responses. As a result, the selection of the ith variable depends on its strongest effect on each one of the responses but not its overall effect. Therefore we expect this method will select more variables than L_2SVS which is confirmed by our simulation study. The **RemMap** (Peng et al. 2010) method imposes both row-wise and element-wise sparsity of $\mathbf{C}$ by solving the following optimization problem:

$$\min_{\mathbf{C}} \|\mathbf{Y} - \mathbf{XC}\|^2 + \lambda_1 \sum_{i=1}^{p} \|\mathbf{C}^i\|_1 + \lambda_2 \sum_{i=1}^{p} \|\mathbf{C}^i\|.$$

All these three methods encourage row-wise sparsity of $\mathbf{C}$. But none of them imposes the reduced-rank structure as we do in SRRR.

Another method is the **SPLS** method (Chun & Keles 2010) that is based on the partial least squares (PLS). A PLS method tries to find the multidimensional direction in the predictor space that explains the maximum multidimensional variance direction in the response space. The SPLS (sparse PLS) encourages sparsity in the direction vector by imposing the L_1 constraint on the optimization criterion for PLS. Because it identifies significant latent components, the SPLS essentially employs a reduced-rank structure though in a different way from the SRRR. Unlike the SRRR, SPLS does not directly target on prediction of the responses and thus we expect SPLS has disadvantage in term of prediction, as we will observe in both simulation and real data examples.

The last method is to apply separate regressions with variable selection using lasso (referred to as **SepLasso**). Unlike the other methods we mentioned above, this method ignores the possible interrelation between the responses and fits each of them separately. Although this method is not expected to be competitive, it serves as a good benchmark for all multivariate methods.

6.2 Adaptive weighting

For a fair comparison, a single penalty parameter is used for SRRR when comparing it with the five methods listed in the previous subsection. The adaptive weighting for SRRR as discussed in Section 3.5 is also evaluated, along with the same adaptive weighting scheme for L_2SVS. They are referred to as adaptive SRRR and adaptive L_2SVS in this paper. The adaptive weighting for other methods is not considered either because there is no existing implementation or because it is not straightforward how to incorporate it in the existing method.

6.3 Simulation setups and evaluation methods

In the simulation study, the data were generated using model (5), $\mathbf{Y} = \mathbf{XBA}^T + \mathbf{E}$. To be specific, the $n \times p$ matrix $\mathbf{X}$ was generated from multivariate normal distribution $\mathcal{N}(0, \mathbf{\Sigma}_x)$ and $\mathbf{\Sigma}_x$ has diagonal elements 1 and off-diagonal elements ρ_x. For the $p \times r$ component matrix $\mathbf{B}$, the elements of its first p_0 rows were generated from $\mathcal{N}(0, 1)$ and the rest $p - p_0$ rows were set to be zero. The elements of $q \times r$ matrix $\mathbf{A}$ were generated from $\mathcal{N}(0, 1)$. The elements of the $n \times q$ random noise matrix $\mathbf{E}$ is generated from $\mathcal{N}(0, \sigma^2\mathbf{\Sigma}_e)$ and $\mathbf{\Sigma}_e$ has diagonal elements 1 and off diagonal elements ρ_e. The magnitude of the noise σ^2 is chosen so that the signal to noise ratio (SNR), $trace(\mathbf{C}^T\mathbf{\Sigma}_x\mathbf{C})/trace(\mathbf{E})$, equals 1. After the matrix $\mathbf{Y}$ was obtained, we centered and standardized it and then applied the proposed SRRR method and the five related methods as well as adaptive SRRR and adaptive L_2SVS. We compared predictive accuracy of these methods in terms of the mean squared error (MSE), defined as

$$MSE = \|\mathbf{X}(\hat{\mathbf{B}}\hat{\mathbf{A}}^T - \mathbf{BA}^T)\|^2/nq, \tag{24}$$

We also compared all the methods in terms of sensitivity and specificity, which are commonly used measures to evaluate the accuracy of variable selection. The sensitivity refers to the ratio between the number of correct selection and the total number of relevant variables (p_0), which measures the ability of detecting the relevant variables. The specificity refers to the ratio between number of correct "deletion" and the total number of irrelevant variables ($p - p_0$), which measures the ability of detecting the irrelevant variables. If a method selects the relevant variables accurately, it will have both high sensitivity and high specificity. If a method tends to over-select, it will produce high sensitivity but low specificity. If a method tends to under-select, it will have low sensitivity

but high specificity. Lastly, we compared how accurate the low rank structure is estimated in terms of the number of selected factors for SRRR (and adaptive SRRR) and SPLS.

In the first set of simulations (Case 1), $n > p$. More specifically $n = 100,\ p = 30,\ p_0 = 10,\ q = 10$, and varying r, ρ_x and ρ_e. The precise setups are described as follows.

Case 1a: We set $r = 3$, $\rho_x = 0$ and $\rho_e = 0$ in this baseline model which has a strong factor structure and this setup favors our method.

Case 1b: We set $r = 10$, and $\rho_x = \rho_e = 0$. In this setup $\mathbf{C}$ is of full rank and we intend to use this case to see if SRRR has robust performance when the low rank structure is violated.

Case 1c: Same as Case 1a but with $\rho_x = 0.5$. In this case, moderate amount of correlations among predictors are introduced.

Case 1d: Same as Case 1a but with $\rho_e = 0.5$. In this case, moderate amount of correlations among error terms are introduced and ideally the weighted least square criterion should be considered. We used this case for robustness check for the ordinary least square criterion used by all methods considered in our comparative study. Rothman, Levina & Zhu (2010) proposed the MRCE method that directly models the error correlation for sparse multivariate regression. Extension of SRRR that models covariance structure may lead to more efficient estimation and is left for future research.

The second set of simulations is concerned about higher dimensional cases with $p \geq n$. In particular, we consider the following two settings for $n,\ p,\ p_0,\ q$. The other parameters are equal to baseline values in Case 1a, that is, $r = 3$, $\rho_x = 0$ and $\rho_e = 0$.

Case 2a: $n = p = 100$, $p_0 = 30$, $q = 10$.

Case 2b: $n = 100$, $p = 300$, $p_0 = 30$, $q = 30$.

In the last case of simulation, the row-wise sparsity assumption in the coefficient matrix $\mathbf{C}$ is violated. We use this setup to test the SRRR method in an unfavorable case.

Case 3: $\mathbf{C}$ has element-wise sparsity with 70% randomly assigned zero elements. The other parameters are set as follows $n = p = 100$, $p_0 = 100$, $r = q = 10$.

6.4 Simulation results

We applied all eight methods to each of the seven setups. We used five-fold cross-validation to choose tuning parameter(s) for all methods. In particular for SRRR and SPLS, the tuning parameters

include λ and the number of factors r. For the adaptive methods, tuning parameters also include γ. The performance of all methods was measured using the criteria detailed in the previous subsection.

The performance of the fitted model was measured using the MSE defined in (24) where $\mathbf{X}$ is based on 1000 test samples. Summary of the MSE for each method is given in Table 1. The first six columns provide fair comparison of our SRRR with five other methods as all of them use the same tuning strategy without adaptive weights. In all cases except case 1b (violation of low rank assumption) and the last one (violation of row-wise sparsity), SRRR produces considerable smaller average MSE among all methods, and the reduction of MSE is usually substantial. For case 1b, there is no factor structure so it is understandable that using SRRR has no clear advantage over using L_2SVS. For the last case where element-wise sparsity is imposed, RemMap performs the best without surprise because its form of penalty accommodates both row-wise and element-wise sparsity structure; our SRRR does a reasonably good job in this case. The separate lasso regression is not competitive with any multivariate methods considered; it gives the largest MSE except in the last case.

Table 2 reports the sensitivity and specificity, for each method. All methods except SPLS tend to over-select indicated by high sensitivity but low specificity, and our SRRR has the highest specificity compared to other methods. It is noted in earlier work (Peng et al. 2010) that using cross validation as tuning criterion to achieve best prediction performance often leads to over-selection, indicated by low specificity. SPLS does a supreme job in variable selection when $n > p$, but it sensitivity becomes the worst among all methods when $p \geq n$. The separate lasso regression and L_∞SVS give the lowest specificity. The SRRR performs towards the top based on the overall evaluation of sensitivity and specificity.

Table 3 shows that the 5-fold CV can select the number of factors r quite accurately for the SRRR and adaptive SRRR method.

The effect of adaptive weighting can be seen by comparing the first and the last two columns of Table 1 and Table 2. Adaptive weighting improves the accuracy of prediction and variable selection for both SRRR and L_2SVS in all the setups where $n > p$. For the cases where $p \geq n$ the adaptive weighting produces similar prediction accuracy as the unweighted version and tends to have lower sensitivity but higher specificity. That the adaptive weighting has the tendency of under-selecting

relevant variables can be explained by the fact that the variables filtered out in the first stage will get infinitely large weights in the penalty term and therefore can not be included in the final model. It is an interesting research topic to explore if use of a different pilot estimator can improve the performance of adaptive weighting in $p \geq n$ cases.

7. REAL DATA EXAMPLE: YEAST CELL CYCLE DATASET

This data analysis is concerned about identifying transcription factors (TF) that regulate the RNA transcript levels of Yeast genes within the eukaryotic cell cycle. Spellman et al. (1998) identified 800 cell cycle-regulated genes by three independent synchronization methods. We use the data generated by the α factor arrest method which are RNA levels measured every 7 minutes for 119 minutes with a total of 18 time points covering two cell cycles ($\mathbf{Y}$). The genes whose RNA levels varied periodically were identified as cell cycle-regulated genes. The chromatin immunoprecipitation (ChIP) data (Lee et al. 2002) contains binding information of these 800 genes for a total of 106 TFs ($\mathbf{X}$). We use a smaller data set analyzed by Chun & Keles (2010) which includes 524 of the 800 genes, after excluding the genes with missing RNA levels or binding information. The data is public available in the R package "spls".

We applied SRRR and adaptive SRRR to this data. We also considered L_2SVS, RemMap, and SPLS — the three other methods that showed competitive performance in our simulation study. The 5-fold CV was used for selecting tuning parameters, including the number of factors r for SRRR, adaptive SRRR, and SPLS. For each 5-fold random partition of the data, we obtained by a trace plot of the minimum CV error for $r = 1,\ 2, ...,\ 10$, where the minimum CV error corresponds to the best penalty parameters selected by CV. Figure 1 shows 20 such traces for 20 random partitions. For SRRR and adaptive SRRR, the trace plots flatten out from $r = 4$ and thus we set $r = 4$ for further analysis. For SPLS, the "elbow" shape is less pronounced and we feel that $r = 6$ is the most natural choice.

For comparison of prediction accuracy, we randomly split the data into two halves, one as the training set and one as the test set. We used the training set to identify the best penalty parameters and then used the corresponding model to predict on the test data. The top row of Table 4 reports the means and standard deviations of squared prediction errors from 50 random splits. All methods

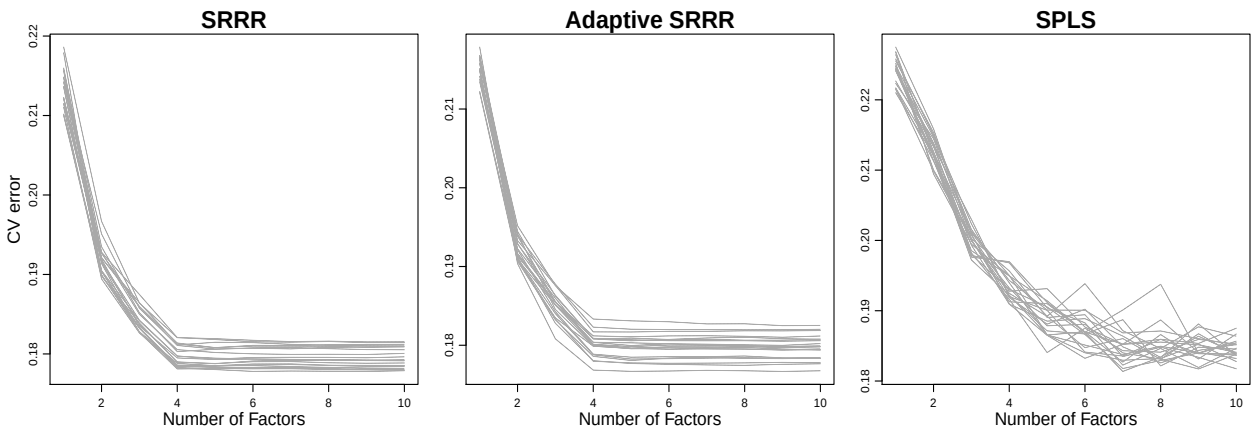

Figure 1: The 5-fold CV error at optimal values of tuning parameters when fixing the number of factors.

except SPLS perform very similarly in terms of prediction errors. The larger prediction error of SPLS is due to the fact that SPLS tends to under-select.

For comparison of variable selection, we fit the entire data using the optimal penalty parameters selected by the 5-fold CV and kept track on whether each of 106 TF's is chosen in the fitted model. To account for the variability introduced by random partitions of CV, this experiment was done 100 times and we identified the TFs which were consistently included or excluded in each model. The "stability" results in Table 4 show how many TF's are selected at least 90 times, between 10 and 90 times, and under 10 times in the 100 runs. The results show that adaptive SRRR selects fewer TFs (60) than regular SRRR (69), and L_2SVS selects similar number of TFs (59) as adaptive SRRR. RemMap selects 94 out of the 106 TFs and SPLS selects the least number 12. The table also reports in parentheses how many of these consistently selected TFs are among the 21 experimentally confirmed TFs that are related to the cell cycle regulation (Wang, Chen & Li 2007). We can see that SRRR, adaptive SRRR and L_2SVS identify similar number of confirmed TFs, RemMap identifies slightly more, while SPLS identifies much less.

We remark on the stability of variable selection for each method. Ideally a TF should be either selected or excluded all the time, not affected by random partition in the cross-validation. The more such TF's fall in [90, 100] or [0, 10] categories, the more stable the model is. The higher number in the middle category (10, 90) indicates less stable selection. We see that L_2SVS and

SPLS are less stable in variable selection. Adaptive SRRR and RemMap have 0 TFs in the middle category and thus are more stable. However, a lower number in the middle category may also be a consequence of over-selection or under-selection indicated by extremely unequal numbers in the other two categories. After taking this into account, we conclude that the adaptive SRRR performs the best in terms of stability of variable selection. Since L_2SVS can be thought as a special case of SRRR without rank restriction, the fact that SRRR and adaptive SRRR outperform L_2SVS indicates that efficiency and stability of the variable selection is improved by enforcing a low rank structure on the coefficient matrix.

To further evaluate the variable selection result, we examined the chance of selecting irrelevant variables for each method when we randomly permuted the order of genes for the transcript level data to break the dependence between the transcript level and the TFs. We used 5-fold CV for selecting tuning parameters and then fit the entire data. We recorded the false positive rate, the ratio of the number of selected TFs to the total number of TFs. The permutation was done 100 times. Table 4 shows the average false positive rates. We see that SRRR methods have the smallest false positive rates.

The effect of the 16 confirmed TFs at 18 time points identified by adaptive SRRR are drawn in Figure 2. We see that except for MET31 and STE12, all TFs show some periodic effect and some of them such as SWI4 and SWI6 show two clear cycles. Lee et al. (2002) identified another 10 TFs which coordinate transcriptional regulation of the cell cycle with other cellular processes (metabolism, environmental responses and development). The adaptive SRRR can identify 8 of these 10 TFs and their effect is shown in the first two rows of Figure 3. Among the 36 additional TFs identified by adaptive SRRR, we rank them by their overall effect throughout the cell cycle and the last two rows in Figure 3 show the effect of the top 8 TFs. Again most of these TFs show some periodic effect and a lot of them show two clear cycles.

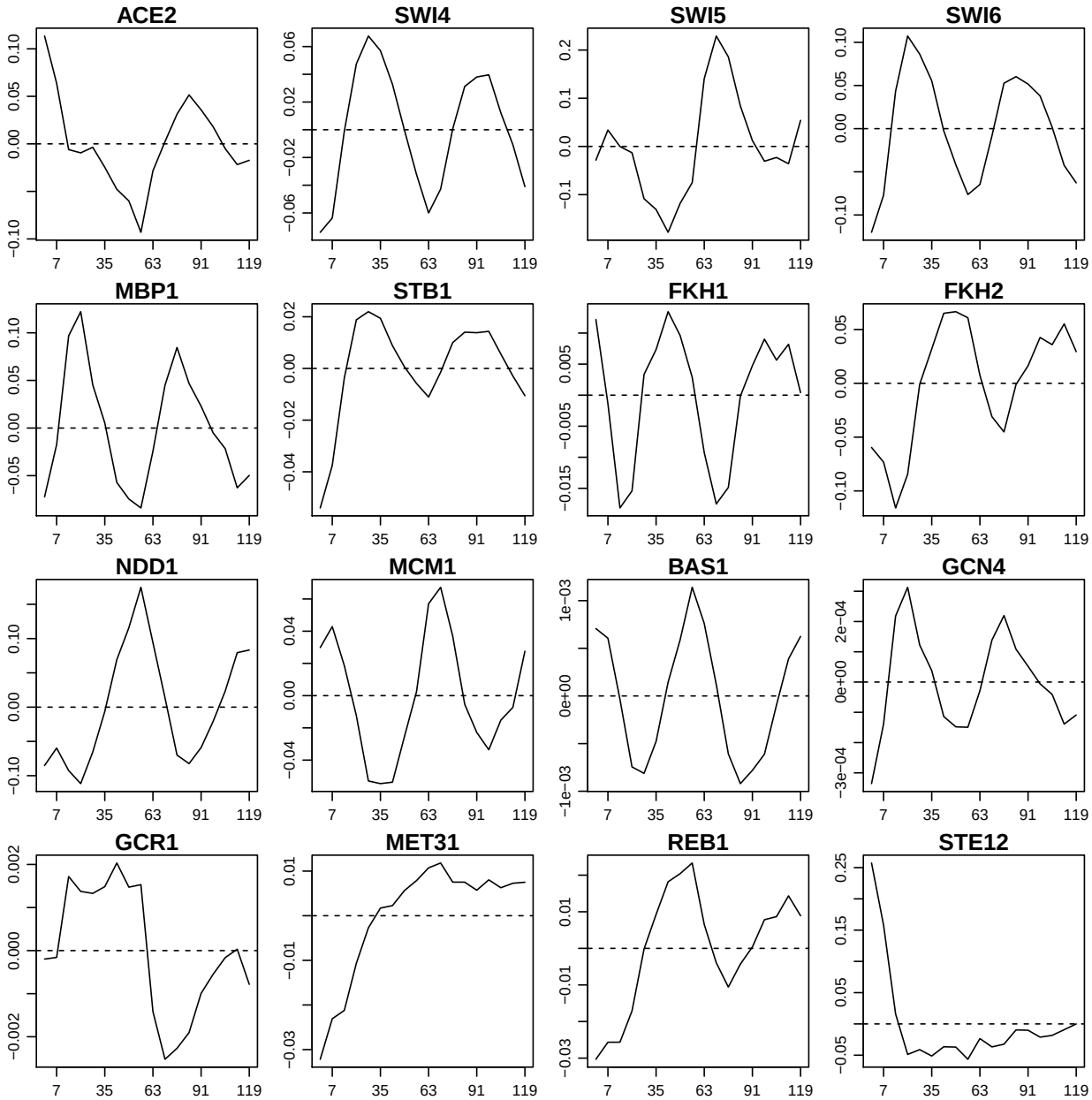

Figure 2: The effect of the 16 confirmed TFs identified by adaptive SRRR. These TFs are confirmed to be related to the cell cycle regulation.

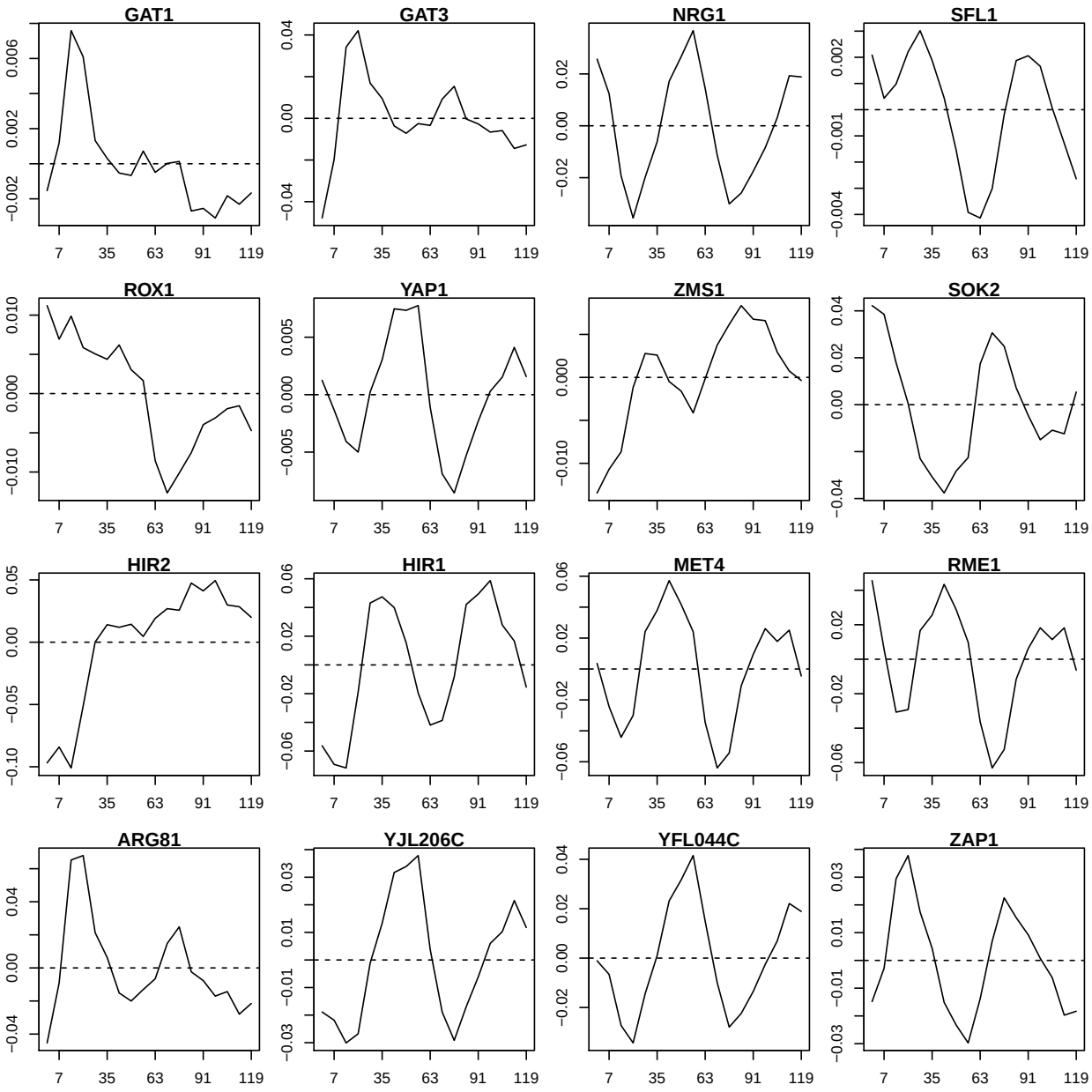

Figure 3: The effect of 16 of the additional 44 TFs identified adaptive SRRR. The first 8 TFs are confirmed TFs with the function of coordinating transcriptional regulation of the cell cycle with other cellular processes. The last 8 TFs have larger overall effect throughout the cell cycle among the rest TFs.

APPENDIX

Proof of Lemma 4. Writing $\mathbf{U} = \mathbf{U}_* + \mathbf{\Delta}_u$ and similarly for $\mathbf{D}$ and $\mathbf{V}$, we obtain the following expansion

$$\begin{aligned}\mathbf{U}\mathbf{D}\mathbf{V}^T - \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T &= (\mathbf{U}_* + \mathbf{\Delta}_u)(\mathbf{D}_* + \mathbf{\Delta}_d)(\mathbf{V}_* + \mathbf{\Delta}_v)^T - \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T \\ &= \mathbf{U}_*\mathbf{D}_*\mathbf{\Delta}_v^T + \mathbf{U}_*\mathbf{\Delta}_d\mathbf{V}_*^T + \mathbf{\Delta}_u\mathbf{D}_*\mathbf{V}_*^T \\ &\quad + \mathbf{\Delta}_u\mathbf{\Delta}_d\mathbf{V}_*^T + \mathbf{\Delta}_u\mathbf{D}_*\mathbf{\Delta}_v^T + \mathbf{U}_*\mathbf{\Delta}_d\mathbf{\Delta}_v^T + \mathbf{\Delta}_u\mathbf{\Delta}_d\mathbf{\Delta}_v^T.\end{aligned} \tag{25}$$

It follows from (17) that $\|\mathbf{\Delta}_u\| = \|\mathbf{U}_*\mathbf{G}_u + \mathbf{H}_u\| + O(\|\mathbf{G}_u\|^2 + \|\mathbf{H}_u\|^2) = O((\|\mathbf{G}_u\|^2 + \|\mathbf{H}_u\|^2)^{1/2})$. Similarly, $\|\mathbf{\Delta}_v\| = O((\|\mathbf{G}_v\|^2 + \|\mathbf{H}_v\|^2)^{1/2})$, and $\|\mathbf{\Delta}_d\| = O(\|\mathbf{K}\|)$. Note that $\|\mathbf{U}_*\|$, $\|\mathbf{D}_*\|$, and $\|\mathbf{V}_*\|$ are all bounded. Thus,

$$\|\mathbf{\Delta}_u\mathbf{\Delta}_d\mathbf{V}_*^T + \mathbf{\Delta}_u\mathbf{D}_*\mathbf{\Delta}_v^T + \mathbf{U}_*\mathbf{\Delta}_d\mathbf{\Delta}_v^T + \mathbf{\Delta}_u\mathbf{\Delta}_d\mathbf{\Delta}_v^T\| = O(\|\mathbf{G}_u\|^2 + \|\mathbf{H}_u\|^2 + \|\mathbf{G}_v\|^2 + \|\mathbf{H}_v\|^2 + \|\mathbf{K}\|^2).$$

Plugging (17)–(19) into (25) and using the anti-symmetry property $\mathbf{G}_v^T = -\mathbf{G}_v$, we obtain the desired result. ■

Proof of Lemma 5. Since $\mathbf{G}_u$ and $\mathbf{G}_v$ are antisymmetric, we can write $\mathbf{G}_u = \mathbf{\Delta} - \mathbf{\Delta}^T$ and $\mathbf{G}_v = \mathbf{\Gamma} - \mathbf{\Gamma}^T$, where $\mathbf{\Delta}$ and $\mathbf{\Gamma}$ are lower triangular square matrices whose diagonal elements are 0's. Therefore,

$$\mathbf{G}_u\mathbf{D}_* - \mathbf{D}_*\mathbf{G}_v = (\mathbf{\Delta}\mathbf{D}_* - \mathbf{D}_*\mathbf{\Gamma}) - (\mathbf{\Delta}^T\mathbf{D}_* - \mathbf{D}_*\mathbf{\Gamma}^T).$$

Since the two terms on the above right-hand side are respectively lower triangular and upper triangular, taking the Frobenius norm yields

$$f \equiv \|\mathbf{G}_u\mathbf{D}_* - \mathbf{D}_*\mathbf{G}_v\|^2 = \|\mathbf{\Delta}\mathbf{D}_* - \mathbf{D}_*\mathbf{\Gamma}\|^2 + \|\mathbf{\Delta}^T\mathbf{D}_* - \mathbf{D}_*\mathbf{\Gamma}^T\|^2.$$

Let δ_{ij} and γ_{ij} be the (i,j) entry of $\mathbf{\Delta}$ and $\mathbf{\Gamma}$ respectively. By the definition of the Frobenius norm and using the lower/upper triangular property of relevant matrices, we obtain that

$$\begin{aligned}f &= \sum_{i>j}(\delta_{ij}d_j - d_i\gamma_{ij})^2 + \sum_{i>j}(d_i\delta_{ij} - \gamma_{ij}d_j)^2 \\ &= \sum_{i>j}(d_j^2\delta_{ij}^2 + d_i^2\gamma_{ij}^2 + d_i^2\delta_{ij}^2 + d_j^2\gamma_{ij}^2 - 2d_id_j\delta_{ij}\gamma_{ij} - 2d_id_j\delta_{ij}\gamma_{ij}).\end{aligned}$$

For a lower bound of f, we use the inequality $2\delta_{ij}\gamma_{ij} \le \delta_{ij}^2 + \gamma_{ij}^2$ and complete the squares $d_i^2 + d_j^2 - 2d_id_j = (d_i - d_j)^2$ to get

$$f \ge \sum_{i>j}\{(d_i - d_j)^2(\delta_{ij}^2 + \gamma_{ij}^2)\} \ge \min_{i>j}(d_i - d_j)^2 \sum_{i>j}(\delta_{ij}^2 + \gamma_{ij}^2) = \frac{1}{2}\min_{i>j}(d_i - d_j)^2\,(\|\mathbf{G}_u\|^2 + \|\mathbf{G}_v\|^2).$$

For an upper bound, we use the inequality $(a-b)^2 \le 2(a^2+b^2)$ to obtain

$$f \le 2\sum_{i>j}(d_i^2+d_j^2)(\delta_{ij}^2+\gamma_{ij}^2) \le 2\max_{i<j}(d_i^2+d_j^2)\sum_{i>j}(\delta_{ij}^2+\gamma_{ij}^2) = 2\max_i d_i^2\,(\|\mathbf{G}_u\|^2+\|\mathbf{G}_v\|^2).$$

The proof of the lemma is complete. ■

Proof of Theorem 1. We first prove part (i). We only need show that (23) holds. Without loss of generality, we assume that $\mathbf{X}\mathbf{X}^T/n \to \mathbf{I}$, because if $\mathbf{X}\mathbf{X}^T/n \to \mathbf{\Sigma}$, then the argument below goes through when $\mathbf{X}$ is replaced by $\mathbf{\Sigma}^{-1/2}\mathbf{X}$. Set $\alpha_n = \alpha/\sqrt{n}$. We need find a uniform lower bound of $Q(\mathbf{\Theta}) - Q(\mathbf{\Theta}_*)$ for all $\mathbf{\Theta}$ with $\|\mathbf{\Theta}\| = \alpha_n$. Denote $\mathbf{E} = \mathbf{Y} - \mathbf{X}\mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T$. We have that

$$Q(\mathbf{\Theta}) - Q(\mathbf{\Theta}_*) = \|\mathbf{Y}-\mathbf{X}\mathbf{U}\mathbf{D}\mathbf{V}^T\|^2 - \|\mathbf{E}\|^2 + \sum_{i=1}^p \lambda_{n,i}\|\mathbf{U}^i\mathbf{D}\| - \sum_{i=1}^p \lambda_{n,i}\|\mathbf{U}_*^i\mathbf{D}_*\|. \tag{26}$$

We first consider the terms that involve penalty functions. Using $\|\mathbf{U}_i^i\mathbf{D}_*\| = 0$ for $i > p_0$ and the inequality $\|\mathbf{a}\| - \|\mathbf{b}\| \ge -\|\mathbf{a}-\mathbf{b}\|$, we obtain that

$$\begin{aligned}\sum_{i=1}^p \lambda_{n,i}\|\mathbf{U}^i\mathbf{D}\| - \sum_{i=1}^p \lambda_{n,i}\|\mathbf{U}_*^i\mathbf{D}_*\| &\ge \sum_{i=1}^{p_0} \lambda_{n,i}\|\mathbf{U}^i\mathbf{D}\| - \sum_{i=1}^{p_0} \lambda_{n,i}\|\mathbf{U}_*^i\mathbf{D}_*\| \\ &\ge -\sum_{i=1}^{p_0} \lambda_{n,i}\|\mathbf{U}^i\mathbf{D} - \mathbf{U}_*^i\mathbf{D}_*\|.\end{aligned} \tag{27}$$

It follows from the Cauchy–Schwarz inequality and a simpler version of Lemma 4 that

$$\begin{aligned}\sum_{i=1}^{p_0} \lambda_{n,i}\|\mathbf{U}^i\mathbf{D} - \mathbf{U}_*^i\mathbf{D}_*\| &\le (\max_{1\le i\le p_0}\lambda_{n,i})\sqrt{p_0}\|\mathbf{U}\mathbf{D} - \mathbf{U}_*\mathbf{D}_*\| \\ &\le \Big(\max_{1\le i\le p_0}\frac{\lambda_{n,i}}{\sqrt{n}}\Big)\sqrt{n}O(\|\mathbf{G}_u\| + \|\mathbf{H}_u\| + \|\mathbf{K}\|) \le O(\sqrt{n}\alpha_n).\end{aligned} \tag{28}$$

Next, consider the first two terms on the right side of (26). Since $\mathbf{Y} - \mathbf{X}\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{Y} - \mathbf{X}\mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T - \mathbf{X}(\mathbf{U}\mathbf{D}\mathbf{V}^T - \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T) = \mathbf{E} - \mathbf{X}(\mathbf{U}\mathbf{D}\mathbf{V}^T - \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T)$, we have that

$$\|\mathbf{Y}-\mathbf{X}\mathbf{U}\mathbf{D}\mathbf{V}^T\|^2 = \|\mathbf{E}\|^2 - 2\langle \mathbf{E}, \mathbf{X}(\mathbf{U}\mathbf{D}\mathbf{V}^T - \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T)\rangle + \|\mathbf{X}(\mathbf{U}\mathbf{D}\mathbf{V}^T - \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T)\|^2. \tag{29}$$

According to Lemma 4, we can write $\mathbf{U}\mathbf{D}\mathbf{V}^T - \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T = \mathbf{F} + \mathbf{T}$, where

$$\mathbf{F} = \mathbf{U}_*(\mathbf{G}_u\mathbf{D}_* - \mathbf{D}_*\mathbf{G}_v)\mathbf{V}_*^T + \mathbf{U}_*\mathbf{D}_*\mathbf{H}_v^T + \mathbf{H}_u\mathbf{D}_*\mathbf{V}_*^T + \mathbf{U}_*\mathbf{D}_*\mathbf{K}\mathbf{V}_*^T,$$

and $\|\mathbf{T}\| = O(\alpha_n^2)$. Lemma 5 and (21) imply that $\|\mathbf{F}\| = O(\alpha_n)$. By expanding the norm squared and using the Cauchy-Schwarz inequality, we obtain that

$$\|\mathbf{U}\mathbf{D}\mathbf{V}^T - \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T\|^2 = \|\mathbf{F}+\mathbf{T}\|^2 = \|\mathbf{F}\|^2 + O(\alpha_n^3).$$

Since $\mathbf{X}\mathbf{X}^T/n \to \mathbf{I}$, it follows that

$$\frac{1}{n}\|\mathbf{X}(\mathbf{U}\mathbf{D}\mathbf{V}^T - \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T)\|^2 = \|\mathbf{F}\|^2 + O(\alpha_n^3). \tag{30}$$

Note that

$$\langle \mathbf{E}, \mathbf{X}(\mathbf{U}\mathbf{D}\mathbf{V}^T - \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T)\rangle = \langle \mathbf{E}, \mathbf{X}\mathbf{F}\rangle + \langle \mathbf{E}, \mathbf{X}\mathbf{T}\rangle. \tag{31}$$

Since $\|\mathbf{X}\|^2/n = \mathrm{tr}(\mathbf{X}^T\mathbf{X})/n \to p$, we have that $\|\mathbf{X}\| = O(\sqrt{n})$. Simple calculation yields that, for any matrix $\mathbf{S}$ with compatible dimensions, $\mathrm{var}(\langle \mathbf{E}, \mathbf{X}\mathbf{S}\rangle) = O(\|\mathbf{X}\|^2\|\mathbf{S}\|^2)$, and thus $|\langle \mathbf{E}, \mathbf{X}\mathbf{S}\rangle| = O_P(\|\mathbf{X}\|\|\mathbf{S}\|) = O_P(\sqrt{n})\|\mathbf{S}\|$. Using this result, we obtain that

$$\langle \mathbf{E}, \mathbf{X}\mathbf{T}\rangle = O_P(\sqrt{n})\|\mathbf{T}\| = O_P(\sqrt{n}\alpha_n^2). \tag{32}$$

Combining (29)–(32) yields

$$\|\mathbf{Y} - \mathbf{X}\mathbf{U}\mathbf{D}\mathbf{V}^T\|^2 = \|\mathbf{E}\|^2 - 2\langle \mathbf{E}, \mathbf{X}\mathbf{F}\rangle + n\|\mathbf{F}\|^2 + O(n\alpha_n^3) + O_P(\sqrt{n}\alpha_n^2). \tag{33}$$

Applying the same argument for (32) gives

$$|\langle \mathbf{E}, \mathbf{X}\mathbf{F}\rangle| = O_P(\sqrt{n})\|\mathbf{F}\|. \tag{34}$$

Consequently,

$$\|\mathbf{Y} - \mathbf{X}\mathbf{U}\mathbf{D}\mathbf{V}^T\|^2 - \|\mathbf{E}\|^2 \geq -O_P(\sqrt{n})\|\mathbf{F}\| + n\|\mathbf{F}\|^2 + O(n\alpha_n^3) + O_P(\sqrt{n}\alpha_n^2). \tag{35}$$

Combining (26)–(28) and (35), we obtain that

$$Q(\mathbf{\Theta}) - Q(\mathbf{\Theta}_*) \geq -O_P(\sqrt{n})\|\mathbf{F}\| + n\|\mathbf{F}\|^2 - O(\sqrt{n}\alpha_n) + O(n\alpha_n^3) + O_P(\sqrt{n}\alpha_n^2).$$

Lemma 5 and (21) together imply that, $M_1\|\mathbf{\Theta}\| \leq \|\mathbf{F}\| \leq M_2\|\mathbf{\Theta}\|$ for some positive constants M_1 and M_2. Thus, if $\|\mathbf{\Theta}\| = \alpha_n = \alpha/\sqrt{n}$,

$$Q(\mathbf{\Theta}) - Q(\mathbf{\Theta}_*) \geq -O_P(\alpha) + M_1\alpha^2 + O(\alpha^3/\sqrt{n}) + O_P(\alpha^2/\sqrt{n}).$$

Therefore, for any $\epsilon > 0$, we can choose a large α such that, $\inf_{\mathbf{\Theta}:\|\mathbf{\Theta}\|=\alpha_n} Q(\mathbf{\Theta}) - Q(\mathbf{\Theta}_*) > 0$ with probability larger than $1-\epsilon$. The proof of part (i) is complete.

To prove part (ii), note that part (i) implies that $\|\widehat{\mathbf{\Theta}}\| = O_P(n^{-1/2})$. Since $\|\widehat{\mathbf{\Theta}}\|^2 = \|\widehat{\mathbf{G}}_u\|^2 + \|\widehat{\mathbf{H}}_u\|^2 + \|\widehat{\mathbf{G}}_v\|^2 + \|\widehat{\mathbf{H}}_v\|^2 + \|\widehat{\mathbf{K}}\|^2$, we obtain that $(\|\widehat{\mathbf{G}}_u\|^2 + \|\widehat{\mathbf{H}}_u\|^2)^{1/2} = O_P(n^{-1/2})$, $(\|\widehat{\mathbf{G}}_v\|^2 + \|\widehat{\mathbf{H}}_v\|^2)^{1/2} = O_P(n^{-1/2})$, and $\|\widehat{\mathbf{K}}\| = O_p(n^{-1/2})$. The desired result then follows from Lemma 3. ■

Proof of Theorem 2.. Denote the diagonal elements of $\mathbf{D}_*$ as $\mathbf{D}_{*kk}$, $1 \le k \le r$, and similarly for $\widehat{\mathbf{D}}$. Because $\mathbf{D}_{*kk} > 0$ and $\widehat{\mathbf{D}}$ is consistent in estimating $\mathbf{D}_*$, we have that $\widehat{\mathbf{D}}_{kk} > 0$ for all k. The optimizing objective function can be written as

$$\|\mathbf{Y}-\mathbf{X}\mathbf{U}\mathbf{D}\mathbf{V}^T\|^2+\sum_{i=1}^{p}\lambda_{n,i}\|\mathbf{U}^i\mathbf{D}\| = \mathrm{tr}(\mathbf{Y}^T\mathbf{Y})+\mathrm{tr}(\mathbf{D}\mathbf{U}^T\mathbf{X}^T\mathbf{X}\mathbf{U}\mathbf{D}-2\mathbf{V}^T\mathbf{Y}^T\mathbf{X}\mathbf{U}\mathbf{D})+\sum_{i=1}^{p}\lambda_{n,i}\|\mathbf{U}^i\mathbf{D}\|.$$

Suppose $\|\widehat{\mathbf{U}}^i\| > 0$. The first order condition for $\widehat{\mathbf{U}}^i$ is

$$2\mathbf{X}_i^T\mathbf{X}\widehat{\mathbf{U}}\widehat{\mathbf{D}}^2 - 2\mathbf{X}_i^T\mathbf{Y}\widehat{\mathbf{V}}\widehat{\mathbf{D}} + \lambda_{n,i}\frac{\widehat{\mathbf{U}}^i\widehat{\mathbf{D}}^2}{\|\widehat{\mathbf{U}}^i\widehat{\mathbf{D}}\|} = \mathbf{0},$$

or equivalently,

$$\frac{2}{\sqrt{n}}\mathbf{X}_i^T(\mathbf{X}\widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^T - \mathbf{Y})\widehat{\mathbf{V}} + \frac{\lambda_{n,i}}{\sqrt{n}}\frac{\widehat{\mathbf{U}}^i\widehat{\mathbf{D}}}{\|\widehat{\mathbf{U}}^i\widehat{\mathbf{D}}\|} = \mathbf{0}. \tag{36}$$

Note that $(1/\sqrt{n})\mathbf{X}_i^T(\mathbf{X}\mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T-\mathbf{Y}) = -(1/\sqrt{n})\mathbf{X}_i^T\mathbf{E} = O_p(1)$, $\widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^T-\mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T = O_p(1/\sqrt{n})$, and $\mathbf{X}^T\mathbf{X}/n = O_P(1)$. Therefore

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\mathbf{X}_i^T(\mathbf{X}\widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^T - \mathbf{Y}) \\
&\quad= \frac{1}{\sqrt{n}}\mathbf{X}_i^T(\mathbf{X}\mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T - \mathbf{Y}) + \frac{1}{\sqrt{n}}\mathbf{X}_i^T\mathbf{X}(\widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^T - \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T) \\
&\quad= O_p(1) + \frac{1}{\sqrt{n}}\mathbf{X}_i^T\mathbf{X}O_p(\frac{1}{\sqrt{n}}) = O_p(1).
\end{aligned}$$

This implies that

$$\frac{1}{\sqrt{n}}\mathbf{X}_i^T(\widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^T - \mathbf{Y})\mathbf{V} = O_p(1). \tag{37}$$

If $\|\widehat{\mathbf{U}}^i\| > 0$ for some $i > p_0$, then letting

$$k^* = \arg\max_{1\le k\le r}|\widehat{\mathbf{U}}_k^i\widehat{\mathbf{D}}_{kk}|,$$

we have that

$$\frac{|\widehat{\mathbf{U}}_{k^*}^i\widehat{\mathbf{D}}_{k^*k^*}|}{\|\widehat{\mathbf{U}}^i\widehat{\mathbf{D}}\|} > \frac{1}{\sqrt{r}}. \tag{38}$$

Considering (37), (38), and the assumption that $\lambda_{n,i}/\sqrt{n} \to \infty$, for $i > p_0$, we obtain that the first order condition (36) won't hold for $i > p_0$. This is a contradiction. Consequently, $\|\widehat{\mathbf{U}}^i\| = 0$ for all $i > p_0$ with probability tending to one. ■

REFERENCES

Anderson, T. W. (1999), "Asymptotic distribution of the reduced rank regression estimator under general conditions," The Annals of Statistics, 27, 1141–1154.

Chun, H., & Keles, S. (2010), "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," Journal of the Royal Statistical Society: Series B, 72, 3–25.

Edelman, A., Arias, T. A., & Smith, S. T. (1998), "The Geometry of algorithms with orthogonality constraints," SIAM Journal on Matrix Analysis and Applications, 20, 303–353.

Friedman, J., Hastie, T., Hofling, H., , & Tibshirani, R. (2007), "Pathwise Coordinate Optimization," The Annals of Applied Statistics, 1, 302–332.

Gower, J. C., & Dijksterhuis, G. B. (2004), Procrustes Problems, New York: Oxford University Press.

Izenman, A. J. (1975), "Reduced-rank regression for the multivariate linear model," Journal of Multivariate Analysis, 5, 248–264.

Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., Simon, I., Zeitlinger, J., Jennings, E., Murray, H., Gordon, D., Ren, B., Wyrick, J., Tagne, J., Volkert, T., Fraenkel, E., Gifford, D., & Young, R. (2002), "Transcriptional regulatory networks in Saccharomyces cerevisiae," Science, 298(5594), 799.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D., Pollack, J. R., & Wang, P. (2010), "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer," Annuls of Applied Statistics, 4, 53–77.

Reinsel, G. C., & Velu, R. P. (1998), Multivariate Reduced-Rank Regression: Theory and Applications, : Springer.

Rothman, A., Levina, E., & Zhu, J. (2010), "Sparse multivariate regression with covariance estimation," Journal of Computational and Graphical Statistics, 19(4), 947–962.

Simila, T., & Tikka, J. (2007), "Input selection and shrinkage in multiresponse linear regression," Computational Statistics & Data Analysis, 52, 406–422.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., & Futcher, B. (1998), "Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," Molecular biology of the cell, 9(12), 3273–3297.

Tibshirani, R. J. (1996), "Regression shrinkage and selection via the lasso," J. Roy. Statist. Soc. Ser. B, 58, 267–288.

Turlach, B., Venables, W., & Wright, S. (2005), "Simultaneous variable selection," Technometrics, 47, 350–363.

Wang, H., & Leng, C. (2008), "A note on adaptive group lasso," Computational Statistics & Data Analysis, 52, 5277–5286.

Wang, L., Chen, G., & Li, H. (2007), "Group SCAD regression analysis for microarray time course gene expression data," Bioinformatics, 23(12), 1486–1494.

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

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

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," Journal of the American Statistical Association, 101, 1418–1429.

Zou, H., & Zhang, H. (2009), "On the adaptive elastic-net with a diverging number of parameters," Annals of statistics, 37(4), 1733.

Table 1: Simulation Examples (Section 6.3). Mean Squared Errors (MSE) for seven setups. Reported are the means and SEs (in parentheses) based on 30 simulation runs.

$n = 100$			Parameters			Mean Squared Error							
p	p_0	q	r	ρ_x	ρ_e	SRRR	L_2SVS	RemMap	SPLS	L_∞SVS	SepLasso	aSRRR	aL_2SVS
30	10	10	3	0	0	0.044	0.087	0.096	0.085	0.100	0.123	0.038	0.064
						(0.002)	(0.002)	(0.003)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
			10	0	0	0.077	0.085	0.092	0.085	0.120	0.111	0.061	0.062
						(0.003)	(0.003)	(0.003)	(0.004)	(0.002)	(0.003)	(0.002)	(0.002)
			3	.5	0	0.045	0.066	0.066	0.079	0.074	0.101	0.044	0.062
						(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
			3	0	.5	0.053	0.09	0.099	0.099	0.104	0.12	0.047	0.069
						(0.002)	(0.003)	(0.003)	(0.006)	(0.003)	(0.003)	(0.002)	(0.002)
100	30	10	3	0	0	0.132	0.208	0.208	0.269	0.236	0.274	0.136	0.207
						(0.005)	(0.005)	(0.004)	(0.009)	(0.004)	(0.004)	(0.006)	(0.005)
300	30	30	3	0	0	0.091	0.223	0.228	0.223	0.253	0.334	0.092	0.181
						(0.002)	(0.005)	(0.004)	(0.009)	(0.004)	(0.004)	(0.006)	(0.004)
100	100	10	10	0	0	0.26	0.261	0.238	0.362	0.353	0.247	0.294	0.306
						(0.003)	(0.003)	(0.004)	(0.004)	(0.002)	(0.004)	(0.004)	(0.005)

Table 2: Simulation Examples (Section 6.3). Average Sensitivity (Se) and Specificity (Sp) over 30 simulation runs.

$n = 100$			Parameters																	
p	p_0	q	r	ρ_x	ρ_e	SRRR		L_2SVS		RemMap		SPLS		L_∞SVS		SepLasso		aSRRR		aL_2SVS	
						Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp
30	10	10	3	0	0	1	0.42	1	0.4	1	0.23	0.94	0.94	1	0.1	1	0.02	1	0.82	1	0.74
			10	0	0	1	0.35	1	0.34	1	0.16	0.97	0.95	1	0.09	1	0.02	1	0.8	1	0.75
			3	.5	0	0.99	0.62	1	0.52	1	0.46	0.9	0.94	1	0.29	1	0.08	0.98	0.69	0.96	0.73
			3	0	.5	1	0.5	1	0.37	1	0.24	0.92	0.92	1	0.13	1	0.03	0.99	0.79	1	0.77
100	30	10	3	0	0	0.92	0.57	0.87	0.51	0.89	0.5	0.56	0.81	0.9	0.46	0.96	0.23	0.84	0.79	0.71	0.88
300	30	30	3	0	0	0.93	0.74	0.83	0.68	0.93	0.41	0.51	0.98	0.85	0.64	0.96	0.25	0.89	0.82	0.68	0.96
100	100	10	10	0	0	0.71	NA	0.77	NA	0.86	NA	0.6	NA	0.68	NA	0.89	NA	0.38	NA	0.41	NA

Table 3: Simulation Examples (Section 6.3). Average number of factors selected for the SRRR, adaptive SRR and SPLS, based on 30 simulation runs.

$n = 100$			Parameters			Number of Factors		
p	p_0	q	r	ρ_x	ρ_e	SRRR	SPLS	aSRRR
30	10	10	3	0	0	3	5	3.1
			10	0	0	7.1	8.6	7.4
			3	.5	0	3.1	3.2	3.4
			3	0	.5	3	5.1	3
100	30	10	3	0	0	3.1	4.2	3
300	30	30	3	0	0	3	7.2	3.1
100	100	10	10	0	0	9.3	5.2	9

Table 4: Yeast Cell Cycle Data. Mean squared prediction error (MSPE) is based on half-splitting the data into training and test sets. Mean and SE of test errors are reported. Stability results report the times of each TF being selected over 100 runs of 5-fold CV. False positive rate is the average percent of false selection over 100 random permutations.

		SRRR	aSRRR	L_2SVS	RemMap	SPLS
MSPE	mean	.189	.188	.190	.189	.197
	(SE)	(.001)	(.001)	(.001)	(.001)	(.001)
Stability	[90, 100]	69	60	59	94	12
		(16)	(16)	(17)	(21)	(7)
	(10, 90)	10	0	36	0	24
	[0, 10]	27	46	11	12	70
False Positive		2.10	1.74	2.91	4.39	7.36