SOFAR: Large-Scale Association Network Learning *

Yoshimasa Uematsu, Yingying Fan, Kun Chen, Jinchi Lv and Wei Lin

March 28, 2019

Abstract

Many modern big data applications feature large scale in both numbers of responses and predictors. Better statistical efficiency and scientific insights can be enabled by understanding the large-scale response-predictor association network structures via layers of sparse latent factors ranked by importance. Yet sparsity and orthogonality have been two largely incompatible goals. To accommodate both features, in this paper we suggest the method of sparse orthogonal factor regression (SOFAR) via the sparse singular value decomposition with orthogonality constrained optimization to learn the underlying association networks, with broad applications to both unsupervised and supervised learning tasks such as biclustering with sparse singular value decomposition, sparse principal component analysis, sparse factor analysis, and spare vector autoregression analysis. Exploiting the framework of convexity-assisted nonconvex optimization, we derive nonasymptotic error bounds for the suggested procedure characterizing the theoretical advantages. The statistical guarantees are powered by an efficient SOFAR algorithm with convergence property. Both computational and theoretical advantages of our procedure are demonstrated with several simulations and real data examples.

Running title: SOFAR

Key words: Big data; Large-scale association network; Simultaneous response and predictor selection; Latent factors; Sparse singular value decomposition; Orthogonality constrained optimization; Nonconvex statistical learning

*Yoshimasa Uematsu is Assistant Professor, Department of Economics and Management, Tohoku University, Sendai 980-8576, Japan (E-mail: *yoshimasa.uematsu.e7@tohoku.ac.jp*). Yingying Fan is Dean's Associate Professor in Business Administration, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089 (E-mail: *fanyingy@marshall.usc.edu*). Kun Chen is Associate Professor, Department of Statistics, University of Connecticut, Storrs, CT 06269 (E-mail: *kun.chen@uconn.edu*). Jinchi Lv is Kenneth King Stonier Chair in Business Administration and Professor, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089 (E-mail: *jinchilv@marshall.usc.edu*). Wei Lin is Assistant Professor, School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, China 100871 (E-mail:*weilin@math.pku.edu.cn*). This work was supported by Grant-in-Aid for JSPS Fellows 26-1905, NIH Grant 1R01GM131407-01, NSF CAREER Awards DMS-0955316 and DMS-1150318, NIH grant U01 HL114494, NSF grant DMS-1613295, a grant from the Simons Foundation, Adobe Data Science Research Award, and NSFC grants 11671018 and 71532001. Most of this work was completed while Uematsu visited USC Marshall as a JSPS Overseas Research Fellow and Postdoctoral Scholar. Part of this work was completed while Fan and Lv visited the Departments of Statistics at University of California, Berkeley and Stanford University. These authors sincerely thank both departments for their hospitality. The authors also would like to thank the Associate Editor and referees for their valuable comments that helped improve the article substantially.

1 Introduction

The genetics of gene expression variation may be complex due to the presence of both local and distant genetic effects and shared genetic components across multiple genes [14, 18]. A useful statistical analysis in such studies is to simultaneously classify the genetic variants and gene expressions into groups that are associated. For example, in a yeast expression quantitative trait loci (eQTLs) mapping analysis, the goal is to understand how the eQTLs, which are regions of the genome containing DNA sequence variants, influence the expression level of genes in the yeast MAPK signaling pathways. Extensive genetic and biochemical analysis has revealed that there are a few functionally distinct signaling pathways of genes [36, 14], suggesting that the association structure between the eQTLs and the genes is of low rank. Each signaling pathway involves only a subset of genes, which are regulated by only a few genetic variants, suggesting that each association between the eQTLs and the genes is sparse in both the input and the output (or in both the responses and the predictors), and the pattern of sparsity should be pathway specific. Moreover, it is known that the yeast MAPK pathways regulate and interact with each other [36]. The complex genetic structures described above clearly call for a joint statistical analysis that can reveal multiple distinct associations between subsets of genes and subsets of genetic variants. If we treat the genetic variants and gene expressions as the predictors and responses, respectively, in a multivariate regression model, the task can then be carried out by seeking a sparse representation of the coefficient matrix and performing predictor and response selection simultaneously. The problem of large-scale response-predictor association network learning is indeed of fundamental importance in many modern big data applications featuring large scale in both numbers of responses and predictors.

Observing n independent pairs $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, \cdots, n$, with $\mathbf{x}_i \in \mathbb{R}^p$ the covariate vector and $\mathbf{y}_i \in \mathbb{R}^q$ the response vector, motivated from the above applications we consider the following multivariate regression model

$$\mathbf{Y} = \mathbf{X}\mathbf{C}^* + \mathbf{E}, \tag{1}$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T \in \mathbb{R}^{n\times q}$ is the response matrix, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times p}$ is the predictor matrix, $\mathbf{C}^* \in \mathbb{R}^{p\times q}$ is the true regression coefficient matrix, and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)^T$ is the error matrix. To model the sparse relationship between the responses and the predictors as in the yeast eQTLs mapping analysis, we exploit the following singular value decomposition (SVD) of the coefficient matrix

$$\mathbf{C}^* = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*T} = \sum_{j=1}^{r} d_j^* \mathbf{u}_j^* \mathbf{v}_j^{*T}, \tag{2}$$

where $1 \le r \le \min(p, q)$ is the rank of matrix $\mathbf{C}^*$, $\mathbf{D}^* = \mathrm{diag}(d_1^*, \ldots, d_r^*)$ is a diagonal matrix of nonzero singular values, and $\mathbf{U}^* = (\mathbf{u}_1^*, \ldots, \mathbf{u}_r^*) \in \mathbb{R}^{p\times r}$ and $\mathbf{V}^* = (\mathbf{v}_1^*, \ldots, \mathbf{v}_r^*) \in \mathbb{R}^{q\times r}$ are the orthonormal matrices of left and right singular vectors, respectively. Here, we assume that $\mathbf{C}^*$ is low-rank with only r nonzero singular values, and the matrices $\mathbf{U}^*$ and $\mathbf{V}^*$ are sparse.

Under the sparse SVD structure (2), model (1) can be rewritten as

$$\widetilde{\mathbf{Y}} = \widetilde{\mathbf{X}}\mathbf{D}^* + \widetilde{\mathbf{E}},$$

where $\widetilde{\mathbf{Y}} = \mathbf{Y}\mathbf{V}^*$, $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{U}^*$, and $\widetilde{\mathbf{E}} = \mathbf{E}\mathbf{V}^* \in \mathbb{R}^{n\times r}$ are the matrices of latent responses, predictors, and random errors, respectively. The associations between the predictors and responses are thus diagonalized under the pairs of transformations specified by $\mathbf{U}^*$ and $\mathbf{V}^*$. When $\mathbf{C}^*$ is of low rank, this provides an appealing *low-dimensional latent model* interpretation for model (1). Further, note that the latent responses and predictors are linear combinations of the original responses and predictors, respectively. Thus, the interpretability of the SVD can be enhanced if we require that the left and right singular vectors be sparse so that each latent predictor/response involves only a small number of the original predictors/responses, thereby performing the task of variable selection among the predictors/responses, as needed in the yeast eQTLs analysis.

The above model (1) with low-rank coefficient matrix has been commonly adopted in the literature. In particular, the reduced rank regression [1, 39, 55] is an effective approach to dimension reduction by constraining the coefficient matrix $\mathbf{C}^*$ to be of low rank. Bunea et al. [15] proposed a rank selection criterion that can be viewed as an L_0 regularization on the singular values of $\mathbf{C}^*$. The popularity of L_1 regularization methods such as the Lasso [60] led to the development of nuclear norm regularization in multivariate regression [66]. Chen et al. [21] proposed an adaptive nuclear norm penalization approach to bridge the gap between L_0 and L_1 regularization methods and combine some of their advantages. With the additional SVD structure (2), Chen et al. [19] proposed a new estimation method with a correctly specified rank by imposing a weighted L_1 penalty on each rank-1 SVD layer for the classical setting of fixed dimensionality. Chen and Huang [22] and Bunea et al. [16] explored a low-rank representation of $\mathbf{C}^*$ in which the rows of $\mathbf{C}^*$ are sparse; however, their approaches do not impose sparsity on the right singular vectors and, hence, are inapplicable to settings with high-dimensional responses where response selection is highly desirable.

Recently, there have been some new developments in sparse and low-rank regression problems. Ma and Sun [50] studied the properties of row-sparse reduced-rank regression model with nonconvex sparsity-inducing penalties, and later Ma et al. [49] extended their work to two-way sparse reduced-rank regression. Chen and Huang [23] extended the row-sparse reduced-rank regression by incorporating covariance matrix estimation, and the authors mainly focused on computational issues. Lian et al. [46] proposed a semiparametric reduced-rank regression with a sparsity penalty on the coefficient matrix itself. Goh et al. [33] studied the Bayesian counterpart of the row/column-sparse reduced-rank regression and established its posterior consistency. However, none of these works considered the possible entry-wise sparsity in the SVD of the coefficient matrix. The sparse and low-rank regression models have also been applied in various fields to solve important scientific problems. To name a few, Chen et al. [20] applied a sparse and low-rank bi-linear model for the task of source-sink reconstruction in marine ecology, Zhu et al. [70] used a Bayesian low-rank model for associating neuroimaging phenotypes and genetic

markers, and Ma et al. [48] used a threshold SVD regression model for learning regulatory relationships in genomics.

In view of the key role that the sparse SVD plays for simultaneous dimension reduction and variable selection in model (1), in this paper we suggest a unified regularization approach to estimating such a sparse SVD structure. Our proposal successfully meets three key methodological challenges that are posed by the complex structural constraints on the SVD. First, sparsity and orthogonality are two largely incompatible goals and would seem difficult to be accommodated within a single framework. For instance, a standard orthogonalization process such as QR factorization will generally destroy the sparsity pattern of a matrix. Previous methods either relaxed the orthogonality constraint to allow efficient search for sparsity patterns [19], or avoided imposing both sparsity and orthogonality requirements on the same factor matrix [22, 16]. To resolve this issue, we formulate our approach as an orthogonality constrained regularization problem, which yields *simultaneously sparse and orthogonal* factor matrices in the SVD. Second, we employ the nuclear norm penalty to encourage sparsity among the singular values and achieve rank reduction. As a result, our method produces a continuous solution path, which facilitates rank parameter tuning and distinguishes it from the L_0 regularization method adopted by Bunea et al. [16]. Third, unlike rank-constrained estimation, the nuclear norm penalization approach makes the estimation of singular vectors more intricate, since one does not know a priori which singular values will vanish and, hence, which pairs of left and right singular vectors are unidentifiable. Noting that the degree of identifiability of the singular vectors increases with the singular value, we propose to penalize the singular vectors weighted by singular values, which proves to be meaningful and effective. Combining these aspects, we introduce *sparse orthogonal factor regression* (SOFAR), a novel regularization framework for high-dimensional multivariate regression. While respecting the orthogonality constraint, we allow the sparsity-inducing penalties to take a general, flexible form, which includes special cases that adapt to the entrywise and rowwise sparsity of the singular vector matrices, resulting in a nonconvex objective function for the SOFAR method.

In addition to the aforementioned three methodological challenges, the nonconvexity of the SOFAR objective function also poses important algorithmic and theoretical challenges in obtaining and characterizing the SOFAR estimator. To address these challenges, we suggest a two-step approach exploiting the framework of convexity-assisted nonconvex optimization (CANO) to obtain the SOFAR estimator. More specifically, in the first step we minimize the L_1-penalized squared loss for the multivariate regression (1) to obtain an initial estimator. Then in the second step, we minimize the SOFAR objective function in an asymptotically shrinking neighborhood of the initial estimator. Thanks to the convexity of its objective function, the initial estimator can be obtained effectively and efficiently. Yet since the finer sparsity structure imposed through the sparse SVD (2) is completely ignored in the first step, the initial estimator meets none of the aforementioned three methodological challenges. Nevertheless, since it is theoretically guaranteed that the initial estimator is not far away from the true coefficient matrix $\mathbf{C}^*$ with asymptotic probability one, searching in an asymptotically shrinking neighborhood of the initial estimator significantly alleviates the nonconvexity issue of the SOFAR objective function. In fact, under

the framework of CANO we derive nonasymptotic bounds for the prediction, estimation, and variable selection errors of the SOFAR estimator characterizing the theoretical advantages. In implementation, to disentangle the sparsity and orthogonality constraints we develop an efficient SOFAR algorithm and establish its convergence properties.

Our suggested SOFAR method for large-scale association network learning is in fact connected to a variety of statistical methods in both unsupervised and supervised multivariate analysis. For example, the sparse SVD and sparse principal component analysis (PCA) for a high-dimensional data matrix can be viewed as unsupervised versions of our general method. Other prominent examples include sparse factor models, sparse canonical correlation analysis [63], and sparse vector autoregressive (VAR) models for high-dimensional time series. See Section 2.2 for more details on these applications and connections.

The rest of the paper is organized as follows. Section 2 introduces the SOFAR method and discusses its applications to several unsupervised and supervised learning tasks. We present the nonasymptotic properties of the method in Section 3. Section 4 develops an efficient optimization algorithm and discusses its convergence and tuning parameter selection. We provide several simulation and real data examples in Section 5. All the proofs of main results and technical details are detailed in the Supplementary Material. An associated R package implementing the suggested method is available at https://cran.r-project.org/package=rrpack.

2 Large-scale association network learning via SOFAR

2.1 Sparse orthogonal factor regression

To estimate the sparse SVD of the true regression coefficient matrix $\mathbf{C}^*$ in model (1), we start by considering an estimator of the form $\mathbf{UDV}^T$, where $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_m) \in \mathbb{R}^{m\times m}$ with $d_1 \geq \cdots \geq d_m \geq 0$ and $1 \leq m \leq \min\{p, q\}$ is a diagonal matrix of singular values, and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_m) \in \mathbb{R}^{p\times m}$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_m) \in \mathbb{R}^{q\times m}$ are orthonormal matrices of left and right singular vectors, respectively. Although it is always possible to take $m = \min(p, q)$ without prior knowledge of the rank r, it is often sufficient in practice to take a small m that is slightly larger than the expected rank (estimated by some procedure such as in Bunea et al. [15]), which can dramatically reduce computation time and space. Throughout the paper, for any matrix $\mathbf{M} = (m_{ij})$ we denote by $\|\mathbf{M}\|_F$, $\|\mathbf{M}\|_1$, $\|\mathbf{M}\|_\infty$, and $\|\mathbf{M}\|_{2,1}$ the Frobenius norm, entrywise L_1-norm, entrywise L_∞-norm, and rowwise $(2,1)$-norm defined, respectively, as $\|\mathbf{M}\|_F = \big(\sum_{i,j} m_{ij}^2\big)^{1/2}$, $\|\mathbf{M}\|_1 = \sum_{i,j} |m_{ij}|$, $\|\mathbf{M}\|_\infty = \max_{i,j} |m_{ij}|$, and $\|\mathbf{M}\|_{2,1} = \sum_i \big(\sum_j m_{ij}^2\big)^{1/2}$. We also denote by $\|\cdot\|_2$ the induced matrix norm (operator norm).

As mentioned in the Introduction, we employ the nuclear norm penalty to encourage sparsity among the singular values, which is exactly the entrywise L_1 penalty on $\mathbf{D}$. Penalization directly on $\mathbf{U}$ and $\mathbf{V}$, however, is inappropriate since the singular vectors are not equally identifiable and should not be subject to the same amount of regularization. Singular vectors corresponding to larger singular values can be estimated more accurately and should contribute more to the regularization, whereas those corresponding to vanishing singular values are unidentifiable and should play no role in the regularization. Therefore,

we propose an *importance weighting* by the singular values and place sparsity-inducing penalties on the weighted versions of singular vector matrices, $\mathbf{UD}$ and $\mathbf{VD}$. Note also that our goal is to estimate not only the low-rank matrix $\mathbf{C}^*$ but also the factor matrices $\mathbf{D}^*$, $\mathbf{U}^*$, and $\mathbf{V}^*$. Taking into account these points, we consider the orthogonality constrained optimization problem

$$(\widehat{\mathbf{D}}, \widehat{\mathbf{U}}, \widehat{\mathbf{V}}) = \underset{\mathbf{D},\mathbf{U},\mathbf{V}}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{XUDV}^T\|_F^2 + \lambda_d\|\mathbf{D}\|_1 + \lambda_a\rho_a(\mathbf{UD}) + \lambda_b\rho_b(\mathbf{VD}) \right\} \quad \text{subject to} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}_m, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}_m, \tag{3}$$

where $\rho_a(\cdot)$ and $\rho_b(\cdot)$ are penalty functions to be clarified later, and $\lambda_d, \lambda_a, \lambda_b \geq 0$ are tuning parameters that control the strengths of regularization. We call this regularization method *sparse orthogonal factor regression* (SOFAR) and the regularized estimator $(\widehat{\mathbf{D}}, \widehat{\mathbf{U}}, \widehat{\mathbf{V}})$ the SOFAR estimator. Note that $\rho_a(\cdot)$ and $\rho_b(\cdot)$ can be equal or distinct, depending on the scientific question and the goals of variable selection. Letting $\lambda_d = \lambda_b = 0$ while setting $\rho_a(\cdot) = \|\cdot\|_{2,1}$ reduces the SOFAR estimator to the sparse reduced-rank estimator of Chen and Huang [22]. In view of our choices of $\rho_a(\cdot)$ and $\rho_b(\cdot)$, although $\mathbf{D}$ appears in all three penalty terms, rank reduction is achieved mainly through the first term, while variable selection is achieved through the last two terms under necessary scalings by $\mathbf{D}$.

We note that one major advantage of SOFAR is that the final estimates satisfy the orthogonality constraints in (3). This is also the major distinction of our method from many existing ones. The orthogonality constraints are motivated from a combination of practical, methodological, and theoretical considerations. On the practical side, the orthogonality constraints maximize the separation of different latent layers, ensure that the importance of these layers can be measured by the magnitudes of diagonals in $\mathbf{D}^*$, and thus enhance the interpretation. On the methodological side, they are a natural, convenient way to ensure the identifiability of the factor matrices $\mathbf{D}^*$, $\mathbf{U}^*$, and $\mathbf{V}^*$ [7]. On the theoretical side, they allow us to establish rigorous error bound inequalities for the estimates $\widehat{\mathbf{D}}$, $\widehat{\mathbf{A}}$, and $\widehat{\mathbf{B}}$. Nevertheless the orthogonality condition among the sparse latent factors may not hold exactly in certain real applications. Thus in Section 5, we will investigate the robustness of our method through simulation studies where the orthogonality condition in the model is violated. It would be interesting to formally study the scenario when the orthogonality condition may hold approximately, which is beyond the scope of the current paper and we leave it for future research.

Note that for simplicity we do not explicitly state the ordering constraint $d_1 \geq \cdots \geq d_m \geq 0$ in optimization problem (3). In fact, when $\rho_a(\cdot)$ and $\rho_b(\cdot)$ are matrix norms that satisfy certain invariance properties, such as the entrywise L_1-norm and rowwise $(2,1)$-norm, this constraint can be easily enforced by simultaneously permuting and/or changing the signs of the singular values and the corresponding singular vectors. The orthogonality constraints are, however, essential to the optimization problem in that a solution cannot be simply obtained through solving the unconstrained regularization problem followed by an orthogonalization process. The interplay between sparse regularization and orthogonality constraints is crucial for achieving important theoretical and practical advantages, which distinguishes our SOFAR method from most previous procedures.

2.2 Applications of SOFAR

The SOFAR method provides a unified framework for a variety of statistical problems in multivariate analysis. We give four such examples, and in each example, briefly review existing techniques and suggest new methods.

2.2.1 Biclustering with sparse SVD

The biclustering problem of a data matrix, which can be traced back to Hartigan [37], aims to simultaneously cluster the rows (samples) and columns (features) of a data matrix into statistically related subgroups. A variety of biclustering techniques, which differ in the criteria used to relate clusters of samples and clusters of features and in whether overlapping of clusters is allowed, have been suggested as useful tools in the exploratory analysis of high-dimensional genomic and text data. See, for example, Busygin et al. [17] for a survey. One way of formulating the biclustering problem is through the mean model

$$\mathbf{X} = \mathbf{C}^* + \mathbf{E}, \tag{4}$$

where the mean matrix $\mathbf{C}^*$ admits a sparse SVD (2) and the sparsity patterns in the left (or right) singular vectors serve as indicators for the samples (or features) to be clustered. Lee et al. [44] proposed to estimate the first sparse SVD layer by solving the optimization problem

$$\begin{aligned}(\hat{d}, \widehat{\mathbf{u}}, \widehat{\mathbf{v}}) = \mathop{\arg\min}_{d,\mathbf{u},\mathbf{v}} \left\{ \frac{1}{2}\|\mathbf{X} - d\mathbf{u}\mathbf{v}^T\|_F^2 + \lambda_a \rho_a(d\mathbf{u}) + \lambda_b \rho_b(d\mathbf{v}) \right\} \\ \text{subject to} \quad \|\mathbf{u}\|_2 = 1, \quad \|\mathbf{v}\|_2 = 1,\end{aligned} \tag{5}$$

and obtain the next sparse SVD layer by applying the same procedure to the residual matrix $\mathbf{X} - \hat{d}\widehat{\mathbf{u}}\widehat{\mathbf{v}}^T$. Clearly, problem (5) is a specific example of the SOFAR problem (3) with $m = 1$ and $\lambda_d = 0$; however, the orthogonality constraints are not maintained during the layer-by-layer extraction process. The orthogonality issue also exists in most previous proposals, for example, Zhang et al. [68].

The multivariate linear model (1) with a sparse SVD (2) can be viewed as a supervised version of the above biclustering problem, which extends the mean model (4) to a general design matrix and can be used to identify interpretable clusters of predictors and clusters of responses that are significantly associated. Applying the SOFAR method to model (4) yields the new estimator

$$\begin{aligned}(\widehat{\mathbf{D}}, \widehat{\mathbf{U}}, \widehat{\mathbf{V}}) = \mathop{\arg\min}_{\mathbf{D},\mathbf{U},\mathbf{V}} \left\{ \frac{1}{2}\|\mathbf{X} - \mathbf{U}\mathbf{D}\mathbf{V}^T\|_F^2 + \lambda_d\|\mathbf{D}\|_1 + \lambda_a \rho_a(\mathbf{U}\mathbf{D}) + \lambda_b \rho_b(\mathbf{V}\mathbf{D}) \right\} \\ \text{subject to} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}_m, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}_m,\end{aligned} \tag{6}$$

which estimates all sparse SVD layers simultaneously while determining the rank by nuclear norm penalization and preserving the orthogonality constraints.

2.2.2 Sparse PCA

A useful technique closely related to sparse SVD is sparse principal component analysis (PCA), which enhances the convergence and improves the interpretability of PCA by introducing sparsity in the loadings of principal components. There has been a fast growing literature on sparse PCA due to its importance in dimension reduction for high-dimensional data. Various formulations coupled with efficient algorithms, notably through L_0 regularization and its L_1 and semidefinite relaxations, have been proposed by Zou et al. [72], d'Aspremont et al. [25], Shen and Huang [57], Johnstone and Lu [40], and Guo et al. [35], among others. Recently, Benidis et al. [9] developed a new method to estimate sparse eigenvectors without trading off their orthogonality based on the eigenvalue decomposition rather than the SVD using the Procrustes reformulation.

We are interested in two different ways of casting sparse PCA in our sparse SVD framework. The first approach bears a resemblance to the proposal of Zou et al. [72], which formulates sparse PCA as a regularized multivariate regression problem with the data matrix $\mathbf{X}$ treated as both the responses and the predictors. Specifically, they proposed to solve the optimization problem

$$\begin{aligned}(\widehat{\mathbf{A}}, \widehat{\mathbf{V}}) = \underset{\mathbf{A},\mathbf{V}}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{V}^T\|_F^2 + \lambda_a \rho_a(\mathbf{A}) \right\} \\ \text{subject to} \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}_m,\end{aligned} \tag{7}$$

and the loading vectors are given by the normalized columns of $\widehat{\mathbf{A}}$, $\widehat{\mathbf{a}}_j/\|\widehat{\mathbf{a}}_j\|_2$, $j = 1, \ldots, m$. However, the orthogonality of the loading vectors, a desirable property enjoyed by the standard PCA, is not enforced by problem (7). Similarly applying the SOFAR method leads to the estimator

$$\begin{aligned}(\widehat{\mathbf{D}}, \widehat{\mathbf{U}}, \widehat{\mathbf{V}}) = \underset{\mathbf{D},\mathbf{U},\mathbf{V}}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{U}\mathbf{D}\mathbf{V}^T\|_F^2 + \lambda_d\|\mathbf{D}\|_1 + \lambda_a \rho_a(\mathbf{U}\mathbf{D}) \right\} \\ \text{subject to} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}_m, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}_m,\end{aligned}$$

which explicitly imposes orthogonality among the loading vectors (the columns of $\widehat{\mathbf{U}}$). One can optionally ignore the nuclear norm penalty and determine the number of principal components by some well-established criterion.

The second approach exploits the connection of sparse PCA with regularized SVD suggested by Shen and Huang [57]. They proposed to solve the rank-1 matrix approximation problem

$$\begin{aligned}(\widehat{\mathbf{u}}, \widehat{\mathbf{b}}) = \underset{\mathbf{u},\mathbf{b}}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{X} - \mathbf{u}\mathbf{b}^T\|_F^2 + \lambda_b \rho_b(\mathbf{b}) \right\} \\ \text{subject to} \quad \|\mathbf{u}\|_2 = 1,\end{aligned} \tag{8}$$

and obtain the first loading vector $\widehat{\mathbf{b}}/\|\widehat{\mathbf{b}}\|_2$. Applying the SOFAR method similarly to the rank-m matrix

approximation problem yields the estimator

$$(\widehat{\mathbf{D}}, \widehat{\mathbf{U}}, \widehat{\mathbf{V}}) = \underset{\mathbf{D},\mathbf{U},\mathbf{V}}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{X} - \mathbf{U}\mathbf{D}\mathbf{V}^T\|_F^2 + \lambda_d\|\mathbf{D}\|_1 + \lambda_b\rho_b(\mathbf{V}\mathbf{D}) \right\}$$
$$\text{subject to} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}_m, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}_m,$$

which constitutes a multivariate generalization of problem (8), with the desirable orthogonality constraint imposed on the loading vectors (the columns of $\widehat{\mathbf{V}}$) and the optional nuclear norm penalty useful for determining the number of principal components.

2.2.3 Sparse factor analysis

Factor analysis plays an important role in dimension reduction and feature extraction for high-dimensional time series. A low-dimensional factor structure is appealing from both theoretical and practical angles, and can be conveniently incorporated into many other statistical tasks, such as forecasting with factor-augmented regression [59] and covariance matrix estimation [28]. See, for example, Bai and Ng [6] for an overview.

Let $\mathbf{x}_t \in \mathbb{R}^p$ be a vector of observed time series. Consider the factor model

$$\mathbf{x}_t = \mathbf{\Lambda}\mathbf{f}_t + \mathbf{e}_t, \quad t = 1, \ldots, T, \tag{9}$$

where $\mathbf{f}_t \in \mathbb{R}^m$ is a vector of latent factors, $\mathbf{\Lambda} \in \mathbb{R}^{p\times m}$ is the factor loading matrix, and $\mathbf{e}_t$ is the idiosyncratic error. Most existing methods for high-dimensional factor models rely on classical PCA [5, 2] or maximum likelihood to estimate the factors and factor loadings [4, 3]; as a result, the estimated factors and loadings are generally nonzero. However, in order to assign economic meanings to the factors and loadings and to further mitigate the curse of dimensionality, it would be desirable to introduce sparsity in the factors and loadings. Writing model (9) in the matrix form

$$\mathbf{X} = \mathbf{F}\mathbf{\Lambda}^T + \mathbf{E}$$

with $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)^T$, $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)^T$, and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_T)^T$ reveals its equivalence to model (4). Therefore, under the usual normalization restrictions that $\mathbf{F}^T\mathbf{F}/T = \mathbf{I}_m$ and $\mathbf{\Lambda}^T\mathbf{\Lambda}$ is diagonal, we can solve for $(\widehat{\mathbf{D}}, \widehat{\mathbf{U}}, \widehat{\mathbf{V}})$ in problem (6) and estimate the sparse factors and loadings by $\widehat{\mathbf{F}} = \sqrt{T}\widehat{\mathbf{U}}$ and $\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{V}}\widehat{\mathbf{D}}/\sqrt{T}$.

2.2.4 Sparse VAR analysis

Vector autoregressive (VAR) models have been widely used to analyze the joint dynamics of multivariate time series; see, for example, Stock and Watson [58]. Classical VAR analysis suffers greatly from the large number of free parameters in a VAR model, which grows quadratically with the dimensionality. Early attempts in reducing the impact of dimensionality have explored reduced rank methods such as

canonical analysis and reduced rank regression [12, 62]. Regularization methods such as the Lasso have recently been adapted to VAR analysis for variable selection [38, 52, 42, 8].

We present an example in which our parsimonious model setup is most appropriate. Suppose we observe the data $(\mathbf{y}_t, \mathbf{x}_t)$, where $\mathbf{y}_t \in \mathbb{R}^q$ is a low-dimensional vector of time series whose dynamics are of primary interest, and $\mathbf{x}_t \in \mathbb{R}^p$ is a high-dimensional vector of informational time series. We assume that $\mathbf{x}_t$ are generated by the VAR equation

$$\mathbf{x}_t = \mathbf{C}^{*T}\mathbf{x}_{t-1} + \mathbf{e}_t,$$

where $\mathbf{C}$ has a sparse SVD (2). This implies a low-dimensional latent model of the form

$$\mathbf{g}_t = \mathbf{D}^*\mathbf{f}_{t-1} + \widetilde{\mathbf{e}}_t,$$

where $\mathbf{f}_t = \mathbf{U}^{*T}\mathbf{x}_t$, $\mathbf{g}_t = \mathbf{V}^{*T}\mathbf{x}_t$, and $\widetilde{\mathbf{e}}_t = \mathbf{V}^{*T}\mathbf{e}_t$. Following the factor-augmented VAR (FAVAR) approach of Bernanke et al. [10], we augment the latent factors $\mathbf{f}_t$ and $\mathbf{g}_t$ to the dynamic equation of $\mathbf{y}_t$ and consider the joint model

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \mathbf{A}^T & \mathbf{B}^T \\ \mathbf{0} & \mathbf{D}^* \end{pmatrix} \begin{pmatrix} \mathbf{y}_{t-1} \\ \mathbf{f}_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \widetilde{\mathbf{e}}_t \end{pmatrix}.$$

We can estimate the parameters $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{D}^*$ by a two-step method: first apply the SOFAR method to obtain estimates of $\mathbf{D}^*$ and $\mathbf{f}_t$, and then estimate $\mathbf{A}$ and $\mathbf{B}$ by a usual VAR since both $\mathbf{y}_t$ and $\mathbf{f}_t$ are of low dimensionality. Our approach differs from previous methods in that we enforce sparse factor loadings; hence, it would allow the factors to be given economic interpretations and would be useful for uncovering the structural relationships underlying the joint dynamics of $(\mathbf{y}_t, \mathbf{x}_t)$.

3 Theoretical properties

We now investigate the theoretical properties of the SOFAR estimator (3) for model (1) under the sparse SVD structure (2). Our results concern nonasymptotic error bounds, where both response dimensionality q and predictor dimensionality p can diverge simultaneously with sample size n. The major theoretical challenges stem from the nonconvexity issues of our optimization problem which are prevalent in nonconvex statistical learning.

3.1 Technical conditions

We begin with specifying a few assumptions that facilitate our technical analysis. To simplify the technical presentation, we focus on the scenario of $p \geq q$ and our proofs can be adapted easily to the case of $p < q$ with the only difference that the rates of convergence in Theorems 1 and 2 will be modified correspondingly. Assume that each column of $\mathbf{X}$, $\widetilde{\mathbf{x}}_j$ with $j = 1, \ldots, p$, has been rescaled such that

$\|\widetilde{\mathbf{x}}_j\|_2^2 = n$. The SOFAR method minimizes the objective function in (3). Since the true rank r is unknown and we cannot expect that one can choose m to perfectly match r, the SOFAR estimates $\widehat{\mathbf{U}}$, $\widehat{\mathbf{V}}$, and $\widehat{\mathbf{D}}$ are generally of different sizes than $\mathbf{U}^*$, $\mathbf{V}^*$, and $\mathbf{D}^*$, respectively. To ease the presentation, we expand the dimensions of matrices $\mathbf{U}^*$, $\mathbf{V}^*$, and $\mathbf{D}^*$ by simply adding columns and rows of zeros to the right and to the bottom of each of the matrices to make them of sizes $p \times q$, $q \times q$, and $q \times q$, respectively. We also expand the matrices $\widehat{\mathbf{D}}$, $\widehat{\mathbf{U}}$, and $\widehat{\mathbf{V}}$ similarly to match the sizes of $\mathbf{D}^*$, $\mathbf{U}^*$, and $\mathbf{V}^*$, respectively. Define $\mathbf{A}^* = \mathbf{U}^*\mathbf{D}^*$ and $\mathbf{B}^* = \mathbf{V}^*\mathbf{D}^*$, and correspondingly $\widehat{\mathbf{A}} = \widehat{\mathbf{U}}\widehat{\mathbf{D}}$ and $\widehat{\mathbf{B}} = \widehat{\mathbf{V}}\widehat{\mathbf{D}}$ using the SOFAR estimates $(\widehat{\mathbf{U}}, \widehat{\mathbf{V}}, \widehat{\mathbf{D}})$.

Definition 1 (Robust spark)**.** *The robust spark κ_c of the $n \times p$ design matrix $\mathbf{X}$ is defined as the smallest possible positive integer such that there exists an $n \times \kappa_c$ submatrix of $n^{-1/2}\mathbf{X}$ having a singular value less than a given positive constant c.*

Condition 1. *(Parameter space) The true parameters $(\mathbf{C}^*, \mathbf{D}^*, \mathbf{A}^*, \mathbf{B}^*)$ lie in $\mathcal{C} \times \mathcal{D} \times \mathcal{A} \times \mathcal{B}$, where $\mathcal{C} = \{\mathbf{C} \in \mathbb{R}^{p\times q} : \|\mathbf{C}\|_0 < \kappa_{c_2}/2\}$, $\mathcal{D} = \{\mathbf{D} = \mathrm{diag}\{d_j\} \in \mathbb{R}^{q\times q} : d_j = 0$ or $|d_j| \geq \tau\}$, $\mathcal{A} = \{\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q} : a_{ij} = 0$ or $|a_{ij}| \geq \tau\}$, and $\mathcal{B} = \{\mathbf{B} = (b_{ij}) \in \mathbb{R}^{q\times q} : b_{ij} = 0$ or $|b_{ij}| \geq \tau\}$ with κ_{c_2} the robust spark of $\mathbf{X}$, $c_2 > 0$ some constant, and $\tau > 0$ asymptotically vanishing.*

Condition 2. *(Constrained eigenvalue) It holds that $\max_{\|\mathbf{u}\|_0<\kappa_{c_2}/2,\ \|\mathbf{u}\|_2=1} \|\mathbf{X}\mathbf{u}\|_2^2 \leq c_3 n$ and $\max_{1\leq j\leq r} \|\mathbf{X}\mathbf{u}_j^*\|_2^2 \leq c_3 n$ for some constant $c_3 > 0$, where $\mathbf{u}_j^*$ is the left singular vector of $\mathbf{C}^*$ corresponding to singular value d_j^*.*

Condition 3. *(Error term) The error term $\mathbf{E} \in \mathbb{R}^{n\times q} \sim N(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Sigma})$ with the maximum eigenvalue $\alpha_{\max}$ of $\mathbf{\Sigma}$ bounded from above and diagonal entries of $\mathbf{\Sigma}$ being σ_j^2's.*

Condition 4. *(Penalty functions) For matrices $\mathbf{M}$ and $\mathbf{M}^*$ of the same size, the penalty functions ρ_h with $h \in \{a, b\}$ satisfy $|\rho_h(\mathbf{M}) - \rho_h(\mathbf{M}^*)| \leq \|\mathbf{M} - \mathbf{M}^*\|_1$.*

Condition 5. *(Relative spectral gap) The nonzero singular values of $\mathbf{C}^*$ satisfy that $d_{j-1}^{*2} - d_j^{*2} \geq \delta^{1/2} d_{j-1}^{*2}$ for $2 \leq j \leq r$ with $\delta > 0$ some constant, and r and $\sum_{j=1}^r (d_1^*/d_j^*)^2$ can diverge as $n \to \infty$.*

The concept of robust spark in Definition 1 was introduced initially in [69] and [30], where the thresholded parameter space was exploited to characterize the global optimum for regularization methods with general penalties. Similarly, the thresholded parameter space and the constrained eigenvalue condition which builds on the robust spark condition of the design matrix in Conditions 1 and 2 are essential for investigating the computable solution to the nonconvex SOFAR optimization problem in (3). By Proposition 1 of [30], the robust spark κ_{c_2} can be at least of order $O\{n/(\log p)\}$ with asymptotic probability one when the rows of $\mathbf{X}$ are independently sampled from multivariate Gaussian distributions with dependency. Although Condition 3 assumes Gaussianity, our theory can in principle carry over to the case of sub-Gaussian errors, provided that the concentration inequalities for Gaussian random variables used in our proofs are replaced by those for sub-Gaussian random variables.

Condition 4 includes many kinds of penalty functions that bring about sparse estimates. Important examples include the entrywise L_1-norm and rowwise $(2,1)$-norm, where the former encourages sparsity among the predictor/response effects specific to each rank-1 SVD layer, while the latter promotes predictor/response-wise sparsity regardless of the specific layer. To see why the rowwise $(2,1)$-norm satisfies Condition 4, observe that

$$\|\mathbf{M}\|_1 \equiv \sum_i \sum_j |m_{ij}| = \sum_i \left(\sum_{j,k} |m_{ij}||m_{ik}| \right)^{1/2} \geq \sum_i \left(\sum_j m_{ij}^2 \right)^{1/2} \equiv \|\mathbf{M}\|_{2,1},$$

which along with the triangle inequality entails that Condition 4 is indeed satisfied. Moreover, Condition 4 allows us to use concave penalties such as SCAD [29] and MCP [67]; see, for instance, the proof of Lemma 1 in [30].

Intuitively, Condition 5 rules out the nonidentifiable case where some nonzero singular values are tied with each other and the associated singular vectors in matrices $\mathbf{U}^*$ and $\mathbf{V}^*$ are identifiable only up to some orthogonal transformation. In particular, Condition 5 enables us to establish the key Lemma 3 in Section B.1 of Supplementary Material, where the matrix perturbation theory can be invoked.

3.2 Main results

Since the objective function of the SOFAR method (3) is nonconvex, solving this optimization problem is highly challenging. To overcome the difficulties, as mentioned in the Introduction we exploit the framework of CANO and suggest a two-step approach, where in the first step we solve the following L_1-penalized squared loss minimization problem

$$\widetilde{\mathbf{C}} = \underset{\mathbf{C} \in \mathbb{R}^{p \times q}}{\arg\min} \left\{ (2n)^{-1} \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_F^2 + \lambda_0 \|\mathbf{C}\|_1 \right\} \tag{10}$$

to construct an initial estimator $\widetilde{\mathbf{C}}$ with $\lambda_0 \geq 0$ some regularization parameter. If $\widetilde{\mathbf{C}} = \mathbf{0}$, then we set the final SOFAR estimator as $\widehat{\mathbf{C}} = \mathbf{0}$; otherwise, in the second step we do a refined search and minimize the SOFAR objective function (3) in an asymptotically shrinking neighborhood of $\widetilde{\mathbf{C}}$ to obtain the final SOFAR estimator $\widehat{\mathbf{C}}$. In the case of $\widetilde{\mathbf{C}} = \mathbf{0}$, our two-step procedure reduces to a one-step procedure. Since Theorem 1 below establishes that $\widetilde{\mathbf{C}}$ can be close to $\mathbf{C}^*$ with asymptotic probability one, having $\widetilde{\mathbf{C}} = \mathbf{0}$ is a good indicator that the true $\mathbf{C}^* = \mathbf{0}$.

Thanks to its convexity, the objective function in (10) in the first step can be solved easily and efficiently. In fact, since the objective function in (10) is separable it follows that the jth column of $\widetilde{\mathbf{C}}$ can be obtained by solving the univariate response Lasso regression

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ (2n)^{-1} \|\mathbf{Y}\mathbf{e}_j - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_0 \|\boldsymbol{\beta}\|_1 \right\},$$

where $\mathbf{e}_j$ is a q-dimensional vector with jth component 1 and all other components 0. The above uni-

variate response Lasso regression has been studied extensively and well understood, and many efficient algorithms have been proposed for solving it. Denote by $(\widetilde{\mathbf{D}}, \widetilde{\mathbf{U}}, \widetilde{\mathbf{V}})$ the initial estimator of $(\mathbf{D}^*, \mathbf{U}^*, \mathbf{V}^*)$ obtained from the SVD of $\widetilde{\mathbf{C}}$, and let $\widetilde{\mathbf{A}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}$ and $\widetilde{\mathbf{B}} = \widetilde{\mathbf{V}}\widetilde{\mathbf{D}}$. Since the bounds for the SVD are key to the analysis of SOFAR estimator in the second step, for completeness we present the nonasymptotic bounds on estimation errors of the initial estimator in the following theorem.

Theorem 1 (Error bounds for initial estimator)**.** *Assume that Conditions 1–3 hold and let* $\lambda_0 = c_0\sigma_{\max}\big(n^{-1}\log(pq)\big)^{1/2}$ *with* $\sigma_{\max} = \max_{1\le j\le q}\sigma_j$ *and* $c_0 > \sqrt{2}$ *some constant. Then with probability at least* $1 - 2(pq)^{1-c_0^2/2}$*, the estimation error is bounded as*

$$\|\widetilde{\mathbf{C}} - \mathbf{C}^*\|_F \le R_n \equiv c(n^{-1}s\log(pq))^{1/2} \tag{11}$$

with $s = \|\mathbf{C}^*\|_0$ *and* $c > 0$ *some constant. Under additional Condition 5, with the same probability bound the following estimation error bounds hold simultaneously*

$$\|\widetilde{\mathbf{D}} - \mathbf{D}^*\|_F \le c(n^{-1}s\log(pq))^{1/2}, \tag{12}$$

$$\|\widetilde{\mathbf{A}} - \mathbf{A}^*\|_F + \|\widetilde{\mathbf{B}} - \mathbf{B}^*\|_F \le c\eta_n(n^{-1}s\log(pq))^{1/2}, \tag{13}$$

where $\eta_n = 1 + \delta^{-1/2}\big(\sum_{j=1}^r (d_1^*/d_j^*)^2\big)^{1/2}$.

For the case of $q = 1$, the estimation error bound (11) is consistent with the well-known oracle inequality for Lasso [11]. The additional estimation error bounds (12) and (13) for the SVD in Theorem 1 are, however, new to the literature. It is worth mentioning that Condition 5 and the latest results in [65] play a crucial role in establishing these additional error bounds.

After obtaining the initial estimator $\widetilde{\mathbf{C}}$ from the first step, we can solve the SOFAR optimization problem in an asymptotically shrinking neighborhood of $\widetilde{\mathbf{C}}$. More specifically, we define $\widetilde{\mathcal{P}}_n = \{\mathbf{C} : \|\mathbf{C} - \widetilde{\mathbf{C}}\|_F \le 2R_n\}$ with R_n the upper bound in (11). Then it is seen from Theorem 1 that the true coefficient matrix $\mathbf{C}^*$ is contained in $\widetilde{\mathcal{P}}_n$ with probability at least $1 - 2(pq)^{1-c_0^2/2}$. Further define

$$\mathcal{P}_n = \widetilde{\mathcal{P}}_n \cap (\mathcal{C} \times \mathcal{D} \times \mathcal{A} \times \mathcal{B}), \tag{14}$$

where sets $\mathcal{C}$, $\mathcal{D}$, $\mathcal{A}$, and $\mathcal{B}$ are defined in Condition 1. Then with probability at least $1 - 2(pq)^{1-c_0^2/2}$, the set $\mathcal{P}_n$ defined in (14) is nonempty with at least one element $\mathbf{C}^*$ by Condition 1. We minimize the SOFAR objective function (3) by searching in the shrinking neighborhood $\mathcal{P}_n$ and denote by $\widehat{\mathbf{C}}$ the resulting SOFAR estimator. Then it follows that with probability at least $1 - 2(pq)^{1-c_0^2/2}$,

$$\|\widehat{\mathbf{C}} - \mathbf{C}^*\|_F \le \|\widehat{\mathbf{C}} - \widetilde{\mathbf{C}}\|_F + \|\widetilde{\mathbf{C}} - \mathbf{C}^*\|_F \le 3R_n,$$

where the first inequality is by the triangle inequality and the second one is by the construction of set $\mathcal{P}_n$ and Theorem 1. Therefore, we see that the SOFAR estimator given by our two-step procedure is guaranteed to have convergence rate at least $O(R_n)$.

Since the initial estimator investigated in Theorem 1 completely ignores the finer sparse SVD structure of the coefficient matrix $\mathbf{C}^*$, intuitively the second step of SOFAR estimation can lead to improved error bounds. Indeed we show in Theorem 2 below that with the second step of refinement, up to some columnwise sign changes the SOFAR estimator can admit estimator error bounds in terms of parameters r, s_a, and s_b with $r = \|\mathbf{D}^*\|_0$, $s_a = \|\mathbf{A}^*\|_0$, and $s_b = \|\mathbf{B}^*\|_0$. When r, s_a, and s_b are drastically smaller than s, these new upper bounds can have better rates of convergence.

Theorem 2 (Error bounds for SOFAR estimator)**.** *Assume that Conditions 1–5 hold,* $\lambda_{\max} \equiv \max(\lambda_d, \lambda_a, \lambda_b) = c_1 \left(n^{-1}\log(pr)\right)^{1/2}$ *with* $c_1 > 0$ *some large constant,* $\log p = O(n^\alpha)$, $q = O(n^{\beta/2})$, $s = O(n^\gamma)$, *and* $\eta_n^2 = o(\min\{\lambda_{\max}^{-1}\tau, n^{1-\alpha-\beta-\gamma}\tau^2\})$ *with* $\alpha, \beta, \gamma \geq 0$, $\alpha + \beta + \gamma < 1$, *and* η_n *as given in Theorem 1. Then with probability at least*

$$1 - \left\{2(pq)^{1-c_0^2/2} + 2(pr)^{-\tilde{c}_2} + 2pr\exp\left(-\tilde{c}_3 n^{1-\beta-\gamma}\tau^2\eta_n^{-2}\right)\right\}, \tag{15}$$

the SOFAR estimator satisfies the following error bounds simultaneously:

$$(a) \quad \|\widehat{\mathbf{C}} - \mathbf{C}^*\|_F \leq c\min\{s, (r+s_a+s_b)\eta_n^2\}^{1/2}\{n^{-1}\log(pq)\}^{1/2}, \tag{16}$$

$$(b) \quad \|\widehat{\mathbf{D}} - \mathbf{D}^*\|_F + \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_F + \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F \leq c\min\{s, (r+s_a+s_b)\eta_n^2\}^{1/2}\eta_n\{n^{-1}\log(pq)\}^{1/2}, \tag{17}$$

$$(c) \quad \|\widehat{\mathbf{D}} - \mathbf{D}^*\|_0 + \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_0 + \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_0 \leq (r+s_a+s_b)[1+o(1)], \tag{18}$$

$$(d) \quad \|\widehat{\mathbf{D}} - \mathbf{D}^*\|_1 + \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_1 + \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_1 \leq c(r+s_a+s_b)\eta_n^2\lambda_{\max}, \tag{19}$$

$$(e) \quad n^{-1}\|\mathbf{X}(\widehat{\mathbf{C}} - \mathbf{C}^*)\|_F^2 \leq c(r+s_a+s_b)\eta_n^2\lambda_{\max}^2, \tag{20}$$

where $c_0 > \sqrt{2}$ *and* $c, \tilde{c}_2, \tilde{c}_3$ *are some positive constants.*

We see from Theorem 2 that the upper bounds in (16) and (17) are the minimum of two rates, one involving $r + s_a + s_b$ (the total sparsity of $\mathbf{D}^*$, $\mathbf{A}^*$, and $\mathbf{B}^*$) and the other one involving s (the sparsity of matrix $\mathbf{C}^*$). The rate involving s is from the first step of Lasso estimation, while the rate involving $r + s_a + s_b$ is from the second step of SOFAR refinement. For the case of $s > (r + s_a + s_b)\eta_n^2$, our two-step procedure leads to enhanced error rates under the Frobenius norm. Moreover, the error rates in (18)–(20) are new to the literature and not shared by the initial Lasso estimator, showing again the advantages of having the second step of refinement. It is seen that our two-step SOFAR estimator is capable of recovering the sparsity structure of $\mathbf{D}^*$, $\mathbf{A}^*$, and $\mathbf{B}^*$ very well.

Let us gain more insights into these new error bounds. In the case of univariate response with $q = 1$, we have $\eta_n = 1 + \delta$, $r = 1$, $s_a = s$, and $s_b = 1$. Then the upper bounds in (16)–(20) reduce to $c\{sn^{-1}\log p\}^{1/2}$, $c\{sn^{-1}\log p\}^{1/2}$, cs, $cs\{n^{-1}\log p\}^{1/2}$, and $cn^{-1}s\log p$, respectively, which are indeed within a logarithmic factor of the oracle rates for the case of high-dimensional univariate response regression. Furthermore, in the rank-one case of $r = 1$ we have $\eta_n = 1 + \delta^{-1/2}$ and $s = s_a s_b$. Correspondingly, the upper bounds in (11)–(13) for the initial Lasso estimator all be-

come $c\{n^{-1}s_a s_b \log(pq)\}^{1/2}$, while the upper bounds in (16)–(20) for the SOFAR estimator become $c\{(s_a+s_b)n^{-1}\log(pq)\}^{1/2}$, $c\{(s_a+s_b)n^{-1}\log(pq)\}^{1/2}$, $c(s_a+s_b)$, $c(s_a+s_b)\{n^{-1}\log(pq)\}^{1/2}$, and $cn^{-1}(s_a+s_b)\log(pq)$, respectively. In particular, we see that the SOFAR estimator can have much improved rates of convergence even in the setting of $r=1$.

4 Implementation of SOFAR

The interplay between sparse regularization and orthogonality constraints creates substantial algorithmic challenges for solving the SOFAR optimization problem (3), for which many existing algorithms can become either inefficient or inapplicable. For example, coordinate descent methods that are popular for solving large-scale sparse regularization problems [32] are not directly applicable because the penalty terms in problem (3) are not separable under the orthogonality constraints. Also, the general framework for algorithms involving orthogonality constraints [26] does not take sparsity into account and hence does not lead to efficient algorithms in our context. Recently, Benidis et al. [9] focused on the unsupervised learning setting and introduced a new algorithm for estimating sparse eigenvectors without trading off their orthogonality based on the eigenvalue decomposition rather than the SVD. To obtain sparse orthogonal eigenvectors, they applied the minorization-maximization framework on the sparse PCA problem, which results in solving a sequence of rectangular Procrustes problems. Inspired by a recently revived interest in the augmented Lagrangian method (ALM) and its variants for large-scale optimization in statistics and machine learning [13], in this section we develop an efficient algorithm for solving problem (3).

4.1 SOFAR algorithm with ALM-BCD

The architecture of the proposed SOFAR algorithm is based on the ALM coupled with block coordinate descent (BCD). The first construction step is to utilize variable splitting to separate the orthogonality constraints and sparsity-inducing penalties into different subproblems, which then enables efficient optimization in a block coordinate descent fashion. To this end, we introduce two new variables $\mathbf{A}$ and $\mathbf{B}$, and express problem (3) in the equivalent form

$$\begin{aligned}(\widehat{\boldsymbol{\Theta}},\widehat{\boldsymbol{\Omega}}) = \underset{\boldsymbol{\Theta},\boldsymbol{\Omega}}{\arg\min}\left\{\frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{U}\mathbf{D}\mathbf{V}^T\|_F^2+\lambda_d\|\mathbf{D}\|_1+\lambda_a\rho_a(\mathbf{A})+\lambda_b\rho_b(\mathbf{B})\right\}\\ \text{subject to}\quad \mathbf{U}^T\mathbf{U}=\mathbf{I}_m,\quad \mathbf{V}^T\mathbf{V}=\mathbf{I}_m,\quad \mathbf{U}\mathbf{D}=\mathbf{A},\quad \mathbf{V}\mathbf{D}=\mathbf{B},\end{aligned} \tag{21}$$

where $\boldsymbol{\Theta}=(\mathbf{D},\mathbf{U},\mathbf{V})$ and $\boldsymbol{\Omega}=(\mathbf{A},\mathbf{B})$. We form the augmented Lagrangian for problem (21) as

$$\begin{aligned}L_\mu(\boldsymbol{\Theta},\boldsymbol{\Omega},\boldsymbol{\Gamma}) = \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{U}\mathbf{D}\mathbf{V}^T\|_F^2+\lambda_d\|\mathbf{D}\|_1+\lambda_a\rho_a(\mathbf{A})+\lambda_b\rho_b(\mathbf{B})+\langle\boldsymbol{\Gamma}_a,\mathbf{U}\mathbf{D}-\mathbf{A}\rangle\\ +\langle\boldsymbol{\Gamma}_b,\mathbf{V}\mathbf{D}-\mathbf{B}\rangle+\frac{\mu}{2}\|\mathbf{U}\mathbf{D}-\mathbf{A}\|_F^2+\frac{\mu}{2}\|\mathbf{V}\mathbf{D}-\mathbf{B}\|_F^2,\end{aligned}$$

Table 1: SOFAR algorithm with ALM-BCD

Parameters: λ_d, λ_a, λ_b, and $\gamma > 1$
Initialize $\mathbf{U}^0$, $\mathbf{V}^0$, $\mathbf{D}^0$, $\mathbf{A}^0$, $\mathbf{B}^0$, $\mathbf{\Gamma}_a^0$, $\mathbf{\Gamma}_b^0$, and μ^0
For $k = 0, 1, \ldots$ do
  update $\mathbf{U}$, $\mathbf{V}$, $\mathbf{D}$, $\mathbf{A}$, and $\mathbf{B}$:
    (a) $\mathbf{U}^{k+1} \leftarrow \underset{\mathbf{U}^T\mathbf{U}=\mathbf{I}_m}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{U}\mathbf{D}^k(\mathbf{V}^k)^T\|_F^2 + \frac{\mu^k}{2}\|\mathbf{U}\mathbf{D}^k - \mathbf{A}^k + \mathbf{\Gamma}_a^k/\mu^k\|_F^2 \right\}$
    (b) $\mathbf{V}^{k+1} \leftarrow \underset{\mathbf{V}^T\mathbf{V}=\mathbf{I}_m}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{U}^{k+1}\mathbf{D}^k\mathbf{V}^T\|_F^2 + \frac{\mu^k}{2}\|\mathbf{V}\mathbf{D}^k - \mathbf{B}^k + \mathbf{\Gamma}_b^k/\mu^k\|_F^2 \right\}$
    (c) $\mathbf{D}^{k+1} \leftarrow \underset{\mathbf{D}\geq\mathbf{0}}{\arg\min} \Big\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{U}^{k+1}\mathbf{D}(\mathbf{V}^{k+1})^T\|_F^2 + \frac{\mu^k}{2}\|\mathbf{U}^{k+1}\mathbf{D} - \mathbf{A}^k + \mathbf{\Gamma}_a^k/\mu^k\|_F^2 + \frac{\mu^k}{2}\|\mathbf{V}^{k+1}\mathbf{D} - \mathbf{B}^k + \mathbf{\Gamma}_b^k/\mu^k\|_F^2 + \lambda_d\|\mathbf{D}\|_1 \Big\}$
    (d) $\mathbf{A}^{k+1} \leftarrow \underset{\mathbf{A}}{\arg\min} \left\{ \frac{\mu^k}{2}\|\mathbf{U}^{k+1}\mathbf{D}^{k+1} - \mathbf{A} + \mathbf{\Gamma}_a^k/\mu^k\|_F^2 + \lambda_a\rho_a(\mathbf{A}) \right\}$
    (e) $\mathbf{B}^{k+1} \leftarrow \underset{\mathbf{B}}{\arg\min} \left\{ \frac{\mu^k}{2}\|\mathbf{V}^{k+1}\mathbf{D}^{k+1} - \mathbf{B} + \mathbf{\Gamma}_b^k/\mu^k\|_F^2 + \lambda_b\rho_b(\mathbf{B}) \right\}$
    (f) optionally, repeat (a)–(e) until convergence
  update $\mathbf{\Gamma}_a$ and $\mathbf{\Gamma}_b$:
    (a) $\mathbf{\Gamma}_a^{k+1} \leftarrow \mathbf{\Gamma}_a^k + \mu^k(\mathbf{U}^{k+1}\mathbf{D}^{k+1} - \mathbf{A}^{k+1})$
    (b) $\mathbf{\Gamma}_b^{k+1} \leftarrow \mathbf{\Gamma}_b^k + \mu^k(\mathbf{V}^{k+1}\mathbf{D}^{k+1} - \mathbf{B}^{k+1})$
  update μ by $\mu^{k+1} \leftarrow \gamma\mu^k$
end

where $\mathbf{\Gamma} = (\mathbf{\Gamma}_a, \mathbf{\Gamma}_b)$ is the set of Lagrangian multipliers and $\mu > 0$ is a penalty parameter. Based on ALM, the proposed algorithm consists of the following iterations:

1. $(\mathbf{\Theta}, \mathbf{\Omega})$-step: $(\mathbf{\Theta}^{k+1}, \mathbf{\Omega}^{k+1}) \leftarrow \arg\min_{\mathbf{\Theta}:\, \mathbf{U}^T\mathbf{U}=\mathbf{V}^T\mathbf{V}=\mathbf{I}_m, \mathbf{\Omega}} L_\mu(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{\Gamma}^k)$;
2. $\mathbf{\Gamma}$-step: $\mathbf{\Gamma}_a^{k+1} \leftarrow \mathbf{\Gamma}_a^k + \mu(\mathbf{U}^{k+1}\mathbf{D}^{k+1} - \mathbf{A}^{k+1})$ and $\mathbf{\Gamma}_b^{k+1} \leftarrow \mathbf{\Gamma}_b^k + \mu(\mathbf{V}^{k+1}\mathbf{D}^{k+1} - \mathbf{B}^{k+1})$.

The $(\mathbf{\Theta}, \mathbf{\Omega})$-step can be solved by a block coordinate descent method [61] cycling through the blocks $\mathbf{U}$, $\mathbf{V}$, $\mathbf{D}$, $\mathbf{A}$, and $\mathbf{B}$. Note that the orthogonality constraints and the sparsity-inducing penalties are now separated into subproblems with respect to $\mathbf{\Theta}$ and $\mathbf{\Omega}$, respectively. To achieve convergence of the SOFAR algorithm in practice, an inexact minimization with a few block coordinate descent iterations is often sufficient. Moreover, to enhance the convergence of the algorithm to a feasible solution we optionally increase the penalty parameter μ by a ratio $\gamma > 1$ at the end of each iteration. This leads to the SOFAR algorithm with ALM-BCD described in Table 1.

We still need to solve the subproblems in algorithm 1. The $\mathbf{U}$-update is similar to the weighted orthogonal Procrustes problem considered by Koschat and Swayne [43]. By expanding the squares and omitting terms not involving $\mathbf{U}$, this subproblem is equivalent to minimizing

$$\frac{1}{2}\|\mathbf{X}\mathbf{U}\mathbf{D}^k\|_F^2 - \mathrm{tr}(\mathbf{U}^T\mathbf{X}^T\mathbf{Y}\mathbf{V}^k\mathbf{D}^k) - \mathrm{tr}(\mathbf{U}^T(\mu^k\mathbf{A}^k - \mathbf{\Gamma}_a^k)\mathbf{D}^k)$$

subject to $\mathbf{U}^T\mathbf{U} = \mathbf{I}_m$. Taking a matrix $\mathbf{Z}$ such that $\mathbf{Z}^T\mathbf{Z} = \rho^2\mathbf{I}_p - \mathbf{X}^T\mathbf{X}$, where ρ^2 is the largest

eigenvalue of $\mathbf{X}^T\mathbf{X}$, we can follow the argument of Koschat and Swayne [43] to obtain the iterative algorithm: for $j = 0, 1, \ldots$, form the $p \times m$ matrix $\mathbf{C}_1 = (\mathbf{X}^T\mathbf{Y}\mathbf{V}^k + \mu^k\mathbf{A}^k - \mathbf{\Gamma}_a^k + \mathbf{Z}^T\mathbf{Z}\mathbf{U}^j\mathbf{D}^k)\mathbf{D}^k$, compute the SVD $\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T = \mathbf{C}_1$, and update $\mathbf{U}^{j+1} = \mathbf{U}_1\mathbf{V}_1^T$. Note that $\mathbf{C}_1$ depends on $\mathbf{Z}^T\mathbf{Z}$ only, and hence the explicit computation of $\mathbf{Z}$ is not needed. The $\mathbf{V}$-update is similar to a standard orthogonal Procrustes problem and amounts to maximizing

$$\mathrm{tr}(\mathbf{V}^T\mathbf{Y}^T\mathbf{X}\mathbf{U}^{k+1}\mathbf{D}^k) + \mathrm{tr}(\mathbf{V}^T(\mu^k\mathbf{B}^k - \mathbf{\Gamma}_b^k)\mathbf{D}^k)$$

subject to $\mathbf{V}^T\mathbf{V} = \mathbf{I}_m$. A direct method for this problem [34, pp. 327–328] gives the algorithm: form the $q \times m$ matrix $\mathbf{C}_2 = (\mathbf{Y}^T\mathbf{X}\mathbf{U}^{k+1} + \mu^k\mathbf{B}^k - \mathbf{\Gamma}_b^k)\mathbf{D}^k$, compute the SVD $\mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^T = \mathbf{C}_2$, and set $\mathbf{V} = \mathbf{U}_2\mathbf{V}_2^T$. Since m is usually small, the SVD computations in the $\mathbf{U}$- and $\mathbf{V}$-updates are cheap. The Lasso problem in the $\mathbf{D}$-update reduces to a standard quadratic program with the nonnegativity constraint, which can be readily solved by efficient algorithms; see, for example, Sha et al. [56]. Note that the $\mathbf{D}$-update may set some singular values to exactly zero; hence, a greedy strategy can be taken to further bring down the computational complexity, by removing the zero singular values and reducing the sizes of the relevant matrices accordingly in subsequent computations. The updates of $\mathbf{A}$ and $\mathbf{B}$ are free of orthogonality constraints and therefore easy to solve. With the popular choices of $\|\cdot\|_1$ and $\|\cdot\|_{2,1}$ as the penalty functions, the updates can be performed by entrywise and rowwise soft-thresholding, respectively.

Following the theoretical analysis for the SOFAR method in Section 3, we employ the SVD of the cross-validated L_1-penalized estimator $\widetilde{\mathbf{C}}$ in (10) to initialize $\mathbf{U}$, $\mathbf{V}$, $\mathbf{D}$, $\mathbf{A}$, and $\mathbf{B}$; the $\mathbf{\Gamma}_a$ and $\mathbf{\Gamma}_b$ are initialized as zero matrices. In practice, for large-scale problems we can further scale up the SOFAR method by performing feature screening with the initial estimator $\widetilde{\mathbf{C}}$, that is, the response variables corresponding to zero columns in $\widetilde{\mathbf{C}}$ and the predictors corresponding to zero rows in $\widetilde{\mathbf{C}}$ could be removed prior to the finer SOFAR analysis.

4.2 Convergence analysis and tuning parameter selection

For general nonconvex problems, an ALM algorithm needs not to converge, and even if it converges, it needs not to converge to an optimal solution. We have the following convergence results regarding the proposed SOFAR algorithm with ALM-BCD.

Theorem 3 (Convergence of SOFAR algorithm)**.** *Assume that* $\sum_{k=1}^{\infty}\{[\Delta L_\mu(\mathbf{U}^k)]^{1/2} + [\Delta L_\mu(\mathbf{V}^k)]^{1/2} + [\Delta L_\mu(\mathbf{D}^k)]^{1/2}\} < \infty$ *and the penalty functions* $\rho_a(\cdot)$ *and* $\rho_b(\cdot)$ *are convex, where* $\Delta L_\mu(\cdot)$ *denotes the decrease in* $L_\mu(\cdot)$ *by a block update. Then the sequence generated by the SOFAR algorithm converges to a local solution of the augmented Lagrangian for problem* (21)*.*

Note that without the above assumption on $(\mathbf{U}^k)$, $(\mathbf{V}^k)$, and $(\mathbf{D}^k)$, we can only show that the differences between two consecutive $\mathbf{U}$-, $\mathbf{V}$-, and $\mathbf{D}$-updates converge to zero by the convergence of the sequence $(L_\mu(\cdot))$, but the sequences $(\mathbf{U}^k)$, $(\mathbf{V}^k)$, and $(\mathbf{D}^k)$ may not necessarily converge. Although

Theorem 3 does not ensure the convergence of algorithm 1 to an optimal solution, numerical evidence suggests that the algorithm has strong convergence properties and the produced solutions perform well in numerical studies.

The above SOFAR algorithm is presented for a fixed triple of tuning parameters $(\lambda_d, \lambda_a, \lambda_b)$. One may apply a fine grid search with K-fold cross-validation or an information criterion such as BIC and its high-dimensional extensions including GIC [31] to choose an optimal triple of tuning parameters and hence a best model. In either case, a full search over a three-dimensional grid would be prohibitively expensive, especially for large-scale problems. Theorem 2, however, suggests that the parameter tuning can be effectively reduced to one or two dimensions. Hence, we adopt a search strategy which is computationally affordable and still provides reasonable and robust performance. To this end, we first estimate an upper bound on each of the tuning parameters by considering the *marginal null model*, where two of the three tuning parameters are fixed at zero and the other is set to the minimum value leading to a null model. We denote the upper bounds thus obtained by $(\lambda_d^*, \lambda_a^*, \lambda_b^*)$, and conduct a search over a one-dimensional grid of values between $(\lambda_d^*, \lambda_a^*, \lambda_b^*)$ and $(\varepsilon\lambda_d^*, \varepsilon\lambda_a^*, \varepsilon\lambda_b^*)$, with $\varepsilon > 0$ sufficiently small (e.g., 10^{-3}) to ensure the coverage of a full spectrum of reasonable solutions. Our numerical experience suggests that this simple search strategy works well in practice while reducing the computational cost dramatically. More flexibility can be gained by adjusting the ratios between λ_d, λ_a, and λ_b if additional information about the relative sparsity levels of $\mathbf{D}$, $\mathbf{A}$, and $\mathbf{B}$ is available.

5 Numerical studies

5.1 Simulation examples

Our Condition 4 in Section 3.1 accommodates a large group of penalty functions including concave ones such as SCAD and MCP. As demonstrated in [73] and [27], nonconvex regularization problems can be solved using the idea of local linear approximation, which essentially reduces the original problem to the weighted L_1-regularization with the weights chosen adaptively based on some initial solution. For this reason, in the simulation study we focus on the entrywise L_1-norm $\|\cdot\|_1$ and the rowwise $(2,1)$-norm $\|\cdot\|_{2,1}$, as well as their adaptive extensions. The use of adaptively weighted penalties has also been explored in the contexts of reduced rank regression [21] and sparse PCA [45]. We next provide more details on the adaptive penalties used in our simulation study. To simplify the presentation, we use the entrywise L_1-norm as an example.

Incorporating adaptive weighting into the penalty terms in problem (21) leads to the adaptive SOFAR estimator

$$(\widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}}) = \underset{\mathbf{\Theta},\mathbf{\Omega}}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{U}\mathbf{D}\mathbf{V}^T\|_F^2 + \lambda_d\|\mathbf{W}_d \circ \mathbf{D}\|_1 + \lambda_a\|\mathbf{W}_a \circ \mathbf{A}\|_1 + \lambda_b\|\mathbf{W}_b \circ \mathbf{B}\|_1 \right\}$$
$$\text{subject to} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}_m, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}_m, \quad \mathbf{U}\mathbf{D} = \mathbf{A}, \quad \mathbf{V}\mathbf{D} = \mathbf{B},$$

where $\mathbf{W}_d \in \mathbb{R}^{m\times m}$, $\mathbf{W}_a \in \mathbb{R}^{p\times m}$, and $\mathbf{W}_b \in \mathbb{R}^{q\times m}$ are weighting matrices that depend on the initial estimates $\widetilde{\mathbf{D}}$, $\widetilde{\mathbf{A}}$, and $\widetilde{\mathbf{B}}$, respectively, and $\circ$ is the Hadamard or entrywise product. The weighting matrices are chosen to reflect the intuition that singular values and singular vectors of larger magnitude should be less penalized in order to reduce bias and improve efficiency in estimation. As suggested in [73], if one is interested in using some nonconvex penalty functions $\rho_a(\cdot)$ and $\rho_b(\cdot)$ then the weight matrices can be constructed by using the first order derivatives of the penalty functions and the initial solution $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{D}})$. In our implementation, for simplification we adopt the alternative popular choice of $\mathbf{W}_d = \mathrm{diag}(\widetilde{d}_1^{-1}, \ldots, \widetilde{d}_m^{-1})$ with $\widetilde{d}_j$ the jth diagonal entry of $\widetilde{\mathbf{D}}$, as suggested in Zou [71]. Similarly, we set $\mathbf{W}_a = (\widetilde{a}_{ij}^{-1})$ and $\mathbf{W}_b = (\widetilde{b}_{ij}^{-1})$ with $\widetilde{a}_{ij}$ and $\widetilde{b}_{ij}$ the (i,j)th entries of $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$, respectively. Extension of the SOFAR algorithm with ALM-BCD in Section 4.1 is also straightforward, with the $\mathbf{D}$-update becoming an adaptive Lasso problem and the updates of $\mathbf{A}$ and $\mathbf{B}$ now performed by adaptive soft-thresholding. A further way of improving the estimation efficiency is to exploit regularization methods in the thresholded parameter space [30] or thresholded regression [69], which we do not pursue in this paper.

We compare the SOFAR estimator with the entrywise L_1-norm (Lasso) penalty (SOFAR-L) or the rowwise $(2,1)$-norm (group Lasso) penalty (SOFAR-GL) with five alternative methods, including three classical methods, namely, the ordinary least squares (OLS), separate adaptive Lasso regressions (Lasso), and reduced rank regression (RRR), and two recent sparse and low rank methods, namely, reduced rank regression with sparse SVD (RSSVD) proposed by Chen et al. [19] and sparse reduced rank regression (SRRR) considered by Chen and Huang [22] (see also the rank constrained group Lasso estimator in Bunea et al. 16). Both Chen et al. [19] and Chen and Huang [22] used adaptive weighted penalization. We thus consider both nonadaptive and adaptive versions of the SOFAR-L, SOFAR-GL, RSSVD, and SRRR methods.

5.1.1 Simulation setups

We consider several simulation settings with various model dimensions and sparse SVD patterns in the coefficient matrix $\mathbf{C}^*$. In all settings, we took the sample size $n = 200$ and the true rank $r = 3$. Models 1 and 2 concern the entrywise sparse SVD structure in $\mathbf{C}^*$. The design matrix $\mathbf{X}$ was generated with i.i.d. rows from $N_p(\mathbf{0}, \mathbf{\Sigma}_x)$, where $\mathbf{\Sigma}_x = (0.5^{|i-j|})$. In model 1, we set $p = 100$ and $q = 40$, and let $\mathbf{C}^* = \sum_{j=1}^3 d_j^* \mathbf{u}_j^* \mathbf{v}_j^{*T}$ with $d_1^* = 20$, $d_2^* = 15$, $d_3^* = 10$, and

$$\begin{gathered}
\tilde{\mathbf{u}}_1 = (\mathrm{unif}(S_u, 5), \mathrm{rep}(0, 20))^T, \quad \tilde{\mathbf{u}}_2 = (\mathrm{rep}(0, 3), -\tilde{u}_{1,4}, \tilde{u}_{1,5}, \mathrm{unif}(S_u, 3), \mathrm{rep}(0, 17))^T, \\
\tilde{\mathbf{u}}_3 = (\mathrm{rep}(0, 8), \mathrm{unif}(S_u, 2), \mathrm{rep}(0, 15))^T, \quad \mathbf{u}_j^* = \tilde{\mathbf{u}}_j / \|\tilde{\mathbf{u}}_j\|_2, \quad j = 1, 2, 3, \\
\tilde{\mathbf{v}}_1 = (\mathrm{unif}(S_v, 5), \mathrm{rep}(0, 10))^T, \quad \tilde{\mathbf{v}}_2 = (\mathrm{rep}(0, 5), \mathrm{unif}(S_v, 5), \mathrm{rep}(0, 5))^T, \\
\tilde{\mathbf{v}}_3 = (\mathrm{rep}(0, 10), \mathrm{unif}(S_v, 5))^T, \quad \mathbf{v}_j^* = \tilde{\mathbf{v}}_j / \|\tilde{\mathbf{v}}_j\|_2, \quad j = 1, 2, 3,
\end{gathered}$$

where $\mathrm{unif}(S, k)$ denotes a k-vector with i.i.d. entries from the uniform distribution on the set S, $S_u = \{-1, 1\}$, $S_v = [-1, -0.5] \cup [0.5, 1]$, $\mathrm{rep}(\alpha, k)$ denotes a k-vector replicating the value α, and $\tilde{u}_{j,k}$ is

the kth entry of $\tilde{\mathbf{u}}_j$. Model 2 is similar to Model 1 except with higher model dimensions, where we set $p = 400$, $q = 120$, and appended 300 and 80 zeros to each $\mathbf{u}_j^*$ and $\mathbf{v}_j^*$ defined above, respectively.

Models 3 and 4 pertain to the rowwise/columnwise sparse SVD structure in $\mathbf{C}^*$. Also, we intend to study the case of approximate low-rankness/sparsity, by not requiring the signals be bounded away from zero. We generated $\mathbf{X}$ with i.i.d. rows from $N_p(\mathbf{0}, \mathbf{\Sigma}_x)$, where $\mathbf{\Sigma}_x$ has diagonal entries 1 and off-diagonal entries 0.5. The rowwise sparsity patterns were generated in a similar way to the setup in Chen and Huang [22] except that we allow also the matrix of right singular vectors to be rowwise sparse, so that response selection may also be necessary. Specifically, we let $\mathbf{C}^* = \mathbf{C}_1\mathbf{C}_2^T$, where $\mathbf{C}_1 \in \mathbb{R}^{p\times r}$ with i.i.d. entries in its first p_0 rows from $N(0,1)$ and the rest set to zero, and $\mathbf{C}_2 \in \mathbb{R}^{q\times r}$ with i.i.d. entries in its first q_0 rows from $N(0,1)$ and the rest set to zero. We set $p = 100$, $p_0 = 10$, $q = q_0 = 10$ in Model 3, and $p = 400$, $p_0 = 10$, $q = 200$, and $q_0 = 10$ in Model 4. We also investigate models with even higher dimensions. In Model 5, we experimented with increasing the dimensions of Model 2 to $p = 1000$ and $q = 400$, by adding more noise variables, i.e., appending zeros to the $\mathbf{u}_j^*$ and $\mathbf{v}_j^*$ vectors.

Finally, we consider Model 6 where the orthogonality among the sparse factors is violated. Specifically, Model 6 is similar to Model 1, except that we modify the true values of $\mathbf{U}^*$ and $\mathbf{V}^*$ as follows,

$$\begin{aligned}
\tilde{\mathbf{u}}_1 &= (\text{unif}(S_u, 5), \text{rep}(0, 20))^T, \quad \tilde{\mathbf{u}}_2 = (\text{rep}(0, 3), \text{unif}(S_u, 5), \text{rep}(0, 17))^T, \\
\tilde{\mathbf{u}}_3 &= (\text{rep}(0, 8), \text{unif}(S_u, 2), \text{rep}(0, 15))^T, \quad \mathbf{u}_j^* = \tilde{\mathbf{u}}_j/\|\tilde{\mathbf{u}}_j\|_2, \quad j = 1, 2, 3, \\
\tilde{\mathbf{v}}_1 &= (\text{unif}(S_v, 5), \text{rep}(0, 10))^T, \quad \tilde{\mathbf{v}}_2 = (\text{rep}(0, 4), \text{unif}(S_v, 5), \text{rep}(0, 6))^T, \\
\tilde{\mathbf{v}}_3 &= (\text{rep}(0, 8), \text{unif}(S_v, 5), \text{rep}(0, 2))^T, \quad \mathbf{v}_j^* = \tilde{\mathbf{v}}_j/\|\tilde{\mathbf{v}}_j\|_2, \quad j = 1, 2, 3.
\end{aligned}$$

Model 7 is similar to Model 6 except with higher model dimensionality, where we set $p = 400$, $q = 120$, and appended 300 and 80 zeros to each $\mathbf{u}_j^*$ and $\mathbf{v}_j^*$ defined above, respectively. We would like to point out that when the sparse factors are not exactly orthogonal, the model in fact can be regarded as close to a two-way row-sparse SOFAR model (similar to Models 3 and 4 where both $\mathbf{U}^*$ and $\mathbf{V}^*$ are orthogonal and row-sparse); this is because if we compute the SVD of the true coefficient matrix, the resulting orthogonal factors will still have sparsity corresponding to the completely irrelevant responses and predictors.

In all the seven settings, we generated the data $\mathbf{Y}$ from the model $\mathbf{Y} = \mathbf{X}\mathbf{C}^* + \mathbf{E}$, where the error matrix $\mathbf{E}$ has i.i.d. rows from $N_q(\mathbf{0}, \sigma^2\mathbf{\Sigma})$ with $\mathbf{\Sigma} = (0.5^{|i-j|})$. In each simulation, σ^2 is computed to control the signal to noise ratio, defined as $\|d_r^*\mathbf{X}\mathbf{u}_r^*\mathbf{v}_r^{*T}\|_F/\|\mathbf{E}\|_F$, to be exactly 1. The simulation was replicated 300 times in each setting.

All methods under comparison except OLS require selection of tuning parameters, which include the rank parameter in RRR, RSSVD, and SRRR and the regularization parameters in SOFAR-L, SOFAR-GL, RSSVD, and SRRR. To reveal the full potential of each method, we chose the tuning parameters based on the predictive accuracy evaluated on a large, independently generated validation set of size 2000. The results with tuning parameters chosen by cross-validation or GIC [31] were similar to those based on a large validation set, and hence are not reported.

The model accuracy of each method is measured by the mean squared error $\|\widehat{\mathbf{C}} - \mathbf{C}^*\|_F^2/(pq)$ for estimation (MSE-Est) and $\|\mathbf{X}(\widehat{\mathbf{C}} - \mathbf{C}^*)\|_F^2/(nq)$ for prediction (MSE-Pred). The variable selection performance is characterized by the false positive rate (FPR%) and false negative rate (FNR%) in recovering the sparsity patterns of the SVD, that is, $\mathrm{FPR} = \mathrm{FP}/(\mathrm{TN} + \mathrm{FP})$ and $\mathrm{FNR} = \mathrm{FN}/(\mathrm{TP} + \mathrm{FN})$, where TP, FP, TN, and FN are the numbers of true nonzeros, false nonzeros, true zeros, and false zeros, respectively. The rank selection performance is evaluated by average estimated rank (Rank) and the percentage of correct rank identification (Rank%). Finally, for the SOFAR-L, SOFAR-GL, and RSSVD methods which explicitly produce an SVD, the orthogonality of estimated factor matrices is measured by $100(\|\widehat{\mathbf{U}}^T\widehat{\mathbf{U}}\|_1 + \|\widehat{\mathbf{V}}^T\widehat{\mathbf{V}}\|_1 - 2r)$ (Orth), which is minimized at zero when exact orthogonality is achieved.

5.1.2 Simulation results

We first compare the performance of nonadaptive and adaptive versions of the four sparse regularization methods. Because of the space constraint, only the results in terms of *MSE-Pred* in high-dimensional models 2 and 4 are presented. The comparisons in other model settings are similar and thus omitted. From Fig. 1, we observe that adaptive weighting generally improves the empirical performance of each method. For this reason, we only consider the adaptive versions of these regularization methods in other comparisons.

The comparison results with adaptive penalty for Models 1 and 2 are summarized in Table 2. The entrywise sparse SVD structure is exactly what the SOFAR-L and RSSVD methods aim to recover. We observe that SOFAR-L performs the best among all methods in terms of both model accuracy and sparsity recovery. Although RSSVD performs only second to SOFAR-L in Model I, it has substantially worse performance in Model 2 in terms of model accuracy. This is largely because the RSSVD method does not impose any form of orthogonality constraints, which tends to cause nonidentifiability issues and compromise its performance in high dimensions. We note further that SOFAR-GL and SRRR perform worse than SOFAR-L, since they are not intended for entrywise sparsity recovery. However, these two methods still provide remarkable improvements over the OLS and RRR methods due to their ability to eliminate irrelevant variables, and over the Lasso method due to the advantages of imposing a low-rank structure. Compared to SRRR, the SOFAR-GL method results in fewer false positives and shows a clear advantage due to response selection.

The simulation results for Models 3 and 4 are reported in Table 3. For the rowwise sparse SVD structure in these two models, SOFAR-GL and SRRR are more suitable than the other methods. All sparse regularization methods result in higher false negative rates than in Models 1 and 2 because of the presence of some very weak signals. In Model 3, where the matrix of right singular vectors is not sparse and the dimensionality is moderate, SOFAR-GL has a slightly worse performance compared to SRRR since response selection is unnecessary. The advantages of SOFAR are clearly seen in Model 4, where the dimension is high and many irrelevant predictors and responses coexist; SOFAR-GL performs slightly better than SOFAR-L, and both methods substantially outperform the other methods. In both

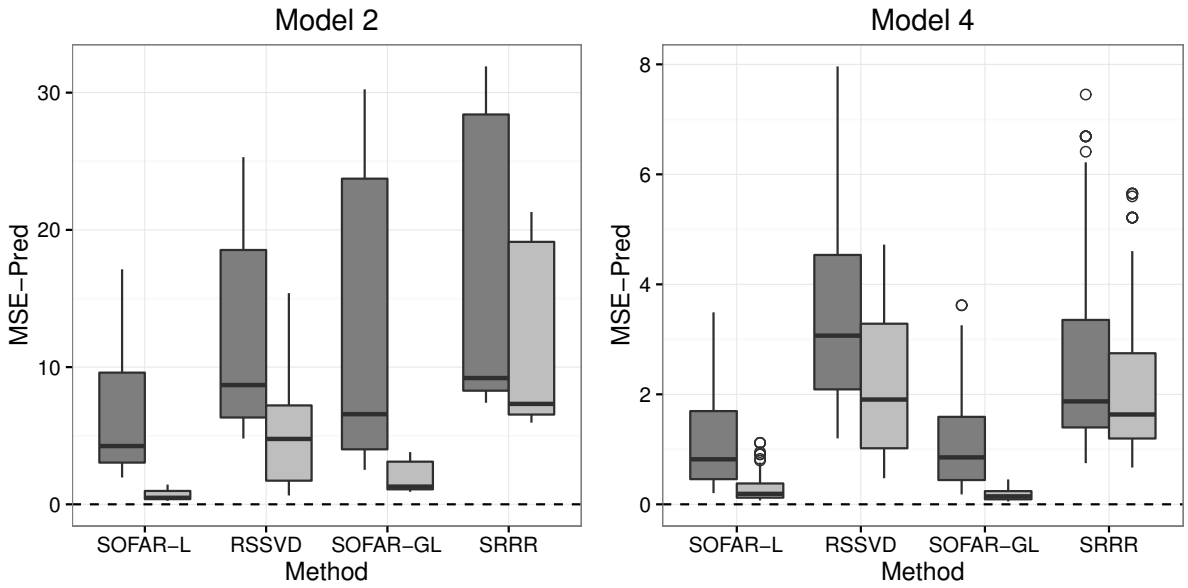

Figure 1: Boxplots of *MSE-Pred* for Models 2 and 4 with nonadaptive (dark gray) and adaptive (light gray) versions of various methods

models, SOFAR-L and RSSVD result in higher false negative rates, since they introduce more parsimony than necessary by encouraging entrywise sparsity in $\mathbf{U}$ and $\mathbf{V}$. Table 4 shows that the SOFAR methods still greatly outperform the others in both estimation and sparse recovery. In contrast, RSSVD becomes unstable and inaccurate; this again shows the effectiveness of enforcing the orthogonality in high-dimensional sparse SVD recovery.

Finally, Table 5 summarizes the results for Models 6 and 7. As expected, in Model 6 where the model dimensionality is low, SOFAR-GL performs the best, followed by RSSVD and SOFAR-L, whose performance is comparable to each other. In Model 7 where the model dimensionality is much higher, SOFAR-GL and SOFAR-L perform much better than RSSVD, as the latter becomes less stable. In both models, SRRR is outperformed by SOFAR since the former does not pursue sparsity in $\mathbf{V}$. We have also experimented with other modified settings, and all the results are consistent with what have been reported. These results confirm that it is still preferable to apply SOFAR even when the underlying sparse factors are not exactly orthogonal, especially so in high-dimensional problems.

5.2 Real data analysis

In genetical genomics experiments, gene expression levels are treated as quantitative traits in order to identify expression quantitative trait loci (eQTLs) that contribute to phenotypic variation in gene expression. The task can be regarded as a multivariate regression problem with the gene expression levels as responses and the genetic variants as predictors, where both responses and predictors are often of high dimensionality. Most existing methods for eQTL data analysis exploit entrywise or rowwise sparsity of the coefficient matrix to identify individual genetic effects or master regulators [54], which

Table 2: Simulation results for Models 1–2 with various methods[1]

Model	*Method*	*MSE-Est*	*MSE-Pred*	*FPR (%)*	*FNR (%)*	*Rank*	*Rank (%)*	*Orth*
1	OLS	250.7 (129.2)	753.8 (392.2)	100	0			
	Lasso	12.7 (5.9)	80.8 (34.1)	3.8	0			
	RRR	14.7 (6.8)	58.6 (29.3)	100	0	3	100	0
	SOFAR-L	0.4 (0.1)	2.8 (1.3)	0	0	3	100	0
	RSSVD	0.5 (0.3)	3.8 (2.3)	0.2	0	3	99.7	1.9
	SOFAR-GL	1.2 (0.5)	8.2 (4.1)	9.8	0	3	100	0
	SRRR	3.2 (1.0)	25.2 (12.6)	35.5	0	3	100	5.1
2	OLS	1013.0 (117.0)	765.6 (407.2)	100	0			
	Lasso	21.3 (7.0)	59.0 (18.1)	1.3	0			
	RRR	756.4 (56.8)	30.2 (15.9)	100	0	3	0	0
	SOFAR-L	0.2 (0.1)	0.7 (0.3)	0	0	3	0	0
	RSSVD	2.5 (2.4)	5.3 (4.1)	1	0.1	3	0	28.4
	SOFAR-GL	0.7 (0.4)	2.0 (1.0)	2.7	0	3	0	0
	SRRR	3.8 (1.5)	12.0 (6.3)	19.8	0	3	0	40.2

[1]Adaptive versions of Lasso, SOFAR-L, RSSVD, SOFAR-GL, and SRRR were applied. Means of performance measures with standard deviations in parentheses over 300 replicates are reported. *MSE-Est* values are scaled by multiplying 10^4 in Model 1 and 10^5 in Model 2, and *MSE-Pred* values are scaled by multiplying 10^3.

Table 3: Simulation results for Models 3–4 with various methods[1]

Model	*Method*	*MSE-Est*	*MSE-Pred*	*FPR (%)*	*FNR (%)*	*Rank*	*Rank (%)*	*Orth*
3	OLS	599.2 (339.2)	1530.1 (870.8)	100	0			
	Lasso	97.6 (50.0)	472.8 (242.7)	15.5	0.6			
	RRR	102.6 (70.2)	291.9 (191.8)	100	0	3	100	0
	SOFAR-L	24.8 (15.3)	129.5 (83.2)	0.3	7.4	3.7	30.3	0
	RSSVD	17.3 (11.3)	96.6 (66.4)	0.6	11	3	100	29
	SOFAR-GL	16.6 (11.4)	94.4 (67.5)	0.4	1.1	3.6	41.7	0
	SRRR	11.0 (6.7)	63.1 (40.2)	0.6	0.3	3	100	14.8
4	OLS	252.3 (78)	126.5 (65.4)	100	0			
	Lasso	37.4 (11.8)	73.2 (24.1)	0.8	2.5			
	RRR	186.6 (51.6)	6.1 (3.9)	100	0	3	99	0
	SOFAR-L	0.1 (0.1)	0.3 (0.2)	0.1	4.8	3	92.7	0.1
	RSSVD	1.0 (0.7)	2.2 (1.3)	0.3	11.5	3	100	40.1
	SOFAR-GL	0.1 (0.0)	0.2 (0.1)	0	0.1	3	100	0
	SRRR	0.8 (0.5)	2.0 (1.2)	24.9	0.2	3	100	31.3

[1]Adaptive versions of Lasso, SOFAR-L, RSSVD, SOFAR-GL, and SRRR were applied. Means of performance measures with standard deviations in parentheses over 300 replicates are reported. *MSE-Est* values are scaled by multiplying 10^4 in Model 3 and 10^5 in Model 4, and *MSE-Pred* values are scaled by multiplying 10^3.

Table 4: Simulation results for Model 5. We use Model 2 with increased dimensions $p = 1000$, $q = 400$ by adding noise variables[1]

Model	*Method*	*MSE-Est*	*MSE-Pred*	*FPR (%)*	*FNR (%)*	*Rank*	*Rank (%)*	*Orth*
5	OLS	151.5 (5.7)	230.1 (122.9)	100	0			
	Lasso	3.9 (1.8)	29.3 (11.8)	0.6	0			
	RRR	146.8 (7.7)	61.5 (77.1)	100	0	2.6	57.7	0
	SOFAR-L	0.1 (0.0)	0.1 (0.0)	0	0	3	100	0
	RSSVD	6.6 (14.4)	2.8 (2.7)	3.1	1	3	99	49.1
	SOFAR-GL	0.1 (0.0)	0.2 (0.1)	0.8	0	3	100	0
	SRRR	0.5 (0.2)	3.6 (1.8)	19.7	0	3	100	55.5

[1]Adaptive versions of Lasso, SOFAR-L, RSSVD, SOFAR-GL, and SRRR were applied. Means of performance measures with standard deviations in parentheses over 300 replicates are reported. *MSE-Est* values are scaled by 10^5 and *MSE-Pred* values are scaled by multiplying 10^3.

Table 5: Simulation results for Models 6–7 when the sparse factors are not exactly orthogonal.[1]

Model	*Method*	*MSE-Est*	*MSE-Pred*	*FPR (%)*	*FNR (%)*	*Rank*	*Rank (%)*	*Orth*
6	OLS	291.7 (133.9)	868.5 (403.1)	100	0			
	Lasso	11.8 (4.7)	71.9 (28.4)	10.3	0			
	RRR	17.6 (7.3)	69.5 (31.2)	100	0	3	100	0
	SOFAR-L	2.2 (0.9)	14.1 (6.3)	8	0.1	3	100	0
	RSSVD	2.0 (0.7)	13.6 (6.1)	7.3	0.1	3	100	22.7
	SOFAR-GL	1.2 (0.5)	8.9 (3.9)	9.5	0	3	100	0
	SRRR	3.5 (1.1)	28.7 (13.1)	36.2	0	3	100	5.6
7	OLS	1045.1 (122.4)	886.5 (403.4)	100	0			
	Lasso	33.2 (12.7)	79.9 (24.4)	3.2	0			
	RRR	754.0 (67.8)	35.2 (15.4)	100	0	3	99.7	0
	SOFAR-L	1.2 (0.5)	3.1 (1.5)	2.4	0.2	3	100	0
	RSSVD	4.8 (4.1)	8.5 (5.1)	2.4	1.7	3	99.7	64.4
	SOFAR-GL	1.0 (0.4)	2.6 (1.1)	3.6	0	3	100	0
	SRRR	4.3 (1.5)	13.9 (6.2)	20	0	3	100	45

[1]Adaptive versions of Lasso, SOFAR-L, RSSVD, SOFAR-GL, and SRRR were applied. Means of performance measures with standard deviations in parentheses over 300 replicates are reported. *MSE-Est* values are scaled by multiplying 10^4 in Model 6 and 10^5 in Model 7, and *MSE-Pred* values are scaled by multiplying 10^3.

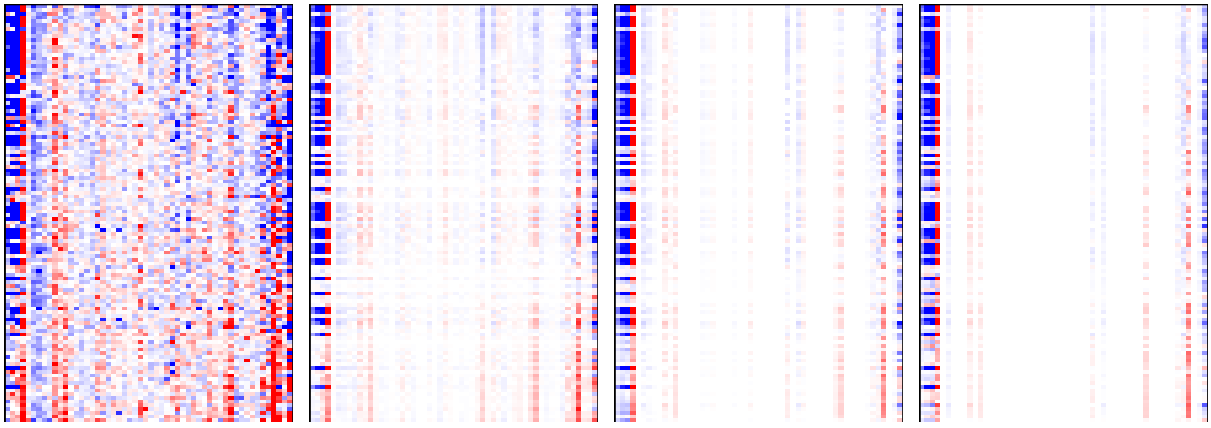

Figure 2: Heat maps of $\mathbf{Y}$ and its estimates by RRR, SOFAR-L, and SOFAR-GL (from left to right)

not only tends to suffer from low detection power for multiple eQTLs that combine to affect a subset of gene expression traits, but also may offer little information about the functional grouping structure of the genetic variants and gene expressions. By exploiting a sparse SVD structure, the SOFAR method is particularly appealing for such applications, and may provide new insights into the complex genetics of gene expression variation. Here the orthogonality can be roughly interpreted as maximum separability, so that different association layers are more likely to reflect different functional pathways.

We illustrate our approach by the analysis of a yeast eQTL data set described by Brem and Kruglyak [14], where $n = 112$ segregants were grown from a cross between two budding yeast strains, BY4716 and RM11-1a. For each of the segregants, gene expression was profiled on microarrays containing 6216 genes, and genotyping was performed at 2957 markers. Similar to Yin and Li [64], we combined the markers into blocks such that markers with the same block differed by at most one sample, and one representative marker was chosen from each block; a marginal gene–marker association analysis was then performed to identify markers that are associated with the expression levels of at least two genes with a p-value less than 0.05, resulting in a total of $p = 605$ markers.

Owing to the small sample size and weak genetic perturbations, we focused our analysis on $q = 54$ genes in the yeast MAPK signaling pathways [41]. We then applied the proposed SOFAR methods with adaptive weighting. Both SOFAR-L and SOFAR-GL methods resulted in a model of rank 3, indicating that dimension reduction is very effective for the data set. Also, the SVD layers estimated by the SOFAR methods are indeed sparse. The SOFAR-L estimates include 140 nonzeros in $\widehat{\mathbf{U}}$, which involve only 112 markers, and 40 nonzeros in $\widehat{\mathbf{V}}$, which involve only 27 genes. The sparse SVD produced by SOFAR-GL involves only 34 markers and 15 genes. The SOFAR-GL method is more conservative since it tends to identify markers that regulate all selected genes rather than a subset of genes involved in a specific SVD layer. We compare the original gene expression matrix $\mathbf{Y}$ and its estimates $\mathbf{X}\widehat{\mathbf{C}}$ by the RRR, SOFAR-L and SOFAR-GL methods using heat maps in Fig. 2. It is seen that the SOFAR methods achieve both low-rankness and sparsity, while still capturing main patterns in the original matrix.

Fig. 3 shows the scatterplots of the latent responses $\mathbf{Y}\widehat{\mathbf{v}}_j$ versus the latent predictors $\mathbf{X}\widehat{\mathbf{u}}_j$ for $j =$

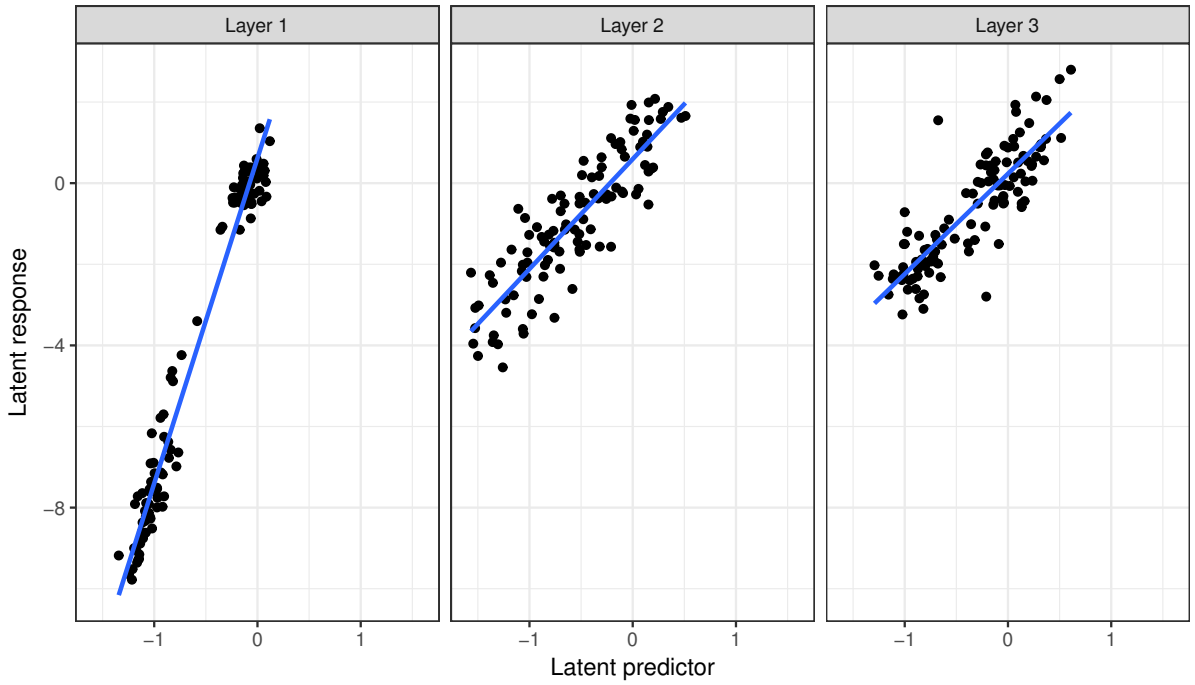

Figure 3: Scatterplots of the latent responses versus the latent predictors in three SVD layers for the yeast data estimated by the SOFAR-L method

$1, 2, 3$, where $\widehat{\mathbf{u}}_j$ and $\widehat{\mathbf{v}}_j$ are the jth columns of $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$, respectively. The plots demonstrate a strong association between each pair of latent variables, with the association strength descending from layer 1 to layer 3. A closer look at the SVD layers reveals further information about clustered samples and genes. The plot for layer 1 indicates that the yeast samples form two clusters, suggesting that our method may be useful for classification based on the latent variables. Also, examining the nonzero entries in $\widehat{\mathbf{v}}_1$ shows that this layer is dominated by four genes, namely, STE3 (-0.66), STE2 (0.59), MFA2 (0.40), and MFA1 (0.22). All four genes are upstream in the pheromone response pathway, where MFA2 and MFA1 are genes encoding mating pheromones and STE3 and STE2 are genes encoding pheromone receptors [24]. The second layer is mainly dominated by CTT1 (-0.93), and other leading genes include SLN1 (0.16), SLT2 (-0.14), MSN4 (-0.14), and GLO1 (-0.13). Interestingly, CTT1, MSN4, and GLO1 are all downstream genes linked to the upstream gene SLN1 in the high osmolarity/glycerol pathway required for survival in response to hyperosmotic stress. Finally, layer 3 includes the leading genes FUS1 (0.81), FAR1 (0.32), STE2 (0.25), STE3 (0.24), GPA1 (0.22), FUS3 (0.18), and STE12 (0.11). These genes consist of two major groups that are downstream (FUS1, FAR1, FUS3, and STE12) and upstream (STE2, STE3, and GPA1) in the pheromone response pathway. Overall, our results suggest that there are common genetic components shared by the expression traits of the clustered genes and clearly reveal strong associations between the upstream and downstream genes on several signaling pathways, which are consistent with the current functional understanding of the MAPK signaling pathways.

To examine the predictive performance of SOFAR and other competing methods, we randomly split the data into a training set of size 92 and a test set of size 20. The model was fitted using the

training set and the predictive accuracy was evaluated on the test set based on the prediction error $\|\mathbf{Y} - \mathbf{X}\widehat{\mathbf{C}}\|_F^2/(nq)$. The splitting process was repeated 50 times. The scaled prediction errors for the RRR, SOFAR-L, SOFAR-GL, RSSVD, and SRRR methods are 3.4 (0.3), 2.6 (0.2), 2.5 (0.2), 2.9 (0.3), and 2.6 (0.2), respectively. The comparison shows the advantages of sparse and low-rank estimation. RSSVD yields higher prediction error and is less stable than the SOFAR methods. Although the SRRR method yielded similar predictive accuracy compared to SOFAR methods on this data set, it resulted in a less parsimonious model and cannot be used for gene selection or clustering.

References

[1] Anderson, T. W. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.*, **22**, 327–351.

[2] Bai, J. (2003) Inferential theory for factor models of large dimensions. *Econometrica*, **71**, 135–171.

[3] Bai, J. and Li, K. (2012) Statistical analysis of factor models of high dimension. *Ann. Statist.*, **40**, 436–465.

[4] — (2016) Maximum likelihood estimation and inference for approximate factor models of high dimension. *Review of Economics and Statistics*, **98**, 298–309.

[5] Bai, J. and Ng, S. (2002) Determining the number of factors in approximate factor models. *Econometrica*, **70**, 191–221.

[6] — (2008) Large dimensional factor analysis. *Foundns Trends Econmetr.*, **3**, 89–163.

[7] — (2013) Principal components estimation and identification of static factors. *Journal of Econometrics*, **176**, 18–29. URL: https://ideas.repec.org/a/eee/econom/v176y2013i1p18-29.html.

[8] Basu, S. and Michailidis, G. (2015) Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, **43**, 1535–1567.

[9] Benidis, K., Sun, Y., Babu, P. and Palomar, D. P. (2016) Orthogonal sparse pca and covariance estimation via procrustes reformulation. *IEEE Trans. on Signal Processing*, **64**, 6211–6226.

[10] Bernanke, B. S., Boivin, J. and Eliasz, P. (2005) Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *Q. J. Econ.*, **120**, 387–422.

[11] Bickel, P., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of lasso and dantzig selector. *Annals of statistics*, **37**, 1705–1732.

[12] Box, G. E. P. and Tiao, G. C. (1977) A canonical analysis of multiple time series. *Biometrika*, **64**, 355–365.

[13] Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundns Trends Mach. Learn.*, **3**, 1–122.

[14] Brem, R. B. and Kruglyak, L. (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natn. Acad. Sci. USA*, **102**, 1572–1577.

[15] Bunea, F., She, Y. and Wegkamp, M. H. (2011) Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.*, **39**, 1282–1309.

[16] — (2012) Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.*, **40**, 2359–2388.

[17] Busygin, S., Prokopyev, O. and Pardalos, P. M. (2008) Biclustering in data mining. *Comput. Oper. Res.*, **35**, 2964–2987.

[18] Cai, T. T., Li, H., Liu, W. and Xie, J. (2013) Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, **100**, 139–156.

[19] Chen, K., Chan, K.-S. and Stenseth, N. C. (2012) Reduced rank stochastic regression with a sparse singular value decomposition. *J. R. Statist. Soc. B*, **74**, 203–221.

[20] — (2014) Source-sink reconstruction through regularized multicomponent regression analysis–with application to assessing whether North Sea cod larvae contributed to local fjord cod in Skagerrak. *Journal of the American Statistical Association*, **109**, 560–573.

[21] Chen, K., Dong, H. and Chan, K.-S. (2013) Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, **100**, 901–920.

[22] Chen, L. and Huang, J. Z. (2012) Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Am. Statist. Ass.*, **107**, 1533–1545.

[23] — (2016) Sparse reduced-rank regression with covariance estimation. *Statistics and Computing*, **26**, 461–470.

[24] Chen, R. E. and Thorner, J. (2007) Function and regulation in MAPK signaling pathways: Lessons learned from the yeast *Saccharomyces Cerevisiae*. *Biochim. Biophys. Acta*, **1773**, 1311–1340.

[25] d'Aspremont, A., El Ghaoui, L., Jordan, M. I. and Lanckriet, G. R. G. (2007) A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.*, **49**, 434–448.

[26] Edelman, A., Arias, T. A. and Smith, S. T. (1998) The geometry of algorithms with orthogonality constraints. *SIAM J. Matrx Anal. Appl.*, **20**, 303–353.

[27] Fan, J., Fan, Y. and Barut, E. (2014) Adaptive robust variable selection. *The Annals of Statistics*, **42**, 324–351.

[28] Fan, J., Fan, Y. and Lv, J. (2008) High dimensional covariance matrix estimation using a factor model. *J. Econmetr.*, **147**, 186–197.

[29] Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348–1360.

[30] Fan, Y. and Lv, J. (2013) Asymptotic equivalence of regularization methods in thresholded parameter space. *J. Am. Statist. Ass.*, **108**, 1044–1061.

[31] Fan, Y. and Tang, C. Y. (2013) Tuning parameter selection in high dimensional penalized likelihood. *J. R. Statist. Soc. B*, **75**, 531–552.

[32] Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1**, 302–332.

[33] Goh, G., Dey, D. K. and Chen, K. (2017) Bayesian sparse reduced rank multivariate regression. *Journal of Multivariate Analysis*, **157**, 14–28.

[34] Golub, G. H. and Van Loan, C. F. (2013) *Matrix Computations*. Baltimore: The Johns Hopkins University Press, 4th edn.

[35] Guo, J., James, G., Levina, E., Michailidis, G. and Zhu, J. (2010) Principal component analysis with sparse fused loadings. *J. Computnl Graph. Statist.*, **19**, 930–946.

[36] Gustin, M. C., Albertyn, J., Alexander, M. and Davenport, K. (1998) Map kinase pathways in the yeast saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews*, **62**, 1264–1300.

[37] Hartigan, J. A. (1972) Direct clustering of a data matrix. *J. Am. Statist. Ass.*, **67**, 123–129.

[38] Hsu, N.-J., Hung, H.-L. and Chang, Y.-M. (2008) Subset selection for vector autoregressive processes using Lasso. *Computnl Statist. Data Anal.*, **52**, 3645–3657.

[39] Izenman, A. J. (1975) Reduced-rank regression for the multivariate linear model. *J. Multiv. Anal.*, **5**, 248–264.

[40] Johnstone, I. M. and Lu, A. Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Statist. Ass.*, **104**, 682–703.

[41] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2014) Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.

[42] Kock, A. and Callot, L. (2015) Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, **186**, 325–344.

[43] Koschat, M. A. and Swayne, D. F. (1991) A weighted Procrustes criterion. *Psychometrika*, **56**, 229–239.

[44] Lee, M., Shen, H., Huang, J. Z. and Marron, J. S. (2010) Biclustering via sparse singular value decomposition. *Biometrics*, **66**, 1087–1095.

[45] Leng, C. and Wang, H. (2009) On general adaptive sparse principal component analysis. *J. Computnl Graph. Statist.*, **18**, 201–215.

[46] Lian, H., Feng, S. and Zhao, K. (2015) Parametric and semiparametric reduced-rank regression with flexible sparsity. *Journal of Multivariate Analysis*, **136**, 163 – 174.

[47] Lv, J. (2013) Impacts of high dimensionality in finite samples. *The Annals of Statistics*, **41**, 2236–2262.

[48] Ma, X., Xiao, L. and Wong, W. H. (2014) Learning regulatory programs by threshold svd regression. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 15675–15680.

[49] Ma, Z., Ma, Z. and Sun, T. (2014) Adaptive estimation in two-way sparse reduced-rank regression. *ArXiv e-prints arXiv:1403.1922*.

[50] Ma, Z. and Sun, T. (2014) Adaptive sparse reduced-rank regression. *ArXiv e-prints arXiv:1403.1922*.

[51] Mirsky, L. (1960) Symmetric gauge functions and unitarily invariant norms. *Quarterly Journal of Mathematics*, **11**, 50–59.

[52] Nardi, Y. and Rinaldo, A. (2011) Autoregressive process modeling via the Lasso procedure. *J. Multiv. Anal.*, **102**, 528–549.

[53] Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012) A unified framework for high-dimensional decomposable regularizers. *Statistical Science*, **27**, 538–557.

[54] Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R. and Wang, P. (2010) Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Statist.*, **4**, 53–77.

[55] Reinsel, G. C. and Velu, R. P. (1998) *Multivariate Reduced-Rank Regression: Theory and Applications*. New York: Springer.

[56] Sha, F., Lin, Y., Saul, L. K. and Lee, D. D. (2007) Multiplicative updates for nonnegative quadratic programming. *Neur. Computn*, **19**, 2004–2031.

[57] Shen, H. and Huang, J. Z. (2008) Sparse principal component analysis via regularized low rank matrix approximation. *J. Multiv. Anal.*, **99**, 1015–1034.

[58] Stock, J. H. and Watson, M. W. (2001) Vector autoregressions. *J. Econ. Perspect.*, **15**, 101–115.

[59] — (2002) Forecasting using principal components from a large number of predictors. *J. Am. Statist. Ass.*, **97**, 1167–1179.

[60] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, **58**, 267–288.

[61] Tseng, P. (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optimizn Theor. Appl.*, **109**, 475–494.

[62] Velu, R. P., Reinsel, G. C. and Wichern, D. W. (1986) Reduced rank models for multiple time series. *Biometrika*, **73**, 105–118.

[63] Witten, D. M., Tibshirani, R. and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

[64] Yin, J. and Li, H. (2011) A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Statist.*, **5**, 2630–2650.

[65] Yu, Y., Wang, T. and Samworth, R. (2015) A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, **102**, 315–323.

[66] Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007) Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Statist. Soc. B*, **69**, 329–346.

[67] Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.

[68] Zhang, Z., Zha, H. and Simon, H. (2002) Low-rank approximations with sparse factors I: Basic algorithms and error analysis. *SIAM J. Matrx Anal. Appl.*, **23**, 706–727.

[69] Zheng, Z., Fan, Y. and Lv, J. (2014) High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society Series B*, **76**, 627–649.

[70] Zhu, H., Khondker, Z., Lu, Z. and Ibrahim, J. G. (2014) Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association*, **109**, 997–990.

[71] Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.

[72] Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *J. Computnl Graph. Statist.*, **15**, 265–286.

[73] Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**, 1509–1533.

Supplementary Material to "SOFAR: Large-Scale Association Network Learning"

Yoshimasa Uematsu[1], Yingying Fan[1], Kun Chen[2], Jinchi Lv[1] and Wei Lin[3]

University of Southern California[1], University of Connecticut[2] and Peking University[3]

This Supplementary Material contains the proofs of Theorems 1–3 and additional technical details.

A Proofs of main results

To ease the technical presentation, we introduce some necessary notation. Recall that $\mathbf{A}^* = \mathbf{U}^*\mathbf{D}^*$, $\mathbf{B}^* = \mathbf{V}^*\mathbf{D}^*$, $\mathbf{A} = \mathbf{U}\mathbf{D}$, and $\mathbf{B} = \mathbf{V}\mathbf{D}$. Denote by $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{C}} - \mathbf{C}^*$, $\widehat{\mathbf{\Delta}}^d = \widehat{\mathbf{D}} - \mathbf{D}^*$, $\widehat{\mathbf{\Delta}}^a = \widehat{\mathbf{A}} - \mathbf{A}^*$, and $\widehat{\mathbf{\Delta}}^b = \widehat{\mathbf{B}} - \mathbf{B}^*$ the different estimation errors, and $\mathrm{FS}(\widehat{\mathbf{M}}) = |\{(i,j) : \mathrm{sgn}(\widehat{m}_{ij}) \neq \mathrm{sgn}(m^*_{ij})\}|$ the total number of falsely discovered signs of an estimator $\widehat{\mathbf{M}} = (\widehat{m}_{ij})$ for matrix $\mathbf{M}^* = (m^*_{ij})$. For $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_m) \in \mathbb{R}^{m\times m}$, we define $\mathbf{D}^-$ as a diagonal matrix with $\mathrm{rank}(\mathbf{D}^-) = \mathrm{rank}(\mathbf{D})$ and jth diagonal entry $d_j^- = d_j^{-1}1\{d_j > 0\}$, and define $\mathbf{D}^{*-}$ based on $\mathbf{D}^*$ similarly. For any matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, denote by $\langle \mathbf{M}_1, \mathbf{M}_2\rangle = \mathrm{tr}(\mathbf{M}_1^T\mathbf{M}_2)$. Hereafter we use c to denote a generic positive constant whose value may vary from line to line.

A.1 Proof of Theorem 1

We prove the bounds in (11)–(13) separately. Recall that $s = \|\mathbf{C}^*\|_0$ and define a space

$$\mathcal{C}_0 = \{\mathbf{M} \in \mathbb{R}^{p\times q} : m_{ij} = 0 \text{ for } (i,j) \neq S\},$$

where S stands for the support of $\mathbf{C}^*$. We also denote by $\mathcal{C}_0^\perp$ the orthogonal complement of $\mathcal{C}_0$.

Part 1: Proof of bound (11). The proof is composed of two steps. We first derive the *deterministic* error bound (11) under the assumption that

$$\|n^{-1}\mathbf{X}^T\mathbf{E}\|_\infty \leq \lambda_0/2 \tag{A.1}$$

holds almost surely in the first step and then verify that condition (A.1) holds with high probability in the second step.

Step 1. Since the objective function is convex, the global optimality of $\widetilde{\mathbf{C}}$ implies

$$(2n)^{-1}\|\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{C}}\|_F^2 + \lambda_0\|\widetilde{\mathbf{C}}\|_1 \leq (2n)^{-1}\|\mathbf{Y} - \mathbf{X}\mathbf{C}^*\|_F^2 + \lambda_0\|\mathbf{C}^*\|_1.$$

Then letting $\widetilde{\mathbf{\Delta}} \equiv \widetilde{\mathbf{C}} - \mathbf{C}^*$, we see that

$$(2n)^{-1}\|\mathbf{X}\widetilde{\mathbf{\Delta}}\|_F^2 \leq \langle n^{-1}\mathbf{X}^T\mathbf{E}, \widetilde{\mathbf{\Delta}}\rangle + \lambda_0(\|\mathbf{C}^*\|_1 - \|\widetilde{\mathbf{\Delta}} + \mathbf{C}^*\|_1). \tag{A.2}$$

By Hölder's inequality and the assumed condition (A.1), it holds that

$$\langle n^{-1}\mathbf{X}^T\mathbf{E}, \widetilde{\mathbf{\Delta}}\rangle \leq \|n^{-1}\mathbf{X}^T\mathbf{E}\|_\infty\|\widetilde{\mathbf{\Delta}}\|_1 \leq 2^{-1}\lambda_0\|\widetilde{\mathbf{\Delta}}\|_1. \tag{A.3}$$

By the triangle inequality, we have

$$\lambda_0(\|\mathbf{C}^*\|_1 - \|\widetilde{\mathbf{\Delta}} + \mathbf{C}^*\|_1) \leq \lambda_0\|\widetilde{\mathbf{\Delta}}\|_1. \tag{A.4}$$

Therefore, (A.2) together with Lemma 4 in Section B.2 and (A.3)–(A.4) entails that

$$2c_2\|\widetilde{\mathbf{\Delta}}\|_F^2 \leq 2n^{-1}\|\mathbf{X}\widetilde{\mathbf{\Delta}}\|_F^2 \leq 6\lambda_0\|\widetilde{\mathbf{\Delta}}\|_1. \tag{A.5}$$

Meanwhile, since $n^{-1}\|\mathbf{X}\widetilde{\mathbf{\Delta}}\|_F^2$ is nonnegative (A.2) is also bounded from below as

$$0 \leq \langle n^{-1}\mathbf{X}^T\mathbf{E}, \widetilde{\mathbf{\Delta}}\rangle + \lambda_0(\|\mathbf{C}^*\|_1 - \|\widetilde{\mathbf{\Delta}} + \mathbf{C}^*\|_1). \tag{A.6}$$

Note that $\mathbf{C}^*_{\mathcal{C}_0^\perp} = \mathbf{0}$ in our model. Hence it follows from the triangle inequality and decomposability of the nuclear norm that

$$\begin{aligned}
\lambda_0(\|\mathbf{C}^*\|_1 - \|\widetilde{\mathbf{\Delta}} + \mathbf{C}^*\|_1) &= \lambda_0(\|\mathbf{C}^*_{\mathcal{C}_0} + \mathbf{C}^*_{\mathcal{C}_0^\perp}\|_1 - \|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0} + \widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp} + \mathbf{C}^*_{\mathcal{C}_0} + \mathbf{C}^*_{\mathcal{C}_0^\perp}\|_1) \\
&\leq \lambda_0(\|\mathbf{C}^*_{\mathcal{C}_0}\|_1 + \|\mathbf{C}^*_{\mathcal{C}_0^\perp}\|_1 - \|\mathbf{C}^*_{\mathcal{C}_0} + \widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1 + \|\mathbf{C}^*_{\mathcal{C}_0^\perp} + \widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1) \\
&= \lambda_0(\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 - \|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1).
\end{aligned} \tag{A.7}$$

Thus by (A.3) and (A.7), we can bound (A.6) from above as

$$\begin{aligned}
0 &\leq 2^{-1}\lambda_0\|\widetilde{\mathbf{\Delta}}\|_1 + \lambda_0(\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 - \|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1) \\
&\leq 2^{-1}\lambda_0(\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 + \|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1) + \lambda_0(\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 - \|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1) \\
&= 2^{-1}\lambda_0(3\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 - \|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1),
\end{aligned}$$

which can be equivalently rewritten as

$$\lambda_0\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1 \leq 3\lambda_0\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1. \tag{A.8}$$

We are now ready to derive the error bound. For a generic positive constant c, (A.5) is bounded from

above by the decomposability of the ℓ_1-norm and (A.8) as

$$c\|\widetilde{\mathbf{\Delta}}\|_F^2 \leq \lambda_0\|\widetilde{\mathbf{\Delta}}\|_1 = \lambda_0\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 + \lambda_0\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1 \leq 4\lambda_0\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1. \tag{A.9}$$

Using the subspace compatibility conditions (see the proof of Theorem 1 of [53]), we can show that

$$\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 \leq s^{1/2}\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_F \leq s^{1/2}\|\widetilde{\mathbf{\Delta}}\|_F.$$

Therefore, with c changed appropriately (A.9) can be further bounded as

$$\|\widetilde{\mathbf{\Delta}}\|_F^2 \leq cs^{1/2}\lambda_0\|\widetilde{\mathbf{\Delta}}\|_F.$$

This consequently yields the desired error bound

$$\|\widetilde{\mathbf{\Delta}}\|_F \leq cs^{1/2}\lambda_0,$$

which completes the first step of the proof.

Step 2. Let $\mathbf{x}_i$ and $\mathbf{e}_j$ denote the ith and jth columns of $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{E} \in \mathbb{R}^{n\times q}$, respectively. Since $\|\mathbf{X}^T\mathbf{E}\|_\infty = \max_{1\leq i\leq p}\max_{1\leq j\leq q}|\mathbf{x}_i^T\mathbf{e}_j|$, using Bonferroni's inequality and the Gaussianity of $\mathbf{e}_j$ we deduce

$$\begin{aligned} P\left(n^{-1}\|\mathbf{X}^T\mathbf{E}\|_\infty \geq \lambda_0\right) &\leq \sum_{i=1}^{p}\sum_{j=1}^{q} P\left(n^{-1}|\mathbf{x}_i^T\mathbf{e}_j| \geq \lambda_0\right) \\ &\leq 2\sum_{i=1}^{p}\sum_{j=1}^{q}\exp\left(-\frac{n^2\lambda_0^2}{2\mathbb{E}|\mathbf{x}_i^T\mathbf{e}_j|^2}\right). \end{aligned} \tag{A.10}$$

Since $\mathbf{e}_j$ is distributed as $N\left(0, \sigma_j^2\mathbf{I}_n\right)$, it holds that

$$\mathbb{E}|\mathbf{x}_i^T\mathbf{e}_j|^2 = \sigma_j^2\mathbf{x}_i^T\mathbf{x}_i \leq \sigma_{\max}^2 n. \tag{A.11}$$

By the assumption $\lambda_0^2 = c_0^2\sigma_{\max}^2 n^{-1}\log(pq)$ and (A.10)–(A.11), the upper bound on the probability in (A.10) can be further bounded from above by

$$2pq\exp\left\{-(c_0^2/2)\log(pq)\right\} = 2(pq)^{1-c_0^2/2},$$

which concludes the proof for bound (11).

Part 2: Proofs of bounds (12) and (13). Both inequalities (12) and (13) are direct consequences of Lemma 3 in Section B.1 and bound (11). This completes the proof of Theorem 1.

A.2 Proof of Theorem 2

Recall that we solve SOFAR in a local neighborhood $\mathcal{P}_n$ of the initial solution $\widetilde{\mathbf{C}}$. It follows that $\|\widehat{\mathbf{\Delta}}\|_F \le \|\widehat{\mathbf{C}} - \widetilde{\mathbf{C}}\|_F + \|\widetilde{\mathbf{C}} - \mathbf{C}^*\|_F \le 3R_n \le cs^{1/2}\lambda_{\max}$, where $\mathcal{P}_n$ is defined in (14), R_n is as in Theorem 1, and c is some generic positive constant. Thus by Lemma 3, we have

$$\|\widehat{\mathbf{\Delta}}^a\|_F + \|\widehat{\mathbf{\Delta}}^b\|_F + \|\widehat{\mathbf{\Delta}}^d\|_F \le c\eta_n\|\widehat{\mathbf{\Delta}}\|_F \tag{A.12}$$

$$\le cs^{1/2}\lambda_{\max}\eta_n, \tag{A.13}$$

where $\eta_n = 1 + \delta^{-1/2}\big(\sum_{j=1}^r (d_1^*/d_j^*)^2\big)^{1/2}$. Note that under Conditions 1 and 2, Lemma 4 and Lemma 1 in Section A.3 entail that

$$\|\widehat{\mathbf{\Delta}}\|_F^2 \le cn^{-1}\|\mathbf{X}\widehat{\mathbf{\Delta}}\|_F^2 \le c\lambda_{\max}\Big(\|\widehat{\mathbf{\Delta}}^d\|_1 + \|\widehat{\mathbf{\Delta}}^a\|_1 + \|\widehat{\mathbf{\Delta}}^b\|_1\Big). \tag{A.14}$$

Furthermore, it follows from the Cauchy–Schwarz inequality and (A.12) that

$$\begin{aligned}
&\|\widehat{\mathbf{\Delta}}^a\|_1 + \|\widehat{\mathbf{\Delta}}^d\|_1 + \|\widehat{\mathbf{\Delta}}^b\|_1 \\
&\le \max\{\|\widehat{\mathbf{\Delta}}^d\|_0, \|\widehat{\mathbf{\Delta}}^a\|_0, \|\widehat{\mathbf{\Delta}}^b\|_0\}^{1/2}\Big(\|\widehat{\mathbf{\Delta}}^a\|_F + \|\widehat{\mathbf{\Delta}}^d\|_F + \|\widehat{\mathbf{\Delta}}^b\|_F\Big) \\
&\le c\eta_n\{\|\widehat{\mathbf{\Delta}}^d\|_0 + \|\widehat{\mathbf{\Delta}}^a\|_0 + \|\widehat{\mathbf{\Delta}}^b\|_0\}^{1/2}\|\widehat{\mathbf{\Delta}}\|_F.
\end{aligned} \tag{A.15}$$

Combining (A.15) and (A.14) leads to

$$\|\widehat{\mathbf{\Delta}}\|_F \le c\lambda_{\max}\eta_n\{\|\widehat{\mathbf{\Delta}}^d\|_0 + \|\widehat{\mathbf{\Delta}}^a\|_0 + \|\widehat{\mathbf{\Delta}}^b\|_0\}^{1/2}. \tag{A.16}$$

We next provide an upper bound for $\|\widehat{\mathbf{\Delta}}^d\|_0 + \|\widehat{\mathbf{\Delta}}^a\|_0 + \|\widehat{\mathbf{\Delta}}^b\|_0$. Since $(\widehat{\mathbf{D}}, \widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ and $(\mathbf{D}^*, \mathbf{A}^*, \mathbf{B}^*)$ are elements in $\mathcal{D} \times \mathcal{A} \times \mathcal{B}$ by Condition 1, we have

$$\mathrm{FS}(\widehat{\mathbf{D}})^{1/2}\tau \le \|\widehat{\mathbf{\Delta}}^d\|_F, \quad \mathrm{FS}(\widehat{\mathbf{A}})^{1/2}\tau \le \|\widehat{\mathbf{\Delta}}^a\|_F, \quad \text{and} \quad \mathrm{FS}(\widehat{\mathbf{B}})^{1/2}\tau \le \|\widehat{\mathbf{\Delta}}^b\|_F. \tag{A.17}$$

By the definition of $\mathrm{FS}(\widehat{\mathbf{A}})$, it holds that $\|\widehat{\mathbf{\Delta}}^a\|_0 \le s_a + \mathrm{FS}(\widehat{\mathbf{A}})$. Similar inequalities hold for $\|\widehat{\mathbf{\Delta}}^b\|_0$ and $\|\widehat{\mathbf{\Delta}}^d\|_0$. Therefore, it follows from (A.17) and (A.12) that

$$\begin{aligned}
\|\widehat{\mathbf{\Delta}}^d\|_0 + \|\widehat{\mathbf{\Delta}}^a\|_0 + \|\widehat{\mathbf{\Delta}}^b\|_0 &\le r + s_a + s_b + \mathrm{FS}(\widehat{\mathbf{D}}) + \mathrm{FS}(\widehat{\mathbf{A}}) + \mathrm{FS}(\widehat{\mathbf{B}}) \\
&\le r + s_a + s_b + \tau^{-2}\Big(\|\widehat{\mathbf{\Delta}}^a\|_F + \|\widehat{\mathbf{\Delta}}^b\|_F + \|\widehat{\mathbf{\Delta}}^d\|_F\Big)^2 \\
&\le r + s_a + s_b + c(\eta_n/\tau)^2\|\widehat{\mathbf{\Delta}}\|_F^2.
\end{aligned} \tag{A.18}$$

Plugging (A.18) into (A.16) yields

$$\|\widehat{\mathbf{\Delta}}\|_F \le c\lambda_{\max}\eta_n\Big(r + s_a + s_b + c(\eta_n/\tau)^2\|\widehat{\mathbf{\Delta}}\|_F^2\Big)^{1/2}.$$

Thus solving for $\|\widehat{\mathbf{\Delta}}\|_F$ gives

$$\|\widehat{\mathbf{\Delta}}\|_F \leq \frac{c(r+s_a+s_b)^{1/2}\lambda_{\max}\eta_n}{\{1-c\lambda_{\max}^2(\eta_n^2/\tau)^2\}^{1/2}}, \tag{A.19}$$

which together with Theorem 1 results in the first inequality in Theorem 2.

Plugging (A.19) into (A.12), we deduce

$$\|\widehat{\mathbf{\Delta}}^a\|_F + \|\widehat{\mathbf{\Delta}}^b\|_F + \|\widehat{\mathbf{\Delta}}^d\|_F \leq \frac{c(r+s_a+s_b)^{1/2}\lambda_{\max}\eta_n^2}{\{1-c\lambda_{\max}^2(\eta_n^2/\tau)^2\}^{1/2}},$$

which along with (A.13) entails the second inequality in Theorem 2. Note that plugging (A.19) into (A.18) and combining terms yield

$$\begin{aligned}\|\widehat{\mathbf{\Delta}}^d\|_0 + \|\widehat{\mathbf{\Delta}}^a\|_0 + \|\widehat{\mathbf{\Delta}}^b\|_0 &\leq (r+s_a+s_b)\left[1+\frac{c\lambda_{\max}^2(\eta_n^2/\tau)^2}{1-c\lambda_{\max}^2(\eta_n^2/\tau)^2}\right] \\ &= (r+s_a+s_b)[1+o(1)],\end{aligned}$$

which gives the third inequality in Theorem 2.

We now plug the above inequality and (A.19) into (A.15). Then it holds that

$$\|\widehat{\mathbf{\Delta}}^a\|_1 + \|\widehat{\mathbf{\Delta}}^d\|_1 + \|\widehat{\mathbf{\Delta}}^b\|_1 \leq \frac{c(r+s_a+s_b)\lambda_{\max}\eta_n^2}{1-c\lambda_{\max}^2(\eta_n^2/\tau)^2}, \tag{A.20}$$

which yileds the fourth inequality in Theorem 2. Finally, it follows from Lemma 1 and (A.20) that

$$n^{-1}\|\mathbf{X}\widehat{\mathbf{\Delta}}\|_F^2 \leq \frac{c(r+s_a+s_b)\lambda_{\max}^2\eta_n^2}{1-c\lambda_{\max}^2(\eta_n^2/\tau)^2},$$

which establishes the fifth inequality in the theorem and concludes the proof of Theorem 2.

A.3 Lemma 1 and its proof

Lemma 1. *Under the conditions of Theorem 2, with at least probability as specified in (15) we have*

$$n^{-1}\|\mathbf{X}\widehat{\mathbf{\Delta}}\|_F^2 \leq c\lambda_{\max}\left(\|\widehat{\mathbf{\Delta}}^d\|_1 + \|\widehat{\mathbf{\Delta}}^a\|_1 + \|\widehat{\mathbf{\Delta}}^b\|_1\right),$$

where c is some positive constant.

Proof of Lemma 1. Denote by $\mathcal{E}_2$ the event on which inequalities (A.25)–(A.27) hold. Then by Lemma 2 in Section A.4, we see that event $\mathcal{E}_2$ holds with probability bound as specified in (15). We will prove Lemma 1 by conditioning on event $\mathcal{E}_2$. Since the SOAR estimator is the minimizer in the

neighborhood $\mathcal{P}_n$ defined in (14), it holds that

$$
\begin{aligned}
&(2n)^{-1}\|\mathbf{Y}-\mathbf{X}\widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^T\|_F^2+\lambda_d\|\widehat{\mathbf{D}}\|_1+\lambda_a\rho_a(\widehat{\mathbf{A}})+\lambda_b\rho_b(\widehat{\mathbf{B}})\\
&\leq (2n)^{-1}\|\mathbf{Y}-\mathbf{X}\mathbf{U}^*\mathbf{D}^*(\mathbf{V}^*)^T\|_F^2+\lambda_d\|\mathbf{D}^*\|_1+\lambda_a\rho_a(\mathbf{A}*)+\lambda_b\rho_b(\mathbf{B}^*).
\end{aligned}
$$

Let $\widehat{\mathbf{\Delta}}=\widehat{\mathbf{C}}-\mathbf{C}^*$. Rearranging terms in the above inequality leads to

$$
\begin{aligned}
(2n)^{-1}\|\mathbf{X}\widehat{\mathbf{\Delta}}\|_F^2 &\leq \langle n^{-1}\mathbf{X}^T\mathbf{E},\widehat{\mathbf{\Delta}}\rangle\\
&+\lambda_d\left(\|\mathbf{D}^*\|_1-\|\widehat{\mathbf{D}}\|_1\right)+\lambda_a\left(\rho_a(\mathbf{A}^*)-\rho_a(\widehat{\mathbf{A}})\right)+\lambda_b\left(\rho_b(\mathbf{B}^*)-\rho_b(\widehat{\mathbf{B}})\right).
\end{aligned}
\tag{A.21}
$$

By the definition of $\mathbf{D}^-$, the estimation error can be decomposed as

$$
\begin{aligned}
\widehat{\mathbf{\Delta}}&\equiv\widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^T-\mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*T}=\widehat{\mathbf{A}}\widehat{\mathbf{D}}^-\widehat{\mathbf{B}}^T-\mathbf{A}^*\mathbf{D}^{*-}\mathbf{B}^{*T}\\
&=\widehat{\mathbf{\Delta}}^a(\widehat{\mathbf{B}}\widehat{\mathbf{D}}^-)^T-\mathbf{U}^*\widehat{\mathbf{\Delta}}^d(\widehat{\mathbf{B}}\widehat{\mathbf{D}}^-)^T+\mathbf{U}^*(\widehat{\mathbf{\Delta}}^b)^T.
\end{aligned}
$$

The above decomposition together with Hölder's inequality entails that the following inequality

$$
\begin{aligned}
&\langle n^{-1}\mathbf{X}^T\mathbf{E},\widehat{\mathbf{\Delta}}\rangle\\
&=\langle n^{-1}\mathbf{X}^T\mathbf{E}\widehat{\mathbf{B}}\widehat{\mathbf{D}}^-,\widehat{\mathbf{\Delta}}^a\rangle-\langle n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E}\widehat{\mathbf{B}}\widehat{\mathbf{D}}^-,\widehat{\mathbf{\Delta}}^d\rangle+\langle n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E},\widehat{\mathbf{\Delta}}^{bT}\rangle\\
&\leq\|n^{-1}\mathbf{X}^T\mathbf{E}\widehat{\mathbf{B}}\widehat{\mathbf{D}}^-\|_\infty\|\widehat{\mathbf{\Delta}}^a\|_1+\|n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E}\widehat{\mathbf{B}}\widehat{\mathbf{D}}^-\|_\infty\|\widehat{\mathbf{\Delta}}^d\|_1+\|n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E}\|_\infty\|\widehat{\mathbf{\Delta}}^b\|_1\\
&\leq\lambda_a\|\widehat{\mathbf{\Delta}}^a\|_1+\lambda_d\|\widehat{\mathbf{\Delta}}^d\|_1+\lambda_b\|\widehat{\mathbf{\Delta}}^b\|_1
\end{aligned}
\tag{A.22}
$$

holds on event $\mathcal{E}_2$.

By the triangle inequality for the ℓ_1-norm and Condition 4, we deduce

$$
\begin{aligned}
&\lambda_d\left(\|\mathbf{D}^*\|_1-\|\widehat{\mathbf{D}}\|_1\right)+\lambda_a\left(\rho_a(\mathbf{A}^*)-\rho_a(\widehat{\mathbf{A}})\right)+\lambda_b\left(\rho_b(\mathbf{B}^*)-\rho_b(\widehat{\mathbf{B}})\right)\\
&\leq\lambda_d\|\widehat{\mathbf{\Delta}}^d\|_1+\lambda_a\|\widehat{\mathbf{\Delta}}^a\|_1+\lambda_b\|\widehat{\mathbf{\Delta}}^b\|_1.
\end{aligned}
\tag{A.23}
$$

Thus plugging (A.22) and (A.23) into (A.21) yields

$$
\begin{aligned}
(cn)^{-1}\|\mathbf{X}\widehat{\mathbf{\Delta}}\|_F^2&\leq\lambda_d\|\widehat{\mathbf{\Delta}}^d\|_1+\lambda_a\|\widehat{\mathbf{\Delta}}^a\|_1+\lambda_b\|\widehat{\mathbf{\Delta}}^b\|_1\\
&\leq\lambda_{\max}\left(\|\widehat{\mathbf{\Delta}}^d\|_1+\|\widehat{\mathbf{\Delta}}^a\|_1+\|\widehat{\mathbf{\Delta}}^b\|_1\right)
\end{aligned}
\tag{A.24}
$$

with $\lambda_{\max}=\max(\lambda_d,\lambda_a,\lambda_b)$, which completes the proof of Lemma 1.

A.4 Lemma 2 and its proof

Lemma 2. *Under the conditions of Theorem 2, with at least probability as specified in (15) the following inequalities hold*

$$\sup_{(\mathbf{B},\mathbf{D})\in\mathcal{P}_n} \|n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty \le \lambda_d, \tag{A.25}$$

$$\sup_{(\mathbf{B},\mathbf{D})\in\mathcal{P}_n} \|n^{-1}\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty \le \lambda_a, \tag{A.26}$$

$$\sup_{(\mathbf{B},\mathbf{D})\in\mathcal{P}_n} \|n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E}\|_\infty \le \lambda_b. \tag{A.27}$$

Proof of Lemma 2. Recall that $\widetilde{\mathcal{P}}_n = \{\mathbf{C} : \|\mathbf{C}-\widetilde{\mathbf{C}}\|_F \le 2R_n\}$, where $\widetilde{\mathbf{C}}$ is the initial Lasso estimator and $R_n = c(n^{-1}s\log(pq))^{1/2}$ is as defined in Theorem 1. It follows from Theorem 1 that the true regression coefficient matrix $\mathbf{C}^*$ falls in the neighborhood $\widetilde{\mathcal{P}}_n$ with probability at least $1-2(pq)^{1-c_0^2/2}$, where $c_0 > \sqrt{2}$ is some constant given in Theorem 1. Note that the neighborhood $\widetilde{\mathcal{P}}_n$ shrinks asymptotically as $n\to\infty$ since $R_n^2 = O(n^{\alpha+\beta/2+\gamma-1})$ and $\alpha+\beta/2+\gamma < \alpha+\beta+\gamma < 1$ holds under our assumptions. In order to deal with the nonconvexity of the objective function, we exploit the framework of convexity-assisted nonconvex optimization (CANO) and solve the SOFAR optimization problem in the shrinking local region $\mathcal{P}_n = \widetilde{\mathcal{P}}_n \cap (\mathcal{C}\times\mathcal{D}\times\mathcal{A}\times\mathcal{B})$ as defined in (14).

Observe that for any $\mathbf{C}\in\widetilde{\mathcal{P}}_n$, by the triangle inequality it holds that

$$\|\mathbf{C}-\mathbf{C}^*\|_F \le \|\mathbf{C}-\widetilde{\mathbf{C}}\|_F + \|\widetilde{\mathbf{C}}-\mathbf{C}^*\|_F \le 3R_n;$$

that is, with probability at least $1-2(pq)^{1-c_0^2/2}$, $\widetilde{\mathcal{P}}_n \subset \{\mathbf{C} : \|\mathbf{C}-\mathbf{C}^*\|_F \le 3R_n\}$. Further, by Lemma 3 we have $\{\mathbf{C} : \|\mathbf{C}-\mathbf{C}^*\|_F \le 3R_n\} \subset \mathcal{E}_1$, where

$$\begin{aligned}\mathcal{E}_1 = \{\mathbf{C}\equiv\mathbf{A}\mathbf{D}^-\mathbf{B} : \|\mathbf{D}-\mathbf{D}^*\|_F \le 3R_n,\\ \|\mathbf{A}-\mathbf{A}^*\|_F + \|\mathbf{B}-\mathbf{B}^*\|_F \le 3c\eta_n R_n\}\end{aligned} \tag{A.28}$$

with $c > 0$ some constant. Combining the above results yields that with probability at least $1-2(pq)^{1-c_0^2/2}$, $\mathcal{P}_n \subset \widetilde{\mathcal{P}}_n \subset \mathcal{E}_1$, which entails

$$P\Big(\mathcal{P}_n \not\subset \mathcal{E}_1\Big) \le 2(pq)^{1-c_0^2/2}. \tag{A.29}$$

We next establish that (A.25)–(A.27) hold with asymptotic probability one. Note that it follows from

the definition of conditional probability and (A.29) that

$$
\begin{aligned}
&P\Big(\sup_{\mathbf{C}\in\mathcal{P}_n}\|n^{-1}\mathbf{U}^*\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty>\lambda_d\Big)\\
&\quad\le P\Big(\sup_{\mathbf{C}\in\mathcal{P}_n}\|n^{-1}\mathbf{U}^*\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty>\lambda_d\Big|\mathcal{P}_n\subset\mathcal{E}_1\Big)+P\Big(\mathcal{P}_n\not\subset\mathcal{E}_1\Big)\\
&\quad\le P\Big(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{U}^*\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty>\lambda_d)+2(pq)^{1-c_0^2/2}.
\end{aligned}
$$

Thus to prove (A.25), we only need to show that

$$
\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{U}^*\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty\le\lambda_d \tag{A.30}
$$

holds with asymptotic probability one. Similarly, to show (A.26) and (A.27) we only need to prove that

$$
\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty\le\lambda_a, \tag{A.31}
$$

$$
\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E}\|_\infty\le\lambda_b \tag{A.32}
$$

hold with asymptotic probability one. We next proceed to prove (A.30)–(A.32) hold with asymptotic probability one.

Denote by $\mathbf{x}_i$ and $\mathbf{e}_j$ the ith and jth columns of $\mathbf{X}\in\mathbb{R}^{n\times p}$ and $\mathbf{E}\in\mathbb{R}^{n\times q}$, respectively. Let $\mathbf{x}_i^*$ and $\mathbf{e}_j^*$ be the ith and jth columns of $\mathbf{X}^*\equiv\mathbf{X}\mathbf{U}^*\in\mathbb{R}^{n\times q}$ and $\mathbf{E}^*\equiv\mathbf{E}\mathbf{V}^*\in\mathbb{R}^{n\times q}$, respectively. It is seen that the last $q-r$ columns of $\mathbf{X}^*$ and $\mathbf{E}^*$ are all zero. First, we show that (A.30) holds with significant probability. The decomposition

$$
\mathbf{B}\mathbf{D}^-=\mathbf{V}^*+\mathbf{\Delta}^b\mathbf{D}^-+\mathbf{V}^*\mathbf{D}^*\mathbf{\Delta}^{d-}
$$

and the triangle inequality lead to

$$
\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty\le\|n^{-1}\mathbf{X}^{*T}\mathbf{E}^*\|_\infty+\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{\Delta}^b\mathbf{D}^-\|_\infty+\|n^{-1}\mathbf{X}^{*T}\mathbf{E}^*\mathbf{D}^*\mathbf{\Delta}^{d-}\|_\infty,
$$

where $\mathbf{\Delta}^{d-}=\mathbf{D}^--\mathbf{D}^{*-}=\mathrm{diag}\{d_j^{-1}-(d_j^*)^{-1}\}$. Thus it holds that

$$
\begin{aligned}
&P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty\ge\lambda_d\right)\le P\left(\|n^{-1}\mathbf{X}^{*T}\mathbf{E}^*\|_\infty\ge\lambda_d/3\right)\\
&+P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{\Delta}^b\mathbf{D}^-\|_\infty\ge\lambda_d/3\right)+P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}^*\mathbf{D}^*\mathbf{\Delta}^{d-})\|_\infty\ge\lambda_d/3\right).
\end{aligned}
\tag{A.33}
$$

Let us consider the first term on the right hand side of (A.33). Since $\mathbf{E}\sim N(\mathbf{0},\mathbf{I}_n\otimes\mathbf{\Sigma})$ by Condition 3, the jth column vector of $\mathbf{E}^*$, $\mathbf{e}_j^*=\mathbf{E}\mathbf{v}_j^*$ with $\mathbf{v}_j^*$ the jth column vector of $\mathbf{V}^*$, is distributed as

$N\left(0, \mathbf{v}_j^{*T}\mathbf{\Sigma}\mathbf{v}_j^* I_n\right)$. Furthermore, note that $\|\mathbf{X}^{*T}\mathbf{E}^*\|_\infty = \max_{1\le i\le q}\max_{1\le j\le q}|\mathbf{x}_i^{*T}\mathbf{e}_j^*|$ and

$$\mathbb{E}|\mathbf{x}_i^{*T}\mathbf{e}_j^*|^2 = \mathbf{v}_j^{*T}\mathbf{\Sigma}\mathbf{v}_j^*\mathbf{x}_i^{*T}\mathbf{x}_i^* \le \alpha_{\max}\mathbf{u}_i^{*T}\mathbf{X}^T\mathbf{X}\mathbf{u}_i^* \le \alpha_{\max}c_3 n \le cn, \tag{A.34}$$

where $\alpha_{\max}$ denotes the maximum eigenvalue of $\mathbf{\Sigma}$ and the second inequality follows from Condition 2 and the fact that $\mathbf{u}_i^* = \mathbf{0}$ for $i = r+1, \cdots, q$. Therefore, it follows from Bonferroni's inequality, the Gaussianity of $\mathbf{e}_j^*$, and (A.34) that for $\lambda_d^2 = c_1^2 n^{-1}\log(pr)$,

$$\begin{aligned} P\left(n^{-1}\|\mathbf{X}^{*T}\mathbf{E}^*\|_\infty \ge \lambda_d/3\right) &\le \sum_{i=1}^r\sum_{j=1}^r P\left(n^{-1}|\mathbf{x}_i^{*T}\mathbf{e}_j^*| \ge \lambda_d/3\right) \\ &\le 2\sum_{i=1}^r\sum_{j=1}^r \exp\left(-\frac{n^2\lambda_d^2/9}{2\mathbb{E}|\mathbf{x}_i^{*T}\mathbf{e}_j^*|^2}\right) \\ &\le 2r^2\exp\left(-\frac{n^2c_1^2n^{-1}\log(pr)}{18cn}\right) \\ &= 2r^2(pr)^{-c_1^2/c}. \end{aligned} \tag{A.35}$$

We now consider the second term on the right hand side of (A.33). Some algebra gives

$$\begin{aligned} \|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{\Delta}^b\mathbf{D}^-\|_\infty &= \|n^{-1}(\mathbf{I}_q\otimes\mathbf{X}^{*T}\mathbf{E})\operatorname{vec}(\mathbf{\Delta}^b\mathbf{D}^-)\|_\infty \\ &\le \max_{1\le i\le r}\sum_{j=1}^q |n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j|\,\|\operatorname{vec}(\mathbf{\Delta}^b\mathbf{D}^-)\|_\infty \\ &\le q\max_{1\le i\le r}\max_{1\le j\le q}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j|\,\|(\mathbf{D}^-\otimes\mathbf{I}_q)\operatorname{vec}(\mathbf{\Delta}^b)\|_\infty \\ &\le q\|\mathbf{D}^-\|_\infty\max_{1\le i\le r}\max_{1\le j\le q}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j|\,\|\operatorname{vec}(\mathbf{\Delta}^b)\|_\infty. \end{aligned}$$

Since we solve SOFAR in the local neighborhood $\mathcal{P}_n$ defined in (14), by Condition 1 we have $\|\mathbf{D}^-\|_\infty \le \tau^{-1}$ for any $\mathbf{C} \equiv \mathbf{A}\mathbf{D}^-\mathbf{B} \in \mathcal{P}_n$. Thus by (A.28), the second term in the upper bound of (A.33) can be bounded as

$$\begin{aligned} \sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{\Delta}^b\mathbf{D}^-\|_\infty &\le (q/\tau)\max_{1\le i\le r}\max_{1\le j\le q}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j|\sup_{\mathcal{E}_1}\|\operatorname{vec}(\mathbf{\Delta}^b)\|_\infty \\ &\le (q/\tau)\max_{1\le i\le r}\max_{1\le j\le q}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j|\sup_{\mathcal{E}_1}\|\mathbf{\Delta}^b\|_F \\ &\le 3c(q/\tau)\eta_n R_n\max_{1\le i\le r}\max_{1\le j\le q}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j|. \end{aligned} \tag{A.36}$$

Similarly to (A.34), we can show that

$$\mathbb{E}|\mathbf{x}_i^{*T}\mathbf{e}_j|^2 \le \sigma_j^2 c_3 n \le \sigma_{\max}^2 c_3 n \le cn. \tag{A.37}$$

Therefore, in view of (A.36), (A.37), $R_n^2 = O(sn^{-1}\log(pq))$, and $p \ge q$, the same inequality as (A.35)

results in

$$
\begin{aligned}
&P\left(\sup_{\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\boldsymbol{\Delta}^{b}\mathbf{D}^{-}\|_\infty \geq \lambda_d/3\right)\\
&= P\left(3c(q/\tau)\eta_n R_n \max_{1\leq i\leq r}\max_{1\leq j\leq q}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j| \geq \lambda_d/3\right)\\
&\leq 2\sum_{i=1}^{r}\sum_{j=1}^{q}\exp\left(-\frac{n^2\lambda_d^2}{81c(q/\tau)^2\eta_n^2 R_n^2\mathbb{E}|\mathbf{x}_i^{*T}\mathbf{e}_j|^2}\right)\\
&= 2qr\exp\left(-\frac{c_1^2 n}{c(q/\tau)^2\eta_n^2 s}\right),
\end{aligned}
\tag{A.38}
$$

where c is some positive constant.

It remains to investigate the third term on the right hand side of (A.33). Since $\mathbf{D}^*\boldsymbol{\Delta}^{d-}$ is a diagonal matrix whose (k,k)th entry is given by $(d_k^* - d_k)/d_k$ with $\mathrm{rank}(\mathbf{D}^*\boldsymbol{\Delta}^{d-}) \leq r$, the last $q-r$ columns of both $\mathbf{X}^*$ and $\mathbf{E}^*$ are zero, and $\mathbf{D}\in\mathcal{D}$, we have

$$
\begin{aligned}
\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}^*\mathbf{D}^*\boldsymbol{\Delta}^{d-}\|_\infty &\leq \max_{1\leq i\leq r}\max_{1\leq j\leq r}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j^*|\sup_{\mathcal{E}}\|\mathbf{D}^*\boldsymbol{\Delta}^{d-}\|_\infty\\
&\leq \tau^{-1}\max_{1\leq i\leq r}\max_{1\leq j\leq r}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j^*|\max_{1\leq k\leq r}|d_k^* - d_k|\\
&\leq \tau^{-1}\max_{1\leq i\leq r}\max_{1\leq j\leq r}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j^*|\|\boldsymbol{\Delta}^{d}\|_F\\
&\leq 3(R_n/\tau)\max_{1\leq i\leq r}\max_{1\leq j\leq r}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j^*|.
\end{aligned}
\tag{A.39}
$$

Then by (A.34) and (A.39), the same inequality yields

$$
\begin{aligned}
P\left(\sup\|n^{-1}\mathbf{X}^{*T}\mathbf{E}^*\mathbf{D}^*\boldsymbol{\Delta}^{d-}\|_\infty \geq \lambda_d/3\right) &\leq P\left(3(R_n/\tau)\max_{1\leq i\leq r}\max_{1\leq j\leq r}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j^*| \geq \lambda_d/3\right)\\
&\leq 2\sum_{i=1}^{r}\sum_{j=1}^{r}\exp\left(-\frac{n^2\lambda_d^2}{81c(R_n/\tau)^2\mathbb{E}|\mathbf{x}_i^{*T}\mathbf{e}_j^*|^2}\right)\\
&\leq 2r^2\exp\left(-\frac{c_1^2 n^2 n^{-1}\log(pr)}{csn^{-1}\log(pq)\tau^{-2}n}\right)\\
&\leq 2r^2\exp\left(-\frac{c_1^2\tau^2 n}{cs}\right).
\end{aligned}
\tag{A.40}
$$

Therefore, combining (A.35), (A.38), and (A.40) with (A.33) gives the probability bound

$$
\begin{aligned}
&P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{B}\mathbf{D}^{-}\|_\infty \geq \lambda_d\right)\\
&\leq 2r^2(pr)^{-c_1^2/c} + 2rq\exp\left(-\frac{c_1^2 n}{c(q/\tau)^2\eta_n^2 s}\right) + 2r^2\exp\left(-\frac{c_1^2\tau^2 n}{cs}\right).
\end{aligned}
\tag{A.41}
$$

We next prove that (A.31) holds with high probability. The arguments are similar to those for proving (A.30) except that $\mathbf{X}^*$ is replaced with $\mathbf{X}$ in the proof of (A.25). More specifically, note that we have

the following decomposition of probability bound

$$P\left(\sup_{\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty \geq \lambda_a\right) \leq P\left(\|n^{-1}\mathbf{X}^T\mathbf{E}^*\|_\infty \geq \lambda_a/3\right) \tag{A.42}$$
$$+ P\left(\sup_{\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}\boldsymbol{\Delta}^b\mathbf{D}^-\|_\infty \geq \lambda_a/3\right) + P\left(\sup_{\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}^*\mathbf{D}^*\boldsymbol{\Delta}^{d-})\|_\infty \geq \lambda_a/3\right).$$

Thus, it suffices to bound the probabilities on the right hand side of (A.42). Let us consider the first term. Observe that

$$\mathbb{E}|\mathbf{x}_i^T\mathbf{e}_j^*|^2 \leq \alpha_{\max}\mathbf{x}_i^T\mathbf{x}_i = \alpha_{\max}n \leq cn,$$

where c is some positive constant. Thus, setting $\lambda_a^2 = c_1^2 n^{-1}\log(pr)$ and noting that $\mathbf{E}^*$ has only r nonzero columns lead to the bound

$$\begin{aligned} P\left(n^{-1}\|\mathbf{X}^T\mathbf{E}^*\|_\infty \geq \lambda_a/3\right) &\leq 2\sum_{i=1}^{p}\sum_{j=1}^{r}\exp\left(-\frac{n^2\lambda_a^2}{8\mathbb{E}|\mathbf{x}_i^T\mathbf{e}_j^*|^2}\right) \\ &\leq 2pr\exp\left(-\frac{c_1^2 n^2 n^{-1}\log(pr)}{cn}\right) \\ &\leq 2(pr)^{1-c_1^2/c}. \end{aligned} \tag{A.43}$$

We next consider the second probability bound on the right hand side of (A.42). Since

$$\mathbb{E}|\mathbf{x}_i^T\mathbf{e}_j|^2 \leq \sigma_{\max}^2\mathbf{x}_i^T\mathbf{x}_i = \sigma_{\max}^2 n \leq cn,$$

by replacing $\max_{1\leq i\leq r}$ in (A.36) and (A.38) with $\max_{1\leq i\leq p}$ we deduce

$$P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}\boldsymbol{\Delta}^b\mathbf{D}^-\|_\infty \geq \lambda_a/3\right) \leq 2pr\exp\left(-\frac{c_1^2 n}{c(q/\tau)^2\eta_n^2 s}\right). \tag{A.44}$$

It remains to study the third probability bound on the right hand side of (A.42). Similarly, replacing $\max_{1\leq i\leq r}$ in (A.39) and (A.40) with $\max_{1\leq i\leq p}$ yields

$$P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}^*\mathbf{D}^*\boldsymbol{\Delta}^{d-}\|_\infty \geq \lambda_a/3\right) \leq 2pr\exp\left(-\frac{c_1^2\tau^2 n}{cs}\right). \tag{A.45}$$

Thus combining (A.43)–(A.45), we can bound (A.42) as

$$\begin{aligned} &P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty \geq \lambda_a\right) \\ &\leq 2(pr)^{1-c_1^2/c} + 2pr\exp\left(-\frac{c_1^2 n}{c(q/\tau)^2\eta_n^2 s}\right) + 2pr\exp\left(-\frac{c_1^2\tau^2 n}{cs}\right). \end{aligned} \tag{A.46}$$

Finally, we show that condition (A.32) holds with large probability. Choosing $\lambda_b^2 = c_1^2 n^{-1}\log(pr)$

results in

$$P\left(n^{-1}\|\mathbf{X}^{*T}\mathbf{E}\|_\infty \geq \lambda_b\right) \leq 2\sum_{i=1}^{r}\sum_{j=1}^{q}\exp\left(-\frac{n^2\lambda_b^2}{2\mathbb{E}|\mathbf{x}_i^{*T}\mathbf{e}_j|^2}\right)$$
$$\leq 2qr\exp\left(-\frac{c_1^2 n^2 n^{-1}\log(pr)}{cn}\right)$$
$$\leq 2qr(pr)^{-c_1^2/c}. \tag{A.47}$$

Consequently, for the given set of regularization parameters $(\lambda_d, \lambda_a, \lambda_b)$ it follows from (A.41), (A.46), and (A.47) that conditions (A.30)–(A.32) hold simultaneously with probability at least

$$1 - \left\{2(pr)^{1-c_1^2/c} + 2pr\exp\left(-\frac{c_1^2 n}{c(q/\tau)^2\eta_n^2 s}\right)\right\},$$

where we have used the facts of $c_1^2 > c$ and $p \geq q \geq 1$. Moreover, to check that the probability bound converges to one, since $c_1^2 > c$ it is sufficient to show that

$$2pr\exp\left(-\frac{c_1^2 n}{c(q/\tau)^2\eta_n^2 s}\right)$$

converges to zero. This follows immediately from the assumptions of $\log p = O(n^\alpha)$, $q = O(n^{\beta/2})$, $s = O(n^\gamma)$, and $\eta_n/\tau = o(n^{(1-\alpha-\beta-\gamma)/2})$, which concludes the proof of Lemma 2.

A.5 Proof of Theorem 3

Recall that the theoretical results for the SOFAR estimator established in the paper hold simultaneously over the set of all local minimizers in a neighborhood of the initial Lasso estimator. Thus we aim to establish the convergence of the SOFAR algorithm when supplied the initial Lasso estimator. Note that the equivalent form of the SOFAR problem (21) with the slack variables $\mathbf{A}$ and $\mathbf{B}$ can be solved using the augmented Lagrangian form with sufficiently large penalty parameter $\mu > 0$. From now on, we fix parameter μ and the set of Lagrangian multipliers $\mathbf{\Gamma}$, and thus work with the objective function $L_\mu(\mathbf{\Theta}, \mathbf{\Omega}; \mathbf{\Gamma})$.

By the nature of the block coordinate descent algorithm applied to $(\mathbf{U}, \mathbf{V}, \mathbf{D}, \mathbf{A}, \mathbf{B})$, the sequence $(L_\mu(\cdot))$ of values of the objective function $L_\mu(\mathbf{\Theta}, \mathbf{\Omega}; \mathbf{\Gamma})$ is decreasing. Clearly the function $L_\mu(\mathbf{\Theta}, \mathbf{\Omega}; \mathbf{\Gamma})$ is bounded from below. Thus the sequence $(L_\mu(\cdot))$ converges. Since the rank parameter m is fixed in the SOFAR algorithm, we assume for simplicity that the diagonal matrix $\mathbf{D}^k$ of singular values has all the diagonal entries bounded away from zero, since otherwise we can solve the SOFAR problem with a smaller rank m.

By assumption, we have

$$\sum_{k=1}^{\infty}[\Delta L_\mu(\mathbf{U}^k)]^{1/2} < \infty,\ \sum_{k=1}^{\infty}[\Delta L_\mu(\mathbf{V}^k)]^{1/2} < \infty,\ \text{ and }\ \sum_{k=1}^{\infty}[\Delta L_\mu(\mathbf{D}^k)]^{1/2} < \infty,$$

where $\Delta L_\mu(\cdot)$ stands for the decrease in $L_\mu(\cdot)$ by a block update. Note that the $\mathbf{U}$-space with constraint $\mathbf{U}^T\mathbf{U} = \mathbf{I}_m$ is a Stiefel manifold which is compact and smooth; see, e.g., [47] for a brief review of the geometry of Stiefel manifold. Since the $\mathbf{D}$-sequence is always positive definite by assumption, the objective function along the $\mathbf{U}$-block with all the other four blocks fixed is convex and has positive curvature bounded away from zero along any direction in the $\mathbf{U}$-space. By definition, $\mathbf{U}^k$ is the minimizer of such a restricted objective function, which entails that the gradient of this function at $\mathbf{U}^k$ on the Stiefel manifold vanishes. Thus it follows easily from the mean value theorem and the fact of positive curvature that $\Delta L_\mu(\mathbf{U}^k)$ is bounded from below by some positive constant δ times $d_g^2(\mathbf{U}^k, \mathbf{U}^{k-1})$, where $d_g(\cdot,\cdot)$ denotes the distance function on the Stiefel manifold. Then it holds that

$$\sum_{k=1}^{\infty} d_g(\mathbf{U}^k, \mathbf{U}^{k-1}) \leq \delta^{-1/2} \sum_{k=1}^{\infty} [\Delta L_\mu(\mathbf{U}^k)]^{1/2} < \infty,$$

which along with the triangle inequality entails that $(\mathbf{U}^k)$ is a Cauchy sequence on the Stiefel manifold. Therefore, the sequence $(\mathbf{U}^k)$ converges to a limit point $\mathbf{U}_*$ on the Stiefel manifold which is a local solution along the $\mathbf{U}$-block. Similarly, we can show that the sequence $(\mathbf{V}^k)$ also converges to a limit point $\mathbf{V}_*$ on the Stiefel manifold that is a local solution along the $\mathbf{V}$-block.

Recall that the diagonal matrix $\mathbf{D}^k$ of singular values is assumed to have all the diagonal entries bounded away from zero. Since we have shown that the sequences $(\mathbf{U}^k)$ and $(\mathbf{V}^k)$ converge to limit points $\mathbf{U}_*$ and $\mathbf{V}_*$ on the Stiefel manifolds, respectively, it follows from the fact that both $\mathbf{U}_*$ and $\mathbf{V}_*$ have full column rank m that as k becomes large, the objective function along the $\mathbf{D}$-block with all the other four blocks fixed is convex and has positive curvature bounded away from zero. Thus an application of similar arguments as above yields that the sequence $(\mathbf{D}^k)$ also converges to a limit point $\mathbf{D}_*$.

With the established convergence results of the sequences $(\mathbf{U}^k)$, $(\mathbf{V}^k)$, and $(\mathbf{D}^k)$, the convergence of the sequences $(\mathbf{A}^k)$ and $(\mathbf{B}^k)$ follows easily from the convergence property of the block coordinate descent algorithm applied to separable convex problems [61], by noting that the objective function with $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{D}$ replaced by their limit points is jointly convex in $\mathbf{A}$ and $\mathbf{B}$ since the penalty functions $\rho_a(\cdot)$ and $\rho_b(\cdot)$ are assumed to be convex. This completes the proof of Theorem 3.

B Additional technical details

B.1 Lemma 3 and its proof

Lemma 3. *Under Condition 5, we have for any matrix* $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ *and* $\mathbf{C}^* = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*T}$ *with* $\|\mathbf{C} - \mathbf{C}^*\|_2 \leq d_1^*$ *that*

$$\|\mathbf{D} - \mathbf{D}^*\|_F \leq \|\mathbf{C} - \mathbf{C}^*\|_F,$$
$$\|\mathbf{A} - \mathbf{A}^*\|_F + \|\mathbf{B} - \mathbf{B}^*\|_F \leq c\eta_n \|\mathbf{C} - \mathbf{C}^*\|_F,$$

where $\eta_n = 1 + \delta^{-1/2}\big(\sum_{j=1}^{r}(d_1^*/d_j^*)^2\big)^{1/2}$ *and* $c > 0$ *is some constant.*

Proof of Lemma 3. It is well known that the inequality

$$\|\mathbf{D} - \mathbf{D}^*\|_F \le \|\mathbf{C} - \mathbf{C}^*\|_F$$

holds; see, for example, [51]. It remains to show the second desired inequality. Recall that $\mathbf{A}^* = \mathbf{U}^*\mathbf{D}^*$. By the decomposition

$$\mathbf{C} - \mathbf{C}^* = (\mathbf{A} - \mathbf{A}^*)\mathbf{V}^T + \mathbf{A}^*(\mathbf{V} - \mathbf{V}^*)^T$$

and the unitary property of the Frobenius norm, we have

$$\|\mathbf{A} - \mathbf{A}^*\|_F \le \|\mathbf{C} - \mathbf{C}^*\|_F + \|\mathbf{D}^*(\mathbf{V} - \mathbf{V}^*)^T\|_F. \tag{A.48}$$

Let us examine the second term on the right hand side of (A.48). To do so, we apply Theorem 3 of [65] to $\mathbf{V} - \mathbf{V}^*$ columnwise to avoid the identifiability issue. When $r = 1$ or 2, it holds that

$$\|\mathbf{v}_1 - \mathbf{v}_1^*\|_2 \le \frac{cd_1^*\|\mathbf{C} - \mathbf{C}^*\|_F}{\delta^{1/2}(d_1^*)^2}, \quad \|\mathbf{v}_r - \mathbf{v}_r^*\|_2 \le \frac{cd_1^*\|\mathbf{C} - \mathbf{C}^*\|_F}{\delta^{1/2}(d_r^*)^2}. \tag{A.49}$$

When $r \ge 3$, in addition to (A.49) we have for $j = 2, \ldots, r-1$,

$$\|\mathbf{v}_j - \mathbf{v}_j^*\|_2 \le \frac{c(2d_1^* + \|\mathbf{C} - \mathbf{C}^*\|_2)\|\mathbf{C} - \mathbf{C}^*\|_F}{\min(d_{j-1}^{*2} - d_j^{*2}, d_j^{*2} - d_{j+1}^{*2})},$$

where $c > 0$ is some constant. Since Condition 5 gives $d_{j-1}^{*2} - d_j^{*2} \ge \delta^{1/2}(d_{j-1}^*)^2 \ge \delta^{1/2}(d_j^*)^2$, it follows from the assumption $\|\mathbf{C} - \mathbf{C}^*\|_2 \le d_1^*$ that the above inequality can be further bounded as

$$\|\mathbf{v}_j - \mathbf{v}_j^*\|_2 \le \frac{c(2d_1^* + \|\mathbf{C} - \mathbf{C}^*\|_2)\|\mathbf{C} - \mathbf{C}^*\|_F}{\min(d_{j-1}^{*2} - d_j^{*2}, d_j^{*2} - d_{j+1}^{*2})} \le \frac{cd_1^*\|\mathbf{C} - \mathbf{C}^*\|_F}{\delta^{1/2}(d_j^*)^2}.$$

Thus these inequalities entail that

$$\|\mathbf{D}^*(\mathbf{V} - \mathbf{V}^*)^T\|_F^2 = \sum_{j=1}^{r} d_j^{*2}\|\mathbf{v}_j - \mathbf{v}_j^*\|_2^2 \le (c/\delta)\|\mathbf{C} - \mathbf{C}^*\|_F^2 \sum_{j=1}^{r}(d_1^*/d_j^*)^2. \tag{A.50}$$

Consequently, combining (A.48) and (A.50) leads to the bound

$$\|\mathbf{A} - \mathbf{A}^*\|_F \le \|\mathbf{C} - \mathbf{C}^*\|_F + (c/\delta^{1/2})\|\mathbf{C} - \mathbf{C}^*\|_F \left\{\sum_{j=1}^{r}(d_1^*/d_j^*)^2\right\}^{1/2}.$$

On the other hand, the bound for $\|\mathbf{B} - \mathbf{B}^*\|_F$ can be obtained by the decomposition $\mathbf{C} - \mathbf{C}^* = \mathbf{U}(\mathbf{B} - \mathbf{B}^*)^T + (\mathbf{U} - \mathbf{U}^*)\mathbf{B}^{*T}$ and similar arguments. Therefore, adding both bounds together and enlarging

the positive constant c conclude the proof of Lemma 3.

B.2 Lemma 4 and its proof

Lemma 4. *Under Conditions 1 and 2, it holds for any* $\mathbf{C} \in \mathcal{C}$ *that*

$$n^{-1}\|\mathbf{X}(\mathbf{C} - \mathbf{C}^*)\|_F^2 \geq c_2\|\mathbf{C} - \mathbf{C}^*\|_F^2.$$

Proof of Lemma 4. Denote by $\mathbf{\Delta} = \mathbf{C} - \mathbf{C}^*$, $\mathbf{W} = \mathbf{I}_q \otimes \mathbf{X}$, and $\boldsymbol{\delta} = \mathrm{vec}(\mathbf{\Delta})$, where $\mathbf{I}_q$ is the $q \times q$ identity matrix. It follows from the triangle inequality and Condition 1 that

$$\begin{aligned}\|\boldsymbol{\delta}\|_0 &= \|\,\mathrm{vec}(\mathbf{C}) - \mathrm{vec}(\mathbf{C}^*)\|_0 \leq \|\,\mathrm{vec}(\mathbf{C})\|_0 + \|\,\mathrm{vec}(\mathbf{C}^*)\|_0 \\ &< \kappa_{c_2}/2 + \kappa_{c_2}/2 = \kappa_{c_2}.\end{aligned}$$

Note that the singular values of $\mathbf{W}$ are the same as those of the original design matrix $\mathbf{X}$ with the multiplicity of each singular value multiplied by q. This entails that the robust spark of $\mathbf{W}$ is equal to that of $\mathbf{X}$, which is κ_{c_2} for a given positive constant c_2. Thus by the definition of the robust spark, we obtain

$$n^{-1}\|\mathbf{X}\mathbf{\Delta}\|_F^2 = n^{-1}\|\mathbf{W}\boldsymbol{\delta}\|_2^2 = n^{-1}\|\mathbf{W}_{\mathrm{supp}(\boldsymbol{\delta})}\boldsymbol{\delta}_{\mathrm{supp}(\boldsymbol{\delta})}\|_2^2 \geq c_2\|\boldsymbol{\delta}\|_2^2 = c_2\|\mathbf{\Delta}\|_F^2,$$

where the subscript $\mathrm{supp}(\boldsymbol{\delta})$ denotes the restriction of the matrix to the corresponding columns or that of the vector to the corresponding components. This completes the proof of Lemma 4.