9-5-2017

On Sequential Estimation of Multivariate Associations

Aditya Mishra
University of Connecticut - Storrs, aditya.mishra@uconn.edu

On Sequential Estimation of Multivariate Associations

Aditya Kumar Mishra, Ph.D.
University of Connecticut, 2017

ABSTRACT

In multivariate analysis, one of the major problems of interest is to model multivariate responses using observed predictors. We often encounter responses of either continuous, binary or count types, or may be of mixed types. Multivariate linear regression (MLR) models the relationship in case of Gaussian outcomes. When outcomes are non-Gaussian or mixed types, i.e., generalized, a possible modeling approach proceeds via maximization of likelihood obtained after assuming conditionally independent observed outcomes are from exponential dispersion family. In high-dimensional setting, responses maybe interrelated and predictors maybe correlated or unimportant. Such dependency can be induced through a low-rank and sparse coefficient matrix which also facilitates model interpretation. Specifically, the structure translates into having co-sparse left and right singular vectors in the singular value decomposition of the coefficient matrix. In this thesis, we have proposed algorithms to recover such matrices. In MLR, we reformulate the problem as a supervised co-sparse factor analysis, and develop an efficient computational procedure, named *sequential factor extraction via co-sparse unit-rank estimation*

(SeCURE). A unit-step in SeCURE extracts a sparse and unit-rank coefficient matrix leading to co-sparsity in corresponding singular vectors. In the generalized setting, motivated by SeCURE, we propose a sequential procedure to recover the desired coefficient matrices, named as *generalized sequential factor extraction via co-sparse unit-rank estimation* (GSeCURE). Because of the complicated likelihood structure, a unit-step of GSeCURE estimates co-sparse singular vectors via iteratively optimizing a surrogate of the objective function. Efficacy of both SeCURE and GSeCURE are demonstrated by simulation studies and various applications.

On Sequential Estimation of Multivariate Associations

Aditya Kumar Mishra

B.S., M.S., Statistics and Informatics, Indian Institute of Technology Kharagpur, WB,

India, 2010

M.S., Statistics, University of Connecticut, CT, USA, 2016

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2017

APPROVAL PAGE

Doctor of Philosophy Dissertation

On Sequential Estimation of Multivariate Associations

Presented by

Aditya Kumar Mishra, M.S. Statistics and Informatics, M.S. Statistics

Major Co-Advisor _____

Kun Chen

Major Co-Advisor _____

Dipak Kumar Dey

Associate Advisor _____

Haim Bar

Associate Advisor _____

Elizabeth Schifano

University of Connecticut

2017

Acknowledgements

This dissertation would not have been possible without the support of many individuals, and it's my pleasure in acknowledging them.

First and foremost, I would like to thank both of my academic advisors, Dr. Kun Chen and Prof. Dipak Dey, for accepting me as their PhD Student. I am especially indebted to Dr. Chen, for being extremely patient with me and helping me to learn statistical method development from its conceptualization to efficient implementation. His phenomenal success in statistics has greatly motivated me to pursue a career in academics, and in future will continue to be a role model. I would like to express my deep gratitude to my other advisor Prof. Dey for giving me the freedom to explore, helping me realize power for critical thinking, teaching me how to develop the new ideas, and at the same time supporting me morally. My sincere thanks must go to committee members of my thesis, Dr. Haim Bar and Dr. Elizabeth Schifano, for being generous with their academic excellence and precious time.

In the end, I would like to dedicate this thesis to my family especially to the two most beautiful women my mother and my wife Jai, for standing by my side in all these years to support me. I have been extremely fortunate in my life to have grandparents whose life taught me about the importance of passion and dedication to realize my dream. A special sense of gratitude to my late parents-in-law, for their unfailing emotional

support and blessings. Last but not least, I wouldn't have made it through without the full support of my beloved wife Jai. Her endless love and unwavering faith kept me motivated throughout the journey, without which realizing my dream would not have been possible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Multivariate analysis problems of large scale are increasingly required and formulated in various fields ranging from computational biology to health care, economics, and finance. The problem pertains to modeling dependency of multivariate responses using observed predictors. For example, in computational biology a problem on yeast cell-cycle data of eukaryotic cell cycle is to identify transcription factors (TF) that regulate RNA levels of yeast genes obtained at 18 different time points; see Chen and Huang [2012b]. In health care, the Framingham Heart Study [FHS, 2017] is being conducted to identify factors affecting cardiovascular diseases using single nucleotide polymorphisms (SNP) data as predictor, and observed phenotypes as the multivariate response of patients. In economic research, a related problem is to model investment risk, demand and spending as the responses, using market conditions as predictors. There are several data sets available on handwritten digit recognition on UCI Machine Learning Repository [Lichman, 2013], where the aim is to develop a model for predicting digits using pixel information. In a longitudinal study of aging [Stanziano et al., 2010], it is of interest to understand dependency of health conditions like memory status, depression, daily activities, cognitive

ability on predictor variables like demographics, medical health records etc.

In the above examples, outcome variables can be either continuous, count, binary types, or may be of mixed types. Also, some entries in the observed responses or predictors could be missing. Moreover, some of these problems are high-dimensional with interrelated responses and correlated predictors. It is quite possible that some predictor variables might be completely unimportant. As pointed out by Fan and Lv [2010], the problem poses challenges pertaining to statistical theory, method development, and its scalable implementation. When all the outcome variables are continuous, and the underlying process generating them is Gaussian, one can use multivariate linear regression (MLR) model to best predict responses using predictors. As we step into the non-Gaussian territory, the modeling problem becomes challenging because of difficulty in specifying a joint distribution of the observed outcomes especially when they are interrelated. One way is to express the joint distribution as the product of a marginal and conditional distributions, using which Cox and Wermuth [1992], Fitzmaurice and Laird [1995] proposed likelihood based approaches for the case of bivariate discrete and continuous response. To handle such outcomes, Prentice and Zhao [1991], Zhao et al. [1992] used the framework of generalized estimating equations for estimation of mean and covariance parameter. The inability to handling missing responses, restricted framework and non-applicability in high dimension make the approach unsuitable.

In the high-dimensional setting, we aim to learn a parsimonious model facilitating

interpretation with good prediction performance. One naive approach for getting parameter estimate of the model is to fit a separate model for each response variable, thus ignoring underlying the multivariate nature of the problem. Existing methods cater to the cases with only one type of outcome variables, and largely most of the work has been for continuous outcome variables, i.e., particularly for MLR. Taking into account high dimensionality of only predictors, multivariate ridge regression [Hoerl and Kennard, 1970, Brown and Zidek, 1980] was proposed. However, the approach ignores interrelation among response variables, and hence we look for improvement. The dimension reduction approach of reduced-rank regression (RRR) [Anderson, 1951, Izenman, 1975, Reinsel and Velu, 1998, Bunea et al., 2011, Chen and Chan, 2014] overcomes this deficiency by inducing dependency through the low-rank coefficient matrix. With the aim to identify relevant predictors and discarding others, the sparse regression [Tibshirani, 1996, Fan and Li, 2001, Zou, 2006, Zhao and Yu, 2006, Zhang and Huang, 2008, Bickel et al., 2009] approach is proposed for performing variable selection in the univariate setting. Using the idea of sparse regression, variable selection in multivariate regression is achieved by having a sparse coefficient matrix estimate [Turlach et al., 2005, Peng et al., 2010, Obozinski et al., 2011]. In a non-Gaussian setting with only one type of outcome, for dimension reduction, Yee and Hastie [2003] generalized RRR and proposed the reduced rank vector generalized linear model (RR-VGLM), and She [2011] further proposed an iterative procedure with guaranteed convergence. Methods for estimating a sparse coefficient matrix are still unexplored because of the complicated model

formulation.

The methods discussed above separately attains the two aspects, i.e., dimension reduction and sparsity, in the problems under consideration. Former induce multivariate dependency and later discard redundant variables. Their procedure mainly proceeds via optimization of penalized likelihood with penalties such as nuclear norm [Chen et al., 2013] and rank for performing dimension reduction, while LASSO [Tibshirani, 1996], adaptive LASSO [Zou, 2006] and SCAD [Fan and Li, 2001] for inducing sparsity. An effective strategy desires to achieve both simultaneously. Multivariate response variables are modeled using predictors via the coefficient matrix, and the two desirable property of dimension reduction and variable selection can be simultaneously attained by having a low-rank and sparse coefficient matrix structure. Specifically, the structure translates into having co-sparse left and right singular vectors in the singular value decomposition (SVD) of the coefficient matrix. Such structure results in the association of response and predictor variables via latent factors, where each of them is constructed using a subset of predictors affecting only a subset of responses.

Several methods are proposed for recovery of required SVD components of the coefficient matrix in MLR. For example, Yuan et al. [2007], Negahban and Wainwright [2011b], Koltchinskii et al. [2011], Chen et al. [2013] proposed the singular value penalization approach to estimate low-rank structure using a nuclear norm penalty on the coefficient matrix; Mukherjee and Zhu [2011] proposed a reduced-rank ridge regression performing similar recovery using ridge penalty along with rank constraint. To perform

variable selection in a reduced rank model, Chen et al. [2012], Chen and Huang [2012b], Bunea et al. [2012], Ma and Sun [2014] proposed low-rank and sparse models which use a sparsity inducing penalty formulated from SVD components of the coefficient matrix.

Alternatively, coefficient matrix of rank r can be expressed as the sum of r unit rank matrices, each of which is the outer-product of left and right singular vectors from its SVD components. RRR is a supervised extension of factor analysis, and the latent factors and loading matrix can be constructed from the unit rank matrix obtained sequentially by performing r unit-rank estimation problems; see Reinsel and Velu [1998]. When each of the unit rank matrices is sparse, coefficient matrix expressed as their sum will be both low-rank and sparse. Thus, a possible extension of the sequential procedure is to solve the unit-rank estimation problem in presence of sparsity inducing penalty. In a contemporary work, for MLR, Bahadori et al. [2016] proposed the sequential approach to obtain a low-rank and sparse coefficient matrix by solving a sparse generalized eigenvalue problems. The framework is a bit restricted when it comes to extending the idea in cases of mixed types/non-Gaussian outcomes.

Motivated by the idea of sequential estimation in RRR, we aim to develop a procedure for low-rank and sparse coefficient matrix estimation for MLR. From a statistical modeling perspective, we formulate the problem as a sparse factor regression and develop an efficient sequential computation procedure, called *sequential factor extraction via constrained unit-rank estimation* (SeCURE). At each sequential step, a latent factor

is constructed as a linear combination of the subset the observed predictors, for predicting the responses after accounting for the effects of the previous factors; each factor is allowed to potentially influence only a subset of responses. This is attained via co-sparse left and right singular vectors of the estimated unit-rank coefficient matrix. Each sequential step reduces to a regularized unit-rank regression in which the orthogonality constraints among the sparse factors become optional rather than necessary, in contrast to alternative joint estimation approach. Coordinate descent and Lagrangian multipliers are utilized to ensure fast computation and algorithmic convergence, even in the presence of missing data. The sequential procedure terminates automatically when the residual signal is not sufficient, thus no need to specify rank of the coefficient matrix. We justify our computation approach by showing that the sequential estimators enjoy the oracle properties for recovering the underlying sparse factor structure. The efficacy of our method is demonstrated by simulation studies and two real applications in genetics.

In multivariate problems, we have seen that it is not always that outcome variables are continuous in nature. They can also be of binary or count types or may be of mixed types. This calls for the development of methodology in the more generalized setting. We assume that underlying process generating such outcomes can be from the exponential dispersion family, e.g., Poisson, Bernoulli or Gaussian. Again in the high-dimensional setting, the interrelated response variables and correlated predictor variables can be modeled via a low-rank and sparse coefficient matrix. We do not find any significant literature solving the required problem. Such problems exist in real life and require

attention.

For multivariate regression problems in the generalized setting, motivated by SeCURE, we propose a sequential procedure for low-rank and sparse coefficient matrix estimation referred as *generalize sequential factor extraction via constrained unit-rank estimation* (GSeCURE). In a unit step, the procedure estimates a sparse unit-rank matrix in terms of co-sparse singular vectors by minimizing a regularized unit-rank constrained likelihood function defined using responses, predictors and offset term accounting for the estimated signal. The optimization problem is challenging because of complicated likelihood structure. Hence, following She [2012a], a surrogate of the objective function is defined using scaled observed predictors minimizing which ensures its monotone descending property. Under certain mild regularity condition, our algorithm converges. Moreover, our formulation can easily handle missing entries in the response matrix while also providing a reasonably acceptable estimate. We have depicted efficacy of GSeCURE through several simulated examples in different scenarios of the generalized multivariate regression setting with $\{0\%, 20\%\}$ missing entries in the response matrix.

In rest of the thesis, we have given details of our proposed methods and their efficient implementation. We have presented our method sequential co-sparse factor regression (SeCURE) in Chapter 2. The greedy algorithm, generalized co-sparse factor regression (GSeCURE), for low-rank and sparse component matrix estimation in the generalized multivariate regression is described in Chapter 3. We have implemented both procedures in R package *secure* using Rcpp and RcppArmadillo. Chapter 4 provides a vignette of

secure package. We conclude and discuss some future research topics in Chapter 5.

Chapter 2

Sequential Co-Sparse Factor Regression

2.1 Introduction

We are interested in studying the predictive relationship between a multivariate response $\mathbf{y} \in \mathbb{R}^q$ and a multivariate predictor $\mathbf{x} \in \mathbb{R}^p$, using the multivariate linear regression model

$$\mathbf{y} = \mathbf{C}^{*\mathrm{T}}\mathbf{x} + \mathbf{e}, \tag{2.1}$$

where $\mathbf{C}^* \in \mathbb{R}^{p\times q}$ is an unknown coefficient matrix, and $\mathbf{e} \in \mathbb{R}^q$ is a random error vector of zero mean. Given n independent observations $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, let $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times q}$ be the response matrix, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ the predictor/design matrix, and $\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times q}$ the error matrix. Then the sample

model can be expressed as

$$\mathbf{Y} = \mathbf{X}\mathbf{C}^* + \mathbf{E}. \tag{2.2}$$

The responses and the predictors are assumed to be centered, so there is no intercept term. In many applications, it is desirable to assume that $\mathbf{C}^*$ admits certain low-dimensional structures, the exploration of which could utilize the multivariate dependency, alleviate the curse of dimensionality, and facilitate model interpretation. In particular, $\mathbf{C}^*$ could be of low rank, i.e., $r^* = \text{rank}(\mathbf{C}^*) \leq \min(r_x, q)$, where $r_x = \text{rank}(\mathbf{X})$.

To recover the potential low-rank structure in $\mathbf{C}^*$, the reduced-rank regression (RRR) [Anderson, 1951, Reinsel and Velu, 1998] was formulated as a rank-constrained estimation,

$$\min_{\mathbf{C}} \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_F^2, \qquad \text{s.t. } \text{rank}(\mathbf{C}) \leq r, \tag{2.3}$$

for $1 \leq r \leq \min(r_x, q)$, where $\|\cdot\|_F$ denotes the Frobnenius norm. This reduces the effective number of regression parameters from $r_x q$ to $(r_x + q - r)r$, and the reduction can be substantial when r is much smaller than both r_x and q. Bunea et al. [2011] formulated the rank-constrained estimation as a regularized regression problem in which the penalty is proportional to the ℓ_0 norm of the singular values of $\mathbf{C}$. The convex nuclear norm penalization approach achieved rank reduction through penalizing the ℓ_1 norm of the

singular values of $\mathbf{C}$ [Yuan et al., 2007, Negahban and Wainwright, 2011a, Lu et al., 2012]. See also, Mukherjee and Zhu [2011], Chen et al. [2013] and Zhou and Li [2014].

It is fair to say that the reduced-rank structure is one of the most critical ingredients in multivariate learning. The RRR naturally connects to factor analysis (FA) and principal component analysis (PCA): both FA/PCA and RRR aim to identify a certain low-dimensional subspace to represent $\mathbf{Y}$, and the main difference is that the latter conducts a supervised search in the column space of $\mathbf{X}$ while the search in the former is unsupervised. Motivated by these connections, a sparse SVD representation of $\mathbf{XC}^*$ is quite appealing,

$$\frac{1}{\sqrt{n}}\mathbf{XC}^* = (\frac{1}{\sqrt{n}}\mathbf{XU}^*)\mathbf{D}^*\mathbf{V}^{*\mathrm{T}}, \qquad \text{s.t. } (\frac{1}{\sqrt{n}}\mathbf{XU}^*)^{\mathrm{T}}(\frac{1}{\sqrt{n}}\mathbf{XU}^*) = \mathbf{V}^{*\mathrm{T}}\mathbf{V}^* = \mathbf{I}_{r^*}, \quad (2.4)$$

where $\mathbf{C}^* = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*\mathrm{T}}$, $\mathbf{I}_{r^*}$ denotes the $r^* \times r^*$ identity matrix, and $\mathbf{D}^* = \mathrm{diag}\{d_1^*, \ldots, d_r^*\}$ is an $r^* \times r^*$ diagonal matrix of the singular values. Here both $\mathbf{U}^* = [\mathbf{u}_1^*, \ldots, \mathbf{u}_{r^*}^*] \in \mathbb{R}^{p \times r^*}$ and $\mathbf{V}^* = [\mathbf{v}_1^*, \ldots, \mathbf{v}_{r^*}^*] \in \mathbb{R}^{q \times r^*}$ are assumed to be sparse, which we refer to as the *"co-sparsity"* structure. There are several prominent features in this co-sparse and low-rank formulation: (I) $(1/\sqrt{n})\mathbf{XU}^*$ gives r^* latent predictors/factors, i.e., $(1/\sqrt{n})\mathbf{Xu}_k^*$, $k = 1, \ldots, r^*$, each of which is constructed as a linear combination of a subset of the original predictors (factor-specific predictor selection); (II) each latent factor is allowed to be related to a subset of the response variables as each $\mathbf{v}_k^*$ is also sparse (factor-specific response selection); (III) the sample latent factors are uncorrelated with each

other, due to the orthogonality of $(1/\sqrt{n})\mathbf{X}\mathbf{U}^*$, and the parameter identifiability is ensured with the orthogonality of $\mathbf{V}^*$; (IV) the singular values d_k^* directly indicate the strengths and importance of the associations between $\mathbf{Y}$ and the latent factors. In recent years, although several works considered sparse and low-rank models, none of them could achieve (I)–(IV) together. In Chen et al. [2012], the coefficient matrix $\mathbf{C}^*$, rather than the regression component $\mathbf{X}\mathbf{C}^*$, was assumed to admit a sparse SVD, so their method failed to achieve (III) and (IV). In Chen and Huang [2012a], Bunea et al. [2012] and Ma and Sun [2014], all the latent factors were constructed from the same subset of predictors and were related to all the responses, so their methods did not fully achieve (I) and did not consider (II). In Chen et al. [2013], the decomposition of the regression components $\mathbf{X}\mathbf{C}^*$ were considered mainly for constructing a class of computationally efficient low-rank estimators, but they did not pursue its potential sparsity. So far, how to efficiently recover the model structure in (2.4) remains a challenging problem, largely due to the simultaneous presence of orthogonality and co-sparsity requirements.

To tackle this problem, we explore the underlying data generating mechanism targeted by the sparse SVD structure in (2.4), which enables us to reformulate the problem as a *co-sparse factor regression* and develop a novel sequential extraction procedure. At each sequential step, the estimation problem nicely reduces to a sparse multivariate regression with a unit-rank constraint, in which the orthogonality constraint is not needed at all. Interestingly, each estimated sparse and unit-rank coefficient matrix automatically leads to co-sparsity in its two singular vectors. As such, each step extracts a new

latent factor as a sparse linear combination of the original predictors, which aims to best predict the responses after controlling for the effects of the previously extracted factors. The proposed method completely bypasses the orthogonality constraints in the recovery of the co-sparse and low-rank structure, which, in contrast, would have been inevitable in any joint estimation approach due to the need of parameter identifiability. Most recently, a similar sparse and low-rank model was considered in Bahadori et al. [2016], in which a two-step estimation procedure was used, i.e., solve sparse generalized eigenvalue problems to obtain initial estimates, and then apply thresholding for refining the co-sparisty pattern. Comparing to their method, our approach achieves factor extraction and co-sparsity recovery together, allowing the two tasks to inform and reinforce each other. Our regression formulation also allows convenient handling of incomplete response data, which is bound to occur in big data problems.

We propose the co-sparse factor regression model in Section 2.2. A sequential extraction procedure for model estimation is presented in Section 2.3. We then develop a co-sparse unit-rank estimation algorithm in Section 2.4 with convergence guarantee. In Section 2.5, the sequentially extracted estimators are shown to enjoy the oracle properties asymptotically, and a non-asymptotic error bound reveals interesting finite-sample behaviors of the estimators. Simulation studies in Section 2.6 and an application in Section 2.7.2 further showcase the effectiveness of the proposed approach.

2.2 Co-Sparse Factor Regression Model

To motivate, we start with the familiar factor analysis (FA) [Anderson and Rubin, 1956]. For a continuous variable $\mathbf{y} \in \mathbb{R}^q$, the model can be expressed as $\mathbf{y} = \widetilde{\mathbf{V}}\mathbf{z} + \mathbf{e}$, where $\mathbf{z} \in \mathbb{R}^{r^*}$ consists of a set of unobserved latent variables, $\widetilde{\mathbf{V}} \in \mathbb{R}^{q \times r^*}$ is an unknown loading matrix, and $\mathbf{e} \in \mathbb{R}^q$ is a random error vector of zero mean. The r^* latent factors satisfy $\mathrm{E}(\mathbf{z}) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{z}) = \mathbf{I}_{r^*}$, the error satisfies $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{e}) = \mathbf{\Sigma}$, usually a diagonal matrix, and $\mathbf{e}$ and $\mathbf{z}$ are independent. Here we assume $\mathrm{E}(\mathbf{y}) = \mathbf{0}$, so there is no intercept. It is common to parameterize $\widetilde{\mathbf{V}} = \mathbf{V}^*\mathbf{D}^*$ such that $\mathbf{V}^{*\mathrm{T}}\mathbf{V}^* = \mathbf{I}_{r^*}$ and $\mathbf{D}^* = \mathrm{diag}\{d_1^*, \ldots, d_{r^*}^*\}$.

Now consider the RRR model in (2.1). Write the coefficient matrix $\mathbf{C}^*$ as $\mathbf{C}^* = \mathbf{U}^*\widetilde{\mathbf{V}}^{\mathrm{T}}$, for some $\mathbf{U}^* \in \mathbb{R}^{p \times r^*}$ and $\widetilde{\mathbf{V}} \in \mathbb{R}^{q \times r^*}$. Let $\mathbf{x} \in \mathbb{R}^p$ be the multivariate predictor variable, with $\mathrm{E}(\mathbf{x}) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{x}) = \mathbf{\Gamma}$. It is immediately clear that RRR can be regarded as a supervised factor analysis in which the latent factors are some linear projections of $\mathbf{x}$, i.e,

$$\mathbf{y} = \mathbf{C}^{*\mathrm{T}}\mathbf{x} + \mathbf{e} = \widetilde{\mathbf{V}}(\mathbf{U}^{*\mathrm{T}}\mathbf{x}) + \mathbf{e}, \tag{2.5}$$

where $\widetilde{\mathbf{V}} = \mathbf{V}^*\mathbf{D}^*$ with $\mathbf{V}^{*\mathrm{T}}\mathbf{V}^* = \mathbf{I}_{r^*}$ and $\mathrm{cov}(\mathbf{U}^{*\mathrm{T}}\mathbf{x}) = \mathbf{U}^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{U}^* = \mathbf{I}_{r^*}$. With n

independent observations, (2.5) then leads to the following sample model,

$$\begin{aligned} \mathbf{Y} =& \mathbf{X}\mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*\mathrm{T}} + \mathbf{E} = \sum_{k=1}^{r^*} d_k^*(\mathbf{X}\mathbf{u}_k^*)\mathbf{v}_k^{*\mathrm{T}} + \mathbf{E} = \sum_{k=1}^{r^*} \mathbf{X}\mathbf{C}_k^* + \mathbf{E}, \\ & \text{s.t. } \mathbf{U}^{*\mathrm{T}}\boldsymbol{\Gamma}\mathbf{U}^* = \mathbf{I}_{r^*}, \mathbf{V}^{*\mathrm{T}}\mathbf{V}^* = \mathbf{I}_{r^*}. \end{aligned} \tag{2.6}$$

where $\mathbf{C}^* = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*\mathrm{T}}$, $\mathbf{U}^* = [\mathbf{u}_1^*, \ldots, \mathbf{u}_{r^*}^*] \in \mathbb{R}^{p\times r^*}$, $\mathbf{V}^* = [\mathbf{v}_1^*, \ldots, \mathbf{v}_{r^*}^*] \in \mathbb{R}^{q\times r^*}$, $\mathbf{D}^* = \mathrm{diag}\{d_1^*, \ldots, d_{r^*}^*\}$, and $\mathbf{C}_k^* = d_k^*\mathbf{u}_k^*\mathbf{v}_k^{*\mathrm{T}}$. The d_k^*s are assumed to be distinct and are placed in a descending order, i.e., $d_1^* > \ldots > d_{r^*}^* > 0$. By further assuming that both $\mathbf{U}^*$ and $\mathbf{V}^*$ in (2.6) are sparse, we arrive at our proposed *co-sparse factor regression* model (SFAR).

It is now clear that the decomposition in (2.6) can be viewed as the population version of the sample decomposition in (2.4). Consequently, the above SFAR model possesses several prominent features, e.g., latent model interpretation, predictor selection, response selection, etc., as discussed in Section 2.1. In particular, the model reveals that there are r^* distinct uncorrelated latent factors of descending importance relating the responses to the predictors. In the kth latent association, $k = 1, \ldots, r^*$, the elements in $\mathbf{u}_k^*$ give the weights for constructing the factor, the elements in $\mathbf{v}_k^*$ provide the relative effects of the kth factor on the responses, and d_k^* indicates the importance of the kth factor. Due to the co-sparsity on $\mathbf{u}_k^*$ and $\mathbf{v}_k^*$, each latent factor may be constructed from only a subset of the predictors and may only influence a subset of the responses.

At the first glance, it is convenient to conduct model estimation by solving

$$\min_{\mathbf{U},\mathbf{D},\mathbf{V}} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}\|_F^2 + \lambda_1\rho_1(\mathbf{U}) + \lambda_2\rho_2(\mathbf{V}) \right\}, \text{ s.t. } \mathbf{U}^{\mathrm{T}}(\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n})\mathbf{U} = \mathbf{I}, \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}, \tag{2.7}$$

where $\mathbf{U} \in \mathbb{R}^{p\times r}$, $\mathbf{V} \in \mathbb{R}^{q\times r}$, and $\mathbf{D} = \mathrm{diag}\{d_1, \ldots, d_r\}$, for some $1 \leq r \leq \min(r_x, q)$. Here ρ_1, ρ_2 are some sparsity-inducing penalty functions [Tibshirani, 1996] with λ_1, λ_2 being their tuning parameters. Since $\mathbf{\Gamma}$ is unknown, it is replaced with $\mathbf{X}^{\mathrm{T}}\mathbf{X}/n$. Unfortunately, there are several difficulties with this joint estimation approach. The rank r has to be specified in advance. The method requires that the sample realizations of the latent factors are exactly uncorrelated, which is necessary to ensure parameter identifiability in (2.7). However, this requirement is too stringent in view of the true model in (2.6). Moreover, to the best of our knowledge, existing optimization methods for solving orthogonality constrained regularized estimation are mostly heuristic and do not scale well to large problems. The simultaneous presence of the high-dimensional low-rank structure, the co-sparsity regularization and the orthogonality constraints makes the estimation challenging. This motivates us to explore new avenues for conducting model estimation in SFAR.

2.3 Co-Sparsity Recovery via Sequential Extraction

2.3.1 Sequential Extraction in Reduced-Rank Estimation

The SFAR model in (2.6) aims to approximate $\mathbf{Y}$ using a few supervised latent factors, which are uncorrelated and ordered in descending importance based on their predictive power. A promising way is to extract the latent factors in a *sequential fashion*. This idea of sequential extraction is not new, but it was mainly studied in unsupervised learning from machine learning perspectives; see, e.g., Mackey [2009] and Journée et al. [2010].

To lay the foundation of our approach, we first show in Proposition 2.1 that the classical RRR can be performed via a sequential unit-rank extraction procedure.

Proposition 2.1. *The reduced-rank estimator* $\widetilde{\mathbf{C}}(r)$*, which solves* (2.3) *for some* $1 \leq r \leq \min(r_x, q)$*, can be obtained by sequentially performing unit-rank regression,*

$$\widetilde{\mathbf{C}}_k = \arg\min_{\mathbf{C}} \|\mathbf{Y}_k - \mathbf{X}\mathbf{C}\|_F^2, \qquad \textit{s.t. rank}(\mathbf{C}) \leq 1, \tag{2.8}$$

each time with $\mathbf{Y}_k$ *being the current residual matrix, i.e.,* $\mathbf{Y}_1 = \mathbf{Y}$*,* $\mathbf{Y}_k = \mathbf{Y} - \mathbf{X}\sum_{h=1}^{k-1}\widetilde{\mathbf{C}}_h$*,* $k = 2, \ldots, r$*. That is,* $\widetilde{\mathbf{C}}(r) = \sum_{k=1}^{r}\widetilde{\mathbf{C}}_k$*.*

Moreover, $\widetilde{\mathbf{C}}_k = \widetilde{d}_k\widetilde{\mathbf{u}}_k\widetilde{\mathbf{v}}_k^{\mathrm{T}}$ *such that* $\widetilde{\mathbf{u}}_k = (1/\widetilde{d}_k)(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\widetilde{\mathbf{v}}_k$*, where* $()^{-}$ *denotes the Moore-Penrose inverse, and* $\widetilde{\mathbf{v}}_k$ *is the normalized eigenvector corresponding to the* k^{th} *largest eigenvalue* $\widetilde{d}_k^2$ *of* $(1/n)\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$*. Therefore, the estimators satisfy* $(1/n)\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widetilde{\mathbf{U}} = \mathbf{I}$ *and* $\widetilde{\mathbf{V}}^{\mathrm{T}}\widetilde{\mathbf{V}} = \mathbf{I}$*, where* $\widetilde{\mathbf{U}} = [\widetilde{\mathbf{u}}_1, \ldots, \widetilde{\mathbf{u}}_r]$ *and* $\widetilde{\mathbf{V}} = [\widetilde{\mathbf{v}}_1, \ldots, \widetilde{\mathbf{v}}_r]$*.*

Proof. The rank-r RRR estimator of $\mathbf{C}$, which minimizes $\|\mathbf{Y} - \mathbf{XC}\|_F^2$ subject to $\text{rank}(\mathbf{C}) = r$, is given by

$$\widetilde{\mathbf{C}}(r) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^{\mathrm{T}},$$

where $\widetilde{\mathbf{V}} = [\widetilde{\mathbf{v}}_1, \ldots, \widetilde{\mathbf{v}}_r]$ and $\widetilde{\mathbf{v}}_k$ is the normalized eigenvector that corresponds to the kth largest eigenvalue $\widetilde{d}_k^2$ of the matrix $\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}/n$ [Reinsel and Velu, 1998]. Then one can write $\widetilde{\mathbf{C}}(r) = \sum_{k=1}^{r} \widetilde{\mathbf{C}}_k$ with $\widetilde{\mathbf{C}}_k = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\widetilde{\mathbf{v}}_k\widetilde{\mathbf{v}}_k^{\mathrm{T}}$.

It then suffices to show that $\widetilde{\mathbf{C}}_k$ is the solution from the kth step in the sequential estimation. In the sequential RRR, the first step minimizes $\|\mathbf{Y} - \mathbf{XC}\|_F^2$ subject to $\text{rank}(\mathbf{C}) = 1$; it is easily seen that the solution is given by $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\widetilde{\mathbf{v}}_1\widetilde{\mathbf{v}}_1^{\mathrm{T}} = \widetilde{\mathbf{C}}_1$. Now, following the sequential fitting procedure, the solution of the second step is given by

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{C}}_1)\mathbf{v}\mathbf{v}^{\mathrm{T}}$$

where $\mathbf{v}$ is the normalized eigenvector that corresponds to the largest eigenvalue of the

matrix $(\mathbf{Y}-\mathbf{X}\widetilde{\mathbf{C}}_1)^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}(\mathbf{Y}-\mathbf{X}\widetilde{\mathbf{C}}_1)/n$, which can be simplified as follows,

$$\begin{aligned}
&(\mathbf{Y}-\mathbf{X}\widetilde{\mathbf{C}}_1)^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}(\mathbf{Y}-\mathbf{X}\widetilde{\mathbf{C}}_1)\\
=&\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}-\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\widetilde{\mathbf{v}}_1\widetilde{\mathbf{v}}_1^{\mathrm{T}}\\
&-\widetilde{\mathbf{v}}_1\widetilde{\mathbf{v}}_1^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}+\widetilde{\mathbf{v}}_1\widetilde{\mathbf{v}}_1^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\widetilde{\mathbf{v}}_1\widetilde{\mathbf{v}}_1^{\mathrm{T}}\\
=&\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}-n\widetilde{d}_1^2\widetilde{\mathbf{v}}_1\widetilde{\mathbf{v}}_1^{\mathrm{T}}-n\widetilde{d}_1^2\widetilde{\mathbf{v}}_1\widetilde{\mathbf{v}}_1^{\mathrm{T}}+n\widetilde{d}_1^2\widetilde{\mathbf{v}}_1\widetilde{\mathbf{v}}_1^{\mathrm{T}}\\
=&\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}-n\widetilde{d}_1^2\widetilde{\mathbf{v}}_1\widetilde{\mathbf{v}}_1^{\mathrm{T}}.
\end{aligned}$$

Since $\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\widetilde{\mathbf{v}}_2/n-\widetilde{d}_1^2\widetilde{\mathbf{v}}_1\widetilde{\mathbf{v}}_1^{\mathrm{T}}\widetilde{\mathbf{v}}_2=\widetilde{d}_2^2\widetilde{\mathbf{v}}_2$, it follows that $\mathbf{v}=\widetilde{\mathbf{v}}_2$, and the solution of the second step is

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}(\mathbf{Y}-\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\widetilde{\mathbf{v}}_1\widetilde{\mathbf{v}}_1^{\mathrm{T}})\widetilde{\mathbf{v}}_2\widetilde{\mathbf{v}}_2^{\mathrm{T}}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\widetilde{\mathbf{v}}_2\widetilde{\mathbf{v}}_2^{\mathrm{T}}=\widetilde{\mathbf{C}}_2.$$

Similarly, we can show that the sequential fitting solution for the kth step is given by $\widetilde{\mathbf{C}}_k$, for $k=1,\ldots,r$; thus $\widetilde{\mathbf{C}}_k$ can be obtained sequentially.

Remaining part of the proof can be easily followed from the expression of $\widetilde{\mathbf{C}}_k$.

□

In Proposition 2.1, there is no orthogonality constraint ever imposed. Interestingly, the sequentially extracted solutions automatically correspond to the decomposition in (2.6) and automatically satisfy the exact orthogonality! Although this "free orthogonality" may no longer hold when sparse regularization on $\mathbf{U}$ and $\mathbf{V}$ is imposed, Proposition

2.1 suggests that it is promising that the co-sparse $\mathbf{U}^*$ and $\mathbf{V}^*$ can be consistently recovered via sequential extraction without explicit orthogonality constraints.

Proposition 2.1 also further reveals that the proposed decomposition in (2.6) is quite special and plays a key role to ensure the potential success of sequential recovery. To understand this, recall that there are infinite number of ways to decompose $\mathbf{C}^*$ to the form $\mathbf{C}^* = \mathbf{U}_c\mathbf{V}_c^{\mathrm{T}}$ for some $\mathbf{U}_c \in \mathbb{R}^{p\times r^*}$ and $\mathbf{V}_c \in \mathbb{R}^{q\times r^*}$, simply because $\mathbf{C}^* = \mathbf{U}_c\mathbf{V}_c^{\mathrm{T}} = (\mathbf{U}_c\mathbf{Q}^{\mathrm{T}})(\mathbf{V}_c\mathbf{Q}^{-1})^{\mathrm{T}}$ for any non-singular matrix $\mathbf{Q} \in \mathbb{R}^{r^*\times r^*}$. Among these infinite possibilities, the decomposition in (2.6) is the special one that can produce a set of uncorrelated latent factors and can be recovered by sequential unit-rank extraction.

2.3.2 Sequential Extraction in SFAR

Motivated by (2.8), to extract the first latent factor and its effects on the responses, we formulate the task as the following unit-rank penalized least squares problem,

$$\widehat{\mathbf{C}}_1^{(\lambda)} = \arg\min_{\mathbf{C}} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_F^2 + \rho(\mathbf{C};\lambda) \right\}, \text{ s.t. } \mathrm{rank}(\mathbf{C}) \le 1, \tag{2.9}$$

where $\rho(\cdot;\lambda)$ is a certain sparsity-inducing penalty term with tuning parameter λ. Comparing to (2.8), the only difference in (2.9) is that additional sparsity regularization on $\mathbf{C}$ is introduced. From a matrix approximation point of view, this new criterion aims to produce the best unit-rank approximation of $\mathbf{Y}$ in the column space of $\mathbf{X}$ subject to a given level of sparsity in $\mathbf{C}$. As $\widehat{\mathbf{C}}_1^{(\lambda)}$ is of unit rank, it can be decomposed as

$\widehat{\mathbf{C}}_1^{(\lambda)} = \widehat{d}_1^{(\lambda)}\widehat{\mathbf{u}}_1^{(\lambda)}\widehat{\mathbf{v}}_1^{(\lambda)\mathrm{T}}$, with $\widehat{d}_1^{(\lambda)} \geq 0$, $\widehat{\mathbf{u}}_1^{(\lambda)\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\mathbf{u}}_1^{(\lambda)}/n = 1$ and $\widehat{\mathbf{v}}_1^{(\lambda)\mathrm{T}}\widehat{\mathbf{v}}_1^{(\lambda)} = 1$. Similar to SVD, this decomposition is unique up to the signs of the vectors as long as $\widehat{\mathbf{C}}_1^{(\lambda)}$ is not a zero matrix.

It is important to realize that the sparsity structure in a unit-rank matrix is very special, in that its sparsity can directly translate to the co-sparsity of its pairs of singular vectors! Proposition 2.2 fully characterizes this correspondence. This simple yet interesting observation is the key that enables us to "kill two birds with one stone": by promoting the entrywise sparsity of the unit-rank coefficient matrix, solving (2.9) yields the sparse factor coefficients $\widehat{\mathbf{u}}_1^{(\lambda)}$ and the sparse factor effects $\widehat{\mathbf{v}}_1^{(\lambda)}$ simultaneously.

Proposition 2.2. *Suppose* $\mathbf{C} \in \mathbb{R}^{p\times q}$ *is of unit rank and* $\mathbf{C} \neq \mathbf{0}$. *Write* $\mathbf{C} = [c_{ij}] = [\mathbf{c}_1, \ldots, \mathbf{c}_q] = [\widetilde{\mathbf{c}}_1^{\mathrm{T}}, \ldots, \widetilde{\mathbf{c}}_p^{\mathrm{T}}]^{\mathrm{T}}$. *Then* $\mathbf{C}$ *can be decomposed as* $\mathbf{C} = d\mathbf{u}\mathbf{v}^{\mathrm{T}}$, *for some* $d > 0$, $\mathbf{u} = [u_i] \in \mathbb{R}^p$ *and* $\mathbf{v} = [v_j] \in \mathbb{R}^q$. *Moreover, for any* $i = 1, \ldots, p$, $j = 1, \ldots, q$, *it holds that (a)* $c_{ij} = 0 \iff$ *Either* $\widetilde{\mathbf{c}}_i = \mathbf{0}$ *or* $\mathbf{c}_j = \mathbf{0}$; *(b)* $u_i = 0 \iff \widetilde{\mathbf{c}}_i = \mathbf{0}$; *and (c)* $v_j = 0 \iff \mathbf{c}_j = \mathbf{0}$.

We use the adaptive elastic net penalty [Zou and Hastie, 2005, Zou and Zhang, 2009],

$$\begin{aligned}\rho(\mathbf{C}; \lambda) = \rho(\mathbf{C}; \mathbf{W}_1, \lambda, \alpha) &= \alpha\lambda\|\mathbf{W}_1 \circ \mathbf{C}\|_1 + (1-\alpha)\lambda\|\mathbf{C}\|_F^2 \\ &= \alpha\lambda\sum_{i=1}^{p}\sum_{j=1}^{q} w_{ij1}|c_{ij}| + (1-\alpha)\lambda\sum_{i=1}^{p}\sum_{j=1}^{q} c_{ij}^2. \end{aligned} \tag{2.10}$$

Here $\|\cdot\|_1$ denotes the ℓ_1 norm, the operator "$\circ$" stands for the Hadamard product, $\mathbf{W}_1 = [w_{ij1}]_{p\times q}$ is a pre-specified weighting matrix, λ is a tuning parameter controlling

the overall amount of regularization, and $\alpha \in (0,1)$ controls the relative weights between the two penalty terms. We set $\mathbf{W}_1 = |\widetilde{\mathbf{C}}_1|^{-\gamma}$ such that $w_{ij1} = w_1^{(d)} w_{i1}^{(u)} w_{j1}^{(v)}$, with

$$w_1^{(d)} = |\widetilde{d}_1|^{-\gamma}, \mathbf{w}_1^{(u)} = [w_{11}^{(u)}, ..., w_{p1}^{(u)}]^{\mathrm{T}} = |\widetilde{\mathbf{u}}_1|^{-\gamma}, \mathbf{w}_1^{(v)} = [w_{11}^{(v)}, ..., w_{q1}^{(v)}]^{\mathrm{T}} = |\widetilde{\mathbf{v}}_1|^{-\gamma}, \quad (2.11)$$

where $\widetilde{\mathbf{C}}_1 = \widetilde{d}_1 \widetilde{\mathbf{u}}_1 \widetilde{\mathbf{v}}_1^{\mathrm{T}}$ is the first set of unit-rank RRR estimator given in Proposition 2.1, and γ is a non-negative constant with $|\cdot|^{-\gamma}$ componentwisely defined. As suggested by Zou [2006], we set $\gamma = 2$. Since we mainly focus on sparse estimation, we fix α as a constant, i.e., $\alpha = 0.95$. Comparing to lasso, one advantage of elastic net is that the additional ridge penalty improves the convexity of the problem and can enhance the stability of optimization (see Section 2.4). For simplicity, we may write $\rho(\mathbf{C}; \mathbf{W}, \lambda, \alpha) = \rho(\mathbf{C}; \mathbf{W}, \lambda)$.

To focus on the sequential extraction procedure, we defer the details of optimization to Section 2.4. Assuming the solution path of (2.9) can be fitted, the tuning parameter λ can be chosen based on either cross validation or some information criterion. This yields the parameter estimates $\widehat{\mathbf{C}}_1$, or equivalently, $(\widehat{d}_1, \widehat{\mathbf{u}}_1, \widehat{\mathbf{v}}_1)$.

The search of the subsequent latent factor proceeds by replacing $\mathbf{Y}$ with the current residual matrix $\mathbf{Y}_k$. Specifically, the estimation of the kth factor, for $k = 2, \ldots, r$, becomes

$$\widehat{\mathbf{C}}_k^{(\lambda)} = \arg\min_{\mathbf{C}} \left\{ \frac{1}{2} \|\mathbf{Y}_k - \mathbf{X}\mathbf{C}\|_F^2 + \rho(\mathbf{C}; \mathbf{W}_k, \lambda) \right\}, \text{ s.t. } \mathrm{rank}(\mathbf{C}) \leq 1, \quad (2.12)$$

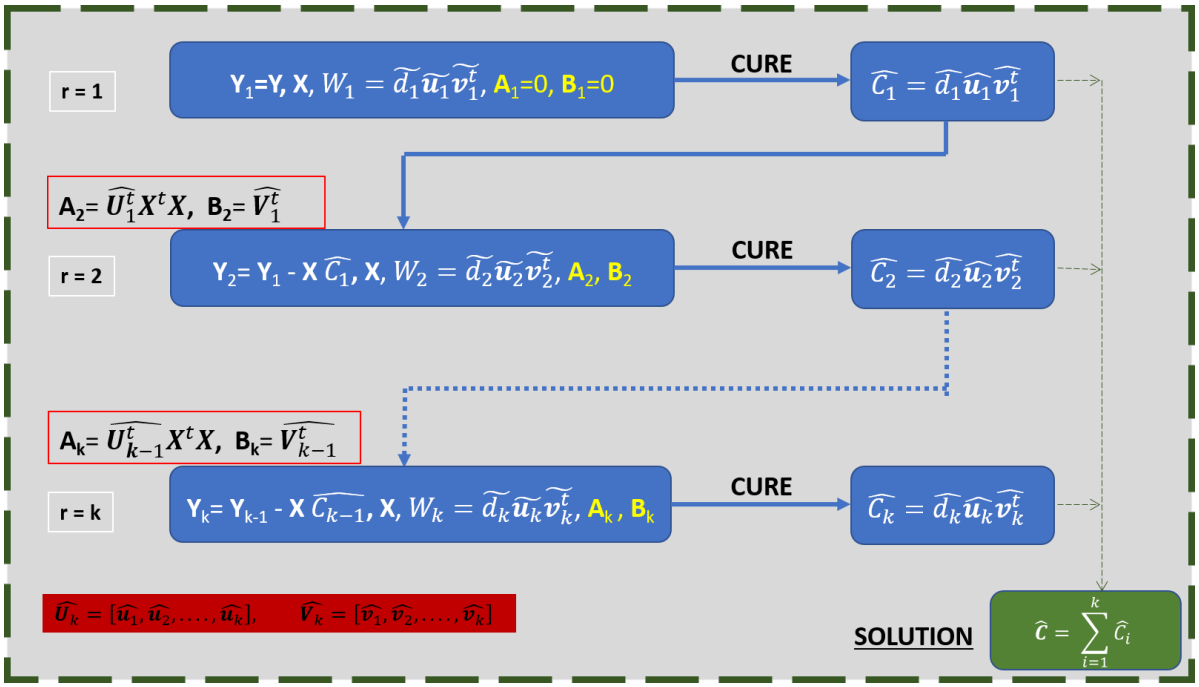
where $\mathbf{W}_k$ is similarly constructed as in (2.11) with $(\widetilde{d}_k, \widetilde{\mathbf{u}}_k, \widetilde{\mathbf{v}}_k)$ from the sequential RRR (see proposition 2.1), and

$$\mathbf{Y}_k = \mathbf{Y} - \sum_{l=1}^{k-1} \mathbf{X}\widehat{\mathbf{C}}_l = \mathbf{Y} - \sum_{l=1}^{k-1} \widehat{d}_l \mathbf{X}\widehat{\mathbf{u}}_l\widehat{\mathbf{v}}_l^{\mathrm{T}}, \tag{2.13}$$

with $\widehat{\mathbf{C}}_l = \widehat{d}_l\widehat{\mathbf{u}}_l\widehat{\mathbf{v}}_l^{\mathrm{T}}$, $\widehat{d}_l \geq 0$, $\widehat{\mathbf{u}}_l^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\mathbf{u}}_l/n = 1$ and $\widehat{\mathbf{v}}_l^{\mathrm{T}}\widehat{\mathbf{v}}_l = 1$. Upon defining $\mathbf{Y}_1 = \mathbf{Y}$, (2.12) also subsumes the first-step problem in (2.9).

We sequentially perform regularized estimation analysis using (2.12) to obtain $\widehat{\mathbf{C}}_k$ or equivalently $(\widehat{d}_k, \widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k)$, for $k = 1, \ldots, r$, $r \leq \min(r_x, q)$. The estimators up to rank r for $\mathbf{C}$ are given by $\widehat{\mathbf{C}}(k) = \sum_{l=1}^{k} \widehat{\mathbf{C}}_l$, for $k = 1, \ldots, r$, which are nested to each other. We refer to the generic problem in (2.12) together with the tuning process as *co-sparse unit-rank estimation* (CURE), i.e., CURE$(\mathbf{C}; \mathbf{Y}_k, \mathbf{X}, \mathbf{W}_k)$. Algorithm 1 and Figure 1 summarizes the proposed computation procedure, i.e., *sequential factor extraction via co-sparse unit-rank estimation* (SeCURE).

Figure 1: SeCURE: Sequential Factor Extraction via Co-Sparse Unit-Rank Estimation



Algorithm 1 Sequential Factor Extraction via Co-Sparse Unit-Rank Estimation (SeCURE)

Initialization: set $k = 1$, and set a desired rank $r \geq 1$.
repeat
(1) Obtain $(\widetilde{d}_k, \widetilde{\mathbf{u}}_k, \widetilde{\mathbf{v}}_k)$ from RRR, and construct $\mathbf{W}_k$ as in (2.11).
(2) Compute the current residual matrix $\mathbf{Y}_k$ as in (2.13).
(3) Perform the CURE$(\mathbf{C}; \mathbf{Y}_k, \mathbf{X}, \mathbf{W}_k)$ analysis via (2.12) (including the tuning process), and obtain $\widehat{\mathbf{C}}_k$ or equivalently $(\widehat{d}_k, \widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k)$.
if $\widehat{d}_k = 0$ **then**
Set $\widehat{d}_h = 0$ for any $k \leq h \leq r$; $k \leftarrow r + 1$.
else
$k \leftarrow k + 1$.
end if
until $k = r + 1$.
return $\widehat{\mathbf{C}}_k$ and $(\widehat{d}_k, \widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k)$ for all $k = 1, \ldots, r$ with $\widehat{d}_k \neq 0$.

2.4 Computation

2.4.1 Co-Sparse Unit-Rank Estimation Algorithm

For simplicity, we drop the index k and write the generic form of the problem in (2.12) as

$$\min_{\mathbf{C}} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_F^2 + \rho(\mathbf{C}; \mathbf{W}, \lambda) \right\}, \text{ s.t. } \operatorname{rank}(\mathbf{C}) \leq 1, \tag{2.14}$$

where $\rho(\mathbf{C}; \mathbf{W}, \lambda)$ is the adaptive elastic net penalty defined in (2.10). When $\lambda = 0$, the problem reduces to (2.8), which admits an explicit solution as given in Proposition 2.1.

For the general case $\lambda > 0$, we reformulate (2.14) by expressing $\mathbf{C} = d\mathbf{u}\mathbf{v}^{\mathrm{T}}$, with $d \geq 0$, $\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u}/n = 1$ and $\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1$, corresponding to (2.6). Then (2.14) is equivalent to

$$\begin{aligned} \min_{(d,\mathbf{u},\mathbf{v})} & \left\{ Q(d, \mathbf{u}, \mathbf{v}; \lambda) = \frac{1}{2}\|\mathbf{Y} - d\mathbf{X}\mathbf{u}\mathbf{v}^{\mathrm{T}}\|_F^2 + \rho(d\mathbf{u}\mathbf{v}^{\mathrm{T}}; \mathbf{W}, \lambda) \right\}, \\ & \text{s.t. } d \geq 0, \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} = n, \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1, \end{aligned} \tag{2.15}$$

where $\rho(d\mathbf{u}\mathbf{v}^{\mathrm{T}}; \mathbf{W}, \lambda) = \alpha\lambda \sum_{i=1}^{p} \sum_{j=1}^{q} w_{ij}|du_i v_j| + (1-\alpha)\lambda \sum_{i=1}^{p} \sum_{j=1}^{q} (du_i v_j)^2$. That is, $(\widehat{d}^{(\lambda)}, \widehat{\mathbf{u}}^{(\lambda)}, \widehat{\mathbf{v}}^{(\lambda)})$ solves (2.15) if and only if $\widehat{\mathbf{C}}^{(\lambda)} = \widehat{d}^{(\lambda)}\widehat{\mathbf{u}}^{(\lambda)}\widehat{\mathbf{v}}^{(\lambda)\mathrm{T}}$ solves (2.14).

We propose to solve (2.15) by a block coordinate descent algorithm with two overlapping blocks of parameters $(d, \mathbf{u})$ and $(d, \mathbf{v})$, motivated by Chen et al. [2012]. Consider

the update of $(d, \mathbf{u})$ with fixed $\mathbf{v}$ satisfying $\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1$. The relevant constraints are $\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} = n$ and $d \geq 0$. The key is to recognize that the objective function is a function of $(d, \mathbf{u})$ only through their product $\check{\mathbf{u}} = d\mathbf{u}$, so both constraints are avoided when optimizing with respect to $\check{\mathbf{u}}$. It boils down to solving the following problem with respect to $\check{\mathbf{u}}$ (referred to as the **U**-step):

$$\min_{\check{\mathbf{u}}} \left\{ Q_{\mathbf{u}}(\check{\mathbf{u}}; \mathbf{v}, \lambda) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}^{(u)}\check{\mathbf{u}}\|_2^2 + \lambda_1^{(u)} \sum_{i=1}^{p} w_i |\check{u}_i| + \lambda_2^{(u)} \sum_{i=1}^{p} \check{u}_i^2 \right\}, \tag{2.16}$$

where $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$, $\mathbf{X}^{(u)} = \mathbf{v} \otimes \mathbf{X}$, $\lambda_1^{(u)} = \alpha\lambda w^{(d)}(\sum_{j=1}^{q} w_j^{(v)}|v_j|)$, and $\lambda_2^{(u)} = (1 - \alpha)\lambda \sum_{j=1}^{q} v_j^2$. Here $\mathrm{vec}(\cdot)$ is the vectorization operator, and $\otimes$ denotes the Kronecker product. This problem is a convex elastic net regression, which can be efficiently solved. Once we obtain the solution $\check{\mathbf{u}}$, we update $d = \|\mathbf{X}\check{\mathbf{u}}\|_2/\sqrt{n}$ and $\mathbf{u} = \check{\mathbf{u}}/d$ to satisfy the constraints whenever $\check{\mathbf{u}} \neq \mathbf{0}$. Here we remark that $\check{\mathbf{u}} \neq \mathbf{0}$ implies that $\|\mathbf{X}\check{\mathbf{u}}\|_2 \neq \mathbf{0}$, because otherwise $Q_{\mathbf{u}}(\mathbf{0}; \mathbf{v}, \lambda) \leq Q_{\mathbf{u}}(\check{\mathbf{u}}; \mathbf{v}, \lambda)$, which contradicts with the optimality of $\check{\mathbf{u}}$. When $\check{\mathbf{u}} = \mathbf{0}$, $\mathbf{u}$ is no longer identifiable; we then update $d = 0$ and terminate the algorithm.

Similarly, for fixed $\mathbf{u}$ satisfying $\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} = n$, the minimization of (2.15) with respect to $(d, \mathbf{v})$ becomes minimization with respect to $\check{\mathbf{v}} = d\mathbf{v}$ (referred to as the **V**-step):

$$\min_{\check{\mathbf{v}}} \left\{ Q_{\mathbf{v}}(\check{\mathbf{v}}; \mathbf{u}, \lambda) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}^{(v)}\check{\mathbf{v}}\|_2^2 + \lambda_1^{(v)} \sum_{j=1}^{q} w_j |\check{v}_j| + \lambda_2^{(v)} \sum_{j=1}^{q} \check{v}_j^2 \right\}, \tag{2.17}$$

where $\mathbf{X}^{(v)} = \mathbf{I}_q \otimes (\mathbf{X}\mathbf{u})$, $\lambda_1^{(v)} = \alpha\lambda w^{(d)}(\sum_{i=1}^p w_i^{(u)}|u_i|)$, and $\lambda_2^{(v)} = (1-\alpha)\lambda\sum_{i=1}^p u_i^2$. Once $\check{\mathbf{v}}$ is obtained, we update $d = \|\check{\mathbf{v}}\|_2$ and $\mathbf{v} = \check{\mathbf{v}}/d$ whenever $\check{\mathbf{v}} \neq \mathbf{0}$. Similarly, when $\check{\mathbf{v}} = \mathbf{0}$, we update $d = 0$ and terminate the algorithm.

Both the **V**-step and **U**-step are a sub-case of a linear constrained adaptive elastic net problem, the solution of which is provided in Appendix A.1. Our CURE algorithm is summarized in Algorithm 2. Without loss of generality, we start from fixed $\mathbf{v} = \mathbf{v}_\lambda^0$. It generates a sequence $\{(\widetilde{d}_\lambda^s, d_\lambda^s, \mathbf{u}_\lambda^s, \mathbf{v}_\lambda^s)\}_{s\in\mathbb{N}}$ until reaching convergence.

Algorithm 2 Co-Sparse Unit-Rank Estimation Algorithm (CURE)

Initialization: Set $\lambda > 0$ and $0 < \alpha < 1$. Obtain initial value $\mathbf{v}_\lambda^0$. Set $s = 0$.
repeat
(1) **U**-step: Given $\mathbf{v} = \mathbf{v}_\lambda^s$, minimize objective function (2.16) to obtain $\check{\mathbf{u}}_\lambda^{s+1}$. Update $\widetilde{d}_\lambda^{s+1} = \|\mathbf{X}\check{\mathbf{u}}_\lambda^{s+1}\|_2/\sqrt{n}$ and $\mathbf{u}_\lambda^{s+1} = \check{\mathbf{u}}_\lambda^{s+1}/\widetilde{d}_\lambda^{s+1}$ when $\check{\mathbf{u}}_\lambda^{s+1} \neq \mathbf{0}$; otherwise return $d_\lambda^{s+1} = \widetilde{d}_\lambda^{s+1} = 0$ and terminate the algorithm.
(2) **V**-step: Given $\mathbf{u} = \mathbf{u}_\lambda^{s+1}$, minimize the objective function (2.17) to obtain $\check{\mathbf{v}}_\lambda^{s+1}$. Update $d_\lambda^{s+1} = \|\check{\mathbf{v}}_\lambda^{s+1}\|_2$ and $\mathbf{v}_\lambda^{s+1} = \check{\mathbf{v}}_\lambda^{s+1}/d_\lambda^{s+1}$ when $\check{\mathbf{v}}_\lambda^{s+1} \neq \mathbf{0}$; otherwise return $d_\lambda^{s+1} = 0$ and terminate the algorithm.
$s \leftarrow s + 1$.
until convergence, e.g., $\|\mathbf{C}_\lambda^{s+1} - \mathbf{C}_\lambda^s\|_F/\|\mathbf{C}_\lambda^s\|_F < \epsilon$, where $\mathbf{C}_\lambda^s = d_\lambda^s\mathbf{u}_\lambda^s\mathbf{v}_\lambda^{s\mathrm{T}}$ and $\epsilon = 10^{-4}$.
return $(\widehat{\mathbf{d}}^{(\lambda)}, \widehat{\mathbf{u}}^{(\lambda)}, \widehat{\mathbf{v}}^{(\lambda)})$ and $\widehat{\mathbf{C}}^{(\lambda)} = \widehat{d}^{(\lambda)}\widehat{\mathbf{u}}^{(\lambda)}\widehat{\mathbf{v}}^{(\lambda)\mathrm{T}}$.

2.4.2 Convergence Analysis and Generation of Solution Paths

Our CURE algorithm is closely connected to the Alternating Convex Search (ACS) method for biconvex optimization [Gorski et al., 2007b, Luenberger et al.]. The main difference between CURE and a standard ACS is that the two blocks of parameters in CURE, $(d, \mathbf{u})$ and $(d, \mathbf{v})$, are overlapping to each other. We are able to establish the

convergence behavior of CURE to some coordinatewise minimum point, as summarized in Theorem 2.3.

Theorem 2.3. *Consider the optimization problem in* (2.15) *with* $\lambda > 0$ *and* $0 < \alpha < 1$. *Assume the weights* $\{w_{ij}\}$ *and the data* $(\mathbf{Y}, \mathbf{X})$ *are finite, and the initial value* $\mathbf{v}_\lambda^0$ *satisfies* $\arg\min_{\check{\mathbf{u}}} Q_{\mathbf{u}}(\check{\mathbf{u}}; \mathbf{v}_\lambda^0, \lambda) \neq \mathbf{0}$, *i.e.,* $\widetilde{d}_\lambda^1 \neq 0$. *Then the sequence* $\{(d_\lambda^s, \mathbf{u}_\lambda^s, \mathbf{v}_\lambda^s)\}_{s\in\mathbb{N}}$ *generated by the CURE algorithm is uniformly bounded and has at least one accumulation point. Morever, all accumulation points are coordinatewise minimum points and have the same objective value, and* $Q(d_\lambda^s, \mathbf{u}_\lambda^s, \mathbf{v}_\lambda^s; \lambda)$ *converges monotonically to* $Q(d^{(\lambda)}, \mathbf{u}^{(\lambda)}, \mathbf{v}^{(\lambda)}; \lambda)$ *for some coordinatewise minimum point* $(d^{(\lambda)}, \mathbf{u}^{(\lambda)}, \mathbf{v}^{(\lambda)})$.

Proof. We acknowledge that our convergence proof mainly follows the framework in Gorski et al. [2007b].

In each step of SeCURE, we solve a CURE problem given by (2.14) which is equivalent to the biconvex optimization problem (2.15).

We may write $\boldsymbol{\theta} = (d, \mathbf{u}, \mathbf{v})$, $Q(d, \mathbf{u}, \mathbf{v}; \lambda) = Q(d, \mathbf{u}, \mathbf{v}) = Q(\boldsymbol{\theta})$ if no confusion arises. Denote the constrained parameter space for (2.15) as

$$\Omega = \{(d, \mathbf{u}, \mathbf{v}); d \geq 0, \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} = n, \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1\}.$$

It is easy to see that $\Omega = \Omega_d \times \Omega_{\mathbf{u}} \times \Omega_{\mathbf{v}}$, where $\Omega_d = \{d; d \geq 0\}$, $\Omega_{\mathbf{u}} = \{\mathbf{u}; \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} = n\}$ and $\Omega_{\mathbf{v}} = \{\mathbf{v}; \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1\}$. Recall that in each iteration of CURE in Algorithm 2, starting from $(d^s, \mathbf{u}^s, \mathbf{v}^s)$, we first update $(d, \mathbf{u})$ for fixed $\mathbf{v} = \mathbf{v}^s$ to obtain $(\tilde{d}^{s+1}, \mathbf{u}^{s+1}, \mathbf{v}^s)$, and

then update $(d, \mathbf{v})$ for fixed $\mathbf{u} = \mathbf{u}^{s+1}$ to obtain $(d^{s+1}, \mathbf{u}^{s+1}, v^{s+1})$. In the following, we consider the case that $\lambda > 0$ and $0 < \alpha < 1$, and assume the weights $\{w_{ij}\}$ and the data $(\mathbf{Y}, \mathbf{X})$ are finite, and the initial value $\mathbf{v}^0$ satisfies $\arg\min_{\check{\mathbf{u}}} Q(\check{\mathbf{u}}; \mathbf{v}^0, \lambda) \neq \mathbf{0}$, i.e., $\widetilde{d}^1 \neq 0$. (We have dropped the subscript λ in the above notations for convenience.) The order of the alternating updates is arbitrary, and all the results still apply for the reversed order.

Definition 2.4. *A feasible solution of* (2.15)*, denoted as* $(d^*, \mathbf{u}^*, \mathbf{v}^*) \in \Omega$*, is said to be a partial optimum if it satisfies*

$$\begin{cases} Q(d^*, \mathbf{u}^*, \mathbf{v}^*) \leq Q(d, \mathbf{u}, \mathbf{v}^*), & \forall (d, \mathbf{u}), (d, \mathbf{u}, \mathbf{v}^*) \in \Omega; \\ Q(d^*, \mathbf{u}^*, \mathbf{v}^*) \leq Q(d, \mathbf{u}^*, \mathbf{v}), & \forall (d, \mathbf{v}), (d, \mathbf{u}^*, \mathbf{v}) \in \Omega. \end{cases} \tag{2.18}$$

From the above definition, a partial optimal must also be a *coordinatewise minimum point*, as the conditions in (2.18) imply that

$$\begin{cases} Q(d^*, \mathbf{u}^*, \mathbf{v}^*) \leq Q(d^*, \mathbf{u}^*, \mathbf{v}), & \forall \mathbf{v}, (d^*, \mathbf{u}^*, \mathbf{v}) \in \Omega; \\ Q(d^*, \mathbf{u}^*, \mathbf{v}^*) \leq Q(d, \mathbf{u}^*, \mathbf{v}^*), & \forall d, (d, \mathbf{u}^*, \mathbf{v}^*) \in \Omega; \\ Q(d^*, \mathbf{u}^*, \mathbf{v}^*) \leq Q(d^*, \mathbf{u}, \mathbf{v}^*), & \forall \mathbf{u}, (d, \mathbf{u}, \mathbf{v}^*) \in \Omega. \end{cases} \tag{2.19}$$

Proposition 2.5. *Each iterative step in the CURE algorithm non-increases the objective in (2.15), i.e.,* $Q(d^{s+1}, \mathbf{u}^{s+1}, \mathbf{v}^{s+1}) \leq Q(\tilde{d}^{s+1}, \mathbf{u}^{s+1}, \mathbf{v}^s) \leq Q(d^s, \mathbf{u}^s, \mathbf{v}^s)$*, for all* $s \geq 0$*. Moreover, the objective function in* (2.15) *is bounded from below, so that the sequence* $\{Q(d^s, \mathbf{u}^s, \mathbf{v}^s)\}_{s \in \mathbb{N}}$ *generated by Algorithm 2 converges monotonically.*

The above result is obvious. We now analyze the behavior of the sequence $\{(d^s, \mathbf{u}^s, \mathbf{v}^s)\}_{s\in\mathbb{N}}$. We start by introducing an algorithmic map, as a point-to-set map, to characterize our CURE algorithm. For any set Ω, we denote its power set as 2^Ω, which is the set of all subsets of Ω.

Definition 2.6. *Let* $\mathbf{z}^s = (\tilde{d}^s, d^s, \mathbf{u}^s, \mathbf{v}^s) = (\tilde{d}^s, \boldsymbol{\theta}^s) \in \Omega_z$, *where* $\Omega_z = \Omega_d \times \Omega_d \times \Omega_{\mathbf{u}} \times \Omega_{\mathbf{v}}$. *The algorithmic map of the CURE algorithm* $\mathcal{A} : \Omega_z \to 2^{\Omega_z}$ *is defined by* $\mathbf{z}^{s+1} \in \mathcal{A}(\mathbf{z}^s)$, *if and only if*

$$\begin{cases} Q(\tilde{d}^{s+1}, \mathbf{u}^{s+1}, \mathbf{v}^s) \le Q(d, \mathbf{u}, \mathbf{v}^s), & \forall (d, \mathbf{u}), (d, \mathbf{u}, \mathbf{v}^s) \in \Omega; \\ Q(d^{s+1}, \mathbf{u}^{s+1}, \mathbf{v}^{s+1}) \le Q(d, \mathbf{u}^{s+1}, \mathbf{v}), & \forall (d, \mathbf{v}), (d, \mathbf{u}^{s+1}, v) \in \Omega. \end{cases} \tag{2.20}$$

Lemma 2.7. *The algorithmic map* $\mathcal{A}$ *is closed, i.e., if* $\mathbf{z}^s$ *and* $\mathbf{z}_1^s$ *are such that* $\mathbf{z}_1^s \in \mathcal{A}(\mathbf{z}^s)$ *with* $\lim_{s\to\infty} \mathbf{z}^s = \mathbf{z}^*$ *and* $\lim_{s\to\infty} \mathbf{z}_1^s = \mathbf{z}_1^*$, *then* $\mathbf{z}_1^* \in \mathcal{A}(\mathbf{z}^*)$.

Proof. Since $\mathbf{z}_1^s \in \mathcal{A}(\mathbf{z}^s)$ we have

$$\begin{cases} Q(\tilde{d}_1^s, \mathbf{u}_1^s, \mathbf{v}^s) \le Q(d, \mathbf{u}, \mathbf{v}^s), & \forall (d, \mathbf{u}), (d, \mathbf{u}, \mathbf{v}^s) \in \Omega; \\ Q(d_1^s, \mathbf{u}_1^s, \mathbf{v}_1^s) \le Q(d, \mathbf{u}_1^s, \mathbf{v}), & \forall (d, \mathbf{v}), (d, \mathbf{u}_1^s, \mathbf{v}) \in \Omega. \end{cases}$$

When $s \to \infty$, by the continuity of Q, we get

$$\begin{cases} Q(\tilde{d}_1^*, \mathbf{u}_1^*, \mathbf{v}^*) = \lim\limits_{s\to\infty} Q(\tilde{d}_1^s, \mathbf{u}_1^s, \mathbf{v}^s) \le \lim\limits_{s\to\infty} Q(d, \mathbf{u}, \mathbf{v}^s) = Q(d, \mathbf{u}, \mathbf{v}^*), & \forall (d, \mathbf{u}), (d, \mathbf{u}, \mathbf{v}^*) \in \Omega \\ Q(d_1^*, \mathbf{u}_1^*, \mathbf{v}_1^*) = \lim\limits_{s\to\infty} Q(d_1^s, \mathbf{u}_1^s, \mathbf{v}_1^s) \le \lim\limits_{s\to\infty} Q(d, \mathbf{u}_1^s, \mathbf{v}) = Q(d, \mathbf{u}_1^*, \mathbf{v}), & \forall (d, \mathbf{v}), (d, \mathbf{u}_1^*, \mathbf{v}) \in \Omega. \end{cases}$$

Therefore, $\mathbf{z}_1^* \in \mathcal{A}(\mathbf{z}^*)$. This completes the proof. □

Lemma 2.8. *Suppose the sequence* $\{\mathbf{z}^s\}_{s\in\mathbb{N}}$ *generated by the CURE algorithm converges to* $\mathbf{z}^* = (\tilde{d}^*, d^*, \mathbf{u}^*, \mathbf{v}^*)$. *Then* $\boldsymbol{\theta}^* = (d^*, \mathbf{u}^*, \mathbf{v}^*)$ *is a partial optimal of* (2.15).

Proof. The sequence $\{\mathbf{z}^s\}_{s\in\mathbb{N}}$ is convergent with limit point $\mathbf{z}^*$. Since the algorithmic map $\mathcal{A}$ is closed by Lemma 2.7 and $\mathbf{z}^{s+1} \in \mathcal{A}(\mathbf{z}^s)$ for all $s \in \mathbb{N}$, then $\mathbf{z}^* = \mathcal{A}(\mathbf{z}^*)$. It follows that

$$\begin{cases} Q(\tilde{d}^*, \mathbf{u}^*, \mathbf{v}^*) \leq Q(d, \mathbf{u}, \mathbf{v}^*), & \forall (d, \mathbf{u}), (d, \mathbf{u}, \mathbf{v}^*) \in \Omega; \\ Q(d^*, \mathbf{u}^*, \mathbf{v}^*) \leq Q(d, \mathbf{u}^*, \mathbf{v}), & \forall (d, \mathbf{v}), (d, \mathbf{u}^*, \mathbf{v}) \in \Omega. \end{cases}$$

Therefore, $\boldsymbol{\theta}^*$ is a partial optimal. □

Lemma 2.9. *The sequence* $\{\mathbf{z}^s\}_{s\in\mathbb{N}}$ *generated by the CURE algorithm is uniformly bounded.*

Proof. By the construction of the algorithm, $\{(\mathbf{u}^s, \mathbf{v}^s)\}$ are always bounded, because $\mathbf{u}^{s\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u}^s = n$ and $\mathbf{v}^{s\mathrm{T}}\mathbf{v}^s = 1$. It suffices to show that $\tilde{d}^s$ and d^s are bounded. We have $\tilde{d}^s = \|\mathbf{X}\check{\mathbf{u}}^s\|_2/\sqrt{n}$, with

$$\check{\mathbf{u}}^s = \arg\min_{\check{\mathbf{u}}} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}^{(u)}\check{\mathbf{u}}\|_2^2 + \alpha\lambda^{(u)} \sum_{i=1}^{p} w_i|\check{u}_i| + (1-\alpha)\lambda \sum_{i=1}^{p} \check{u}_i^2 \right\}, \tag{2.21}$$

where $\mathbf{X}^{(u)} = \mathbf{v}^{s-1} \otimes \mathbf{X}$ with $\|\mathbf{v}^{s-1}\|_2 = 1$, and $\lambda^{(u)} = \lambda w^{(d)}(\sum_{j=1}^{q} w_j^{(v)}|v_j^{s-1}|)$. To see $\tilde{d}^s$

is bounded, let

$$\begin{aligned}\check{\mathbf{u}}_R^s &= \arg\min_{\check{\mathbf{u}}} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}^{(u)}\check{\mathbf{u}}\|_2^2 + (1-\alpha)\lambda \sum_{i=1}^{p} \check{u}_i^2 \right\} \\ &= \{\mathbf{X}^{\mathrm{T}}\mathbf{X} + \sqrt{(1-\alpha)\lambda}\mathbf{I}\}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v}^{s-1},\end{aligned}$$

which corresponds to the solution of (2.21) when $\alpha = 0$. It follows that $\|\check{\mathbf{u}}^s\|_1 \leq \|\check{\mathbf{u}}_R^s\|_1$. Then

$$\begin{aligned}\tilde{d^s} &\leq \frac{d_1(\mathbf{X})}{\sqrt{n}}\|\check{\mathbf{u}}^s\|_2 \\ &\leq \frac{d_1(\mathbf{X})}{\sqrt{n}}\|\check{\mathbf{u}}^s\|_1 \\ &\leq \frac{d_1(\mathbf{X})}{\sqrt{n}}\|\check{\mathbf{u}}_R^s\|_1 \\ &\leq \frac{d_1(\mathbf{X})\sqrt{p}}{\sqrt{n}}\|\check{\mathbf{u}}_R^s\|_2 \\ &\leq \frac{d_1(\mathbf{X})\sqrt{p}}{\sqrt{n}}\|\{\mathbf{X}^{\mathrm{T}}\mathbf{X} + \sqrt{(1-\alpha)\lambda}\mathbf{I}\}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\|_F.\end{aligned}$$

For finite $(\mathbf{Y}, \mathbf{X})$, the right hand side is finite. Similarly, the uniform boundedness of $\{d^s\}$ can be shown. This completes the proof.

□

Lemma 2.10. *Let* $\mathbf{z}^{s+1} = \mathcal{A}(\mathbf{z}^s)$. *Assume* $\lambda > 0$, $0 < \alpha < 1$, *and the initial value* $\mathbf{v}^0$ *satisfies* $\arg\min_{\check{\mathbf{u}}} Q(\check{\mathbf{u}}; \mathbf{v}^0, \lambda) \neq \mathbf{0}$. *Then the optimal solutions of both the* $\mathbf{U}$*-step with* $\mathbf{v} = \mathbf{v}^s$ *and the* $\mathbf{V}$*-step with* $\mathbf{u} = \mathbf{u}^{s+1}$ *are unique. Moreover,* $\boldsymbol{\theta}^s \neq \boldsymbol{\theta}^{s+1}$ *implies that*

$Q(d^{s+1}, \mathbf{u}^{s+1}, \mathbf{v}^{s+1}) < Q(d^s, \mathbf{u}^s, \mathbf{v}^s)$.

Proof. The problem in either the **U**-step or the **V**-step are strictly convex in either $\check{\mathbf{u}}$ or $\check{\mathbf{v}}$ whenever $\lambda > 0$ and $0 < \alpha < 1$; consequently they always produce unique solutions $\check{\mathbf{u}}^s$ or $\check{\mathbf{v}}^s$ for $s \geq 1$. By Theorem 2.11, $\tilde{d}^1 \neq 0$ ensures that $\check{\mathbf{u}}^s \neq \mathbf{0}$ and $\check{\mathbf{v}}^s \neq \mathbf{0}$ for any $s \geq 1$. Therefore, the solutions in terms of the pair $(d, \mathbf{u})$ or $(d, \mathbf{v})$ are also always unique.

From $\mathbf{z}^{s+1} = \mathcal{A}(\mathbf{z}^s)$, it holds that

$$Q(\boldsymbol{\theta}^{s+1}) = Q(d^{s+1}, \mathbf{u}^{s+1}, \mathbf{v}^{s+1}) \leq Q(\tilde{d}^{s+1}, \mathbf{u}^{s+1}, \mathbf{v}^s) \leq Q(d^s, \mathbf{u}^s, \mathbf{v}^s) = Q(\boldsymbol{\theta}^s).$$

Now suppose $Q(d^{s+1}, \mathbf{u}^{s+1}, \mathbf{v}^{s+1}) = Q(\tilde{d}^{s+1}, \mathbf{u}^{s+1}, \mathbf{v}^s) = Q(d^s, \mathbf{u}^s, \mathbf{v}^s)$. Then due to the uniqueness of the solution in the **U**-step, we must have $(\tilde{d}^{s+1}, \mathbf{u}^{s+1}) = (d^s, \mathbf{u}^s)$. Similarly, due to the uniqueness of the solution in the **V**-step, we must have $(\tilde{d}^{s+1}, \mathbf{v}^s) = (d^{s+1}, \mathbf{v}^{s+1})$. It then follows that $\boldsymbol{\theta}^{s+1} = \boldsymbol{\theta}^s$. This completes the proof.

□

With the above results, we now prove Theorem 2.3. By Lemma 2.9, $\{\mathbf{z}^s\}_{s\in\mathbb{N}}$ is bounded and thus has at least one accumulation point $\mathbf{z}^*$. It follows that we have a convergent subsequence $\{\mathbf{z}^k\}_{k\in\mathbb{K}}$ with $\mathbb{K} \subset \mathbb{N}$ that converges to $\mathbf{z}^*$. Similarly, $\{\mathbf{z}^{k+1}\}_{k\in\mathbb{K}}$ has a convergent subsequence that converges to an accumulation point $\mathbf{z}^+$, and $\{\mathbf{z}^{k-1}\}_{k\in\mathbb{K}}$ has a convergent subsequence that converges to an accumulation point $\mathbf{z}^-$. By the closedness of the algorithmic map shown in Lemma 2.7, it holds that $\mathbf{z}^+ \in \mathcal{A}(\mathbf{z}^*)$ and $\mathbf{z}^* \in \mathcal{A}(\mathbf{z}^-)$. It follows from Proposition 2.5 that $Q(\boldsymbol{\theta}^+) = Q(\boldsymbol{\theta}^*) = Q(\boldsymbol{\theta}^-)$. Now suppose

$\boldsymbol{\theta}^*$ is not a partial optimal, then by Lemmas 2.8 and 2.10, $Q(\boldsymbol{\theta}^+) < Q(\boldsymbol{\theta}^*) < Q(\boldsymbol{\theta}^-)$, which leads to contradiction. Therefore, $\boldsymbol{\theta}^*$ must be a partial optimal. The proof is completed.

□

Due to the nonconvex nature of the problem, Theorem 2.3 does not guarantee any global optimality of the solution. Nevertheless, the established convergence results are adequate to ensure stable computation for practical use [Luenberger et al., Bunea et al., 2012].

In Theorem 2.3, it is required that the initial value $\mathbf{v}_\lambda^0$ does not produce a null solution in the first step of the algorithm. The reasons and implications are as follows. One key to establish the convergence results is to ensure that each sub-problem in either (2.16) or (2.17) always produces a unique solution in terms of the pair $(d, \mathbf{u})$ or $(d, \mathbf{v})$ along the iterations. Owing to the use of the elastic net penalty, these sub-problems are strictly convex in either $\check{\mathbf{u}} = d\mathbf{u}$ or $\check{\mathbf{v}} = d\mathbf{v}$ whenever $\lambda > 0$ and $0 < \alpha < 1$; consequently they always produce unique solutions of $d\mathbf{u}$ or $d\mathbf{v}$. However, the complication occurs when $\check{\mathbf{u}} = \mathbf{0}$ or $\check{\mathbf{v}} = \mathbf{0}$, where the solution in terms of the pair $(d, \mathbf{u})$ or $(d, \mathbf{v})$ are not unique anymore. Therefore, the occurrence of the null solution during iterations needs special care. Nicely, we are able to show that as long as the initial value $\mathbf{v}_\lambda^0$ does not produce a null solution, it is guaranteed that the null solution will not occur later along the iterations. Moreover, suppose $d^{(\lambda)} \neq 0$ where $(d^{(\lambda)}, \mathbf{u}^{(\lambda)}, \mathbf{v}^{(\lambda)})$ is the solution of (2.15) with tuning parameter λ. Then we show that when solving (2.15) with a smaller tuning

parameter than λ, the null solution can again be avoided by using $\mathbf{v}^{(\lambda)}$ as the initial value. Theorem 2.11 summarizes these results.

Theorem 2.11. *Consider solving* (2.15) *with* $\lambda > 0$ *and* $0 < \alpha < 1$ *using the CURE algorithm. If* $\arg\min_{\check{\mathbf{u}}} Q_{\mathbf{u}}(\check{\mathbf{u}}; \mathbf{v}_\lambda^0, \lambda) \neq \mathbf{0}$*, i.e.,* $\widetilde{d}_\lambda^1 \neq 0$*, then* $\widetilde{d}_\lambda^s \neq 0$ *and* $d_\lambda^s \neq 0$ *for any* $s \geq 1$*. Let* $(d^{(\lambda)}, \mathbf{u}^{(\lambda)}, \mathbf{v}^{(\lambda)})$ *denote the solution of* (2.15) *with tuning parameter* λ*. If* $d^{(\lambda_1)} \neq 0$*, then* $d^{(\lambda_2)} \neq 0$ *for any* $\lambda_2 \leq \lambda_1$*. Moreover, if* $d^{(\lambda_1)} \neq 0$*, then setting* $\mathbf{v}_{\lambda_2}^0 = \mathbf{v}^{(\lambda_1)}$ *ensures that* $\widetilde{d}_{\lambda_2}^s \neq 0$ *and* $d_{\lambda_2}^s \neq 0$ *for any* $s \geq 1$*.*

Proof. We assume $\lambda > 0$, $0 < \alpha < 1$, and $\arg\min_{\check{\mathbf{u}}} Q(\check{\mathbf{u}}; \mathbf{v}_\lambda^0, \lambda) \neq \mathbf{0}$, i.e., $\widetilde{d}_\lambda^1 \neq 0$. Let after some **U**-step, $\widetilde{d}_\lambda^{s+1} = 0$. Then $Q(0, \mathbf{u}^{s+1}, \mathbf{v}_\lambda^s; \lambda) \leq Q(d, \mathbf{u}, \mathbf{v}_\lambda^s; \lambda)$, for any $\mathbf{u}^{s+1}$ and any $(d, \mathbf{u})$. It follows that $Q(0, \mathbf{u}_\lambda^s, \mathbf{v}_\lambda^s; \lambda) \leq Q(d_\lambda^s, \mathbf{u}_\lambda^s, \mathbf{v}_\lambda^s; \lambda) \leq Q(d, \mathbf{u}_\lambda^s, \mathbf{v})$ for any $(d, \mathbf{v})$. Therefore, it must be true that $d_\lambda^s = 0$ in the previous step, by the optimality of the **V**-step. Similarly, assume after some **V**-step, $d_\lambda^{s+1} = 0$. Then $Q(0, \mathbf{u}_\lambda^{s+1}, \mathbf{v}^{s+1}; \lambda) \leq Q(d, \mathbf{u}_\lambda^{s+1}, \mathbf{v}; \lambda)$, for any $\mathbf{v}^{s+1}$ and any $(d, \mathbf{v})$. It then follows that $Q(0, \mathbf{u}_\lambda^{s+1}, \mathbf{v}_\lambda^s; \lambda) \leq Q(\widetilde{d}_\lambda^{s+1}, \mathbf{u}_\lambda^{s+1}, \mathbf{v}_\lambda^s; \lambda) \leq Q(d, \mathbf{u}, \mathbf{v}_\lambda^s; \lambda)$ for any $(d, \mathbf{u})$. Therefore, it must be true that $\widetilde{d}_\lambda^{s+1} = 0$, by the optimality of the **U**-step. By induction, whenever $\widetilde{d}_\lambda^{s+1} = 0$ or $d_\lambda^{s+1} = 0$ for any $s \geq 0$, it must be true that $\widetilde{d}_\lambda^1 = 0$. This completes the first part of the proof.

By $d^{(\lambda_1)} \neq 0$, we have $Q(d^{(\lambda_1)}, \mathbf{u}^{(\lambda_1)}, \mathbf{v}^{(\lambda_1)}; \lambda_1) \leq Q(0, \mathbf{u}^{(\lambda_1)}, \mathbf{v}^{(\lambda_1)}; \lambda_1) = \|\mathbf{Y}\|_F^2/2$. It follows that $Q(d^{(\lambda_1)}, \mathbf{u}^{(\lambda_1)}, \mathbf{v}^{(\lambda_1)}; \lambda_2) \leq Q(d^{(\lambda_1)}, \mathbf{u}^{(\lambda_1)}, \mathbf{v}^{(\lambda_1)}; \lambda_1) \leq \|\mathbf{Y}\|_F^2/2 = Q(0, \mathbf{u}, \mathbf{v}; \lambda_2)$, for any $(\mathbf{u}, \mathbf{v})$. Therefore, $(d^{(\lambda_1)}, \mathbf{u}^{\lambda_1}, \mathbf{v}^{\lambda_1})$ can achieve a smaller objective value than the zero solution when the tuning parameter becomes λ_2. That is, (2.15)

with tuning parameter λ_2 can not have a null solution. In particular, when setting $\mathbf{v}^0_{\lambda_2} = \mathbf{v}^{(\lambda_1)}$, we have $Q(d^{(\lambda_1)}, \mathbf{u}^{(\lambda_1)}, \mathbf{v}^0_{\lambda_2}; \lambda_2) \leq Q(0, \mathbf{u}, \mathbf{v}^0_{\lambda_2}; \lambda_2)$ for any $\mathbf{u}$, so that $\widetilde{d}^0_{\lambda_2} \neq 0$. The result then follows using the first part.

□

It is now in order to consider the generation of the solution paths using CURE. Theorem 2.11 clearly suggests to fit the model for a sequence of descending λ values with a warm-start strategy, i.e., the estimate from the previous larger λ value is used as the initial value for the next smaller λ value. It then remains to determine the relevant range of λ values, for which it boils down to determine $\lambda_{\max}$, the smallest λ value for which (2.15) produces the null solution. Nicely, similar to Chen et al. [2012], both $\lambda_{\max}$ and the first set of non-zero solution on the paths can be explicitly determined. Denote $\mathbf{Y} = [\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(q)}]$ and $\mathbf{X} = [\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(p)}]$, and let $(i^0, j^0) = \arg\max_{(i,j)}\{|\mathbf{x}^{\mathrm{T}}_{(i)}\mathbf{y}_{(j)}|/w_{ij}\}$, assuming no ties. Then, for any $0 < \alpha \leq 1$, we have $\lambda_{\max} = |\mathbf{x}^{\mathrm{T}}_{(i^0)}\mathbf{y}_{(j^0)}|/(\alpha w_{i^0 j^0})$, and the initial $\mathbf{v}$ vector, denoted as $\mathbf{v}^0 \in \mathbb{R}^q$, is a standard basis vector with $v^0_{j^0} = 1$ and $v^0_j = 0$ for any $j \neq j^0$.

In practice, we fit the model for a sequence of 100 descending λ values equally spaced on the log-scale between $[\lambda_{\max} - \epsilon, \lambda_{\min}]$, to produce a spectrum of co-sparsity patterns. Here ϵ is a small positive number, and $\lambda_{\min}$ is taken as either 0 or a fraction of $\lambda_{\max}$ at which the model has excessive number of non-zero coefficients (we set $\lambda_{\min} = \lambda_{\max} \times 10^{-5}$). With the above strategies, for each fixed λ within the range, we do not observe any early termination of the algorithm due to the production of null solution,

and we always observe a unique limit point. The computation of the solution paths is efficient and stable.

2.4.3 Tuning and Rank Selection

For the genetic problem in (2.14) or (2.15), once the solution path is obtained, we need to choose the optimal solution along the path. For small-scale problems, λ can be chosen by cross validation [Stone, 1974]. Alternatively, various information criteria have been widely used due to their computational efficiency. In all our numerical studies, we use the following hybrid information criterion because of its superior performance in sparse learning,

$$\mathrm{IC}(\lambda) = \begin{cases} \log\{\mathrm{SSE}(\lambda)\} + \{\log(qn)/(nq)\} df(\lambda) & \text{when } p < n; \\ \log(\mathrm{SSE}(\lambda)) + \{\log\log(nq)\log(pq)/(nq)\} df(\lambda) & \text{when } p \geq n, \end{cases}$$

where $\mathrm{SSE}(\lambda) = \|\mathbf{Y} - \widehat{d}^{(\lambda)}\mathbf{X}\widehat{\mathbf{u}}^{(\lambda)}\widehat{\mathbf{v}}^{(\lambda)\mathrm{T}}\|_F^2$, and $df(\lambda)$ is the model degrees of freedom, which is estimated by $\widehat{df}(\lambda) = \sum_{i=1}^{p} \mathrm{I}(\widehat{u}_i^{(\lambda)} \neq 0) + \sum_{j=1}^{q} \mathrm{I}(\widehat{v}_j^{(\lambda)} \neq 0) - 1$ with $\mathrm{I}(\cdot)$ being the indicator function. The criterion corresponds to the familiar Bayesian Information Criterion (BIC) when $p < n$ and the generalized information criterion proposed by Fan and Tang [2013] when $p \geq n$. The SeCURE does not require the rank or the number of factors to be specified in advance. The sequential unit-rank extraction proceeds until $\widehat{d}_k = 0$ for some k.

2.4.4 Extensions to Incomplete Data and Exact Orthogonality

Here we explore some useful extensions of SeCURE. Let $\mathcal{H} = \{(i,j); y_{ij} \text{ is observed }, i = 1,\ldots,n, j = 1,\ldots,q.\}$ be an index set collecting the indices of all the observed entries in $\mathbf{Y}$. Let $\widetilde{\mathbf{Y}} = P_{\mathcal{H}}(\mathbf{Y})$ be the projection of $\mathbf{Y}$ onto $\mathcal{H}$ such that $\widetilde{y}_{ij} = y_{ij}$ for any $(i,j) \in \mathcal{H}$ and $\widetilde{y}_{ij} = 0$ otherwise. We extend SeCURE by modifying (2.12) as

$$\min_{\mathbf{C}} \left\{ \frac{1}{2} \|P_{\mathcal{H}}(\mathbf{Y}_k) - P_{\mathcal{H}}(\mathbf{X}\mathbf{C})\|_F^2 + \rho(\mathbf{C}; \mathbf{W}_k, \lambda) \right\}, \text{ s.t. } \operatorname{rank}(\mathbf{C}) \leq 1, \mathbf{A}_k\mathbf{C} = \mathbf{0}, \mathbf{B}_k\mathbf{C}^{\mathrm{T}} = \mathbf{0}.$$

where $\mathbf{A}_1 = \mathbf{0}$, $\mathbf{B}_1 = \mathbf{0}$ and for $k \geq 2$, $\mathbf{A}_k$ and $\mathbf{B}_k$ are formed from the estimated factors as $\mathbf{A}_k = \widehat{\mathbf{U}}_{k-1}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\mathbf{B}_k = \widehat{\mathbf{V}}_{k-1}^{\mathrm{T}}$, with $\widehat{\mathbf{U}}_{k-1} = [\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_{k-1}]$ and $\widehat{\mathbf{V}}_{k-1} = [\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_{k-1}]$, and all the other terms are defined the same as before. Only the observed entries contribute to the loss function, which follows the same spirit as matrix completion [Candès and Recht, 2009]. The linear constraints enforce the orthogonality among the estimated factors.

Accordingly, the generic CURE problem in (2.15) can be modified to the form

$$\min_{d,\mathbf{u},\mathbf{v}} \left\{ \frac{1}{2} \|P_{\mathcal{H}}(\mathbf{Y}) - P_{\mathcal{H}}(d\mathbf{X}\mathbf{u}\mathbf{v}^{\mathrm{T}})\|_F^2 + \rho(d\mathbf{u}\mathbf{v}^{\mathrm{T}}; \mathbf{W}, \lambda, \alpha) \right\},$$
$$\text{s.t. } d \geq 0, \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} = n, \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1; \mathbf{A}\mathbf{u} = \mathbf{0}, \mathbf{B}\mathbf{v} = \mathbf{0}.$$

Nicely, the CURE optimization can still be conducted via the block coordinate descent algorithm. The problem in either the **U**-step or the **V**-step becomes a linear-constrained adaptive elastic net regression, which remains convex and admits efficient computation [James et al., 2013, Lin et al., 2014b]; see Appendix A.1 for solution. Implementations of the proposed methods are publicly available in the R package `secure` [R Development Core Team, 2014], which can be accessed at https://CRAN.R-project.org/package=secure.

2.5 Theoretical Properties

2.5.1 Asymptotic Results

In our asymptotic analysis, we consider the model in (2.5) and (2.6) with fixed p, q. The following assumptions on the design, the random error, and the singular values are used.

A1. $\|(1/n)\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{\Gamma}\|_F = O_p(1/\sqrt{n})$, where $\mathbf{\Gamma}$ is a fixed, positive definite matrix.

A2. The error vectors $\mathbf{e}_i$, $i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) normal random variables with $\mathrm{E}(\mathbf{e}_i) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{e}_i) = \mathbf{\Sigma}$, a positive definite matrix.

A3. $d_1^* > \ldots > d_{r^*}^* > 0$.

First, Proposition 2.12 shows that the sequential RRR estimators are consistent. This justifies our approach of using them for constructing the adaptive weights.

Proposition 2.12. *[Reinsel and Velu, 1998] Consider the sequential reduced-rank estimators* $(\widetilde{d}_k, \widetilde{\mathbf{u}}_k, \widetilde{\mathbf{v}}_k)$, $k = 1, \ldots, r$, *presented in Proposition 2.1. Under conditions* ***A1****–****A3****, as* $n \to \infty$,

i. $\sqrt{n}(\widetilde{d}_k - d_k^*) = O_p(1)$, $\sqrt{n}(\widetilde{\mathbf{u}}_k - \mathbf{u}_k^*) = O_p(1)$ *and* $\sqrt{n}(\widetilde{\mathbf{v}}_k - \mathbf{v}_k^*) = O_p(1)$, *for* $k = 1, \ldots, r^*$.

ii. $\sqrt{n}\widetilde{d}_k = O_p(1)$, *for* $k = r^* + 1, \ldots, r$.

Recall that we have parameterized the true coefficient matrix $\mathbf{C}^*$ in (2.6); we have also written $\mathbf{C}_k^* = d_k^* \mathbf{u}_k^* \mathbf{v}_k^{*\mathrm{T}}$ and $\mathbf{C}^* = \sum_{k=1}^{r^*} \mathbf{C}_k^*$. To simplify the problem, we consider a reparameterization of $\mathbf{C}^*$, by letting each singular value d_k^* be absorbed to each pair of vectors $(\mathbf{u}_k^*, \mathbf{v}_k^*)$. Specifically, for each $\mathbf{v}_k^*$, $k = 1, \ldots, r^*$, there must exist at least one entry that is nonzero, i.e., $v_{\ell_k k}^* \neq 0$ for some $1 \leq \ell_k \leq q$. Accordingly, we reparameterize $\mathbf{C}_k^*$ as $\mathbf{C}_k^* = \mathbf{u}_k^* \mathbf{v}_k^{*\mathrm{T}}$, with $v_{\ell_k k}^* = 1$. Here, with some abuse of notation, $(\mathbf{u}_k^*, \mathbf{v}_k^*)$ have been re-defined, and they are re-scaled versions of their original counterparts, i.e., the new pair now satisfies $(\mathbf{u}_k^{*\mathrm{T}} \boldsymbol{\Gamma} \mathbf{u}_k^*)(\mathbf{v}_k^{*\mathrm{T}} \mathbf{v}_k^*) = d_k^*$, $k = 1, \ldots, r^*$. Consequently, we parameterize $\mathbf{C}^*$ as

$$\mathbf{C}^* = \mathbf{U}^* \mathbf{V}^{*\mathrm{T}}, \text{ s.t. } \mathbf{U}^{*\mathrm{T}} \boldsymbol{\Gamma} \mathbf{U}^* \text{ and } \mathbf{V}^{*\mathrm{T}} \mathbf{V}^* \text{ are diagonal}; v_{\ell_k k}^* = 1, k = 1, \ldots, r^*. \quad (2.22)$$

Now all the elements in $\mathbf{U}^*$ and $\mathbf{V}^*$ are uniquely defined and hence identifiable. For $k > r^*$, it is obvious that $\mathbf{C}_k^* = \mathbf{0}$ and its corresponding singular vectors become unidentifiable; as such, we can set $\mathbf{u}_k^* = \mathbf{0}$ but still choose $\mathbf{v}_k^*$ to be a nonzero vector with some

$v^*_{\ell_k k} = 1$.

For the sequential reduced-rank estimators, let $\widetilde{d}_k$ be absorbed to $(\widetilde{\mathbf{u}}_k, \widetilde{\mathbf{v}}_k)$ similar to (2.22), such that $\widetilde{\mathbf{C}}_k = \widetilde{d}_k \widetilde{\mathbf{u}}_k \widetilde{\mathbf{v}}_k^{\mathrm{T}}$ in Proposition 2.1 now becomes $\widetilde{\mathbf{C}}_k = \widetilde{\mathbf{u}}_k \widetilde{\mathbf{v}}_k^{\mathrm{T}}$ with $\widetilde{v}_{\ell_k k} = 1$. With some abuse of notation, here we have still used $(\widetilde{\mathbf{u}}_k, \widetilde{\mathbf{v}}_k)$ to denote the rescaled versions. Then, as $n \to \infty$, the consistency results in Proposition 2.12 are alternatively expressed as

i. $\sqrt{n}(\widetilde{\mathbf{u}}_k - \mathbf{u}^*_k) = O_p(1)$ and $\sqrt{n}(\widetilde{\mathbf{v}}_k - \mathbf{v}^*_k) = O_p(1)$, for $k = 1, \ldots, r^*$.

ii. $\sqrt{n}\widetilde{d}_k = O_p(1)$, where $\widetilde{d}_k = (1/n)(\widetilde{\mathbf{u}}_k^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \widetilde{\mathbf{u}}_k)(\widetilde{\mathbf{v}}_k^{\mathrm{T}} \widetilde{\mathbf{v}}_k)$, for $k = r^* + 1, \ldots, r$.

Now, consider the proposed SeCURE approach. Under the new parameterization in (2.22), the objective function in the kth step of SeCURE can be written as

$$Q_k^{(n)}(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \|\mathbf{Y}_k - \mathbf{X}\mathbf{u}\mathbf{v}^{\mathrm{T}}\|_F^2 + \alpha \lambda_k^{(n)} \sum_{i=1}^{p} \sum_{j=1}^{q} w_{ijk} |u_i v_j| + (1 - \alpha) \lambda_k^{(n)} \sum_{i=1}^{p} \sum_{j=1}^{q} (u_i v_j)^2, \tag{2.23}$$

where $\mathbf{u} \in \mathbb{R}^p$, $\mathbf{v} \in \mathbb{R}^q$ with $v_{\ell_k} = 1$, and $\mathbf{Y}_1 = \mathbf{Y}$, $\mathbf{Y}_k = \mathbf{Y} - \sum_{1 \le h \le k-1} \mathbf{X} \widehat{\mathbf{u}}_h \widehat{\mathbf{v}}_h^{\mathrm{T}}$, where $(\widehat{\mathbf{u}}_h, \widehat{\mathbf{v}}_h) = \arg\min Q_h^{(n)}(\mathbf{u}, \mathbf{v})$. Here $w_{ijk} = w_{ik} w_{jk}$, where $w_{ik} = |\widetilde{u}_{ik}|^{-\gamma}$ and $w_{jk} = |\widetilde{v}_{jk}|^{-\gamma}$ with some fixed $\gamma > 0$. Here $\alpha \in (0, 1)$ is considered as a fixed constant.

We show that our SeCURE estimators enjoy consistency, asymptotic normality, variable selection consistency and rank selection consistency. Our main results are summarized in Theorems 2.13–2.15 below. The following assumption on the choice of $\lambda_k^{(n)}$ is required.

A4. $\frac{\lambda_k^{(n)}}{\sqrt{n}} \to 0$ and $\frac{\lambda_k^{(n)}}{\sqrt{n}} n^{\gamma/2} \to \infty$ as $n \to \infty$, for $k = 1, \ldots, r^*$. Also, $\frac{\lambda_k^{(n)}}{\sqrt{n}} n^{\gamma/2} \to \infty$ as $n \to \infty$, for $k = r^* + 1, \ldots, r$.

Theorem 2.13. *(Existence of Local Minimum). Suppose **A1**–**A3** are satisfied, and* $\lambda_k^{(n)}/\sqrt{n} \to \lambda_k \geq 0$ *as* $n \to \infty$. *Then there is a local minimizer* $(\widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k)$ *of* $Q_k^{(n)}(\mathbf{u}, \mathbf{v})$, *such that*

i. $\|\widehat{\mathbf{u}}_k - \mathbf{u}_k^*\| = O_p(n^{-1/2})$ *and* $\|\widehat{\mathbf{v}}_k - \mathbf{v}_k^*\| = O_p(n^{-1/2})$, *for* $k = 1, \ldots, r^*$.

ii. $|\widehat{d}_k| = O_p(n^{-1/2})$ *where* $\widehat{d}_k = (1/n)(\widehat{\mathbf{u}}_k^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \widehat{\mathbf{u}}_k)(\widehat{\mathbf{v}}_k^{\mathrm{T}} \widehat{\mathbf{v}}_k)$, *for* $k = r^* + 1, \ldots, r$.

Proof. Before we prove the result, let us define some notations. Suppose $\mathbf{Z}$ is an arbitrary matrix, and $\mathcal{A}$ and $\mathcal{B}$ are subsets of the collection of row and column indices of $\mathbf{Z}$, respectively. We let $\mathbf{Z}_{\mathcal{A}\mathcal{B}}$ denote a sub-matrix of $\mathbf{Z}$ whose rows and columns are chosen from $\mathbf{Z}$ according to the index sets $\mathcal{A}$ and $\mathcal{B}$, respectively. For simplicity, we write $\mathbf{Z}_{\mathcal{A}\mathcal{A}} = \mathbf{Z}_{\mathcal{A}}$ when $\mathbf{Z}$ is a square matrix, $\mathbf{Z}_{\mathcal{A}\mathcal{B}} = \mathbf{Z}_{\cdot\mathcal{B}}$ ($\mathbf{Z}_{\mathcal{A}\cdot}$) when $\mathcal{A}$ ($\mathcal{B}$) consists of all the row (column) indices, and $\mathbf{Z}_{\mathcal{A}\cdot} = \mathbf{Z}_{\mathcal{A}}$ when $\mathbf{Z}$ is a vector. For the kth column of matrix $\mathbf{Z}$ given by $\mathbf{z}_k$, we represent element in set $\mathcal{A}$ by notation $\mathbf{z}_{k\mathcal{A}}$.

Now, define

$$\Omega_k = \left\{ \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}} : \mathbf{u}_k \in \mathbb{R}^p \text{ and } \mathbf{v}_k \in \mathbb{R}^q \text{ with } v_{\ell_k k} = 1 \right\}.$$

We first prove the result for $k = 1$. Consider a neighborhood of $\mathbf{C}_1^* = \mathbf{u}_1^* \mathbf{v}_1^{*\mathrm{T}}$ of radius

$h > 0$,

$$\mathcal{N}(\mathbf{C}_1^*, h) = \left\{ (\mathbf{u}_1^* + \mathbf{a}/\sqrt{n})(\mathbf{v}_1^* + \mathbf{b}/\sqrt{n})^{\mathrm{T}}; \|\mathbf{\Gamma}^{1/2}\mathbf{a}\| \leq h, \mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^q, \|\mathbf{b}\| \leq h, b_{\ell_1} = 0 \right\}.$$

We claim that for any $\epsilon > 0$, there exists a large enough h such that

$$P\left\{ \inf_{\|\Gamma^{1/2}\mathbf{a}\| = \|\mathbf{b}\| = h} Q_1^{(n)}(\mathbf{u}_1^* + \mathbf{a}/\sqrt{n}, \mathbf{v}_1^* + \mathbf{b}/\sqrt{n}) > Q_1^{(n)}(\mathbf{u}_1^*, \mathbf{v}_1^*) \right\} \geq 1 - \epsilon. \tag{2.24}$$

The claim implies that with probability at least $1 - \epsilon$ there exists local minimum $\widehat{\mathbf{u}}_1 \widehat{\mathbf{v}}_1^{\mathrm{T}}$ in the interior of $\mathcal{N}(\mathbf{C}_1^*, h)$, i.e. there exists a local minimizer such that $\|\widehat{\mathbf{u}}_1 - \mathbf{u}_1^*\| = O_p(n^{-1/2})$ and $\|\widehat{\mathbf{v}}_1 - \mathbf{v}_1^*\| = O_p(n^{-1/2})$.

It remains to verify (2.24). Define

$$\Psi_1^{(n)}(\mathbf{a}, \mathbf{b}) = Q_1^{(n)}(\mathbf{u}_1^* + \mathbf{a}/\sqrt{n}, \mathbf{v}_1^* + \mathbf{b}/\sqrt{n}) - Q_1^{(n)}(\mathbf{u}_1^*, \mathbf{v}_1^*) = T_1 + T_2 + T_3, \tag{2.25}$$

where

$$\left.\begin{aligned} T_1 &= \frac{1}{2}\|\mathbf{Y} - \mathbf{X}(\mathbf{u}_1^* + \mathbf{a}/\sqrt{n})(\mathbf{v}_1^* + \mathbf{b}/\sqrt{n})^{\mathrm{T}}\|_F^2 - \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{u}_1^*\mathbf{v}_1^{*\mathrm{T}}\|_F^2, \\ T_2 &= \alpha\lambda_1^{(n)} \left\{ \|\mathbf{W}_1 \circ (\mathbf{u}_1^* + \mathbf{a}/\sqrt{n})(\mathbf{v}_1^* + \mathbf{b}/\sqrt{n})^{\mathrm{T}}\|_1 - \|\mathbf{W}_1 \circ \mathbf{u}_1^*\mathbf{v}_1^{*\mathrm{T}}\|_1 \right\}, \\ T_3 &= (1 - \alpha)\lambda_1^{(n)} \{ \|(\mathbf{u}_1^* + \mathbf{a}/\sqrt{n})(\mathbf{v}_1^* + \mathbf{b}/\sqrt{n})^{\mathrm{T}}\|_F^2 - \|\mathbf{u}_1^*\mathbf{v}_1^{*\mathrm{T}}\|_F^2 \}. \end{aligned}\right\}. \tag{2.26}$$

To simplify T_1, we write $T_1 = T_{11} + T_{12}$, where

$$
\begin{aligned}
T_{11} &= -\operatorname{tr}\left\{(\mathbf{Y} - \mathbf{X}\mathbf{u}_1^*\mathbf{v}_1^{*\mathrm{T}})^{\mathrm{T}}\frac{\mathbf{X}}{\sqrt{n}}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})\right\} \\
&= -\operatorname{tr}\left\{(\sum_{l=2}^{r^*}\mathbf{X}\mathbf{u}_l^*\mathbf{v}_l^{*\mathrm{T}} + \mathbf{E})^{\mathrm{T}}\frac{\mathbf{X}}{\sqrt{n}}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})\right\} \\
&= -\operatorname{tr}\left\{\frac{\mathbf{E}^{\mathrm{T}}\mathbf{X}}{\sqrt{n}}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})\right\} - \sum_{l=2}^{r^*}\operatorname{tr}\left\{\mathbf{v}_l^*\mathbf{u}_l^{*\mathrm{T}}\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{\sqrt{n}}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})\right\}, \\
T_{12} &= \operatorname{tr}\left\{(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})^{\mathrm{T}}\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})\right\}.
\end{aligned}
$$

Using the orthogonality of the true parameters in (2.22), we have $\mathbf{u}_l^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_1^* = 0$ and $\operatorname{tr}(\mathbf{v}_l^*\mathbf{u}_l^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{a}\mathbf{v}_1^{*\mathrm{T}}) = \operatorname{tr}(\mathbf{v}_1^{*\mathrm{T}}\mathbf{v}_l^*\mathbf{u}_l^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{a}) = 0$ for any $l > 1$. Thus, for large n,

$$
\left.
\begin{aligned}
T_{11} &= -\operatorname{tr}\left\{\frac{\mathbf{E}^{\mathrm{T}}\mathbf{X}}{\sqrt{n}}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}})\right\} - \sum_{l>1}^{r^*}\mathbf{a}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_l^*\mathbf{b}^{\mathrm{T}}\mathbf{v}_l^* + O_p(\frac{1}{\sqrt{n}}), \\
T_{12} &= \operatorname{tr}\left\{(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}})^{\mathrm{T}}\mathbf{\Gamma}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}})\right\} + O_p(\frac{1}{\sqrt{n}}).
\end{aligned}
\right\} \tag{2.27}
$$

Similar to Chen et al. [2012], it is seen that as $n \to \infty$,

$$
T_2 \geq \alpha\frac{\lambda_1^{(n)}}{\sqrt{n}}\sum_i\sum_j w_{ij1}\operatorname{sign}(u_{i1}^*v_{j1}^*)(u_{i1}^*b_j + a_iv_{j1}^*) + O_p(\frac{1}{\sqrt{n}}), \tag{2.28}
$$

$$
T_3 = (1-\alpha)\frac{\lambda_1^{(n)}}{\sqrt{n}}\sum_i\sum_j 2u_{i1}^*v_{j1}^*(u_{i1}^*b_j + a_iv_{j1}^*) + O_p(\frac{1}{\sqrt{n}}). \tag{2.29}
$$

Now combining (2.26)–(2.29), we have

$$\Psi_1^{(n)}(\mathbf{a},\mathbf{b}) \geq -\mathbf{z}^{\mathrm{T}}\mathrm{vec}(\frac{\mathbf{X}^{\mathrm{T}}\mathbf{E}}{\sqrt{n}}) + \frac{1}{2}\mathbf{z}^{\mathrm{T}}(\mathbf{I}_q \otimes \mathbf{\Gamma})\mathbf{z} - \sum_{l>1}^{r^*} \mathbf{a}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_l^*\mathbf{b}^{\mathrm{T}}\mathbf{v}_l^* + O_p(\frac{1}{\sqrt{n}})$$
$$+ \frac{\alpha\lambda_1^{(n)}}{\sqrt{n}} \sum_{i,j} w_{ij1}\mathrm{sign}(u_{i1}^* v_{j1}^*)(u_{i1}^* b_j + a_i v_{j1}^*) + \frac{(1-\alpha)\lambda_1^{(n)}}{\sqrt{n}} \sum_{i,j} 2u_{i1}^* v_{j1}^*(u_{i1}^* b_j + a_i v_{j1}^*),$$

where $\mathbf{z} = \mathrm{vec}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}})$. Now we show that, the quadratic term of $\mathbf{a}$ and $\mathbf{b}$ is $O(h^2)$ and positive, thus dominates the other terms of order $O(h)$, for sufficiently large h. Let us refer the set $\{j = 1,\ldots,q; j \neq \ell_1\}$ as ℓ_1^c, and define $\Psi_{11}^{(n)}(\mathbf{a},\mathbf{b}_{\ell_1^c}) = \frac{1}{2}\mathbf{z}^{\mathrm{T}}(\mathbf{I}_q \otimes \mathbf{\Gamma})\mathbf{z} - \sum_{l>1}^{r^*} \mathbf{a}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_l^*\mathbf{b}^{\mathrm{T}}\mathbf{v}_l^*$; it is noted that $b_{\ell_1} = 0$ always.

It suffices to show $\Psi_{11}^{(n)}(\mathbf{a},\mathbf{b}_{\ell_1^c}) = O(h^2)$ and is positive for large enough h. For fixed $\mathbf{b}_{\ell_1^c}$ we can minimize $\Psi_{11}^{(n)}(\mathbf{a},\mathbf{b}_{\ell_1^c})$ with $\widehat{\mathbf{a}} = \mathbf{N}^{-1}\mathbf{t}_a$ where $\mathbf{N} = \|\mathbf{v}_k^*\|_2^2\mathbf{\Gamma}$ and $\mathbf{t}_a = \sum_{l>1}^{r^*} \mathbf{\Gamma}\mathbf{u}_l^*\mathbf{v}_{l\ell_1^c}^{*\mathrm{T}}\mathbf{b}_{\ell_1^c} - \mathbf{\Gamma}\mathbf{u}_1^*\mathbf{v}_{1\ell_1^c}^{*\mathrm{T}}\mathbf{b}_{\ell_1^c}$. Then

$$\Psi_{11}^{(n)}(\widehat{\mathbf{a}},\mathbf{b}_{\ell_1^c}) = \mathbf{b}_{\ell_1^c}^{\mathrm{T}}\mathbf{M}\mathbf{b}_{\ell_1^c}/2,$$

where $\mathbf{M} = \mathbf{u}_1^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_1^*\mathbf{I}_{q-1} - (\sum_{l=1}^{r^*} \mathbf{u}_l^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_l^*\mathbf{v}_{l\ell_1^c}^*\mathbf{v}_{l\ell_1^c}^{*\mathrm{T}})/\|\mathbf{v}_1^*\|_2^2$. Now it suffices to show that $\mathbf{M}$ is positive definite. Given that $\mathbf{u}_k^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_k^*\|\mathbf{v}_k^*\|_2^2 = d_k^{*2}$, we can write $\mathbf{M}$ as

$$\mathbf{M} = \frac{d_1^{*2}}{\|\mathbf{v}_1^*\|_2^2}\left\{\mathbf{I}_{q-1} - \sum_{l=1}^{r^*} \frac{d_l^{*2}}{d_1^{*2}} \frac{\mathbf{v}_{l\ell_1^c}^*\mathbf{v}_{l\ell_1^c}^{*\mathrm{T}}}{\|\mathbf{v}_l^*\|_2^2}\right\}.$$

The rest of the proof for showing $\mathbf{M}$ is positive definite is similar to Chen et al. [2012];

in the proof of Theorem 2.14 below, we will also revisit this result. We thus omit the details here.

Now we use mathematical induction to finish the proof. Suppose $\|\widehat{\mathbf{u}}_l - \mathbf{u}_l^*\| = O_p(n^{-1/2})$ and $\|\widehat{\mathbf{v}}_l - \mathbf{v}_l^*\| = O_p(n^{-1/2})$, $l = 1, \ldots, k-1$, for some $k \geq 2$. Define

$$\mathcal{N}(\mathbf{C}_k^*, h) = \left\{ (\mathbf{u}_k^* + \mathbf{a}/\sqrt{n})(\mathbf{v}_k^* + \mathbf{b}/\sqrt{n})^{\mathrm{T}}; \|\mathbf{\Gamma}^{1/2}\mathbf{a}\| \leq h, \mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^q, \|\mathbf{b}\| \leq h, b_{\ell_k} = 0 \right\}.$$

We claim that for any $\epsilon > 0$, there exists large enough constant h such that

$$P\left\{ \inf_{\|\Gamma^{1/2}\mathbf{a}\| = \|\mathbf{b}\| = h} Q_k^{(n)}(\mathbf{u}_k^* + \mathbf{a}/\sqrt{n}, \mathbf{v}_k^* + \mathbf{b}/\sqrt{n}) > Q_k^{(n)}(\mathbf{u}_k^*, \mathbf{v}_k^*) \right\} \geq 1 - \epsilon.$$

Define

$$\Psi_k^{(n)}(\mathbf{a}, \mathbf{b}) = Q_k^{(n)}(\mathbf{u}_k^* + \mathbf{a}/\sqrt{n}, \mathbf{v}_k^* + \mathbf{b}/\sqrt{n}) - Q_k^{(n)}(\mathbf{u}_k^*, \mathbf{v}_k^*) = T_1 + T_2 + T_3, \qquad (2.30)$$

where T_1, T_2, T_3 are similar defined as in (2.26), by replacing $(\mathbf{Y}, \mathbf{u}_1^*, \mathbf{v}_1^*, \mathbf{W}_1, \lambda_1^{(n)})$ with

$(\mathbf{Y}_k, \mathbf{u}_k^*, \mathbf{v}_k^*, \mathbf{W}_k, \lambda_k^{(n)})$. Again, we write $T_1 = T_{11} + T_{12}$, where

$$\begin{aligned}
T_{11} &= -\operatorname{tr}\left\{(\mathbf{Y}_k - \mathbf{X}\mathbf{u}_k^*\mathbf{v}_k^{*\mathrm{T}})^{\mathrm{T}}\frac{\mathbf{X}}{\sqrt{n}}(\mathbf{u}_k^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_k^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})\right\} \\
&= \operatorname{tr}\left\{(\mathbf{X}\sum_{l=1}^{k-1}(\widehat{\mathbf{u}}_l\widehat{\mathbf{v}}_l^{\mathrm{T}} - \mathbf{u}_l^*\mathbf{v}_l^{*\mathrm{T}}) - \sum_{l\geq k}\mathbf{X}\mathbf{u}_l^*\mathbf{v}_l^{*\mathrm{T}} - \mathbf{E})^{\mathrm{T}}\frac{\mathbf{X}}{\sqrt{n}}(\mathbf{u}_k^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_k^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})\right\} \\
&= \operatorname{tr}\left\{\sum_{l=1}^{k-1}\sqrt{n}(\widehat{\mathbf{u}}_l\widehat{\mathbf{v}}_l^{\mathrm{T}} - \mathbf{u}_l^*\mathbf{v}_l^{*\mathrm{T}})^{\mathrm{T}}\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n}(\mathbf{u}_k^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_k^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})\right\} - \sum_{l>k}^{r^*}\mathbf{a}^{\mathrm{T}}\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n}\mathbf{u}_l^*\mathbf{b}^{\mathrm{T}}\mathbf{v}_l^* \\
&\quad - \operatorname{tr}\left\{\frac{\mathbf{E}^{\mathrm{T}}\mathbf{X}}{\sqrt{n}}(\mathbf{u}_k^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_k^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})\right\},
\end{aligned}$$

and

$$T_{12} = \operatorname{tr}\left\{(\mathbf{u}_k^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_k^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})^{\mathrm{T}}\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n}(\mathbf{u}_k^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_k^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})\right\}.$$

We know that $\sqrt{n}(\widehat{\mathbf{u}}_l\widehat{\mathbf{v}}_l^{\mathrm{T}} - \mathbf{u}_l^*\mathbf{v}_l^{*\mathrm{T}}) = O_p(1)$ for any $l < k$. Now as $n \to \infty$

$$\begin{aligned}
T_{11} &= -\operatorname{tr}\left\{\frac{\mathbf{E}^{\mathrm{T}}\mathbf{X}}{\sqrt{n}}(\mathbf{u}_k^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_k^{*\mathrm{T}})\right\} - \sum_{l>k}^{r^*}\mathbf{a}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{u}_l^*\mathbf{b}^{\mathrm{T}}\mathbf{v}_l^* + O_p(1) \\
T_{12} &= \operatorname{tr}\left\{(\mathbf{u}_k^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_k^{*\mathrm{T}})^{\mathrm{T}}\boldsymbol{\Gamma}(\mathbf{u}_k^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_k^{*\mathrm{T}})\right\} + O_p(\frac{1}{\sqrt{n}}).
\end{aligned}$$

The rest of the proof is similar to the case of $k = 1$. This completes the proof.

□

Theorem 2.14. *(Asymptotic Normality) Suppose* ***A1-A4*** *are satisfied. Let* $\mathcal{A}_k =$

$\{i; u^*_{ik} \neq 0\}$ and $\mathcal{B}_k = \{j; v^*_{jk} \neq 0, j \neq \ell_k\}$, for $k = 1, \ldots, r^*$; denote their complements as $\mathcal{A}_k^c$ and $\mathcal{B}_k^c$. Accordingly, define the subvectors $\mathbf{u}^*_{k\mathcal{A}_k}$, $\widehat{\mathbf{u}}_{k\mathcal{A}_k}$, $\mathbf{v}^*_{k\mathcal{B}_k}$, $\widehat{\mathbf{v}}^*_{k\mathcal{B}_k}$, etc. Then

i. *$\sqrt{n}(\widehat{\mathbf{u}}_{k\mathcal{A}_k} - \mathbf{u}^*_{k\mathcal{A}_k})$ and $\sqrt{n}(\widehat{\mathbf{v}}_{k\mathcal{B}_k} - \mathbf{v}^*_{k\mathcal{B}_k})$ are jointly asymptotically normally distributed with zero mean.*

ii. *$\sqrt{n}(\widehat{\mathbf{u}}_{k\mathcal{A}_k^c} - \mathbf{u}^*_{k\mathcal{A}_k^c}) \to_d 0$ and $\sqrt{n}(\widehat{\mathbf{v}}_{k\mathcal{B}_k^c} - \mathbf{v}^*_{k\mathcal{B}_k^c}) \to_d 0$ as $n \to \infty$.*

Proof. Consider the objective function $\Psi_k^{(n)}(\mathbf{a}, \mathbf{b})$ defined in (2.30) with $Q_k^{(n)}(\mathbf{a}, \mathbf{b})$ defined in (2.23). We first derive the limiting objective function. The limit of T_1 has been discussed in the proof of Theorem 2.13. To deal with T_2, consider the following three cases.

(I) When $u^*_{ik} v^*_{jk} \neq 0$, we have

(i) $w_{ijk} \to_p |u^*_{ik} v^*_{jk}|^{-\gamma}$;

(ii) $\sqrt{n}(|u^*_{ik} v^*_{jk} + \frac{1}{\sqrt{n}} u^*_{ik} b_j + \frac{1}{\sqrt{n}} a_i v^*_{jk} + \frac{1}{n} a_i b_j| - |u^*_{ik} v^*_{jk}|) \to \mathrm{sgn}(u^*_{ik} v^*_{jk})(u^*_{ik} b_j + a_i v^*_{jk})$;

(iii) $\lambda_k^{(n)} / \sqrt{n} \to 0$.

Then $\frac{\lambda_k^{(n)}}{\sqrt{n}} w_{ijk} \sqrt{n}(|u^*_{ik} v^*_{jk} + \frac{1}{\sqrt{n}} u^*_{ik} b_j + \frac{1}{\sqrt{n}} a_i v^*_{jk} + \frac{1}{n} a_i b_j| - |u^*_{ik} v^*_{jk}|) \to_p 0$.

(II) When $u_{ik}^* = 0, v_{jk}^* \neq 0$, we have

$$\text{(i) } \frac{\lambda_k^{(n)}}{\sqrt{n}} w_{ijk} = \frac{\lambda_k^{(n)}}{\sqrt{n}} n^{\frac{\gamma}{2}} |\sqrt{n}\tilde{c}_{ijk}|^{-\gamma} \to_p \infty;$$

$$\text{(ii) } \sqrt{n}(|u_{ik}^* v_{jk}^* + \frac{1}{\sqrt{n}} u_{ik}^* b_j + \frac{1}{\sqrt{n}} a_i v_{jk}^* + \frac{1}{n} a_i b_j| - |u_{ik}^* v_{jk}^*|) \to |a_i v_{jk}^*|.$$

Then $\frac{\lambda_k^{(n)}}{\sqrt{n}} w_{ijk} \sqrt{n}(|u_{ik}^* v_{jk}^* + \frac{1}{\sqrt{n}} u_{ik}^* b_j + \frac{1}{\sqrt{n}} a_i v_{jk}^* + \frac{1}{n} a_i b_j| - |u_{ik}^* v_{jk}^*|) \to_p \infty$, if $a_i \neq 0$.

(III) When $u_{ik}^* \neq 0$, $v_{jk}^* = 0$, similarly we have $\frac{\lambda_k^{(n)}}{\sqrt{n}} w_{ijk} \sqrt{n}(|u_{ik}^* v_{jk}^* + \frac{1}{\sqrt{n}} u_{ik}^* b_j + \frac{1}{\sqrt{n}} a_i v_{jk}^* + \frac{1}{n} a_i b_j| - |u_{ik}^* v_{jk}^*|) \to_p \infty$, if $b_j \neq 0$.

It is easy to verify that $T_3 \to_p 0$ under **A4**. Therefore,

$$\Psi_k^{(n)}(\mathbf{a}, \mathbf{b}) \to_d \Psi_k(\mathbf{a}, \mathbf{b})$$

$$= \begin{cases} -\mathbf{z}^{\mathrm{T}}\mathbf{w} + \frac{1}{2}\mathbf{z}^{\mathrm{T}}(\mathbf{I}_q \otimes \mathbf{\Gamma})\mathbf{z} - \sum_{l>k}^{r^*} \mathbf{a}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_l^*\mathbf{b}^{\mathrm{T}}\mathbf{v}_l^*, & a_i = 0 \text{ if } i \notin \mathcal{A}_k; b_j = 0 \text{ if } j \notin \mathcal{B}_k, \\ \infty, & \text{otherwise.} \end{cases}$$

where $\mathbf{z} = \mathrm{vec}(\mathbf{u}_k^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_k^{*\mathrm{T}})$, and $\mathbf{w} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{\Gamma})$.

Next we show that $\Psi_k(\mathbf{a}, \mathbf{b})$ has a unique minimum denoted as $(\widehat{\mathbf{a}}, \widehat{\mathbf{b}})$. Obviously, $\forall i \notin \mathcal{A}_k$, $\widehat{a}_i = 0$, and $\forall k \notin \mathcal{B}_k$, $\widehat{b}_j = 0$. Because $\mathrm{vec}(\mathbf{a}\mathbf{v}_k^{*\mathrm{T}}) = (\mathbf{v}_k^* \otimes \mathbf{I}_p)\mathbf{a}$ and $\mathrm{vec}(\mathbf{u}_k^*\mathbf{b}^{\mathrm{T}}) =$

$(\mathbf{I}_q \otimes \mathbf{u}_k^*)\mathbf{b}$, we have

$$\Psi_k(\mathbf{a},\mathbf{b}) = -\mathbf{a}^{\mathrm{T}}(\mathbf{v}_k^{*\mathrm{T}} \otimes \mathbf{I}_p)\mathbf{w} + \frac{1}{2}\mathbf{a}^{\mathrm{T}}(\|\mathbf{v}_k^*\|_2^2 \otimes \mathbf{\Gamma})\mathbf{a} + \mathbf{a}^{\mathrm{T}}(\mathbf{v}_k^{*\mathrm{T}} \otimes \mathbf{\Gamma}\mathbf{u}_k^*)\mathbf{b} - \mathbf{b}^{\mathrm{T}}(\mathbf{I}_q \otimes \mathbf{u}_k^{*\mathrm{T}})\mathbf{w} + \frac{1}{2}\mathbf{b}^{\mathrm{T}}(\mathbf{I}_q \otimes \mathbf{u}_k^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_k^*)\mathbf{b} - \sum_{l>k}^{r^*} \mathbf{a}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_l^*\mathbf{b}^{\mathrm{T}}\mathbf{v}_l^*.$$

Now define the following conditional objective function,

$$\Psi_{\mathcal{A}_k\mathcal{B}_k}(\mathbf{a}_{\mathcal{A}_k}, \mathbf{b}_{\mathcal{B}_k}) = \Psi_k(\mathbf{a},\mathbf{b} \mid \forall i \notin \mathcal{A}_k, a_i = 0; \forall j \notin \mathcal{B}_k, b_j = 0).$$

We first fix $\mathbf{b}_{\mathcal{B}_k}$ to optimize the above function with respect to $\mathbf{a}_{\mathcal{A}_k}$,

$$\Psi_{\mathcal{A}_k\mathcal{B}_k}(\mathbf{a}_{\mathcal{A}_k}, \mathbf{b}_{\mathcal{B}_k} \mid \mathbf{b}_{\mathcal{B}_k}) = -\mathbf{a}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{p}_a + \frac{1}{2}\mathbf{a}_{\mathcal{A}_k}^{\mathrm{T}}\mathbf{N}\mathbf{a}_{\mathcal{A}_k} + \text{const},$$

where $\mathbf{p}_a = \left\{\mathbf{v}_k^{*\mathrm{T}} \otimes (\mathbf{I}_p)_{\mathcal{A}_k\cdot}\right\}\mathbf{w} + \tilde{\mathbf{T}}\mathbf{b}_{\mathcal{B}_k}$ with $\tilde{\mathbf{T}} = \sum_{l>k}^{r^*} \mathbf{\Gamma}_{\mathcal{A}_k\cdot}\mathbf{u}_l^*\mathbf{v}_{l\mathcal{B}_k}^{*\mathrm{T}} - \mathbf{\Gamma}_{\mathcal{A}_k\cdot}\mathbf{u}_k^*\mathbf{v}_{k\mathcal{B}_k}^{*\mathrm{T}}$ and $\mathbf{N} = \|\mathbf{v}_k^*\|_2^2\mathbf{\Gamma}_{\mathcal{A}_k}$. $\mathbf{N}$ is obviously positive definite, so the unique minimizer is $\widehat{\mathbf{a}}_{\mathcal{A}_k} = \mathbf{N}^{-1}\mathbf{p}_a$.

We then substitute the expression of $\widehat{\mathbf{a}}_{\mathcal{A}_k}$ in $\Psi_{\mathcal{A}_k\mathcal{B}_k}(\mathbf{a}_{\mathcal{A}_k}, \mathbf{b}_{\mathcal{B}_k})$. After some algebra, we have

$$\Psi_{\mathcal{A}_k\mathcal{B}_k}(\mathbf{a}_{\mathcal{A}_k}, \mathbf{b}_{\mathcal{B}_k} \mid \mathbf{a}_{\mathcal{A}_k} = \widehat{\mathbf{a}}_{\mathcal{A}_k}) = -\mathbf{b}_{\mathcal{B}_k}^{\mathrm{T}}\mathbf{p}_b + \frac{1}{2}\mathbf{b}_{\mathcal{B}_k}^{\mathrm{T}}\mathbf{M}\mathbf{b}_{\mathcal{B}_k} + \text{const},$$

where

$$\mathbf{M} = (\mathbf{u}_k^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_k^*)\mathbf{I}_{\mathcal{B}_k} - \tilde{\mathbf{T}}^{\mathrm{T}}\mathbf{N}^{-1}\tilde{\mathbf{T}},$$

and

$$\mathbf{p}_b = \{(\mathbf{I}_q)_{\mathcal{B}_k.} \otimes \mathbf{u}_k^{*\mathrm{T}}\}\mathbf{w} + \tilde{\mathbf{T}}^{\mathrm{T}}\mathbf{N}^{-1}\left\{\mathbf{v}_k^{*\mathrm{T}} \otimes (\mathbf{I}_p)_{\mathcal{A}_k.}\right\}\mathbf{w}.$$

It remains to verify that $\mathbf{M}$ is positive-definite. For $l \neq k$ we can write $\mathbf{u}_k^{*\mathrm{T}}\mathbf{\Gamma}_{.\mathcal{A}_k}\mathbf{\Gamma}_{\mathcal{A}_k}^{-1}\mathbf{\Gamma}_{\mathcal{A}_k.}\mathbf{u}_l^* = \mathbf{u}_{k\mathcal{A}_k}^{*\mathrm{T}}\mathbf{\Gamma}_{\mathcal{A}_k}\mathbf{\Gamma}_{\mathcal{A}_k}^{-1}\mathbf{\Gamma}_{\mathcal{A}_k.}\mathbf{u}_l^* = \mathbf{u}_k^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_l^* = 0$. Hence $\mathbf{u}_k^{*\mathrm{T}}\mathbf{\Gamma}_{.\mathcal{A}_k}\mathbf{N}^{-1}\mathbf{\Gamma}_{\mathcal{A}_k.}\mathbf{u}_l^* = (\mathbf{u}_k^{*\mathrm{T}}\mathbf{\Gamma}_{.\mathcal{A}_k}\mathbf{\Gamma}_{\mathcal{A}_k}^{-1}\mathbf{\Gamma}_{\mathcal{A}_k.}\mathbf{u}_l^*)/\|\mathbf{v}_k^*\|_2^2 = 0$. We can write $\mathbf{M}$ as:

$$\begin{aligned}\mathbf{M} &= \mathbf{u}_k^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_k^*\mathbf{I}_{\mathcal{B}_k} - \sum_{l=k}^{r^*}\mathbf{v}_{l\mathcal{B}_k}^*\mathbf{u}_l^{*\mathrm{T}}\mathbf{\Gamma}_{.\mathcal{A}_k}\mathbf{N}^{-1}\mathbf{\Gamma}_{\mathcal{A}_k.}\mathbf{u}_l^*\mathbf{v}_{l\mathcal{B}_k}^{*\mathrm{T}} \\ &= \mathbf{u}_k^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_k^*\mathbf{I}_{\mathcal{B}_k} - \frac{1}{\|\mathbf{v}_k^*\|_2^2}\sum_{l=k}^{r^*}\mathbf{v}_{l\mathcal{B}_k}^*\mathbf{u}_l^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_l^*\mathbf{v}_{l\mathcal{B}_k}^{*\mathrm{T}} + \frac{1}{\|\mathbf{v}_k^*\|_2^2}\sum_{l=k}^{r^*}\mathbf{v}_{l\mathcal{B}_k}^*\mathbf{u}_l^{*\mathrm{T}}(\mathbf{\Gamma} - \mathbf{\Gamma}_{.\mathcal{A}_k}\mathbf{\Gamma}_{\mathcal{A}_k}^{-1}\mathbf{\Gamma}_{\mathcal{A}_k.})\mathbf{u}_l^*\mathbf{v}_{l\mathcal{B}_k}^{*\mathrm{T}}\end{aligned}$$

The third term is non-negative definite (n.n.d) as $\mathbf{\Gamma} - \mathbf{\Gamma}_{.\mathcal{A}_k}\mathbf{\Gamma}_{\mathcal{A}_k}^{-1}\mathbf{\Gamma}_{\mathcal{A}_k.}$ is n.n.d.. We show that

$$\mathbf{M}_1 = \mathbf{u}_k^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_k^*\mathbf{I}_{\mathcal{B}_k} - \frac{1}{\|\mathbf{v}_k^*\|_2^2}\sum_{l=k}^{r^*}\mathbf{v}_{l\mathcal{B}_k}^*\mathbf{u}_l^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_l^*\mathbf{v}_{l\mathcal{B}_k}^{*\mathrm{T}}$$

is positive definite. Using $(\mathbf{u}_l^{*\mathrm{T}}\mathbf{\Gamma}\mathbf{u}_l^*)\|\mathbf{v}_l^*\|_2^2 = d_l^{*2}$, we can write $\mathbf{M}_1 = (d_k^{*2}/\|\mathbf{v}_k^*\|_2^2)\mathbf{H}$, where

$$\mathbf{H} = \mathbf{I}_{\mathcal{B}_k} - \sum_{l=k}^{r^*}\frac{d_l^{*2}}{d_k^{*2}}\frac{\mathbf{v}_{l\mathcal{B}_k}^*\mathbf{v}_{l\mathcal{B}_k}^{*\mathrm{T}}}{\|\mathbf{v}_l^*\|_2^2}.$$

Now it is sufficient to show that $\mathbf{H}$ is positive definite. Let $\mathcal{K} = \{k, \ldots, r^*\}$, we write $\mathbf{H}$

as

$$\mathbf{H} = \mathbf{I}_{\mathcal{B}_k} - \mathbf{V}^*_{\mathcal{B}_k\mathcal{K}}\mathbf{D}^*_{\mathcal{K}}(\mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}})^{-1}\mathbf{D}^*_{\mathcal{K}}\mathbf{V}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}$$

where $\mathbf{D}^*_{\mathcal{K}} = \mathrm{diag}\{d^*_k/d^*_k, \ldots, d^*_{r^*}/d^*_k\}$ and $\mathbf{V}^*_{\mathcal{B}_k\mathcal{K}}$, $\mathbf{V}^*_{\cdot\mathcal{K}}$ are sub-matrix of $\mathbf{V}^*$. We define $\widetilde{\mathbf{V}}^*_{\mathcal{B}_k\mathcal{K}} = \mathbf{V}^*_{\mathcal{B}_k\mathcal{K}}\mathbf{D}^*_{\mathcal{K}}(\mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}})^{-1/2}$ so that $\mathbf{H} = \mathbf{I}_{\mathcal{B}_k} - \widetilde{\mathbf{V}}^*_{\mathcal{B}_k\mathcal{K}}\widetilde{\mathbf{V}}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}$. For any non-zero $\mathbf{s} \in \mathbb{R}^{|\mathcal{B}_k|}$ that is orthogonal to the column space of $\widetilde{\mathbf{V}}^*_{\mathcal{B}_k\mathcal{K}}$, we have $\mathbf{Hs} = \mathbf{s}$, yielding a positive eigenvalue. It then remains to show that for any $\mathbf{s} \in \mathbb{R}^{|\mathcal{K}|}$, $\mathbf{H}\widetilde{\mathbf{V}}^*_{\mathcal{B}_k\mathcal{K}}\mathbf{s} = \mathbf{0}$ only if $\widetilde{\mathbf{V}}^*_{\mathcal{B}_k\mathcal{K}}\mathbf{s} = \mathbf{0}$. Note that $\mathbf{H}\widetilde{\mathbf{V}}^*_{\mathcal{B}_k\mathcal{K}}\mathbf{s} = \widetilde{\mathbf{V}}^*_{\mathcal{B}_k\mathcal{K}}(\mathbf{I} - \widetilde{\mathbf{V}}^{*\mathrm{T}}{}_{\mathcal{B}_k\mathcal{K}}\widetilde{\mathbf{V}}^*_{\mathcal{B}_k\mathcal{K}})\mathbf{s}$ and

$$\begin{aligned}\mathbf{I} - \widetilde{\mathbf{V}}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\widetilde{\mathbf{V}}^*_{\mathcal{B}_k\mathcal{K}} &= \mathbf{I} - (\mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}})^{-1/2}\mathbf{D}^*_{\mathcal{K}}\mathbf{V}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\mathbf{V}^*_{\mathcal{B}_k\mathcal{K}}\mathbf{D}^*_{\mathcal{K}}(\mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}})^{-1/2} \\ &= (\mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}})^{-1/2}(\mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}} - \mathbf{D}^*_{\mathcal{K}}\mathbf{V}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\mathbf{V}^*_{\mathcal{B}_k\mathcal{K}}\mathbf{D}^*_{\mathcal{K}})(\mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}})^{-1/2},\end{aligned}$$

in which

$$\begin{aligned}\mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}} - \mathbf{D}^*_{\mathcal{K}}\mathbf{V}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\mathbf{V}^*_{\mathcal{B}_k\mathcal{K}}\mathbf{D}^*_{\mathcal{K}} &= \mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}} - \mathbf{D}^{*2}_{\mathcal{K}}\mathbf{V}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\mathbf{V}^*_{\mathcal{B}_k\mathcal{K}} \\ &= \mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}} - \mathbf{D}^{*2}_{\mathcal{K}}(\mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}} - \mathbf{V}^{*\mathrm{T}}_{\mathcal{B}^c_k\mathcal{K}}\mathbf{V}^*_{\mathcal{B}^c_k\mathcal{K}} - \mathbf{V}^{*\mathrm{T}}_{\ell_k\mathcal{K}}\mathbf{V}^*_{\ell_k\mathcal{K}}) \\ &= (\mathbf{I} - \mathbf{D}^{*2}_{\mathcal{K}})\mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}} + \mathbf{D}^{*2}_{\mathcal{K}}\mathbf{V}^{*\mathrm{T}}_{\ell_k\mathcal{K}}\mathbf{V}^*_{\ell_k\mathcal{K}} + \mathbf{D}^{*2}_{\mathcal{K}}\mathbf{V}^{*\mathrm{T}}_{\mathcal{B}^c_k\mathcal{K}}\mathbf{V}^*_{\mathcal{B}^c_k\mathcal{K}}.\end{aligned}$$

The decomposition simply shows the fact that $\mathbf{H}$ is n.n.d.. It should be noted that the first diagonal element in the diagonal matrix $(\mathbf{I} - \mathbf{D}^{*2}_{\mathcal{K}})\mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^*_{\cdot\mathcal{K}}$ is zero but since the first

diagonal element of the symmetric matrix $\mathbf{D}^{*2}_{\mathcal{K}}\mathbf{V}^{*\mathrm{T}}_{\ell_k\mathcal{K}}\mathbf{V}^{*}_{\ell_k\mathcal{K}}$ is non-zero, as a result overall we can claim $(\mathbf{I}-\mathbf{D}^{*2}_{\mathcal{K}})\mathbf{V}^{*\mathrm{T}}_{\cdot\mathcal{K}}\mathbf{V}^{*}_{\cdot\mathcal{K}}+\mathbf{D}^{*2}_{\mathcal{K}}\mathbf{V}^{*\mathrm{T}}_{\ell_k\mathcal{K}}\mathbf{V}^{*}_{\ell_k\mathcal{K}}$ to be positive definite. This result is true only when $d^*_k > d^*_{k+1} > \ldots > d^*_{r^*}$. Hence $\mathbf{I}-\widetilde{\mathbf{V}}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}$ is invertible. Now suppose $\mathbf{H}\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}\mathbf{s}=\mathbf{0}$, we have

$$
\begin{aligned}
\mathbf{0} &= \mathbf{H}\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}\mathbf{s} \\
&= \widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}\mathbf{s}-\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}(\mathbf{I}-\widetilde{\mathbf{V}}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}})^{-1}(\mathbf{I}-\widetilde{\mathbf{V}}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}})\widetilde{\mathbf{V}}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}\mathbf{s} \\
&= \widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}\mathbf{s}-\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}(\mathbf{I}-\widetilde{\mathbf{V}}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}})^{-1}\widetilde{\mathbf{V}}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}\left\{\mathbf{I}-\widetilde{\mathbf{V}}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}\right\}\mathbf{s} \\
&= \widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}\mathbf{s}-\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}(\mathbf{I}-\widetilde{\mathbf{V}}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}})^{-1}\widetilde{\mathbf{V}}^{*\mathrm{T}}_{\mathcal{B}_k\mathcal{K}}\mathbf{H}\widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}\mathbf{s} \\
&= \widetilde{\mathbf{V}}^{*}_{\mathcal{B}_k\mathcal{K}}\mathbf{s}-\mathbf{0}.
\end{aligned}
$$

Thus, $\mathbf{H}$ is indeed positive-definite. The unique minimum of $\mathbf{b}_{\mathcal{B}_k}$ is given by $\widehat{\mathbf{b}}_{\mathcal{B}_k} = \mathbf{M}^{-1}\mathbf{p}_b$.

We can then substitute the expression of $\widehat{\mathbf{b}}_{\mathcal{B}_k}$ in $\widehat{\mathbf{a}}_{\mathcal{A}_k}$ to obtain the final expression of $\widehat{\mathbf{a}}_{\mathcal{A}_k}$. It can be seen that both $\widehat{\mathbf{a}}_{\mathcal{A}_k}$ and $\widehat{\mathbf{b}}_{\mathcal{B}_k}$ are linear functions of $\mathbf{w}$ which follows a multivariate normal distribution with zero mean. The asymptomatic normality results follow by noting that $\widehat{\mathbf{a}}^{(n)}_{\mathcal{A}_k} = \sqrt{n}(\widehat{\mathbf{u}}^{(n)}_{k\mathcal{A}_k} - \mathbf{u}^*_{k\mathcal{A}_k})$ and $\widehat{\mathbf{b}}^{(n)}_{\mathcal{B}_k} = \sqrt{n}(\widehat{\mathbf{v}}^{(n)}_{k\mathcal{B}_k} - \mathbf{v}^*_{k\mathcal{B}_k})$, and invoking the Argmax theorem [Page 81, van der Vaart, 2000]. For similicity we omit the explicit expression of the asymptotic covariance matrices.

□

Theorem 2.15. *(Selection Consistency and Rank Consistency) Suppose* ***A1-A4*** *are*

satisfied. Let $\mathcal{A} = \{(i,k) : u^*_{ik} \neq 0, k = 1,\ldots,r^*\}$ *and* $\mathcal{B} = \{(j,k) : v^*_{jk} \neq 0, (j,k) \neq (\ell_k,k), k = 1,\ldots,r^*\}$*, and let* $\mathcal{A}^{(n)} = \{(i,k) : \widehat{u}_{ik} \neq 0, k = 1,\ldots,r^*\}$ *and* $\mathcal{B}^{(n)} = \{(j,k) : \widehat{v}_{jk} \neq 0, (j,k) \neq (\ell_k,k), k = 1,\ldots,r^*\}$*. Let* $\widehat{d_k} = (1/n)(\widehat{\mathbf{u}}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\mathbf{u}}_k)(\widehat{\mathbf{v}}_k^{\mathrm{T}}\widehat{\mathbf{v}}_k)$ *be the* k*th estimated singular value, for* $k = r^*+1,\ldots,r$*. Then*

i. $P(\mathcal{A}^{(n)} = \mathcal{A}) \to 1$ *and* $P(\mathcal{B}^{(n)} = \mathcal{B}) \to 1$ *as* $n \to \infty$*.*

ii. $P(\widehat{d_k} = 0) \to 1$ *as* $n \to \infty$*, for* $k = r^*+1,\ldots,r$*.*

Proof. According to the results in Theorem 2.14, $\widehat{\mathbf{u}}_{k\mathcal{A}_k} \to_p \mathbf{u}^*_{k\mathcal{A}}$ and $\widehat{\mathbf{v}}_{k\mathcal{B}} \to_p \mathbf{v}^*_{k\mathcal{B}}$; thus $\forall(i,k) \in \mathcal{A}$, $P((i,k) \in \mathcal{A}^{(n)}) \to 1$, and $\forall(j,k) \in \mathcal{B}$, $P((j,k) \in \mathcal{B}^{(n)}) \to 1$ as $n \to \infty$. Then it suffices to show that $\forall(i,k) \notin \mathcal{A}$, $P((i,k) \in \mathcal{A}^{(n)}) \to 0$, and $\forall(j,k) \notin \mathcal{B}$, $P((j,k) \in \mathcal{B}^{(n)}) \to 0$.

$\forall(j,k) \notin \mathcal{B}$, consider the event $(j,k) \in \mathcal{B}^{(n)}$. Using the KKT optimality conditions on (2.23),

$$\frac{1}{\sqrt{n}}\mathbf{x}_j^{(v)\mathrm{T}}(\mathbf{y}_k - \mathbf{X}^{(v)}\widehat{\mathbf{v}}_k) = \frac{1}{\sqrt{n}}\lambda^{(v)}w_{jk}^{(v)} + 2(1-\alpha)\frac{1}{\sqrt{n}}\lambda_k^{(n)}\widehat{v}_{ik}\sum\widehat{u}_{ik}^2 \tag{2.31}$$

where $\mathbf{X}^{(v)} = (\mathbf{x}_1^{(v)},\ldots,\mathbf{x}_q^{(v)}) = \mathbf{I}_q \otimes \mathbf{X}\widehat{\mathbf{u}}_k$, $\lambda^{(v)} = \alpha\lambda_k^{(n)}\sum_{i=1}^p w_{ik}^{(u)}|\widehat{u}_{ik}|$, $w_{ik}^{(u)} = |\tilde{u}_{ik}|^{-\gamma}$,

$w_{jk}^{(v)} = |\tilde{v}_{jk}|^{-\gamma}$, and $\mathbf{y}_k = \mathrm{vec}(\mathbf{Y} - \mathbf{X}\sum_{l<k}\widehat{\mathbf{u}}_l\widehat{\mathbf{v}}_l^{\mathrm{T}})$. Consider the left hand side of (2.31):

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\mathbf{x}_j^{(v)\mathrm{T}}(\mathbf{y}_k - \mathbf{X}^{(v)}\widehat{\mathbf{v}}_k) \\
=&\frac{1}{\sqrt{n}}\left\{\widehat{\mathbf{u}}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\sum_{l=k}^{r^*}\mathbf{u}_l^* v_{jl}^* + \widehat{\mathbf{u}}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\sum_{l<k}\mathbf{u}^*{}_l v_{jl}^* - \sum_{l<k}\widehat{\mathbf{u}}_l\widehat{v}_{jl}) + \widehat{\mathbf{u}}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{e}_{(j)} - \widehat{\mathbf{u}}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\mathbf{u}}_k\widehat{v}_{jk}\right\} \\
=&\widehat{\mathbf{u}}_k^{\mathrm{T}}\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n}\sqrt{n}(\sum_{l<k}\mathbf{u}_l^* v_{jl}^* - \sum_{l<k}\widehat{\mathbf{u}}_l\widehat{v}_{jl}) + \frac{\widehat{\mathbf{u}}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{e}_{(j)}}{\sqrt{n}} - \frac{\widehat{\mathbf{u}}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\mathbf{u}}_k}{n}\sqrt{n}(\widehat{v}_{jk} - v_{jk}^*) \\
&+ \frac{1}{\sqrt{n}}\widehat{\mathbf{u}}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\sum_{l>k}^{r^*}\mathbf{u}_l^* v_{jl}^* = O_p(1).
\end{aligned}
$$

On the other hand, the right hand side of (2.31) equals:

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\lambda^{(v)}w_{jk}^{(v)} + 2(1-\alpha)\frac{1}{\sqrt{n}}\lambda_k^{(n)}\widehat{v}_{ik}\sum\widehat{u}_{ik}^2 = \frac{\lambda_k^{(n)}}{\sqrt{n}}|\widehat{\mathbf{u}}_k|^{\mathrm{T}}\mathbf{w}_k^{(u)}w_{jk}^{(v)} + 2(1-\alpha)\frac{1}{\sqrt{n}}\lambda_k^{(n)}\widehat{v}_{ik}\sum\widehat{u}_{ik}^2 \\
=&\frac{\lambda_k^{(n)}}{\sqrt{n}}n^{\frac{\gamma}{2}}|\widehat{\mathbf{u}}_k|^{\mathrm{T}}|\sqrt{n}\tilde{v}_{jk}\tilde{\mathbf{u}}_k|^{-\gamma} + 2(1-\alpha)\frac{1}{\sqrt{n}}\lambda_k^{(n)}\widehat{v}_{ik}\sum\widehat{u}_{ik}^2 \to_p \infty
\end{aligned}
$$

as $n \to \infty$. Therefore, as $n \to \infty$,

$$
P((j,k) \in \mathcal{B}^{(n)}) \le P\{\frac{1}{\sqrt{n}}\mathbf{x}_j^{(v)\mathrm{T}}(\mathbf{y}_k - \mathbf{X}^{(v)}\widehat{\mathbf{v}}_k) = \frac{1}{\sqrt{n}}\lambda^{(v)}w_{jk}^{(v)} + 2(1-\alpha)\frac{1}{\sqrt{n}}\lambda_k^{(n)}\widehat{\mathbf{v}}_{ik}\sum\widehat{\mathbf{u}}_{ik}^2\} \to 0.
$$

$\forall (i,k) \notin \mathcal{A}$, consider the event $(i,k) \in \mathcal{A}^{(n)}$. Using the KKT optimality conditions on (2.23),

$$
\frac{1}{\sqrt{n}}\mathbf{x}_i^{(u)\mathrm{T}}(\mathbf{y}_k - \mathbf{X}^{(u)}\widehat{\mathbf{u}}_k) = \frac{1}{\sqrt{n}}\lambda^{(u)}w_{ik}^{(u)} + 2(1-\alpha)\frac{1}{\sqrt{n}}\lambda_k^{(n)}\widehat{u}_{ik}\sum\widehat{v}_{jk}^2, \tag{2.32}
$$

where $\mathbf{X}^{(u)} = (\mathbf{x}_1^{(u)}, \ldots, \mathbf{x}_p^{(u)}) = \widehat{\mathbf{v}}_k \otimes \mathbf{X}$, $\lambda^{(u)} = \lambda_k^{(n)} \sum_{j=1}^{q} w_{jk}^{(v)} |\widehat{v}_{jk}|$, $w_{ik}^{(u)} = |\tilde{u}_{ik}|^{-\gamma}$, $w_{jk}^{(v)} = |\tilde{v}_{jk}|^{-\gamma}$, and $\mathbf{y}_k = \mathrm{vec}(\mathbf{Y} - \mathbf{X} \sum_{l<k} \widehat{\mathbf{u}}_l \widehat{\mathbf{v}}_l^{\mathrm{T}})$. Consider the left hand side of (2.32):

$$
\begin{aligned}
&\frac{1}{\sqrt{n}} \mathbf{x}_i^{(u)\mathrm{T}} (\mathbf{y}_k - \mathbf{X}^{(u)} \widehat{\mathbf{u}}_k) \\
=&\frac{1}{\sqrt{n}} \widehat{\mathbf{v}}_k^{\mathrm{T}} \left\{ (\sum_{l=k}^{r^*} \mathbf{v}_l^* \mathbf{u}_l^{*\mathrm{T}} - \widehat{\mathbf{v}}_k \widehat{\mathbf{u}}_k^{\mathrm{T}} + \sum_{l<k} \mathbf{v}_l^* \mathbf{u}_l^{*\mathrm{T}} - \sum_{l<k} \widehat{\mathbf{v}}_l \widehat{\mathbf{u}}_l^{\mathrm{T}}) \mathbf{X}^{\mathrm{T}} \mathbf{x}_{(i)} + \mathbf{E}^{\mathrm{T}} \mathbf{x}_{(i)} \right\} \\
=&\widehat{\mathbf{v}}_k^{\mathrm{T}} \sqrt{n} (\sum_{l<k} \mathbf{v}_l^* \mathbf{u}_l^{*\mathrm{T}} - \sum_{l<k} \widehat{\mathbf{v}}_l \widehat{\mathbf{u}}_l^{\mathrm{T}}) \frac{\mathbf{X}^{\mathrm{T}} \mathbf{x}_{(i)}}{n} \\
&+ \left\{ \sqrt{n} (\widehat{\mathbf{v}}_k - \mathbf{v}_k^*)^{\mathrm{T}} \mathbf{v}_k^* \mathbf{u}_k^{*\mathrm{T}} - \widehat{\mathbf{v}}_k^{\mathrm{T}} \widehat{\mathbf{v}}_k \sqrt{n} (\widehat{\mathbf{u}}_k - \mathbf{u}_k^*)^{\mathrm{T}} - \sqrt{n} (\widehat{\mathbf{v}}_k^{\mathrm{T}} \widehat{\mathbf{v}}_k - \mathbf{v}_k^{*\mathrm{T}} \mathbf{v}_k^*) \mathbf{u}_k^{*\mathrm{T}} \right\} \frac{\mathbf{X}^{\mathrm{T}} \mathbf{x}_{(i)}}{n} \\
&+ \sqrt{n} \widehat{\mathbf{v}}_k^{\mathrm{T}} \sum_{l>k} \mathbf{v}_l^* \mathbf{u}_l^{*\mathrm{T}} \frac{\mathbf{X}^{\mathrm{T}} \mathbf{x}_{(i)}}{n} + \frac{\widehat{\mathbf{v}}_k^{\mathrm{T}} \mathbf{E}^{\mathrm{T}} \mathbf{x}_{(i)}}{\sqrt{n}} \\
=&O_p(1).
\end{aligned}
$$

But the right hand side of (2.32) equals:

$$
\begin{aligned}
&\frac{1}{\sqrt{n}} \lambda^{(u)} w_{ik}^{(u)} + 2(1-\alpha) \frac{1}{\sqrt{n}} \lambda_k^{(n)} \widehat{u}_{ik} \sum \widehat{v}_{jk}^2 = \frac{\lambda_k^{(n)}}{\sqrt{n}} |\widehat{\mathbf{v}}_k|^{\mathrm{T}} \mathbf{w}_k^{(v)} w_{ik}^{(u)} + 2(1-\alpha) \frac{1}{\sqrt{n}} \lambda_k^{(n)} \widehat{u}_{ik} \sum \widehat{v}_{jk}^2 \\
=&\frac{\lambda_k^{(n)}}{\sqrt{n}} n^{\frac{\gamma}{2}} |\widehat{\mathbf{v}}_k|^{\mathrm{T}} |\sqrt{n} \tilde{u}_{ik} \tilde{\mathbf{v}}_k|^{-\gamma} + 2(1-\alpha) \frac{1}{\sqrt{n}} \lambda_k^{(n)} \widehat{u}_{ik} \sum \widehat{v}_{jk}^2 \to_p \infty
\end{aligned}
$$

as $n \to \infty$. Therefore, as $n \to \infty$,

$$
P((i,k) \in \mathcal{A}^{(n)}) \leq P(\frac{1}{\sqrt{n}} \mathbf{x}_i^{(u)\mathrm{T}} (\mathbf{y}_k - \mathbf{X}^{(u)} \widehat{\mathbf{u}}_k) = \frac{1}{\sqrt{n}} \lambda^{(u)} w_{ik}^{(u)} + 2(1-\alpha) \frac{1}{\sqrt{n}} \lambda_k^{(n)} \widehat{u}_{ik} \sum \widehat{v}_{jk}^2) \to 0.
$$

To prove the result in part (ii), we re-define the objective function in (2.23) according

to the original parameterization in (2.6),

$$Q_k^{(n)}(d_k;\mathbf{u}_k,\mathbf{v}_k)=\frac{1}{2}\|\mathbf{Y}_k-d_k\mathbf{X}\mathbf{u}_k\mathbf{v}_k^{\mathrm{T}}\|_F^2+\alpha\lambda_k^{(n)}\sum_{i=1}^{p}\sum_{j=1}^{q}d_k w_{ijk}|u_{ik}v_{jk}| \\ +(1-\alpha)\lambda_k^{(n)}\sum_{i=1}^{p}\sum_{j=1}^{q}(d_k u_{ik}v_{jk})^2, \tag{2.33}$$

such that $\mathbf{u}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u}_k/n=1$ and $\mathbf{v}_k^{\mathrm{T}}\mathbf{v}_k=1$. Now for any $k\in\{r^*+1,\ldots,r\}$, the KKT optimality condition for $\widehat{d}_k$ gives

$$\frac{1}{\sqrt{n}}\mathrm{vec}^{\mathrm{T}}(\mathbf{X}\widehat{\mathbf{u}}_k\widehat{\mathbf{v}}_k^{\mathrm{T}})(\mathbf{y}_k-\widehat{d}_k\mathrm{vec}(\mathbf{X}\widehat{\mathbf{u}}_k\widehat{\mathbf{v}}_k^{\mathrm{T}}))=\frac{\lambda_k^{(n)}}{\sqrt{n}}\sum_{i=1}^{p}\sum_{j=1}^{q}w_{ijk}|\widehat{u}_{ik}\widehat{v}_{jk}|+\frac{2(1-\alpha)\lambda_k^{(n)}}{\sqrt{n}}\widehat{d}_k\|\widehat{\mathbf{u}}_k\|_2^2\|\widehat{\mathbf{v}}_k\|_2^2, \tag{2.34}$$

where $\mathbf{y}_k=\mathrm{vec}(\mathbf{Y}-\mathbf{X}\sum_{l<k}\widehat{d}_l\widehat{\mathbf{u}}_l\widehat{\mathbf{v}}_l^{\mathrm{T}})$. Taking left hand side of (2.34), we get

$$\begin{aligned}\text{LHS}&=\mathrm{vec}^{\mathrm{T}}(\widehat{\mathbf{u}}_k\widehat{\mathbf{v}}_k^{\mathrm{T}})\Big(\mathbf{I}_q\otimes\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n}\Big)\left\{\sqrt{n}\sum_{l=1}^{r^*}\mathrm{vec}(d_l^*\mathbf{u}_l^*\mathbf{v}_l^{*\mathrm{T}}-\widehat{d}_l\widehat{\mathbf{u}}_l\widehat{\mathbf{v}}_l^{\mathrm{T}})-\sqrt{n}\sum_{l=r^*+1}^{k}\widehat{d}_l\mathrm{vec}(\widehat{\mathbf{u}}_l\widehat{\mathbf{v}}_l^{\mathrm{T}})\right\}\\&\quad+\mathrm{vec}^{\mathrm{T}}(\widehat{\mathbf{u}}_k\widehat{\mathbf{v}}_k^{\mathrm{T}})\mathrm{vec}(\frac{\mathbf{X}^{\mathrm{T}}\mathbf{E}}{\sqrt{n}})\\&\quad=O_p(1).\end{aligned}$$

But as $n\to\infty$, the right hand side of (2.34)

$$\begin{aligned}\text{RHS} &= \frac{1}{\sqrt{n}}\lambda_k^{(n)}\sum_{i=1}^{p}\sum_{j=1}^{q} w_{ijk}|\widehat{u}_{ik}\widehat{v}_{jk}| + (1-\alpha)\frac{2}{\sqrt{n}}\lambda_k^{(n)}\widehat{d}_k\|\widehat{\mathbf{u}}_k\|_2^2\|\widehat{\mathbf{v}}_k\|_2^2 \\ &= \frac{\lambda_k^{(n)}}{\sqrt{n}}n^{\gamma/2}\sum_{i=1}^{p}\sum_{j=1}^{q}|\sqrt{n}O_p(\widetilde{d}_k)|^{-\gamma}|\widehat{u}_{ik}\widehat{v}_{jk}| + (1-\alpha)\frac{2}{\sqrt{n}}\lambda_k^{(n)}\widehat{d}_k\|\widehat{\mathbf{u}}_k\|_2^2\|\widehat{\mathbf{v}}_k\|_2^2 \\ &\to_p \infty.\end{aligned}$$

Hence, $P(\widehat{d}_k = 0) \to 0$ as $n \to \infty$, for $k = r^* + 1, \ldots, r$ This completes the proof. □

2.5.2 A Non-Asymptotic Error Bound

We briefly discuss the non-asymptotic behaviors of the SeCURE estimators. The asymptotic convergence of $\mathbf{X}^{\mathrm{T}}\mathbf{X}/n$ in **A1** is no longer assumed. Consequently, we consider the model

$$\mathbf{Y} = \mathbf{X}\mathbf{C}^* + \mathbf{E} = \mathbf{X}\mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*\mathrm{T}} + \mathbf{E}, \qquad \text{s.t. } \mathbf{U}^{*\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{U}^*/n = \mathbf{I}_{r^*}, \mathbf{V}^{*\mathrm{T}}\mathbf{V}^* = \mathbf{I}_{r^*}. \tag{2.35}$$

To focus on the fundamentals, we study SeCURE estimation using the non-adaptive elastic net penalty, i.e., $w_{ijk} = 1$. Let $\mathbf{C}^* = \sum_{k=1}^{r^*}\mathbf{C}_k^*$ with $\mathbf{C}_k^* = d_k^*\mathbf{u}_k^*\mathbf{v}_k^{*\mathrm{T}}$, and denote the corresponding SeCURE estimators by solving (2.12) as $\widehat{\mathbf{C}}_k$. Denote $\lambda_1 = \lambda\alpha$ and $\lambda_2 = \lambda(1-\alpha)$. Let $\mathbf{P}_{\mathbf{x}}$ denote the projection matrix unto the column space of $\mathbf{X}$. Let $\lambda_{\min}(\mathbf{H})$ and $\lambda_{\max}(\mathbf{H})$ denote the minimum and maximum eigenvalues of a square matrix

$\mathbf{H}$. Let $d_k(\mathbf{H})$ denote the kth largest singular value of a matrix $\mathbf{H}$. We consider the following conditions.

B1. $b \leq \lambda_{\min}(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n) \leq \lambda_{\max}(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n) \leq B$, where b and B are two positive constants.

B2. The error matrix $\mathbf{E}$ has independent $\mathrm{N}(0, \sigma^2)$ entries.

Our main result is presented in Theorem 2.16.

Theorem 2.16. *Suppose **B1** and **A3** are satisfied. For $k = 1, \ldots, r^*$,*

$$\begin{aligned}\|\widehat{\mathbf{C}}_k - \mathbf{C}_k^*\|_F \leq & \frac{\lambda_1\sqrt{pq} + 2\sqrt{2}d_1(\mathbf{P}_{\mathbf{x}}\mathbf{E})\sqrt{\lambda_{\max}} + 2\sqrt{2}\sqrt{n}d_{k+1}^*\sqrt{\lambda_{\max}} + 2\sqrt{2}\lambda_2\|\mathbf{C}_k^*\|_F}{\lambda_{\min} + \lambda_2} \\ & + \frac{2\sqrt{2}\lambda_{\max}}{\lambda_{\min} + \lambda_2}(\sum_{h=1}^{k-1}\|\widehat{\mathbf{C}}_h - \mathbf{C}_h^*\|_F). \end{aligned} \tag{2.36}$$

Proof. By definition,

$$\|\mathbf{Y} - \mathbf{X}\widehat{\mathbf{C}}_1\|_F^2 + \lambda_1\|\widehat{\mathbf{C}}_1\|_1 + \lambda_2\|\widehat{\mathbf{C}}_1\|_F^2 - \|\mathbf{Y} - \mathbf{X}\mathbf{C}_1^*\|_F^2 - \lambda_1\|\mathbf{C}_1^*\|_1 - \lambda_2\|\mathbf{C}_1^*\|_F^2 \leq 0,$$

which implies that

$$\begin{aligned} & \|\mathbf{Y} - \mathbf{X}\widehat{\mathbf{C}}_1\|_F^2 + \lambda_2\|\widehat{\mathbf{C}}_1\|_F^2 - \|\mathbf{Y} - \mathbf{X}\mathbf{C}_1^*\|_F^2 - \lambda_2\|\mathbf{C}_1^*\|_F^2 \\ \leq & \lambda_1(\|\mathbf{C}_1^*\|_1 - \|\widehat{\mathbf{C}}_1\|_1) \\ \leq & \lambda_1\sqrt{pq}\|\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*\|_F. \end{aligned}$$

The left hand side (LHS) can be organized as

$$\begin{aligned}\text{LHS} =&\text{tr}\{(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2\mathbf{I})(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*) - 2\mathbf{C}_1^{*\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2\mathbf{I})\mathbf{C}_1^{*\mathrm{T}} \\ &+ 2\mathbf{C}_1^{*\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2\mathbf{I})\widehat{\mathbf{C}}_1 - 2\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)\}.\end{aligned}$$

Thus we have,

$$\begin{aligned}&\text{tr}\{(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2\mathbf{I})(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)\} \\ \leq\ &\lambda_1\sqrt{pq}\|\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*\|_F - 2\text{tr}\{\mathbf{C}_1^{*\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2\mathbf{I})(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)\} + 2\text{tr}\{\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)\} \\ \leq\ &2\text{tr}\{\mathbf{E}^{\mathrm{T}}\mathbf{X}(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)\} + \lambda_1\sqrt{pq}\|\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*\|_F - 2\lambda_2\text{tr}\{\mathbf{C}_1^{*\mathrm{T}}(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)\} \\ &+ 2\text{tr}\{\mathbf{C}_{-1}^{*\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)\}. \qquad (2.37)\end{aligned}$$

Here we denote $\mathbf{C}_{-k}^* = \sum_{h=k+1}^{r^*}\mathbf{C}_h^*$ and we have used the fact that $\mathbf{Y} = \mathbf{X}\sum_{k=1}^{r^*}\mathbf{C}_k^* + \mathbf{E}$.

The LHS of (2.37) can be lower bounded by

$$(\lambda_{\min} + \lambda_2)\|(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)\|_F^2 \leq \text{tr}\{(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2\mathbf{I})(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)\}. \qquad (2.38)$$

Consider the terms on the ride hand side (RHS) of (2.37). The stochastic term is

bounded by

$$\begin{aligned}\mathrm{tr}\{\mathbf{E}^{\mathrm{T}}\mathbf{X}(\widehat{\mathbf{C}}_1-\mathbf{C}_1^*)\} &= \mathrm{tr}\{(\mathbf{P}_{\mathbf{x}}\mathbf{E})^{\mathrm{T}}\mathbf{X}(\widehat{\mathbf{C}}_1-\mathbf{C}_1^*)\}\\ &\le d_1(\mathbf{P}_{\mathbf{x}}\mathbf{E})\|\mathbf{X}(\widehat{\mathbf{C}}_1-\mathbf{C}_1^*)\|_*\\ &\le \sqrt{2}d_1(\mathbf{P}_{\mathbf{x}}\mathbf{E})\|\mathbf{X}(\widehat{\mathbf{C}}_1-\mathbf{C}_1^*)\|_F\\ &\le \sqrt{2}d_1(\mathbf{P}_{\mathbf{x}}\mathbf{E})\sqrt{\lambda_{\max}}\|(\widehat{\mathbf{C}}_1-\mathbf{C}_1^*)\|_F. \qquad (2.39)\end{aligned}$$

Here we have used the inequalities $\mathrm{tr}(\mathbf{H}_1\mathbf{H}_2) \le d_1(\mathbf{H}_1)\|\mathbf{H}_2\|_*$, $\|\mathbf{H}\|_* \le \sqrt{r(\mathbf{H})}\|\mathbf{H}\|_F$ and the fact that $r(\widehat{\mathbf{C}}_1-\mathbf{C}_1^*) \le 2$. For the remaining terms on the RHS of (2.37), we have

$$\mathrm{tr}\{\mathbf{C}_1^{*\mathrm{T}}(\widehat{\mathbf{C}}_1-\mathbf{C}_1^*)\} \le \sqrt{2}\|\mathbf{C}_1^*\|_F\|(\widehat{\mathbf{C}}_1-\mathbf{C}_1^*)\|_F, \qquad (2.40)$$

and

$$\begin{aligned}\mathrm{tr}\{\mathbf{C}_{-1}^{*\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\widehat{\mathbf{C}}_1-\mathbf{C}_1^*)\} &\le d_1(\mathbf{X}\mathbf{C}_{-1}^*)\|\mathbf{X}(\widehat{\mathbf{C}}_1-\mathbf{C}_1^*)\|_*\\ &= \sqrt{n}d_2^*\|\mathbf{X}(\widehat{\mathbf{C}}_1-\mathbf{C}_1^*)\|_*\\ &\le \sqrt{2}\sqrt{n}d_2^*\sqrt{\lambda_{\max}}\|(\widehat{\mathbf{C}}_1-\mathbf{C}_1^*)\|_F. \qquad (2.41)\end{aligned}$$

Combining (2.38)–(2.41), we obtain

$$\|\widehat{\mathbf{C}}_1-\mathbf{C}_1^*\|_F \le \frac{\lambda_1\sqrt{pq}+2\sqrt{2}d_1(\mathbf{P}_{\mathbf{x}}\mathbf{E})\sqrt{\lambda_{\max}}+2\sqrt{2}\sqrt{n}d_2^*\sqrt{\lambda_{\max}}+2\sqrt{2}\lambda_2\|\mathbf{C}_1^*\|_F}{\lambda_{\min}+\lambda_2}.$$

For subsequent steps when $k \geq 2$, the main difference is that $\mathbf{Y}$ is replaced with $\mathbf{Y} - \sum_{j=1}^{k-1} \mathbf{X}\widehat{\mathbf{C}}_j$. Following similar derivation, we obtain

$$\begin{aligned}\mathrm{tr}\{(\widehat{\mathbf{C}}_k - \mathbf{C}_k^*)^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2\mathbf{I})(\widehat{\mathbf{C}}_k - \mathbf{C}_k^*)\} \leq\ & 2\mathrm{tr}\{\mathbf{E}^{\mathrm{T}}\mathbf{X}(\widehat{\mathbf{C}}_k - \mathbf{C}_k^*)\} + \lambda_k\|\mathbf{W}_k\|_F\|\widehat{\mathbf{C}}_k - \mathbf{C}_k^*\|_F \\ & + 2\mathrm{tr}\{\mathbf{C}_{-k}^{*\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\widehat{\mathbf{C}}_1 - \mathbf{C}_1^*)\} - 2\lambda_2\mathrm{tr}\{\mathbf{C}_k^{*\mathrm{T}}(\widehat{\mathbf{C}}_k - \mathbf{C}_k^*)\} \\ & + 2\mathrm{tr}\{\sum_{h=1}^{k-1}(\mathbf{C}_h^* - \widehat{\mathbf{C}}_h)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\widehat{\mathbf{C}}_k - \mathbf{C}_k^*)\}. \end{aligned} \tag{2.42}$$

The treatments of the terms on the RHS are similar, except for the last term, which can be bounded as

$$\begin{aligned}\mathrm{tr}\{\sum_{h=1}^{k-1}(\mathbf{C}_h^* - \widehat{\mathbf{C}}_h)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\widehat{\mathbf{C}}_k - \mathbf{C}_k^*)\} &\leq \sum_{j=1}^{k-1} d_1(\mathbf{X}\mathbf{C}_j^* - \mathbf{X}\widehat{\mathbf{C}}_j)\|\mathbf{X}(\widehat{\mathbf{C}}_k - \mathbf{C}_k^*)\|_* \\ &\leq \sqrt{2}\lambda_{\max}(\sum_{h=1}^{k-1}\|\mathbf{C}_h^* - \widehat{\mathbf{C}}_h\|_F)\|\widehat{\mathbf{C}}_k - \mathbf{C}_k^*\|_F.\end{aligned}$$

The rest of the proof is similar to the case $k = 1$ and hence is omitted.

□

Corollary 2.17. *Suppose **B1**, **B2** and **A3** are satisfied. Also assume*

(a) $\lim_{n\to\infty} \lambda_2/n = 0$, $\lim_{n\to\infty} \lambda_1/\sqrt{n} = 0$;

(b) $\lim_{n\to\infty} \log(pq)/\log(n) = \nu$, $0 < \nu < 1$; $q = O(p)$; $d_k(\mathbf{X}\mathbf{C}^*) = o(\sqrt{pq})$.

Then $\widehat{\mathbf{C}}_k$ *is* $\sqrt{n}/\sqrt{pq}$ *consistent for* $\mathbf{C}_k^*$, *for* $k = 1, \ldots, r^*$.

Proof. The following lemma treats the stochastic term $d_1(\mathbf{P_x E})$, which is directly from Bunea et al. [2011] and Chen et al. [2013].

Lemma 2.18. *Suppose **B2** holds. Then* $\mathbb{E}\{d_1(\mathbf{P_x E})\} \leq \sigma(\sqrt{r_x}+\sqrt{q})$, *and for any* $t > 0$, $P\{d_1(\mathbf{P_x E}) \geq \mathbb{E}\{d_1(\mathbf{P_x E})\} + \sigma t\} \leq \exp(-t^2/2)$.

The results in Corollary 2.17 can be obtained easily by verifying $(\sqrt{n}/\sqrt{pq})\|\widehat{\mathbf{C}}_k - \mathbf{C}_k^*\|_F = o_p(1)$ using the results in Theorem 2.16 and Lemma 2.18. □

This bound reveals some interesting properties of SeCURE estimators. First, it shows that $\sqrt{n}d_{k+1}^*$ plays similar role as the stochastic term $d_1(\mathbf{P_x E})$ in the estimation of $\mathbf{C}_k^*$. The former term measures the size of the left-over signal in the model, while the latter term measures the size of the stochastic error; they both appear in the error bound because they both act as "errors" in the estimation of the kth factor. Moreover, the second term in the bound shows the phenomenon of "noise accumulation" that arises naturally due to the sequential procedure itself. This is caused by the fact in each step the ideal response matrix to use, i.e., $\mathbf{Y} - \sum_{h=1}^{k-1} \mathbf{X}\mathbf{C}_h^*$, is clearly not available, which has to be estimated by $\mathbf{Y} - \sum_{h=1}^{k-1} \mathbf{X}\widehat{\mathbf{C}}_h$. The stochastic term $d_1(\mathbf{P_x E})$ can be treated similarly as in Bunea et al. [2011]. Theorem 2.16 can then be used to establish the consistency of SeCURE with diverging model dimensions.

2.6 Simulation

2.6.1 Setups

We compare estimation, prediction and co-sparsity recovery performance of ordinary least squares (OLS), reduced-rank regression (RRR), sparse reduced-rank regression using adaptive group lasso (SRRR) by Chen and Huang [2012a], reduced-rank regression with sparse singular value decomposition using adaptive lasso (RSSVD) by Chen et al. [2012], and several versions of our proposed SeCURE approach. We consider both adaptive elastic net and adaptive lasso penalties, and the corresponding methods are denoted as SeCURE(AE) and SeCURE(AL), respectively. We also show that SeCURE can optionally enforce the exact orthogonality; the corresponding methods are denoted as SeCURE(AE*) and SeCURE(AL*).

The true rank of $\mathbf{C}^* \in \mathbb{R}^{p\times q}$ is set as $r^* = 3$, and we set $d_1^* = 20$, $d_2^* = 15$, $d_3^* = 10$. Model I is a low dimensional example with $p = 50$, $q = 25$ and $n = 400$. The $\mathbf{u}_k^*$ is generated as $\mathbf{u}_k^* = \check{\mathbf{u}}_k/\|\check{\mathbf{u}}_k\|$, where $\check{\mathbf{u}}_1 = [\text{unif}(\mathcal{A}_u, 8), \text{rep}(0, 42)]^{\mathrm{T}}$, $\check{\mathbf{u}}_2 = [\text{rep}(0, 5), \text{unif}(\mathcal{A}_u, 9), \text{rep}(0, 36)]^{\mathrm{T}}$, and $\check{\mathbf{u}}_3 = [\text{rep}(0, 11), \text{unif}(\mathcal{A}_u, 9), \text{rep}(0, 30)]^{\mathrm{T}}$; $\mathbf{v}_k^*$ is generated as $\mathbf{v}_k^* = \check{\mathbf{v}}_k/\|\check{\mathbf{v}}_k\|$, where $\check{\mathbf{v}}_1 = [\text{unif}(\mathcal{A}_v, 5), \text{rep}(0, 20)]^{\mathrm{T}}$, $\check{\mathbf{v}}_2 = [\text{rep}(0, 5), \text{unif}(\mathcal{A}_v, 5), \text{rep}(0, 15)]^{\mathrm{T}}$, and $\check{\mathbf{v}}_3 = [\text{rep}(0, 10), \text{unif}(\mathcal{A}_v, 5), \text{rep}(0, 10)]^{\mathrm{T}}$. The notation $\text{unif}(\mathcal{A}, b)$ denotes a vector of length b whose entries are i.i.d. uniformly distributed on set $\mathcal{A}$; we use $\mathcal{A}_u = \pm 1$, $\mathcal{A}_v = [-1, -0.3] \cup [0.3, 1]$. The notation $\text{rep}(a, b)$

denotes a vector of length b, whose entries are all equal to a. Model II is a high dimensional example with $p = 500$, $q = 200$ and $n = 400$, in which the $\mathbf{U}^*$ and $\mathbf{V}^*$ are obtained by appending the $\mathbf{u}_k^*$ and $\mathbf{v}_k^*$ vectors generated from Model I with zeros.

The predictor matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ is generated from the SFAR model in (2.5) and (2.6). Specifically, let $\mathbf{x} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Gamma})$ where $\mathbf{\Gamma} = [\gamma_{ij}]_{p\times p}$ with $\gamma_{ij} = 0.5^{|i-j|}$. Denote $\mathbf{x}_1 = \mathbf{U}^{\mathrm{T}}\mathbf{x}$; then $\mathbf{x}_1 \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{r^*})$. As such, we first generate $\mathbf{X}_1 \in \mathbb{R}^{n\times r^*}$ by drawing n random samples from $\mathrm{N}(\mathbf{0}, \mathbf{I}_{r^*})$. Given a generated $\mathbf{U}^*$, we can find a $\mathbf{U}_\perp^* \in \mathbb{R}^{p\times(p-r^*)}$ such that $\mathbf{P} = [\mathbf{U}^*, \mathbf{U}_\perp^*] \in \mathbb{R}^{p\times p}$ and $\mathrm{rank}(\mathbf{P}) = p$. It follows that $\mathbf{P}^{\mathrm{T}}\mathbf{x} \sim \mathrm{N}(\mathbf{0}, \mathbf{P}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{P})$. Let $\mathbf{x}_2 = \mathbf{U}_\perp^{*\mathrm{T}}\mathbf{x}$; we can then generate $\mathbf{X}_2 \in \mathbb{R}^{n\times(p-r^*)}$ by drawing n random samples from the conditional distribution of $\mathbf{x}_2$ given $\mathbf{x}_1$. Finally, we obtain $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]\mathbf{P}^{-1}$.

The rows of the error matrix $\mathbf{E}$ are generated as i.i.d samples from $\mathrm{N}(0, \sigma^2\mathbf{\Delta})$, where $\mathbf{\Delta} = [\delta_{ij}]_{q\times q}$ with $\delta_{ij} = \rho^{|i-j|}$ and we experiment with $\rho \in \{0, 0.3, 0.5\}$. The response matrix $\mathbf{Y}$ is then generated by $\mathbf{Y} = \mathbf{X}\mathbf{C}^* + \mathbf{E}$. We set σ according to a given signal to noise ratio (SNR) defined as $\mathrm{SNR} = \|d_{r^*}\mathbf{X}\mathbf{u}_{r^*}^*\mathbf{v}_{r^*}^{*\mathrm{T}}\|_2/\|\mathbf{E}\|_F$, i.e., $\mathrm{SNR} \in \{0.25, 0.5\}$. The missing entries in $\mathbf{Y}$ are randomly selected, and we consider missing proportions $\mathrm{M}\% \in \{0\%, 15\%, 30\%\}$. The experiment was replicated 300 times under each setup.

The estimation accuracy is measured by $\mathrm{Er}(\mathbf{C}) = \|\widehat{\mathbf{C}} - \mathbf{C}\|_F^2/(pq)$, $\mathrm{Er}(\mathbf{X}\mathbf{C}) = \|\mathbf{\Gamma}^{\frac{1}{2}}(\widehat{\mathbf{C}}-\mathbf{C})\|_F^2/(nq)$. The sparsity recovery in the decomposition of the coefficient matrix is characterized by the false positive rate (FPR) and the false negative rate (FNR), calculated from comparing the sparsity pattern of $(\widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k)$ to that of $(\mathbf{u}_k^*, \mathbf{v}_k^*)$, $k = 1, \ldots, r^*$. We evaluate rank estimation performance using the (relative) percentage of signal left

in the (r^*+1)th estimated component, i.e., $\mathrm{R}\% = 100(\widehat{d}^2_{r^*+1}/\sum_{k=1}^{r^*+1}\widehat{d}^2_k)$. Finally, the orthogonality of the SeCURE estimates is measured by $\mathrm{ORT} = \|\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\mathbf{U}}/n\|_1 + \|\widehat{\mathbf{V}}^{\mathrm{T}}\widehat{\mathbf{V}}\|_1 - 2\widehat{r}$.

2.6.2 Simulation Results

Tables 1 – 6 report the simulation results for Models I and II, under the setting that $\mathrm{SNR} = \{0.25, 0.50\}$ and $\rho = \{0, 0.3, 0.5\}$. Consider first the results in case of complete data. SeCURE methods outperform all other methods in both predictive accuracy and sparse recovery. As expected, LS performs the worst; RRR improves LS greatly but is outperformed by other sparse and low-rank approaches. Both RSSVD and SRRR do not perform well in sparsity recovery, because the former searches for the sparsity pattern according to the SVD of $\mathbf{C}^*$ rather than that of $\mathbf{XC}^*$, and the latter only considers rowwise sparsity in $\mathbf{C}^*$. All the regularized methods perform well in reducing the rank, as R% is mostly close to zero.

Neither RSSVD nor SRRR can handle incomplete data, so they are dropped from the comparison when $\mathrm{M}\% \in \{15\%, 30\%\}$. As a benchmark, we implemented an algorithm for fitting RRR with incomplete data using the idea of matrix completion, which can be regarded as a special case of SeCURE for setting $\lambda = 0$. As expected, the performance of all methods deteriorates when the percentage of missing increases, but SeCURE still performs well and always greatly outperforms RRR.

In general, the performance of SeCURE using adaptive lasso is comparable to that

of using adaptive elastic net; the former tends to perform slightly better in estimation and prediction. We still recommend to use elastic net in practice, because it better ensures the stability in optimization as shown in our convergence analysis. Also, when computing power is adequate, tuning α may further improve the elastic net.

Figure 2 and 3 compares different versions of SeCURE in terms of the prediction and the orthogonality measures using Model I and II, respectively. Here we consider SeCURE using either non-adaptive elastic net or adaptive elastic net, and with or without orthogonality constraints. (The lasso versions are omitted). The adaptive penalization in general improves model estimation. The exact orthogonality, when desired, can be efficiently achieved by SeCURE. However, as seen from Tables 1–6, enforcing orthogonality of the sample factors, in general, may hurt both model estimation and sparsity recovery. This clearly demonstrates the advantage of SeCURE, as it is able to bypass the orthogonality in the recovery of the co-sparsity pattern.

SeCURE is more computationally efficient than its closest competitor RSSVD, and the efficiency gain increases with the model dimension. We have also experimented with different SNR values, error correlations and missing proportions, and the results are all consistent with those reported herein. As expected, the performance of the methods improves when SNR becomes higher, the missing proportions decreases or the error correlation weakens; the error correlation impacts both the accuracy and the variability in model estimation slightly.

Table 1: Simulation: results of Model I with $\rho = 0$. For better presentation, Er($\mathbf{C}$) and Er($\mathbf{XC}$) are scaled by multiplying 10^4.

Method	Er(C)	Er(XC)	FPR	FNR	R%
	Missing= 0%, SNR = 0.25				
LS	2994.8 (191.0)	228.4 (13.9)	100.0 (0.0)	0.0 (0.0)	–
RRR	757.0 (75.9)	61.7 (5.8)	100.0 (0.0)	0.0 (0.0)	4.4 (0.7)
SeCURE(AL*)	115.4 (33.3)	10.4 (2.8)	1.2 (0.8)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	93.5 (27.9)	8.2 (2.3)	0.2 (0.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	119.8 (36.0)	10.7 (3.0)	1.5 (1.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	98.4 (29.8)	8.6 (2.5)	0.3 (0.4)	0.0 (0.0)	0.0 (0.0)
RSSVD	308.5 (109.3)	27.5 (9.5)	8.3 (5.8)	2.9 (3.5)	0.0 (0.0)
SRRR	366.9 (48.2)	32.5 (4.2)	53.9 (2.1)	0.0 (0.0)	2.2 (0.5)
	Missing= 0%, SNR = 0.5				
LS	748.7 (47.8)	57.1 (3.5)	100.0 (0.0)	0.0 (0.0)	–
RRR	186.3 (18.7)	15.2 (1.4)	100.0 (0.0)	0.0 (0.0)	1.2 (0.2)
SeCURE(AL*)	34.9 (11.6)	3.4 (1.2)	1.5 (1.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	19.0 (5.9)	1.7 (0.5)	0.1 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	34.8 (11.3)	3.3 (1.1)	1.6 (1.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	19.7 (6.4)	1.7 (0.5)	0.1 (0.3)	0.0 (0.0)	0.0 (0.0)
RSSVD	92.3 (24.0)	8.1 (2.1)	15.0 (5.3)	0.0 (0.0)	0.0 (0.0)
SRRR	90.9 (11.2)	7.9 (1.0)	51.1 (0.0)	0.0 (0.0)	0.6 (0.1)
	Missing= 15%, SNR = 0.25				
RRR	906.7 (92.3)	73.4 (7.4)	100.0 (0.0)	0.0 (0.0)	2.9 (0.3)
SeCURE(AL*)	138.4 (40.1)	12.4 (3.5)	1.4 (0.9)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	115.6 (35.5)	10.1 (2.9)	0.5 (0.6)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	142.7 (41.9)	12.8 (3.6)	1.7 (1.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	119.9 (35.9)	10.6 (3.1)	0.7 (0.7)	0.0 (0.0)	0.0 (0.0)
	Missing= 15%, SNR = 0.5				
RRR	223.2 (23.2)	18.0 (1.8)	100.0 (0.0)	0.0 (0.0)	0.8 (0.1)
SeCURE(AL*)	39.1 (12.5)	3.8 (1.2)	1.7 (1.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	23.3 (7.4)	2.1 (0.7)	0.2 (0.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	39.2 (12.6)	3.8 (1.2)	1.6 (1.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	24.0 (7.7)	2.2 (0.7)	0.2 (0.3)	0.0 (0.0)	0.0 (0.0)
	Missing= 30%, SNR = 0.25				
RRR	1123.6 (119.2)	91.1 (9.2)	100.0 (0.0)	0.0 (0.0)	3.5 (0.3)
SeCURE(AL*)	174.8 (52.0)	15.8 (4.6)	1.9 (1.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	153.3 (45.7)	13.5 (3.8)	0.9 (0.8)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	177.0 (49.6)	16.2 (4.6)	2.3 (1.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	157.8 (46.7)	13.9 (3.9)	1.1 (0.8)	0.0 (0.0)	0.0 (0.0)
	Missing= 30%, SNR = 0.5				
RRR	274.5 (29.6)	22.2 (2.2)	100.0 (0.0)	0.0 (0.0)	1.0 (0.1)
SeCURE(AL*)	44.9 (13.1)	4.4 (1.4)	1.9 (1.4)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	29.9 (8.8)	2.8 (0.9)	0.4 (0.7)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	45.1 (13.3)	4.4 (1.4)	2.0 (1.4)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	30.3 (9.0)	2.8 (0.9)	0.5 (0.6)	0.0 (0.0)	0.0 (0.0)

Table 2: Simulation: results of Model I with $\rho = 0.3$. For better presentation, Er($\mathbf{C}$) and Er($\mathbf{XC}$) are scaled by multiplying 10^4.

Method	Er(C)	Er(XC)	FPR	FNR	R%
	Missing= 0%, SNR = 0.25				
LS	2994.4 (208.0)	227.7 (14.2)	100.0 (0.0)	0.0 (0.0)	–
RRR	810.7 (96.4)	65.6 (7.1)	100.0 (0.0)	0.0 (0.0)	5.1 (0.8)
SeCURE(AL*)	117.7 (38.7)	10.5 (3.2)	1.2 (0.8)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	92.9 (29.8)	8.1 (2.5)	0.3 (0.4)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	121.1 (39.1)	10.8 (3.3)	1.5 (1.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	96.4 (30.4)	8.4 (2.5)	0.5 (0.6)	0.0 (0.0)	0.0 (0.0)
RSSVD	317.8 (113.6)	28.6 (10.4)	8.2 (5.7)	3.4 (4.3)	0.0 (0.0)
SRRR	388.2 (55.5)	34.2 (4.6)	55.4 (3.0)	0.0 (0.0)	2.5 (0.5)
	Missing= 0%, SNR = 0.5				
LS	748.6 (52.0)	56.9 (3.5)	100.0 (0.0)	0.0 (0.0)	–
RRR	198.8 (23.5)	16.1 (1.7)	100.0 (0.0)	0.0 (0.0)	1.4 (0.2)
SeCURE(AL*)	34.9 (12.5)	3.4 (1.2)	1.5 (1.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	19.2 (6.5)	1.7 (0.5)	0.1 (0.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	35.1 (12.6)	3.4 (1.2)	1.6 (1.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	19.9 (7.1)	1.7 (0.6)	0.1 (0.3)	0.0 (0.0)	0.0 (0.0)
RSSVD	92.8 (24.5)	8.3 (2.2)	14.9 (5.3)	0.0 (0.0)	0.0 (0.0)
SRRR	95.0 (12.5)	8.2 (1.0)	51.1 (0.0)	0.0 (0.0)	0.7 (0.1)
	Missing= 15%, SNR = 0.25				
RRR	951.8 (113.3)	76.8 (8.5)	100.0 (0.0)	0.0 (0.0)	3.3 (0.4)
SeCURE(AL*)	137.9 (43.1)	12.7 (4.1)	1.5 (1.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	113.3 (36.4)	10.2 (3.3)	0.6 (0.7)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	140.0 (42.3)	12.8 (3.9)	1.7 (1.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	118.3 (37.0)	10.7 (3.4)	0.7 (0.7)	0.0 (0.0)	0.0 (0.0)
	Missing= 15%, SNR = 0.5				
RRR	232.9 (27.6)	18.8 (2.0)	100.0 (0.0)	0.0 (0.0)	0.9 (0.1)
SeCURE(AL*)	39.0 (13.2)	3.8 (1.3)	1.6 (1.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	23.3 (8.0)	2.1 (0.7)	0.2 (0.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	38.9 (12.9)	3.8 (1.3)	1.7 (1.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	24.1 (8.6)	2.2 (0.7)	0.2 (0.4)	0.0 (0.0)	0.0 (0.0)
	Missing= 30%, SNR = 0.25				
RRR	1151.7 (118.3)	93.8 (9.7)	100.0 (0.0)	0.0 (0.0)	3.8 (0.4)
SeCURE(AL*)	173.1 (51.4)	15.9 (4.8)	2.0 (1.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	153.4 (50.7)	13.6 (4.3)	0.9 (0.8)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	176.9 (52.0)	16.1 (4.8)	2.3 (1.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	154.9 (47.4)	14.0 (4.3)	1.3 (1.0)	0.0 (0.0)	0.0 (0.0)
	Missing= 30%, SNR = 0.5				
RRR	280.6 (29.0)	22.8 (2.3)	100.0 (0.0)	0.0 (0.0)	1.0 (0.1)
SeCURE(AL*)	45.3 (14.0)	4.4 (1.4)	2.0 (1.4)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	29.9 (9.4)	2.7 (0.9)	0.5 (0.7)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	45.7 (14.4)	4.4 (1.4)	2.1 (1.5)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	30.6 (9.8)	2.8 (0.9)	0.5 (0.7)	0.0 (0.0)	0.0 (0.0)

Table 3: Simulation: results of Model I with $\rho = 0.5$. For better presentation, Er($\mathbf{C}$) and Er($\mathbf{XC}$) are scaled by multiplying 10^4.

Method	Er(C)	Er(XC)	FPR	FNR	R%
	Missing= 0%, SNR = 0.25				
LS	2992.9 (233.5)	227.6 (15.5)	100.0 (0.0)	0.0 (0.0)	–
RRR	917.5 (127.7)	73.9 (9.7)	100.0 (0.0)	0.0 (0.0)	6.3 (1.2)
SeCURE(AL*)	121.0 (44.9)	10.8 (3.8)	1.3 (1.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	94.1 (35.8)	8.2 (3.0)	0.3 (0.4)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	124.0 (44.9)	11.1 (4.0)	1.5 (1.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	98.8 (37.5)	8.6 (3.1)	0.6 (0.7)	0.0 (0.0)	0.0 (0.0)
RSSVD	341.4 (123.4)	31.2 (11.3)	8.6 (5.7)	3.9 (4.6)	0.0 (0.0)
SRRR	436.0 (72.1)	38.2 (5.9)	58.1 (3.5)	0.0 (0.0)	3.1 (0.7)
	Missing= 0%, SNR = 0.5				
LS	748.2 (58.4)	56.9 (3.9)	100.0 (0.0)	0.0 (0.0)	–
RRR	223.6 (31.4)	17.9 (2.3)	100.0 (0.0)	0.0 (0.0)	1.8 (0.4)
SeCURE(AL*)	34.5 (12.7)	3.3 (1.2)	1.5 (1.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	19.2 (7.2)	1.7 (0.6)	0.2 (0.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	34.7 (12.6)	3.4 (1.2)	1.3 (0.8)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	19.9 (7.9)	1.8 (0.7)	0.2 (0.3)	0.0 (0.0)	0.0 (0.0)
RSSVD	93.5 (24.9)	8.2 (2.1)	15.1 (5.4)	0.0 (0.0)	0.0 (0.0)
SRRR	103.2 (16.3)	8.9 (1.3)	51.1 (0.0)	0.0 (0.0)	0.8 (0.2)
	Missing= 15%, SNR = 0.25				
RRR	1055.1 (143.0)	84.4 (10.6)	100.0 (0.0)	0.0 (0.0)	4.0 (0.5)
SeCURE(AL*)	140.1 (46.9)	12.9 (4.3)	1.4 (1.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	118.6 (46.1)	10.7 (4.2)	0.6 (0.7)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	143.9 (47.1)	13.2 (4.4)	1.7 (1.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	123.1 (47.1)	10.9 (4.2)	0.8 (0.8)	0.0 (0.0)	0.0 (0.0)
	Missing= 15%, SNR = 0.5				
RRR	256.8 (35.0)	20.5 (2.6)	100.0 (0.0)	0.0 (0.0)	1.1 (0.2)
SeCURE(AL*)	38.5 (13.1)	3.8 (1.4)	1.7 (1.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	23.3 (8.7)	2.1 (0.8)	0.2 (0.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	38.7 (12.9)	3.8 (1.4)	1.7 (1.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	24.1 (9.3)	2.2 (0.8)	0.2 (0.4)	0.0 (0.0)	0.0 (0.0)
	Missing= 30%, SNR = 0.25				
RRR	1247.5 (148.6)	100.7 (11.5)	100.0 (0.0)	0.0 (0.0)	4.4 (0.6)
SeCURE(AL*)	179.8 (61.4)	16.6 (5.7)	2.0 (1.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	158.4 (57.4)	14.3 (5.1)	1.0 (0.9)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	181.3 (60.2)	16.8 (5.7)	2.4 (1.5)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	159.7 (53.0)	14.5 (5.0)	1.4 (1.1)	0.0 (0.0)	0.0 (0.0)
	Missing= 30%, SNR = 0.5				
RRR	302.6 (35.6)	24.3 (2.7)	100.0 (0.0)	0.0 (0.0)	1.2 (0.2)
SeCURE(AL*)	45.8 (15.1)	4.4 (1.5)	2.0 (1.5)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	30.0 (10.1)	2.8 (1.0)	0.5 (0.7)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	45.6 (14.7)	4.4 (1.5)	2.1 (1.5)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	30.7 (10.5)	2.8 (1.0)	0.5 (0.7)	0.0 (0.0)	0.0 (0.0)

Table 4: Simulation: results of Model II with $\rho = 0$. For better presentation, Er($\mathbf{C}$) and Er($\mathbf{XC}$) are scaled by multiplying 10^4.

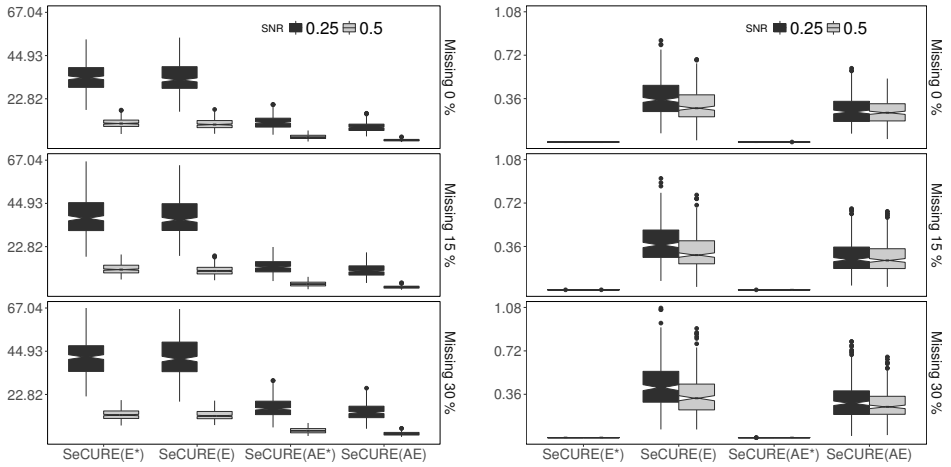
Method	Er(C)	Er(XC)	FPR	FNR	R%
	Missing= 0%, SNR = 0.25				
LS	1016.4 (38.8)	882.3 (34.5)	100.0 (0.0)	0.0 (0.0)	–
RRR	44.6 (3.1)	38.7 (2.5)	100.0 (0.0)	0.0 (0.0)	13.5 (1.1)
SeCURE(AL*)	0.4 (0.2)	0.4 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.1 (0.0)	0.1 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.4 (0.2)	0.4 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.2 (0.0)	0.1 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
RSSVD	1.9 (0.8)	1.6 (0.7)	1.3 (0.6)	0.7 (1.5)	0.7 (1.0)
SRRR	2.5 (0.2)	2.5 (0.3)	30.1 (0.0)	0.0 (0.0)	7.5 (3.7)
	Missing= 0%, SNR = 0.5				
LS	264.4 (9.8)	229.3 (8.9)	100.0 (0.0)	0.0 (0.0)	–
RRR	21.2 (1.5)	18.1 (1.3)	100.0 (0.0)	0.0 (0.0)	5.2 (0.8)
SeCURE(AL*)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
RSSVD	0.8 (0.3)	0.7 (0.2)	1.4 (0.4)	0.0 (0.0)	0.0 (0.0)
SRRR	0.6 (0.1)	0.6 (0.1)	30.1 (0.0)	0.0 (0.0)	0.4 (0.1)
	Missing= 15%, SNR = 0.25				
RRR	52.3 (3.5)	45.4 (3.1)	100.0 (0.0)	0.0 (0.0)	3.4 (0.2)
SeCURE(AL*)	0.5 (0.2)	0.5 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.2 (0.1)	0.2 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.5 (0.2)	0.5 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.2 (0.1)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Missing= 15%, SNR = 0.5				
RRR	23.1 (1.5)	19.7 (1.3)	100.0 (0.0)	0.0 (0.0)	0.9 (0.1)
SeCURE(AL*)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.0 (0.0)	0.0 (0.0)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.0 (0.0)	0.0 (0.0)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)
	Missing= 30%, SNR = 0.25				
RRR	72.2 (7.9)	62.9 (6.9)	100.0 (0.0)	0.0 (0.0)	4.0 (0.2)
SeCURE(AL*)	0.6 (0.2)	0.6 (0.2)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.3 (0.1)	0.3 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.7 (0.2)	0.6 (0.2)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.4 (0.1)	0.3 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Missing= 30%, SNR = 0.5				
RRR	29.5 (3.8)	25.4 (3.4)	100.0 (0.0)	0.0 (0.0)	1.3 (0.4)
SeCURE(AL*)	0.2 (0.1)	0.2 (0.1)	0.3 (0.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.1 (0.0)	0.1 (0.0)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.2 (0.1)	0.2 (0.1)	0.3 (0.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.1 (0.0)	0.1 (0.0)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)

Table 5: Simulation: results of Model II with $\rho = 0.3$. For better presentation, Er($\mathbf{C}$) and Er($\mathbf{XC}$) are scaled by multiplying 10^4.

Method	Er(C)	Er(XC)	FPR	FNR	R%
	Missing= 0%, SNR = 0.25				
LS	1015.8 (40.2)	882.6 (34.7)	100.0 (0.0)	0.0 (0.0)	–
RRR	48.1 (4.1)	41.8 (3.5)	100.0 (0.0)	0.0 (0.0)	13.7 (1.4)
SeCURE(AL*)	0.4 (0.2)	0.4 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.1 (0.1)	0.1 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.4 (0.2)	0.4 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.2 (0.1)	0.1 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
RSSVD	1.9 (0.9)	1.7 (0.8)	1.3 (0.6)	0.6 (1.4)	0.2 (0.4)
SRRR	2.5 (0.3)	2.5 (0.3)	30.1 (0.0)	0.0 (0.0)	11.0 (1.5)
	Missing= 0%, SNR = 0.5				
LS	264.4 (10.4)	229.2 (9.1)	100.0 (0.0)	0.0 (0.0)	–
RRR	22.0 (1.6)	18.8 (1.3)	100.0 (0.0)	0.0 (0.0)	6.4 (0.9)
SeCURE(AL*)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
RSSVD	0.8 (0.3)	0.7 (0.2)	1.4 (0.4)	0.0 (0.0)	0.0 (0.0)
SRRR	0.6 (0.1)	0.6 (0.1)	30.1 (0.0)	0.0 (0.0)	0.5 (0.1)
	Missing= 15%, SNR = 0.25				
RRR	55.2 (4.2)	48.2 (3.6)	100.0 (0.0)	0.0 (0.0)	3.9 (0.2)
SeCURE(AL*)	0.5 (0.2)	0.5 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.2 (0.1)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.5 (0.2)	0.5 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.2 (0.1)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Missing= 15%, SNR = 0.5				
RRR	23.8 (1.6)	20.4 (1.4)	100.0 (0.0)	0.0 (0.0)	1.1 (0.1)
SeCURE(AL*)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Missing= 30%, SNR = 0.25				
RRR	75.1 (9.1)	65.4 (7.8)	100.0 (0.0)	0.0 (0.0)	4.5 (0.2)
SeCURE(AL*)	0.6 (0.3)	0.6 (0.2)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.3 (0.1)	0.3 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.7 (0.3)	0.6 (0.2)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.4 (0.1)	0.3 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Missing= 30%, SNR = 0.5				
RRR	28.7 (2.9)	24.7 (2.7)	100.0 (0.0)	0.0 (0.0)	1.2 (0.1)
SeCURE(AL*)	0.2 (0.1)	0.2 (0.1)	0.3 (0.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.1 (0.0)	0.1 (0.0)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.2 (0.1)	0.2 (0.1)	0.3 (0.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.1 (0.0)	0.1 (0.0)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)

Table 6: Simulation: results of Model II with $\rho = 0.5$. For better presentation, Er($\mathbf{C}$) and Er($\mathbf{XC}$) are scaled by multiplying 10^4.

Method	Er(C)	Er(XC)	FPR	FNR	R%
	Missing= 0%, SNR = 0.25				
LS	1017.1 (41.6)	883.1 (36.5)	100.0 (0.0)	0.0 (0.0)	–
RRR	55.1 (5.8)	47.9 (4.8)	100.0 (0.0)	0.0 (0.0)	13.3 (2.0)
SeCURE(AL*)	0.4 (0.2)	0.4 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.1 (0.1)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.4 (0.2)	0.4 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.2 (0.1)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
RSSVD	2.6 (1.5)	2.2 (1.1)	2.0 (1.2)	2.0 (3.0)	1.7 (2.7)
SRRR	2.6 (0.3)	2.6 (0.4)	30.2 (0.2)	0.0 (0.0)	11.7 (1.3)
	Missing= 0%, SNR = 0.5				
LS	264.5 (10.5)	229.4 (9.2)	100.0 (0.0)	0.0 (0.0)	–
RRR	23.6 (1.9)	20.1 (1.5)	100.0 (0.0)	0.0 (0.0)	8.4 (1.1)
SeCURE(AL*)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
RSSVD	0.8 (0.4)	0.7 (0.3)	1.4 (0.5)	0.0 (0.0)	0.0 (0.0)
SRRR	0.7 (0.1)	0.6 (0.1)	30.1 (0.0)	0.0 (0.0)	0.7 (0.2)
	Missing= 15%, SNR = 0.25				
RRR	62.0 (6.0)	54.1 (5.0)	100.0 (0.0)	0.0 (0.0)	5.0 (0.3)
SeCURE(AL*)	0.5 (0.2)	0.5 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.2 (0.1)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.5 (0.2)	0.5 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.2 (0.1)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Missing= 15%, SNR = 0.5				
RRR	25.3 (1.9)	21.7 (1.6)	100.0 (0.0)	0.0 (0.0)	1.4 (0.1)
SeCURE(AL*)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.0 (0.0)	0.0 (0.0)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.2 (0.1)	0.2 (0.1)	0.3 (0.2)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.0 (0.0)	0.0 (0.0)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)
	Missing= 30%, SNR = 0.25				
RRR	82.0 (9.9)	71.6 (8.7)	100.0 (0.0)	0.0 (0.0)	5.4 (0.3)
SeCURE(AL*)	0.6 (0.3)	0.6 (0.2)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.3 (0.2)	0.3 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.7 (0.3)	0.6 (0.2)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.4 (0.2)	0.3 (0.1)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)
	Missing= 30%, SNR = 0.5				
RRR	29.7 (2.8)	25.6 (2.5)	100.0 (0.0)	0.0 (0.0)	1.5 (0.1)
SeCURE(AL*)	0.2 (0.1)	0.2 (0.1)	0.3 (0.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AL)	0.1 (0.0)	0.1 (0.0)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE*)	0.2 (0.1)	0.2 (0.1)	0.3 (0.3)	0.0 (0.0)	0.0 (0.0)
SeCURE(AE)	0.1 (0.0)	0.1 (0.0)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)



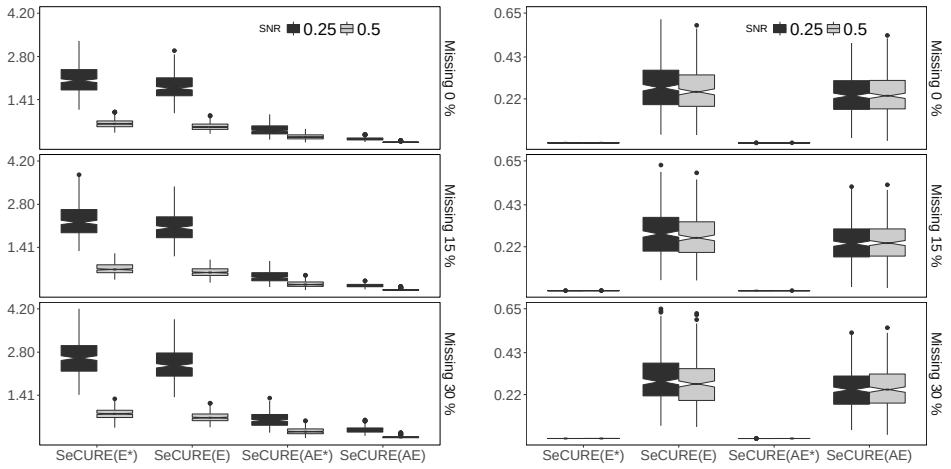
(a) Prediction error (b) Orthogonality measure

Figure 2: Simulation: Boxplots of scaled predictive measure Er($\mathbf{XC}$) (left panel) and the orthogonality measure ORT (right panel) in Model I with $\rho = 0.3$. SeCURE(E*) and SeCURE(E) denote SeCURE using non-adaptive elastic net penalty with and without orthogonality constraints, respectively.

2.7 Application

2.7.1 Biclustering with Chemotherapy Survival Data

When $\mathbf{X} = \mathbf{I}_n$, SeCURE performs sequential sparse and unit-rank approximation of a data matrix $\mathbf{Y}$, which can serve as an unsupervised learning tool for biclustering [Lee et al., 2010]. To demonstrate the effectiveness of SeCURE in biclustering, we consider a gene expression dataset from the patients with diffuse large-B-cell lymphoma (DLBCL) after chemotherapy [Rosenwald et al., 2002]. The gene expression profiles can act as molecular predictors that influence survival of patient after chemotherapy. Previous work by Hoshida et al. [2007] found three subtypes among the subjects, e.g., OxPhos (oxidative phosphorylation), BCR (Bcell response) and HR (host response). The data



(a) Prediction error (b) Orthogonality measure

Figure 3: Simulation: Boxplots of scaled Er($\mathbf{XC}$) (left panel) and ORT (right panel) in Model II and $\rho = 0.3$. SeCURE(E*) and SeCURE(E) denote SeCURE using non-adaptive elastic net penalty with and without orthogonality constraints, respectively.

consists of expression levels of $q = 661$ genes from $n = 180$ patients. Among the patients, 42, 51 and 87 of them were classified to OxPhos, BCR and HR groups, respectively. The data thus form an $n \times q$ matrix $\mathbf{Y}$ whose rows represent the subjects and columns correspond to the genes [Rosenwald et al., 2002]. Here we apply SeCURE to conduct a biclustering analysis of both the genes and the patients, to explore sets of relevant genes that can distinguish the three subtypes as well as to explore for new subtypes.

Using SeCURE, the estimated rank is $\widehat{r} = 5$. We mainly analyze the first three estimated factor components, as they explain more than 92.0% of the total variance of the first 5 components, i.e., $(\sum_{j=1}^{3} \widehat{d}_j^2)/(\sum_{j=1}^{5} \widehat{d}_j^2) \approx 0.920$. Heatmaps of the original gene expression matrix, its rank-3 approximation from SeCURE and the three estimated unit-rank components are shown in Figure 5. We include the selected genes from SeCURE and 100 randomly chosen unselected genes as reference. The genes and the subjects are properly sorted to better reveal the bicluster patterns. Overall SeCURE captures the

main patterns hidden in the original gene expression matrix. Interestingly, the first unit-rank component shows a clear contrast between a subset of patients from the HR group and a subset of patients from the OxPhos and BCR groups, while the second component appears to be mainly a contrast between the rest of the patients from the HR group and the rest of the subjects from OxPhos and BCR groups. The third component becomes ever more sparse, and the identified genes show a contrast between the OxPhos group and the BCR group. Therefore, the first two components show that there is a clear difference between the HR group and the other two groups, while the third component finds genes that differentiate the OxPhos group and the BCR group. Our findings agree well with the results from Rosenwald et al. [2002]. In addition, SeCURE reveals that the three identified groups may form smaller subgroups.

For comparison, we have also conducted biclustering via the sparse singular value decomposition (SSVD) analysis [Lee et al., 2010, Chen et al., 2012]. It turns out that SSVD leads to a much less sparse decomposition, and consequently the biclusters are not as well separated as in SeCURE. (Neither RRR nor SRRR is capable of performing biclustering). To test SeCURE with incomplete data, we have introduced 30% missing values in $\mathbf{Y}$ randomly and repeated the analysis; the patterns in the original matrix is still successfully captured. The scree plot of SeCURE, the heatmaps from SSVD and SeCURE on complete data, and the heatmaps from the incomplete data analysis with SeCURE are shown in Figures 4–6, respectively.

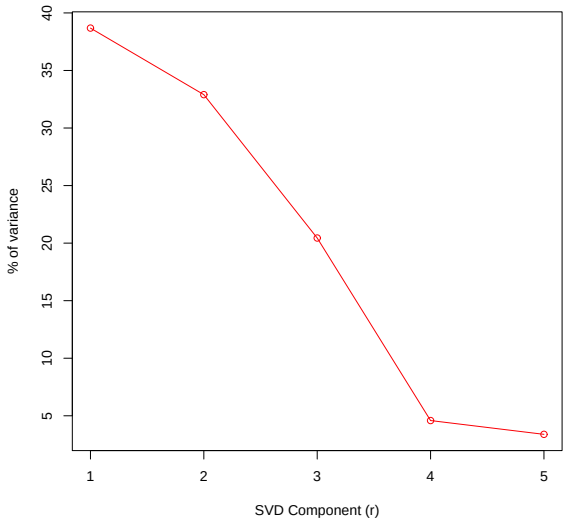


Figure 4: Chemotherapy Survival Data: Scree plot showing the relative % variances explained by the estimated latent factors from SeCURE. The first 3 factors explains 38.68%, 32.90%, 20.44% of the total variance in the first 5 estimated nonzero components, respectively.

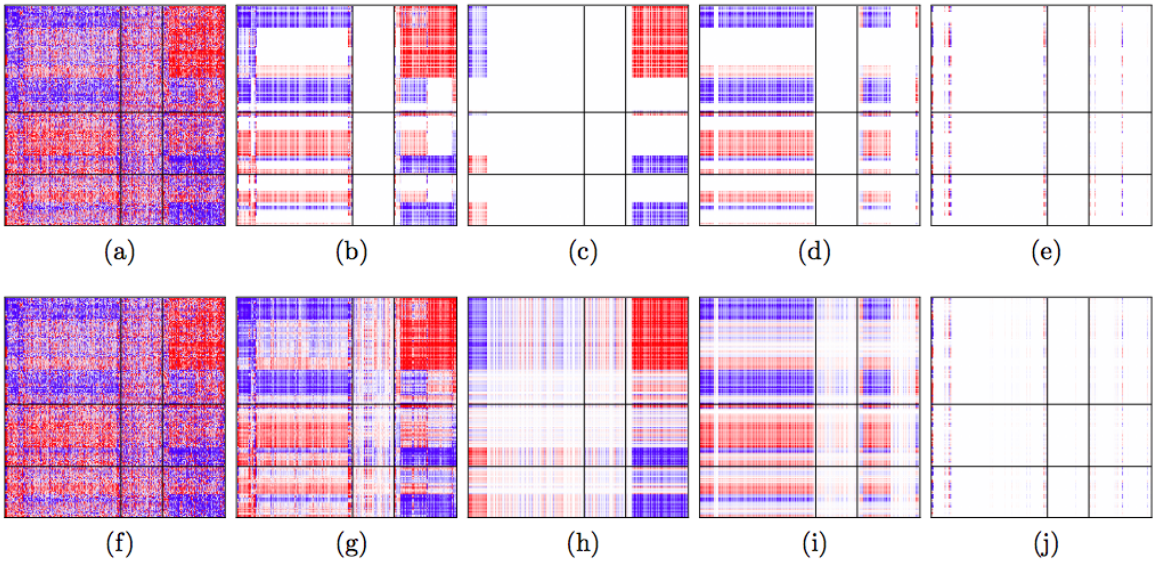


Figure 5: Chemotherapy Survival Data: Comparison of SeCURE and SSVD. Both (a) and (f) show the original gene expression matrix. (b)–(e) in the upper row show results from SeCURE: (b) the rank-3 approximation from SeCURE, and (c)–(e) the three latent components from SeCURE which sum up to (b). (g)–(j) in the lower row show results from SSVD: (g) the rank-3 approximation from SSVD, and (h)–(j) the three latent components from SSVD which sum up to (g). The horizontal line in each panel represents the three classes of patients, HR, BCR and OxPhos, from the top to the bottom. The two vertical lines indicate 100 unselected genes in between.

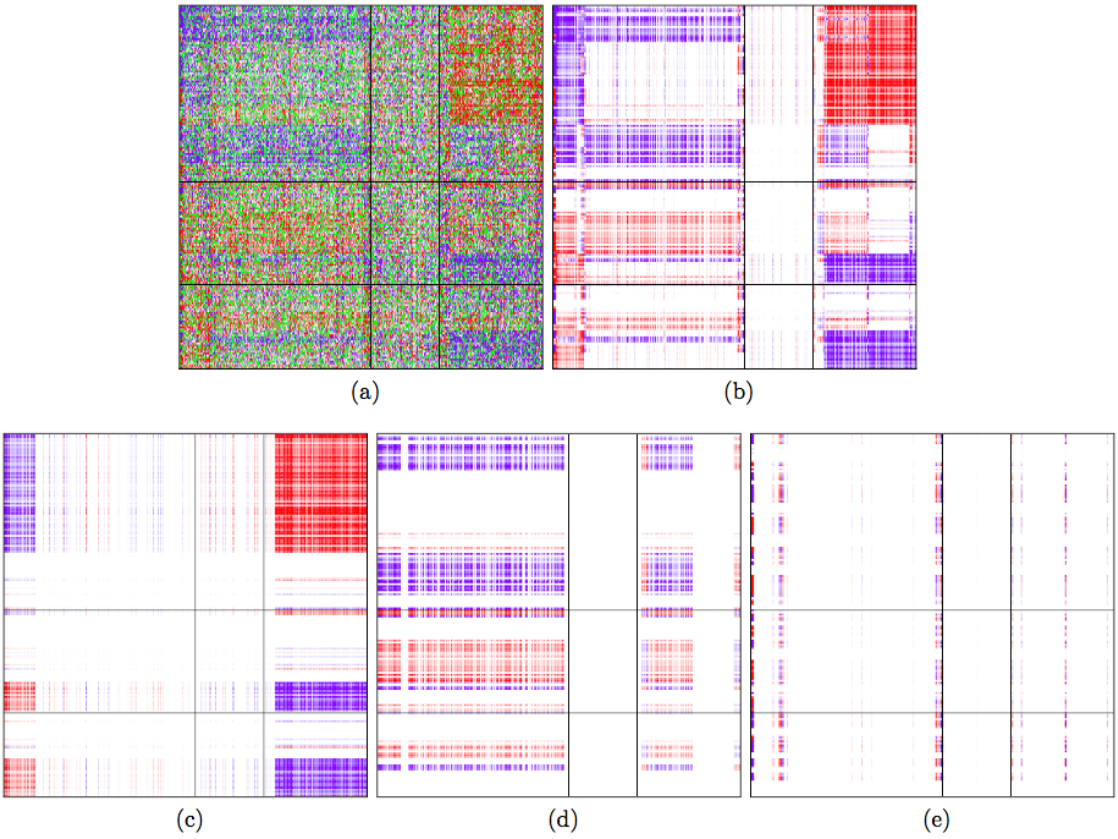


Figure 6: Chemotherapy Survival Data: Heatmaps produced by SeCURE from incomplete data with 30% entrywise missing values. For better comparison, the same set of genes of the same order are shown as in Figure 5. (a) the original gene expression matrix with 30% values missing, (b) its rank-3 approximation from SeCURE, and (c)–(e) the three latent components from SeCURE which sum up to (b). The horizontal line in each panel represents the three classes of patients, HR, BCR and OxPhos, from the top to the bottom. The two vertical lines indicate 100 unselected genes in between.

2.7.2 Yeast Cell Cycle Data

Here we consider a yeast cell cycle analysis, for identifying transcription factors regulating the RNA transcript levels of yeast genes within the eukaryotic cell cycle. The Eukariotic cell cycle data were generated using α factor arrest method, consisting of

RNA levels measured every 7 minutes for 119 minutes with a total of 18 time points covering two cell cycle of 6178 genes. The chromatin immunoprecipitation (ChIP) data ($\mathbf{X}$) [Lee et al., 2002] contain complete binding information of a subset of 1790 genes for a total of 113 transcription factors. However, the RNA data corresponding to these genes still contain about 2% missing values. Spellman et al. [1998] identified 800 genes responsible for cell cycle regulation. Chun and Kele**c**s [2010] and Chen and Huang [2012a] analyzed 524 of the 800 genes after excluding the genes with missing RNA levels and/or binding information. Since SeCURE can directly handle missing values in the response, we use all the $n = 1790$ genes, to examine the association between the RNA levels along the $q = 18$ time points and the binding information of $p = 113$ transcription factors (TF).

The estimated rank by SeCURE is $\widehat{r} = 4$. We mainly discuss the first three factors as they explain 93.7% of the variation of the four factors. Our SeCURE approach selected 51, 57, 13 TFs in the three factors, respectively, with 83 distinct TFs in total. Of these, 17 are among the 21 experimentally confirmed TFs that relate to cell cycle [Wang et al., 2007]. The SRRR and the sparse partial least squares (SPLS) selected 60 and 12 TFs, of which 16 and 7 are among the confirmed TFs, respectively [Chen and Huang, 2012a]. We have also fitted the RSSVD method with imputed data by filling the blanks using column means, and the method selected 24 TFs, of which 11 are among the confirmed ones. Figure 7 shows the estimated effects (rows in $\widehat{\mathbf{C}}$) of experimentally confirmed TFs identified by SeCURE. As expected, the effects are mostly periodic and the two cycles

are clearly seen.

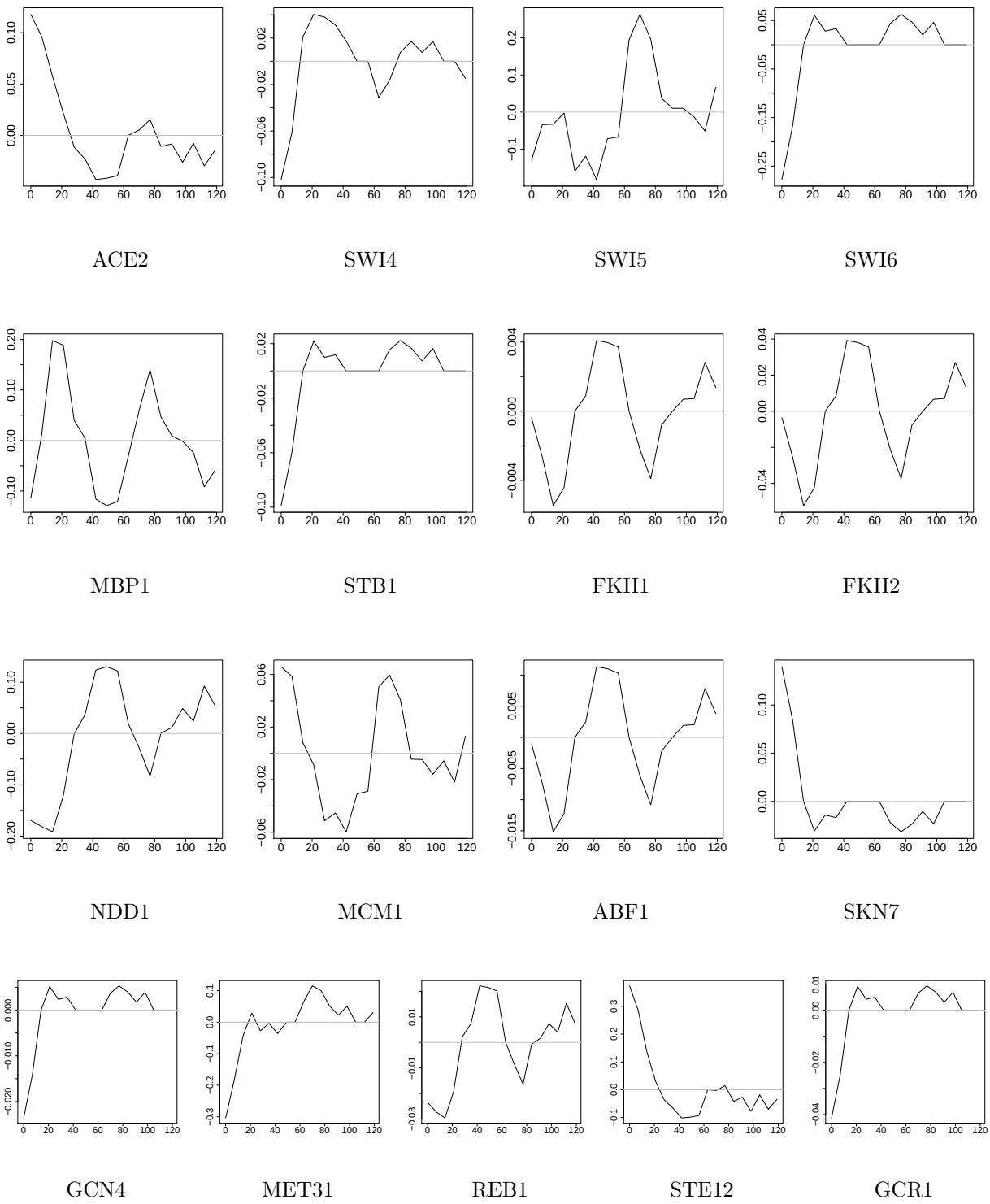


Figure 7: Yeast Cell Cycle Data: Estimated transcriptional effects of 17 experimentally confirmed TFs identified by SeCURE. Three experimentally confirmed TFs, GCR2, CBF1, BAS1 and LEU3, are not selected by SeCURE.

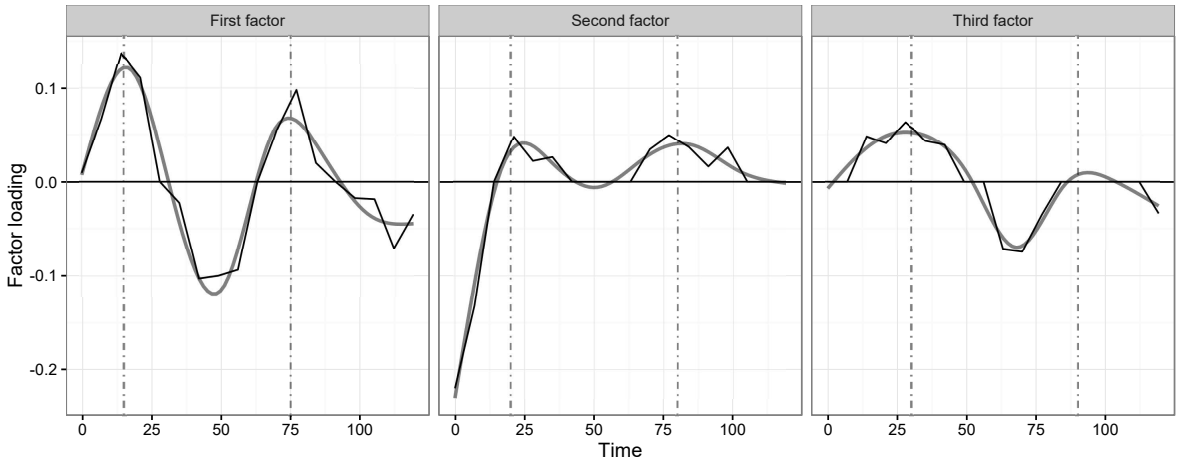


Figure 8: Yeast Cell Cycle Data: Estimated loadings of the RNAs of 18 time points on the three identified latent factors from the TFs. The fitted curves using kernel smoothing are added. The two vertical lines are drawn at 15 and 75 in the first panel, at 20 and 80 in the second panel and at 30 and 90 in the third panel.

The SeCURE analysis identified a few latent factors from the 113 TFs, given by the columns in $\mathbf{X}\widehat{\mathbf{U}}$. It is interesting to check how the RNA levels at 18 time points load on these factors, which is revealed by examining the columns of $\widehat{\mathbf{V}}\widehat{\mathbf{D}}$. Figure 8 plots the first three columns of $\widehat{\mathbf{V}}\widehat{\mathbf{D}}$ together with fitted curves using kernel smoothing. The periodic effects are very clear. Interestingly, the three curves appear to be at slightly different phases: the first curve peaks around 15 and 75 minutes, the second peaks around 20 and 80 minutes, and the third peaks around 30 and 90 minutes. In either of the first two curves, the two cycles are similar to each other in shape and magnitude. In contrast, in the third curve the two cycles are apparently different in magnitude and the second cycle almost disappears. Further examining the TFs involved in these factors may shed more light on the differential roles played by the TFs.

Chapter 3

A Greedy Algorithm for Generalized Sparse and Low-rank Recovery

3.1 Introduction

Technological advancement has lead to the collection of high-dimensional data in various fields of science and technology ranging from health and molecular biology to economics and finance. An associated problem of interest is to model dependency of multivariate outcomes/responses using observed predictors/features. Outcome variables can be either continuous, count, binary, or may be of mixed types. Also, some entries in the observed response or predictors maybe missing. The multivariate linear regression (MLR) model caters to the case when underlying process generating continuous outcome variables is assumed to be Gaussian. In other cases, when the outcome variables are either not continuous or mixed types, modeling relationship among responses and predictors is challenging because of difficulty ranging from model formulation to its estimation. This difficulty may arise due to interrelated response, complicated likelihood structure, lack

of efficient computational algorithm, inability to handle missing entries and many more.

Moreover, in high-dimensional setting, regardless of the types of response, they maybe interrelated, and also predictor variables maybe correlated or unimportant. Here, we assume that underlying process generating each of them belongs to exponential dispersion family like Gaussian, Poisson or Bernoulli. In such scenario, an *effective strategy* to induce multivariate dependency is possible through *dimension reduction*, and to discard redundant variables through *variable selection*. The desirable objectives can be attained by having a low-rank and sparse coefficient matrix.

For MLR, we developed a sequential approach SeCURE in Chapter 2, to recover such low-rank and sparse coefficient matrix. Now, consider the case of similar or mixed types of outcomes where former refers outcomes to be of same types. Motivated by SeCURE, we have defined the model to be *generalized co-sparse factor regression*, and proposed a greedy sequential algorithm for estimation of such low-rank and sparse coefficient matrix, referred to as *generalized sequential extraction via constrained unit rank estimation* (GSeCURE). In each step of the sequential procedure, a latent factor is constructed as the linear combination of subset of predictors (called generalized latent factor) affecting only a subset of response variables via co-sparse singular vectors of the coefficient matrix. The parameter estimation in a unit sequential step proceeds via minimization of unit-rank constrained regularized negative log-likelihood function. The problem is non-convex, and the parameter estimation is challenging because of the complicated

likelihood structure. Hence, following She [2012a], the estimation proceeds via a surrogate of the objective function. We have shown that with suitable scaling of observed predictors, minimizing surrogate ensures the monotone descending property of the objective function. Moreover, our formulation can efficiently handle missing entries in the response matrix while also providing a reasonably good parameter estimates.

In rest of the chapter, generalized sequential co-sparse factor regression is proposed in Section 3.2. The optimization problem in a unit-step of the sequential procedure is solved in Section 3.3. We have discussed a consistency result of the estimator in Section 3.4. Efficacy of the proposed procedure is demonstrated via simulation studies in Section 3.5. Details of the proofs of all the relevant Theorems are provided in Appendix B.

3.2 Generalized Co-Sparse Factor Regression Model

3.2.1 Model Setup

Given n instance of independent observations, define response matrix as $\mathbf{Y} = [y_{ik}] = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times q}$, predictor matrix as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times p}$, and control variable matrix as $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times p_z}$ where first column of $\mathbf{Z}$ equals $\mathbf{1}_n$ corresponding to intercept term. To deal with the case of missing entries in $\mathbf{Y}$, define index set

$$\Omega = \{(i,k); y_{ik} \text{ is observed}, i = 1, \ldots, n, k = 1, \ldots, q\},$$

using which, we have $\widetilde{\mathbf{Y}} = \mathcal{P}_\Omega(\mathbf{Y})$ denoting the projection of $\mathbf{Y}$ onto Ω with $\tilde{y}_{ik} = y_{ik}$ for any $(i,k) \in \Omega$ otherwise $\tilde{y}_{ik} = 0$.

Consider each of the outcome variables belong to a specific exponential dispersion family [Jorgensen, 1987], e.g., Gaussian, Bernoulli or Poisson. For the ith instance of observed responses, the probability density function of an outcome y_{ik} corresponding to kth response variable is defined as

$$f(y_{ik}; \theta_{ik}, \phi) = \exp\left\{ \frac{y_{ik}\theta_{ik} - b_k(\theta_{ik})}{a_k(\phi)} + c_k(y_{ik}; \phi) \right\}, \tag{3.1}$$

where θ_{ik} is the natural parameter, ϕ_k is the dispersion parameter of the kth outcome variable, and model functions $a_k(\cdot)$, $b_k(\cdot)$, $c_k(\cdot)$ are known for a specific distribution; see Table 7 for details. Without loss of generality, we apply the canonical link function $g_k = (b_k')^{-1}$ for each outcome variable, so that $\mathbb{E}(y_{ik}) = b_k'(\theta_{ik}) = g_k^{-1}(\theta_{ik})$, where $b_k'(\cdot)$ denotes the derivative function of $b_k(\cdot)$. For the kth outcome, in terms of the observed predictor and control variables, we define an element θ_{ik} of the natural parameter matrix $\Theta \in \mathbb{R}^{n \times q}$ as

$$\theta_{ik} = o_{ik} + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\beta}_k + \mathbf{x}_i^{\mathrm{T}} \mathbf{c}_k, \qquad \forall \qquad i = 1, \ldots, n, k = 1, \ldots, q, \tag{3.2}$$

where o_{ik}s are known offset terms, $\boldsymbol{\beta}_k$s are unknown coefficient vectors corresponding

to control variables, and $\mathbf{c}_k$s are unknown coefficient vectors corresponding to the high-dimensional predictors. Now, define $\mathbf{O} = [o_{ik}] \in \mathbb{R}^{n\times q}$ as offset term matrix, $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_q] = [\tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_p]^{\mathrm{T}} \in \mathbb{R}^{p\times q}$ as predictor variable coefficient matrix, and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q] = [\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_{p_z}]^{\mathrm{T}} \in \mathbb{R}^{p_z\times q}$ as control variable coefficient matrix with intercept term given by first row $\tilde{\boldsymbol{\beta}}_1$. In terms of model variables $\{\mathbf{O}, \mathbf{C}, \boldsymbol{\beta}\}$, we define the natural parameter matrix as,

$$\Theta = \Theta(\mathbf{C}, \boldsymbol{\beta}) = \mathbf{O} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{X}\mathbf{C}, \tag{3.3}$$

where $\Theta = [\theta_{ik}] \in \mathbb{R}^{n\times q}$. Now, assume y_{ik}s are conditionally independent given θ_{ik}. The negative log-likelihood of the observed outcomes is given by

$$\mathcal{L}(\Theta, \Phi) = \mathcal{L}(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{O}) \equiv -\sum_{(i,k)\in\Omega} \ell_k(\mathbf{c}_k, \boldsymbol{\beta}_k, \phi_k; \mathbf{x}_i, \mathbf{z}_i, y_{ik}) = -\sum_{(i,k)\in\Omega} \ell_k(\theta_{ik}, \phi_k), \tag{3.4}$$

where

$$\ell_k(\theta_{ik}, \phi_k) = \frac{y_{ik}\theta_{ik} - b_k(\theta_{ik})}{a_k(\phi_k)} + c_k(\tilde{y}_{ik}, \phi_k).$$

For convenience, we define some notations. Corresponding to kth column of Θ, vector function $\mathbf{b}_k(\Theta_{\cdot k}) = [b_k(\theta_{ik}), \ldots, b_k(\theta_{nk})]^{\mathrm{T}}$, and its derivative as $\mathbf{b}_k'(\Theta_{\cdot k}) =$

Table 7: Some common distributions in the exponential dispersion family.

Distribution	Mean	Variance	θ	ϕ	$a(\phi)$	$b(\theta)$	$c(y;\phi)$
Bernoulli(p)	p	$p(1-p)$	$\log\{p(1-p)^{-1}\}$	1	1	$\log(1+e^{\theta})$	0
Poisson(λ)	λ	λ	$\log\lambda$	1	1	e^{θ}	$-\log y!$
Normal(μ, σ^2)	μ	σ^2	μ	σ^2	ϕ	$\theta^2/2$	$-(y^2\phi^{-1}+\log 2\pi)/2$

$[b_k^{'}(\theta_{ik}),\ldots,b_k^{'}(\theta_{nk})]^{\mathrm{T}}$. Then, the matrix function and its derivative are given by

$$\mathbf{B}(\Theta)=[\mathbf{b}_1(\Theta_{.1}),\ldots,\mathbf{b}_q(\Theta_{.q})],\qquad\text{and}\qquad\mathbf{B}^{'}(\Theta)=[\mathbf{b}_1^{'}(\Theta_{.1}),\ldots,\mathbf{b}_q^{'}(\Theta_{.q})].\tag{3.5}$$

Similarly, one can define any rth derivative of matrix function $\mathbf{B}(\Theta)$. Also, using the dispersion parameters of q outcome variables, we define the diagonal matrix $\Phi=\mathrm{diag}[a_1(\phi_1),\ldots,a_q(\phi_q)]$.

Using the notation defined, we re-write the negative log-likelihood $\mathcal{L}(\cdot)$ (3.4) as

$$\mathcal{L}(\mathbf{C},\boldsymbol{\beta},\Phi;\mathbf{O})=\mathcal{L}(\Theta,\Phi)=-\langle\widetilde{\mathbf{Y}},\Theta\Phi^{-1}\rangle+\langle\widetilde{\mathbf{J}},\mathbf{B}(\Theta)\Phi^{-1}\rangle,\tag{3.6}$$

where $\mathbf{J}=\mathbf{1}_{n\times q}$ and $\widetilde{\mathbf{J}}=\mathcal{P}_{\Omega}(\mathbf{J})$, and $\langle\mathbf{A},\mathbf{B}\rangle=\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{B})$ is the *trace/dot product* operator on the two involved matrices. From here onwards, for the convenience of presentation, we have used $\mathbf{Y}$ instead of $\widetilde{\mathbf{Y}}$ in model formulation and estimation.

3.2.2 Generalized Co-sparse Factor Regression

Consider the model (3.1) and its negative log-likelihood function (3.6) defined using exponential family density function. A possible approach to estimate model parameters

is by solving separate generalized linear model (GLM) problems for each response variable. When response variables are interrelated and predictor variables correlated, the approach ignores multivariate nature of the problem. Our strategy here is to induce multivariate dependency by associating responses with predictors via generalized latent factors constructed from few unknown linear combinations of predictors. We implement the strategy by imposing rank constraints on coefficient matrix $\mathbf{C}$. Thus, the optimization problem, in terms of the negative log-likelihood function $\mathcal{L}(\cdot)$ (3.6), to estimate the model parameters is given by

$$\min_{\mathbf{C},\boldsymbol{\beta},\Phi} \mathcal{L}(\Theta, \Phi) \qquad \text{s.t.} \qquad \text{rank}(\mathbf{C}) \leq r, \tag{3.7}$$

where $r = 1, \ldots, p \wedge q$. The rank constrained model defined in (3.7) is referred to as *mixed-response reduced rank regression* (mRRR).

Another approach to perform dimension reduction is through the composite structure comprises of SVD component of the coefficient matrix $\mathbf{C}$; (see Chen et al. [2012], Chen and Huang [2012b], Bunea et al. [2012], Ma and Sun [2014], Mishra et al. [2017]). Motivated by SeCURE, the SVD of the coefficient matrix $\mathbf{C}$ of rank r is given by

$$\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}, \qquad \text{s.t.} \qquad \mathbf{U}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{U}/n = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_r, \tag{3.8}$$

where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \in \mathbb{R}^{p\times r}$ is left singular vector matrix, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_r] \in \mathbb{R}^{q\times r}$

is right singular vector matrix, and $\mathbf{D} = \text{diag}\{d_1, \ldots, d_r\} \in \mathbb{R}^{r \times r}$ is diagonal matrix of singular values. The orthogonality constraints ensure the SVD of $\mathbf{C}$ to be identifiable. Using the composite structure, we obtain *generalized latent factors* $\mathbf{XU}/\sqrt{n}$ through which responses and predictors are associated.

In terms of the SVD components of the coefficient matrix (3.8), we express $\mathbf{C}$ as the sum of r constituent unit rank matrices, i.e., $\mathbf{C} = \sum_{i=i}^{r} \mathbf{C}_i$, where $\mathbf{C}_i = d_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}$. Motivated by SeCURE, we propose a sequential greedy approach to recover unit rank constituent, $\mathbf{C}_i$'s, of coefficient matrix $\mathbf{C}$, referred to as *sequential mRRR*. To begin with, for $i = 1$, an estimate of the constituent matrix $\mathbf{C}_1$ is obtained by solving the optimization problem given by,

$$(\widetilde{\mathbf{C}}_1, \widetilde{\boldsymbol{\beta}}_1, \widetilde{\Phi}_1) \equiv \underset{\mathbf{C}, \boldsymbol{\beta}, \Phi}{\arg\min}\, \mathcal{L}(\Theta, \Phi; \widetilde{\mathbf{O}}^{(1)}) \qquad \text{s.t.} \qquad \text{rank}(\mathbf{C}) \leq 1, \tag{3.9}$$

where $\widetilde{\mathbf{O}}^{(1)} = \mathbf{O}$ is offset matrix and $(\widetilde{\boldsymbol{\beta}}_1, \widetilde{\Phi}_1)$ are estimate of $(\boldsymbol{\beta}, \Phi)$ in the first unit-step of the sequential procedure. The joint estimation of the unknown parameters $(\mathbf{C}_1, \boldsymbol{\beta}, \Phi)$ are nontrivial. Thus, the algorithm involves a $\mathbf{C}$-step, $\boldsymbol{\beta}$-step and Φ-step where parameter $\mathbf{C}_1$, $\boldsymbol{\beta}$ and Φ are estimated respectively while keeping other parameters fixed. The details of the estimation steps are relegated to Section 3.3, where we have proposed the procedure for a more generalized setting and the solution of the optimization problem (3.9) corresponds to a specific case.

The subsequent unit-rank constituent matrices $\mathbf{C}_i$, for $i = 2, \ldots, r$, are obtained by

solving the optimization problem given by,

$$(\widetilde{\mathbf{C}}_i, \widetilde{\boldsymbol{\beta}}_i, \widetilde{\Phi}_i) \equiv \min_{\mathbf{C},\boldsymbol{\beta},\Phi} \mathcal{L}(\Theta, \Phi; \widetilde{\mathbf{O}}^{(i)}) \qquad \text{s.t.} \qquad \text{rank}(\mathbf{C}) \leq 1, \tag{3.10}$$

where $(\widetilde{\boldsymbol{\beta}}_i, \widetilde{\Phi}_i)$ are estimates of $(\boldsymbol{\beta}, \Phi)$ in the ith unit-step of the sequential procedure with

$$\mathbf{O}^{(i)} = \mathbf{O}^{(i-1)} + \mathbf{X}\widetilde{\mathbf{C}}_{i-1}, \qquad i = 2, \ldots, r \tag{3.11}$$

accounting for estimated signal obtained from previous sequential steps. Using the estimates of $\widetilde{\mathbf{C}}_i$, for $i = 1, \ldots, r$, we obtain $\widetilde{\mathbf{C}} = \sum_{i=1}^{r} \widetilde{\mathbf{C}}_i$. It is then trivial to obtain the specific SVD decomposition of coefficient matrix $\widetilde{\mathbf{C}}$ with orthogonality constraint $\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widetilde{\mathbf{U}}/n = \mathbf{1}$ and $\widetilde{\mathbf{V}}^{\mathrm{T}}\widetilde{\mathbf{V}} = \mathbf{1}$ where

$$\widetilde{\mathbf{U}} = [\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_r], \;\; \widetilde{\mathbf{V}} = [\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_r], \;\; \widetilde{\mathbf{D}} = \text{diag}[\tilde{d}_1, \ldots, \tilde{d}_r], \tag{3.12}$$

with $\widetilde{\mathbf{C}}_i = \tilde{d}_i \tilde{\mathbf{u}}_i \tilde{\mathbf{v}}_i^{\mathrm{T}}$. The parameters $(\boldsymbol{\beta}, \Phi)$ are estimated as $(\widetilde{\boldsymbol{\beta}}_r, \widetilde{\Phi}_r)$, the values obtained in the last sequential step, i.e., for $i = r$. Assuming solutions of (3.9) and (3.10) are known, we present the algorithm for *sequential mRRR*; see Algorithm 3 for details.

In addition to dimension reduction through a low-rank assumption on the coefficient matrix $\mathbf{C}$, we desire to perform variable selection as well. In terms of the composite

Algorithm 3 Sequential mRRR

Initialize: $\boldsymbol{\beta}^{(0)}$, $\Phi^{(0)}$, and set a desired rank $r \geq 1$.
for $k \leftarrow 1$ to r **do**
  Compute the current offset matrix $\mathbf{O}^{(k)}$ as in (3.11).
  Initialize $\mathbf{C}_k^{(0)}$, and set $\boldsymbol{\beta}^{(0)} = \widetilde{\boldsymbol{\beta}}_{k-1}$, $\Phi^{(0)} = \widetilde{\Phi}_{k-1}$, estimates from previous step.
  repeat
    (1) **C**-step: $\mathbf{C}_k^{(t+1)} = \mathbb{T}^{(1)}(\mathbf{C}_k^{(t)} + \mathbf{X}^{\mathrm{T}}\{\mathbf{Y} - \boldsymbol{\mu}(\mathbf{O}_k, \mathbf{C}_k^{(t)}, \boldsymbol{\beta}^{(t)})\}\Phi^{(t)-1})$ where $\mathbb{T}^{(1)}(\mathbf{M})$ extract 1st SVD component of matrix $\mathbf{M}$. (see (3.24))
    (2) $\boldsymbol{\beta}$-step: $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \mathbf{Z}^{\mathrm{T}}\{\mathbf{Y} - \boldsymbol{\mu}(\mathbf{O}_k, \mathbf{C}_k^{(t+1)}, \boldsymbol{\beta}^{(t)})\}\Phi^{(t)-1}$,
    (3) Φ-step: $\Phi^{(t+1)} = \arg\min_{\Phi} \sum_{i,k} \mathcal{L}(\mathbf{C}_k^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \Phi)$,
    $t \leftarrow t + 1$.
  until convergence,
    e.g., $\|[\mathbf{C}^{(t+)}\ \boldsymbol{\beta}^{(t+1)}] - [\mathbf{C}^{(t)}\ \boldsymbol{\beta}^{(t)}]\|_F / \|[\mathbf{C}^{(t)}\ \boldsymbol{\beta}^{(t)}]\|_F \leq \epsilon$ with $\epsilon = 10^{-6}$.
  return $\widetilde{\mathbf{C}}_k$, $\widetilde{\boldsymbol{\beta}}_k$, $\widetilde{\Phi}_k$.
end for
return $\widetilde{\mathbf{C}} = \sum_{i=1}^{r} \widetilde{\mathbf{C}}_i$, $\widetilde{\boldsymbol{\beta}}_r$, $\widetilde{\Phi}_r$.

structure defined in (3.8), variable selection can be achieved by having co-sparse singular vectors. We refer it as *generalized co-sparse factor regression* model (gSFAR). The desired low-rank and sparse structure facilitates model interpretation and improves model accuracy. A possible optimization problem to attain both low-rank and sparse coefficient matrix is given by

$$\min_{\mathbf{C},\boldsymbol{\beta},\Phi} \quad \mathcal{L}(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{O}) + \lambda_1 \rho_1(\mathbf{U}) + \lambda_2 \rho_2(\mathbf{V}), \ \text{ s.t. } \mathbf{U}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{U}/n = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_r \tag{3.13}$$

where $\rho_1(\cdot)$ and $\rho_2(\cdot)$ are sparsity-inducing penalty functions with λ_1 and λ_2 as their corresponding tuning parameters. Unfortunately, there are several difficulties with this joint estimation approach. The rank r has to be specified in advance. The orthogonality

constraint solves identifiability issue, thus cannot be compromised. The simultaneous presence of the high-dimensional low-rank structure, orthogonality constraint and presence of two tunning parameters in the sparsity regularization makes the estimation challenging. Thus, we need to look into other avenues for efficient recovery of low-rank and sparse structure in gSFAR.

We overcome this challenge again by using the composite structure (3.8) of the SVD components of the coefficient matrix $\mathbf{C}$. Motivated by SeCURE, we propose a greedy sequential approach that proceeds via *generalized constrained unit rank estimation* (G-CURE) step to obtain sparse unit-rank constituent matrices in order of importance. In a unit step, a latent factor is constructed as linear combination of a subset of predictors affecting only a subset of responses because of "co-sparse" left and right singular vectors. We call the algorithm as *generalized sequential extraction via constrained unit rank estimation* (GSeCURE) and the latent factor as *generalize latent factor*.

To begin with, for $i = 1$, estimates of the constituent matrix $\mathbf{C}_1$ and unknown parameters $(\boldsymbol{\beta}, \Phi)$ are obtained by solving the optimization problem given by

$$(\widehat{\mathbf{C}}_1, \widehat{\boldsymbol{\beta}}_1, \widehat{\Phi}_1) \equiv \underset{\mathbf{C},\boldsymbol{\beta},\Phi}{\arg\min} \left\{ \mathcal{L}(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{O}^{(1)}) + \rho(\mathbf{C}; \lambda) \right\}, \quad \text{s.t.} \quad \text{rank}(\mathbf{C}) \leq 1, \tag{3.14}$$

where constraint $\rho(\mathbf{C})$ is a sparsity inducing penalty function, $\mathbf{O}^{(1)} = \mathbf{0}$ is an offset matrix. Here, $(\widehat{\boldsymbol{\beta}}_1, \widehat{\Phi}_1)$ are estimates of $(\boldsymbol{\beta}, \Phi)$ from first step of the proposed sequential procedure. In the optimization problem (3.14), unit rank constraint implies that $\mathbf{C} =$

$d\mathbf{u}\mathbf{v}^{\mathrm{T}}$ with $d \geq 0$, $\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u}/n = 1$ and $\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1$ where $d \in \mathbb{R}$, $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{v} \in \mathbb{R}^q$. Similar to SVD, this decomposition is unique up to the signs of the vectors as long as $\mathbf{C}$ is not a zero matrix.

We use the adaptive elastic net penalty [Zou and Hastie, 2005, Zou and Zhang, 2009],

$$\begin{aligned}\rho(\mathbf{C};\lambda) = \rho(\mathbf{C};\mathbf{W}_1,\lambda,\alpha) &= \alpha\lambda\|\mathbf{W}_1 \circ \mathbf{C}\|_1 + (1-\alpha)\lambda\|\mathbf{C}\|_F^2 \\ &= \alpha\lambda\sum_{i=1}^{p}\sum_{j=1}^{q} w_{ij1}|c_{ij}| + (1-\alpha)\lambda\sum_{i=1}^{p}\sum_{j=1}^{q} c_{ij}^2. \end{aligned} \tag{3.15}$$

Here $\|\cdot\|_1$ denotes the ℓ_1 norm, the operator "$\circ$" stands for the Hadamard product, $\mathbf{W}_1 = [w_{ij1}]_{p\times q}$ is a pre-specified weighting matrix, λ is a tuning parameter controlling the overall amount of regularization, and $\alpha \in (0,1)$ controls the relative weights between the two penalty terms. We set $\mathbf{W}_1 = |\widetilde{\mathbf{C}}_1|^{-\gamma}$ such that $w_{ij1} = w_1^{(d)} w_{i1}^{(u)} w_{j1}^{(v)}$, with

$$w_1^{(d)} = |\widetilde{d}_1|^{-\gamma}, \mathbf{w}_1^{(u)} = [w_{11}^{(u)},...,w_{p1}^{(u)}]^{\mathrm{T}} = |\widetilde{\mathbf{u}}_1|^{-\gamma}, \mathbf{w}_1^{(v)} = [w_{11}^{(v)},...,w_{q1}^{(v)}]^{\mathrm{T}} = |\widetilde{\mathbf{v}}_1|^{-\gamma}, \tag{3.16}$$

where $\widetilde{\mathbf{C}}_1 = \widetilde{d}_1\widetilde{\mathbf{u}}_1\widetilde{\mathbf{v}}_1^{\mathrm{T}}$ is estimate of the first constituent matrix obtained for sequential mRRR using Algorithm 3, and γ is a non-negative constant with $|\cdot|^{-\gamma}$ componentwisely defined. As suggested by Zou [2006], we set $\gamma = 2$. Since we mainly focus on sparse estimation, we fix α as a constant, i.e., $\alpha = 0.95$. Comparing to lasso, one advantage of elastic net is that the additional ridge penalty improves the convexity of the problem and can enhance the stability of optimization. For simplicity, we may write $\rho(\mathbf{C};\mathbf{W},\lambda,\alpha) =$

$\rho(\mathbf{C}; \mathbf{W}, \lambda)$.

To focus on the sequential extraction procedure, we defer the details of computation obtaining $(\widehat{\mathbf{C}}_1, \widehat{\boldsymbol{\beta}}_1, \widehat{\Phi}_1)$ to Section 3.3. Assuming the solution path of (3.14) can be fitted, the tuning parameter λ can be chosen based on either cross validation or some information criterion. The unit-rank matrix $\mathbf{C}$ is estimated as $\widehat{\mathbf{C}}_1$, or equivalently, $(\widehat{d}_1, \widehat{\mathbf{u}}_1, \widehat{\mathbf{v}}_1)$.

To estimate subsequent unit-rank constituent $\mathbf{C}_k$ of coefficient matrix $\mathbf{C}$, for $k = 2, \ldots, r$, we solve the optimization problem,

$$(\widehat{\mathbf{C}}_k, \widehat{\boldsymbol{\beta}}_k, \widehat{\Phi}_k) \equiv \underset{\mathbf{C}, \boldsymbol{\beta}, \Phi}{\arg\min} \left\{ \mathcal{L}(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{O}^{(k)}) + \rho(\mathbf{C}; \mathbf{W}_k) \right\}, \quad \text{s.t.} \quad \text{rank}(\mathbf{C}) \leq 1, \tag{3.17}$$

where $(\widehat{\boldsymbol{\beta}}_k, \widehat{\Phi}_k)$ are estimates of $(\boldsymbol{\beta}, \Phi)$ from kth step of the proposed sequential procedure with

$$\mathbf{O}^{(k)} = \mathbf{O}^{(k-1)} + \mathbf{X}\widehat{\mathbf{C}}_{k-1} \tag{3.18}$$

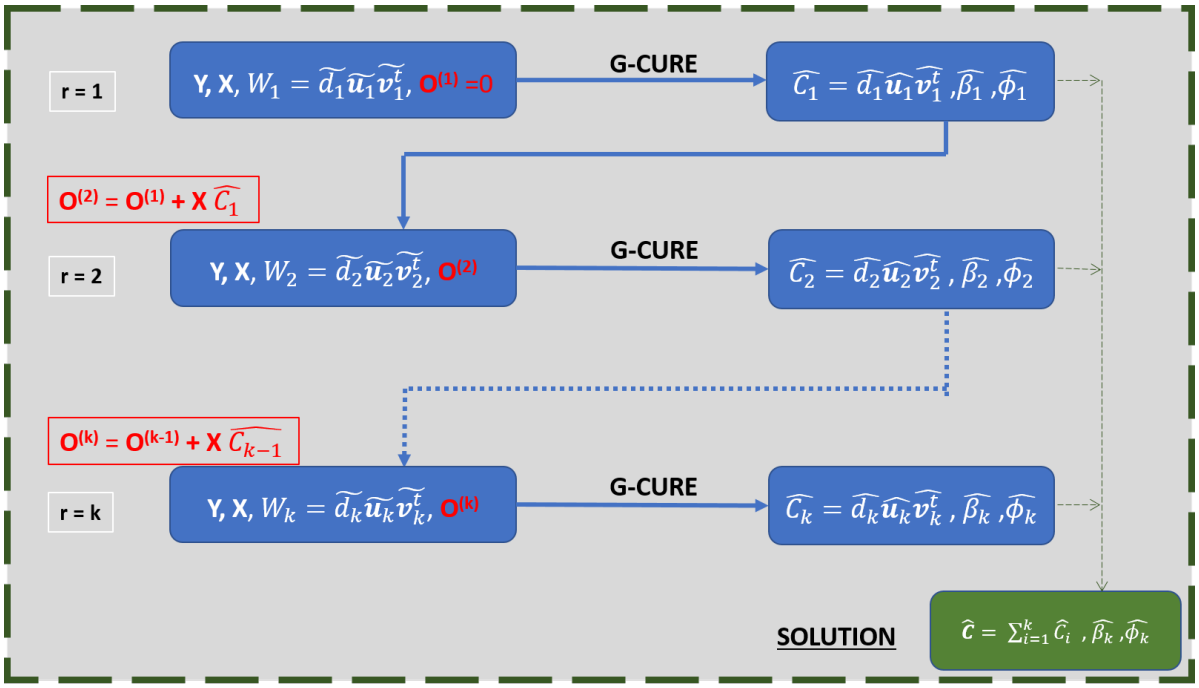
accounting for signal estimated from previous steps. The penalty term $\rho(\mathbf{C}; \mathbf{W}_k)$ is defined in term of the weight matrix $\mathbf{W}_k$. The construction of $\mathbf{W}_k$ follows (3.16) using estimate $\widetilde{\mathbf{C}}_k = \tilde{d}_k \tilde{\mathbf{u}}_k \tilde{\mathbf{v}}_k^{\mathrm{T}}$ obtained as solution of (3.10) in sequential mRRR using Algorithm 3.

We sequentially perform regularized estimation using (3.17) to obtain $\widehat{\mathbf{C}}_k$ or equivalently $(\widehat{d}_k, \widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k)$, for $k = 1, \ldots, r$, $r \leq \min(r_x, q)$ where $r_x = \text{rank}(\mathbf{X})$, and $(\widehat{\boldsymbol{\beta}}_k, \widehat{\Phi}_k)$. We refer to the generic problem in (3.17) together with the tuning process as *generalized constrained unit-rank estimation* (G-CURE), i.e., G-CURE$(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{Y}, \mathbf{X}, \mathbf{W}, \mathbf{O})$, where $\mathbf{W} = w^{(d)}\mathbf{w}^{(u)}\mathbf{w}^{(v)\text{T}}$ refers to generic weight matrix obtained after dropping subscript 1 from its definition given in (3.16), $\mathbf{C} = d\mathbf{u}\mathbf{v}^{\text{T}}$ is unit-rank matrix and $\mathbf{O}$ to generic offset term matrix. The solution of generic problem G-CURE is relegated to Section 3.3. Assuming its solution known, Algorithm 4 and Figure 9 summarizes the proposed computation procedure, i.e., *Generalized Sequential Extraction via Constrained Unit-rank Estimation* (GSeCURE).

Algorithm 4 Generalized Sequential Co-sparse Factor Regression (GSeCURE)

Initialize: $\boldsymbol{\beta}^{(0)}$, $\Phi^{(0)}$, and set a desired rank $r \geq 1$.
for $k \leftarrow 1$ to r **do**
  Compute the current offset matrix $\mathbf{O}^{(k)}$ as in (3.18).
  Initialize $\mathbf{C}_k^{(0)}$, and set $\boldsymbol{\beta}^{(0)} = \widehat{\boldsymbol{\beta}}_{k-1}$, $\Phi^{(0)} = \widehat{\Phi}_{k-1}$, estimates from previous step.
  repeat
    Perform the G-CURE$(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{Y}, \mathbf{X}, \mathbf{W}_k, \mathbf{O}^{(k)})$ analysis via (3.17) (including the tuning process), and obtain $\widehat{\mathbf{C}}_k$ or equivalently $(\widehat{d}_k, \widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k)$, $\widehat{\boldsymbol{\beta}}_k$ and $\widehat{\Phi}_k$.
  until convergence
  if $\widehat{d}_k = 0$ **then**
    Set $\widehat{d}_h = 0$ for any $k \leq h \leq r$;
  end if
end for
return $\widehat{\mathbf{C}} = \sum_{i=1}^{r} \widehat{\mathbf{C}}_i$, $\widehat{\boldsymbol{\beta}}_r$, $\widehat{\Phi}_r$.

Figure 9: GSeCURE: Generalized Sequential Factor Extraction via Co-Sparse Unit-Rank Estimation



3.3 Computation

3.3.1 Generalized Constrained Unit-rank Estimation (G-CURE)

Here we provide details of the procedure required to solve the optimization problem for a unit step of GSeCURE, i.e., G-CURE(·). For the ease of notation, we are dropping subscript k from G-CURE(·) defined in (3.17) and write the generic problem as

$$(\widehat{\mathbf{C}}, \widehat{\boldsymbol{\beta}}, \widehat{\Phi}) \equiv \underset{\mathbf{C},\boldsymbol{\beta},\Phi}{\arg\min} \left\{ F(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{O}) \equiv \mathcal{L}(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{O}) + \rho(\mathbf{C}; \mathbf{W}) \right\}, \quad \text{s.t.} \quad \text{rank}(\mathbf{C}) \leq 1, \tag{3.19}$$

where $\mathbf{O}$ and $\mathbf{W}$ are corresponding offset and weight matrix respectively. The generic weight matrix $\mathbf{W}$ is constructed from $(\widetilde{d}, \widetilde{\mathbf{u}}, \widetilde{\mathbf{v}})$ using its definition (3.16). Let unit rank matrix $\mathbf{C} = d\mathbf{u}\mathbf{v}^{\mathrm{T}}$ where $d \in \mathbb{R}$, $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{v} \in \mathbb{R}^q$. The constraint in equation (3.8) leads to $\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} = n, \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1$. In terms of defined parameters, we reformulate the G-CURE problem (3.19) as

$$(\widehat{\mathbf{C}}, \widehat{\boldsymbol{\beta}}, \widehat{\Phi}) \equiv \underset{\mathbf{C},\boldsymbol{\beta},\Phi}{\arg\min} \left\{ F(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{O}) \equiv \mathcal{L}(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{O}) + \rho(\mathbf{C}; \mathbf{W}, \lambda) \right\}, \tag{3.20}$$

$$\text{s.t. } \mathbf{C} = d\mathbf{u}\mathbf{v}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} = n, \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1.$$

The optimization problem (3.20) is non-convex, and joint estimation is intractable. Thus, we propose an iterative stepwise strategy where estimation proceeds via a $\mathbf{C}$-step, $\boldsymbol{\beta}$-step and Φ-step to update unknown parameters $\mathbf{C}, \boldsymbol{\beta}$ and Φ respectively.

We begin with analysis of the problem in the $\mathbf{C}$-step for updating unknown parameter $\mathbf{C}$ with others held fixed. Given the complicated structure of negative likelihood function $\mathcal{L}(\cdot)$ (3.6), simplification of the corresponding objective function is difficult. Following She [2012b], we consider a surrogate of the objective function $F(\mathbf{C}, \boldsymbol{\beta}, \Phi, \mathbf{O})$ (3.20). In terms of unit-rank matrix variable $\mathbf{A} \in \mathbb{R}^{p\times q}$, the function is given by

$$\begin{aligned} G(\mathbf{A}; \mathbf{C}, \boldsymbol{\beta}, \Phi, \mathbf{O}) =& \mathcal{L}(\mathbf{A}, \boldsymbol{\beta}, \Phi; \mathbf{O}) + \frac{1}{2}\|\mathbf{A} - \mathbf{C}\|_F^2 + \langle \mathbf{B}^{'}(\Theta), \mathbf{X}(\mathbf{A} - \mathbf{C})\Phi^{-1}\rangle \\ & - \langle \mathbf{J}, [\mathbf{B}(\Theta(\mathbf{A}, \boldsymbol{\beta})) - \mathbf{B}(\Theta)]\Phi^{-1}\rangle + \rho(\mathbf{A}; \mathbf{W}, \lambda), \end{aligned} \tag{3.21}$$

where $\Theta(\mathbf{A}, \boldsymbol{\beta})$ is defined for parameter $(\mathbf{A}, \boldsymbol{\beta})$ according to definition (3.3). It can be verified that surrogate function $G(\mathbf{C}; \mathbf{C}, \boldsymbol{\beta}, \Phi, \mathbf{O}) = F(\mathbf{C}, \boldsymbol{\beta}, \Phi, \mathbf{O})$. After some algebra, G can be simplified as

$$\begin{aligned} G(\mathbf{A}; \mathbf{C}, \boldsymbol{\beta}, \Phi) =& \langle \mathbf{B}^{'}(\Theta) - \mathbf{Y}, \mathbf{X}\mathbf{A}\Phi^{-1}\rangle + \frac{1}{2}\|\mathbf{A} - \mathbf{C}\|_F^2 + \rho(\mathbf{A}; \mathbf{W}, \lambda) + \text{const}, \\ =& \frac{1}{2}\|\mathbf{A} - \mathbf{C} - \mathbf{X}^{\mathrm{T}}\{\mathbf{Y} - \boldsymbol{\mu}(\mathbf{C}, \boldsymbol{\beta})\}\Phi^{-1}\|_F^2 + \rho(\mathbf{A}; \mathbf{W}, \lambda) + \text{const}, \end{aligned} \tag{3.22}$$

where "const" represents any reminder constant term that does not depend on $\mathbf{A}$. Here we have used $\boldsymbol{\mu}(\mathbf{C}, \boldsymbol{\beta}) = \mathbf{B}^{'}(\Theta)$. One core setup in our algorithm is based on minimizing

$G(\mathbf{A}; \mathbf{C}, \boldsymbol{\beta}, \Phi)$ with respect to $\mathbf{A}$, and with proper scaling of the observed predictor and control variables, it ensures the monotone descending property of the objective function (see the proof of Theorem 3.1). The sparsity inducing penalty $\rho(\mathbf{A}; \mathbf{W}, \lambda)$ is defined according to equation (3.15). Now, we focus on minimizing the objective function $G(\mathbf{A}; \mathbf{C}, \boldsymbol{\beta}, \Phi, \mathbf{O})$ under unit-rank constraint of matrix $\mathbf{A}$.

First, let us consider a case of $\lambda = 0$ through which we intend to solve problem in (3.9) and (3.10). On dropping the subscript i from optimization problem (3.10), the generic problem is given by

$$(\widetilde{\mathbf{C}}, \widetilde{\boldsymbol{\beta}}, \widetilde{\Phi}) \equiv \underset{\mathbf{C}, \boldsymbol{\beta}, \Phi}{\arg\min}\, \mathcal{L}(\Theta, \Phi; \mathbf{O}) \qquad \text{s.t.} \qquad \text{rank}(\mathbf{C}) \leq 1. \tag{3.23}$$

Unknown parameters $(\widetilde{\mathbf{C}}, \widetilde{\boldsymbol{\beta}}, \widetilde{\Phi})$ are updated in a $\mathbf{C}$-step, $\boldsymbol{\beta}$-step and Φ-step respectively. The procedure for updating in $\mathbf{C}$-step in (3.23) differs from that of problem in (3.20). The remaining steps, i.e., $\boldsymbol{\beta}$-step and Φ-step are similar for the two problems. From (3.22), the surrogate function for updating $\widetilde{\mathbf{C}}$ in the $\mathbf{C}$-step is given by

$$\frac{1}{2}\|\mathbf{A} - \widetilde{\mathbf{C}} - \mathbf{X}^{\mathrm{T}}\{\mathbf{Y} - \boldsymbol{\mu}(\widetilde{\mathbf{C}}, \boldsymbol{\beta})\}\Phi^{-1}\|_F^2,$$

and the estimate of unit rank matrix $\mathbf{A}$ minimizing it is then given by

$$\widehat{\mathbf{A}} = \mathbb{T}^{(1)}(\widetilde{\mathbf{C}} + \mathbf{X}^{\mathrm{T}}\{\mathbf{Y} - \boldsymbol{\mu}(\widetilde{\mathbf{C}}, \boldsymbol{\beta})\}\Phi^{-1}), \tag{3.24}$$

where operator $\mathbb{T}^{(1)}(\mathbf{M})$ refers to the first SVD component of matrix $\mathbf{M}$. This results in updating $\widetilde{\mathbf{C}}$ with $\widehat{\mathbf{A}}$ in the $\mathbf{C}$-step.

Now, for $\lambda \neq 0$, we present an approach to minimize the surrogate function $G(\cdot)$ with respect to unit-rank matrix $\mathbf{A}$. The problem is non-convex and a closed form solution is not possible. To overcome, we decompose unit-rank matrix $\mathbf{A}$ as $\mathbf{u}_a\mathbf{d}_a\mathbf{v}_a^{\mathrm{T}}$ where the vector $\mathbf{u}_a \in \mathbb{R}^p$, $\mathbf{v}_a \in \mathbb{R}^q$ and $d_a \in \mathbb{R}$. Corresponding to the constraint in problem (3.20), we set $\|\mathbf{v}_a\|_2 = 1$ and $\|\mathbf{X}\mathbf{u}_a/\sqrt{n}\|_2 = 1$. Motivated by Chen et al. [2012], the unit-rank matrix $\mathbf{A}$ is estimated in terms of block variables $(d_a, \mathbf{u}_a)$ and $(d_a, \mathbf{v}_a)$. In the first step, consider $\check{\mathbf{v}} = d_a\mathbf{v}_a$. On substituting $\mathbf{A} = \mathbf{u}_a\check{\mathbf{v}}^{\mathrm{T}}$ in (3.22), for fixed $\mathbf{u}_a$, the optimization problem to estimate unknown parameter $\check{\mathbf{v}}$ is given by

$$\min_{\check{\mathbf{v}}} \left\{ G^{(v)}(\check{\mathbf{v}}; \mathbf{u}_a, \mathbf{C}, \boldsymbol{\beta}, \Phi) = \frac{1}{2}\|\mathbf{u}_a\check{\mathbf{v}}^{\mathrm{T}} - \mathbf{C}^{(\mathbf{v})}\|_F^2 + \rho(\mathbf{u}_a\check{\mathbf{v}}^{\mathrm{T}}; \mathbf{W}, \lambda) \right\}, \tag{3.25}$$

where $\mathbf{C}^{(\mathbf{v})} = \mathbf{C} + \mathbf{X}^{\mathrm{T}}\{\mathbf{Y} - \boldsymbol{\mu}(\mathbf{C}, \boldsymbol{\beta})\}\Phi^{-1}$. The closed form solution of the unknown parameter $\check{\mathbf{v}}$ equals $\mathbf{S}(\mathbf{C}^{(\mathbf{v})\mathrm{T}}\mathbf{u}_a, \alpha\lambda\mathbf{u}_a^{\mathrm{T}}\mathbf{w}^{(u)}w^{(d)}\mathbf{w}^{(v)})/\{1 + (1-\alpha)\lambda\}\|\mathbf{u}_a\|_2^2$ (called V-step), where $\mathbf{S}(\cdot, \lambda)$ is soft thresholding operator on $\cdot$. Using estimate of $\check{\mathbf{v}}$ and constraint $\|\mathbf{v}_a\|_2 = 1$, we extract and then updated the block variable $(d_a, \mathbf{v}_a)$ as $(\tilde{d}_a,\ \hat{\mathbf{v}}_a)$. Using $(d_a, \mathbf{v}_a)$, unit-rank matrix $\mathbf{C}$ is updated as $\overline{\mathbf{C}} = \mathbf{u}_a\check{\mathbf{v}}^{\mathrm{T}}$.

In the second step, with $\mathbf{v}_a = \hat{\mathbf{v}}_a$ fixed, we perform estimation in terms of block variable $(d_a, \mathbf{u}_a)$, for which we consider $\check{\mathbf{u}} = d_a\mathbf{u}_a$. On substituting $\mathbf{A} = \check{\mathbf{u}}\mathbf{v}_a^{\mathrm{T}}$ in (3.22), we obtain the surrogate objective function for estimating $\check{\mathbf{u}}$; for convenience refer it as

$G^{(u)}(\check{\mathbf{u}}; \mathbf{v}_a, \mathbf{C}, \boldsymbol{\beta}, \Phi)$. The optimization problem to estimate the unknown parameter $\check{\mathbf{u}}$ (called U-step) is given by,

$$\min_{\check{\mathbf{u}}} \left\{ G^{(u)}(\check{\mathbf{u}}; \mathbf{v}_a, \mathbf{C}, \boldsymbol{\beta}, \Phi) = \frac{1}{2}\|\check{\mathbf{u}}\mathbf{v}_a^{\mathrm{T}} - \mathbf{C}^{(\mathbf{u})}\|_F^2 + \rho(\check{\mathbf{u}}\mathbf{v}_a^{\mathrm{T}}; \mathbf{W}, \lambda) \right\}, \tag{3.26}$$

where $\mathbf{C}^{(\mathbf{u})} = \overline{\mathbf{C}} + \mathbf{X}^{\mathrm{T}}\{\mathbf{Y} - \boldsymbol{\mu}(\overline{\mathbf{C}}, \boldsymbol{\beta})\}\Phi^{-1}$. Like the V-step, the solution for unknown parameter $\check{\mathbf{u}}$ equals $\mathbf{S}(\mathbf{C}^{(\mathbf{u})}\mathbf{v}_a, \alpha\lambda\mathbf{v}_a^{\mathrm{T}}\mathbf{w}^{(v)}w^{(d)}\mathbf{w}^{(u)})/\{1 + (1-\alpha)\lambda\}\|\mathbf{v}_a\|_2^2$ (called U-step). We extract and then update the block variables $(d_a, \mathbf{u}_a)$ as $(\hat{d}_a, \hat{\mathbf{u}}_a)$ using constraint $\|\mathbf{X}\mathbf{u}_a/\sqrt{n}\|_2 = 1$.

Thus, the $\mathbf{C}$-step involves performing a U-step and V-step to estimate the constituents of unit-rank matrix $\mathbf{A}$. At the end tth iteration, let unit-rank matrix $\mathbf{C} = \mathbf{C}^{(t)} = \mathbf{u}^{(t)}d^{(t)}\mathbf{v}^{(t)\mathrm{T}}$, control variable coefficient matrix $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$ and dispersion parameter $\Phi = \Phi^{(t)}$. In V-step of $(t+1)$th iteration, we set $\mathbf{u}_a = \mathbf{u}^{(t)}$ and estimate block variable $(d_a, \mathbf{v}_a)$ as $(\tilde{d}_a,\ \hat{\mathbf{v}}_a)$. Now, assign $(\tilde{d}_a,\ \hat{\mathbf{v}}_a)$ to $(\tilde{d}^{(t+1)}, \mathbf{v}^{(t+1)})$, using which define $\widetilde{\mathbf{C}}^{(t+1)} = \tilde{d}^{(t+1)}\mathbf{u}^{(t)}\mathbf{v}^{(t+1)\mathrm{T}}$. Similarly, in U-step of $(t+1)$th iteration, we set $\mathbf{v}_a = \hat{\mathbf{v}}_a = \mathbf{v}^{(t+1)}$ and estimate block variable $(d_a, \mathbf{u}_a)$ as $(\hat{d}_a,\ \hat{\mathbf{u}}_a)$. Also, assign $(\hat{d}_a,\ \hat{\mathbf{u}}_a)$ to $(d^{(t+1)}, \mathbf{u}^{(t+1)})$ and obtain $\mathbf{C}^{(t+1)} = d^{(t+1)}\mathbf{u}^{(t+1)}\mathbf{v}^{(t+1)\mathrm{T}}$.

In the $\boldsymbol{\beta}$-step, when $\mathbf{C}$ and Φ are held fixed, it can be seen that solving (3.20) with respect to $\boldsymbol{\beta}$ reduces to a set of univariate GLM problems. In the presence of non-Gaussian outcomes, the corresponding GLM problems can be solved using an iterative algorithm which could be very time consuming. Alternatively, we take a majorization

approach to get an one-step update of $\boldsymbol{\beta}$, similar to the previous method of updating $\mathbf{C}$. Define surrogate function

$$\begin{aligned} H(\boldsymbol{\alpha}; \mathbf{C}, \boldsymbol{\beta}, \Phi) =& \mathcal{L}(\mathbf{C}, \boldsymbol{\alpha}, \Phi; \mathbf{O}) + \frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_F^2 + \langle \mathbf{B}^{'}(\Theta), \mathbf{X}(\boldsymbol{\alpha} - \boldsymbol{\beta})\Phi^{-1} \rangle \\ & - \langle \mathbf{J}, [\mathbf{B}(\Theta(\mathbf{C}, \boldsymbol{\alpha})) - \mathbf{B}(\Theta)]\Phi^{-1} \rangle + \text{const.} \end{aligned} \tag{3.27}$$

Then minimizing H with respect to $\boldsymbol{\alpha}$ is the same as

$$\min_{\boldsymbol{\alpha}} \ \|\boldsymbol{\alpha} - \boldsymbol{\beta} - \mathbf{Z}^{\mathrm{T}}\{\mathbf{Y} - \boldsymbol{\mu}(\mathbf{C}, \boldsymbol{\beta})\}\Phi^{-1}\|_F^2,$$

which is a simple least squares problem. So when $\mathbf{C} = \mathbf{C}^{(t+1)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$, and $\Phi = \Phi^{(t)}$, the minimizer

$$\widehat{\boldsymbol{\alpha}} = \boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \mathbf{Z}^{\mathrm{T}}\{\mathbf{Y} - \boldsymbol{\mu}(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)})\}\Phi^{(t)-1}. \tag{3.28}$$

Once $\mathbf{C}$ and $\boldsymbol{\beta}$ are updated, we can then update Φ by maximizing the log-likelihood function. With solutions for updating parameters $(\mathbf{C}, \boldsymbol{\beta}, \Phi)$ through a $\mathbf{C}$-step, $\boldsymbol{\beta}$-step and Φ-step respectively, we present the G-CURE in Algorithm 5.

The analysis in G-CURE proceeds via surrogate functions. Hence, it is important to establish the fact that steps in G-CURE Algorithm 5 lead to the monotone descending property of original objective function $F(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{O})$ in (3.20). In the tth iteration, let the current parameter estimate be given by $(\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)})$. The coefficient matrix is

Algorithm 5 Generalized Constrained Unit-Rank Estimation: G-CURE

Initialize $\mathbf{C}$, $\boldsymbol{\beta}^{(0)}$ and $\Phi^{(0)}$. Set $t \leftarrow 0$.
repeat
(1) **C**-step:

V-step (3.25): block variables $(\tilde{d}^{(t+1)}, \mathbf{v}^{(t+1)})$ are recovered from $\check{\mathbf{v}}$ using constraint defined in (3.20); define $\widetilde{\mathbf{C}}^{(t+1)} = \tilde{d}^{(t+1)}\mathbf{u}^{(t)}\mathbf{v}^{(t+1)\mathrm{T}}$.

U-step (3.26): block variables $(d^{(t+1)}, \mathbf{u}^{(t+1)})$ are recovered from $\check{\mathbf{u}}$ using constraint defined in (3.20); define $\mathbf{C}^{(t+1)} = d^{(t+1)}\mathbf{u}^{(t+1)}\mathbf{v}^{(t+1)\mathrm{T}}$.

(2) $\boldsymbol{\beta}$-step: $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \mathbf{Z}^{\mathrm{T}}\{\mathbf{Y} - \boldsymbol{\mu}(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)})\}\Phi^{(t)-1}$,
(3) Φ-step: $\Phi^{(t+1)} = \arg\max_{\Phi} \mathcal{L}(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \Phi)$

$t \leftarrow t + 1$.
until convergence,
e.g., $\|[\mathbf{C}^{(t+)}\ \boldsymbol{\beta}^{(t+1)}] - [\mathbf{C}^{(t)}\ \boldsymbol{\beta}^{(t)}]\|_F / \|[\mathbf{C}^{(t)}\ \boldsymbol{\beta}^{(t)}]\|_F \leq \epsilon$ with $\epsilon = 10^{-6}$.
return $\widehat{\mathbf{C}}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\Phi}$.

updated to $\widetilde{\mathbf{C}}^{(t+1)}$ and $\mathbf{C}^{(t+1)}$ after the V-step and U-step respectively in the $(t+1)$ iteration. The $\boldsymbol{\beta}$-step then results in updated parameter $\boldsymbol{\beta}^{(t+1)}$. Before we state a theorem ensuring the monotone descending property of the G-CURE Algorithm 5, we define some notations. For $\widetilde{\xi}_c^{(t+1)} = \{a\mathbf{C}^{(t)} + (1-a)\widetilde{\mathbf{C}}^{(t+1)}; 0 < a < 1\}$,

$$\gamma_1 = \max_{1 \leq k \leq q} \sup_{a \in (0,1)} \|\mathbf{I}(\Theta_{\cdot k}(\widetilde{\xi}_c^{(t+1)}, \boldsymbol{\beta}^{(t)}), \phi_k^{(t)})\|_2, \tag{3.29}$$

with

$$\mathbf{I}(\Theta_{\cdot k}(\widetilde{\xi}_c^{(t+1)}, \boldsymbol{\beta}^{(t)})) = \mathbf{X}^{\mathrm{T}}\zeta(\Theta_{\cdot k}(\widetilde{\xi}_c^{(t+1)}, \boldsymbol{\beta}^{(t)}), \phi_k^{(t)})\mathbf{X}, \tag{3.30}$$

and

$$\zeta(\Theta_{.k}(\widetilde{\xi}_c^{(t+1)}, \boldsymbol{\beta}^{(t)}), \phi_k^{(t)}) = \text{diag}[\mathbf{B}''_{.k}(\Theta_{.k}(\widetilde{\xi}_c^{(t+1)}, \boldsymbol{\beta}^{(t)}))]/\phi_k^{(t)}. \tag{3.31}$$

Similarly, using definitions in (3.30) and (3.31), for $\xi_c^{(t+1)} = \{a\widetilde{\mathbf{C}}^{(t+1)} + (1-a)\mathbf{C}^{(t+1)}; 0 < a < 1\}$,

$$\gamma_2 = \max_{1\le k\le q} \sup_{a\in(0,1)} \|\mathbf{I}(\Theta_{.k}(\xi_c^{(t+1)}, \boldsymbol{\beta}^{(t)}), \phi_k^{(t)})\|_2, \tag{3.32}$$

and for $\xi_\beta^{(t+1)} \in \{a\boldsymbol{\beta}^{(t)} + (1-a)\boldsymbol{\beta}^{(t+1)}; 0 < a < 1\}$

$$\gamma_3 = \max_{1\le k\le q} \sup_{a\in(0,1)} \|\mathbf{I}(\Theta_{.k}(\mathbf{C}^{(t+1)}, \xi_\beta^{(t+1)}), \phi_k^{(t)})\|_2, \tag{3.33}$$

such that $\mathbf{I}(\Theta_{.k}(\mathbf{C}^{(t+1)}, \xi_\beta^{(t+1)}), \phi_k^{(t)}) = \mathbf{Z}^{\mathrm{T}}\zeta(\Theta_{.k}(\mathbf{C}^{(t+1)}, \xi_\beta^{(t+1)}), \phi_k^{(t)})\mathbf{Z}$. Now, we summarize monotone decreasing property of G-CURE algorithm in the following Theorem.

Theorem 3.1. *The sequence* $\{\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}\}$ *produced by Algorithm 5 satisfies,*

$$\begin{aligned} &F(\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}) - F(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \Phi^{(t+1)}) \\ &\qquad \ge \frac{\kappa_1}{2}\|\tilde{\mathbf{C}}^{(t+1)} - \mathbf{C}^{(t)}\|_F^2 + \frac{\kappa_2}{2}\|\mathbf{C}^{(t+1)} - \tilde{\mathbf{C}}^{(t+1)}\|_F^2 + \frac{\kappa_3}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_F^2, \end{aligned}$$

where $\kappa_1 = 1 + \eta_1(1 + (1-\alpha)\lambda) - \gamma_1$, $\kappa_2 = 1 + \eta_1(1 + (1-\alpha)\lambda) - \gamma_2$, *and* $\kappa_3 = 2 - \gamma_3$ *with parameter* $\eta_1 = \max(0, 1 - L_1)$ *for constant* $L_1 \in [0, 1]$ *fixed for a threshold rule* $\mathbf{S}$

(see definition 2.1 She [2012a]). Moreover, $\kappa_1 \geq 0$, $\kappa_2 \geq 0$ *and* $\kappa_3 \geq 0$ *can be ensured by proper scaling of* $\mathbf{X}$ *and* $\mathbf{Z}$.

The proof of the Theorem 3.1 is relegated to Appendix Section B.1. According to Theorem 3.1, $\kappa_1 \geq 0$, $\kappa_2 \geq 0$ and $\kappa_3 \geq 0$ ensures the monotone decreasing property of the algorithm. It should be noted that, in Gaussian responses $b_k''(x) = 1$ and $a_k(\phi_k) = \sigma_k^2$ resulting in $\gamma_1 \leq \|\mathbf{X}\|_2^2 / \min(\sigma_k^2)$; in Bernoulli responses $b_k''(x) = e^x/(1+e^x)^2 \leq 1/4$ and $a_k(\phi_k) = 1$ resulting in $\gamma_1 \leq \|\mathbf{X}\|_2^2/4$; and in Poisson responses $b_k''(x) = e^x$ and $a_k(\phi_k) = 1$ which results in lack of universal bound of γ_1. Thus, the scaling factor κ_1^* for Gaussian and Bernoulli responses are given by $\|\mathbf{X}\|_2 / \min(\sigma_k)$ and $\|\mathbf{X}\|_2/2$, respectively. To deal with Poisson responses case, we could choose large enough κ_1^* such that monotone descending property of G-CURE is ensured. In mixed types responses cases, either choose maximum of scaling factor κ_1^* obtained for each type of outcomes or empirically set large enough value ensuring monotone descending property. For $\kappa_2 \geq 0$, choice of corresponding scaling parameter κ_2^* is same as that of κ_1^*. To ensure $\kappa_3 \geq 0$, we perform proper scaling of $\mathbf{Z}$ with scaling parameter κ_3^* derived in similar way as that of κ_1^*. Finally, the unknown dispersion parameters are estimated based on maximizing the log-likelihood, so it is guaranteed to non-increase the objective function.

As we mainly focus on sparse estimation, we fix $\alpha = 0.95$ and only tune λ in all our numerical studies. In practice, to initialize the G-CURE algorithm, we set $\boldsymbol{\beta}^{(0)}$ as the model estimate of $\boldsymbol{\beta}$ with only control variables, $\Phi^{(0)} = \mathbf{1}$, $\mathbf{u}_j^{(0)} = 1$ for j equals the row-index of maximum unit in $|\mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \boldsymbol{\mu}(\mathbf{O})|$ else $\mathbf{u}_j^{(0)} = 0$, and obtain $(d^{(0)}, \mathbf{v}^{(0)})$ as

the beginning step. A sequence of λ ranging from λ_{max} to λ_{min} is generated to obtain a solution path. When the model is fitted for a sequence of λ values, the warm start strategy is adopted, i.e., using the solution from previous fit as the initial value for the next λ value. For model selection, we use generalize information criteria [Fan and Tang, 2013]. The details of generating a sequence of λ for a solution path and model selection criteria are relegated to Section 3.3.3 where we have discussed the two issues in detail.

3.3.2 Convergence Analysis of G-CURE

We define the *generalized constrained unit rank estimation* (G-CURE) problem in (3.20), and estimated the unknown parameters $(d, \mathbf{u}, \mathbf{v}, \boldsymbol{\beta}, \Phi)$ using Algorithm 5. We equivalently write (3.20) as

$$F(\mathbf{u}, d, \mathbf{v}, \boldsymbol{\beta}, \Phi; \mathbf{O}) = \mathcal{L}(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{O}) + \alpha\lambda \sum_{(i,j)}^{(p,q)} w_{ij}|du_i v_j| + (1-\alpha)\lambda \sum_{(i,j)}^{(p,q)} (du_i v_j)^2,$$

$$\text{s.t. } d \geq 0, \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} = n, \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1. \tag{3.34}$$

Jointly, in terms of the unknown parameters, the problem is non-convex, and a globally optimal solution is not possible. Thus, using G-CURE Algorithm 5, we obtain a partial optimal solution; for detail see definition 2.4 in Chapter 2. Using Theorem 3.2, convergence of the estimates of unknown parameters to an accumulation point [Gorski et al., 2007b] is guaranteed. Before providing details of the theorem, we define a sequence of set $\{\text{Ł}^s\}_{s\in\mathbb{N}} = (d_\lambda^{(s)}, \mathbf{u}_\lambda^{(s)}, \mathbf{v}_\lambda^{(s)}, \boldsymbol{\beta}_\lambda^{(s)}, \Phi_\lambda^{(s)})$ for unknown parameters estimates in sth iteration

of G-CURE algorithm corresponding to the tunning parameter λ.

Theorem 3.2. *Consider the optimization problem in* (3.34) *with* $\lambda > 0$ *and* $0 < \alpha < 1$. *Assume the weights* $\{w_{ij}\}$ *and the data* $(\mathbf{Y}, \mathbf{X})$ *are finite, and the initial value* $\mathbf{u}_\lambda^0$ *satisfies* $\arg\min_{\check{\mathbf{v}}} F(\check{\mathbf{v}}; \mathbf{u}_\lambda^0, \boldsymbol{\beta}^0, \lambda) \neq \mathbf{0}$, *i.e.,* $\widetilde{d}_\lambda^0 \neq 0$. *Then the sequence* $\{L^s\}_{s\in\mathbb{N}}$ *generated by the G-CURE algorithm is uniformly bounded and has at least one accumulation point. Moreover, all accumulation points are coordinatewise minimum points and have the same objective value, and* $F(L^s; \lambda)$ *converges monotonically to* $F(L^*; \lambda)$ *where* $L^* = (d^{(\lambda)}, \mathbf{u}^{(\lambda)}, \mathbf{v}^{(\lambda)}, \boldsymbol{\beta}^{(\lambda)}, \Phi^{(\lambda)})$ *is coordinatewise minimum point.*

We acknowledge that our convergence proof mainly follows the framework developed in proving Theorem 2.3 of Chapter 2; see also Gorski et al. [2007b]. Details of the proof are relegated to Appendix Section B.2. It is important to note here that Theorem 3.2 is subjected to condition of $\widetilde{d}_\lambda^0 \neq 0$, and Proposition 3.3 (for proof see Theorem 2.11 of Chapter 2) ensures that required condition is met.

Proposition 3.3. *Consider solving* (3.34) *with* $\lambda > 0$ *and* $0 < \alpha < 1$ *using the G-CURE algorithm. If* $\arg\min_{\check{\mathbf{v}}} F(\check{\mathbf{v}}; \mathbf{u}_\lambda^0, \boldsymbol{\beta}^0, \lambda) \neq \mathbf{0}$, *i.e.,* $\widetilde{d}_\lambda^0 \neq 0$, *then* $\widetilde{d}_\lambda^s \neq 0$ *and* $d_\lambda^{s+1} \neq 0$ *for any* $s \geq 0$. *Let* $(d^{(\lambda)}, \mathbf{u}^{(\lambda)}, \mathbf{v}^{(\lambda)}, \boldsymbol{\beta}^{(\lambda)}, \Phi^{(\lambda)})$ *be the solution of* (3.34) *with tuning parameter* λ. *If* $d^{(\lambda_1)} \neq 0$, *then* $d^{(\lambda_2)} \neq 0$ *for any* $\lambda_2 \leq \lambda_1$. *Moreover, if* $d^{(\lambda_1)} \neq 0$, *then setting* $\mathbf{v}_{\lambda_2}^0 = \mathbf{v}^{(\lambda_1)}$ *ensures that* $\widetilde{d}_{\lambda_2}^s \neq 0$ *and* $d_{\lambda_2}^{s+1} \neq 0$ *for any* $s \geq 0$.

3.3.3 Tuning and Rank Selection

For G-CURE analysis using its generic formulation (3.20), we fit the model for a grid of 100 λ values equally spaced on the log-scale between $[\lambda_{\max}, \lambda_{\min}]$, to cover a spectrum of sparsity patterns in $\mathbf{u}$ and $\mathbf{v}$. Here $\lambda_{\min}$ is taken as a fraction of $\lambda_{\max}$ at which the model has excessive number of non-zero coefficients (in our numerical studies we set $\lambda_{\min} = \lambda_{\max} \times 10^{-6}$), and $\lambda_{\max}$ is the smallest λ at which the estimated singular value becomes zero. She [2012b] proposed the value of $\lambda_{\max}$ to be $\|\mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \boldsymbol{\mu}(\mathbf{0}))\|_{\infty}$.

For a given tunning parameter λ, we obtain estimates of the parameters ($\mathbf{C}$, $\boldsymbol{\beta}$, Φ) using G-CURE Algorithm 5. Let us specify the estimated values corresponding to tunning parameter λ as ($\widehat{\mathbf{C}}^{(\lambda)}$, $\widehat{\boldsymbol{\beta}}^{(\lambda)}$, $\widehat{\Phi}^{(\lambda)}$). In G-CURE setup, the estimated natural parameter matrix $\Theta^{(\lambda)} = \Theta(\widehat{\mathbf{C}}^{(\lambda)}, \widehat{\boldsymbol{\beta}}^{(\lambda)})$ defined according to definition (3.3). Once the solution path is obtained, we need to choose the optimal solution along the path. For small-scale problems, λ can be chosen by cross validation [Stone, 1974]. Alternatively, various information criteria have been widely used due to their computational efficiency. In our numerical studies, we use *generalize information criterion(GIC)* [Fan and Tang, 2013] because of its superior performance in sparse learning, defined as

$$\mathrm{GIC}(\lambda) = \mathbf{D}(\widehat{\boldsymbol{\mu}}_{\lambda}; \mathbf{Y})/(nq) + \{\log\log(nq)\log(pq)/(nq)\}df(\lambda)$$

such that deviance measuring model fit is given by

$$\mathbf{D}(\widehat{\boldsymbol{\mu}}_\lambda; \mathbf{Y}) = 2\{\mathcal{L}(\widehat{\boldsymbol{\mu}}_\lambda; \mathbf{Y}) - \mathcal{L}(\mathbf{Y}; \mathbf{Y})\}. \tag{3.35}$$

Here $\mathcal{L}(\widehat{\boldsymbol{\mu}}_\lambda; \mathbf{Y})$ is the negative log-likelihood of a multivariate GLM with canonical link function, calculated using mean structure given by $\widehat{\boldsymbol{\mu}}_\lambda = \mathbf{B}^{'}(\Theta^{(\lambda)})$ and response matrix $\mathbf{Y}$. Similarly, $\mathcal{L}(\mathbf{Y}; \mathbf{Y})$ is the negative log-likelihood of the corresponding saturated model. $df(\lambda)$ is the model degrees of freedom, which is estimated by $\widehat{df}(\lambda) = q(p_z + 1) + \sum_{i=1}^{p} \mathrm{I}(\widehat{u}_i^{(\lambda)} \neq 0) + \sum_{j=1}^{q} \mathrm{I}(\widehat{v}_j^{(\lambda)} \neq 0) - 1$ with $\mathrm{I}(\cdot)$ being the indicator function.

3.4 Theoretical Properties

3.4.1 Asymptotic Results

We model dependency of responses $\mathbf{Y} \in \mathbb{R}^{n \times q}$ using control variables $\mathbf{Z} \in \mathbb{R}^{n \times p_z}$ and predictors $\mathbf{X} \in \mathbb{R}^{n \times p}$, and for this assume that there exists a regression model given by

$$\mathbf{Y} = \mathbf{B}^{'}(\Theta^*) + \mathbf{E}, \tag{3.36}$$

where matrix $\mathbf{E} \in \mathbb{R}^{n \times q}$ is a sub-exponential random error matrix [Xu, 2016], and $\mathbf{B}^{'}(\cdot)$ (see (3.5)) is a function operator defined on the natural parameter matrix $\Theta^* = \mathbf{X}\mathbf{C}^* + \mathbf{Z}\boldsymbol{\beta}^*$ for unknown predictor and control variable coefficient matrix $\mathbf{C}^* \in \mathbb{R}^{p \times q}$

and $\boldsymbol{\beta}^* \in \mathbb{R}^{p_z \times q}$ respectively. We consider the following standard assumptions on the design matrices $(\mathbf{X}, \mathbf{Z})$ and the random error matrix $\mathbf{E}$.

A1. $(1/n)\mathbf{X}^{\mathrm{T}}\mathbf{X} \xrightarrow{a.s} \boldsymbol{\Gamma}_1$, $(1/n)\mathbf{Z}^{\mathrm{T}}\mathbf{Z} \xrightarrow{a.s} \boldsymbol{\Gamma}_2$ and $(1/n)\mathbf{X}^{\mathrm{T}}\mathbf{Z} \xrightarrow{a.s} \mathbf{0}$ as $n \to \infty$, where $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ are fixed, positive definite matrix.

A2. The elements of error matrix $\mathbf{E}$, i.e., e_{ik}, $i = 1, \ldots, n, k = 1, \ldots, q$, are independent and identically distributed (i.i.d.) with $\mathrm{E}(\mathbf{e}_i) = 0$, and defined using sub-exponential parameter (σ^2, b) [Xu, 2016].

Now consider the sparse factor structure in the true coefficient matrix $\mathbf{C}^*$. Let $r^* = \mathrm{rank}(\mathbf{C}^*) \leq \min(p, q)$ be rank of the true model. Let us assume that there exists a decomposition of the low-rank coefficient matrix $\mathbf{C}^*$ given by

$$\mathbf{C}^* = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*\mathrm{T}}, \qquad \text{s.t. } \mathbf{U}^{*\mathrm{T}}\boldsymbol{\Gamma}_1\mathbf{U}^* = \mathbf{I}_{r^*}, \mathbf{V}^{*\mathrm{T}}\mathbf{V}^* = \mathbf{I}_{r^*}, \tag{3.37}$$

where $\mathbf{I}_{r^*}$ denotes the $r^* \times r^*$ identity matrix, $\mathbf{U}^* = [\mathbf{u}_1^*, \ldots, \mathbf{u}_{r^*}^*] \in \mathbb{R}^{p \times r^*}$ and $\mathbf{V}^* = [\mathbf{v}_1^*, \ldots, \mathbf{v}_{r^*}^*] \in \mathbb{R}^{q \times r^*}$ are the singular vector matrices, and $\mathbf{D}^* = \mathrm{diag}\{d_1^*, \ldots, d_{r^*}^*\}$ is a diagonal matrix of singular values $d_1^*, \ldots, d_{r^*}^* > 0$. Both $\mathbf{U}^*$ and $\mathbf{V}^*$ are assumed to be sparse matrices. We write $\mathbf{C}_k^* = d_k^*\mathbf{u}_k^*\mathbf{v}_k^{*\mathrm{T}}$ and $\mathbf{C}^* = \sum_{k=1}^{r^*} \mathbf{C}_k^*$. Without much loss of generality, we assume that all the non-zero singular values are distinct.

A3. $d_1^* > \ldots > d_{r^*}^* > 0$.

Then each $\mathbf{C}_k^*$ is fully identifiable, and each triplet $(d_k^*, \mathbf{u}_k^*, \mathbf{v}_k^*)$ is identifiable up to

the signs of the two vectors, i.e., $\mathbf{u}_k^* \mathbf{v}_k^{*\mathrm{T}} = (-\mathbf{u}_k^*)(-\mathbf{v}_k^{*\mathrm{T}})$. Additionally, we make an assumption on double derivative of $b_k(\theta_{ik})$ given by

A4.

$$\max_{1\le i\le n, 1\le k\le q} \sup_{\boldsymbol{\beta},\mathbf{C}} |b_k''(\theta_{ik})| \le \bar{\gamma}, \qquad \text{and} \qquad \min_{1\le i\le n, 1\le k\le q} \inf_{\boldsymbol{\beta},\mathbf{C}} |b_k''(\theta_{ik})| \ge \gamma.$$

Now, to simplify the problem setup for facilitating our asymptotic analysis, we let each singular value d_k^* be absorbed to each pair $(\mathbf{u}_k^*, \mathbf{v}_k^*)$ in the decomposition (3.37). Specially, for each $\mathbf{v}_k^*$ vector, $k = 1, \ldots, r^*$, there must exist at least one entry that is nonzero, i.e., $v_{\ell_k k}^* \neq 0$ for some $1 \le \ell_k \le q$. Accordingly, we reparameterize $\mathbf{C}_k^*$ as

$$\mathbf{C}_k^* = \mathbf{u}_k^* \mathbf{v}_k^{*\mathrm{T}}, \qquad \text{s.t.} \qquad v_{\ell_k k}^* = 1.$$

Here, with some abuse of notation, $(\mathbf{u}_k^*, \mathbf{v}_k^*)$ are re-defined, and they are re-scaled versions of their original counterparts in order to absorb the singular value, i.e., the new pair satisfies

$$(\mathbf{u}_k^{*\mathrm{T}} \boldsymbol{\Gamma} \mathbf{u}_k^*)(\mathbf{v}_k^{*\mathrm{T}} \mathbf{v}_k^*) = d_k^*.$$

Consequently, we parameterize the coefficient matrix $\mathbf{C}^*$ as

$$\mathbf{C}^* = \mathbf{U}^* \mathbf{V}^{*\mathrm{T}}, \qquad \text{s.t. } \mathbf{U}^{*\mathrm{T}} \boldsymbol{\Gamma} \mathbf{U}^* \text{ and } \mathbf{V}^{*\mathrm{T}} \mathbf{V}^* \text{ are both diagonal matrices,} \qquad (3.38)$$

$$v_{\ell_k k}^* = 1, k = 1, \ldots, r^*.$$

Now all the elements in $\mathbf{U}^*$ and $\mathbf{V}^*$ are uniquely defined and fully identifiable. For $k > r^*$, it is obvious that $\mathbf{C}_k^* = \mathbf{0}$ and its corresponding singular vectors become unidentifiable; as such, we can set $\mathbf{u}_k^* = \mathbf{0}$ and still choose $\mathbf{v}_k^*$ to be a nonzero vector with a unit entry.

Now, consider our proposed G-CURE approach. Corresponding to the new parameterization (3.38), using generic optimization problem (3.20), the objective function in the kth step of G-CURE can be written as

$$F_k^{(n)}(\mathbf{u}, \mathbf{v}, \boldsymbol{\beta}, \Phi) = \mathcal{L}(\mathbf{C}, \boldsymbol{\beta}, \Phi; \mathbf{O}_k) + \alpha\lambda_k^{(n)} \sum_{i=1}^{p}\sum_{j=1}^{q} w_{ijk}|u_i v_j| + (1-\alpha)\lambda_k^{(n)} \sum_{i=1}^{p}\sum_{j=1}^{q} (u_i v_j)^2, \tag{3.39}$$

where $\mathbf{u} \in \mathbb{R}^p$, $\mathbf{v} \in \mathbb{R}^q$ with $v_{\ell_k} = 1$, $\mathbf{C} = \mathbf{u}\mathbf{v}^{\mathrm{T}}$ and $\mathbf{O}_1 = \mathbf{O}$, $\mathbf{O}_k = \mathbf{O} + \sum_{1 \le h \le k-1} \mathbf{X}\widehat{\mathbf{u}}_h\widehat{\mathbf{v}}_h^{\mathrm{T}}$, where $(\widehat{\mathbf{u}}_h, \widehat{\mathbf{v}}_h, \widehat{\boldsymbol{\beta}}, \widehat{\Phi}) = \arg\min F_h^{(n)}(\mathbf{u}, \mathbf{v}, \boldsymbol{\beta}, \Phi)$. For simplicity assume that $\mathbf{O} = \mathbf{0}$. Here $w_{ijk} = w_{ik}w_{jk}$, where $w_{ik} = |\widetilde{u}_{ik}|^{-\gamma}$ and $w_{jk} = |\widetilde{v}_{jk}|^{-\gamma}$ with some $\gamma > 0$. The regularization parameter $\lambda_k^{(n)}$ is a function of the sample size, but $0 < \alpha \le 1$ is considered as a fixed constant. For the purpose of proof, we assume that $\Phi = \mathbf{I}$.

We define some notations. Suppose $\mathbf{Z}$ is an arbitrary matrix, and $\mathcal{A}$ and $\mathcal{B}$ are subsets of the collection of row and column indices of $\mathbf{Z}$, respectively. We let $\mathbf{Z}_{\mathcal{A}\mathcal{B}}$ denote a submatrix of $\mathbf{Z}$ whose rows and columns are chosen from $\mathbf{Z}$ according to the index sets $\mathcal{A}$ and $\mathcal{B}$, respectively. For simplicity, we write $\mathbf{Z}_{\mathcal{A}\mathcal{A}} = \mathbf{Z}_{\mathcal{A}}$ when $\mathbf{Z}$ is a square matrix, $\mathbf{Z}_{\mathcal{A}\mathcal{B}} = \mathbf{Z}_{\cdot\mathcal{B}}$ ($\mathbf{Z}_{\mathcal{A}\cdot}$) when $\mathcal{A}$ ($\mathcal{B}$) consists of all the row (column) indies, and $\mathbf{Z}_{\mathcal{A}\cdot} = \mathbf{Z}_{\mathcal{A}}$ when $\mathbf{Z}$ is a vector. For kth column of matrix $\mathbf{Z}$ given by $\mathbf{z}_k$, we represent element in

set $\mathcal{A}$ by notation $\mathbf{z}_{k\mathcal{A}}$.

Theorem 3.4. *(Existence of Local Minimum). Suppose **A1**–**A4** are satisfied, and suppose that* $\lambda_k^{(n)}/\sqrt{n} \to \lambda_k \geq 0$ *as* $n \to \infty$ *along with fixed dispersion parameter* $\Phi = \mathbf{I}$. *Then there is a local minimizer* $(\widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k, \widehat{\boldsymbol{\beta}})$ *of* $F_k^{(n)}(\mathbf{u}, \mathbf{v}, \boldsymbol{\beta}, \Phi)$, *such that*

i. $\|\widehat{\mathbf{u}}_k - \mathbf{u}_k^*\| = O_p(n^{-1/2})$, $\|\widehat{\mathbf{v}}_k - \mathbf{v}_k^*\| = O_p(n^{-1/2})$, *and* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| = O_p(n^{-1/2})$ *for* $k = 1, \ldots, r^*$.

ii. $|\widehat{d}_k| = O_p(n^{-1/2})$ *where* $\widehat{d}_k = (1/n)(\widehat{\mathbf{u}}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\mathbf{u}}_k)(\widehat{\mathbf{v}}_k^{\mathrm{T}}\widehat{\mathbf{v}}_k)$, *for* $k = r^* + 1, \ldots, r$.

Proof of the theorem is relegated to Appendix Section (B.3).

3.5 Simulation

3.5.1 Simulation Setting

We have considered several simulation settings which include cases with similar or mixed type of multivariate response variables. The underling distribution in the former case mainly covers response variables from either Gaussian, binomial or Poisson family of distributions, and in latter case their combinations. Simulated examples in mixed type case have a) Gaussian/Bernoulli and b) Gaussian/Poisson outcome variables. In all the cases, the true rank of $\mathbf{C}^* \in \mathbb{R}^{p\times q}$ is $r^* = 3$ with $\mathbf{D}^* = s \times \mathrm{diag}(d_1, d_2, d_{r^*})$ where $d_i^* = s \times d_i$ for $i = 1, .., r^*$. We set $d_1 = 6$, $d_2 = 5$, $d_3 = 3$, and vary s to specify *signal strength* in different cases. In our simulation, Model I refers to a low dimensional

example with $p = 30$, $q = q_1 + q_2 + q_3$ and $n = 200$ where $\{q_1, q_2, q_3\}$ are number of {Gaussian, binomial, Poisson} outcome variables. Simulated models with all Gaussian responses have $q_1 = 30$, all Bernoulli responses have $q_2 = 30$, all Poisson responses have $q_3 = 30$, Gaussian/Bernoulli responses have $q_1 = 15, q_2 = 15$, and Gaussian/Poisson responses have $(q_1 = 15, q_3 = 15)$.

For the similar response case, $\mathbf{u}_k^*$ is generated as $\mathbf{u}_k^* = \check{\mathbf{u}}_k / \|\check{\mathbf{u}}_k\|$, where $\check{\mathbf{u}}_1 = [\text{unif}(\mathcal{A}_u, 8), \text{rep}(0, p - 8)]^{\text{T}}$, $\check{\mathbf{u}}_2 = [\text{rep}(0, 5), \text{unif}(\mathcal{A}_u, 9), \text{rep}(0, p - 14)]^{\text{T}}$, and $\check{\mathbf{u}}_3 = [\text{rep}(0, 11), \text{unif}(\mathcal{A}_u, 9), \text{rep}(0, p - 20)]^{\text{T}}$; $\mathbf{v}_k^*$ is generated as $\mathbf{v}_k^* = \check{\mathbf{v}}_k / \|\check{\mathbf{v}}_k\|$, where $\check{\mathbf{v}}_1 = [\text{unif}(\mathcal{A}_v, 5), \text{rep}(0, q - 5)]^{\text{T}}$, $\check{\mathbf{v}}_2 = [\text{rep}(0, 5), \text{unif}(\mathcal{A}_v, 5), \text{rep}(0, q - 10)]^{\text{T}}$, and $\check{\mathbf{v}}_3 = [\text{rep}(0, 10), \text{unif}(\mathcal{A}_v, 5), \text{rep}(0, q - 15)]^{\text{T}}$. To depict usefulness of joint learning using GSeCURE in the mixed type response case, we generate $\mathbf{v}_k^*$ using $\check{\mathbf{v}}_k = [\overline{\mathbf{v}}_k, \overline{\mathbf{v}}_k]^{\text{T}}$ for $k = 1, 2, 3$, where $\overline{\mathbf{v}}_1 = [\text{unif}(\mathcal{A}_u, 5), \text{rep}(0, q/2 - 5)]$, $\overline{\mathbf{v}}_2 = [\text{rep}(0, 3), \overline{v}_{14}, -\overline{v}_{15}, \text{unif}(\mathcal{A}_u, 3), \text{rep}(0, q/2 - 8)]$, and $\overline{\mathbf{v}}_3 = [\overline{v}_{11}, -\overline{v}_{12}, \text{rep}(0, 4), \overline{v}_{27}, -\overline{v}_{28}, \text{unif}(\mathcal{A}_u, 2), \text{rep}(0, q - 10)]$. The notation $\text{unif}(\mathcal{A}, b)$ denotes a vector of length b whose entries are i.i.d. uniformly distributed on set $\mathcal{A}$; we use $\mathcal{A}_u = \pm 1$, $\mathcal{A}_v = [-1, -0.3] \cup [0.3, 1]$. Notation $\text{rep}(a, b)$ denotes a vector of length b, whose entries are all equal to a. Control variable $\mathbf{Z} = \mathbf{1}_n$ has only an intercept term with coefficient $\boldsymbol{\beta} = [\text{rep}(0.5, q)]^{\text{T}}$.

Model II is a high-dimensional model with $q = 30$ and $p = 300$. Model parameter $\mathbf{u}_k^*$ is generated by appending $\mathbf{u}_k^*$ from Model I with zeros such that $p = 300$. Generation of $\mathbf{v}_k^*$ remains unchanged. The predictor matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ is generated such that latent

factors in $\mathbf{XU}^*/\sqrt{n}$ are orthogonal; for details of generation mechanism, see Section 2.6.1 of Chapter 2. With specific signal strength s for a simulated example, the natural parameter matrix is then constructed as $\boldsymbol{\theta}^* = [\theta^*_{ik}] = \mathbf{Z}\boldsymbol{\beta}^* + \mathbf{XC}^*$. The response matrix $\mathbf{Y}$ is then generated by model (3.1) with dispersion parameter for all Gaussian responses set to be error variance σ^2 obtained at SNR = 0.5; for the definition of SNR see Section 2.6.1 of Chapter 2. In addition, GSeCURE can efficiently handle missing entries in response matrix $\mathbf{Y}$, and to demonstrate that we have considered examples with 20% missing entries in $\mathbf{Y}$.

3.5.2 Methods and Evaluation Criteria

We compare the performance of the GSeCURE algorithm in terms of estimation, prediction and sparsity recovery to that of several other modeling strategies for a given type of multivariate responses, i.e., Gaussian, non-Gaussian, mixed. A simple procedure is to model responses marginally, i.e., each column in the response matrix $\mathbf{Y}$ is fitted with a univariate generalized linear model uGLM. We have used its penalized version available in *glmnet* R package with elastic-net penalty. A recent unpublished manuscript by Chen et al. [2017] for *mixed-response reduced-rank regression* (mRRR) model propose a joint estimation approach to recover rank r coefficient matrix via a) nuclear norm and b) rank penalty on $\mathbf{C}$, denoted by mRRR.n and mRRR.r, respectively. Also, one can sequentially perform r unit-rank constrained mRRR.r analysis to obtain rank-r coefficient matrix, referred to as mRRR.rs. Another possible approach in mixed type responses cases is to

use the GSeCURE algorithm to obtain model estimates for each type separately, and then combine the results obtained to get estimate of coefficient matrix, referred to as GSeCURE.s. To depict effectiveness of the GSeCURE algorithm in the case with missing entries in response matrix $\mathbf{Y}$, we use the available version of mRRR.n, mRRR.r, mRRR.rs and GSeCURE.s dealing with missing entries. uGLM is not applicable for such scenario, but a possible approach is to obtain marginal model estimate only on the basis of available observations in columns of $\mathbf{Y}$. The missing entries in Y are randomly selected, and we consider missing proportions M% $= \{0\%; 20\%\}$. The experiment was replicated 100 times under each setup.

Model estimation is compared in terms of error in coefficient matrix $\text{Er}(\widehat{\mathbf{C}}) = \|\widehat{\mathbf{C}} - \mathbf{C}^*\|_F/(pq)$, and error in dispersion parameter $\text{Er}(\widehat{\Phi}) = \|\widehat{\Phi} - \Phi^*\|_2/q$. Error in the natural parameter Θ estimate is shown using $\text{Er}(\mathbf{Y}) = \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_F/(nq)$ to depict prediction accuracy. In the mixed type response case especially, prediction performance for Gaussian responses and non-Gaussian responses are shown by $\text{Er}_g(\mathbf{Y}) = \|\widehat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g\|_F/(nq_1)$ and $\text{Er}_{ng}(\mathbf{Y}) = \|\widehat{\boldsymbol{\theta}}_{ng} - \boldsymbol{\theta}_{ng}\|_F/\{n(q-q_1)\}$ respectively, where $\{\boldsymbol{\theta}_g, \boldsymbol{\theta}_{ng}\}$ and its estimate $\{\widehat{\boldsymbol{\theta}}_g, \widehat{\boldsymbol{\theta}}_{ng}\}$ are the natural parameter matrices corresponding to their respective outcomes. The sparsity recovery in decomposition of the coefficient matrix is characterized by the false positive rate (FPR) and the false negative rate (FNR), calculated from comparing the sparsity pattern of $(\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k)$ to that of $(\mathbf{u}_k^*, \mathbf{v}_k^*)$ for $k = 1, \ldots, r$. Estimated ranks are reported in metric r. The rank estimation performance is compared using amount of signal left in (r^*+1)th to rth component via $R\% = 100(\sum_{i=r^*+1}^{r} d_i^2)/(\sum_{i=1}^{r} d_i^2)$.

3.5.3 Simulation Results

We have compared various model fitting strategies and reported their performances in Table 8 – Table 17. Boxplots in figures 10 – 12 compare model prediction Er($\mathbf{Y}$) and estimation Er($\widehat{\mathbf{C}}$) performance for Model I and Model II in the case of similar type responses. A similar comparison is made in Figure 13 and 14 for mixed type response case. In similar type response cases, except in Bernoulli multivariate outcomes Model II, we observe that GSeCURE shows better results in terms of a) estimation accuracy by low value of Er($\widehat{\mathbf{C}}$), b) prediction in term of low value of Er($\mathbf{Y}$), c) efficient sparsity recovery by low value of FPR and FNR, and d) consistent rank selection r. Bernoulli outcomes multivariate models have less information/signal strength in comparison with other models, thus they require more instances of observed data to have enough signal. In high-dimensional setting with $p > n$, due lack of enough signal strength efficient recovery is not possible sequentially, thus resulting in such under performance of GSeCURE.

In the mixed type response model, we again see superior performance of GSeCURE in comparison to other models. It should be noted that joint modeling of multivariate mixed outcomes through GSeCURE learns better model estimates in comparison with GSeCURE.s in which separate model estimation is performed for each type. Thus, GSeCURE efficiently induces learning across each type of outcome. To understand importance of joint/induced learning, it should be noted that with only Bernoulli outcomes in Model II, GSeCURE do not perform well; see Figure 11. But, in the same high dimensional setting Model II with Gaussian/Bernoulli outcomes, we observed significantly

better performance due to induced learning; see Figure 14. GSeCURE can efficiently handle missing entries in response matrix $\mathbf{Y}$, and results corresponding to 20% missing cases are also shown in Table 8 – Table 17. In such a setting, performance in all models deteriorates, but GSeCURE still shows superior performance in comparison with other methods.

Table 8: Simulation: results of Model I with Gaussian responses at signal strength s = 1 .

	Er($\widehat{\mathbf{C}}$)	Er($\mathbf{Y}$)	Er(Φ)	FPR	FNR	R%	r
				M% = 0			
GSeCURE	7.42 (1.85)	1.82 (0.42)	0.01 (0.01)	0.15 (0.35)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.r	18.77 (2.28)	4.08 (0.51)	0.02 (0.01)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.rs	18.73 (2.27)	4.07 (0.50)	0.02 (0.01)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.n	51.25 (9.00)	8.50 (1.20)	0.03 (0.01)	100.00 (0.00)	0.00 (0.00)	0.14 (0.08)	6.98 (1.15)
uGLM	57.21 (7.68)	10.61 (1.28)	0.02 (0.01)	100.00 (0.00)	0.00 (0.00)	2.80 (0.39)	26.39 (1.01)
				M% = 20			
GSeCURE	9.73 (2.50)	2.37 (0.57)	0.01 (0.01)	0.26 (0.47)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.r	23.65 (3.09)	5.10 (0.61)	0.04 (0.01)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.rs	24.03 (3.12)	5.22 (0.62)	0.04 (0.01)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.n	75.51 (14.60)	12.56 (2.08)	0.03 (0.01)	100.00 (0.00)	0.00 (0.00)	0.07 (0.05)	5.56 (0.86)
uGLM	73.12 (10.45)	13.59 (1.70)	0.02 (0.01)	100.00 (0.00)	0.00 (0.00)	3.57 (0.50)	26.72 (1.29)

Table 9: Simulation: results of Model I with Bernoulli responses at signal strength s = 1.5.

	Er($\widehat{\mathbf{C}}$)	Er($\mathbf{Y}$)	FPR	FNR	R%	r
			M% = 0			
GSeCURE	97.19 (30.14)	24.64 (7.14)	3.56 (4.69)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.r	168.16 (23.22)	38.93 (5.25)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.rs	158.77 (22.52)	36.40 (5.73)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.n	421.75 (44.15)	100.22 (11.76)	100.00 (0.00)	0.00 (0.00)	2.47 (0.53)	12.24 (0.69)
uGLM	435.11 (44.42)	89.88 (8.47)	100.00 (0.00)	0.00 (0.00)	7.83 (0.81)	26.63 (1.28)
			M% = 20			
GSeCURE	128.13 (41.55)	33.99 (12.04)	2.90 (3.92)	0.38 (0.80)	0.00 (0.00)	3.00 (0.00)
mRRR.r	218.30 (32.69)	50.95 (8.67)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.rs	203.52 (31.03)	46.07 (7.23)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.n	488.29 (56.24)	116.63 (15.23)	100.00 (0.00)	0.00 (0.00)	3.27 (1.03)	12.31 (1.59)
uGLM	541.65 (56.19)	112.66 (11.67)	100.00 (0.00)	0.00 (0.00)	9.66 (1.05)	26.37 (1.35)

Table 10: Simulation: results of Model I with Poisson responses at signal strength s = 0.5 .

	$\text{Er}(\widehat{\mathbf{C}})$	$\text{Er}(\mathbf{Y})$	FPR	FNR	R%	r
	M% = 0					
GSeCURE	3.81 (1.42)	0.93 (0.27)	4.03 (4.82)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.r	26.43 (5.42)	4.35 (0.60)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.rs	22.62 (4.29)	3.91 (0.59)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.n	35.27 (4.91)	5.95 (0.68)	100.00 (0.00)	0.00 (0.00)	1.01 (0.69)	9.42 (2.80)
uGLM	22.58 (1.96)	4.36 (0.29)	100.00 (0.00)	0.00 (0.00)	4.32 (0.53)	26.46 (0.99)
	M% = 20					
GSeCURE	4.87 (1.79)	1.16 (0.32)	2.65 (3.50)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.r	37.21 (9.53)	7.58 (3.13)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.rs	34.44 (11.49)	7.04 (3.40)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.n	44.28 (5.30)	7.72 (0.81)	100.00 (0.00)	0.00 (0.00)	1.39 (1.12)	9.43 (3.19)
uGLM	29.07 (2.81)	5.61 (0.38)	100.00 (0.00)	0.00 (0.00)	5.47 (0.74)	26.42 (1.34)

Table 11: Simulation: results of Model I with Gaussian/Bernoulli responses at signal strength s = 1.5.

	Er($\widehat{\mathbf{C}}$)	Er($\mathbf{Y}$)	Er_g(Y)	Er_{ng}(Y)	Er(Φ)	FPR	FNR	R%	r
GSeCURE	123.79 (29.32)	35.43 (7.85)	6.12 (1.94)	64.92 (15.04)	1.07 (0.34)	0.44 (0.81)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
GSeCURE.s	203.06 (54.84)	50.80 (12.38)	6.88 (2.02)	95.07 (25.47)	1.07 (0.34)	51.52 (1.37)	0.00 (0.00)	16.03 (27.47)	6.00 (0.00)
mRRR.r	171.21 (18.06)	39.87 (5.00)	12.38 (1.65)	67.41 (9.63)	1.02 (0.33)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.rs	202.83 (20.20)	48.37 (5.71)	13.78 (1.86)	82.96 (10.86)	1.07 (0.34)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.n	374.57 (23.79)	84.53 (6.87)	25.96 (3.19)	143.10 (12.48)	1.21 (0.40)	100.00 (0.00)	0.00 (0.00)	0.49 (0.18)	8.78 (1.14)
uGLM	302.07 (31.77)	59.42 (6.10)	23.92 (2.46)	95.12 (11.90)	0.78 (0.27)	100.00 (0.00)	0.00 (0.00)	5.47 (0.61)	24.79 (1.46)
					M% = 20				
GSeCURE	142.12 (38.58)	40.00 (10.05)	7.86 (2.01)	71.51 (18.48)	1.07 (0.36)	0.80 (1.21)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
GSeCURE.s	287.90 (93.97)	71.41 (22.48)	9.91 (2.97)	132.03 (43.80)	1.08 (0.37)	51.45 (1.67)	0.00 (0.00)	21.17 (33.41)	5.76 (0.43)
mRRR.r	210.07 (21.39)	50.83 (6.26)	15.29 (2.23)	86.31 (12.06)	0.25 (0.11)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.rs	234.40 (22.79)	57.17 (6.69)	17.09 (2.31)	97.09 (12.64)	0.27 (0.12)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.n	493.26 (39.40)	113.69 (10.10)	36.65 (5.46)	190.72 (16.91)	0.45 (0.17)	100.00 (0.00)	0.00 (0.00)	0.26 (0.14)	6.52 (0.92)
uGLM	382.55 (44.34)	76.49 (9.24)	31.64 (3.74)	121.93 (18.63)	0.70 (0.25)	100.00 (0.00)	0.00 (0.00)	6.86 (0.81)	24.89 (1.48)

Table 12: Simulation: results of Model I with Gaussian/Poisson responses at signal strength s = 0.3 .

	Er($\widehat{\mathbf{C}}$)	Er($\mathbf{Y}$)	Er_g(Y)	Er_{ng}(Y)	Er(Φ)	FPR	FNR	R%	r
					M% = 0				
GSeCURE	2.25 (0.65)	0.59 (0.15)	0.25 (0.07)	0.94 (0.27)	0.05 (0.00)	0.44 (0.72)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
GSeCURE.s	33.71 (10.73)	9.03 (3.78)	14.33 (7.27)	3.73 (1.46)	0.03 (0.01)	40.58 (16.50)	20.36 (18.79)	1.20 (1.80)	3.78 (1.65)
mRRR.r	42.63 (4.28)	11.53 (0.98)	11.04 (1.01)	12.08 (1.03)	0.03 (0.00)	33.61 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)
mRRR.rs	42.63 (4.28)	11.53 (0.98)	11.04 (1.01)	12.08 (1.03)	0.03 (0.00)	33.61 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)
mRRR.n	51.04 (8.63)	12.81 (3.36)	11.45 (3.88)	14.17 (2.86)	0.03 (0.00)	75.78 (23.61)	24.98 (21.14)	0.00 (0.00)	2.22 (0.64)
uGLM	14.77 (1.58)	2.80 (0.27)	0.96 (0.09)	4.63 (0.53)	0.05 (0.00)	100.00 (0.00)	0.00 (0.00)	7.92 (0.92)	24.89 (1.59)
					M% = 20				
GSeCURE	3.18 (1.00)	0.82 (0.23)	0.34 (0.10)	1.33 (0.44)	0.05 (0.00)	0.90 (1.21)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
GSeCURE.s	33.15 (11.94)	8.47 (4.03)	11.78 (7.79)	5.01 (1.14)	0.03 (0.01)	39.29 (16.17)	20.59 (19.54)	1.62 (1.90)	4.03 (1.69)
mRRR.r	43.87 (4.23)	11.89 (0.98)	11.08 (0.97)	12.70 (0.96)	0.03 (0.00)	33.61 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)
mRRR.rs	43.87 (4.23)	11.89 (0.98)	11.08 (0.97)	12.70 (0.96)	0.03 (0.00)	33.61 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)
mRRR.n	48.10 (10.17)	12.05 (3.70)	9.82 (4.33)	14.16 (2.95)	0.04 (0.01)	84.59 (26.78)	16.98 (23.24)	0.00 (0.00)	2.47 (0.72)
uGLM	18.79 (1.76)	3.62 (0.31)	1.24 (0.14)	5.97 (0.60)	0.05 (0.00)	100.00 (0.00)	0.00 (0.00)	9.86 (1.19)	24.67 (1.77)

Table 13: Simulation: results of Model II with Gaussian responses at signal strength s = 1 .

	$\text{Er}(\widehat{\mathbf{C}})$	$\text{Er}(\mathbf{Y})$	$\text{Er}(\Phi)$	FPR	FNR	R%	r
				M% = 0			
GSeCURE	1.66 (0.75)	3.13 (1.11)	0.01 (0.01)	0.15 (0.16)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.r	51.47 (8.00)	93.29 (37.14)	3.16 (1.98)	44.28 (16.46)	54.61 (16.17)	0.00 (0.00)	1.33 (0.49)
mRRR.rs	51.32 (8.05)	93.05 (37.49)	3.16 (1.98)	44.28 (16.46)	54.61 (16.17)	0.00 (0.00)	1.33 (0.49)
mRRR.n	56.13 (6.86)	84.58 (31.50)	2.28 (1.76)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
uGLM	15.22 (1.96)	23.72 (2.48)	0.07 (0.02)	93.12 (2.74)	0.00 (0.00)	5.53 (0.84)	27.09 (1.48)
				M% = 20			
GSeCURE	2.70 (1.36)	5.07 (2.22)	0.02 (0.02)	0.18 (0.14)	0.60 (1.08)	0.00 (0.00)	3.00 (0.00)
mRRR.r	59.77 (6.88)	122.32 (8.86)	2.53 (0.57)	33.26 (0.00)	65.31 (0.00)	0.00 (0.00)	1.00 (0.00)
mRRR.rs	59.78 (6.86)	122.40 (8.88)	2.53 (0.57)	33.26 (0.00)	65.31 (0.00)	0.00 (0.00)	1.00 (0.00)
mRRR.n	69.97 (5.59)	168.54 (44.35)	4.12 (1.96)	86.00 (23.28)	15.63 (21.55)	0.00 (0.00)	2.54 (0.66)
uGLM	19.83 (2.34)	31.27 (2.93)	0.10 (0.02)	92.23 (2.39)	0.00 (0.00)	7.14 (0.88)	26.72 (1.54)

Table 14: Simulation: results of Model II with Bernoulli responses at signal strength s = 4 .

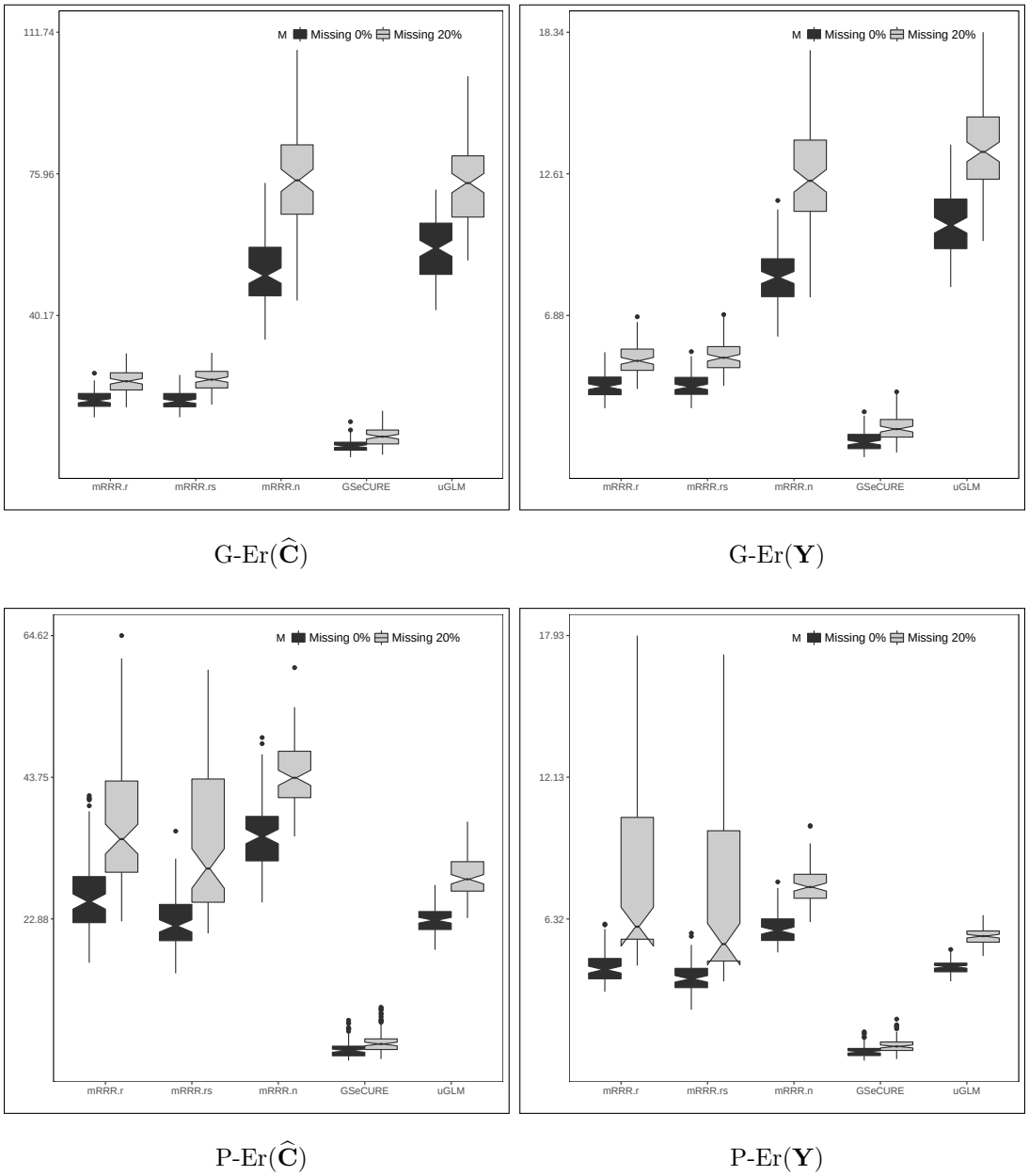
	$\text{Er}(\widehat{\mathbf{C}})$	$\text{Er}(\mathbf{Y})$	FPR	FNR	R%	r
			M% = 0			
GSeCURE	804.17 (116.01)	2074.80 (300.35)	3.37 (1.27)	12.78 (5.44)	0.00 (0.00)	3.00 (0.00)
mRRR.r	769.12 (26.96)	955.13 (89.64)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.rs	924.91 (40.01)	1404.81 (146.10)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.n	995.29 (15.02)	2202.81 (164.83)	100.00 (0.00)	0.00 (0.00)	7.35 (0.94)	17.81 (0.83)
uGLM	700.03 (53.30)	1698.70 (155.62)	92.27 (2.58)	0.00 (0.00)	7.32 (0.80)	27.13 (1.39)
			M% = 20			
GSeCURE	851.71 (93.51)	2212.68 (275.50)	2.71 (1.26)	15.78 (8.04)	0.00 (0.00)	3.00 (0.00)
mRRR.r	797.04 (26.78)	1063.74 (95.86)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.rs	915.63 (35.68)	1420.78 (152.93)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
mRRR.n	1019.74 (16.01)	2312.97 (174.42)	100.00 (0.00)	0.00 (0.00)	7.61 (1.55)	16.43 (1.46)
uGLM	804.86 (52.99)	1968.41 (177.74)	88.28 (2.80)	0.00 (0.00)	10.00 (1.20)	27.09 (1.44)

Table 15: Simulation: results of Model II with Poisson responses at signal strength s = 0.5 .

	$\text{Er}(\widehat{\mathbf{C}})$	$\text{Er}(\mathbf{Y})$	FPR	FNR	R%	r
			M% = 0			
GSeCURE	2.00 (1.34)	4.33 (3.38)	0.95 (0.97)	2.26 (2.63)	0.00 (0.00)	3.00 (0.00)
mRRR.r	13.89 (1.18)	14.61 (1.94)	93.03 (28.16)	14.97 (17.67)	0.00 (0.00)	2.71 (0.77)
mRRR.rs	11.57 (1.11)	11.49 (4.08)	84.71 (18.06)	14.97 (17.67)	0.00 (0.00)	2.71 (0.77)
mRRR.n	13.08 (0.48)	12.95 (0.99)	100.00 (0.00)	0.00 (0.00)	3.73 (1.97)	5.69 (1.35)
uGLM	6.06 (0.58)	10.38 (0.69)	91.50 (2.58)	0.00 (0.00)	8.69 (1.38)	26.71 (1.74)
			M% = 20			
GSeCURE	2.46 (1.49)	5.11 (3.56)	1.37 (1.27)	3.14 (2.84)	0.00 (0.00)	3.00 (0.00)
mRRR.r	14.60 (1.16)	18.26 (1.19)	86.77 (31.15)	21.77 (18.14)	0.00 (0.00)	2.35 (0.60)
mRRR.rs	12.47 (1.07)	14.92 (3.95)	77.75 (18.54)	21.77 (18.14)	0.00 (0.00)	2.35 (0.60)
mRRR.n	13.96 (0.49)	16.35 (1.27)	100.00 (0.00)	0.00 (0.00)	3.49 (2.10)	5.27 (1.08)
uGLM	7.82 (0.84)	13.54 (0.97)	89.75 (3.23)	0.00 (0.00)	11.31 (1.62)	26.68 (1.74)

Table 16: Simulation: results of Model II with Gaussian/Bernoulli responses at signal strength s = 2 .

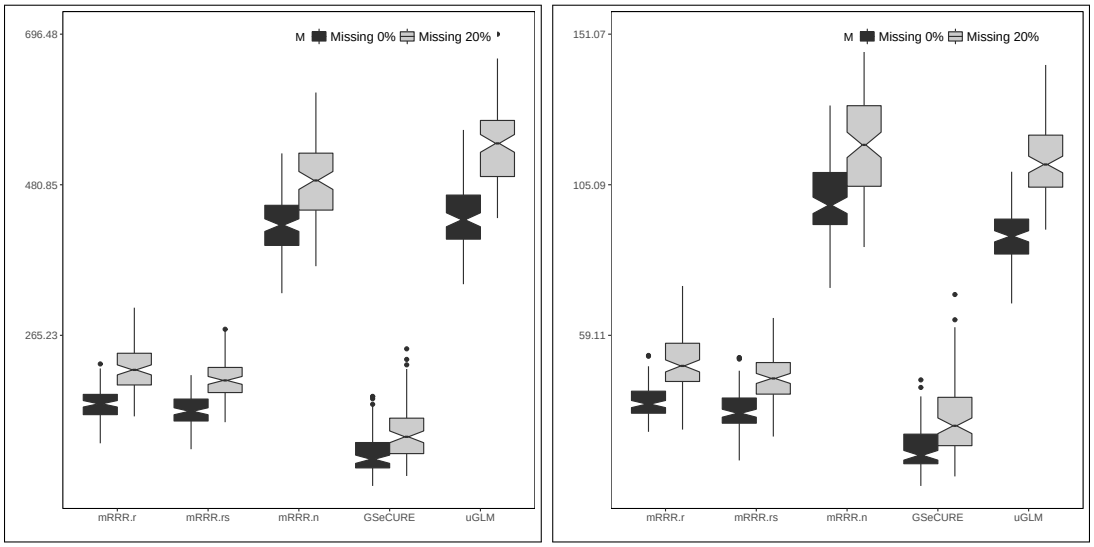
	Er($\widehat{\mathbf{C}}$)	Er($\mathbf{Y}$)	Er_g(Y)	Er_{ng}(Y)	Er(Φ)	FPR	FNR	R%	r
				M% = 0					
GSeCURE	47.10 (11.33)	118.36 (21.34)	22.54 (11.32)	215.02 (35.72)	7.26 (2.01)	0.43 (0.43)	0.53 (1.06)	0.00 (0.00)	3.00 (0.00)
GSeCURE.s	133.51 (23.40)	350.31 (79.45)	31.57 (12.32)	666.17 (156.76)	7.57 (1.97)	7.24 (0.94)	10.21 (9.45)	3.33 (9.49)	4.39 (0.78)
mRRR.r	262.01 (44.09)	533.31 (119.65)	492.56 (119.62)	585.00 (132.79)	69.52 (31.86)	33.37 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)
mRRR.rs	260.06 (47.21)	543.20 (112.21)	509.36 (104.65)	596.71 (126.46)	70.67 (31.77)	33.37 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)
mRRR.n	264.35 (13.69)	575.80 (98.01)	540.12 (124.83)	611.49 (73.59)	90.90 (32.88)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
uGLM	127.29 (10.52)	240.35 (20.76)	104.30 (10.63)	376.13 (38.46)	4.40 (1.35)	92.32 (2.26)	0.00 (0.00)	8.17 (0.90)	25.62 (1.59)
				M% = 20					
GSeCURE	57.66 (12.29)	141.75 (23.26)	35.18 (17.66)	249.31 (36.20)	7.62 (2.23)	0.42 (0.35)	1.41 (1.77)	0.00 (0.00)	3.00 (0.00)
GSeCURE.s	142.38 (21.28)	364.55 (72.78)	53.90 (18.72)	670.09 (147.57)	8.46 (2.47)	7.09 (1.01)	8.57 (7.92)	1.09 (0.73)	4.48 (0.75)
mRRR.r	296.26 (53.02)	572.45 (178.79)	540.30 (189.81)	627.87 (176.62)	40.51 (23.69)	45.43 (19.27)	53.55 (20.73)	0.00 (0.00)	1.38 (0.60)
mRRR.rs	308.64 (78.95)	596.12 (201.12)	575.16 (240.18)	650.20 (169.50)	45.46 (33.78)	45.64 (20.43)	52.93 (21.62)	0.00 (0.00)	1.38 (0.60)
mRRR.n	292.33 (10.33)	768.78 (82.99)	761.94 (97.50)	778.35 (73.45)	91.33 (20.41)	80.54 (20.77)	18.78 (19.34)	0.00 (0.00)	2.38 (0.62)
uGLM	152.00 (11.07)	294.52 (19.80)	143.44 (15.11)	445.17 (35.78)	4.77 (1.65)	89.26 (3.30)	0.00 (0.00)	10.41 (1.31)	25.59 (0.99)



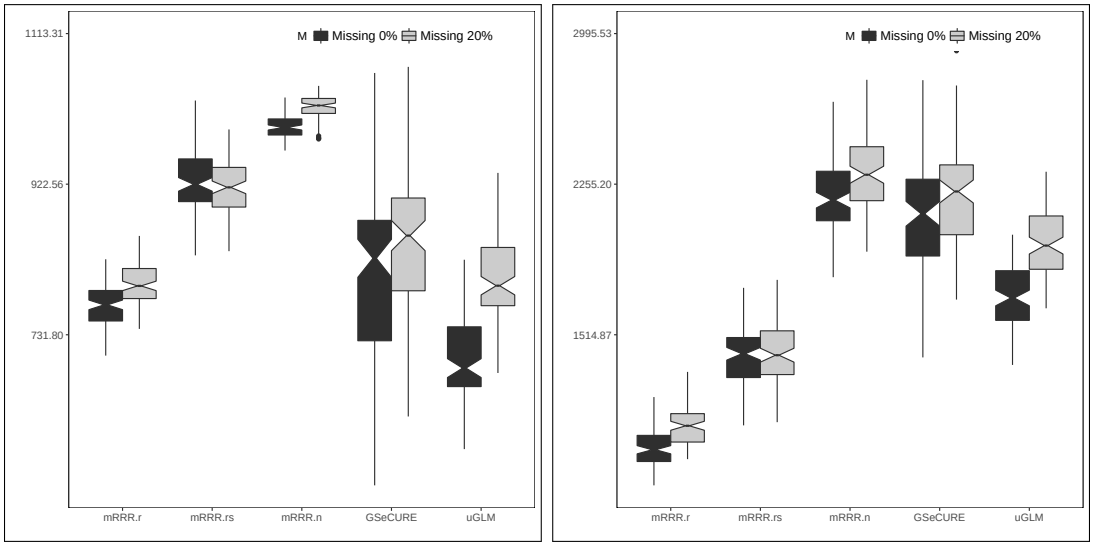
G-Er($\widehat{\mathbf{C}}$) G-Er($\mathbf{Y}$)

P-Er($\widehat{\mathbf{C}}$) P-Er($\mathbf{Y}$)

Figure 10: Simulation – Similar Responses: Notched boxplots of estimation accuracy Er($\widehat{\mathbf{C}}$) and prediction accuracy Er($\mathbf{Y}$) for either Gaussian (G) or Poisson (P) type of multivariate responses in Model I.

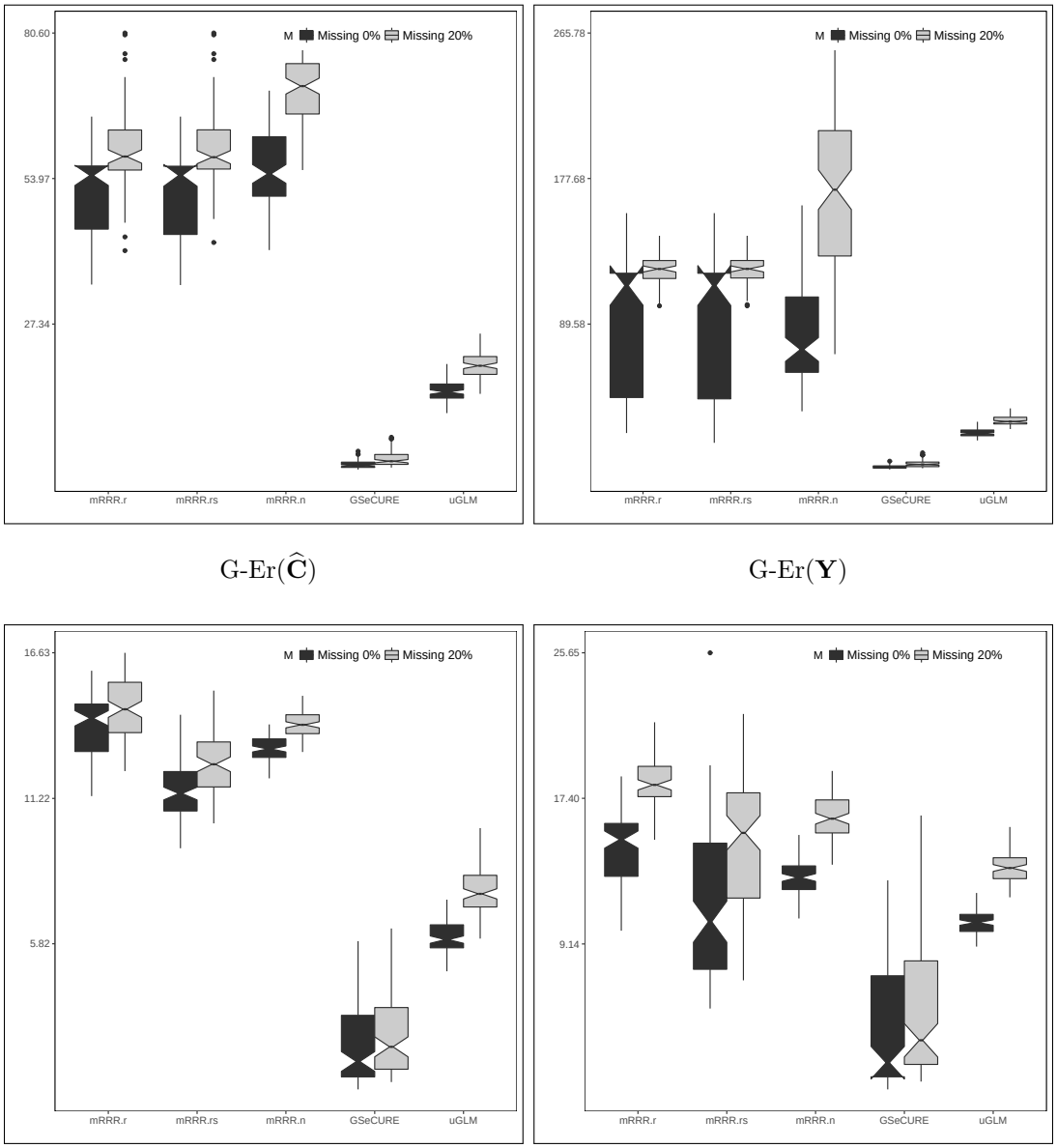


B-Er($\widehat{\mathbf{C}}$) B-Er($\mathbf{Y}$)



B-Er($\widehat{\mathbf{C}}$) B-Er($\mathbf{Y}$)

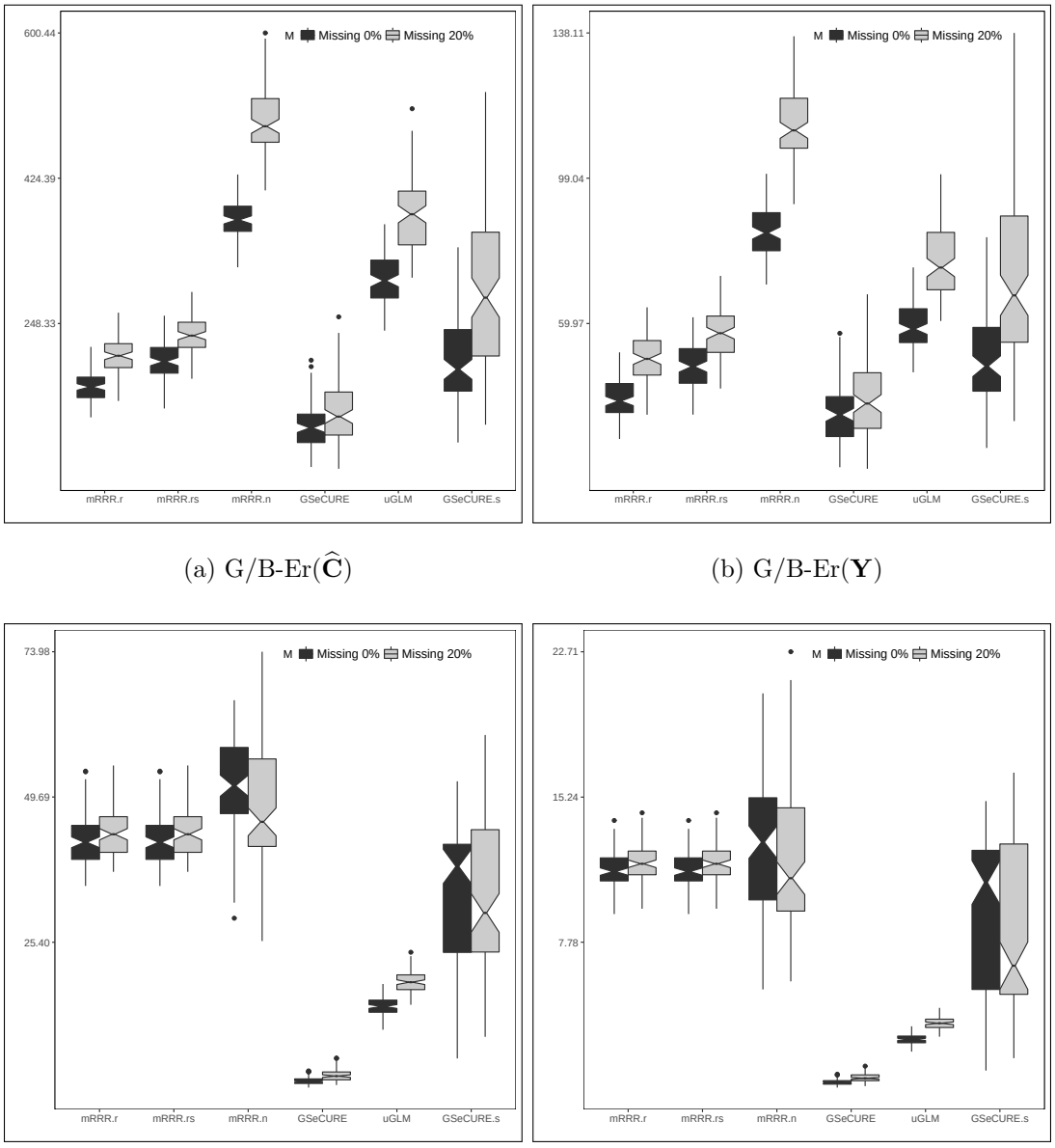
Figure 11: Simulation – Similar Responses: Notched boxplots of estimation accuracy Er($\widehat{\mathbf{C}}$) and prediction accuracy Er($\mathbf{Y}$) for Bernoulli (B) type of multivariate responses in Model I and Model II. Model I in Figure (a)–(b) and Model II in Figure (c)–(d).



G-Er($\widehat{\mathbf{C}}$) G-Er($\mathbf{Y}$)

P-Er($\widehat{\mathbf{C}}$) P-Er($\mathbf{Y}$)

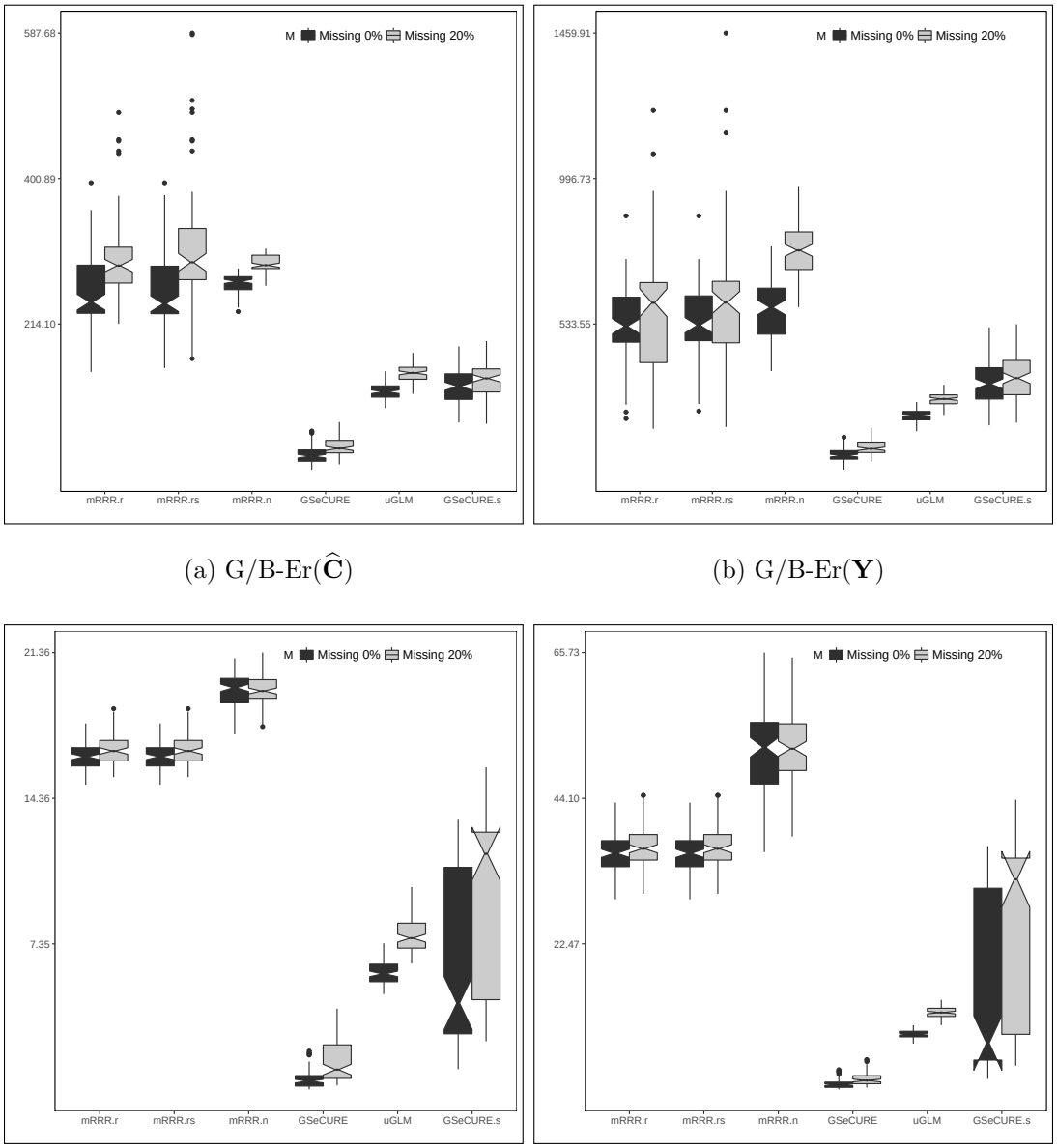
Figure 12: Simulation – Similar Responses: Notched boxplots of estimation accuracy Er($\widehat{\mathbf{C}}$) and prediction accuracy Er($\mathbf{Y}$) for either Gaussian (G) or Poisson (P) type of multivariate responses in Model II.



(a) G/B-Er($\widehat{\mathbf{C}}$) (b) G/B-Er($\mathbf{Y}$)

(c) G/P-Er($\widehat{\mathbf{C}}$) (d) G/P-Er($\mathbf{Y}$)

Figure 13: Simulation – Mixed Responses: Notched boxplots of estimation accuracy Er($\widehat{\mathbf{C}}$) and prediction accuracy Er($\mathbf{Y}$) for Gaussian/Bernoulli (G/B) and Gaussian/Poisson (G/P) mixed type response cases where figure (a)–(d) corresponds to Model I.



(a) G/B-Er($\widehat{\mathbf{C}}$) (b) G/B-Er($\mathbf{Y}$)

(c) G/P-Er($\widehat{\mathbf{C}}$) (d) G/P-Er($\mathbf{Y}$)

Figure 14: Simulation – Mixed Responses: Notched boxplots of estimation accuracy Er($\widehat{\mathbf{C}}$) and prediction accuracy Er($\mathbf{Y}$) for Gaussian/Bernoulli (G/B) and Gaussian/Poisson (G/P) mixed type response cases where figure (a)–(d) corresponds to Model II.

Table 17: Simulation: results of Model II with Gaussian/Poisson responses at signal strength s = 0.5 .

	Er($\widehat{\mathbf{C}}$)	Er($\mathbf{Y}$)	Er_g(Y)	Er_{ng}(Y)	Er(Φ)	FPR	FNR	R%	r
					M% = 0				
GSeCURE	0.85 (0.44)	1.68 (0.72)	1.25 (0.59)	2.04 (0.83)	0.06 (0.01)	0.46 (0.57)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)
GSeCURE.s	6.46 (4.05)	15.61 (12.62)	24.01 (25.25)	7.08 (3.53)	0.16 (0.16)	7.90 (2.08)	7.04 (7.92)	5.06 (10.38)	4.72 (1.52)
mRRR.r	16.38 (0.61)	35.98 (2.87)	37.20 (3.10)	34.84 (2.91)	0.13 (0.03)	33.37 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)
mRRR.rs	16.38 (0.61)	35.98 (2.87)	37.20 (3.10)	34.84 (2.91)	0.13 (0.03)	33.37 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)
mRRR.n	19.57 (0.77)	51.06 (5.75)	57.02 (4.92)	45.25 (6.74)	0.35 (0.08)	55.56 (17.61)	43.57 (17.97)	0.00 (0.00)	1.64 (0.48)
uGLM	6.01 (0.62)	9.10 (0.59)	6.48 (0.66)	11.71 (1.00)	0.09 (0.01)	94.31 (1.76)	0.00 (0.00)	8.34 (0.97)	25.33 (1.59)
					M% = 20				
GSeCURE	1.68 (1.02)	2.42 (1.01)	1.93 (0.94)	2.91 (1.18)	0.06 (0.01)	0.49 (0.44)	0.49 (0.91)	0.00 (0.00)	3.00 (0.00)
GSeCURE.s	9.44 (4.12)	24.37 (12.94)	37.99 (25.36)	10.34 (3.87)	0.25 (0.18)	7.16 (2.47)	18.95 (16.41)	0.39 (0.72)	3.89 (1.77)
mRRR.r	16.72 (0.68)	36.98 (3.00)	37.46 (3.11)	36.42 (2.98)	0.07 (0.01)	33.37 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)
mRRR.rs	16.72 (0.68)	36.98 (3.00)	37.46 (3.11)	36.42 (2.98)	0.07 (0.01)	33.37 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)
mRRR.n	19.55 (0.74)	51.61 (5.78)	56.48 (5.27)	46.61 (6.68)	0.18 (0.04)	54.91 (18.45)	44.22 (19.11)	0.00 (0.00)	1.63 (0.53)
uGLM	7.77 (0.82)	12.27 (0.89)	8.83 (0.99)	15.74 (1.71)	0.10 (0.01)	92.42 (2.42)	0.00 (0.00)	10.90 (1.33)	25.44 (1.68)

Chapter 4

R package `secure`

We have developed software implementing `SeCURE` in `R` package `secure`, which is available on the Comprehensive R Archive Network (CRAN). With the aim to understand its usability, here we have demonstrated the functions and their features.

4.1 SeCURE Implementation

We have implemented sequential co-sparse factor regression in the `secure.path` function of the package `secure`. The approach is described in Chapter 2 of the thesis. With the aim to enhance computation speed, we have implemented the computationally intensive part of the function using `Rcpp` and `RcppArmadillo`.

The function requires the user to provide the multivariate response matrix $\mathbf{Y}$ and predictor matrix $\mathbf{X}$. As a sub-case of SeCURE when predictor $\mathbf{X}$ is not specified, the function assumes it to be identity matrix and performs biclustering [Lee et al., 2010] on observed response $\mathbf{Y}$. The desired rank or maximum number of latent factors can be specified using variable `nrank`. In a unit step, `secure` performs *constrained unit-rank estimation*(CURE) in which the solution path is obtained corresponding to maximum

`nlambda` λ.

The function takes in a set of `control` values which we set using the `secure.control` function in the package. Controlling parameters passed to the function include; `mu`, penalty parameter used in enforcing orthogonality; `nu`, ncremental rate of `mu`; `MMerr`, tolerance in the *majorization maximization* (MM) algorithm for computing initial values when missing values occur; `MMiter`, maximum number iterations in the MM algorithm; `outTol`, tolerance of convergence of outer loop in CURE; `outMaxIter`, maximum number of outer loop iterations in CURE; `inMaxIter`, maximum number of inner loop iteration in CURE; `inTol`, tolerance value required for convergence of inner loop in CURE; `lamMaxFac`, a multiplier of calculated λ_{max}; `lamMinFac`, a multiplier for determining λ_{min} as a fraction of λ_{max}; `gamma0`, power parameter in the adaptive weights; `elnetAlpha`, elastic net penalty parameter; `spU`, maximum proportion of nonzero elements in each column of estimated $\mathbf{U}$; and `spV`, maximum proportion of nonzero elements in each column of $\mathbf{V}$.

We call the function using following command:

```
secure.path(Y, X = NULL, nrank = 3, nlambda = 100, U0 = NULL, V0 = NULL,
            D0 = NULL, orthXU = FALSE, orthV = FALSE, keepPath = TRUE,
            control = list(),  ic = c("GIC", "BICP", "AIC")[1])
```

The function performs model selection using various information criteria `ic` given by Generalize Information Criteria (GIC), Bayesian Information Criteria (BIC) and Akaike

Information Criteria (AIC). Sparsity is induced using the adaptive elastic net penalty, defined in equation, (2.10), and ($\mathbf{U0}, \mathbf{D0}, \mathbf{V0}$) provides input parameters for such construction. The orthogonality constraint is optional which can be chosen using `orthXU` and `orthV`. It should be noted here that the `secure.path` function does not generate the entire solution path, rather it terminates at a point where the required proportion (specified by `spU` and `spV`) of nonzero elements in estimates of $\mathbf{U}$ and $\mathbf{V}$ are met.

The package offers function `secure.sim` which generates multivariate responses and predictor data as suited for our function `secure.path`. The user needs to specify following parameters: `U`, specified value of U; `D`, specified value of D; `V`, specified value of V; `n`, sample size; `snr`, signal to noise ratio; `Xsigma`, covariance matrix for generating sample of X; and `rho`, parameter defining correlated error. The function can be called using following command:

```
secure.sim(U, D, V, n, snr, Xsigma, rho = 0)
```

Another additional function offered by `secure` is `rrr.fit` for fitting reduced rank regression model of rank r. User need to input response matrix $\mathbf{Y}$, predictor matrix $\mathbf{X}$ and rank r. The function can be called using following command:

```
rrr.fit(Y, X, nrank = nrank)
```

4.2 Example

Here we demonstrate a single simulation for model fitting using `secure.path`. Before we generate data, we first define model parameters.

```
# Simulate data from a sparse factor regression model

p <- 100; q <- 100; n <- 200

xrho <- 0.5; nlambda <- 100

nrank <- 3

U <- matrix(0,ncol=nrank ,nrow=p);  V <- matrix(0,ncol=nrank ,nrow=q)

U[,1]<-c(sample(c(1,-1),8,replace=TRUE),rep(0,p-8))

U[,2]<-c(rep(0,5),sample(c(1,-1),9,replace=TRUE),rep(0,p-14))

U[,3]<-c(rep(0,11),sample(c(1,-1),9,replace=TRUE),rep(0,p-20))

V[,1]<-c(sample(c(1,-1),5,replace=TRUE)*runif(5,0.3,1),rep(0,q-5))

V[,2]<-c(rep(0,5),sample(c(1,-1),5,replace=TRUE)*runif(5,0.3,1),rep(0,q-10))

V[,3]<-c(rep(0,10),sample(c(1,-1),5,replace=TRUE)*runif(5,0.3,1),rep(0,q-15))

U[,1:3]<- apply(U[,1:3],2,function(x)x/sqrt(sum(x^2)))

V[,1:3]<- apply(V[,1:3],2,function(x)x/sqrt(sum(x^2)))

D <- diag(c(20,15,10))

C <- U%*%D%*%t(V)

Xsigma <- xrho^abs(outer(1:p, 1:p,FUN="-"))
```

We simulate data of sample size $n = 200$ with q=100 responses and p=100 predictors.

With simulated model parameters in hand, we generate n samples of response and predictors using

```
sim.sample <- secure.sim(U,D,V,n,snr = 0.25,Xsigma,rho=0.3)
Y <- sim.sample$Y;
X <- sim.sample$X;
```

Now, we are ready for model fitting. Before we fit the model, we define maximum `rank.ini` and `control`. Set maximum rank to be equal to 4, and specify sparsity proportion to be 0.25 in both **U** and **V**.

```
rank.ini <- 4  # Set maximum rank to be 4.
# Set largest model to about 25% sparsity
# See secure.control for setting other parameters
control <- secure.control(spU=0.25, spV=0.25)
```

The `secure.path` function provides flexibility to fit the model with or without orthogonality constraint. To fit model in such scenarios we have:

```
# Fit secure without orthogonality
fit.orthF <- secure.path(Y,X,nrank=rank.ini,nlambda = nlambda,
control=control)

# Fit secure with orthogonality if desired. It takes longer time.
fit.orthT <- secure.path(Y,X,nrank=rank.ini,nlambda = nlambda,
orthXU=TRUE,orthV=TRUE,control=control)
```

With model estimates in hand, we check orthogonality in both the case using:

```
crossprod(X%*%fit.orthF$U)/n
crossprod(X%*%fit.orthT$U)/n
```

To demonstrate capability of `secure.path` in handling missing data in multivariate responses, we drop 15% entries in response matrix $\mathbf{Y}$, and save in new matrix $\mathbf{Ym}$.

```
# 15% missing case
miss <- 0.15
t.ind <- sample.int(n*q, size = miss*n*q)
y <- as.vector(Y); y[t.ind] <- NA;  Ym <- matrix(y,n,q)
```

To fit the model with or without orthogonality constraint in such scenarios, the functions are given by:

```
fit.orthF.miss <- secure.path(Ym, X, nrank = rank.ini, nlambda = nlambda,
control = control)
fit.orthT.miss <- secure.path(Ym, X, nrank = rank.ini, nlambda = nlambda,
orthXU=TRUE,orthV=TRUE, control = control)
```

Let `fit` be object storing model output. It would be interesting to see how latent factor loadings estimated $\mathbf{VD}$.

```
VD <- data.frame(fit$V %*% fit$D)
VD$Index <- 1:nrow(fit$V)
require(ggplot2);  require(reshape2)
df <- melt(VD, id=c("Index"))
levels(df$variable) <- c('First','Second','Third')
ggplot(df, aes(Index,value)) + geom_line()  + facet_grid(. ~  variable)+
theme_bw()+ylab("Factor loading")+ geom_hline(aes(yintercept=0))
```

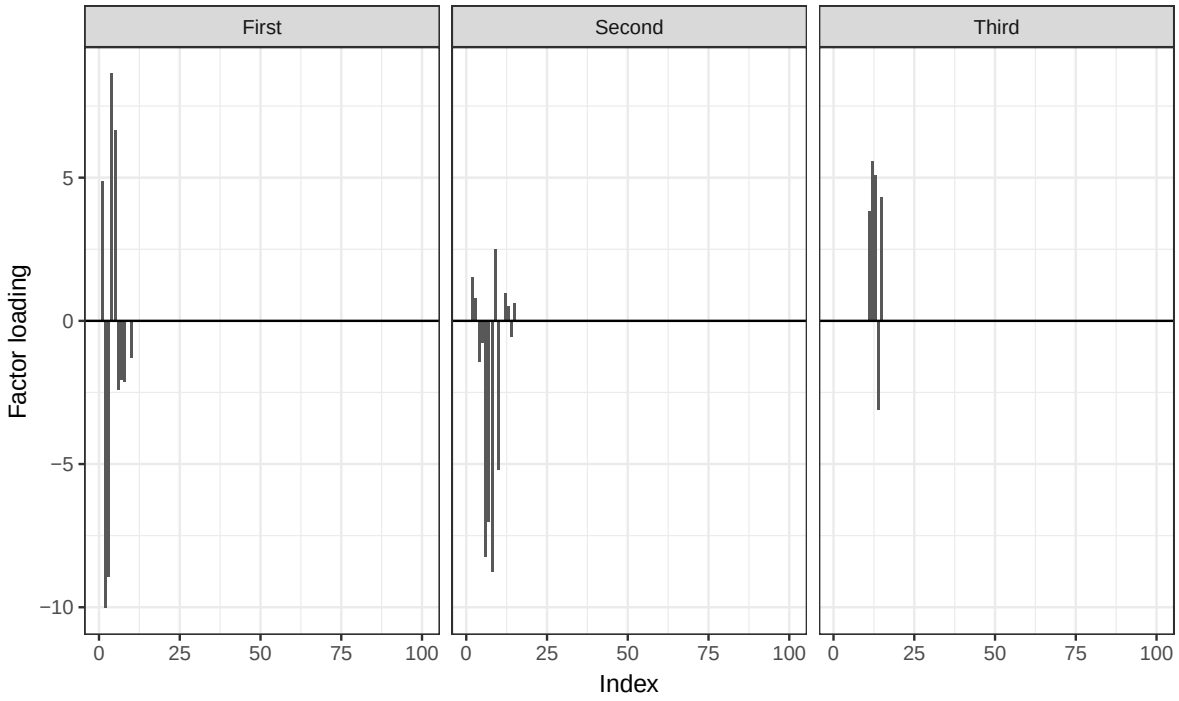


Figure 15: An example of SeCURE: Loadings of first, second and third latent factors.

Chapter 5

Discussion

We have developed a sequential estimation procedure to recover an appealing co-sparse and low-rank factor structure in the multivariate models with outcomes either of continuous, count, binary types or may be of mixed types. A subclass of the model is the multivariate linear regression (MLR), and we have proposed SeCURE for model estimations. To deal with the modeling problem in the generalized setting where multivariate responses are either similar/non-Gaussian types or may be mixed types, we have proposed the greedy algorithm GSeCURE. The models allow each latent factor to be constructed from a subset of predictors, and may potentially influence only a subset of responses. In SeCURE, we show that the orthogonality constraints that are required in joint estimation can be avoided in sequential estimation, and without them, computational efficiency and estimation consistency can both be achieved. Motivated by SeCURE, in GSeCURE such a constraint is not enforced. Because of complicated objective function in a unit-step of GSeCURE, to ensure its monotone descending property the estimation proceeds via a surrogate function suitably defined using scaled predictors. Moreover, both the procedures can efficiently handle missing entries in multivariate responses.

There are several directions for future research. A priority is to perform a comprehensive non-asymptotic analysis to fully understand the finite sample behaviors of the SeCURE estimators. In our sequential procedure, the residual matrix is used in each

step to deflate the information found from the previous steps; it would be interesting to investigate other deflation methods. Also, none of the proposed procedure can handle missing entries in the predictor matrix, thus the problem requires our attention.

Many multivariate analysis (MA) methods like canonical correlation analysis (CCA), principal component analysis (PCA), partial least square (PLS), RRR and linear discriminant analysis (LDA) determines the relation between centered multivariate response and centered multivariate predictor. They do so by finding subspace of both predictor and response in their respective setting. MA technique like PLS, CCA, PCA, RRR can be formulated as generalized eigen decomposition (GED) problem; see technical report by [Borga et al., 1997]. To efficiently solve the GED for high-dimensional data, Borga proposed to use a gradient-descent algorithm on a Rayleigh quotient. In the high dimensional scenario, we have MA technique sparse version as well, like sparse PCA, sparse PLS, sparse CCA etc. The challenge with GED lies in the estimation of sparse eigen vector in rank deficient cases which occurs quite often in high dimension. It is thus interesting to develop a unified sequential estimation procedure that performs co-sparse generalized eigen-decomposition [Bahadori et al., 2016].

[De la Torre, 2012] proposed a least square framework referred as *weighted reduced rank regression* (WRRR) which brings a wider class of multivariate analysis problem into a unified model setup. In some cases, MA problem formulation is quite straight forward while in other cases equivalent GED setup of MA technique is used to establish its connection with WRRR. The least square formulation yields efficient numerical schemes to solve MA techniques, and also it opens up the possibility extending the model in the regularized setting. Thus, to obtain co-sparse generalized eigen-decomposition in MA problems, it would be interesting to explore the avenues of sequential estimation in regularized WRRR like SeCURE.

We have noted that sufficient signal to noise ratio (SNR) is required for efficient recovery of the unit rank matrix in a unit step of GSeCURE. The sequential approach through the unit-rank estimation step provides us flexibility in inducing co-sparsity in singular vectors which are important for performing variable selection in the generalized setting. Hence, we look for a procedure with the unit-rank estimation step which can handle problems with low SNR in the generalized setting. A possible approach is to perform estimation of the unit-rank matrix in a unit step via exclusive extraction strategy (EEA), as proposed in Chen et al. [2012] for MLR.

Like SeCURE, a unit step in the EEA approach also estimates unit-rank matrix expressed as the outer product of singular vectors of the coefficient matrix. The two procedure differs in the type of recovered SVD decomposition of the coefficient matrix. The EEA approach is an iterative procedure that proceeds via unit-rank matrix estimation step which provides us the flexibility to induce co-sparsity in singular vectors. In a unit step, the sparse unit-rank matrix estimated is the one which best predict exclusive layer using observed predictors. The exclusive layer accounts for signal left obtained after subtracting estimated signal in remaining layers from the given signal, i.e., observed responses. It would be interesting to apply the strategy in the generalized setting of the multivariate problems to recover the low-rank and sparse coefficient matrix.

In the generalized setting, when we will be performing the model estimation via EEA approach, rather than obtaining the exclusive layer, in a unit step the estimated signal in remaining layers can be incorporated via the offset term matrix. We will face similar challenge because of complicated likelihood structure as in a unit-step of GSeCURE. Again, the analysis through surrogate of the objective function based on properly scaled predictors will lead us to a computationally efficient algorithm for the model parameters estimation.

Appendix A

Sequential Co-Sparse Factor Regression

A.1 Linear Constrained Elastic Regression

Consider a generic form of the linear constrained adaptive elastic net regression,

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{J(\boldsymbol{\beta})\equiv\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda_1\sum_{j=1}^{p}w_j|\beta_j|+\frac{\lambda_2}{2}\sum_{j=1}^{p}\beta_j^2\right\},\qquad\text{s.t. }\mathbf{A}\boldsymbol{\beta}=\mathbf{b},\quad\text{(A.1)}$$

where $\mathbf{y}\in\mathbb{R}^n$, $\mathbf{X}\in\mathbb{R}^{n\times p}$, $\mathbf{A}=(\mathbf{a}_1,\ldots,\mathbf{a}_p)\in\mathbb{R}^{h\times p}$, $\mathbf{b}\in\mathbb{R}^{h\times 1}$, $\mathbf{w}=[w_1,...,w_p]^{\mathrm{T}}$ are some predetermined weights, and λ_1 and λ_2 are tuning parameters.

This is an equality-constrained convex optimization problem. The augmented Lagrangian function is

$$L_\mu(\boldsymbol{\beta},\mathbf{c})=J(\boldsymbol{\beta})+\mathbf{c}^{\mathrm{T}}(\mathbf{A}\boldsymbol{\beta}-\mathbf{b})+\frac{\mu}{2}\|\mathbf{A}\boldsymbol{\beta}-\mathbf{b}\|_2^2,$$

where $\mathbf{c}$ is the Lagrange multiplier, and $\mu>0$ is a penalty parameter. The Bregman iterative method or the method of multipliers [Bregman, 1967, Boyd et al., 2011, Lin

et al., 2014a] yield the following iterative steps to solve the problem,

$$\left.\begin{aligned}\boldsymbol{\beta}^{(s+1)} &= \arg\min_{\boldsymbol{\beta}} L_\mu(\boldsymbol{\beta}, \mathbf{c}^{(s)}) \\ &= \arg\min_{\boldsymbol{\beta}} \{J(\boldsymbol{\beta}) + \frac{\mu}{2}\|\mathbf{A}\boldsymbol{\beta} - \mathbf{b} - \frac{\mathbf{c}^{(s)}}{\mu}\|_2^2\}, \\ \mathbf{c}^{(s+1)} &= \mathbf{c}^{(s)} - (\mathbf{A}\boldsymbol{\beta}^{(s+1)} - \mathbf{b})\mu.\end{aligned}\right\} \tag{A.2}$$

The method converges under very general conditions. As the iteration proceeds, the residual $\mathbf{A}\boldsymbol{\beta}^{s+1} - b$ converges to zero, yielding optimality.

Following (A.2), the key is to minimize

$$\boldsymbol{\beta}^{(s+1)} = \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^{p} w_j|\beta_j| + \frac{\lambda_2}{2}\|\boldsymbol{\beta}\|_2^2 + \frac{\mu^{(s)}}{2}\|\mathbf{A}\boldsymbol{\beta} - \mathbf{b} - \frac{\mathbf{c}^{(s)}}{\mu^{(s)}}\|_2^2 \right\}. \tag{A.3}$$

Here the penalty parameter $\mu^{(s)}$ can be updated along the iterations; letting $\mu \to \infty$ or increase with small increments can in general improve the speed of convergence [Goldstein and Osher, 2009]. The above problem can be efficiently minimized by a coordinate descent algorithm. Suppose all the β_ks are fixed except β_j, and denote $\mathbf{r}_j = \mathbf{y} - \sum_{k\neq j} \mathbf{x}_k\beta_k$. The objective function with respect to β_j becomes

$$\frac{1}{2}\|\mathbf{r}_j - \mathbf{x}_j\beta_j\|_2^2 + \lambda_1 w_j|\beta_j| + \frac{\mu^{(s)}}{2}\|\mathbf{a}_j\beta_j + \sum_{k\neq j}\mathbf{a}_k\beta_k - \mathbf{b} - \frac{\mathbf{c}^{(s)}}{\mu^{(s)}}\|_2^2 + \frac{\lambda_2}{2}\beta_j^2 + \text{const.}$$

Then it can be easily verified that

$$\widehat{\beta}_j = \frac{\mathcal{S}\left((\mathbf{y} - \sum_{i\neq j}\beta_i\mathbf{x}_i)^{\mathrm{T}}\mathbf{x}_j + \mu^{(s)}\{(\frac{\mathbf{c}^{(s)}}{\mu^{(s)}} + \mathbf{b})^{\mathrm{T}}\mathbf{a}_j - \sum_{i\neq j}\beta_i\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_j\}, \lambda w_j\right)}{\lambda_2 + \mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j + \mu^{(s)}\mathbf{a}_j^{\mathrm{T}}\mathbf{a}_j}, \tag{A.4}$$

where $\mathcal{S}(m, \lambda) = \text{sign}(m)(|m| - \lambda)_+$ is the soft-thresholding operator. (A.3) can then be solved by iteratively updating each β_j, $j = 1, \ldots, p$, by (A.4) until convergence. Our proposed algorithm is presented in Algorithm 6.

Algorithm 6 Bregman Coordinate Descent Algorithm (BCDA)

Initialization: $s = 0$, $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^p$, $\mathbf{c}^{(0)} = \mathbf{0}$, $\mu^{(0)} = 1$, and $\rho \geq 1$.
repeat
(1) Use coordinate descent to obtain $\boldsymbol{\beta}^{(s+1)}$, by iteratively updating β_js using (A.4) until convergence.
(2) $\mathbf{c}^{(s+1)} = \mathbf{c}^{(s)} - (\mathbf{A}\boldsymbol{\beta}^{(s+1)} - \mathbf{b})\mu^{(s)}$.
(3) $\mu^{(s+1)} = \mu^{(s)}\rho$.
$s \leftarrow s + 1$.
until convergence, i.e., $\|\boldsymbol{\beta}^{(s+1)} - \boldsymbol{\beta}^{(s)}\|/\|\boldsymbol{\beta}^{(s)}\| < \epsilon$.
return $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$.

We remark that the problem in (A.1) can also be solved using an alternating direction method of multipliers [Boyd et al., 2011], by reformulating the problem as

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \{f(\boldsymbol{\beta}) + g(\boldsymbol{\gamma})\}, \qquad \text{s.t. } \mathbf{A}\boldsymbol{\beta} = \mathbf{b}, \boldsymbol{\beta} = \boldsymbol{\gamma}, \tag{A.5}$$

where $f(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, and $g(\boldsymbol{\gamma}) = \lambda_1 \sum_{j=1}^p w_j|\gamma_j| + \frac{\lambda_2}{2}\sum_{j=1}^p \gamma_j^2$. We omit the details but refer the readers to Boyd et al. [2011] for details.

Appendix B

A Greedy Algorithm for Generalized Sparse and Low-rank Recovery

B.1 Proof of Theorem 3.1

Proof: We show that Algorithm 5 admits desirable convergence properties. Let the parameter in any tth iteration be given by $Ł^{(t)} = (d^{(t)}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)})$. Then unit-rank matrix $\mathbf{C}^{(t)} = d^{(t)}\mathbf{u}^{(t)}\mathbf{v}^{(t)\mathrm{T}}$. For the ease of presentation, we will denote the natural parameter matrix for $\mathbf{C}^{(t)}$ and $\boldsymbol{\beta}^{(t)}$ by

$$\Theta(\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}) = \Theta(d^{(t)}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}, \boldsymbol{\beta}^{(t)}) = \mathbf{O} + \mathbf{X}\mathbf{C}^{(t)} + \mathbf{Z}\boldsymbol{\beta}^{(t)}. \tag{B.1}$$

In terms of $Ł^{(t)}$, the surrogate function for the V-step, $G^{(v)}(\check{\mathbf{v}}; \mathbf{u}_a, \mathbf{C}, \boldsymbol{\beta}, \Phi)$ defined in (3.25), is given by $G^{(v)}(\check{\mathbf{v}}; d^{(t)}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}) = G^{(v)}(\check{\mathbf{v}}; Ł^{(t)})$ where $\check{\mathbf{v}} = d\mathbf{v}$, $\mathbf{u}_a = \mathbf{u}^{(t)}$, $\mathbf{C} = \mathbf{C}^{(t)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$ and $\Phi = \Phi^{(t)}$. Then, $\check{\mathbf{v}}^{(t)} = \mathbf{v}^{(t)}d^{(t)}$. Let $\check{\mathbf{v}}^{(t+1)} = \tilde{d}^{(t+1)}\mathbf{v}^{(t+1)}$ minimize surrogate function $G^{(v)}(\check{\mathbf{v}}; Ł^{(t)})$. As shown by She [2009], for any $\delta_1 \in \mathbb{R}^q$,

$$G^{(v)}(\check{\mathbf{v}}^{(t+1)} + \delta_1; Ł^{(t)}) - G^{(v)}(\check{\mathbf{v}}^{(t+1)}; Ł^{(t)}) \geq \frac{\eta_1}{2}(1 + (1-\alpha)\lambda)\|\delta_1\|_2^2\|\mathbf{u}^{(t)}\|_2^2,$$

where $\eta_1 = \max(0, 1 - L_1)$ with constant $L_1 \in [0, 1]$ is fixed for a threshold rule $\mathbf{S}$ (see definition 2.1 She [2012a]) defined corresponding to penalty function $\rho(\cdot)$ and $1 - L_1$

bounds below $d\mathbf{S}^{-1}(u;\lambda)/du$. The constant L_1 can be explicitly computed for several commonly used penalty forms. Define $\widetilde{\mathbf{C}}^{(t+1)} = \tilde{d}^{(t+1)}\mathbf{u}^{(t)}\mathbf{v}^{(t+1)\mathrm{T}}$. Choosing $\delta_1 = \check{\mathbf{v}}^{(t)} - \check{\mathbf{v}}^{(t+1)}$, we get

$$G^{(v)}(\check{\mathbf{v}}^{(t)};\mathrm{Ł}^{(t)}) - G^{(v)}(\check{\mathbf{v}}^{(t+1)};\mathrm{Ł}^{(t)}) \geq \frac{\eta_1}{2}(1+(1-\alpha)\lambda)\|\widetilde{\mathbf{C}}^{(t+1)} - \mathbf{C}^{(t)}\|_2^2. \tag{B.2}$$

It should be noted that $G^{(v)}(\check{\mathbf{v}}^{(t)};\mathrm{Ł}^{(t)}) = G(\mathbf{C}^{(t)};\mathbf{C}^{(t)},\boldsymbol{\beta}^{(t)},\Phi^{(t)}) = F(\mathbf{C}^{(t)},\boldsymbol{\beta}^{(t)},\Phi^{(t)})$ and $G^{(v)}(\check{\mathbf{v}}^{(t+1)};\mathrm{Ł}^{(t)}) = G(\tilde{\mathbf{C}}^{(t+1)};\mathbf{C}^{(t)},\boldsymbol{\beta}^{(t)},\Phi^{(t)})$. Using the obtained result in (B.2) and part of the proof in Theorem 2.1 of She [2011], we can say that

$$\begin{aligned} F(\mathbf{C}^{(t)},\boldsymbol{\beta}^{(t)},\Phi^{(t)}) - F(\tilde{\mathbf{C}}^{(t+1)},\boldsymbol{\beta}^{(t)},\Phi^{(t)}) &\geq \frac{1+\eta_1(1+(1-\alpha)\lambda)}{2}\|\tilde{\mathbf{C}}^{(t+1)} - \mathbf{C}^{(t)}\|_F^2 \\ &\quad - \frac{1}{2}\langle \mathbf{B}''(\Theta(\widetilde{\xi}_c^{(t+1)},\boldsymbol{\beta}^{(t)})), (\mathbf{X}\tilde{\mathbf{C}}^{(t+1)} - \mathbf{X}\mathbf{C}^{(t)}) \circ (\mathbf{X}\tilde{\mathbf{C}}^{(t+1)} - \mathbf{X}\mathbf{C}^{(t)})\Phi^{-1}\rangle \\ &\geq \frac{1+\eta_1(1+(1-\alpha)\lambda)}{2}\|\tilde{\mathbf{C}}^{(t+1)} - \mathbf{C}^{(t)}\|_F^2 - \frac{\gamma_1}{2}\|\tilde{\mathbf{C}}^{(t+1)} - \mathbf{C}^{(t)}\|_F^2 \end{aligned}$$

where $\xi_v^{(t+1)} \in \{a\check{v}^{(t)} + (1-a)\check{v}^{(t+1)}; 0 < a < 1\}$ and $\widetilde{\xi}_c^{(t+1)} = \mathbf{u}^{(t)}\xi_v^{(t+1)\mathrm{T}}$. About γ_1, we have

$$\gamma_1 = \max_{1\leq k\leq q}\sup_{a\in(0,1)} \|\mathbf{I}(\Theta_{\cdot k}(\widetilde{\xi}_c^{(t+1)},\boldsymbol{\beta}^{(t)}),\phi_k^{(t)})\|_2, \tag{B.3}$$

with

$$\mathbf{I}(\Theta_{\cdot k}(\widetilde{\xi}_c^{(t+1)},\boldsymbol{\beta}^{(t)})) = \mathbf{X}^{\mathrm{T}}\zeta(\Theta_{\cdot k}(\widetilde{\xi}_c^{(t+1)},\boldsymbol{\beta}^{(t)}),\phi_k^{(t)})\mathbf{X}, \tag{B.4}$$

and

$$\zeta(\Theta_{\cdot k}(\widetilde{\xi}_c^{(t+1)},\boldsymbol{\beta}^{(t)}),\phi_k^{(t)}) = \mathrm{diag}[\mathbf{B}''_{\cdot k}(\Theta_{\cdot k}(\widetilde{\xi}_c^{(t+1)},\boldsymbol{\beta}^{(t)}))]/\phi_k^{(t)}. \tag{B.5}$$

Hence, we obtained that in V-step

$$F(\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}) - F(\tilde{\mathbf{C}}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}) \geq \frac{\kappa_1}{2} \|\tilde{\mathbf{C}}^{(t+1)} - \mathbf{C}^{(t)}\|_F^2, \tag{B.6}$$

where $\kappa_1 = 1 + \eta_1(1 + (1-\alpha)\lambda) - \gamma_1$.

As long as $\kappa_1 \geq 0$, the monotone descending property of the **C**-step is guaranteed. The upper bound of γ_1 can be determined for several distributions. For example, for Gaussian responses, $b_k''(x) = 1$ for any x, and $a_k(\phi_k) = \sigma_k^2$, where σ_k^2 is the variance parameter; therefore, $\gamma_1 \leq \|\mathbf{X}\|_2^2 / \min(\sigma_k^2)$, where the minimum is taken over the indices of all the Gaussian responses. It then suffices to scale $\mathbf{X}$ by some $\kappa_1^* \geq \|\mathbf{X}\|_2 / \min(\sigma_k)$, so that after scaling, $\gamma_1 \leq 1$ and $\kappa_1 \geq 1 - L_1$. In practice, σ_k can be taken as some initial estimator from GLM. For binary responses, $b_k''(x) = e^x/(1+e^x)^2 \leq 1/4$ for any x and $a_k(\phi_k) = 1$, so that $\gamma_1 \leq \|\mathbf{X}\|_2^2/4$. Then it suffices to have $\kappa_1^* = \|\mathbf{X}\|_2/2$. For Poisson responses, however, $b_k''(x) = e^x$ does not have a universal bound; nevertheless, in practice we could empirically choose a large enough κ_1^* to ensure the descending of the **C**-step. In the mixed response setting, κ_1^* can be chosen as the maximum of such quantities for different distributions.

In $(t+1)$th iteration, the parameter after the V-step will be $\widetilde{\mathrm{L}}^{(t+1)} = (\tilde{d}^{(t+1)}, \mathbf{u}^{(t)}, \mathbf{v}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)})$. In terms of updated parameters, the surrogate function of the U-step, $G^{(u)}(\check{\mathbf{u}}; \mathbf{v}_a, \mathbf{C}, \boldsymbol{\beta}, \Phi)$ defined in (3.26), is given by $G^{(u)}(\check{\mathbf{u}}; \widetilde{\mathrm{L}}^{(t+1)})$ where $\check{\mathbf{u}} = d\mathbf{u}$, $\mathbf{v}_a = \mathbf{v}^{(t+1)}$, $\mathbf{C} = \widetilde{\mathbf{C}}^{(t+1)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$ and $\Phi = \Phi^{(t)}$. Then $\check{\mathbf{u}}^{(t)} = d^{(t)}\mathbf{u}^{(t)}$. Let $\check{\mathbf{u}}^{(t+1)}$ minimize surrogate function $G^{(u)}(\check{\mathbf{u}}; \widetilde{\mathrm{L}}^{(t+1)})$. Define $\mathbf{C}^{(t+1)} = \check{\mathbf{u}}^{(t+1)}\mathbf{v}^{(t+1)\mathrm{T}}$. As shown by

She [2009], for any $\delta_2 \in \mathbb{R}^p$,

$$G^{(u)}(\check{\mathbf{u}}^{(t+1)} + \delta_2; \widetilde{\mathrm{L}}^{(t+1)}) - G^{(u)}(\check{\mathbf{u}}^{(t+1)}; \widetilde{\mathrm{L}}^{(t+1)}) \geq \frac{\eta_1}{2}(1 + (1-\alpha)\lambda)\|\delta_2\|_2^2 \|\mathbf{v}^{(t+1)}\|_2^2.$$

On setting $\delta_2 = \tilde{d}^{(t+1)}\mathbf{u}^{(t)} - d^{(t+1)}\mathbf{u}^{(t+1)}$ we have

$$G^{(u)}(\tilde{d}^{(t+1)}\mathbf{u}^{(t)}; \widetilde{\mathrm{L}}^{(t+1)}) - G^{(u)}(\check{\mathbf{u}}^{(t+1)}; \widetilde{\mathrm{L}}^{(t+1)}) \geq \frac{\eta_1}{2}(1 + (1-\alpha)\lambda)\|\widetilde{\mathbf{C}}^{(t+1)} - \mathbf{C}^{(t+1)}\|_2^2. \tag{B.7}$$

Here

$$G^{(u)}(\tilde{d}^{(t+1)}\mathbf{u}^{(t)}; \widetilde{\mathrm{L}}^{(t+1)}) = G(\widetilde{\mathbf{C}}^{(t+1)}; \widetilde{\mathbf{C}}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}) = F(\widetilde{\mathbf{C}}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)})$$
$$G^{(u)}(\check{\mathbf{u}}^{(t+1)}; \widetilde{\mathrm{L}}^{(t+1)}) = G(\mathbf{C}^{(t+1)}; \widetilde{\mathbf{C}}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}).$$

Similar to result (B.6) in V-step, for U-step we have

$$F(\widetilde{\mathbf{C}}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}) - F(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}) \geq \frac{\kappa_2}{2}\|\tilde{\mathbf{C}}^{(t+1)} - \mathbf{C}^{(t+1)}\|_F^2, \tag{B.8}$$

where $\kappa_2 = 1 + \eta_1(1 + (1-\alpha)\lambda) - \gamma_2$ and

$$\gamma_2 = \max_{1 \leq k \leq q} \sup_{a \in (0,1)} \|\mathbf{I}(\Theta_{\cdot k}(\xi_c^{(t+1)}, \boldsymbol{\beta}^{(t)}), \phi_k^{(t)})\|_2, \tag{B.9}$$

such that $\mathbf{I}(\Theta_{\cdot k}(\xi_c^{(t+1)}, \boldsymbol{\beta}^{(t)}))$ is defined according to (B.4) for $\xi_c^{(t+1)} = \xi_u^{(t+1)}\mathbf{v}^{(t+1)\mathrm{T}}$ with $\xi_u^{(t+1)} \in \{a\check{u}^{(t)} + (1-a)\check{u}^{(t+1)}; 0 < a < 1\}$. To define $\mathbf{I}(\Theta_{\cdot k}(\xi_c^{(t+1)}, \boldsymbol{\beta}^{(t)}))$, we require $\zeta(\Theta_{\cdot k}(\xi_c^{(t+1)}, \boldsymbol{\beta}^{(t)}), \phi_k^{(t)})$ which is defined in (B.5).

In $\boldsymbol{\beta}$-step, $\boldsymbol{\beta}^{(t+1)}$ minimize the objective function $H(\cdot)$ (3.27). Similar to (B.2) in V-step,

we get,

$$H(\boldsymbol{\beta}^{(t)};\mathbf{C}^{(t+1)},\boldsymbol{\beta}^{(t)},\Phi^{(t)})-H(\boldsymbol{\beta}^{(t+1)};\mathbf{C}^{(t+1)},\boldsymbol{\beta}^{(t)},\Phi^{(t)})\geq\frac{1}{2}\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^{(t+1)}\|_F^2,$$

by the triangular inequality. Similar to result (B.6) in V-step, it follows that for $\boldsymbol{\beta}$-step we can say that

$$F(\mathbf{C}^{(t+1)},\boldsymbol{\beta}^{(t)},\Phi^{(t)})-F(\mathbf{C}^{(t+1)},\boldsymbol{\beta}^{(t+1)},\Phi^{(t)})\geq\frac{\kappa_3}{2}\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\|_F^2, \tag{B.10}$$

where $\kappa_3=2-\gamma_3$ and

$$\gamma_3=\max_{1\leq k\leq q}\sup_{a\in(0,1)}\|\mathbf{I}(\Theta_{\cdot k}(\mathbf{C}^{(t+1)},\xi_\beta^{(t+1)}),\phi_k^{(t)})\|_2, \tag{B.11}$$

where $\xi_\beta^{(t+1)}\in\{a\boldsymbol{\beta}^{(t)}+(1-a)\boldsymbol{\beta}^{(t+1)};0<a<1\}$. Use (B.4), (B.5) to define $\mathbf{I}(\Theta_{\cdot k}(\mathbf{C}^{(t+1)},\xi_\beta^{(t+1)}),\phi_k^{(t)})$. Therefore, with $\kappa_3\geq 0$, the non-increasing property of the $\boldsymbol{\beta}$-step of algorithm is guaranteed. Again, this can be achieved by proper scaling of $\mathbf{Z}$. Finally, the unknown dispersion parameters are estimated based on maximizing the log-likelihood, so it is guaranteed to non-increase the objective function. Thus, on adding results in (B.6), (B.8) and (B.10), the result in Theorem 3.1 easily follows.

□

B.2 Proof of Theorem 3.2

Proof: In the given scenario, let us denote the parameter set by $Ł=(d,\mathbf{u},\mathbf{v},\boldsymbol{\beta},\Phi)$, and express the objective function interchangeably as $F(d,\mathbf{u},\mathbf{v},\boldsymbol{\beta},\Phi;\mathbf{O})=F(d,\mathbf{u},\mathbf{v},\boldsymbol{\beta},\Phi)=$

$F(\text{Ł})$. Define the constrained parameter space for (3.34) as

$$\Omega = \{(d, \mathbf{u}, \mathbf{v}, \boldsymbol{\beta}, \Phi); d \geq 0, \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} = n, \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1, \boldsymbol{\beta} \in \mathbb{R}^{p_z \times q}, \Phi > 0\}.$$

It is easy to see that $\Omega = \Omega_d \times \Omega_{\mathbf{u}} \times \Omega_{\mathbf{v}} \times \Omega_\beta \times \Omega_\Phi$, where $\Omega_d = \{d; d \geq 0\}$, $\Omega_{\mathbf{u}} = \{\mathbf{u}; \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u} = n\}$, $\Omega_{\mathbf{v}} = \{\mathbf{v}; \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1\}$, Ω_β; $\boldsymbol{\beta} \in \mathbb{R}^{p_z \times q}$ and $\Omega_\Phi; \Phi > \mathbf{0}$. G-CURE algorithm (5) proceeds via $\mathbf{C}$-step, $\boldsymbol{\beta}$-step and Φ-step. The V-step and U-step are two sub-steps of the $\mathbf{C}$-step in which block variables $(d, \mathbf{v})$ and $(d, \mathbf{u})$ are updated respectively, keeping other parameters in Ł fixed. Starting from $\text{Ł}^s = (d^s, \mathbf{u}^s, \mathbf{v}^s, \boldsymbol{\beta}^s, \Phi^s)$, we first update $(d, \mathbf{v})$ for fixed $(\mathbf{u} = \mathbf{u}^s, \boldsymbol{\beta} = \boldsymbol{\beta}^s, \Phi = \Phi^s)$ to obtain $(\tilde{d}^{s+1}, \mathbf{u}^s, \mathbf{v}^{s+1}, \boldsymbol{\beta}^s, \Phi^s)$, and then similarly update block variable $(d, \mathbf{u})$ for fixed $(\mathbf{v} = \mathbf{v}^{s+1}, \boldsymbol{\beta}^s, \Phi = \Phi^s)$ to obtain $(d^{s+1}, \mathbf{u}^{s+1}, v^{s+1}, \boldsymbol{\beta}^s, \Phi^s)$. Then subsequent steps update $\boldsymbol{\beta}$ and Φ to obtain Ł^{s+1}. In the following, we consider the case that $\lambda > 0$ and $0 < \alpha < 1$, and assume the weights $\{w_{ij}\}$ and the data $(\mathbf{Y}, \mathbf{X})$ are finite, and the initial value $(\mathbf{u}^0, \boldsymbol{\beta}^0, \Phi^0)$ satisfies $\arg\min_{\check{\mathbf{V}}} F(\check{\mathbf{v}}; \mathbf{u}^0, \boldsymbol{\beta}^0, \Phi^0, \lambda) \neq \mathbf{0}$, i.e, $\widetilde{d^1} \neq 0$. (We have dropped subscript λ in the above notations for convenience.)

G-CURE algorithm is non-increasing which has been shown in Theorem 3.1. It has been established that a condition under which $(\kappa_1, \kappa_2, \kappa_3)$ are non-negative ensures monotone decreasing property of G-CURE, i.e., $F(\text{Ł}^{s+1}) \leq F(\text{Ł}^s)$. Moreover, the objective function in (3.34) is bounded from below, so that the sequence $\{F(\text{Ł}^s)\}_{s \in \mathbb{N}}$ generated by Algorithm 5 converges monotonically.

In proving the result, we have considered the case of $\Phi = \mathbf{I}$ and control variable parameter $\boldsymbol{\beta} = \mathbf{0}$. It is trivial to extend the result for the general case. Our convergence proof utilizes the framework developed to prove convergence in Theorem 2.3 of Chapter 2. Using the framework, it is sufficient to prove that sequence $\{\text{Ł}^s\}_{s \in \mathbb{N}}$ is uniformly

bounded.

Use Theorem 3.1 to write result for $t = 0, \ldots, n$, and then on addition we get

$$F(\Theta^{(0)}) - F(\Theta^{(n+1)}) \tag{B.12}$$

$$\geq \sum_{t=0}^{n} \frac{\kappa_1}{2} \|\tilde{\mathbf{C}}^{(t+1)} - \mathbf{C}^{(t)}\|_F^2 + \frac{\kappa_2}{2} \|\mathbf{C}^{(t+1)} - \tilde{\mathbf{C}}^{(t+1)}\|_F^2 + \frac{\kappa_3}{2} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_F^2. \tag{B.13}$$

For bounded $Ł^0$, we have $F(\Theta^{(0)})$ bounded as well, and sequence $F(\Theta^{(t)})$ is bounded below and monotone decreasing (3.1). Hence, for any n, the RHS of (B.13) is upper bounded. This results in parameters of set $Ł^t$ to be bounded for $t = 1, \ldots, n$. Using induction, we can say that the parameters in the set $Ł^t$ to be bounded when n tends to infinity.

□

B.3 Proof of Theorem 3.4

Proof: A unit step of sequential approach aims to extract unit rank matrix $\mathbf{C}_k$ and control variable parameter $\boldsymbol{\beta}$. Unit-rank matrix $\mathbf{C}_k$ can be written $\mathbf{C}_k = \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}}$ where $\mathbf{u}_k \in \mathbb{R}^p$ and $\mathbf{v}_k \in \mathbb{R}^q$. To have unique decomposition, let's assume for an arbitrary non-zero index ℓ_k in $\mathbf{v}_k$ we will have $\mathbf{v}_{k\ell_k} = 1$. Now define

$$\Omega_k = \left\{ (\mathbf{u}_k \mathbf{v}_k^{\mathrm{T}}, \boldsymbol{\beta}) : \mathbf{u}_k \in \mathbb{R}^p \text{ and } \mathbf{v}_k \in \mathbb{R}^q \text{ with } v_{\ell_k k} = 1, \boldsymbol{\beta} \in \mathbb{R}^{pz \times q} \right\}.$$

For any kth step, parameters estimate $(\widehat{\mathbf{C}}_k, \widehat{\boldsymbol{\beta}}) \in \Omega_k$. We first prove the result for $k = 1$. Consider a neighborhood of $(\mathbf{C}_1^*, \boldsymbol{\beta}^*)$ where $\mathbf{C}_1^* = \mathbf{u}_1^* \mathbf{v}_1^{*\mathrm{T}}$ of radius $h > 0$,

$$\begin{aligned}\mathcal{N}(\mathbf{C}_1^*, \boldsymbol{\beta}^*, h) =& \{(\mathbf{u}_1^* + \mathbf{a}/\sqrt{n})(\mathbf{v}_1^* + \mathbf{b}/\sqrt{n})^{\mathrm{T}}, \boldsymbol{\beta}^* + \mathcal{A}/\sqrt{n}\}; \\ & \text{s.t.} \qquad \|\boldsymbol{\Gamma}^{1/2}\mathbf{a}\| \leq h, \mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^q, \|\mathbf{b}\| \leq h, b_{\ell_1} = 0, \|\mathcal{A}\| \leq h.\end{aligned}$$

We claim that for any $\epsilon > 0$, there exists a large enough h such that

$$P\left\{\inf_{\|\Gamma^{1/2}\mathbf{a}\| = \|\mathbf{b}\| = \|\mathcal{A}\| = h} F_1^{(n)}(\mathbf{u}_1^* + \mathbf{a}/\sqrt{n}, \mathbf{v}_1^* + \mathbf{b}/\sqrt{n}, \boldsymbol{\beta}^* + \mathcal{A}/\sqrt{n}) > F_1^{(n)}(\mathbf{u}_1^*, \mathbf{v}_1^*, \boldsymbol{\beta}^*)\right\} \geq 1 - \epsilon. \tag{B.14}$$

The claim implies that with probability at least $1-\epsilon$ there exists local minimum $(\widehat{\mathbf{C}}_1, \widehat{\boldsymbol{\beta}})$ in the interior of $\mathcal{N}(\mathbf{C}_1^*, \boldsymbol{\beta}^*, h)$, i.e. there exists a local minimizer such that $\|\widehat{\mathbf{u}}_1 - \mathbf{u}_1^*\| = O_p(n^{-1/2})$, $\|\widehat{\mathbf{v}}_1 - \mathbf{v}_1^*\| = O_p(n^{-1/2})$ and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| = O_p(n^{-1/2})$.

Let $\widehat{\mathbf{u}}_1 = \mathbf{u}_1^* + \mathbf{a}/\sqrt{n}$, $\widehat{\mathbf{v}}_1 = \mathbf{v}_1^* + \mathbf{b}/\sqrt{n}$ and $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \mathcal{A}/\sqrt{n}$. For $\widehat{\mathbf{C}}_1 = \widehat{\mathbf{u}}_1 \widehat{\mathbf{v}}_1^{\mathrm{T}}$, define natural parameter matrix estimate $\widehat{\Theta}_1 = \mathbf{X}\widehat{\mathbf{C}}_1 + \mathbf{Z}\widehat{\boldsymbol{\beta}}$. Similarly, define corresponding true natural parameter matrix $\Theta_1^* = \mathbf{X}\mathbf{C}_1^* + \mathbf{Z}\boldsymbol{\beta}^*$.

It remains to verify (B.14). Define

$$\begin{aligned}\Psi_1^{(n)}(\mathbf{a}, \mathbf{b}, \mathcal{A}) &= F_1^{(n)}(\mathbf{u}_1^* + \mathbf{a}/\sqrt{n}, \mathbf{v}_1^* + \mathbf{b}/\sqrt{n}, \boldsymbol{\beta}^* + \mathcal{A}/\sqrt{n}) - F_1^{(n)}(\mathbf{u}_1^*, \mathbf{v}_1^*, \boldsymbol{\beta}^*) \\ &= F_1^{(n)}(\widehat{\Theta}_1) - F_1^{(n)}(\Theta_1^*) \\ &= T_1 + T_2 + T_3, \end{aligned} \tag{B.15}$$

where

$$\left.\begin{aligned} T_1 &= -\langle \mathbf{Y}, \widehat{\Theta}_1 - \Theta_1^* \rangle + \langle \mathbf{J}, \mathbf{B}(\widehat{\Theta}_1) - \mathbf{B}(\Theta_1^*) \rangle \\ T_2 &= \alpha \lambda_1^{(n)} \left\{ \|\mathbf{W}_1 \circ (\mathbf{u}_1^* + \mathbf{a}/\sqrt{n})(\mathbf{v}_1^* + \mathbf{b}/\sqrt{n})^{\mathrm{T}}\|_1 - \|\mathbf{W}_1 \circ \mathbf{u}_1^* \mathbf{v}_1^{*\mathrm{T}}\|_1 \right\}, \\ T_3 &= (1-\alpha)\lambda_1^{(n)} \{ \|(\mathbf{u}_1^* + \mathbf{a}/\sqrt{n})(\mathbf{v}_1^* + \mathbf{b}/\sqrt{n})^{\mathrm{T}}\|_F^2 - \|\mathbf{u}_1^* \mathbf{v}_1^{*\mathrm{T}}\|_F^2 \}. \end{aligned}\right\} . \tag{B.16}$$

We observe that terms T_2 and T_3 are of $O(h)$; for details see proof of Theorem 2.13 in Chapter 2. Now, we focus on simplifying T_1. On using second order Taylor expansion of $\mathbf{B}(\widehat{\Theta}_1)$ and assumption **A4**, we get

$$\langle \mathbf{J}, \mathbf{B}(\widehat{\Theta}_1) - \mathbf{B}(\Theta_1^*) \rangle \;\geq\; \langle \mathbf{B}^{'}(\Theta_1^*), \widehat{\Theta}_1 - \Theta_1^* \rangle + \frac{\gamma}{2} \|\widehat{\Theta}_1 - \Theta_1^*\|_F^2,$$

and write

$$\begin{aligned} T_1 &\geq -\langle \mathbf{Y}, \widehat{\Theta}_1 - \Theta_1^* \rangle + \langle \mathbf{B}^{'}(\Theta_1^*), \widehat{\Theta}_1 - \Theta_1^* \rangle + \frac{\gamma}{2} \|\widehat{\Theta}_1 - \Theta_1^*\|_F^2 \\ &= -T_{11} + T_{12}, \end{aligned} \tag{B.17}$$

where

$$\begin{aligned} T_{11} &= \langle \mathbf{E}, \Theta_1^* - \widehat{\Theta}_1 \rangle \\ T_{12} &= \langle \mathbf{B}^{'}(\Theta_1^*) - \mathbf{B}^{'}(\Theta^*), \widehat{\Theta}_1 - \Theta_1^* \rangle + \frac{\gamma}{2} \|\widehat{\Theta}_1 - \Theta_1^*\|_F^2. \end{aligned}$$

Using Theorem 7.1 of Xu [2016], we can say that

$$T_{11} \sim Sub - exponential(\sigma_*, b_*),$$

where $\sigma_*^2 = \|\Theta_1^* - \widehat{\Theta}_1\|_F^2 \sigma^2$ and $b_* = b\|\Theta_1^* - \widehat{\Theta}_1\|_\infty$. On applying tail-bound inequality

on T_{11}, for every $\widetilde{\epsilon} > 0$, the term is lower bounded by $\min\{\sigma_* \sqrt{-2\log(\widetilde{\epsilon})}, -2b_* \log(\widetilde{\epsilon})\}$. Hence, T_{11} is of $O(h)$ as well. On combining the results obtained so far, we get

$$\Psi_1^{(n)}(\mathbf{a}, \mathbf{b}, \mathcal{A}) \geq T_{12} + O(h) + O_p(1/\sqrt{n}).$$

Now, we shift our focus on simplification of T_{12}. In terms of the first order Taylor expansion of $\mathbf{B}^{'}(\Theta_1^*)$ near $\mathbf{B}^{'}(\Theta^*)$, we write

$$\langle \mathbf{B}^{'}(\Theta_1^*) - \mathbf{B}^{'}(\Theta^*), \widehat{\Theta}_1 - \Theta_1^* \rangle \geq \gamma \langle \Theta_1^* - \Theta^*, \widehat{\Theta}_1 - \Theta_1^* \rangle$$

which results in

$$T_{12} \geq \gamma \langle \Theta_1^* - \Theta^*, \widehat{\Theta}_1 - \Theta_1^* \rangle + \frac{\gamma}{2} \|\widehat{\Theta}_1 - \Theta_1^*\|_F^2.$$

On replacing the $\Theta^* = \mathbf{X}\mathbf{C}^* + \mathbf{Z}\boldsymbol{\beta}^*, \Theta_1^* = \mathbf{X}\mathbf{C}_1^* + \mathbf{Z}\boldsymbol{\beta}^*$ and $\widehat{\Theta}_1 = \mathbf{X}\widehat{\mathbf{C}}_1 + \mathbf{Z}\widehat{\boldsymbol{\beta}}$, we get

$$T_{12} \geq \gamma \langle -\mathbf{X}\mathbf{C}_{-1}^*, \frac{\mathbf{Z}\mathcal{A}}{\sqrt{n}} + \frac{\mathbf{X}}{\sqrt{n}}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}}) \rangle + \frac{\gamma}{2n} \|\mathbf{Z}\mathcal{A} + \mathbf{X}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}})\|_F^2, \tag{B.18}$$

where $\mathbf{C}_{-1}^* = \sum_{l>1}^r \mathbf{C}_l^*$. Using assumption **A1** and the orthogonality constraint defined in (3.37), the terms in R.H.S of (B.18) can be simplified as

$$\begin{aligned}
&\frac{1}{2n}\|\mathbf{Z}\mathcal{A} + \mathbf{X}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}})\|_F^2 = \frac{1}{2n}(\|\mathbf{Z}\mathcal{A}\|_F^2 + \|\mathbf{X}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}})\|_F^2), \\
&\langle -\mathbf{X}\mathbf{C}_{-1}^*, \frac{\mathbf{X}}{\sqrt{n}}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}}) \rangle = -\sum_{l>1}^{r^*} \mathbf{a}^{\mathrm{T}}\boldsymbol{\Gamma}_1\mathbf{u}_l^*\mathbf{b}^{\mathrm{T}}\mathbf{v}_l^*, \\
&\langle -\mathbf{X}\mathbf{C}_{-1}^*, \frac{\mathbf{Z}\mathcal{A}}{\sqrt{n}} \rangle = 0
\end{aligned}$$

Thus, we have

$$T_{12} \geq \gtrless\Big(\frac{1}{2n}\|\mathbf{Z}\mathcal{A}\|_F^2 + \frac{1}{2n}\|\mathbf{X}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}})\|_F^2 - \sum_{l>1}^{r^*} \mathbf{a}^{\mathrm{T}}\mathbf{\Gamma}_1\mathbf{u}_l^*\mathbf{b}^{\mathrm{T}}\mathbf{v}_l^*\Big).$$

In the proof of Theorem 2.14 Chapter 2, it has been shown that the term

$$\frac{1}{2n}\|\mathbf{X}(\mathbf{u}_1^*\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_1^{*\mathrm{T}})\|_F^2 - \sum_{l>1}^{r^*} \mathbf{a}^{\mathrm{T}}\mathbf{\Gamma}_1\mathbf{u}_l^*\mathbf{b}^{\mathrm{T}}\mathbf{v}_l^* \geq 0,$$

and of $O(h^2)$ under the specified assumptions. The other term $\frac{1}{2n}\|\mathbf{Z}\mathcal{A}\|_F^2$ is also $O(h^2)$. From this, we conclude that the quadratic terms of $\mathbf{a}$, $\mathbf{b}$ and $\mathcal{A}$ are of $O(h^2)$ and positive, thus dominate the other terms of order $O(h)$, for sufficiently large h. Hence, $\Psi_1^{(n)}(\mathbf{a}, \mathbf{b}, \mathcal{A}) > 0$.

Now, we prove result for any kth SVD constituent estimate $(\widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k)$. Suppose $\|\widehat{\mathbf{u}}_l - \mathbf{u}_l^*\| = O_p(n^{-1/2})$ and $\|\widehat{\mathbf{v}}_l - \mathbf{v}_l^*\| = O_p(n^{-1/2})$, $l = 1, \ldots, k-1$, for some $k \geq 2$. Define

$$\begin{aligned}\mathcal{N}(\mathbf{C}_k^*, \boldsymbol{\beta}^*, h) =& \{(\mathbf{u}_k^* + \mathbf{a}/\sqrt{n})(\mathbf{v}_k^* + \mathbf{b}/\sqrt{n})^{\mathrm{T}}, \boldsymbol{\beta}^* + \mathcal{A}/\sqrt{n}\};\\ &\text{s.t.} \qquad \|\mathbf{\Gamma}^{1/2}\mathbf{a}\| \leq h, \mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^q, \|\mathbf{b}\| \leq h, b_{\ell_k} = 0, \|\mathcal{A}\| \leq h.\end{aligned}$$

We claim that for any $\epsilon > 0$, there exists a large enough h such that

$$P\left\{\inf_{\|\Gamma^{1/2}\mathbf{a}\|=\|\mathbf{b}\|=\|\mathcal{A}\|=h} F_k^{(n)}(\mathbf{u}_k^* + \mathbf{a}/\sqrt{n}, \mathbf{v}_k^* + \mathbf{b}/\sqrt{n}, \boldsymbol{\beta}^* + \mathcal{A}/\sqrt{n}) > F_k^{(n)}(\mathbf{u}_k^*, \mathbf{v}_k^*, \boldsymbol{\beta}^*)\right\} \geq 1 - \epsilon, \tag{B.19}$$

where offset term $\mathbf{O}_k = \mathbf{X}\sum_{l=1}^{k-1}\widehat{\mathbf{C}}_l$ to be used in defining $F_k^{(n)}(\cdot)$.

Let $\widehat{\mathbf{u}}_k = \mathbf{u}_k^* + \mathbf{a}/\sqrt{n}$, $\widehat{\mathbf{v}}_k = \mathbf{v}_k^* + \mathbf{b}/\sqrt{n}$ and $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \mathcal{A}/\sqrt{n}$. For $\widehat{\mathbf{C}}_k = \widehat{\mathbf{u}}_k\widehat{\mathbf{v}}_k^{\mathrm{T}}$, define natural parameter matrix estimate $\widehat{\Theta}_k = \mathbf{O}_k + \mathbf{X}\widehat{\mathbf{C}}_k + \mathbf{Z}\widehat{\boldsymbol{\beta}}$. Similarly, define

corresponding true natural parameter matrix $\Theta_k^* = \mathbf{O}_k + \mathbf{X}\mathbf{C}_k^* + \mathbf{Z}\boldsymbol{\beta}^*$. Define

$$\begin{aligned}\Psi_k^{(n)}(\mathbf{a}, \mathbf{b}, \mathcal{A}) &= F_k^{(n)}(\mathbf{u}_k^* + \mathbf{a}/\sqrt{n}, \mathbf{v}_k^* + \mathbf{b}/\sqrt{n}, \boldsymbol{\beta}^* + \mathcal{A}/\sqrt{n}) - F_k^{(n)}(\mathbf{u}_k^*, \mathbf{v}_k^*, \boldsymbol{\beta}^*) \\ &= F_k^{(n)}(\widehat{\Theta}_k) - F_k^{(n)}(\Theta_k^*) \\ &= T_1 + T_2 + T_3, \end{aligned} \tag{B.20}$$

where T_1, T_2, T_3 are similar defined as in (B.16), by replacing $(\mathbf{O}_1, \mathbf{u}_1^*, \mathbf{v}_1^*, \mathbf{W}_1, \lambda_1^{(n)})$ with $(\mathbf{O}_k, \mathbf{u}_k^*, \mathbf{v}_k^*, \mathbf{W}_k, \lambda_k^{(n)})$. Again, we write $T_1 = T_{11} + T_{12}$, where

$$\begin{aligned} T_{11} &= \langle \mathbf{E}, \Theta_k^* - \widehat{\Theta}_k \rangle \\ T_{12} &= \langle \mathbf{B}^{'}(\Theta_k^*) - \mathbf{B}^{'}(\Theta^*), \widehat{\Theta}_k - \Theta_k^* \rangle + \frac{\gamma}{2} \|\widehat{\Theta}_k - \Theta_k^*\|_F^2. \end{aligned}$$

Each of the terms T_{11}, T_2 and T_3 are of $O(h)$ in a similar way as we have for $k = 1$.

Again using a Taylor expansion of $\mathbf{B}^{'}(\Theta_k^*)$ near $\mathbf{B}^{'}(\Theta^*)$, we can write

$$\langle \mathbf{B}^{'}(\Theta_k^*) - \mathbf{B}^{'}(\Theta^*), \widehat{\Theta}_k - \Theta_k^* \rangle \geq \gamma \langle \Theta_k^* - \Theta^*, \widehat{\Theta}_k - \Theta_k^* \rangle$$

which results in

$$T_{12} \geq \gamma(\langle \Theta_k^* - \Theta^*, \widehat{\Theta}_k - \Theta_k^* \rangle + \frac{1}{2} \|\widehat{\Theta}_k - \Theta_k^*\|_F^2).$$

On replacing the Θ_k^*, Θ^* and $\widehat{\Theta}_k$, we get

$$\begin{aligned} T_{12} \geq \gamma(\langle \mathbf{X} \sum_{i=1}^{k-1} (\widehat{\mathbf{C}}_i - \mathbf{C}_i^*) - \mathbf{X}\mathbf{C}_{-k}^*, \frac{\mathbf{Z}\mathcal{A}}{\sqrt{n}} + \frac{\mathbf{X}}{\sqrt{n}} (\mathbf{u}_k^* \mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_k^{*\mathrm{T}} + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}}) \rangle \\ + \frac{1}{2n} \|\mathbf{Z}\mathcal{A} + \mathbf{X}(\mathbf{u}_k^* \mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}_k^{*\mathrm{T}})\|_F^2), \end{aligned}$$

where $\mathbf{C}^*_{-k} = \sum_{l>k}^{r} \mathbf{C}^*_l$. Using assumption **A1** and given that $\widehat{\mathbf{C}}_i$ is $\sqrt{n}$ consistent for $i < k$, we can say that

$$\frac{1}{2n}\|\mathbf{Z}\mathcal{A} + \mathbf{X}(\mathbf{u}^*_k\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}^{*\mathrm{T}}_k)\|^2_F = \frac{1}{2n}(\|\mathbf{Z}\mathcal{A}\|^2_F + \|\mathbf{X}(\mathbf{u}^*_k\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}^{*\mathrm{T}}_k)\|^2_F),$$

$$\langle -\mathbf{X}\mathbf{C}^*_{-k}, \frac{\mathbf{X}}{\sqrt{n}}(\mathbf{u}^*_k\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}^{*\mathrm{T}}_k + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})\rangle = -\sum_{l>k}^{r^*} \mathbf{a}^{\mathrm{T}}\mathbf{\Gamma}_1\mathbf{u}^*_l\mathbf{b}^{\mathrm{T}}\mathbf{v}^*_l,$$

$$\langle -\mathbf{X}\mathbf{C}^*_{-k}, \frac{\mathbf{Z}\mathcal{A}}{\sqrt{n}}\rangle = 0,$$

$$\langle \mathbf{X}\sum_{i=1}^{k-1}(\widehat{\mathbf{C}}_i - \mathbf{C}^*_i), , \frac{\mathbf{Z}\mathcal{A}}{\sqrt{n}}\rangle = 0,$$

$$\langle \mathbf{X}\sum_{i=1}^{k-1}(\widehat{\mathbf{C}}_i - \mathbf{C}^*_i), \frac{\mathbf{X}}{\sqrt{n}}(\mathbf{u}^*_k\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}^{*\mathrm{T}}_k + \frac{\mathbf{a}\mathbf{b}^{\mathrm{T}}}{\sqrt{n}})\rangle = O(h).$$

Thus, we have

$$T_{12} \geq \gtrsim(\frac{1}{2n}\|\mathbf{Z}\mathcal{A}\|^2_F + \frac{1}{2n}\|\mathbf{X}(\mathbf{u}^*_k\mathbf{b}^{\mathrm{T}} + \mathbf{a}\mathbf{v}^{*\mathrm{T}}_k)\|^2_F - \sum_{l>k}^{r^*} \mathbf{a}^{\mathrm{T}}\mathbf{\Gamma}_1\mathbf{u}^*_l\mathbf{b}^{\mathrm{T}}\mathbf{v}^*_l + O(h)).$$

The rest of the proof is similar to the case of $k = 1$ where result is proved using the proof of Theorem 2.14 in Chapter 2. This completes the proof.

□

Bibliography

Framingham heart study, 2017. URL `https://www.framinghamheartstudy.org/`.

T. W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, 22:327–351, 1951.

T. W. Anderson and Herman Rubin. Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, pages 111–150, Berkeley, Calif., 1956. University of California Press.

Mohammad Taha Bahadori, Zemin Zheng, Yan Liu, and Jinchi Lv. Scalable interpretable multi-response regression via seed. *arXiv preprint arXiv:1608.03686*, 2016.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

M Borga, T Landelius, and H Knutsson. A unified approach to *PCA, PLS, MLR and CCA*. Technical report, Technical Report, 1997.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, volume 3. 2011. ISBN 978-1-60198-461-6. doi: 10.1561/2200000016.

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200 – 217, 1967. ISSN 0041-5553. doi: 10.1016/0041-5553(67)90040-7.

P. J. Brown and J. V. Zidek. Adaptive multivariate ridge regression. *Annals of Statistics*, 8(1):pp. 64–74, 1980. ISSN 00905364.

F. Bunea, Y. She, and M. Wegkamp. Joint variable and rank selection for parsimonious estimation of high dimensional matrices. *Annals of Statistics*, 40(5): 2359–2388, 2012.

Florentina Bunea, Yiyuan She, and Marten Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics*, 39(2):1282–1309, 2011.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, December 2009. ISSN 1615-3375.

Kun Chen and Kung-Sik Chan. On rank reduction and variable selection in multivariate regression. *Journal of Statistical Theory and Practice*, 2014. Under review.

Kun Chen, Kung-Sik Chan, and Nils Chr. Stenseth. Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B*, 74(2):203–221, 2012.

Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013.

Kun Chen, Chongliang Luo, Jian Liang, Gen Li, Fei Wang, Changshui Zhang, and Dipak Dey. Leveraging mixed and incomplete outcomes via a mixed-response reduced rank regression. 2017.

Lisha Chen and Jianhua Z. Huang. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545, 2012a.

Lisha Chen and Jianhua Z Huang. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545, 2012b.

Hyonho Chun and Sündüz Kele**c**s. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

David R Cox and Nanny Wermuth. Response models for mixed binary and quantitative variables. *Biometrika*, 79(3):441–461, 1992.

Fernando De la Torre. A least-squares framework for component analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(6):1041–1055, 2012.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360, 2001. ISSN 01621459. doi: 10.2307/3085904.

Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010.

Yingying Fan and Cheng Yong Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B*, 75 (3):531–552, 2013.

Garrett M Fitzmaurice and Nan M Laird. Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American statistical Association*, 90(431):845–852, 1995.

Tom Goldstein and Stanley Osher. The split bregman method for l1-regularized problems. *SIAM J. Img. Sci.*, 2(2):323–343, April 2009. ISSN 1936-4954. doi: 10.1137/080725891.

Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, 2007a.

Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, 2007b.

Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970.

Yujin Hoshida, Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Subclass mapping: identifying common subtypes in independent disease data sets. 2007.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

Gareth M James, Courtney Paulson, and Paat Rusmevichientong. Penalized and constrained regression. 2013.

Bent Jorgensen. Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 127–162, 1987.

Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(Feb):517–553, 2010.

V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, 39(5): 2302–2329, 2011.

Mihee Lee, Haipeng Shen, Jianhua Z. Huang, and J. S. Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 66:1087–1095, 2010.

Tong Ihn Lee, Nicola J Rinaldi, Franccois Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, Christopher T Harbison, Craig M Thompson, Itamar Simon, et al. Transcriptional regulatory networks in saccharomyces cerevisiae. *science*, 298(5594):799–804, 2002.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

W. Lin, K. Chen, and J. Lv. Sparse orthogonal factor regression. *Technical Report*, 2014a. Under review.

Wei Lin, Pixu Shi, Rui Feng, and Hongzhe Li. Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797, 2014b.

Zhaosong Lu, Renato D. C. Monteiro, and Ming Yuan. Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Math. Program.*, 131(1-2):163–194, 2012.

David G Luenberger, Yinyu Ye, et al. *Linear and nonlinear programming*, volume 2. Springer.

Z. Ma and T. Sun. Adaptive sparse reduced-rank regression. *ArXiv e-prints*, March 2014. URL `http://arxiv.org/abs/1403.1922`.

Lester W Mackey. Deflation methods for sparse pca. In *Advances in neural information processing systems*, pages 1017–1024, 2009.

Aditya Mishra, Dipak K Dey, and Kun Chen. Sequential co-sparse factor regression. *Journal of Computational and Graphical Statistics*, (just-accepted), 2017.

Ashin Mukherjee and Ji Zhu. Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining*, 4(6):612–622, 2011.

Sahand Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011a.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011b.

Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 39(1): 1–47, 2011.

Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R. Pollack, and Pei Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.*, 4(1):53–77, 2010.

Ross L Prentice and Lue Ping Zhao. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, pages 825–839, 1991.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. ISBN 3-900051-07-0.

Gregory C. Reinsel and Palani Velu. *Multivariate reduced-rank regression: theory and applications*. New York: Springer, 1998.

Andreas Rosenwald, George Wright, Wing C Chan, Joseph M Connors, Elias Campo, Richard I Fisher, Randy D Gascoyne, H Konrad Muller-Hermelink, Erlend B Smeland, Jena M Giltnane, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947, 2002.

Yiyuan She. Thresholding-based iterative selection procedures for model selection and shrinkage. *Electron. J. Statist.*, 3:384–415, 2009.

Yiyuan She. Reduced Rank Multivariate Generalized Linear Models for Feature Extraction. 2011. URL `http://arxiv.org/abs/1007.3098`.

Yiyuan She. An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics & Data Analysis*, 56(10):2976 – 2990, 2012a. ISSN 0167-9473. Special Issue on Small Area Estimation The 3rd Special Issue on Optimization Heuristics in Estimation and Modelling Problems.

Yiyuan She. An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics & Data Analysis*, 56(10):2976–2990, 2012b.

Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 1998.

Damian C Stanziano, Michael Whitehurst, Patricia Graham, and Bernard A Roos. A review of selected longitudinal studies on aging: past findings and future directions. *Journal of the American Geriatrics Society*, 58(s2), 2010.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–147, 1974.

R. J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.

Berwin A Turlach, William N Venables, and Stephen J Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005. doi: 10.1198/004017005000000139.

A. W. van der Vaart. *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, 2000. ISBN 0521784506.

Lifeng Wang, Guang Chen, and Hongzhe Li. Group *SCAD* regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494, 2007.

Jiaming Xu. Topics in high-dimensional data analysis - concentration inequalities, 2016. URL `http://web.ics.purdue.edu/~xu972/lec07.pdf`.

Thomas Yee and Trevor J. Hastie. Reduced rank vector generalized linear models. *Statistical Modeling*, (3):367–378, 2003.

Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B*, 69(3):329–346, 2007.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

Lue Ping Zhao, Ross L Prentice, and Steven G Self. Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 805–811, 1992.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006. ISSN 1533-7928.

Hua Zhou and Lexin Li. Regularized matrix regression. *Journal of the Royal Statistical Society: Series B*, 76(2):463–483, 2014.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

Hui Zou and Trevor J. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x.

Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37:1733–1751, 2009.