ESTIMATION OF LARGE-SCALE CROSS-COVARIANCE MATRIX WITH GROUP INFORMATION

TEO GUO CI

NATIONAL UNIVERSITY OF SINGAPORE

2015

ESTIMATION OF LARGE-SCALE CROSS-COVARIANCE MATRIX WITH GROUP INFORMATION

TEO GUO CI

(B.Sc. Nanyang Technological University)

SUPERVISED BY

Choi Hyungwon & Johan Lim

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2015

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Teo Guo Ci

October 2015

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Choi Hyungwon, for his guidance and encouragement. I am grateful for his effort and time spent in supervising my research work in the past four years and I am honored to be his student.

I am also grateful to Dr. Johan Lim for guiding me on the theoretical basis of our methodology proposed in this thesis.

I would also like to thank faculty members, staffs and my classmates in the Department of Statistics and Applied Probability for helping me during the course of my study.

Finally, I appreciate the examiners for reviewing this thesis.

Contents

List of Figures

List of Tables

Summary

Shrinkage estimation of large-scale sparse covariance matrix is an important technique in the exploratory analysis of high-dimensional data sets. We propose a two-step procedure to estimate large-scale covariance matrix for one set of variables $\mathbf{X}$ with known group information, followed by estimation of covariance matrix between $\mathbf{X}$ and another set of variables $\mathbf{Y}$, denoted by $\mathbf{\Sigma}_{\mathbf{XX}}$ and $\mathbf{\Sigma}_{\mathbf{XY}}$ respectively. The covariance matrix $\mathbf{\Sigma}_{\mathbf{XY}}$ is estimated as the product of two components, namely $\mathbf{\Sigma}_{\mathbf{XX}}\mathbf{B}$. Similar to the idea of Fan et al. (2013)'s Principal Orthogonal complEment Thresholding (POET) estimator, $\mathbf{\Sigma}_{\mathbf{XX}}$ is decomposed into a systematic factor component and an idiosyncratic component $\mathbf{\Sigma}_{\mathbf{XX}} = \mathbf{\Sigma}_{\mathbf{f}} + \mathbf{\Sigma}_{\boldsymbol{v}}$ in the first step, where the former explains the variability associated with known groups and the latter the residual variability in $\mathbf{X}$. $\mathbf{B}$ is estimated using the group lasso based on the selected groups from the first step. We present the asymptotic properties of the two-step estimator with appropriate conditions. We illustrate the methodology using simulations and the mRNA expression data from The Cancer Genome Atlas (TCGA), focusing on the covariance structure between kinases and substrates.

CHAPTER 1

Introduction

1.1 Motivation

A plethora of statistical methods have been developed for the analysis of high-dimensional molecular data over the past decade, creating a large body of statistical methods for sample classification and prediction, multiple testing correction (Benjamini and Hochberg, 1995; Storey, 2002), and integrative data analysis of multi-platform molecular data sets (Lê Cao et al., 2009; Sass et al., 2013; Shen et al., 2009, 2013; Troyanskaya et al., 2003). An important area of application that has received attention of late is the inference of gene-to-gene association or gene regulatory network inference, aiming to identify gene modules, which reflect that those molecules within a module are functionally related in the context of the given molecular study

(Barabási et al., 2011; Mitra et al., 2013; Oldham et al., 2008; Segal et al., 2003a,b, 2004; Stuart et al., 2003; Vidal et al., 2011; Zhang and Horvath, 2005). Such methods are applicable to both controlled experiments over distinct conditions in model systems and population studies of humans, and the analysis often reveals context-specific gene-gene interactions associated with relevant biological functions.

In the recent statistics literature, this problem has been formulated as large-scale covariance matrix estimation under sparsity assumptions. The premise that the correlation structure is sparse bodes well with the biological reality, in which a small number of functionally related molecules leads to sparse specification of non-zero elements in the covariance matrix. Development of this methodology initially started with element-wise shrinkage estimators in scenarios where the general correlation structure is known (e.g. banding estimator in time series data (Bickel and Levina, 2008b)) and where such structure is unknown (Rothman et al., 2009). Similar to the development of the Least Absolute Shrinkage and Selection Operator (lasso) (Tibshirani, 1996) regression techniques, a more adaptive estimator has also been developed (Cai and Liu, 2011). The theoretical properties such as the convergence rates and selection/consistency have been well established (Bunea et al., 2007; Knight and Fu, 2000). A more relevant development to our work here was that of the Principal Orthogonal complEment Thresholding (POET) estimator, a shrinkage estimator of large covariance matrix by thresholding principal orthogonal complements, the idea of which we extend in our proposal later (Fan et al., 2013).

A common assumption in the existing shrinkage estimation methods is that each covariance term, between two molecular features such as two genes in a microarray experiment, is perceived as an independent unit on its own and thus element-wise shrinkage can be applied to each element independently to yield sparsity in the entire covariance matrix. In biological applications, however, a large collection of experimentally validated relational data are now available and their coverage of the entire "interactome" has substantially increased, and therefore the knowledge of protein complexes, pathways, and even the functional annotations such as gene ontology, can be further utilized to better identify groups of more than several functionally related covariances, i.e. beyond a pair of molecules. Accordingly, it will be a desirable development if the existing group information can be incorporated into the sparse covariance matrix estimation, which is expected to guide more precise identification of co-regulated gene modules that are coherent with current biological literature.

1.2 Cross-covariance Matrix Estimation in Biological Applications

This thesis is primarily concerned with two-stage estimation of the large-scale *cross-covariance* matrix between two sets of variables (molecules) **X** and **Y**, where a part of the estimator consists of full covariance matrix estimation for **X** using group information. This formulation is motivated by real biological questions where one set of variables $\mathbf{X}$ represents upstream regulatory molecules such as those in the upstream signaling cascade and the other set of variables $\mathbf{Y}$ is their downstream regulatory targets or substrates. In this situation, the covariance between the two variables are estimated in a way that incorporates the group structure between the elements of $\mathbf{X}$ or that of $\mathbf{Y}$. We reformulate the problem into the familiar regression setup, in which each element of $\mathbf{Y}$ is expressed as a linear function of $\mathbf{X}$ and a noise component, and the covariates $\mathbf{X}$ are expected to be highly correlated in the high-dimensional regression.

First stage estimation

When the covariates are correlated, a popular solution in the regression is to use a class of factor models, where each factor is a weighted linear combination of elements of $\mathbf{X}$. One example is the principal component regression (Jolliffe, 1982), where the principal components, computed as the eigenvectors of the sample covariance matrix, are used as regressors. Despite great utility as a dimension reduction technique, it is well known that the practical problem with the principal component regression approach is that the factors (e.g. principal components) hardly lend themselves to straightforward interpretation and it involves arbitrary thresholding to assign associated genes to each factor. An ideal scenario is where a major factor is defined by a fixed set of variables (genes) with known biological annotation such as pathway or gene functions, yet it is difficult to have clear-cut assignment of one gene to one group of variables in complex systems because individual genes are often involved in more than one functions in a multifactorial manner.

From a biologist's point of view, the most informative analysis is to directly decompose the correlation of $\mathbf{X}$, denoted by $\mathbf{\Sigma_{XX}}$, into two components, one representing the correlation that is attributable to known biological relationship (or grouping information hereafter), and the other representing the correlation associated with unknown sources. Borrowing the definitions from the POET estimator, we shall call them the *systematic* component $\mathbf{\Sigma_f}$ and the *idiosyncratic* component $\mathbf{\Sigma}_{\boldsymbol{v}}$, respectively. The underlying assumption is that the factors associated with known group information are sparse, and that with the shrinkage estimate of the residual covariance matrix, the covariance matrix of $\mathbf{X}$ can be estimated consistently and the semi-positive definiteness is guaranteed for the final estimate of $\mathbf{\Sigma_{XX}}$ under mild conditions.

Second stage estimation

Our ultimate goal is to estimate the cross-covariance matrix $\mathbf{\Sigma_{XY}}$ and to deduce how much of the total variability is attributable to the known biological relationship and how much is not. The covariance between $\mathbf{X}$ and $\mathbf{Y}$ is expressed as $\mathbf{\Sigma_{XX}B}$, where $\mathbf{B}$ is the regression coefficient matrix and each column $\mathbf{b}_j$ of $\mathbf{B}$ is the regression coefficient of $\mathbf{y}_j \sim \mathbf{X}$. Here, the selected factors from the first stage estimation is utilized as the "groups" in the second stage estimation of the regression coefficient matrix $\mathbf{B}$. Since the group information explicitly specifies which variables in $\mathbf{X}$ are associated with each factor, the selected groups can be used in forming grouped lasso regression between $\mathbf{X}$ and $\mathbf{Y}$, to yield a sparse linear model that incorporates the known biological grouping information in the covariates contributing to $\mathbf{\Sigma_{XX}}$. The outcome of the analysis is the decomposition of the cross-covariance due to the systematic component $\mathbf{\Sigma_f B}$ and the idiosyncratic component $\mathbf{\Sigma}_{\boldsymbol{v}}\mathbf{B}$.

1.3 Literature Review

Before we introduce the method, we first review the relevant statistics literature in the field of shrinkage estimation and large-scale covariance matrix estimation.

1.3.1 Review of shrinkage methods

Consider a linear regression model, $Y = X\beta + \epsilon$, where X is a $n \times p$ matrix, Y and ϵ are vectors of length n, and without loss of generality, let us assume Y and each column of X has zero mean. An ordinary least squares (OLS) estimate is obtained by minimizing the residual squared error:

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \|Y - X\beta\|_2^2$$

In high dimensional data, however, the coefficients are non-zero for all OLS estimates even when only a small subset of variables are indeed significant predictors. While the OLS estimates are unbiased, the prediction error will not be properly minimized if all variables are retained since the variance of the estimated response will be large with increasing model complexity.

The lasso (Tibshirani, 1996) is a popular method that addresses the drawback of OLS, which embodies a penalized least squares with L1-penalty. Instead of minimizing the residual squared error, a penalty is imposed on the regression coefficients β along with the loss function, i.e.

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

The entire solution path (varying λ) of the lasso can be either approximated by the LARS algorithm (Tibshirani et al., 2004), or explicitly obtained through efficient algorithms such as coordinate descent (Höfling and Tibshirani, 2007). It has a continuous piecewise linear solution path and coefficients are set to zero as the shrinkage increases. In particular, Knight and Fu (2000) shows that when the true regression coefficient is zero, its lasso estimate has a positive probability at zero, which allows for automatic variable selection and efficient computation of the optimal tuning parameter. However, the lasso has some limitations itself. As pointed out in Zou and Hastie (2005), if a group of highly correlated predictor variables exists, it tends to select only one variable among the group. In addition, the lasso performs variable selection for individual predictors, and thus the method is not directly applicable if there exists a natural grouping in the predictors and such group structures are desired to be reflected in the selection procedure.

The group lasso (Yuan and Lin, 2006), defined as

$$\hat{\beta}_{grouplasso} = \arg\min_{\beta} \left\| Y - \sum_{j=1}^{J} X_j \beta_j \right\|_2^2 + \lambda \sum_{j=1}^{J} \sqrt{p_j} \, \|\beta_j\|_2$$

was proposed to do group variable selection to address situations where there exists a natural grouping in the regression coefficients. Unlike the lasso, the solution path of group lasso is not piecewise linear in general and therefore finding the optimal tuning parameter requires intensive computation over a suitably fine grid of values. For a fixed tuning parameter, Yuan and Lin proposed a block coordinate-wise minimization method to compute the solutions. Each coordinate descent step is fast, with an explicit formula yielding the minimum for a coordinate. Theoretical properties of the group lasso for linear models has been first established in Nardi and Rinaldo (2008). In addition to the theoretical conditions for some optimality properties, Nardi and Rinaldo showed that those conditions are valid for the double-asymptotic scenario in which the dimension of the parameter space grows with the sample size.

We also remark that computational methods for shrinkage methods is a well-researched topic. Specialized methods such as LARS are efficient, but cannot be applied when the regression type is changed, for example from linear model to generalized linear model, or when the penalty function is replaced with another. Pathwise coordinate optimization are shown to be versatile in Höfling and Tibshirani (2007) for various shrinkage methods and the authors showed by comparing run times of programs for the algorithms, that coordinate-wise descent is very competitive with the LARS algorithm. The authors later extended their work for general linear models in Friedman et al. (2010). Our work utilizes the group lasso penalty to identify significant factors in the predictor variables. We have also used a block coordinate decent algorithm by Yuan and Lin (2006) to find the solution path.

1.3.2 Review of Sparse Covariance Matrix Estimation

Element-wise shrinkage of estimates was first proposed in Bickel and Levina (2008b). In their work, they developed shrinkage estimator of covariance matrix or precision matrix (inverse) for a specific class of covariance matrices, where the variables are ordered in a way that the true covariance matrix is a banded matrix (e.g. time series data). The shrinkage estimation was applied by banding, i.e. penalizing the covariance terms between variables that are not close enough, or banding the Cholesky decompositions of the covariance matrix as such. The authors have proven the optimality of the estimators under the operator norm for a family of covariance matrices thereafter.

However, banding of estimates is not applicable to problems where an ordering is not available. In their follow-up paper, Bickel and Levina (2008a) also investigated element-wise hard-thresholding estimator in a more generalized setting. In a similar way, Rothman et al. (2009) proposed a more generalized class of thresholding operators with element-wise shrinkage based on various penalty functions, and showed that their method is consistent and "sparsistent", implying that the locations of zero and nonzero elements are estimated correctly with probability tending to 1. In their simulations, they demonstrated that, when the matrix is sparse, thresholding based on the smoothly clipped absolute deviation (SCAD) (Antoniadis and Fan, 2001) penalty function, i.e.

$$p_\lambda(z) = \begin{cases} \mathrm{sgn}(z)(|z| - \lambda) & \text{when } |z| \leq 2\lambda \\ \frac{(a-1)z - a\lambda \mathrm{sgn}(z)}{a-2} & \text{when } 2\lambda < |z| \leq a\lambda \\ z & \text{otherwise} \end{cases}$$

performed the best in terms of sparsity. More recently, Cai and Liu (2011) proposed an adaptive version of element-wise shrinkage estimator that takes into account the variability of the estimate of each entry in the matrix, i.e. for heteroscedastic problems. In their method, the shrinkage parameter is tailored for each entry, rendering their thresholding rule non-universal (or uniform), and the choice of thresholds is completely data-dependent. Cai and Liu showed that their estimator achieves superior optimal rate of convergence or a wide class of sparse covariance matrices.

Lastly, Fan et al. (2013) proposed the principal orthogonal complement thresholding (POET) method. The method depends only on the sample covariance matrix, which is expressed as the sum of systematic component and idiosyncratic component. The former is first identified by the first $\hat{K}$ principal components.

$$\hat{\Sigma}_{\text{sam}} = \sum_{i=1}^{\hat{K}} \hat{\lambda}_i \hat{\xi}_i \hat{\xi}_i^{\mathrm{T}} + \sum_{i=\hat{K}+1}^{p} \hat{\lambda}_i \hat{\xi}_i \hat{\xi}_i^{\mathrm{T}}$$

where optimal $\hat{K}$ can be estimated from the data. A element-wise thresholding procedure is applied to attain regularized residual covariance matrix. Our proposal for sparse cross-covariance estimator borrows the idea of the POET estimator. While principal components are the entities being selected in POET, we form explicit groups of variables as factors reflecting the biological relationship and select significant factors as part of the decomposition of the covariance matrix into the systematic component. Using the knowledge of systematic component, we establish the relationship between the sparse set of variables (with specific biological functions) with the outcome variables via group lasso.

CHAPTER 2

The Cancer Genome Atlas Breast Cancer Data

2.1 Introduction

Integrative analysis of multiple high-throughput data is becoming increasingly popular in clinical studies, best exemplified by The Cancer Genome Atlas (TCGA) project (http://cancergenome.nih.gov/). Unlike conventional genomic studies relying on a single platform, TCGA is the first project providing multi-omics data sets across all popular platforms including massive parallel sequencing, mRNA expression arrays, DNA methylation, microRNA expression, reverse-phase protein assay (RPPA), and more recently mass spectrometry-based proteomics. In the invasive breast cancer cohort (Koboldt et al., 2012), for example, these technologies have been performed on the tumor samples from >800 patients in four major basic therapeutic groups previously defined by PAM50 mRNA expression clusters (Parker et al., 2009), including the luminal A and B groups consisting of estrogen receptor and progesterone positive group, the HER2 enriched group, and the basal-like group mostly represented by tumors with low expression of all three hormones. This project provides by far the most comprehensive molecular landscape of this therapeutically diverse disease.

Integration of multiple data sources can be done in many different ways, depending on the purpose of study. For example, expression quantitative trait loci (eQTL) studies aim to identify local and distant genetic regulation of mRNA transcript expression (Chun and Keles, 2009; Kendziorski and Wang, 2006; Rockman and Kruglyak, 2006). The most popular statistical approach in this area is multivariate regression with shrinkage estimation (L_1 penalty) between RNA data (response) and genetic loci (as predictors). On a similar line of methods, Peng et al. (2010) has developed RemMap, a method for evaluating the dependence of RNA transcript abundance on the DNA copy numbers in cancer genomics data via multivariate regression that can incorporate prior kwowledge to avoid penalization of known regulatory relationships. More recently, the so-called proteogenomics approaches surfaced in the proteomics literature, investigating the correlation between mRNA transcript and their corresponding protein abundance, the final gene product of the RNA (Zhang et al., 2014).

However, their analysis focuses on RNA-protein correlation within the same gene and the association between splice variants and protein isoforms rather than correlation of quantitative data between different genes (Zhang et al., 2014).

While estimation of invertible, sparse covariance matrix is generally of primary interest in high-dimensional data analysis, cross-covariance (or correlation) matrix, which is a subset of the entire variance-covariance matrix, is also a biologically important quantity of interest in integrative data analysis setting since it represents the association between distinct molecular types. Representative examples are when one molecular type is a precursor to the other in the central dogma of molecular biology (DNA to RNA, RNA to protein), or when one data type captures regulatory information to the other. For example, DNA methylation data coupled with mRNA transcriptomics data allows us decipher the transcriptional repression of the latter by the former. Likewise, expression data for non-coding RNAs known as microRNAs coupled with proteomics data will also reveal the degree of post-transcriptional regulation. Hence robust estimation of the cross-covariance matrices is of great importance in understanding the dependence between precursors and gene products or functionally active regulatory information in multi-omics data sets such as TCGA.

Estimators of large-scale sparse covariance matrices have typically been developed under sparsity assumption, and it is reasonable to expect this condition to hold true in biological applications with high-throughput experimental data. Hence the major computational approach is to regularize individual elements of the covariance matrix by thresholding, embodying shrinkage estimation techniques. These methods have the obvious advantage of easy implementation and is known to perform well in numerical simulations, further strengthened by theoretical properties with sufficiently sharp convergence rates (Bickel and Levina, 2008b; Cai and Liu, 2011; Rothman et al., 2009). Despite the technical developments over the years, it is nevertheless rare to find a case where a shrinkage estimator of large-scale covariance matrix yielded novel biological insights in a real biological or clinical problem. Moreover, with a few exceptions, the examples in the published articles are typically demonstrated with no more than a few hundreds of genes, which diminishes the applicability to data sets with typically thousands of features or variables.

One possible way to broaden the applicability of these estimators in biological problems is to utilize gene group information in the estimation procedure. We use the terms 'gene group' and 'pathway' interchangeably hereafter. To the best of our knowledge, the only development using group information for shrinkage estimation of large-scale covariance matrix is the work of Levina et al. (2008) in the case where there is natural ordering in the variables. In population genomic studies, such intervention based on curated gene groups will benefit the estimation since we can impose sparsity at the level of biological functions, not individual genes. This is an attractive solution given that the pairwise sample covariance terms not only represent the direct outcome of differential regulation of important biological pathways, but they also reflect the indirect consequences not associated with the regulation of pathways. In such situations, guiding the shrinkage estimation procedure by known gene grouping information can enable decomposition of the non-zero covariances due to the pathway-level regulation (systematic component) and the residual covariances (idiosyncratic component). This is the main motivation for our proposal to develop a novel procedure for cross-covariance matrix estimation.

2.2 Approach

As mentioned earlier, we analyze the DNA methylation and mRNA expression data from the invasive breast cancer cohort (BRCA) of TCGA. The TCGA-BRCA published in 2012 (Koboldt et al., 2012) has reported significant heterogeneity within and across different subtypes of breast cancer at multiple molecular levels. The single nucleotide polymorphism (SNP) arrays and whole exome sequencing revealed three highly common mutation-harboring genes such as TP53, PIK3CA, and GATA3, and highlighted numerous subtype-specific mutations. In addition, the availability of multiple -omics platform data and reverse-phase protein assay allowed them to discover additional subtypes within the previously defined luminal subtype (ER^+ and PR^+) and HER2 enriched subtype.

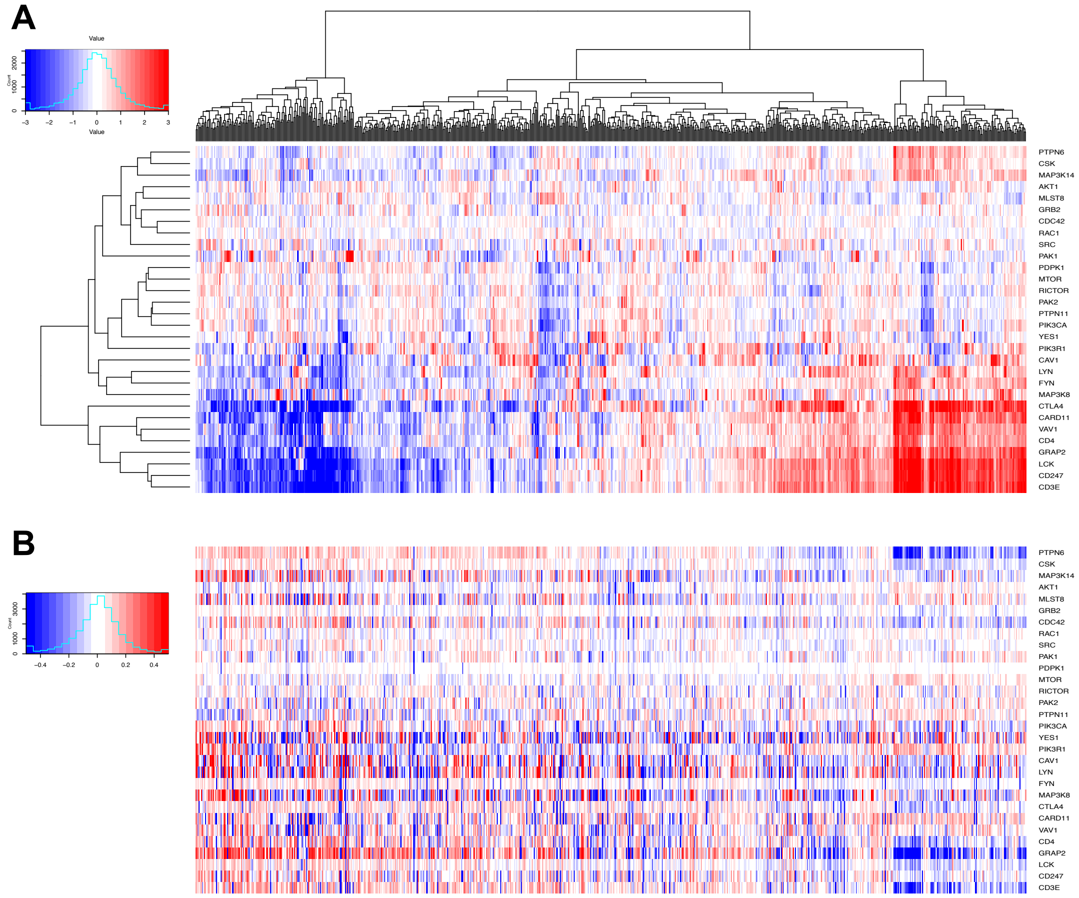

Figure 2.1 (A) mRNA transcript expression and (B) DNA methylation data for kinases and their substrate genes involved in T cell co-stimulation, which involves T cell receptors and mTORC2 complex. Rows and columns are ordered according to the hierarchical clustering of the mRNA data.

In our analysis, we aim to identify epigenetic regulation of mRNA transcript expression via DNA methylation at the pathway-level. Figure 2.1 shows the heat map of both molecular data for 30 genes in 766 primary tumor samples that meet two criteria: (i) each gene has to be either a kinase and kinase substrate (227 kinases and 764 non-kinase substrates) and (ii) it is involved in a particular biological process, T cell co-stimulation in this case. The diagram clearly shows that these genes are correlated at the mRNA transcript with a few exceptions (top), and the expression pattern is negatively correlated with the DNA methylation data (bottom). In other words, although there are other mechanisms of transcriptional regulation, DNA methylation seems to play an influential role in determining mRNA concentration levels for genes involved in T cell activation within the kinase signalling network.

To identify pathway-specific regulatory patterns of DNA methylation and their association with mRNA expression, we build our cross-covariance matrix between the two molecular data in two stages as follows. In the first stage, we detect pathway-specific factors in DNA methylation by estimating the variance-covariance matrix of the data. Following the approach of POET estimator (Fan et al., 2013), we decompose the matrix as the sum of a systematic component explained by pathway-specific factors and an idiosyncratic component representing the residuals. Both components are estimated with group penalty and element-wise penalty respectively, to yield sparse estimates. In the second stage, the gene group information from the pathways selected with non-zero factors in the first stage is used again to estimate the multivariate regression model between DNA methylation and mRNA expression with group lasso (Yuan and Lin, 2006), where we slightly modify the estimation procedure to allow overlapping groups. The two components naturally yield decomposition of the cross-covariance matrix into a systematic and an idiosyncratic one, with the former representing pathway-specific regulation of mRNA expression via DNA methylation.

Note that we forgo a straightforward element-by-element calculation of covariance between DNA methylation of one gene and mRNA of another gene (including self-to-self). This more cumbersome estimation routine is motivated by the ability of the two-stage estimation procedure to proactively utilize the existing gene group information to distinguish concerted regulation of gene expression at the pathway level. This decomposition will help us to determine how much of the cross covariance structure is a representation of coordinated biological functions and also to evaluate how complete or incomplete the existing knowledge of biological functions is. Moreover, the applicability of the same methodology is quite wide as far as integration of multiple molecular data is concerned.

CHAPTER 3

Estimation

Suppose the two sets of variables $\mathbf{X}_i$ and $\mathbf{Y}_i$, vectors of dimensions $(p \times 1)$ and $(q \times 1)$ respectively, are observed for subject $i = 1, \ldots, n$. We arrange them into their respective matrix form $\mathbf{X}_{(n \times p)}$ and $\mathbf{Y}_{(n \times q)}$ where the rows of the matrices correspond to the data for individual subjects. In other words,

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^{\mathrm{T}} \\ \mathbf{X}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{X}_n^{\mathrm{T}} \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1^{\mathrm{T}} \\ \mathbf{Y}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{Y}_n^{\mathrm{T}} \end{pmatrix}. \tag{3.1}$$

where the superscript T denotes transpose of a vector of a matrix. We aim to estimate the cross-covariance matrix $\mathbf{\Sigma_{XY}}$ as the product of two components, i.e. $\mathbf{\Sigma_{XX}B}$, and we thus construct the estimator for $\hat{\mathbf{\Sigma}}_{\mathbf{XX}}$ and

$\hat{\mathbf{B}}$ respectively. We assume that the observed data for subject i, denoted by $\left(\mathbf{X}_i, \mathbf{Y}_i\right)^{\mathrm{T}}$, follows the distribution with mean and variance

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{X}} \\ \boldsymbol{\mu}_{\mathbf{Y}} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{XX}} & \boldsymbol{\Sigma}_{\mathbf{XY}} \\ \boldsymbol{\Sigma}_{\mathbf{YX}} & \boldsymbol{\Sigma}_{\mathbf{YY}} \end{pmatrix}, \tag{3.2}$$

where $\mathbf{X}_i = \left(X_i^1, X_i^2, \ldots, X_i^p\right)^{\mathrm{T}}$ and $\mathbf{Y}_i = \left(Y_i^1, Y_i^2, \ldots, Y_i^q\right)^{\mathrm{T}}$. The mean vectors $\boldsymbol{\mu}_{\mathbf{X}}$, $\boldsymbol{\mu}_{\mathbf{Y}}$ and covariance matrices $\boldsymbol{\Sigma}_{\mathbf{XX}}$, $\boldsymbol{\Sigma}_{\mathbf{XY}}$, and $\boldsymbol{\Sigma}_{\mathbf{YY}}$ are defined as $(p \times p)$, $(p \times q)$, and $(q \times q)$ submatrices of $(p+q) \times (p+q)$ variance-covariance matrix $\boldsymbol{\Sigma}$. Further, we let $(p \times G)$ matrix $\mathbf{A} = \{a_{\ell g}\}$ indicating the group assignment of p variables into G groups, with $a_{\ell g} = 1$ if the variable ℓ is a member of group g and $a_{\ell g} = 0$ otherwise. Since the mean parameters can be estimated from the data, we assume zero mean hereafter without loss of generality.

Given that the variable grouping $\mathbf{A}$ is known, we assume that the elements of $\mathbf{X}_i$ are written as a linear combination of group specific latent factors $\{f_{ig}\}_{g=1}^{G}$, i.e.

$$\begin{aligned} \mathbf{X}_i = \begin{pmatrix} X_i^1 \\ X_i^2 \\ \vdots \\ X_i^p \end{pmatrix} &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1G} \\ a_{21} & a_{22} & \cdots & a_{2G} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pG} \end{pmatrix} \begin{pmatrix} f_{i1} \\ f_{i2} \\ \vdots \\ f_{iG} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{ip} \end{pmatrix} & (3.3) \\ &\equiv \mathbf{A}\mathbf{f}_i + \boldsymbol{\upsilon}_i. & (3.4) \end{aligned}$$

Rearranging the data for all subjects $i = 1, \ldots, n$, we can represent the entire model for $\mathbf{X}_i$ in a vector form:

$$\vec{\mathbf{X}} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A} \end{pmatrix} \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_n \end{pmatrix} + \begin{pmatrix} \boldsymbol{\upsilon}_1 \\ \boldsymbol{\upsilon}_2 \\ \vdots \\ \boldsymbol{\upsilon}_n \end{pmatrix} \tag{3.5}$$

$$= \mathbf{A}_n\mathbf{f} + \boldsymbol{\Upsilon}, \tag{3.6}$$

where $\mathbf{1}_{np}$ is $np-$ dimensional vector of ones, $\mathbf{A}_n$ is the block diagonal matrix consisting of n numbers of $(p \times G)$ matrix $\mathbf{A}$, and $\boldsymbol{\Upsilon}$ is $np-$dimensional distribution with mean $\mathbf{0}$ and covariance matrix as the block diagonal matrix of $\boldsymbol{\Sigma}_{\boldsymbol{\upsilon}}$. Then for all i, we can write the variance-covariance matrix of $\mathbf{X}$ as

$$\mathrm{Var}\big(\mathbf{X}_i\big) = \boldsymbol{\Sigma}_{\mathbf{XX}} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{f}}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\upsilon}}. \tag{3.7}$$

3.1 Estimation in Stage 1

The first stage estimation is finding a set of random factors $\mathbf{f}$ that minimizes

$$\tfrac{1}{2}\big\|\vec{\mathbf{X}} - \mathbf{A}_n\mathbf{f}\big\|_2^2 + \tfrac{1}{2}\tfrac{1}{\log(np)}\textstyle\sum_{k=1}^{G}\mathbf{f}_{.k}^2 + (np)\mathrm{P}_{\lambda_{\mathbf{f}}}\big(\mathbf{f}_{.k}\big) \tag{3.8}$$

where $\mathbf{f}_{.k} = \big(\sum_{i=1}^{n} f_{ik}^2\big)^{1/2}$ and the penalty term is defined as

$$\mathrm{P}_{\lambda_{\mathbf{f}}}\big(\mathbf{f}_{.k}\big) = \lambda_{\mathbf{f}}\mathbf{f}_{.k}. \tag{3.9}$$

Solving the above minimization problem with respect to $\mathbf{f}$ is equivalent to minimizing

$$\tfrac{1}{2}\big\|\vec{\mathbf{X}} - \mathbf{A}_n\mathbf{f}\big\|_2^2 + \tfrac{1}{2}\tfrac{1}{\log(np)}\Big\{\sum_{k=1}^{G}\mathbf{f}_{.k}^2 + 2(np)\cdot\log(np)\cdot\lambda_{\mathbf{f}}\sum_{k=1}^{G}\mathbf{f}_{.k}\big)\Big\}. \tag{3.10}$$

Since typically fewer than G factors will be retained as a result of penalized regression, we use G_0 to denote the number of selected random factors. After selecting G_0 random factors, we reiterate solving the least squares problem

$$\text{minimize}_{\ \mathbf{f}}\ \big\|\vec{\mathbf{X}} - \mathbf{A}_n\mathbf{f}\big\|_2^2 \tag{3.11}$$

to get the predictive value $\widehat{\mathbf{f}}$ of $\mathbf{f}$ to avoid extreme shrinkage and estimate $\mathbf{\Sigma}_{\mathbf{f}}$ by the sample covariance matrix of $\widehat{\mathbf{f}}_1, \widehat{\mathbf{f}}_2, \ldots, \widehat{\mathbf{f}}_n$:

$$\widehat{\mathbf{\Sigma}}_{\mathbf{f}} = \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{\mathbf{f}}_i - \overline{\widehat{\mathbf{f}}}\big)\big(\widehat{\mathbf{f}}_i - \overline{\widehat{\mathbf{f}}}\big). \tag{3.12}$$

where $\overline{\widehat{\mathbf{f}}}$ is the sample mean of $\widehat{\mathbf{f}}_1, \widehat{\mathbf{f}}_2, \ldots, \widehat{\mathbf{f}}_n$. If the factors are orthogonal, then $\mathbf{\Sigma}_{\mathbf{f}} = \mathrm{diag}\big(\tau_1^2, \tau_2^2, \cdots, \tau_{G_0}^2\big)$, we use $\mathrm{diag}\big(\widehat{\mathbf{\Sigma}}_{\mathbf{f}}\big)$ as an estimator of $\mathbf{\Sigma}_{\mathbf{f}}$. This completes the estimation of the systematic component $\mathbf{A}\mathbf{\Sigma}_{\mathbf{f}}\mathbf{A}^{\mathrm{T}}$ of $\mathbf{\Sigma}_{\mathbf{XX}}$.

To estimate the idiosyncratic component of $\boldsymbol{\Sigma}_{\boldsymbol{v}}$, we apply the generalized thresholding to the sample covariance matrix of $\vec{\mathbf{X}} - \mathbf{A}_n\widehat{\mathbf{f}}$. Various estimators are proposed in Antoniadis and Fan (2001), Rothman et al. (2009), and Cai and Liu (2011). In particular, we use the adaptive thresholding procedure of Cai and Liu (2011) using the sample covariance estimate using the residuals in the previous step, i.e.

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{v}} = s_{\lambda_{\boldsymbol{v}}}\left(\frac{1}{n - G_0 - 1}\sum_{i=1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}\right), \tag{3.13}$$

where $\boldsymbol{\xi}_i = \mathbf{X}_i - \mathbf{A}\widehat{\mathbf{f}}_i$ and $s_{\lambda_{\boldsymbol{v}}}(\mathbf{M})$ denotes the adaptive shrinkage operator of matrix $\mathbf{M}$.

3.2 Estimation of B by grouped lasso

The second stage of the estimation is fitting the regression model

$$\mathbf{Y}_i = \mathbf{B}^{\mathrm{T}}\mathbf{X}_i + \mathbf{u}_i \tag{3.14}$$

for $i = 1, \cdots, n$, where $\mathbf{Y}_i$ and $\mathbf{u}_i$ are q-dimensional vectors, and $\mathbf{B}$ is $p \times q$ matrix of coefficients. Note that, by rearranging $\{\mathbf{Y}_i\}_{i=1}^n$ and $\{\mathbf{u}_i\}_{i=1}^n$ into matrices of q columns and $\{\mathbf{X}_i\}_{i=1}^n$ into a matrix of p columns, both row-wise, we recover the original matrix notations $\mathbf{Y}$, $\mathbf{U}$, and $\mathbf{X}$ respectively and

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}. \tag{3.15}$$

Here $\mathbf{Y}$ is $(n \times q)$ matrix where each row is $\mathbf{Y}_i$, $\mathbf{U}$ is a matrix whose $i-$th column is $\mathbf{u}_i$ following the multivariate distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma_{YY}} - \mathbf{\Sigma_{YX}}\mathbf{\Sigma_{XX}}\mathbf{\Sigma_{XY}}$.

Although it is known that $\mathbf{B} = \mathbf{\Sigma_{XX}}^{-1}\mathbf{\Sigma_{XY}}$, here we consider the grouping information of $\mathbf{X}$ in the estimation of $\mathbf{B}$. To achieve this, we propose to formulate this as a grouped lasso regression (Peng et al., 2010; Yuan and Lin, 2006). The estimate of $\mathbf{B}$, say $\widehat{\mathbf{B}}$, is the solution to

$$\text{minimize}_{\mathbf{B}} \;\; \tfrac{1}{2}\big\|\mathbf{Y} - \mathbf{X}\mathbf{B}\big\|_2^2 + \mathrm{P}\big(\mathbf{B}\big), \tag{3.16}$$

where $\mathrm{P}\big(\mathbf{B}\big)$ is grouped lasso penalty. The group lasso penalty imposed in this regression comes from the matrix $\mathbf{A}$ used in the first stage estimation of factors, especially for the *selected* factors. Denote the index set of all the features of $\mathbf{X}$ belonging to a selected factor g by $\mathscr{G}_g$. Then the penalty term can be written as

$$\mathrm{P}\big(\mathbf{B}\big) = \lambda_{\mathbf{B}} \sum_{j=1}^{q} \sum_{g=1}^{G} \sqrt{p_g \sum_{\ell \in \mathscr{G}_g} \mathbf{b}_{\ell j}^2} \tag{3.17}$$

where $\mathbf{b}_{ij}$ denotes the elements of $\mathbf{B}$. Upon obtaining $\widehat{\mathbf{B}}$, the final estimate of $\mathbf{\Sigma_{XY}}$ becomes

$$\widehat{\mathbf{\Sigma}}_{\mathbf{XY}} := \widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\widehat{\mathbf{B}}. \tag{3.18}$$

3.3 Numerical Algorithm

Stage 1: Selection and Estimation of Factors

Without loss of generality, let us assume that each column of $\mathbf{X}$ is centered. Let $\rho = \frac{1}{\log{(np)}}$. The first stage aims to minimize

$$\frac{1}{2}\left\|\vec{\mathbf{X}} - \mathbf{A}_n\mathbf{f}\right\|_2^2 + \frac{1}{2}\rho\sum_{k=1}^{G}\mathbf{f}_{\cdot k}^2 + \lambda_{\mathbf{f}}\sum_{k=1}^{G}\mathbf{f}_{\cdot k} \tag{3.19}$$

with respect to $\{\mathbf{f}_i\}_{i=1}^n$. Following Zou and Hastie (2005), we merge the ridge penalty into the loss function by augmenting $\mathbf{A}_n$ with a diagonal matrix and $\mathbf{X}$ with a zero vector

$$\left\|\vec{\mathbf{X}} - \mathbf{A}_n\mathbf{f}\right\|_2^2 + \rho\sum_{k=1}^{G}\mathbf{f}_{\cdot k}^2 = \left\|\vec{\mathbf{X}}^* - \mathbf{A}^*\mathbf{f}\right\|_2^2 \tag{3.20}$$

where

$$\mathbf{A}^*_{(np+Gn)\times Gn} = \begin{pmatrix} \mathbf{A}_n \\ \\ \sqrt{\rho}\mathbf{I}_{Gn} \end{pmatrix}, \tag{3.21}$$

$$\vec{\mathbf{X}}^* = \begin{pmatrix} \vec{\mathbf{X}} \\ \\ \mathbf{0} \end{pmatrix}. \tag{3.22}$$

We can then restate the objective function as

$$\frac{1}{2}\left\|\vec{\mathbf{X}}^{*}-\mathbf{A}_{n}^{*}\mathbf{f}\right\|_{2}^{2}+\lambda_{\mathbf{f}}\sum_{k=1}^{G}\mathbf{f}_{\cdot k}=\frac{1}{2}\left\|\vec{\mathbf{X}}^{*}-\sum_{j=1}^{G}\mathbf{A}_{j}^{*}\mathbf{f}_{*,j}\right\|_{2}^{2}+\lambda_{\mathbf{f}}\sum_{k=1}^{G}\mathbf{f}_{\cdot k} \tag{3.23}$$

$$=\frac{1}{2}\left\|\vec{\mathbf{X}}^{*}-\sum_{j=1}^{G}\left(c_{j}\mathbf{A}_{j}^{*}\right)\left(\frac{\mathbf{f}_{*,j}}{c_{j}}\right)\right\|_{2}^{2}+\lambda_{\mathbf{f}}\sum_{k=1}^{G}\mathbf{f}_{\cdot k} \tag{3.24}$$

where

$$c_{j}=1/\sqrt{\left\|\mathbf{A}_{*,j}\right\|_{2}^{2}+\rho},$$

$$\mathbf{f}_{*,j}=\left(\mathbf{f}_{1j},\mathbf{f}_{2j},\ldots,\mathbf{f}_{nj}\right)^{\mathrm{T}},$$

$\mathbf{A}_{*,j}$ is the j^{th} column of $\mathbf{A}$, $\mathbf{A}_{j}$ is the j^{th} column of $\mathbf{A}_{n}$, and apply the pathwise coordinate optimization for the group lasso (Höfling and Tibshirani, 2007; Yuan and Lin, 2006). Define $\gamma_{j}=\mathbf{f}_{*,j}/c_{j}$. Then we update each γ_{j} by

$$\gamma_{j}\leftarrow\left(\left\|S_{j}\right\|_{2}-\lambda_{\mathbf{f}}\right)_{+}\left(\frac{S_{j}}{\left\|S_{j}\right\|_{2}}\right) \tag{3.25}$$

where

$$S_{j}=c_{j}\left(\mathbf{A}_{j}^{*}\right)^{\mathrm{T}}\left(\vec{\mathbf{X}}^{*}-\sum_{k\neq j}\left(c_{k}\mathbf{A}_{k}^{*}\right)\gamma_{j}\right), \tag{3.26}$$

c_{j} is set such that

$$\left(c_{j}\mathbf{A}_{j}^{*}\right)^{\mathrm{T}}\left(c_{j}\mathbf{A}_{j}^{*}\right)=\mathbf{1} \tag{3.27}$$

for all $j=1,\ldots,G$. Note that the formula for S_{j} can be simplified to avoid the use of $\mathbf{A}^{*}$ and $\vec{\mathbf{X}}^{*}$:

$$S_{j}=c_{j}\mathbf{A}_{j}^{\mathrm{T}}\left(\vec{\mathbf{X}}-\sum_{k\neq j}\left(c_{k}\mathbf{A}_{k}\right)\gamma_{j}\right). \tag{3.28}$$

To set the value of $\lambda_{\mathbf{f}}$, we use the Akaike Information Criterion (AIC) (Akaike, 1973) defined by $AIC = 2k - 2\ln(L)$, where k is the number of non-zero estimates of the parameters and L is the maximized value of the likelihood. If we denote the number of selected groups for each gene by G_i and SSE_i is the sum of squared residuals of gene i of the fitted model, then

$$AIC = 2n\left(\sum_{i=1}^{p} G_i\right) + n\sum_{i=1}^{p}\log\left(\text{SSE}_i/n\right) \tag{3.29}$$

When all coefficients are zero, γ_j will remain as zero after the update as shown in (3.25) if $\|S_j\|_2 = c_j \left\|\mathbf{A}_j^{\text{T}}\mathbf{X}\right\|_2 \le \lambda_{\mathbf{f}}$. Thus we let $\lambda_{\mathbf{f}} = \max_j \left(c_j \left\|\mathbf{A}_j^{\text{T}}\mathbf{X}\right\|_2\right)$.

Stage 2: Group lasso with overlapping groups

We first define the loss function and the penalty term as

$$\left\|\mathbf{Y} - \mathbf{XB}\right\|_2^2 + \text{P}\big(\mathbf{B}\big) \equiv \mathbf{L} + \mathbf{P}. \tag{3.30}$$

Note here that each row of $\mathbf{B}$ is independent of each other in the objective function, and thus the optimization can be carried out in parallel.

For the group lasso regression in the second stage, we perform pathwise coordinate gradient descent (Höfling and Tibshirani, 2007). Denote the matrix created by combining the columns of $\mathbf{A}$ selected in Stage 1 as $\mathbf{A}_0$. Each predictor variable with the same row pattern in $\mathbf{A}_0$ are classified

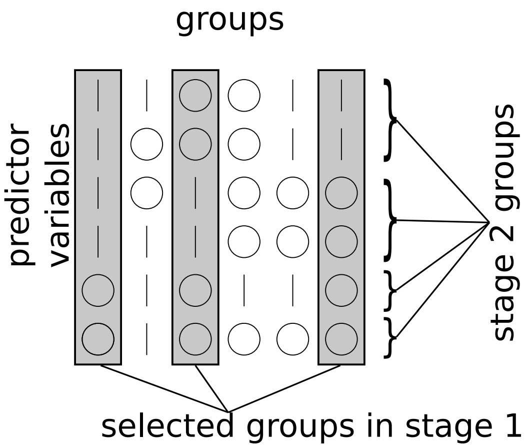

Figure 3.1 Grouping for stage 2. The 1's in matrix $\mathbf{A}$ represents a group membership.

as a group. However, predictor variable with a row pattern of all zeros are regarded as a singleton. Figure 3.1 illustrates the grouping for stage 2. We then partition $\mathbf{X}_i$ by its columns according to the gene groups, and orthonormalize each matrix $\mathbf{X}_i^g$ in the resulting set of matrix. The algorithm in Stage 1 is used to find $\mathbf{B}$. Let B_{jg} be the elements in group g in the j-th column of $\mathbf{B}$. Iteratively update the estimate of $\mathbf{B}$ with

$$B_{jg} \leftarrow \left(\left\|S_{jg}\right\|_2 - \lambda_{\mathbf{B}}\sqrt{p_g}\right)_+ \left(\frac{S_{jg}}{\left\|S_{jg}\right\|_2}\right) \tag{3.31}$$

where

$$S_{jg} = \sum_{i=1}^{n} X_i^{g\prime} \left(\mathbf{Y}_i - \sum_{k \neq g} X_i^k B_{jg}\right) \tag{3.32}$$

for each j. The optimal value of $\lambda_{\mathbf{B}}$ is obtained based on the AIC as in Stage 1.

CHAPTER 4

Theoretical Properties

In this section, we show that the proposed cross-covariance estimator $\widehat{\Sigma}_{\mathbf{XY}}$ converges to $\Sigma_{\mathbf{XY}}$ in the Frobenius norm and also show that both $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{f}}$ have consistency in selecting non-zero elements of $\mathbf{B}$ and $\mathbf{f}$.

4.1 Notations

We fist introduce necessary notations to be used in the theorems and their proofs. For easy comparison with the results in the literature, we rearrange the data $\mathbf{X}$ and $\mathbf{Y}$ introduced in Section 3 into variable forms of vectors or matrices wherever it is deemed to facilitate the statement and proof (especially for the two estimation stages). To begin with, we rewrite the model (3.14) for the second stage estimation using the vectorized notation for $\mathbf{Y}$ as follows. Denoting ℓ-th column of $\mathbf{B}$ by $\mathbf{b}_{.\ell}$,

$$\vec{\mathbf{Y}} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1q} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2q} \\ \vdots \\ \vdots \\ Y_{n1} \\ Y_{n2} \\ \vdots \\ Y_{nq} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1^{\mathrm{T}} & & & \\ & \mathbf{X}_1^{\mathrm{T}} & & \\ & & \ddots & \\ & & & \mathbf{X}_1^{\mathrm{T}} \\ \mathbf{X}_2^{\mathrm{T}} & & & \\ & \mathbf{X}_2^{\mathrm{T}} & & \\ & & \ddots & \\ & & & \mathbf{X}_2^{\mathrm{T}} \\ \ddots & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \ddots \\ \mathbf{X}_n^{\mathrm{T}} & & & \\ & \mathbf{X}_n^{\mathrm{T}} & & \\ & & \ddots & \\ & & & \mathbf{X}_n^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \mathbf{b}_{.1} \\ \mathbf{b}_{.2} \\ \vdots \\ \mathbf{b}_{.q} \end{pmatrix} + \begin{pmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1q} \\ u_{21} \\ u_{22} \\ \vdots \\ u_{2q} \\ \vdots \\ \vdots \\ u_{n1} \\ u_{n2} \\ \vdots \\ u_{nq} \end{pmatrix} \equiv \tilde{\mathbf{X}}\boldsymbol{\beta} + \vec{\mathbf{u}}; \tag{4.1}$$

where $\vec{\mathbf{Y}}$ is $(nq) \times 1$ vector, $\tilde{\mathbf{X}}$ is $(nq) \times (pq)$ matrix, $\boldsymbol{\beta}$ is the $(pq) \times 1$ vector of $\left(\mathbf{b}_{1.}^{\mathrm{T}}, \mathbf{b}_{2.}^{\mathrm{T}}, \ldots, \mathbf{b}_{q.}^{\mathrm{T}}\right)^{\mathrm{T}}$, and $\vec{\mathbf{u}} = \left(u_1, u_2, \ldots, u_{nq}\right)^{\mathrm{T}}$. In addition to the expanded form of the matrix $\tilde{\mathbf{X}}$, we recall the factor model for $\mathbf{X}$ in (3.5), which is the form of

$$\vec{\mathbf{X}} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} = \mathbf{A}_n \mathbf{f} + \boldsymbol{\Upsilon}, \tag{4.2}$$

where $\mathbf{1}_{np}$ is a $np-$dimensional vector of ones, $\mathbf{A}_n$ is the block diagonal matrix by n numbers of $\mathbf{A}$, and $\boldsymbol{\Upsilon}$ is $np-$dimensional distribution with mean 0 and covariance matrix as the block diagonal matrix of $\boldsymbol{\Sigma_\nu}$.

We next define the set of indexes of groups of variables. First, related with the regression model (4.1) with true $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, we let $\mathscr{H}_0 = \big\{g \mid \big\|\boldsymbol{\beta}_g^0\big\| \neq 0,\ g = 1, 2, \ldots, G\big\}$ and $G_0 = \big|\mathscr{H}_0\big|$, which are the index sets of groups with non-zero coefficients and its cardinality. Elementwise, we let $\mathscr{L} = \big\{j \mid 1 \leq j \leq pq\big\}$ and $\mathscr{L}_0 = \big\{j \in \mathscr{L} \big| \beta_j^0 \neq 0\big\}$. For the $g-$th group, $g = 1, 2, \ldots, G$, we let $\mathscr{I}_g$ be the set of indexes of its elements, $d_g = \big|\mathscr{I}_g\big|$, $\mathscr{J}_{g,h} = \mathscr{I}_g \bigcap \mathscr{I}_h$ and $d_{g,h} = \big|\mathscr{J}_{g1,g2}\big|$. Using these notations, the grouped lasso penalty is written as $\lambda_\beta \sum_{g=1}^G \lambda_g \big\|\boldsymbol{\beta}_g\big\|$, where β_g the vector of $\big\{\beta_j,\ j \in \mathscr{I}_g\big\}$ and the tuning parameter was renamed from $\lambda_{\mathbf{B}}$ to λ_β due to the rearrangement of regression coefficients into the vector form.

The element-wise index set A introduces $\tilde{\mathbf{X}}_A$, a submatrix with column vectors $\tilde{\mathbf{X}}_{\cdot j}$ for $j \in A$. This defines $\tilde{\mathbf{X}}_g$ for $g = 1, 2, \ldots, G$, and $\tilde{\mathbf{X}}_{g1,g2} = \big(\tilde{\mathbf{X}}_{\cdot j}, j \in \mathscr{I}_{g1} \bigcup \mathscr{I}_{g2}\big)$ for $g1, g2 \in \{1, 2, \ldots, G\}$. In particular, the set $\mathscr{L}_0$ partitions the design matrix $\tilde{\mathbf{X}}$ into $\tilde{\mathbf{X}}_0$ and $\tilde{\mathbf{X}}_1$, where $\tilde{\mathbf{X}}_0$ is an $(np) \times |\mathscr{L}_0|$

matrix with column vectors $\tilde{\mathbf{X}}_{.j}$ for $j \in \mathscr{L}_0$ and $\tilde{\mathbf{X}}_1$ is an $(np) \times \big(pq - |\mathscr{L}_0|\big)$ matrix with column vectors $\tilde{\mathbf{X}}_{.j}$, $j \in \mathscr{L}_0^c$.

For the factor model (4.2), we assume that, for $j = G_0+1, G_0+2, \ldots, G$, f_{ij} are degenerated to 0 for every $i = 1, 2, \ldots, n$. In the same way we defined $\mathscr{L}_0$, we let

$$\mathscr{I}_0 = \Big\{ j \mid j = (i-1) \cdot G + k,\ i = 1, 2, \ldots, n,\ k = 1, 2, \ldots, G_0 \Big\}, \quad (4.3)$$

which is the set (of indexes) of non-zero elements. Along with the given index sets $\mathscr{I}_0$, let $\mathbf{f}_{i.1} = \big(f_{i1}, f_{i2}, \ldots, f_{iG_0}\big)^{\mathrm{T}}$, $\mathbf{f}_{i.1+} = \big(\mathbf{f}_{i.1}^{\mathrm{T}}, \mathbf{0}_{1\times(G-G_0+1)}\big)^{\mathrm{T}}$, and $\mathbf{f}_0$ be the $(np) \times 1$ random vector whose j-th element is equal to that of $\mathbf{f}$ if $j \in \mathscr{I}_0$ and 0 otherwise. Thus,

$$\Sigma_{\mathbf{f}} = \begin{pmatrix} \Sigma_{\mathbf{f}.11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad (4.4)$$

where $\Sigma_{\mathbf{f}.11}$ is the $G_0 \times G_0$ sub-matrix of $\Sigma_{\mathbf{f}}$.

4.2 Consistency of $\widehat{\Sigma}_{\mathbf{XY}}$

Next, we show the convergence of $\widehat{\Sigma}_{\mathbf{XY}}$ to $\Sigma_{\mathbf{XY}}$ in the Frobenius norm. To show this, we first find

$$
\begin{aligned}
\left\|\widehat{\Sigma}_{\mathbf{XY}} - \Sigma_{\mathbf{XY}}\right\|_{\mathrm{F}} &= \left\|\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\widehat{\mathbf{B}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{B}\right\|_{\mathrm{F}} \\
&\leq \left\|\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\widehat{\mathbf{B}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\widehat{\mathbf{B}}\right\|_{\mathrm{F}} + \left\|\boldsymbol{\Sigma}_{\mathbf{XX}}\widehat{\mathbf{B}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{B}\right\|_{\mathrm{F}} \\
&\leq \left\|\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\right\|_{\mathrm{F}}\left\|\widehat{\mathbf{B}}\right\|_{\mathrm{F}} + \left\|\boldsymbol{\Sigma}_{\mathbf{XX}}\right\|_{\mathrm{F}}\left\|\widehat{\mathbf{B}} - \mathbf{B}\right\|_{\mathrm{F}} \\
&\leq p\left\{\left\|\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\right\|_{\mathrm{F}}\left\|\widehat{\mathbf{B}}\right\|_{\infty} + \left\|\boldsymbol{\Sigma}_{\mathbf{XX}}\right\|_{\infty}\left\|\widehat{\mathbf{B}} - \mathbf{B}\right\|_{\mathrm{F}}\right\} \\
&\leq p\left\{\left\|\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\right\|_{\mathrm{F}}\left\{\left\|\mathbf{B}\right\|_{\infty} + o_p(1)\right\} + \left\|\boldsymbol{\Sigma}_{\mathbf{XX}}\right\|_{\infty}\left\|\widehat{\mathbf{B}} - \mathbf{B}\right\|_{\mathrm{F}}\right\},
\end{aligned}
$$

where the infinity matrix norm for matrix $\mathbf{B} = (b_{ij})$ is defined as $\left|\mathbf{B}\right|_{\infty} := \max_{ij}|b_{ij}|$. Thus, we show the convergence of $\left\|\widehat{\mathbf{B}} - \mathbf{B}\right\|_{\mathrm{F}}$ (Theorem 1) and $\left\|\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\right\|_{\mathrm{F}}$ (Theorem 2) along with the boundedness assumptions on $\left|\mathbf{B}\right|_{\infty}$ and $\left|\boldsymbol{\Sigma}_{\mathbf{XX}}\right|_{\infty}$.

We first show the convergence of $\widehat{\mathbf{B}}$ which is the grouped lasso estimator of a multivariate multiple regression in the second stage. The asymptotic properties of the the grouped lasso estimator are studied by Nardi and Rinaldo (2008) when the groups are mutually exclusive. Theorem 1 below is a modification of Theorem 4.5 of Nardi and Rinaldo (2008) for the estimator allowing overlapping variable groups. Its proof mainly depends on the inequalities from Lemma 1 in the Appendix of Bunea et al. (2007).

The assumptions we make for the theorem are as follows. First, we assume $\tilde{u}_i$s are independent and identically distributed from the normal distribution with mean 0 and variance σ_u^2. Second, we assume the restricted eigenvalue (RE) condition which is known to be necessary for the ℓ_2 consistency of the estimator (Bickel et al., 2009). Define a set of sub-vectors of $\boldsymbol{\beta}$ which is

$$C\big(k, \alpha\big) = \big\{\boldsymbol{\beta} \in \mathscr{R}^{pq} : \big\|\boldsymbol{\beta}_{S^c}\big\| \le \alpha\big\|\boldsymbol{\beta}_S\big\|,\ \forall|S| = k\big\}. \tag{4.5}$$

The RE condition states that, for every $\boldsymbol{\beta} \in C\big(G_0, 3\big)$, there exists $\gamma > 0$ such that

$$\frac{1}{nq}\boldsymbol{\beta}^{\mathrm{T}}\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}\boldsymbol{\beta} = \frac{1}{nq}\big\|\tilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta}\big\|_2^2 \ge \gamma^2\big\|\boldsymbol{\beta}\big\|_2^2. \tag{4.6}$$

Along with the RE condition, one additional assumption is needed related with the overlapping groups. For some constant κ,

$$\max_{1\le j\le pq}\left\{\sum_{g:j\in\mathscr{I}_g}\lambda_g\right\} \le \kappa\min_g \lambda_g. \tag{4.7}$$

The last assumption we make is that

$$\min_g\left\{\frac{nq}{\sigma_u^2}\lambda^2\lambda_g - d_g\right\} - \log G \to \infty, \tag{4.8}$$

which is the assumption (A) of Nardi and Rinaldo (2008), assuring that the event

$$\mathbf{E}_1 = \bigcap_g\left\{\frac{2}{\sqrt{nq}}\big\|\tilde{\mathbf{X}}_g^{\mathrm{T}}\mathbf{u}\big\|_2 < \sqrt{nq}\,\lambda\,\lambda_g\right\} \tag{4.9}$$

occurs with probability tending to 1.

Now we state our Theorem 1 on the convergence of $\widehat{\mathbf{B}}$ to $\mathbf{B}$.

Theorem 1. Under the four assumptions above,

(1) if $\mathbf{E}_1$ satisfies, then

$$\big\|\widehat{\mathbf{B}} - \mathbf{B}^0\big\|_{\mathrm{F}}^2 = \big\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\big\|_2^2 \leq \frac{1}{\gamma^2}\Big\{16G_0\kappa^2\big(\min_g \lambda_g^2\big)\Big\} \cdot \lambda^2, \tag{4.10}$$

where $\mathbf{B}^0$ and $\boldsymbol{\beta}^0$ are the true values of $\mathbf{B}$ and $\boldsymbol{\beta}$.

(2) $\mathrm{P}\big(\mathbf{E}_1\big)$ converges to 1, as $n \to \infty$.

4.3 Proof of Theorem 1

Proof. The proof of the theorem is an extension of Theorem 4.5 of Nardi and Rinaldo (2008) to the grouped lasso estimator with overlapping groups. We briefly summarize their proof and explain the changes induced by allowing the overlapping groups. The main step of the proof is based on the following inequalities:

$$\begin{aligned}
\big\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\big\|_2 &\leq \frac{1}{\min_g \lambda_g} \sum_g \lambda_g \big\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g^0\big\|_2 \\
&\leq \frac{4}{\min_g \lambda_g} \sum_{g \in \mathscr{H}_0} \lambda_g \big\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g^0\big\|_2 && (4.11) \\
&\leq \frac{4}{\min_g \lambda_g} \sqrt{G_0} \sqrt{\sum_{g \in \mathscr{H}_0} \lambda_g^2 \big\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g^0\big\|_2^2}, && (4.12)
\end{aligned}$$

where the last is bounded from the RE condition to $\eta = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ (shown later).

The inequality (4.11) is obtained by modifying Lemma 1 of Bunea et al. (2007) for the grouped lasso penalty. The modified Lemma 1 (which is the same with Lemma 6.1 of Nardi and Rinaldo (2008)) states that, on the event $\mathbf{E}_1$, for any $\boldsymbol{\beta} \in \mathbb{R}^{pq}$ with $\mathscr{H}' = \{g \mid \boldsymbol{\beta}_g \neq 0\}$,

$$\begin{aligned}(1/n)\big\|\tilde{\mathbf{X}}\widehat{\boldsymbol{\beta}} - \tilde{\mathbf{X}}\boldsymbol{\beta}^0\big\|_2^2 + \lambda \sum_{g\in\mathscr{H}} \lambda_g \big\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\big\|_2 \\ \leq (1/n)\big\|\tilde{\mathbf{X}}\boldsymbol{\beta} - \tilde{\mathbf{X}}\boldsymbol{\beta}^0\big\|_2^2 + 4\lambda \sum_{g\in\mathscr{H}'} \lambda_g \big\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\big\|_2 \qquad (4.13)\end{aligned}$$

The extension to cases with overlapping groups can be achieved by replacing the lasso penalty with the grouped lasso penalty and using the following fact:

$$\begin{aligned}\frac{2}{nq}\sum_{i=1}^{nq} \tilde{u}_i \tilde{\mathbf{X}}_{i.}\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big) &= \frac{2}{nq}\sum_{i=1}^{nq} \tilde{u}_i \sum_{j=1}^{pq} X_{ij}\big(\widehat{\beta}_j - \beta_j\big) = \sum_{j=1}^{pq}\frac{2}{nq}\sum_{i=1}^{nq} X_{ij}\tilde{u}_i\big(\widehat{\beta}_j - \beta_j\big) \\ &\leq \sum_{g=1}^{G} \frac{2}{nq}\big\|\tilde{\mathbf{X}}_g^{\mathrm{T}}\vec{\mathbf{u}}\big\|_2 \big\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\big\|_2 \leq \lambda \sum_{g} \lambda_g \big\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\big\|_2.\end{aligned}$$

In the above equation, the first inequality is due to the followings:

$$\begin{aligned}\sum_{j=1}^{pq}\sum_{i=1}^{nq} X_{ij}\tilde{u}_i\big(\widehat{\beta}_j - \beta_j\big) &= \sum_{g=1}^{G}\sum_{j\in\mathscr{H}_g} \frac{1}{\big|\{g|j \in \mathscr{H}_g\}\big|}\sum_{i=1}^{nq} X_{ij}\tilde{u}_i\big(\widehat{\beta}_j - \beta_j\big) \\ &= \sum_{g=1}^{G}\left(\sum_{j\in\mathscr{H}_g} a_j^2\right)^{1/2} \big\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big\|_2,\end{aligned}$$

where

$$\sum_{j\in\mathscr{H}_g} a_j^2 \;=\; \sum_{j\in\mathscr{H}_g}\left\{\frac{1}{\left|\{g|j\in\mathscr{H}_g\}\right|^2}\left(\sum_{i=1}^{nq} X_{ij}\tilde{u}_i\right)^2\right\} \le \|\tilde{\mathbf{X}}_g^{\mathrm{T}}\vec{\mathbf{u}}\|_2^2.$$

The inequality in Equation (4.13) with $\boldsymbol{\beta} = \boldsymbol{\beta}^0$ gives the inequality (4.11) and also

$$\sum_{g\in\left(\mathscr{H}_0\right)^c} \lambda_g\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\|_2 \le 3\sum_{g\in\mathscr{H}_0}\lambda_g\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g^0\|_2. \tag{4.14}$$

The inequality (4.12) is obtained by applying Cauchy-Schwartz inequality to (4.11) and further we have another bound for (4.13) as

$$\begin{aligned}&\big(1/n\big)\|\tilde{\mathbf{X}}\widehat{\boldsymbol{\beta}} - \tilde{\mathbf{X}}\boldsymbol{\beta}^0\|_2^2 + \lambda\sum_g \lambda_g\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\|_2\\ &\quad\le \big(1/n\big)\|\tilde{\mathbf{X}}\boldsymbol{\beta} - \tilde{\mathbf{X}}\boldsymbol{\beta}^0\|_2^2 + 4\lambda\sqrt{|\mathscr{H}'|}\sqrt{\sum_{g\in\mathscr{H}'}\lambda_g^2\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\|_2^2}.\end{aligned}$$

Again by substituting $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, we have

$$\begin{aligned}\big(1/n\big)\|\tilde{\mathbf{X}}\widehat{\boldsymbol{\beta}} - \tilde{\mathbf{X}}\boldsymbol{\beta}^0\|_2^2 &\le 4\lambda\sqrt{G_0}\sqrt{\sum_{g\in\mathscr{H}_0}\lambda_g^2\|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g^0\|_2^2}\\ &\le 4\lambda\sqrt{G_0}\kappa\big(\min_g\lambda_g\big)\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 \qquad (4.15)\end{aligned}$$

which is obtained by noting that

$$\begin{aligned}\sum_{g\in\mathscr{H}_0}\lambda_g^2\big\|\widehat{\boldsymbol{\beta}}_g-\boldsymbol{\beta}_g\big\|_2^2 &\leq \Big(\sum_{g\in\mathscr{H}_0}\lambda_g^2\Big)\sum_{g\in\mathscr{H}_0}\big\|\widehat{\boldsymbol{\beta}}_g-\boldsymbol{\beta}_g\big\|_2^2\\ &\leq \Big(\sum_{g\in\mathscr{H}_0}\lambda_g\Big)^2\sum_{g\in\mathscr{H}_0}\big\|\widehat{\boldsymbol{\beta}}_g-\boldsymbol{\beta}_g\big\|_2^2 \leq \kappa^2\big(\min_g\lambda_g\big)^2\sum_{g\in\mathscr{H}_0}\big\|\widehat{\boldsymbol{\beta}}_g-\boldsymbol{\beta}_g\big\|_2^2.\end{aligned}$$

Finally, the inequalities (4.14) and (4.15) and the RE condition for $\eta=\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0$ provides an upper bound of $\big\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\big\|_2^2$, which is

$$\frac{1}{\gamma^4}\Big\{16G_0\kappa^2\big(\min_g\lambda_g^2\big)\Big\}\cdot\lambda^2. \qquad (4.16)$$

This completes the proof. □

In showing the convergence of $\widehat{\Sigma}_{\mathbf{XY}}$, we need an additional result for $\widehat{\boldsymbol{\beta}}$ (or $\widehat{\mathbf{B}}$) that is $\big\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\big\|_\infty=\big|\widehat{\mathbf{B}}-\mathbf{B}\big|_\infty=o_p(1)$. This is a byproduct of Theorem 1 because, on the event $\mathbf{E}_1$, we have

$$\begin{aligned}\big|\widehat{\mathbf{B}}-\mathbf{B}\big|_\infty^2 &= \big\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big\|_\infty^2\\ &\leq \big\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big\|_2^2=\frac{1}{\gamma^4}\Big\{16G_0\kappa^2\big(\min_g\lambda_g^2\big)\Big\}\cdot\lambda^2,\end{aligned}$$

which converges to 0.

We next prove the convergence of $\big\|\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}-\boldsymbol{\Sigma}_{\mathbf{XX}}\big\|_{\mathrm{F}}$, which is equivalent to the convergences of $\big\|\widehat{\Sigma}_{\mathbf{f}}-\Sigma_{\mathbf{f}}\big\|_{\mathrm{F}}$ and $\big\|\widehat{\Sigma}_{\boldsymbol{\nu}}-\Sigma_{\boldsymbol{\nu}}\big\|_{\mathrm{F}}$. To show the convergence, we need a few lemmas, which are stated below with proofs provided.

Recall that we estimate $\widehat{\mathbf{f}}_i$, $i = 1, 2, \ldots, n$, by solving

$$\frac{1}{2}\big\|\vec{\mathbf{X}} - \mathbf{A}_n\mathbf{f}\big\|_2^2 + \frac{1}{2\log(np)}\sum_{k=1}^{G}\mathbf{f}_{.k}^2 + \big(np\big)\lambda\sum_{k=1}^{G}\mathbf{f}_{.k}, \tag{4.17}$$

where $\mathbf{f}_{.k} = \sum_{i=1}^{n} f_{ik}^2$. In addition, let $\tilde{\mathbf{f}}$ be the solution to the oracle procedure which is

$$\frac{1}{2}\big\|\vec{\mathbf{X}} - \mathbf{A}_n\mathbf{f}_0\big\|_2^2 + \frac{1}{2\log(np)}\sum_{k=1}^{G_0}\mathbf{f}_{.k}^2. \tag{4.18}$$

Suppose that we let $\tilde{\mathbf{f}} = \big(\tilde{\mathbf{f}}_1^{\mathrm{T}}, \tilde{\mathbf{f}}_2^{\mathrm{T}}, \ldots, \tilde{\mathbf{f}}_n^{\mathrm{T}}\big)^{\mathrm{T}}$ and $\tilde{\mathbf{f}}_1 = \big(\tilde{\mathbf{f}}_{1.1}^{\mathrm{T}}, \tilde{\mathbf{f}}_{2.1}^{\mathrm{T}}, \ldots, \tilde{\mathbf{f}}_{n.1}^{\mathrm{T}}\big)^{\mathrm{T}}$, where $\tilde{\mathbf{f}}_i = \big(\tilde{f}_{i1}, \tilde{f}_{i2}, \ldots, \tilde{f}_{iG}\big)^{\mathrm{T}}$ and $\tilde{\mathbf{f}}_{i.1} = \big(\tilde{f}_{i1}, \tilde{f}_{i2}, \ldots, \tilde{f}_{iG_0}\big)^{\mathrm{T}}$. The first lemma is that the predicted value of $\mathbf{f}_0$, particularly $\tilde{\mathbf{f}}_1 = \big(\tilde{\mathbf{f}}_{1.1}^{\mathrm{T}}, \tilde{\mathbf{f}}_{2.1}^{\mathrm{T}}, \ldots, \tilde{\mathbf{f}}_{n.1}^{\mathrm{T}}\big)^{\mathrm{T}}$. We show that its sample variance consistently estimate $\Sigma_{\mathbf{f}.11}$ which defines $\boldsymbol{\Sigma}_{\mathbf{XX}}$.

Lemma 1. Under the assumption that $\big(1/n\big)\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}$ converges in probability to $\mathbf{A}\Sigma_{\mathbf{f}}\mathbf{A}^{\mathrm{T}} + \Sigma_{\Upsilon}$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{f}}_{i.1}\tilde{\mathbf{f}}_{i.1}^{\mathrm{T}} \text{ and } \frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{f}}_i\tilde{\mathbf{f}}_i^{\mathrm{T}} \to \Sigma_{\mathbf{f}.11} \text{ and } \Sigma_{\mathbf{f}} \text{ in probability,} \tag{4.19}$$

respectively.

Proof of Lemma 1

Proof. Let $\mathbf{A}_0$ be the matrix by the first G_0 columns of the $p \times G$ matrix $\mathbf{A}$ corresponding to $\mathbf{f}_{i.1}$. The proof is simply by noting

$$\tilde{\mathbf{f}}_{i.1} = \left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0 + c_n \mathrm{I}_{G_0}\right)^{-1}\mathbf{A}_0^{\mathrm{T}}\mathbf{X}_i, \tag{4.20}$$

where $c_n = 1\big/\log(np)$. Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{f}}_{i.1}\tilde{\mathbf{f}}_{i.1}^{\mathrm{T}} = \left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0 + c_n \mathrm{I}_{G_0}\right)^{-1}\mathbf{A}_0^{\mathrm{T}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)\mathbf{A}_0\left(\mathbf{A}_0^{\mathrm{T}}A_0 + c_n \mathrm{I}_{G_0}\right)^{-1} \tag{4.21}$$

has the limit as

$$\begin{aligned}
&\left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0\right)^{-1}\mathbf{A}_0^{\mathrm{T}}\left(\mathbf{A}\Sigma_{\mathbf{f}}\mathbf{A}^{\mathrm{T}} + \Sigma_{\boldsymbol{\nu}}\right)\mathbf{A}_0\left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0\right)^{-1} \\
&= \left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0\right)^{-1}\mathbf{A}_0^{\mathrm{T}}\left(\mathbf{A}_0\Sigma_{\mathbf{f.11}}\mathbf{A}_0^{\mathrm{T}} + \Sigma_{\boldsymbol{\nu}}\right)\mathbf{A}_0\left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0\right)^{-1} \\
&= \Sigma_{\mathbf{f.11}} + \left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0\right)^{-1}\mathbf{A}_0^{\mathrm{T}}\Sigma_{\boldsymbol{\nu}}\mathbf{A}_0\left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0\right)^{-1},
\end{aligned}$$

where $\left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0\right)^{-1}\mathbf{A}_0^{\mathrm{T}}\Sigma_{\boldsymbol{\nu}}\mathbf{A}_0\left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0\right)^{-1}$ is the variance of $\left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0\right)^{-1}\mathbf{A}_0^{\mathrm{T}}\boldsymbol{\xi}$. Here, the term $\left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0\right)^{-1}\mathbf{A}_0^{\mathrm{T}}\boldsymbol{\xi}$ converges 0 almost surely as p increases by the strong law of large numbers. Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{f}}_{i.1}\tilde{\mathbf{f}}_{i.1}^{\mathrm{T}}$$

converges to $\Sigma_{\mathbf{f}.11}$ in probability. □

The second lemma is an analogy of Theorem 3 of Fan and Li (2012), which shows the consistency of $\widehat{\mathbf{f}}$ in selecting non-zero random factors from $\mathbf{f}_{.1}, \mathbf{f}_{.2}, \ldots, \mathbf{f}_{.G}$.

Lemma 2. Under the assumption that:

(i) The smallest and the largest eigenvalues of $\left(1/p\right)\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0$ and $\Sigma_{\mathbf{f}.11}$ are bounded from below and above, respectively.

(ii) The grouping matrices $\mathbf{A}$ $(p \times G)$ and $\mathbf{A}_0$ $(p \times G_0)$ satisfy

$$\left\| \left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0 + \frac{1}{\log(np)} \mathrm{I}_{G_0} \right)^{-1} \right\|_{\infty} \leq \frac{\sqrt{n}}{\left(np\right)^{1+\delta}\lambda} \tag{4.22}$$

for $\delta \in (0, 1/2)$, and

$$\max_{G_0+1 \leq k \leq G} \left\| \mathbf{a}_{.k}^{\mathrm{T}}\mathbf{A}_0 \left(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0 + \frac{1}{\log(np)} \mathrm{I}_{G_0} \right)^{-1} \right\|_2 < 1, \tag{4.23}$$

where $\mathbf{a}_{.k}$ is the $k-$th column vector of $\mathbf{A}$.

(iii) As both $\max\left\{\mathrm{diag}\left(\Sigma_{\mathbf{f}.11}\right)\right\}\left(np\right)^{\delta}/\sqrt{n}$ and $\lambda^2 np \Big/ \left\{G_0(\log n)\right\}$ increases to ∞ with $n \to \infty$,

we have, with probability tending to 1,

$$\Big\{ j \big| j = (i-1)\cdot G + k,\ \widehat{f}_{ik} \neq 0, i = 1, 2, \ldots, n, k = 1, 2, \ldots, G \Big\} = \mathscr{J}_0 \tag{4.24}$$

and

$$\max_{1 \leq k \leq G_0} \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{f}_{ik} - \tilde{f}_{ik}\right)^2 \leq n^{-\delta}, \tag{4.25}$$

for δ defined in the assumption (ii).

Proof of Lemma 2

Proof. The proof of the lemma is a direct application of Theorem 3 of Fan and Li (2012). The theorem requires the regularity conditions on (1) the penalty function, (2) the eigenvalues of the matrices related with the design matrix $\mathbf{A}$, and (3) the conditions for λ, n, and p. In this section, we use the grouped lasso regression where its penalty function automatically satisfies (1). Thus, we do not need it. The conditions (i) and (ii) of our lemma are the simplified version of the condition for $\mathbf{A}_n$ required by Condition 3 of Fan and Li (2012). The simplification is done using the fact that $\mathbf{A}_n$ is the Kronecker product of the $p \times G$ matrix $\mathbf{A}$ and the $n \times n$ identity matrix I_n [1]. Finally, Fan and Li (2012) requires three condition for the samples sizes (n and p) and λ. With the notations of our problem, they are (a) $\max\left\{\operatorname{diag}\left(\Sigma_{\mathbf{f}.11}\right)\right\}\left(np\right)^{\delta}\Big/\sqrt{n} \to \infty$, (b) $\lambda^2 np\Big/\left\{G_0(\log n)\right\} \to$ and (c) $\lambda^2 np\Big/G_0 \to \infty$. Here, (b) implies (c) and the condition (iii) of the lemma are equivalent to (a) and (b). □

[1] To facilitate the interpretation, we provide matching notations between ours and Fan and Li (2012). The feft is the notation in Fan and Li and the right is the notation of ours.

$$\begin{aligned} n = \sum_{i=1}^{N} n_i &= np \\ N &= n \\ n_i &= p \quad \text{for all } i \\ m_n = \max_{i=1}^{N} n_i &= p \\ s_{2n} &= G_0 \\ q_n &= G. \end{aligned}$$

As shown in Lemma 2, since the probability of the event

$$\mathbf{E}_2 = \left\{ \left\{ j \middle| \widehat{f}_j \neq 0,\ 1 \leq j \leq np \right\} = \mathscr{J}_0 \right\} \tag{4.26}$$

approaches 1, we restrict our discussion to the event $\mathbf{E}_2$ below. In addition, for notational simplicity, we let $G_0 = G$, $\mathbf{A} = \mathbf{A}_0$, $\widehat{\mathbf{f}} = \tilde{\mathbf{f}}$, $\widehat{\mathbf{f}}_i = \tilde{\mathbf{f}}_i = \tilde{\mathbf{f}}_{i.1}$ and $\Sigma_{\mathbf{f}} = \Sigma_{\mathbf{f}.11}$. With this simplified notations, the following lemma (Lemma 3) provides a representation to $\widehat{\mathbf{f}}$ which plays a key role in showing the convergence of $\mathbf{\Sigma_{XX}}$.

Lemma 3. On the event $\mathbf{E}_2$, for $i = 1, 2, \ldots, n$, we have the identity

$$\widehat{\mathbf{f}}_i - \mathbf{f}_i = \left\{ \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \frac{1}{\log(np)} \mathrm{I}_G \right)^{-1} \mathbf{A}^{\mathrm{T}}\mathbf{A} - \mathrm{I}_G \right\} \mathbf{f}_i + \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \frac{1}{\log(np)} \mathrm{I}_G \right)^{-1} \mathbf{A}^{\mathrm{T}} \boldsymbol{\xi}_i. \tag{4.27}$$

and

$$\left\| \widehat{\mathbf{f}}_i - \mathbf{f}_i \right\|_2^2 = O_p\left(\frac{1}{p} \right) \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \left\| \widehat{\mathbf{f}}_i - \mathbf{f}_i \right\|_2^2 = O_p\left(\frac{1}{\sqrt{np}} \right). \tag{4.28}$$

Proof of Lemma 3

Proof. The identity (4.27) is straightforward by the definition of $\widehat{\mathbf{f}}_i$. In addition, the first term of the identity

$$\left\{ \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \frac{1}{\log(np)}\mathrm{I}_G \right)^{-1} \mathbf{A}^{\mathrm{T}}\mathbf{A} - \mathrm{I}_G \right\} \mathbf{f}_i$$

$$= -\frac{1}{\log(np)} \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \frac{1}{\log(np)}\mathrm{I}_G \right)^{-1} \mathbf{f}_i = O_p\left(\frac{1}{\log(np)\cdot p} \right),$$

where the order is obtained by observing $\frac{1}{p}\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0$ has the limit whose smallest and the largest eigenvalues are bounded below and the above. The second term is

$$\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \frac{1}{\log(np)}\mathrm{I}_G \right)^{-1} \mathbf{A}^{\mathrm{T}}\boldsymbol{\xi}_i = O_p\left(\frac{1}{\sqrt{p}} \right),$$

as n and p increases. The order of the second term is again from the assumption of the eigenvalues of $\big(1/p\big)\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0$ and the central limit theorem on $\frac{1}{\sqrt{p}}\mathbf{A}^{\mathrm{T}}\boldsymbol{\xi}_i$. Therefore, $\left\|\widehat{\mathbf{f}}_i - \mathbf{f}_i\right\|_2^2 = O_p\big(1/p\big)$ and similarly we have

$$\frac{1}{n}\sum_{i=1}^{n} \left\|\widehat{\mathbf{f}}_i - \mathbf{f}_i\right\|_2^2 = O_p\left(\frac{1}{\sqrt{np}} \right).$$

□

The theorem below (Theorem 2) is analogous to Theorem 5 of Fan et al. (2013), which shows the convergence of $\widehat{\Sigma}_\Upsilon$ to Σ_Υ in the Frobenius norm. Recall that $\widehat{\Sigma}_\Upsilon$ is the general thresholding (GT) (Rothman et al., 2009) of the sample covariance matrix of residuals

$$\widehat{\boldsymbol{\xi}}_i = \mathbf{X}_i - \mathbf{A}\widehat{\mathbf{f}}_i, \qquad i = 1, 2, \ldots, n.$$

Theorem 5 of Fan et al. (2013) states that, if the above residuals $\widehat{\boldsymbol{\xi}}_i$ are close to $\boldsymbol{\xi}_i$ in the sense that $\max_{1\leq j\leq G}\sum_{i=1}^{n}\left\|\widehat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i\right\|_2^2 = o_p\left(a_n^2\right)$ with $a_n = o(1)$ and $\max_{1\leq i\leq n, 1\leq j\leq G}\left|\widehat{\xi}_{ij} - \xi_{ij}\right| = o_p(1)$, the GT of the sample covariance matrix, denoted by $\widehat{\Sigma}_{\Upsilon}$ converges to the covariance matrix of $\boldsymbol{\xi}_i$, denoted by Σ_{Υ}, in the spectral norm under regularity conditions on Σ_{Υ}, ξ_{ij}, and f_{ij}. The theorem below shows that $\widehat{\boldsymbol{\xi}}_i$ are close enough to $\boldsymbol{\xi}_i$ to show the convergence of $\widehat{\Sigma}_{\Upsilon}$ in both the spectral and the Frobenius norm. We remark that we assume the adaptive thresholding by Cai and Liu (2011), not that in Fan et al. (2013). As claimed by Fan et al. (2013) , Theorem 5 and other results in their section 3.2 are still true for the adaptive thresholding by Cai and Liu (2011). The difference between two are from thresholding value w_n; Cai and Liu (2011) assumes $w_n = \sqrt{\frac{\log p}{n}}$, whereas Fan et al. (2013) uses $w_n = \sqrt{\frac{1}{p}} + \sqrt{\frac{\log p}{n}}$. Here, we use the version of Theorem 5 of Fan et al. (2013) for $w_n = \sqrt{\frac{\log p}{n}}$ without repeating their proof.

Theorem 2. Suppose we assume that:

(i) The smallest and the largest eigenvalues of $\Sigma_{\boldsymbol{\nu}}$ is bounded by positive constants $C_{\min,\boldsymbol{\nu}}$ and $C_{\max,\boldsymbol{\nu}}$ from below and above, respectively.

(ii) The variables ϵ_{ij} and f_{ij} satisfy the sub-Gaussianality in the sense that there are $r_1, r_2 > 0$ and $b_1, b_2 > 0$ such that

$$\mathrm{P}\big(|\epsilon_{ij}| > s\big) \leq \exp\big\{-\big(s/b_1\big)^{r_1}\big\} \quad \text{and} \quad \mathrm{P}\big(|f_{ij}| > s\big) \leq \exp\big\{-(s/b_2)^{r_2}\big\}. \tag{4.29}$$

Then, on the event $\mathbf{E}_2$, we have

$$\big\|\widehat{\Sigma}_{\boldsymbol{v}} - \Sigma_{\boldsymbol{v}}\big\|_F^2 = O_p\left(\frac{p \log p}{n} \cdot C_p^2\right), \tag{4.30}$$

where $C_p = \max_{1 \leq J \leq p} \big|\big\{1 \leq k \leq p \mid \mathrm{cov}\big(\xi_{ij}, \xi_{ik}\big) \neq 0\big\}\big|$.

Proof of Theorem 2

Proof. The proof of the theorem is mainly same with that of Theorem 5 of Fan et al. (2013). We simply show that $\widehat{\boldsymbol{\xi}}_i$ are close to $\boldsymbol{\xi}_i$. From the identity in Lemma 3, we have

$$\widehat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i = \mathbf{A}\big(\mathbf{f}_i - \widehat{\mathbf{f}}_i\big)$$

and thus

$$\frac{1}{n}\sum_{i=1}^{n} \big\|\widehat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i\big\|_2^2 \leq \big\|\mathbf{A}\big\|_{\mathrm{F}}^2 \frac{1}{n}\sum_{i=1}^{n} \big\|\widehat{\mathbf{f}}_i - \mathbf{f}_i\big\|_2^2 = O_p\left(\frac{1}{\sqrt{n}}\right).$$

On the other hand,

$$\widehat{\xi}_{ij} - \xi_{ij} = -\mathbf{a}_{j.}\big(\widehat{\mathbf{f}}_i - \mathbf{f}_i\big),$$

where $\mathbf{a}_{j\cdot}$ is the j-th row vector of the matrix $\mathbf{A}$. Using the Cauchy-Schwarz inequality,

$$\left(\widehat{\xi}_{ij} - \xi_{ij}\right)^2 \leq \|\mathbf{a}_{j\cdot}\|_2^2 \|\widehat{\mathbf{f}}_i - \mathbf{f}_i\|_2^2$$

and

$$\max_{1 \leq j \leq p} \left(\widehat{\xi}_{ij} - \xi_{ij}\right)^2 \leq \|\widehat{\mathbf{f}}_i - \mathbf{f}_i\|_2^2 \max_{1 \leq j \leq p} \|\mathbf{a}_{j\cdot}\|_2^2 = O_p\left(\frac{1}{p}\right) = o_p(1),$$

as both n and p increases.

We next find that

$$\begin{aligned}
\|\widehat{\Sigma}_{\boldsymbol{\nu}} - \Sigma_{\boldsymbol{\nu}}\|_{\mathrm{F}}^2 &= \sum_{j=1}^{p}\sum_{k=1}^{p} \left(\widehat{\Sigma}_{\boldsymbol{\nu}}(j,k) - \Sigma_{\boldsymbol{\nu}}(j,k)\right)^2 \\
&\leq p \max_{1 \leq j \leq p} \sum_{k=1}^{p} \left(\widehat{\Sigma}_{\boldsymbol{\nu}}(j,k) - \Sigma_{\boldsymbol{\nu}}(j,k)\right)^2,
\end{aligned}$$

which, using the same argument of the proof of Theorem 5 of Fan et al. (2013), is an order of

$$O_p\Big(p \cdot C_p^2 w_n^2\Big) = O_p\left(\frac{p \log p}{n} \cdot C_p^2\right).$$

□

Finally, in Theorem 3, we prove the convergence of $\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}$ to $\boldsymbol{\Sigma}_{\mathbf{XX}}$ in the Frobenius norm. This, together with Theorem 2, show the convergence of $\widehat{\Sigma}_{\mathbf{XY}}$ to $\Sigma_{\mathbf{XY}}$ in the Frobenius norm.

Theorem 3. Under the assumptions of Lemma 2 and Theorem 2, on the set $\mathbf{E}_2$, we have

$$\left\|\hat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\right\|_{\mathrm{F}}^2 = O_p\left(\frac{p\log p}{n}C_p^2\right). \tag{4.31}$$

Proof of Theorem 3

Proof. First, we have

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{f}_i\mathbf{f}_i^{\mathrm{T}} - \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{f}}_i\widehat{\mathbf{f}}_i^{\mathrm{T}}\right\|_{\mathrm{F}} \\
&\quad\le \left\|\frac{1}{n}\sum_{i=1}^n \left(\mathbf{f}_i - \widehat{\mathbf{f}}_i\right)\mathbf{f}_i^{\mathrm{T}}\right\|_{\mathrm{F}} + \left\|\frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{f}}_i\left(\widehat{\mathbf{f}}_i^{\mathrm{T}} - \mathbf{f}_i^{\mathrm{T}}\right)\right\|_{\mathrm{F}} \\
&\quad\le \left(\frac{1}{n}\sum_{i=1}^n \left\|\mathbf{f}_i - \widehat{\mathbf{f}}_i\right\|_2^2 \cdot \frac{1}{n}\sum_{i=1}^n \left\|\mathbf{f}_i\right\|_2^2\right)^{1/2} + \left(\frac{1}{n}\sum_{i=1}^n \left\|\mathbf{f}_i - \widehat{\mathbf{f}}_i\right\|_2^2 \cdot \frac{1}{n}\sum_{i=1}^n \left\|\widehat{\mathbf{f}}_i\right\|_2^2\right)^{1/2} \\
&\quad= O_p\left(\frac{1}{\sqrt{np}}\right).
\end{aligned}$$

Using the above and Theorem 2, we have

$$\begin{aligned}
\left\|\hat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\right\|_{\mathrm{F}} &= \left\|\frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{f}}_i\widehat{\mathbf{f}}_i^{\mathrm{T}} - \Sigma_{\mathbf{f}} + \widehat{\Sigma}_{\boldsymbol{\nu}} - \Sigma_{\boldsymbol{\nu}}\right\|_{\mathrm{F}} \\
&\le \left\|\frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{f}}_i\widehat{\mathbf{f}}_i^{\mathrm{T}} - \Sigma_{\mathbf{f}}\right\|_{\mathrm{F}} + \left\|\widehat{\Sigma}_{\boldsymbol{\nu}} - \Sigma_{\boldsymbol{\nu}}\right\|_{\mathrm{F}} \\
&\le \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{f}_i\mathbf{f}_i^{\mathrm{T}} - \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{f}}_i\widehat{\mathbf{f}}_i^{\mathrm{T}}\right\|_{\mathrm{F}} + \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{f}_i\mathbf{f}_i^{\mathrm{T}} - \Sigma_{\mathbf{f}}\right\|_{\mathrm{F}} + \left\|\widehat{\Sigma}_{\boldsymbol{\nu}} - \Sigma_{\boldsymbol{\nu}}\right\|_{\mathrm{F}} \\
&= O_p\left(\frac{1}{\sqrt{np}} + \frac{1}{\sqrt{n}} + \sqrt{\frac{p\log p}{n}}\cdot C_p\right) = O_p\left(\sqrt{\frac{p\log p}{n}}\cdot C_p\right)
\end{aligned}$$

and this completes the proof. □

4.4 Selection consistency of $\widehat{\mathbf{f}}$ and $\widehat{\mathbf{B}}$

The consistency of $\widehat{f}$ in selecting degenerated elements in $\mathbf{f}$ is shown in Lemma 2. In this section, we only focus on the selection consistency of $\widehat{\mathbf{B}}$. We make two assumptions to show the result. First, we assume the smallest and the largest eigenvalues of $\big(1/np\big)\big(\tilde{\mathbf{X}}_0^{\mathrm{T}}\tilde{\mathbf{X}}_0\big)$ are bounded below and above by two positive constants $C_{\min,\tilde{\mathbf{X}}_0}$ and $C_{\max,\tilde{\mathbf{X}}_0}$, respectively. Second, we assume that

$$\max_{j\in\mathscr{L}_0^c}\Big\|\mathbf{X}_{\cdot j}^{\mathrm{T}}\tilde{\mathbf{X}}_0\big(\tilde{\mathbf{X}}_0^{\mathrm{T}}\tilde{\mathbf{X}}_0\big)^{-1}\Big\|_\infty < \frac{\min_{j\in\mathscr{L}_0^c}\Big\{\sum_{g^*:j\in\mathscr{I}_{g^*}}\lambda_{g^*}\Big\}}{\max_{j\in\mathscr{L}_0}\Big\{\sum_{g^*:j\in\mathscr{I}_{g^*}}\lambda_{g^*}\Big\}}\big(1-\epsilon\big). \qquad (4.32)$$

Let $\mathbf{E}_3$ be the event that there exists a solution $\widehat{\boldsymbol{\beta}}$ such that $\big\|\widehat{\boldsymbol{\beta}}_g\big\|_2 > 0$ for all $g\in\mathscr{H}_0$ and $\widehat{\boldsymbol{\beta}}_g = 0$ for all $g\in\mathscr{H}_0^c$.

Theorem 4. Under the above assumptions, $\mathrm{P}\big(\mathbf{E}_3\big)$ converges to 1 as $n\to\infty$.

4.5 Proof of Theorem 4

Proof. The proof of the selection consistency starts with the sub-gradient of

$$\text{minimize}_{\ \mathbf{B}}\quad \ell(\boldsymbol{\beta}) = \tfrac{1}{2}\big\|\mathbf{Y}-\mathbf{X}\mathbf{B}\big\|_{\mathrm{F}}^2 + \mathrm{P}\big(\mathbf{B}\big),$$

where $\mathrm{P}(\mathbf{B})$ is grouped lasso penalty. The sub-gradients are

$$\left[\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j},\ j = 1, 2, 3 \ldots, pq\right] = \left[\mathbf{X}_{.j}^{\mathrm{T}}\Big(\vec{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\Big),\ j = 1, 2, \ldots, pq\right] + \widehat{\boldsymbol{\eta}},$$

where $\widehat{\boldsymbol{\eta}} = \left(\widehat{\eta}_1, \widehat{\eta}_2, \ldots, \widehat{\eta}_{pq}\right)^{\mathrm{T}}$ and

$$\widehat{\eta}_j \quad = \quad \begin{cases} \sum_{g^*: j \in \mathscr{I}_{g^*}} \lambda_{g^*} \frac{\widehat{\beta}_j}{\left\|\widehat{\boldsymbol{\beta}}_{g^*}\right\|_2}, & j \in \mathscr{L}_0, \\ \sum_{g^*: j \in \mathscr{I}_{g^*}} \lambda_{g^*} z_{g^*}, & j \in \mathscr{L}_0^c \end{cases},$$

where z_{g^*} are generic vectors such that $\left\|z_{g^*}\right\|_2 \leq 1$ for all g^*.

The event $\mathbf{E}_3$ holds if and only if

$$\widehat{\boldsymbol{\beta}}_{\mathscr{L}_0} = \boldsymbol{\beta}_{\mathscr{L}_0} + \left(\frac{1}{np}\tilde{\mathbf{X}}_0^{\mathrm{T}}\tilde{\mathbf{X}}_0\right)^{-1}\left(\frac{1}{np}\tilde{\mathbf{X}}_0^{\mathrm{T}}\boldsymbol{\epsilon} - \widehat{\boldsymbol{\eta}}_{\mathscr{L}_0}\right),$$

and

$$\lambda\widehat{\boldsymbol{\eta}}_{\mathscr{L}_0^c} = \frac{1}{np}\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\epsilon} + \frac{1}{np}\mathbf{X}_1^{\mathrm{T}}\tilde{\mathbf{X}}_0\left(\frac{1}{np}\tilde{\mathbf{X}}_0^{\mathrm{T}}\tilde{\mathbf{X}}_0\right)^{-1}\left(\lambda\widehat{\boldsymbol{\eta}}_{\mathscr{L}_0} - \frac{1}{np}\tilde{\mathbf{X}}_0^{\mathrm{T}}\boldsymbol{\epsilon}\right).$$

We will show that (a):

$$\lim_{n\to\infty} \mathrm{P}\left(\left\|\widehat{\boldsymbol{\beta}}_{\mathscr{L}_0} - \boldsymbol{\beta}_{\mathscr{L}_0}^0\right\|_\infty > \alpha\right) = 1$$

where $\alpha = \min_{h \in \mathscr{H}_0}\left\|\boldsymbol{\beta}_h\right\|_\infty$ and (b): the probability of the event, for every $j \in \mathscr{L}_0^c$,

$$\left\|\widehat{\eta}_j\right\| < \sum_{h^*: j \in \mathscr{H}_{h^*}} \lambda_{h^*}. \tag{4.33}$$

converges to 1 also as n increases.

First, we show (a):

$$\begin{aligned}
\mathrm{P}\left(\left\|\widehat{\boldsymbol{\beta}}_{\mathscr{L}_0}-\boldsymbol{\beta}^0_{\mathscr{L}_0}\right\|_\infty>\alpha\right) &\leq \frac{1}{\alpha}\mathrm{E}\left\|\widehat{\boldsymbol{\beta}}_{\mathscr{L}_0}-\boldsymbol{\beta}^0_{\mathscr{L}_0}\right\|_\infty \\
&\leq \frac{1}{\alpha}\left(\mathrm{E}\left\|\mathbf{Z}_{\mathscr{L}_0}\right\|_\infty+\lambda\left\|\Sigma_0^{-1}\widehat{\eta}_{\mathscr{L}_0}\right\|_\infty\right) \\
&\leq \frac{1}{\alpha}\left[3\sigma\sqrt{\frac{\log d_0}{npC_{\min,\tilde{\mathbf{X}}_0}}}+\lambda\frac{\sqrt{|\mathscr{L}_0|}}{C_{\min,\tilde{\mathbf{X}}_0}}\max_{g\in\mathscr{H}_0}d_g(\lambda_g,\right]
\end{aligned} \tag{4.34}$$

where $\mathbf{Z}_{\mathscr{L}_0}=\left(\frac{1}{np}\tilde{\mathbf{X}}_0^{\mathrm{T}}\tilde{\mathbf{X}}_0\right)^{-1}\frac{1}{np}\tilde{\mathbf{X}}_0^{\mathrm{T}}\boldsymbol{\epsilon}$ and $d_0=\sum_{g\in\mathscr{H}_0}d_g$.

Second, we show (b): For $j\in\mathscr{L}_0^c$,

$$\begin{aligned}
\left|\widehat{\boldsymbol{\eta}}_j\right| &= \left|\frac{1}{np}\mathbf{X}_{.j}^{\mathrm{T}}\boldsymbol{\epsilon}+\frac{1}{np}\mathbf{X}_{.j}^{\mathrm{T}}\tilde{\mathbf{X}}_0\left(\frac{1}{np}\tilde{\mathbf{X}}_0^{\mathrm{T}}\tilde{\mathbf{X}}_0\right)^{-1}\left(\lambda\widehat{\boldsymbol{\eta}}_{\mathscr{L}_0}-\frac{1}{np}\tilde{\mathbf{X}}_0^{\mathrm{T}}\boldsymbol{\epsilon}\right)\right| \\
&\leq \left|\frac{1}{n}\mathbf{X}_{.j}^{\mathrm{T}}\tilde{\mathbf{X}}_0\left(\frac{1}{n}\tilde{\mathbf{X}}_0^{\mathrm{T}}\tilde{\mathbf{X}}_0\right)^{-1}\widehat{\eta}_{\mathscr{L}_0}\right| \\
&\quad+\left|\mathbf{X}_{.j}^{\mathrm{T}}\left[\mathrm{I}-\tilde{\mathbf{X}}_0\left(\frac{1}{n}\tilde{\mathbf{X}}_0^{\mathrm{T}}\tilde{\mathbf{X}}_0\right)^{-1}\tilde{\mathbf{X}}_0^{\mathrm{T}}\right]\frac{1}{n}\boldsymbol{\epsilon}\right|.
\end{aligned} \tag{4.35}$$

In the above, the former

$$\max_{j\in\mathscr{L}_0^c}\left|\frac{1}{n}\mathbf{X}_{.j}^{\mathrm{T}}\tilde{\mathbf{X}}_0\left(\frac{1}{n}\tilde{\mathbf{X}}_0^{\mathrm{T}}\tilde{\mathbf{X}}_0\right)^{-1}\widehat{\eta}_{\mathscr{L}_0}\right| \leq \max_{j\in\mathscr{L}_0^c}\left\|\frac{1}{n}\mathbf{X}_{.j}^{\mathrm{T}}\tilde{\mathbf{X}}_0\left(\frac{1}{n}\tilde{\mathbf{X}}_0^{\mathrm{T}}\tilde{\mathbf{X}}_0\right)^{-1}\right\|_\infty\left\|\widehat{\eta}_{\mathscr{L}_0}\right\|_\infty, \tag{4.36}$$

where

$$\begin{aligned} \|\widehat{\eta}_{\mathscr{L}_0}\|_\infty &= \max_{j\in\mathscr{L}_0}\left|\sum_{g^*:j\in\mathscr{H}_{g^*}} \lambda_{g^*}\frac{\widehat{\beta}_j}{\|\widehat{\boldsymbol{\beta}}_{g^*}\|_2}\right| \\ &\le \max_{j\in\mathscr{L}_0}\left\{\sum_{g^*:j\in\mathscr{H}_{g^*}} \lambda_{g^*}\frac{|\widehat{\beta}_j|}{\|\widehat{\boldsymbol{\beta}}_{g^*}\|_2}\right\} \\ &\le \max_{j\in\mathscr{L}_0}\left\{\sum_{g^*:j\in\mathscr{H}_{g^*}} \lambda_{g^*}\right\}. \end{aligned}$$

Therefore, the RHS of (4.36) is smaller than

$$\begin{aligned} &\max_{j\in\mathscr{L}_0^c}\left\|\frac{1}{n}\mathbf{X}_{\cdot j}^{\mathrm{T}}\tilde{\mathbf{X}}_0\left(\frac{1}{n}\tilde{\mathbf{X}}_0^{\mathrm{T}}\tilde{\mathbf{X}}_0\right)^{-1}\right\|_\infty \max_{j\in\mathscr{L}_0}\left\{\sum_{g^*:j\in\mathscr{H}_{g^*}} \lambda_{g^*}\right\} \\ &\le \min_{j\in\mathscr{L}_0^c}\left\{\sum_{g^*:j\in\mathscr{H}_{g^*}} \lambda_{g^*}\right\}\cdot(1-\epsilon) \end{aligned}$$

by using the second assumption.

In the second term in (4.36), let $\mathbf{W} = \left(W_1, W_2, \ldots, W_{\left|\mathscr{L}_0^c\right|}\right)^{\mathrm{T}}$, where

$$W_j = \mathbf{X}_{\cdot j}^{\mathrm{T}}\left[\mathrm{I} - \tilde{\mathbf{X}}_0\left(\frac{1}{n}\tilde{\mathbf{X}}_0^{\mathrm{T}}\tilde{\mathbf{X}}_0\right)^{-1}\tilde{\mathbf{X}}_0^{\mathrm{T}}\right]\frac{1}{n}\boldsymbol{\epsilon}.$$

Then, using the same arguments in Nardi and Rinaldo (2008), we have

$$\mathrm{E}\left\|\mathbf{W}\right\|_\infty \le 3\sigma_\epsilon\sqrt{\frac{\log d}{np}}\max_{j\in\mathscr{L}_0^c}\left\|\mathbf{X}_{\cdot j}\right\|_2,$$

and

$$\begin{aligned}
&\mathrm{P}\left(\frac{1}{\min_{j\in\mathscr{H}_g}\left\{\sum_{g^*:j\in\mathscr{H}_{g^*}}\lambda_{g^*}\right\}}\left\|\mathbf{W}\right\|_\infty > \frac{\epsilon}{2}\right)\\
&\qquad\leq \frac{6\sigma_\epsilon}{\epsilon\min_{j\in\mathscr{H}_g}\left\{\sum_{g^*:j\in\mathscr{H}_{g^*}}\lambda_{g^*}\right\}}\sqrt{\frac{\log d}{np}}\max_{j\in\mathscr{L}_0^c}\left\|\mathbf{X}_{\cdot j}\right\|_2,
\end{aligned}$$

which converges to 0 as n increases. This concludes the proof.

□

CHAPTER 5

Simulation Studies

5.1 Data generation

We first conducted extensive simulation studies to evaluate the performance of the method. We generated the data based on Equations (3.5) and (3.14) with $n = 1000, p = 110, G = 10$, with two distinct group matrix $\mathbf{A}$ of different degrees of overlap between the groups. The first group matrix corresponds to a structure with significant overlapping between groups and the other an almost mutually exclusive grouping structure. We call these two group matrices $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ respectively. Specifically, we set

$$\mathbf{A}^{(1)}_{\ell,g} = \begin{cases} 1 & \text{if } \ell \in [10(g-1)+1, 10\,(g+1)] \\ \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

and

$$\mathbf{A}^{(2)}_{\ell,g} = \begin{cases} 1 & \text{if } g = 1, \ell \leq 14 \\ 1 & \text{if } g = 10, \ell \geq 95 \\ 1 & \text{if } 1 < g < 10, \ell \in [10\,(g-1)+4, 10\,(g+1)-5] \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

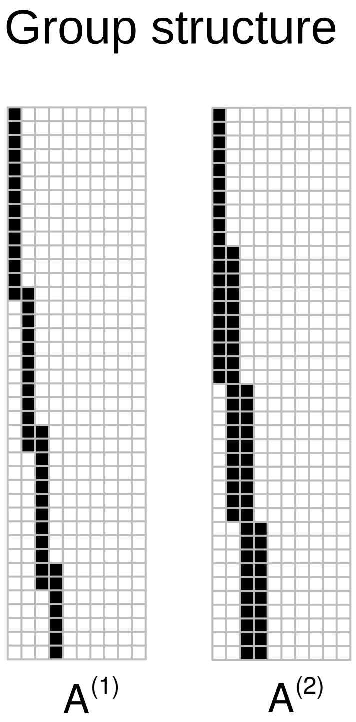

Figure 5.1 The group structure in $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ (large overlap between groups).

Note that most p elements of $\mathbf{X}$ belong to 2 groups in both cases. Next, we randomly selected 5 groups of variables to be associated with non-zero factors, and simulated their factors independently from

$$\mathbf{f}_{i,g} \sim N\,(0,1) \,\text{independently}\ 6 \leq g \leq 10, \forall i. \tag{5.3}$$

Lastly, we added the noise component for ϵ_i^ℓ from Normal distribution

$$\epsilon_\ell^i \sim N\left(0, \sigma^2\right) \tag{5.4}$$

independently for $i = 1, \ldots, n$ and $\ell = 1, \ldots, p$, where σ is the noise parameter for Stage 1 estimation.

For the simulation of the response variables $\mathbf{Y}$, we used a univariate response ($q = 1$) for the simplicity of simulation. Hence the regression matrix $\mathbf{B}$ has a dimension of $p \times 1$ and we simulated $\mathbf{Y}$ from

$$\mathbf{Y}_i = \mathbf{X}_i^{\mathrm{T}}\mathbf{B} + \mathbf{u}_i. \tag{5.5}$$

where

$$\mathbf{B}_\ell = \begin{cases} 1 & \text{for } 100 \leq \ell \leq 110 \\ \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

We simulated $\mathbf{u}_i$ from Normal distribution

$$\mathbf{u}_{i\ell} \sim \mathrm{N}\left(0, \tau^2\right), \tag{5.7}$$

for $i = 1, \ldots, n$ and all ℓ, where τ is therefore the noise parameter for Stage 2 estimation. For the two grouping structures as described above, we have generated 100 simulation data sets for the following four pairs of stage-specific noise levels in $\mathbf{X}$ and $\mathbf{Y}$: $(\sigma, \tau) = (1, 1), (1, 2), (2, 1), (2, 2)$. Using these data sets, we will evaluate our method comparatively with soft thresholding in terms of two key asymptotic properties of the our estimator, namely sparsistency and consistency.

5.2 Result

Since the data was generated by emulating the systematic component for the Stage 1 model, we used the sensitivity and specificity in the detection of non-zero elements in $\hat{\mathbf{\Sigma}}^{\mathbf{0}}_{\mathbf{XX}} := \mathbf{A}\hat{\mathbf{\Sigma}}_{\mathbf{f}}\mathbf{A}'$ and $\hat{\mathbf{\Sigma}}^{\mathbf{0}}_{\mathbf{XY}} := \mathbf{A}\hat{\mathbf{\Sigma}}_{\mathbf{f}}\mathbf{A}'\mathbf{B}$. For the Stage 1 evaluation, we calculated:

$$\begin{aligned}\text{Sensitivity} &= \frac{\text{number of factors correctly estimated to be non-zero}}{\text{number of factors that are non-zero}}\\ \text{Specificity} &= \frac{\text{number of factors correctly estimated to be zero}}{\text{number of factors that are zero}}.\end{aligned}$$

Likewise for Stage 2 evaluation, we calculated

$$\begin{aligned}\text{Sensitivity} &= \frac{\text{number of elements in } \hat{\mathbf{\Sigma}}^{\mathbf{0}}_{\mathbf{XY}} \text{ correctly estimated to be non-zero}}{\text{number of elements that are non-zero}}\\ \text{Specificity} &= \frac{\text{number of elements in } \hat{\mathbf{\Sigma}}^{\mathbf{0}}_{\mathbf{XY}} \text{ correctly estimated to be zero}}{\text{number of elements that are zero}}.\end{aligned}$$

Across all simulation settings, accounting for large and small overlap between groups and variable degree of noise, the proposed method that incorporated the true group structure achieved significantly better results than element-wise shrinkage estimator with soft-thresholding operator in terms of sensitivity and specificity. First, Figures 5.2, 5.3,5.6 and 5.7 showed that Stage 1 estimation identified non-zero and zero factors with high sensitivity and specificity. The results shown is for the case with the highest noise levels $(\sigma, \tau) = (2, 2)$ and grouping structure with a high degree of overlap $\mathbf{A}^{(2)}$, and the performance improved as we lowered the level of noise and groups sharing fewer genes ($\mathbf{A}^{(1)}$). Figures 5.4, 5.5, 5.8, 5.9 is showing the sensitivity and specificity for detecting non-zero elements in the final estimate of $\hat{\mathbf{\Sigma}}^{\mathbf{0}}_{\mathbf{XY}}$, in which the Stage 2 estimation of the proposed estimator outperforms the element-wise shrinkage estimator at all degrees of shrinkage represented by the Frobenius norm of the difference between the true cross-covariance matrix and the estimated matrix. Overall, the simulations under all parameter settings indicated that the proposed estimator recovers the non-zero factors of $\mathbf{f}$ and non-zero elements in the systematic component of cross-covariance matrix $\mathbf{\Sigma}_{\mathbf{XY}}$.

Group structure $\mathbf{A}^{(1)}$

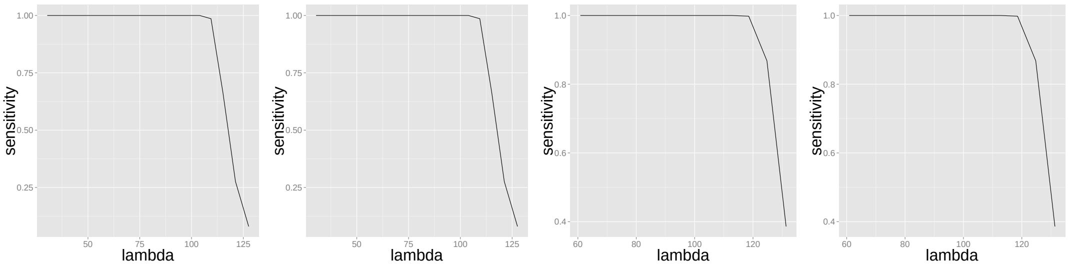

Figure 5.2 Sensitivity as a function of the tuning parameter in Stage 1 estimation under four scenarios: $(\sigma, \tau) = (1, 1)$, $(\sigma, \tau) = (1, 2)$, $(\sigma, \tau) = (2, 1)$, $(\sigma, \tau) = (2, 2)$.

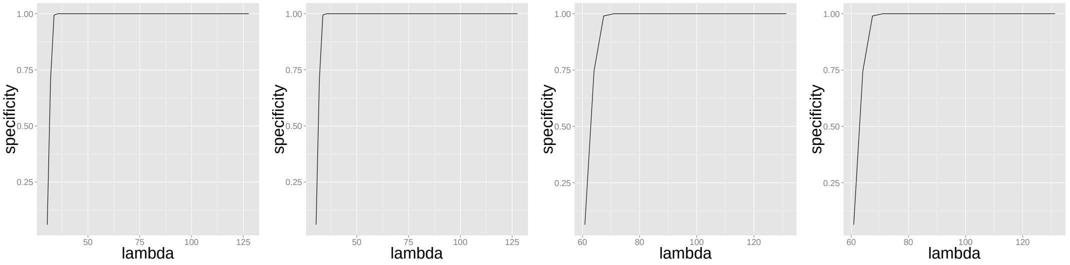

Figure 5.3 Specificity as a function of the tuning parameter in Stage 1.

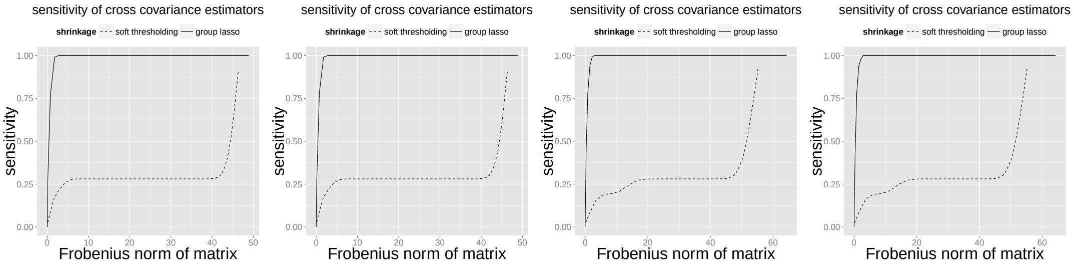

Figure 5.4 Specificity of the overall estimate as a function of the Frobenius norm of the estimated cross covariance matrix.

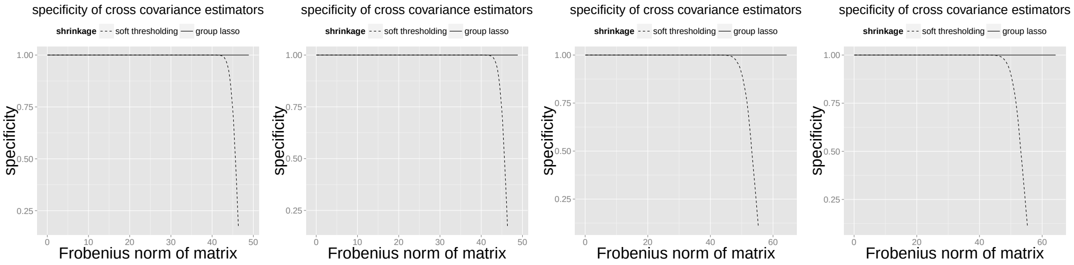

Figure 5.5 Specificity of the overall estimate as a function of the Frobenius norm of the estimated cross covariance matrix.

Group structure $\mathbf{A}^{(2)}$

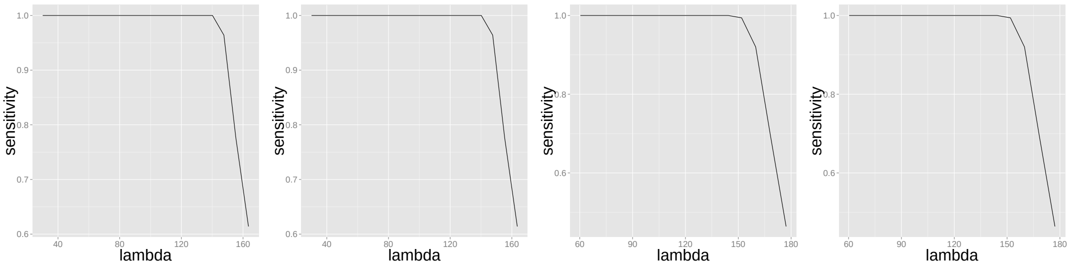

Figure 5.6 Sensitivity as a function of the tuning parameter in Stage 1 estimation under four scenarios: $(\sigma, \tau) = (1, 1)$, $(\sigma, \tau) = (1, 2)$, $(\sigma, \tau) = (2, 1)$, $(\sigma, \tau) = (2, 2)$.

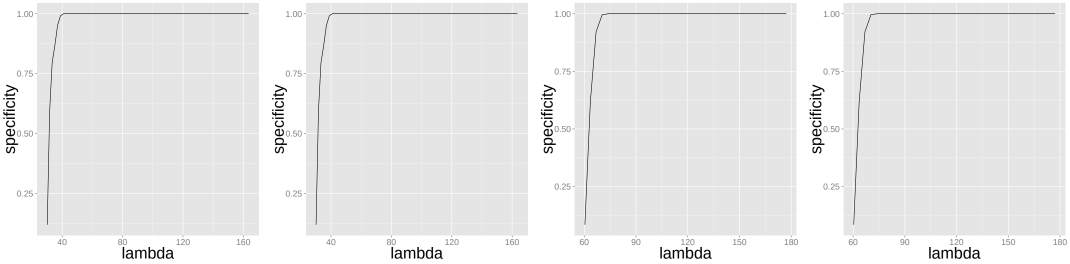

Figure 5.7 Specificity as a function of the tuning parameter in Stage 1.

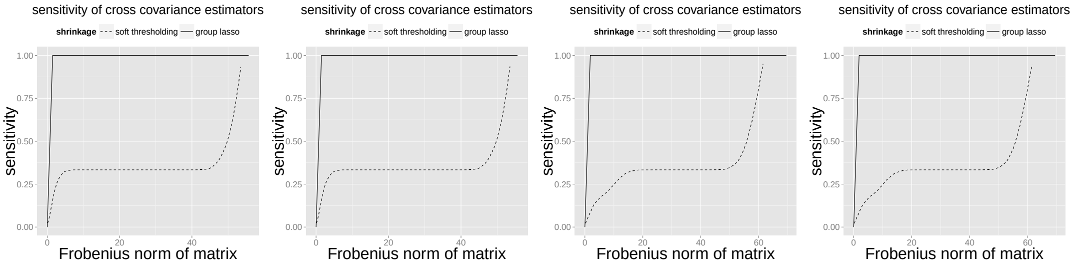

Figure 5.8 Specificity of the overall estimate as a function of the Frobenius norm of the estimated cross covariance matrix.

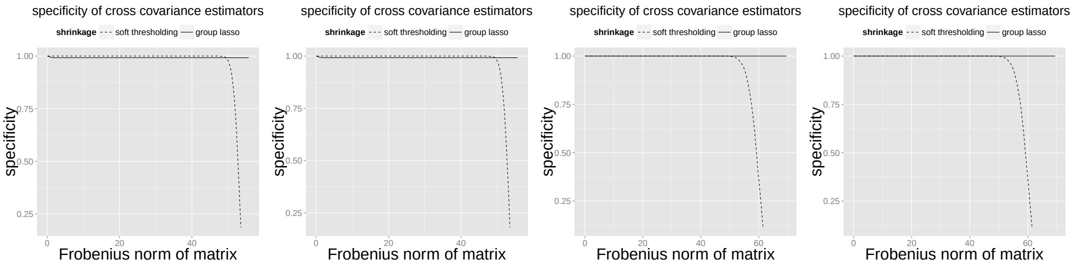

Figure 5.9 Specificity of the overall estimate as a function of the Frobenius norm of the estimated cross covariance matrix.

CHAPTER 6

Analysis of DNA methylation and mRNA expression in TCGA data

We applied the proposed method to the integrative analysis of DNA methylation and mRNA transcript expression data in TCGA-BRCA cohort. We used Gene Ontology terms (Ashburner et al., 2000) as gene group information, especially the set of GO terms containing 10 to 50 genes in the definition. The choice of GO terms is deliberate since GO is a relational database with hierarchical structure: each GO term has a parent term that is larger in size and less specific in definition. By limiting the GO term size in this range, we can reduce the number of shared genes between the GO terms, which has to be controlled to a degree to meet the minimal overlap requirement for the estimators in both stages. To focus on the biological functions implicated in different tumor subtypes, we further reduced the data to the genes involved in the kinase signalling networks, consisting of 227 kinases and 764 non-kinase substrates that are experimentally validated and appear in the iRefIndex database (Razick et al., 2008). All 991 genes were present in the mRNA data (RNA-seq), and 958 genes were present in the DNA methylation data (microarray) for 766 subjects in total.

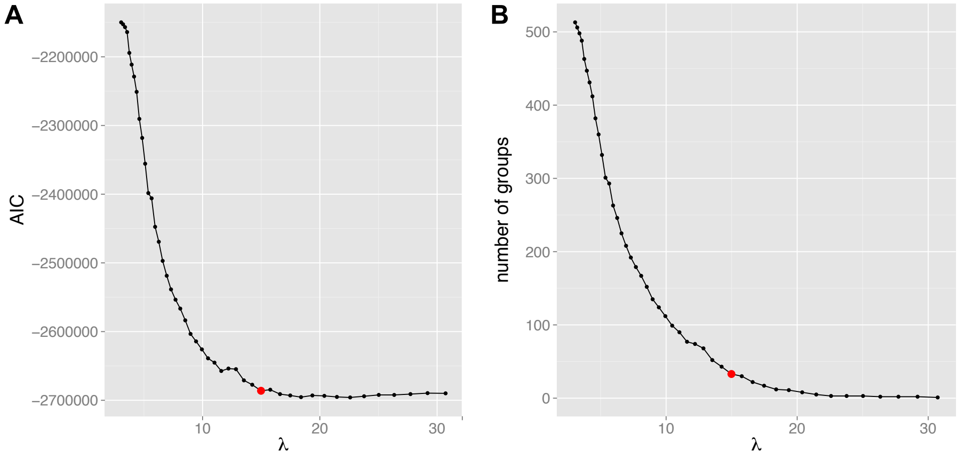

Figure 6.1 Model selection via AIC for the first stage estimation in the K2S analysis. (A) AIC versus shrinkage parameter. (B) The number of selected groups at the corresponding shrinkage parameter.

Figure 6.1 shows the AIC curve and the number of selected groups with non-zero factors in the stage 1 estimation. We selected the optimal threshold $\lambda = 15$ where the curve became flat, which yielded 33 non-zero factors reported in Table 6.1. As we focused on the genes involved in the kinase signalling network, the selected groups included many phosphorylation-mediated gene expression regulation terms for signal transduction as well as other functions such as immune response, inflammatory response, angiogenesis that are commonly implicated in genomic studies of tumors. Interestingly, estimated factors $\widehat{\mathbf{f}}_i$ were mostly positively correlated with the

GO ID	GO term name
GO:0001525	Angiogenesis
GO:0001934	positive regulation of protein phosphorylation
GO:0002576	platelet degranulation
GO:0004872	receptor activity
GO:0005102	receptor binding
GO:0005200	**structural constituent of cytoskeleton**
GO:0005768	endosome
GO:0005783	endoplasmic reticulum
GO:0005882	**intermediate filament**
GO:0006919	activation of cystein-type endopeptidase activity involved in apoptotic process
GO:0006954	inflammatory response
GO:0006955	immune response
GO:0007155	cell adhesion
GO:0007169	transmembrane receptor protein tyrosine kinase signaling
GO:0007264	small GTPase mediated signal transduction
GO:0007267	cell-cell signaling
GO:0007268	synaptic transmission
GO:0007568	aging
GO:0007611	learning or memory
GO:0008083	growth factor activity
GO:0009897	external side of plasma membrane
GO:0009967	positive regulation of signal transduction
GO:0010628	positive regulation of gene expression
GO:0016021	integral component of membrane
GO:0016324	apical plasma membrane
GO:0016477	cell migration
GO:0019904	protein domain specific binding
GO:0030335	positive regulation of cell migration
GO:0031295	**T cell costimulation**
GO:0034220	ion transmembrane transport
GO:0042110	**T cell activation**
GO:0043410	positive regulation of MAPK cane
GO:0051092	positive regulation of NF-kappaB transcription factor activity

Table 6.1 The selected factors (GO terms) in the first stage of the kinase-to-substrate analysis. The GO terms in bold are the ones in which the estimated factors from DNA methylation data were clearly negatively correlated with average mRNA expression data.

average DNA methylation patterns, but not in all GO terms. As shown in some panels of Figure 6.2A, the estimated factors did not exactly match the average DNA methylation profiles in all GO terms, especially in the GO terms that share many genes with at least one other term(s), e.g. T

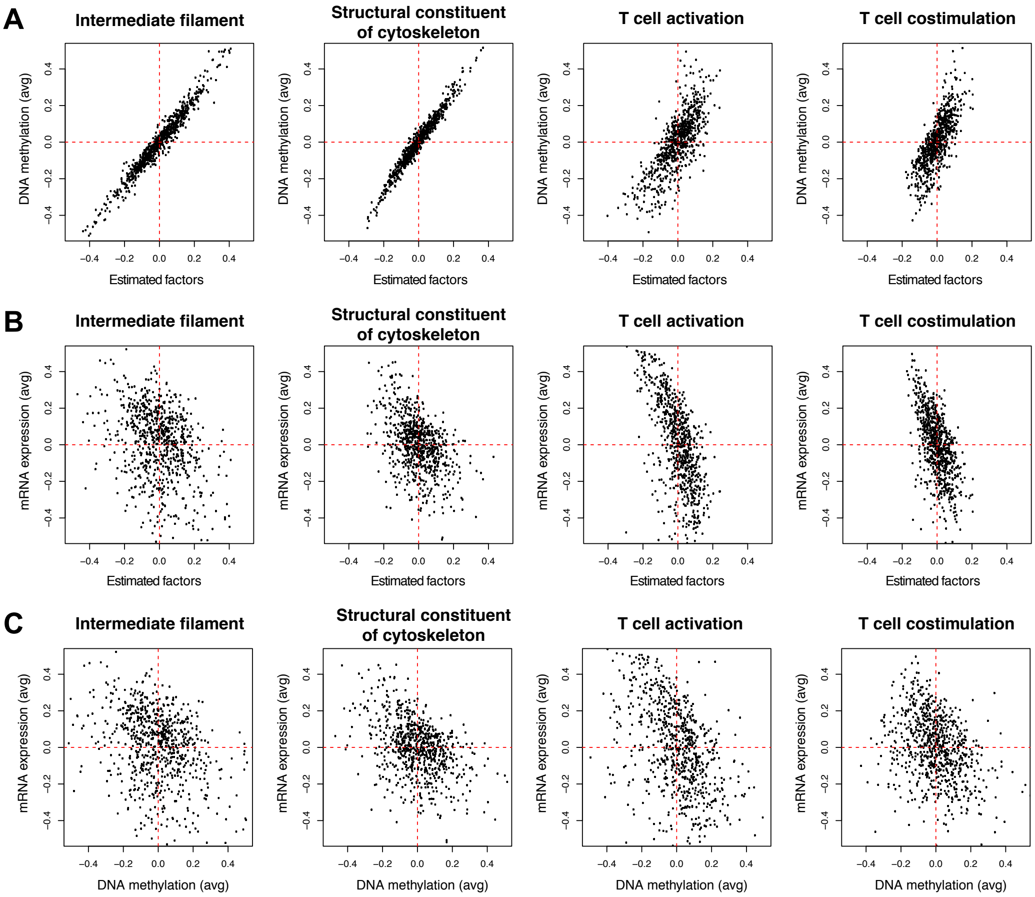

Figure 6.2 (A) Average DNA methylation levels versus estimated factors representing four GO terms, which are positively correlated as expected. (B) Average mRNA expression levels versus estimated factors representing the GO terms. (C) Average mRNA data versus average DNA methylation data for the same GO terms.

cell activation and T cell co-stimulation sharing many member genes with immune response term. In these cases, the factors tend to be shrunken toward zero compared to the average methylation levels, but the correlation with the average mRNA expression within the same term tended to be more negative, as we shall discuss later.

As mentioned above, the proposed method allows us to express the variance-covariance matrix of DNA methylation data as the sum of a systematic and an idiosyncratic component. Figures 6.3A and 6.3B show the

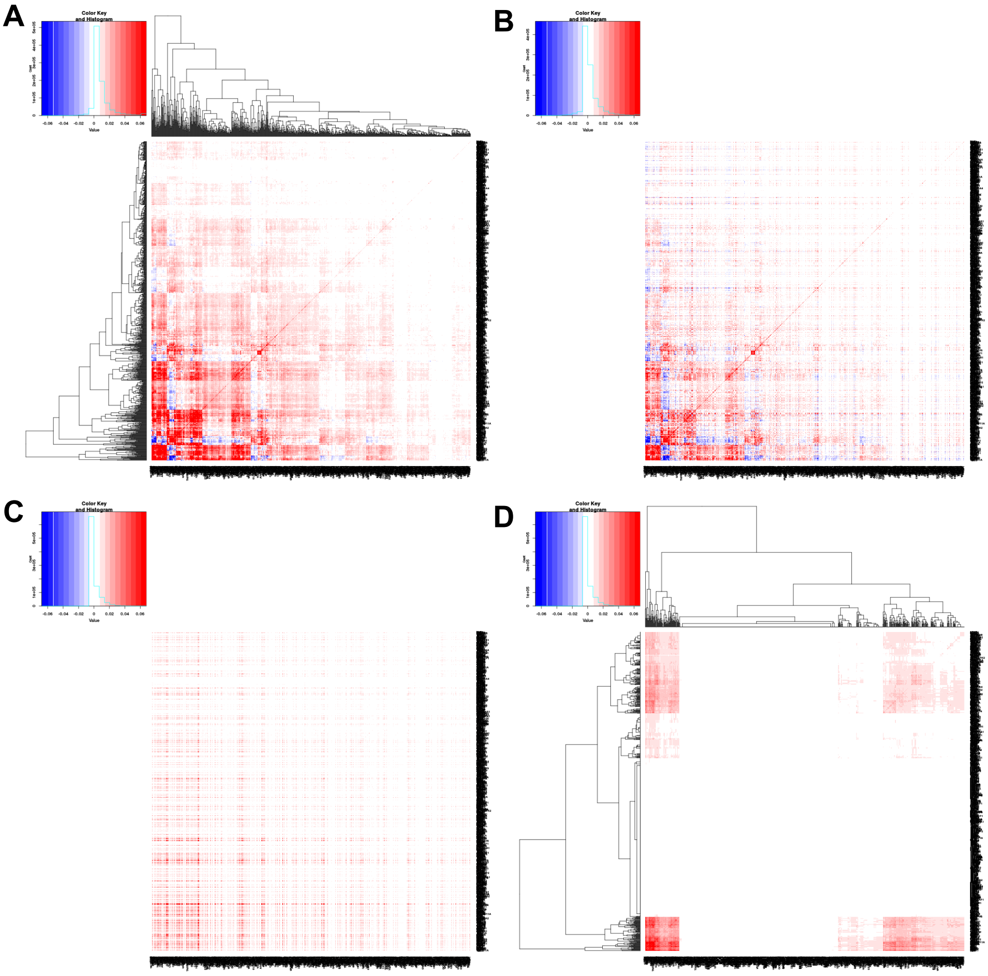

Figure 6.3 (A) Sample covariance matrix of DNA methylation data, indicating largely positive correlation between different genes. (B) Estimated covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{XX}}$ in Stage 1. (C) The systematic component $\mathbf{A}\hat{\mathbf{\Sigma}}_{\mathbf{f}}\mathbf{A}^{\mathrm{T}}$ in Stage 1 estimation. (D) The same matrix in (C) after hierarchical clustering.

sample covariance matrix and the one obtained by the two-stage estimation, respectively. Note that the matrix shown in Figure 6.3B is the sum of the two matrices, the systematic component associated with 33 GO terms shown in Figure 6.3C and the residual component (not shown, Figure 6.3D is the same matrix after hierarchical clustering). A striking realization in this decomposition is the sparsity of the systematic component shown in the last two panels, suggesting that the GO terms we used as grouping

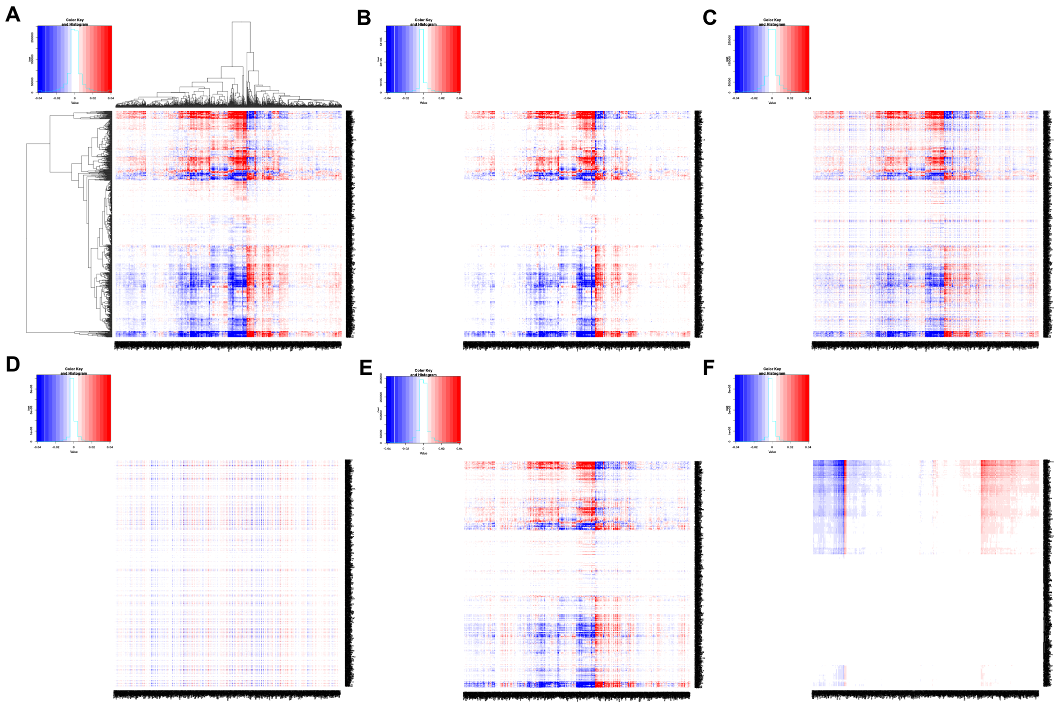

Figure 6.4 (A) Sample cross-covariance matrix between DNA methylation data and mRNA expression data. (B) Cross-covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{XY}}$ with a soft-thresholding estimator. (C) Cross-covariance estimated by the two-stage estimator $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\widehat{\mathbf{B}}$. (D) The systematic component $\mathbf{A}\hat{\mathbf{\Sigma}}_{\mathbf{f}}\mathbf{A}^{\mathrm{T}}\widehat{\mathbf{B}}$ of the cross-covariance matrix. (E) The residual cross-covariance matrix $\hat{\mathbf{\Sigma}}_{\boldsymbol{v}}$.(F) The systematic component in (D) after hierarchical clustering.

information explain a relatively small proportion of total variability in the DNA methylation data. Although indeed there does not exist a concerted regulatory program of DNA methylation at the pathway level, another plausible explanation is that the GO terms used in this analysis does not capture all functional clusters of genes due to lack of discoveries or incompleteness of the database. Moreover, as shown in the example diagram of T cell co-stimulation in Figure 2.1, it may be difficult to expect all 10 to 50 genes in a pathway to be co-regulated to be represented by a common factor. Hence a more careful curation of co-regulated gene groups from the GO terms can improve the proportion of variability explained by the systematic component.

The ultimate goal of our analysis was to identify the pathways in which DNA methylation is the major driver of mRNA expression regulation mechanism, i.e. the cross-covariance matrix estimation. Figures 6.4A, 6.4B, 6.4C are the sample cross-covariance matrix with no shrinkage, with shrinkage by soft-thresholding operator with cross validation-based tuning parameter selection, and the estimate from the proposed two-stage estimator, respectively. The real advantage of the proposed method is the decomposition $\hat{\boldsymbol{\Sigma}}_{\mathbf{XY}} = \mathbf{A}\hat{\boldsymbol{\Sigma}}_{\mathbf{f}}\mathbf{A}^{\mathrm{T}}\widehat{\mathbf{B}} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}}\widehat{\mathbf{B}}$, where the former is shown in Figure 6.4D and 6.4F before and after hierarchical clustering respectively and the latter is shown in Figure 6.4E. The upper left corner (blue) of Figure 6.4F represents the groups of genes in which DNA methylation was negatively correlated with mRNA expression amongst themselves, indicating that DNA methylation played a significant regulatory role on the mRNA expression in those genes at the pathway level. The upper right corner (red) represents two different gene sets for which DNA methylation of one set of genes is positively correlated with mRNA levels of the other set of genes, which carries no biologically meaningful covariances with respect to methylation-mediated RNA expression.

With regard to the GO terms selected in Stage 1, the scatter plot of estimated factors against average mRNA expression patterns in those GO terms (Figure 6.2B) indicates negative correlation with the mRNA data, suggesting the repressive role of methylation on the transcript output. Interestingly, the comparison between Figure 6.2B and Figure 6.2C suggests that the estimated factors are much better correlated with the average methylation levels, and this verifies that the shrunken estimates of factor components are able to reveal the regulatory structure with improved clarity. These findings are also corroborated by previous breast cancer oncogenomic reports, such as methylation-driven intermediate filament dynamics (Noetzel et al., 2010) and cytoskeletal component (Ulirsch et al., 2013), and T lymphocyte infiltration of the breast tumors (Dedeurwaerder et al., 2011).

CHAPTER 7

Discussion

In this work, we have developed a two-stage estimator of cross-covariance matrix, which takes advantage of existing group information between the variables. Despite the detour we take in obtaining shrunken estimates, the advantage of our method lies in its ability to tease out biologically relevant signals from the residual effects, thereby facilitating meaningful interpretation of data. In addition, we have provided theoretical properties such as estimation consistency and model selection consistency on both stages with appropriate conditions, using slight modifications of existing work. Our extensive simulation studies have demonstrated that these properties are valid even when the group information is incomplete, since the residual component captures the rest of the variability unexplained by the grouped variables (factors in the systematic components). Compared to the POET estimator, the major difference in the first stage estimator was that the factors are defined by the previously defined gene grouping information, as opposed to numerically derived linear combinations of variables, i.e. the first K principal components in their method. The second stage estimator of multivariate linear regression model is a modification of group lasso allowing overlap of group membership (Jacob et al., 2009; Li et al., 2015; Simon et al., 2013), which has been previously proposed. We have formally specified the restrictions on group overlap (in Equation 4.7). Our method handles overlapping by reducing such groups to a set of non overlapping groups. This approach is different from the method in Li et al. (2015), where their algorithm directly accounts for the group overlaps.

Our analysis of the human kinase network in TCGA BRCA data recovered previously known hypermethylation activities in the cancer genomes that were validated in independent study populations outside TCGA, suggesting the validity of our approach and increased opportunity of further discoveries of gene expression regulation activities through multi-omics data sets. This can be achieved through application of our method to broader gene sets (e.g. outside kinase signaling networks), or analysis of different data sources such as microRNA paired with protein expression data. We have also illustrated that element-wise shrinkage estimation, in spite of the ease of implementation and numerical optimality of estimation procedures, is likely to capture indirect correlations that are not biologically relevant in the context of joint analysis of DNA methylation and mRNA expression. As biological systems are operated by densely connected networks of molecular machineries, i.e. biological functions or pathways, utilizing previously characterized gene group information is expected to improve the biological relevance of selected covariance terms.

The proposed method is far from flawless nonetheless, which warrants further improvement. First, we assumed that the factors $\{\mathbf{f}_{\cdot g}\}_{g=1}^{G}$ represent the common effects shared by all members of the individual groups, and this assumption can be rigid when the group size is large since there can be subgroups of genes that are regulated differently within the group. This is best exemplified in the PI3K-AKT1-mTORC2 complex genes included in Figure 2.1, located in the middle rows of the two heat maps, which clearly indicate those genes violate our common factor assumption. In addition, the factors can be un-estimable as the group definitions share too many common genes between one another. Therefore it is crucial to screen the gene group definition before fitting the group lasso.

Second, although the simulation studies showed that the systematic component can be estimated consistently, our TCGA data analysis showed that the proportion of cross-covariance explained by the systematic component was small. The main reasons for this outcome can be two-fold. Since the proposed estimation in the first stage is sequential, first applied to the factors and subsequently applied to the residuals, it is possible that the factors could have been underestimated because it was estimated by shrinkage first without simultaneously estimating penalizing residuals. In addition, it is possible that the AIC was suboptimal for model selection, especially when the AIC curves tended to be monotone decreasing (rather than U-shaped) as the shrinkage parameter increased. This observation likely has to do with the fact that the number of factor terms increases along with the sample size, and thus it may be necessary to devise a new model selection criteria for this type of problems. Besides using the AIC for model selection, we have also tried using the Bayesian information criterion (BIC) (Schwarz, 1978) and C_p statistic (Mallows, 1973) for model selection but the results are not shown here. Neither of this two performed reasonably in the simulation studies. In most cases, either all or none of the features were selected. The cause of such failures might be worth investigating in future work.

Lastly, our current penalty structure yields either all zero or all non-zero estimates in each gene group. However, it is possible that further shrinkage on the factor estimates $\{\mathbf{f}_{\cdot g}\}_{g=1}^{G}$ for some subjects but not all, e.g. L_1 penalty on the factors of each individual, can be imposed and it can provide more interpretable results. For example, DNA methylation-mediated mRNA regulation is neither the only mechanism nor universal in every individual in a study such as TCGA, which profiles tumor samples of various molecular types and thus one type of gene expression regulation mechanism is turned on or off in a subset of tumor specimens only. However, we consider these potential refinements beyond the scope of this work and leave them to future research.

Bibliography

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akadémiai Kiado. 3.3

Anestis Antoniadis and Jianqing Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):pp. 939–955, 2001. ISSN 01621459. URL `http://www.jstor.org/stable/2670237`. 1.3.2, 3.1

M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, 2000. 6

Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, January 2011. ISSN 1471-0056. doi: 10.1038/nrg2918. URL `http://www.nature.com/nrg/journal/v12/n1/full/nrg2918.html`. 1.1

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):pp. 289–300, 1995. ISSN 00359246. URL `http://www.jstor.org/stable/2346101`. 1.1

Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, December 2008a. ISSN 0090-5364. doi: 10.1214/08-AOS600. URL `http://projecteuclid.org/euclid.aos/1231165180`. 1.3.2

Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36:199–227, February 2008b. ISSN 0090-5364. doi: 10.1214/009053607000000758. URL `http://projecteuclid.org/euclid.aos/1201877299`. 1.1, 1.3.2, 2.1

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of LASSO and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. 4.2

Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1(0): 169–194, 2007. ISSN 1935-7524. doi: 10.1214/07-EJS008. URL http://projecteuclid.org/euclid.ejs/1179759718. 1.1, 4.2, 4.3

Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106 (494):672–684, June 2011. ISSN 0162-1459, 1537-274X. doi: 10.1198/jasa.2011.tm10560. URL http://www.tandfonline.com/doi/abs/10.1198/jasa.2011.tm10560. 1.1, 1.3.2, 2.1, 3.1, 4.3

H. Chun and S. Keles. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*, 182(1): 79–90, 2009. 2.1

Sarah Dedeurwaerder, Christine Desmedt, Emilie Calonne, Sandeep K Singhal, Benjamin Haibe-Kains, Matthieu Defrance, Stefan Michiels, Michael Volkmar, Rachel Deplus, Judith Luciani, et al. Dna methylation profiling reveals a predominant immune component in breast cancers. *EMBO molecular medicine*, 3(12):726–741, 2011. 6

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, September 2013. ISSN 13697412. doi: 10.1111/rssb.12016. URL http://doi.wiley.com/10.1111/rssb.12016. (document), 1.1, 1.3.2, 2.2, 4.3, 4.3, 4.3

Yingying Fan and Runze Li. Variable selection in linear mixed effects models. *Ann. Stat.*, 40(4):2043–2068, 2012. ISSN 0090-5364; 2168-8966/e. doi: 10.1214/12-AOS1028. 4.3, 4.3, 4.3, 1

Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2 2010. ISSN 1548-7660. URL `http://www.jstatsoft.org/v33/i01`. 1.3.1

Jerome Friedman; Trevor Hastie; Holger Höfling and Robert Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007. ISSN 1932-6157. doi: 10.1214/07-AOAS131. 1.3.1, 3.3, 3.3

Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009. 7

Ian T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31 (3):pp. 300–303, 1982. ISSN 00359254. URL `http://www.jstor.org/stable/2348005`. 1.2

C. Kendziorski and P. Wang. A review of statistical methods for expression quantitative trait loci mapping. *Mamm. Genome*, 17:509–517, 2006. 2.1

Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5):1356–1378, 10 2000. doi: 10.1214/aos/1015957397. URL `http://dx.doi.org/10.1214/aos/1015957397`. 1.1, 1.3.1

Daniel C. Koboldt, Robert S. Fulton, Michael D. McLellan, Heather Schmidt, Joelle Kalicki-Veizer, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, September 2012. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature11412. URL `http://www.nature.com/doifinder/10.1038/nature11412`. 2.1, 2.2

Kim-Anh Lê Cao, Ignacio González, and Sébastien Déjean. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics (Oxford, England)*, 25(21):2855–2856, November 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp515. 1.1

E. Levina, A.J. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *Ann. Appl. Stat.*, 2(1):245–263, 2008. 2.1

Yanming Li, Bin Nan, and Ji Zhu. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 2015. 7

C. L. Mallows. Some Comments on C P. *Technometrics*, 15(4):661, November 1973. ISSN 00401706. doi: 10.2307/1267380. URL `http://www.jstor.org/stable/1267380?origin=crossref`. 7

Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732, October 2013. ISSN 1471-0056. doi: 10.1038/nrg3552. URL `http://www.nature.com/nrg/journal/v14/n10/full/nrg3552.html`. 1.1

Yuval Nardi and Alessandro Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2(0):605–633, 2008. ISSN 1935-7524. doi: 10.1214/08-EJS200. URL `http://projecteuclid.org/euclid.ejs/1217450797`. 1.3.1, 4.2, 4.2, 4.3, 4.3, 4.5

E Noetzel, M Rose, E Sevinc, RD Hilgers, A Hartmann, A Naami, R Knüchel, and E Dahl. Intermediate filament dynamics and breast cancer: aberrant promoter methylation of the synemin gene is associated with early tumor relapse. *Oncogene*, 29(34):4814–4825, 2010. 6

Michael C. Oldham, Genevieve Konopka, Kazuya Iwamoto, Peter Langfelder, Tadafumi Kato, Steve Horvath, and Daniel H. Geschwind. Functional organization of the transcriptome in human brain. *Nature Neuroscience*, 11(11):1271–1282, November 2008. ISSN 1546-1726. doi: 10.1038/nn.2207. 1.1

Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160–1167, 2009. 2.1

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. *The Annals of Applied Statistics*, 4(1):53–77, March 2010. ISSN 1932-6157. doi: 10.1214/09-AOAS271. URL http://projecteuclid.org/euclid.aoas/1273584447. 2.1, 3.2

Sabry Razick, George Magklaras, and Ian M. Donaldson. iRefIndex: a consolidated protein interaction database with provenance. *BMC bioinformatics*, 9:405, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-405. 6

M.V. Rockman and L. Kruglyak. Genetics of global gene expression. *Nature Reviews Genetics*, 7:862–872, 2006. 2.1

Adam J. Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, March 2009. ISSN 0162-1459, 1537-274X. doi: 10.1198/jasa.2009.0101. URL http://www.tandfonline.com/doi/abs/10.1198/jasa.2009.0101. 1.1, 1.3.2, 2.1, 3.1, 4.3

Steffen Sass, Florian Buettner, Nikola S. Mueller, and Fabian J. Theis. A modular framework for gene set analysis integrating multilevel omics data. *Nucleic Acids Research*, 41(21):9622–9633, November 2013. ISSN 1362-4962. doi: 10.1093/nar/gkt752. 1.1

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, March 1978. ISSN 0090-5364. doi: 10.1214/aos/1176344136. URL http://projecteuclid.org/euclid.aos/1176344136. 7

E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19(suppl 1):i273–i282, July 2003a. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btg1038. URL `http://bioinformatics.oxfordjournals.org/content/19/suppl_1/i273`. 1.1

Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, June 2003b. ISSN 1061-4036. doi: 10.1038/ng1165. 1.1

Eran Segal, Nir Friedman, Daphne Koller, and Aviv Regev. A module map showing conditional activity of expression modules in cancer. *Nature Genetics*, 36(10):1090–1098, October 2004. ISSN 1061-4036. doi: 10.1038/ng1434. 1.1

R. Shen, A. B. Olshen, and M. Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, November 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp543. URL `http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp543`. 1.1

Ronglai Shen, Sijian Wang, and Qianxing Mo. Sparse integrative clustering of multiple omics data sets. *The Annals of Applied Statistics*, 7(1):269–294, March 2013. ISSN 1932-6157. doi: 10.1214/12-AOAS578. URL http://projecteuclid.org/euclid.aoas/1365527199. 1.1

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013. 7

John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, August 2002. ISSN 13697412, 14679868. doi: 10.1111/1467-9868.00346. URL http://doi.wiley.com/10.1111/1467-9868.00346. 1.1

Joshua M. Stuart, Eran Segal, Daphne Koller, and Stuart K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)*, 302(5643):249–255, October 2003. ISSN 1095-9203. doi: 10.1126/science.1087447. 1.1

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):pp. 267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178. 1.1, 1.3.1

Robert Tibshirani, Iain Johnstone, Trevor Hastie, and Bradley Efron. Least angle regression. *The Annals of Statistics*, 32 (2):407–499, April 2004. ISSN 0090-5364. doi: 10.1214/009053604000000067. URL http://projecteuclid.org/Dienst/getRecord?id=euclid.aos/1083178935/. 1.3.1

Olga G. Troyanskaya, Kara Dolinski, Art B. Owen, Russ B. Altman, and David Botstein. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences of the United States of America*, 100(14):8348–8353, July 2003. ISSN 0027-8424. doi: 10.1073/pnas.0832373100. 1.1

Jacob Ulirsch, Cheng Fan, George Knafl, Ming Jing Wu, Brett Coleman, Charles M Perou, and Theresa Swift-Scanlan. Vimentin dna methylation predicts survival in breast cancer. *Breast cancer research and treatment*, 137(2):383–396, 2013. 6

Marc Vidal, Michael E. Cusick, and Albert-László Barabási. Interactome Networks and Human Disease. *Cell*, 144(6):986–998, March 2011. ISSN 0092-8674. doi: 10.1016/j.cell.2011.02.016. URL `http://www.cell.com/article/S0092867411001309/abstract`. 1.1

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00532.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x`. 1.3.1, 2.2, 3.2, 3.3

B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M.C. Chambers, L.J. Zimmerman, K.F. Shaddox, S. Kim, S.R. Davies, S. Wang, P. Wang, C.R. Kinsinger, R.C. Rivers, H. Rodriguez, R.R. Townsend, M.J. Ellis, S.A. Carr, D.L. Tabb, R.J. Coffey, R.J. Slebos, D.C. Liebler, and NCI CPTAC. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–387, 2014. 2.1

Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4:Article17, 2005. ISSN 1544-6115. doi: 10.2202/1544-6115.1128. 1.1

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x`. 1.3.1, 3.3