이학석사학위논문

Estimation of Sparse Cross Correlation Matrix

고차원 희소 교차상관행렬의 추정

2022 년 2 월

서울대학교 대학원

통계학과

Cao Yin

Estimation of Sparse Cross Correlation Matrix

고차원 희소 교차상관행렬의 추정

지도교수 임 요 한

이 논문을 이학석사 학위논문으로 제출함

2021 년 10 월

서울대학교 대학원

통계학과

Cao Yin

Cao Yin의 이학석사 학위논문을 인준함

2021 년 12 월

위 원 장	오희석	(인)
부위원장	임요한	(인)
위 원	PARK JUN YONG	(인)

Abstract

Estimation of Sparse Cross Correlation Matrix

Cao Yin

The Department of Statistics

The Graduate School

Seoul National University

In this thesis, we are motivated by an integrative study of multi-omics data and are interested in estimating the cross correlation matrix of two high dimensional random vectors. We rewrite the problem to a multiple testing problem and propose a new method to estimate it by testing individual components of the matrix simultaneously. We apply the proposed method to the integrative analysis of the protein expression data ($\mathbf{X}$) and the mRNA expression data ($\mathbf{Y}$) in TCGA breast cancer cohort.

Keywords: cross-correlation matrix, integrative analysis, local false discovery rate, multiple testing, multi-omics data

Student Number: 2020-26660

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The occurrence of high-dimensional data in a large amount of applications has prompted sustained interest in statistics in recent years. Statistical analysis of such high-dimensional data sometimes requires knowledge of covariance or correlation matrices with dimension far greater than the sample size. Examples include microarray analysis (Jaeger et al., 2003; Shedden and Taylor, 2004; Qiu and Yakovlev, 2007), financial risk management (Fan et al., 2008), and machine learning (Hastie et al., 2009). All of these applications include estimating variance-covariance matrices of one variable vector, but a lot of times researchers are more interested in finding the association between two mutually exclusive sets of variables. Estimation of cross correlation matrix $\mathbf{R_{XY}}$, the off-diagonal submatrix of correlation matrix, is highly involved in data integration problems, especially in the context of multi-omics studies. A typical example is measuring the same gene at two different molecular levels, with one set of data measure the molecular template synthesis of the other set of data (DNA to RNA, or RNA to protein). Using expression data for non-coding RNAs such

as microRNAs, coupled with mRNA and proteomics data, to reveal the degree of post-transcriptional regulation is another common scenario (Cheng et al., 2005). In this paper, we consider estimation and multiple testing of cross correlation matrix with the structural assumption - sparse cross correlation matrix, that is, most entries are zero (Bickel and Levina, 2008; Rothman et al., 2009; Cai and Liu, 2011; Wang and Fan, 2017).

Multiple testing of covariance structures is a widely used methodology in analysis of high-dimensional data. Liu (2013) considers multiple testing for partial correlations under a Gaussian graphical model. Cai and Liu (2016) proposed methods for simultaneous testing of correlations. Xia et al. (2015) proposed methods for differential network analysis. Aimed for detecting significant correlations between variables, large-scale multiple testing for correlations is an important area in statistics with a wide range of applications including gene expression (Carter et al., 2004; Dubois et al., 2010), spatial epidemiology (Elliott and Wartenberg, 2004), and brain imaging (Bennett et al., 2009; Lindquist and Mejia, 2005). The null hypotheses are usually

$$H_{0jk} : \rho_{jk} = 0,$$

where ρ_{jk} is the correlation between variable X_j and Y_k for $1 \leqslant j \leqslant p$, and $1 \leqslant k \leqslant q$. With thousands or even millions of tests to perform at the same time, it becomes challenging to control the overall Type I error rate while maintaining the desired power due to complicated dependence structures. In high-dimensional studies, controlling the false discovery rate (FDR), the proportion of falsely rejected hypotheses among all rejected hypotheses, becomes a common goal.

Methods of controlling FDR has been developed by a lot of researchers since its first proposal by Benjamini and Hochberg in 1995. Under the assumption

that test statistics are independent, the BH step-up procedure (Benjamini and Hochberg, 1995) controls FDR by thresholding the p-values of each individual test. Storey (2002) introduced the q-value which estimates the FDR for a given cutoff value. Efron (2004) proposed an empirical Bayes analysis method to examine the local false discovery rate. However, in the presence of strong correlation, particularly when the matrices are sparse, the situation becomes more difficult. Multiple testing procedures are very unstable when test statistics are correlated because they have a high variability of the number of false and true discoveries from sample to sample (Qiu et al., 2005). Some multiple testing adjustment methods dealing with certain dependence types include Benjamini and Yekutieli (2001) and Fan et al. (2012).

In this paper, we propose a multiple testing procedure for cross correlations. We start from the sample correlation coefficient r_{jk} and use Fisher's z-transformation to construct the test statistic z_{jk} for testing an individual hypothesis H_{0jk}. We then use local false discovery rate procedure to perform multiple testing. As a comparison of simulation performance, we apply both our procedure and procedure proposed by Cai and Liu (2016) to breast cancer cohorts with paired proteomic data ($\mathbf{X}$) and transcriptomic data ($\mathbf{Y}$). We identify significant correlation pairs for both procedures. The resulting cross correlation matrix of our procedure has a higher coverage rate of known transcription regulatory networks catalogued in the cancer cell biology literature.

The rest of the paper is organized as follows. In Section 2, we review the large-scale multiple testing procedure proposed by Cai and Liu (2016) as well as some other FDR control procedure. In Section 3, we give a detailed description of our procedure. A comparison between the method proposed and that of Cai and Liu (2016) numerically using breast cancer data is also discussed in this section. We conclude the paper with a few remarks for the proposed procedure

in Section 4.

Chapter 2

Review

2.1 Cross covariance matrix and correlation matrix

Suppose for subject $i = 1, ..., n$, we observed a vector pair $(\mathbf{X}_i, \mathbf{Y}_i)$, where $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top$ and $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iq})^\top$ are two random vectors with dimension p and q, respectively. We assume the data $\mathbf{Z}_i = (\mathbf{X}_i^\top, \mathbf{Y}_i^\top)^\top$ for each subject follows the multivariate normal distribution with mean and variance

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_\mathbf{X} \\ \boldsymbol{\mu}_\mathbf{Y} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{XX}} & \boldsymbol{\Sigma}_{\mathbf{XY}} \\ \boldsymbol{\Sigma}_{\mathbf{YX}} & \boldsymbol{\Sigma}_{\mathbf{YY}} \end{pmatrix}.$$

The mean vectors $\boldsymbol{\mu}_\mathbf{X}$ and $\boldsymbol{\mu}_\mathbf{Y}$ have length p and q, respectively. The covariance matrices $\boldsymbol{\Sigma}_{\mathbf{XX}}$, $\boldsymbol{\Sigma}_{\mathbf{XY}}$ and $\boldsymbol{\Sigma}_{\mathbf{YY}}$ are of size $p \times p$, $p \times q$ and $q \times q$ respectively. We further arrange $\mathbf{X}_i$ of all subjects into one matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ so that each row of $\mathbf{X}$ contains data $\mathbf{X}_i^\top$ for subject i. Similarly for $\mathbf{Y}_i$, we have matrix $\mathbf{Y} \in \mathbb{R}^{n \times q}$.

The resulting matrices can be represented as follows

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \\ \vdots \\ \mathbf{X}_n^\top \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix},$$

and

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1^\top \\ \mathbf{Y}_2^\top \\ \vdots \\ \mathbf{Y}_n^\top \end{pmatrix} = \begin{pmatrix} Y_{11} & Y_{12} & \cdots & Y_{1q} \\ Y_{21} & Y_{22} & \cdots & Y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nq} \end{pmatrix}.$$

We are interested in the simultaneous correlation tests between X_j and Y_k,

$$H_{0jk} : \mathrm{cov}(X_j, Y_k) = 0 \quad \text{versus} \quad H_{1jk} : \mathrm{cov}(X_j, Y_k) \neq 0,$$

for $1 \leqslant j \leqslant p$ and $1 \leqslant k \leqslant q$. That is to say, we will apply multiple testing procedure to find non-zero covariance pairs while controlling the false discovery rate, the proportion of falsely rejected hypotheses among all rejected hypotheses at given level α, at the same time.

2.2 Procedure by Cai and Liu (2016)

Cai and Liu (2011; 2016) proposed an adaptive thresholding method for sparse covariance matrix estimation and a large-scale multiple testing procedure for correlations in one sample case. In order to use their method, we rewrite the paired vector data $(\mathbf{X}_i, \mathbf{Y}_i)$ as $\mathbf{Z}_i = (\mathbf{X}_i^\top, \mathbf{Y}_i^\top)^\top$, a single vector of length $p+q$. The procedure simultaneously tests the hypotheses

$$H_{0jk} : \sigma_{jk} = 0 \quad \text{versus} \quad H_{1jk} : \sigma_{jk} \neq 0,$$

for $1 \leqslant j < k \leqslant p+q$. They suggest using the test statistic

$$T_{jk} = \frac{\sum_{i=1}^{n}(Z_{ij} - \bar{Z}_j)(Z_{ik} - \bar{Z}_k)}{\sqrt{n\hat{\theta}_{jk}}},$$

where

$$\bar{Z}_j = \frac{1}{n}\sum_{i=1}^{n} Z_{ij},$$

$$\hat{\theta}_{jk} = \frac{1}{n}\sum_{i=1}^{n}[(Z_{ij} - \bar{Z}_j)(Z_{ik} - \bar{Z}_k) - \hat{\sigma}_{jk}]^2,$$

$$\hat{\sigma}_{jk} = \frac{1}{n}\sum_{i=1}^{n}(Z_{ij} - \bar{Z}_j)(Z_{ik} - \bar{Z}_k).$$

Let $0 < \alpha < 1$, the threshold level is defined as

$$\hat{t} = \inf\left\{0 \leqslant t \leqslant \sqrt{4\log p - 2\log\log p} : \frac{G(t)(p^2-p)/2}{\max\{\sum_{1\leqslant j<k\leqslant p+q} I(|T_{jk}| \geqslant t), 1\}} \leqslant \alpha\right\},$$

where $G(t) = 2 - 2\Phi(t)$. If $\hat{t}$ does not exist, they set $\hat{t} = \sqrt{4\log p}$. The procedure rejects H_{0jk} whenever $|T_{jk}| \geqslant \hat{t}$.

2.3 Multiple testing

Multiple testing is a statistical analysis involving a set of tests simultaneously. In general, if m mutually independent tests are each conducted at α level, the probability of making at least one Type I error is $1 - (1-\alpha)^m$. As the number of tests being conducted increases, the probability of at least one Type I error increases. Over the years, different strategies have been proposed to address for the problem of multiplicity. These methods usually require a stringent significance level with which each individual hypothesis can be rejected.

The family-wise error rate (FWER), defined as FWER = $\mathrm{P}(V \geqslant 1)$, has been widely used to account for the problem of multiplicity. The Bonferroni correction provides the classic FWER control method. It tests each hypothesis

	Null is true	Alternative is true	Total
Declared significant	V	S	R
Declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

Table 2.1 Classification of tested hypotheses

at level α/m so that the FWER is guaranteed not exceed the prespecified level α. A review of FWER procedures is give by Hochberg and Tamhane (1987) and Shaffer (1995).

2.3.1 False discovery rate

As the number of tests increases, the power to reject an alternative hypothesis while controlling FWER at the same time is greatly reduced. The false discovery rate (FDR), or expected proportion of false rejections among all rejections, is an alternative to FWER in multiple testing control. It has been showed that the FDR has greater power to find true discoveries while still controlling the proportion of Type I errors at α. Using the notation in Table 2.1, the FDR is defined as

$$\text{FDR} = E(\frac{V}{R}|R > 0).$$

FDR is zero when no hypothesis is rejected.

2.3.2 BH step-up procedure

A common technique for controlling the FDR is provided by Benjamini and Hochberg (1995). Consider testing simultaneously m null hypotheses $H_1, H_2, ..., H_m$ with $p_1, p_2, ..., p_m$ their corresponding p-values. Let $p_{(1)} \leqslant p_{(2)} \leqslant ... \leqslant p_{(m)}$ be ordered p-values, and denote $H_{(i)}$ the null hypothesis corresponding to $p_{(i)}$. The

BH procedure is the following step-up procedure:

$$\text{Let } k = \max\{i : p_{(i)} \leqslant i\alpha/m\}, \text{then reject all } H_{(i)} \text{for } i = 1, 2, \ldots, k.$$

When the test statistics are independent, the BH procedure controls the FDR at level α. The procedure does not need any assumption of p-value distribution; it controls the FDR regardless of the distribution of p-values. However, without the distribution information in the sample, BH (2000) argued that the procedure is conservative when some of the hypotheses are from non-null distributions. In fact, the BH step-up procedure controls the FDR at level $(1-p)\alpha$, where p is the proportion of non-nulls.

2.3.3 Storey's q-value procedure

Realizing the conservativeness of the BH step-up procedure, Storey (2002) introduced the positive False Discovery Rate (pFDR) and the q-value. Storey's approach uses the information of p and estimated the FDR for a given cutoff, contrary to the BH step-up procedure, where level α is fixed and cutoff values are estimated.

Let p be the proportion of non-nulls and G be the marginal distribution of the p-value. For a given p-value cutoff λ, the pFDR is defined as

$$\text{pFDR}(\lambda) = E(\frac{V}{R}|R > 0) = \frac{(1-p)\lambda}{G(\lambda)}.$$

For a set of m hypotheses with independent p-values and rejection region $[0, \gamma]$, the q-value is the minimum pFDR level such that a hypothesis with p-value p_i is just rejected, that is

$$q(p_i) = \inf_{\gamma \geqslant p_i} \{\text{pFDR}(\gamma)\} = \inf_{\gamma \geqslant p_i} \left\{ \frac{(1-p)\gamma}{G(\gamma)} \right\}.$$

Chapter 3

Estimation sparse correlation matrix

3.1 Procedure

In this section, we propose a large-scale multiple testing procedure for estimating sparse cross correlation matrices. We first construct a test statistic for testing no correlation between each pair (X_j, Y_k), $H_{0jk} : \sigma_{jk} = 0$, so that the constructed test statistic asymptotically follows a standard normal distribution under the null hypothesis H_{0jk}. Then we use the local false discovery rate to handle the problem of multiplicity when testing a large number of hypotheses. The overall FDR is controlled under given level α.

The typical statistic for correlation detection is the sample correlation coefficient, r_{jk}, which is defined as

$$r_{jk} = \frac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(Y_{ik} - \bar{Y}_k)}{\sqrt{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}\sqrt{\sum_{i=1}^{n}(Y_{ik} - \bar{Y}_k)^2}},$$

where $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$, and $\bar{Y}_k = \frac{1}{n}\sum_{i=1}^{n} Y_{ik}$.

Since the variance of sample correlation coefficient becomes smaller as the population correlation coefficient gets closer to ± 1, we use the variance stabilization method, Fisher's z-transformation, so that the resulting variable approximately follows a normal distribution with a variance that is stable for different values of correlation. Fisher's z-transformation of r_{jk} is defined as

$$F(r_{jk}) = \frac{1}{2}\ln\frac{1+r_{jk}}{1-r_{jk}},$$

where r_{jk} is the sample correlation coefficient. Under the condition of $(\mathbf{X}_i^\top, \mathbf{Y}_i^\top)^\top$ follows a multivariate normal distribution, it has been showed that $F(r_{jk})$ approximately follows a normal distribution (for large samples, $n > 50$) with mean $\mu = 0$ and standard deviation $\sigma = \frac{1}{\sqrt{n-3}}$,where n is the sample size.

Using the approximation, the following statistic is standardized normal

$$z_{jk} = \frac{F(r_{jk}) - \mu}{\sigma} = \sqrt{n-3}F(r_{jk}) \to N(0,1).$$

We will use z_{jk} as the test statistics and then apply local false discovery rate procedure to those z values.

3.1.1 Local false discovery rate

The traditional FDR calculates a rate applying generally to all hypotheses in the same rejection region. In practical application, the fact that some test statistics are much more extreme than others, or to say, that not all hypotheses are equally likely to contribute the false discoveries makes the FDR a somewhat unsatisfying metric.

The local false discovery rate proposed by Efron (2004) extends the concept of FDR to give a posterior probability at the single hypothesis level. It is a Bayes version of Benjamini and Hochberg (1995)'s procedure focusing on densities rather than tail areas.

Suppose m null hypotheses, each with its own test statistic, are test simultaneously

$$\text{Null hypotheses:} \quad H_{01}, H_{02}, \ldots, H_{0i}, \ldots, H_{0m}$$

$$\text{Test statistics:} \quad z_1, z_2, \ldots, z_i, \ldots, z_m.$$

Assume each of m hypotheses is either null with prior probability p_0 and density $f_0(z)$ or non-null with prior probability $p_1 = 1 - p_0$ and density $f_1(z)$

$$p_0 = \Pr(\text{null is true}) \qquad \text{density} = f_0(z) \quad \text{if null}$$

$$p_1 = \Pr(\text{non-null is true}) \qquad \text{density} = f_1(z) \quad \text{if non-null.}$$

Define the mixture density

$$f(z) = p_0 f_0(z) + p_1 f_1(z).$$

The local false discovery rate is the posterior probability that a case is null given that we observed test statistic z. Using Bayes rule, it can be expressed as

$$\text{fdr}(z) = \text{P}(\text{null} \mid z) = \frac{p_0 f_0(z)}{f(z)}.$$

In our procedure, the test statistics are z_{jk}'s for $j = 1, 2, ..., p$ and $k = 1, 2, ..., q$. The usual cutoff threshold is fdr $\leqslant 0.2$.

3.1.2 fdr Estimation

Mixture Density Estimation

Assume the distribution of z values are smooth, Efron (2005) estimate the mixture density $f(z)$ with Poisson regression using Lindsey's method. The range of the sample $z_1, \ldots, z_m$ is divided into K equal intervals, with s_k being the number of z values in interval k, and $z_{(k)}$ being the midpoint of interval k. The Lindsey's method assumes counts s_k follow an independent Poisson distribution,

$$s_k \overset{ind}{\sim} Poi(\lambda_k) \quad k = 1, 2, \ldots, K$$

with

$$\lambda_k = m\Delta f(z_{(k)}),$$

where Δ is the width of interval.

The method estimates $log(\lambda_k)$ with a pth degree polynomial function of $z_{(k)}$, so that the mixture density $f(z)$ can be estimated by maximum likelihood of the following function

$$f(z) = exp\Big\{ \sum_{j=0}^{p} \beta_j z^j \Big\}$$

satisfying $\int f(z) = 1$.

Efron (2005) also remarked that Lindsey's method with a Poisson regression is almost efficient for estimating $f(z)$ when z_i's are independent. Although under most cases z_i's are dependent and over dispersed, Lindsey's method will still be nearly unbiased at the cost of losing estimating efficiency.

Empirical Null Estimation

The theoretical null distribution $z_i \sim N(0, 1)$ is usually used in individual hypothesis test. With thousands of z values to exam at once, the conventional theoretical null may be inappropriate for the situation in large-scale hypothesis testing. Estimating the empirical null distribution adjusts the theoretical null for the dataset at hand.

Efron and Hastie (2016) assume the two-class model with $f_0(z)$ normal

$$f_0(z) \sim N(\delta_0, \sigma_0^2).$$

To estimate the three parameters $(\delta_0, \sigma_0, p_0)$, the mean and standard deviation of the null density and the proportion of null cases, Efron and Hastie (2016) make the zero assumption that p_0 is large, and that most of the z_i near 0 are

null cases. R-package *locfdr* (Efron et al., 2005; Efron, 2016) uses the following steps to estimate the null distribution: let $\mathcal{A}_0$ be the set near 0, and let

$$\boldsymbol{z_0} = \{z_i : z_i \in \mathcal{A}_0, i = 1, 2, \ldots, m\},$$

$$\mathcal{I}_0 = \{i : z_i \in \mathcal{A}_0, i = 1, 2, \ldots, m\},$$

$$m_0 = |z_0|.$$

Define

$$\phi_{\delta_0,\sigma_0}(z) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(z-\delta_0)^2}{2\delta_0^2}},$$

$$P(\delta_0, \sigma_0) = \int_{\mathcal{A}_0} \phi_{\delta_0,\sigma_0}(z)dz,$$

and

$$\theta = p_0 P(\delta_0, \sigma_0).$$

Then the density of $\boldsymbol{z_0}$ is the product of two terms: probability of having m_0 of z_i in $\mathcal{A}_0$, and conditional probability of those z_i in $\mathcal{A}_0$,

$$f_{\delta_0,\sigma_0,p_0}(\boldsymbol{z_0}) = \left[\binom{m}{m_0}\theta^{m_0}(1-\theta)^{m-m_0}\right]\left[\prod_{\mathcal{I}_0}\frac{\phi_{\delta_0,\sigma_0}(z_i)}{P(\delta_0,\sigma_0)}\right].$$

Maximum likelihood based on the above density gives the empirical null estimates $(\hat{\delta}_0, \hat{\sigma}_0)$. $\hat{\theta} = \frac{m_0}{m}$ can be obtained from the first binomial probability term, so then $\hat{p}_0 = \frac{\hat{\theta}}{P(\hat{\delta}_0,\hat{\sigma}_0)}$.

3.2 Data

We next applied the proposed method to integrative analysis of the protein expression data ($\mathbf{X}$) and the mRNA expression data ($\mathbf{Y}$) in TCGA breast cancer cohort, with group information representing the co-regulation of gene expression by complexes of transcription factor proteins. In total, 76 subjects have both transcriptomics and proteomics data as distributed through the data portals

of TCGA and Clinical Proteomic Tumor Analysis Consortium (CPTAC). In invasive ductal carcinomas, the gene expression variation across patients is well known to be determined by the expression level of the estrogen receptor (ER) protein in the tumor (Rosato et al., 2018), which in turn acts as a nuclear transcription factor and drives gene expression program for cell proliferation. As a benchmark analysis, we first aimed to verify that the non-zero elements of the cross covariance matrix between the transcription factor and co-activator proteins (denoted by TFA hereafter) and the mRNA expression levels of their target genes are the most pronounced variation in the data.

We capitalized on the fact that the TFAs are assembled into protein complexes while in action, and thus hypothesized that utilizing the protein-protein interaction will allow us to first identify the TFA groups associated with large variation in the proteomics data, and their target gene expression levels should be consistently reflected in the transcriptomics data. To this end, we collected *bona fide* protein-protein interaction data from credible sources (Razick et al., 2008; Huttin et al., 2015) for the human TFA proteins (1195 proteins), which have been known to regulate as many as 3114 target genes according to the TF and regulatory element databases such as TRED (Zhao et al., 2005), ITFP (Zheng et al., 2008), ENCODE, and TRRUST (Han et al., 2015).

3.3 Results

Figure 3.1 shows the histogram of the $1195 \times 3114 = 3,121,230$ z-values. The green curve, $f(z)$, is the Poisson regression fit to the histogram counts. Curve $f(z)$ emphasizes the central peak around $z = 0$, showing that a large proportion of (TFA, mRNA) pairs are not correlated. The blue dashed curve is the density $p_0 f_0$ estimated by MLE. Both the MLE and central matching estimates (CME)

give nearly close approximation of null distribution $N(0, 1)$.

Our procedure of estimating cross correlation matrix uses fdr cutoff value 0.1. More than 99.9% of the entries are penalized to zero, resulting in a sparse estimate of correlation matrix. A total of 60,693 (TFA, mRNA) pairs have non-zero correlation, with more than 89% pairs having correlation values less than $|0.5|$ and around one hundred pairs having large correlations.

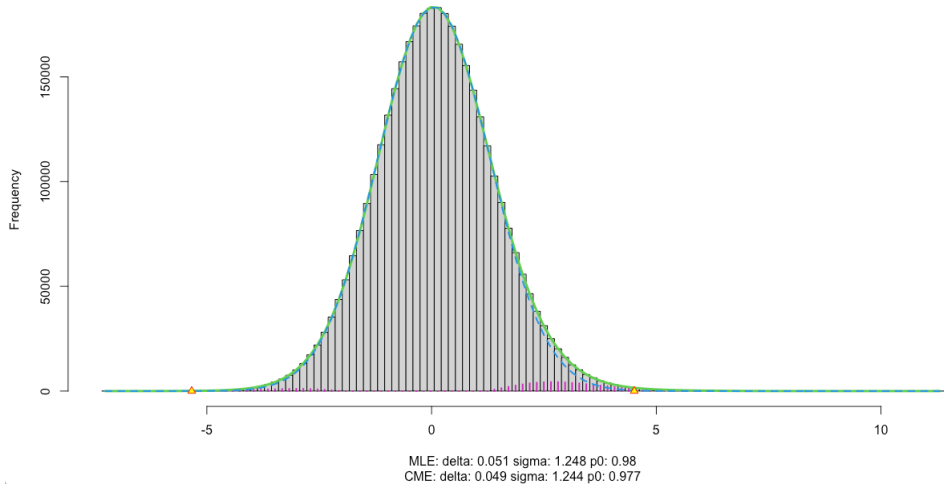

Figure 3.1 Histogram of z-values

We also estimated sparse cross correlation matrix using the adaptive thresholding procedure proposed by Cai and Liu (2016). Since their procedure is designed for testing correlation between elements of one vector from one sample, we put together the transcriptomics data ($q = 3114$) and the proteomics data ($p = 1195$) of all subjects into a single matrix $\mathbf{Z}$, and estimated sparse variance-covariance matrix of the entire data first, and took the submatrix corresponding to the cross covariance matrix after the whole estimation process. A total of 163,726 pairs are found significant or have non-zero covariance. Among non-zero values, more than 59% are between -0.1 and 0.1, suggesting that the

cross covariance matrix has relatively small values compared to the values of variance-covariance matrix. The problem of over-penalization of cross correlation matrix arises: we only need a $p \times q$ part of the variance-covariance matrix but in fact we used values of the whole matrix when deciding thresholds.

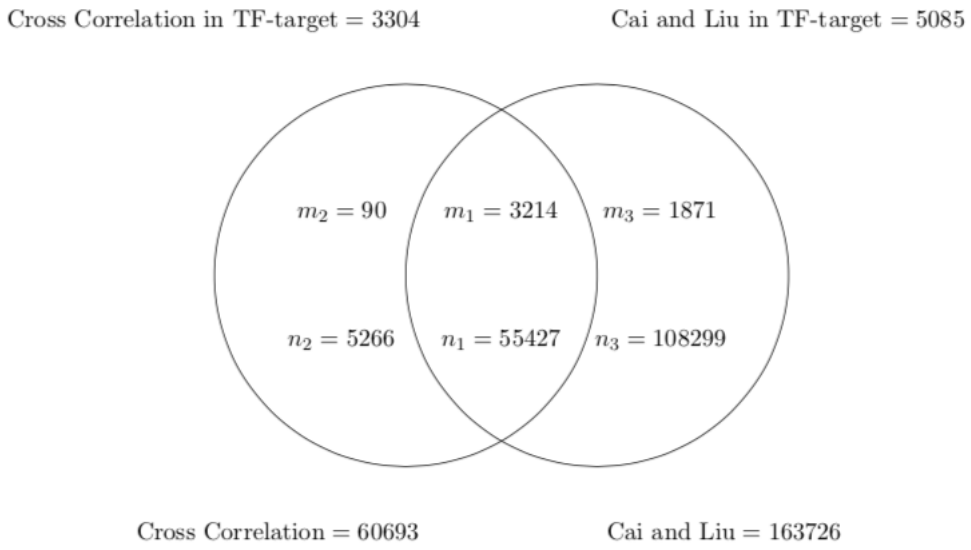

Figure 3.2 Venn Diagram

As a part of procedure accuracy measurement process, we benchmark (TFA, mRNA) pairs with non-zero correlation against the known transcription regulatory networks, and compare the coverage rate between two procedures. The Venn diagram showing the number of non-zero correlation (TFA, mRNA) pairs with and without benchmark for both procedures is given by Figure 3.2. The TF-target pairs are benchmark pairs used. For our procedure, for example, a total of 60,693 (TFA, mRNA) pairs have non-zero correlation, and among these 3304 pairs are also in the known transcription regulatory literature database. The pairs with non-zero correlation founded using our procedure have a higher

proportion that overlaps literature-based regulation, almost two times than the overlap rate of adaptive thresholding procedure. The adaptive thresholding procedure produced a substantial amount of unique non-zero correlation pairs ($n_3 = 108,299$), more than 60% ($\frac{n_3}{n_1+n_3}$) of its all non-zero correlation pairs compared to about 10% ($\frac{n_2}{n_1+n_2}$) using our procedure. However, the proportion of unique non-zero correlation pairs under benchmark among all unique non-zero correlation pairs ($\frac{m_2}{n_2}$ and $\frac{m_3}{n_3}$) are nearly the same, around 1.7%, suggesting that the adaptive thresholding procedure is not efficient in finding unique pairs.

Chapter 4

Conclusion

In this thesis, we propose a new method to estimate the cross-correlation matrix of $\mathbf{R_{XY}}$ of two random vectors $\mathbf{X}$ and $\mathbf{Y}$ based on a multiple testing procedure. The new method rewrites the problem as a multiple testing problem, and estimate the support by testing individual hypotheses on ρ_{jk}s. In doing so, we adapt the Efron's local false discovery rate procedure (Efron, 2004) to test the hypotheses simultaneously. Using the analysis of breast cancer data in TCGA, we show the procedure performs better than Cai and Liu (2016)'s procedure. However, with the recent advances in multiple testing literature, we may be able to refine our procedure in this thesis. We leave this as our next step.

Bibliography

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, **25**, 60–83.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

Bennett, C. M., Wolford, G. L., and Miller, M. B. (2009). The principled control of false positives in neuroimaging. *Social Cognitive and Affecive Neuroscience*, **4**, 417–422.

Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, **36**, 2577–2604.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, **106**, 672–684.

Cai, T. and Liu, W. (2016). Large-scale multiple testing of correlations. *Journal of the American Statistical Association*, **111**, 229–240.

Carter, S. L., Brechbühler, C. M., Griffin, M., and Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, **20**, 2242–2250.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., et al. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–1154.

Dubois, P. C., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G. A. R., et al.(2010). Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, **42**, 295–302.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99**, 96–104.

Efron, B. (2005). Local false discovery rates. URL: `http://statweb.stanford.edu/ ckirby/brad/papers/2005LocalFDR.pdf`.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, New York.

Efron, B., Turnbull, B., Narasimhan, B., and Strimmer, K. (2005). locfdr: Computes Local False Discovery Rates. URL: `https://CRAN.R-project.org/package=locfdr`.

Elliott, P. and Wartenberg, D. (2004). Review Spatial epidemiology: current approaches and future challenges. *Environmental Health Perspectives*, **112**, 998–1006.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, **147**, 186–197.

Fan, J., Han, X., and Gu, W.(2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*,**107**, 1019–1035.

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., et al. (2015). TRRUST: a reference database of human transcriptional regulatory interactions. *Scientific Reports*, **5**, 11432.

Hastie, T. J., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed). Springer, New York.

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., et al. (2015). The bioplex network: A systematic exploration of the human interactome. *Cell*, **162**, 425–440.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley & Sons, New York.

Jaeger, J., Sengupta, R., and Ruzzo, W. L. (2003). Improved gene selection for classification of microarrays. *Pacific Symposium on Biocomputing*, **8**, 53–64.

Lindquist, M. A. and Mejia, A. (2015). Zen and the art of multiple comparisons. *Psychosomatic Medicine*,**77**, 114–125.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, **41**, 2948–2978.

Qiu, X., Klebanov, L., and Yakovlev, A. Y. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding

differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*, **4**, 1–32.

Qiu, X. and Yakovlev, A. (2007). Comments on probabilistic models behind the concept of false discovery rate. *Journal of Bioinformatics and Computational Biology*, **5**, 963–975.

Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex: A consolidated protein interaction database with provenance. *BMC Bioinformatics*, **9**, 405.

Rosato, A., Tenori, L., Cascante, M., De Atauri Carulla, P. R., Martins Dos Santos, V. A., and Saccenti, E. (2018). From correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics*, **14**, 37.

Rothman, A., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, **104**, 177–186.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, **46**, 561–584.

Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., Evans, A., Rapoport, J., and Giedd, J. (2006). Intellectual ability and cortical development in children and adolescents. *Nature*, **440**, 676–679

Shedden, K. and Taylor, J. (2004). Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas. *Methods of Microarray Data Analysis IV*. Springer, New York.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479–498.

Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, **45**, 1342–1374.

Xia, Y., Cai, T., and Cai, T. T. (2015). Testing differential networks with applications to detecting gene-by-gene interactions. *Biometrika*, **102**, 247–266.

Zhao, F., Xuan, Z., Liu, L., and Zhang, M. Q. (2005). TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. *Nucleic Acids Research*, **33**, D103–D107.

Zheng, G., Tu, K., Yang, Q., Xiong, Y.,Wei, C., Xie, L., Zhu, Y., and Li, Y. (2008). ITFP: an integrated platform of mammalian transcription factors. *Bioinformatics*, **24**, 2416–2417.

국문초록

이 논문에서, 우리는 다중 오믹스 데이터에 대한 통합 연구를 통해 동기를 부여받았으며 두 개의 고차원 무작위 벡터의 교차 상관 행렬을 추정하는 데 관심이 있다. 우리는 문제를 다중 테스트 문제로 다시 작성하고 매트릭스의 개별 구성 요소를 동시에 테스트하여 추정하는 새로운 방법을 제안한다. 제안된 방법을 TCGA 유방암 코호트에서 단백질 발현 데이터($\mathbf{X}$)와 mRNA 발현 데이터($\mathbf{Y}$)의 통합 분석에 적용한다.

주요어: cross-correlation matrix, integrative analysis, local false discovery rate, multiple testing, multi-omics data

학번: 2020-26660