Communications for Statistical Applications and Methods
2022, Vol. 29, No. 6, 655–664

https://doi.org/10.29220/CSAM.2022.29.6.655
Print ISSN 2287-7843 / Online ISSN 2383-4757

Estimation of high-dimensional sparse cross correlation matrix

Yin Cao[a], Kwangok Seo[a], Soohyun Ahn[1,b], Johan Lim[a]

[a]Department of Statistics, Seoul National University, Korea;
[b]Department of Mathematics, Ajou University, Korea

Abstract

On the motivation by an integrative study of multi-omics data, we are interested in estimating the structure of the sparse cross correlation matrix of two high-dimensional random vectors. We rewrite the problem as a multiple testing problem and propose a new method to estimate the sparse structure of the cross correlation matrix. To do so, we test the correlation coefficients simultaneously and threshold the correlation coefficients by controlling FRD at a predetermined level α. Further, we apply the proposed method and an alternative adaptive thresholding procedure by Cai and Liu (2016) to the integrative analysis of the protein expression data (X) and the mRNA expression data (Y) in TCGA breast cancer cohort. By varying the FDR level α, we show that the new procedure is consistently more efficient in estimating the sparse structure of cross correlation matrix than the alternative one.

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

1. Introduction

The occurrence of high-dimensional data in a various field has increased and sustained interest in Statistics in recent years. Statistical analysis of such high-dimensional data often requires the knowledge of covariance or correlation matrix, whose dimension p be sometimes much larger than the sample size n. There are various examples including microarray analysis (Jaeger *et al.*, 2003; Shedden and Taylor, 2005), financial risk management (Fan *et al.*, 2008), and brain functional connectivity analysis (Shaw *et al.*, 2006). All these applications include estimating the covariance or correlation matrix of one variable vector, but a lot of researchers are more interested in finding the association between two mutually exclusive sets of variables. In the correlation matrix, the off-diagonal submatrix $\mathbf{R}_{XY}$ is called cross correlation matrix and its estimation 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 (e.g., DNA and RNA, or RNA and protein). Using expression data for non-coding RNAs such as microRNAs, to reveal the degree of post-transcriptional regulation is another common scenario (Cheng *et al.*, 2005).

In this paper, we consider the estimation of cross correlation matrix with the sparsity assumption under which most entries are zero (Bickel and Levina, 2008; Cai and Liu, 2011; Wang and Fan, 2017). The problem of estimation of a sparse cross correlation matrix can be devided into two parts: (i) the

Yin Cao and Kwangok Seo contributed equally to this research.

This research was supported by the National Research Foundation of Korea (NRF-2019R1F1A1056779).

[1] Corresponding author: Department of Mathematics, Ajou University, 206 World cup-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16499, Korea. E-mail: shahn@ajou.ac.kr

Published 30 November 2022 / journal homepage: http://csam.or.kr

estimation of the elemants of the cross correlation matrix and (ii) the estimation of the sparse structure of the cross correlation matrix. In general, the estimation of the elements of the cross correlation matrix is based on the sample correlation. Thus, we accept this convention and focus on the estimation of the sparse structure.

Aimed for detecting significant correlations between variables, large-scale multiple testing for correlation structure is an important area in Statistics and 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) propose a method for simultaneous testing of correlations. Xia *et al.* (2015) and Yu *et al.* (2018) propose a method for differential network analysis. There are a wide range of applications including gene expression (Dubois *et al.*, 2010), spatial epidemiology (Elliott and Wartenberg, 2004), and brain imaging (Bennett *et al.*, 2009). In those studies, the null hypotheses are usually

$$H_{0jk} : \rho_{jk} = 0, \tag{1.1}$$

where ρ_{jk} is the correlation coefficient between variables 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 simultaneously, 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 many researchers since it was first introduced by Benjamini and Hochberg (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) introduces the q-value which estimates FDR for a given cutoff value. Efron (2004) proposes an empirical Bayes method to examine the local false discovery rate. On the other hand, There are some multiple testing adjustment methods dealing with certain dependence assumptions including Benjamini and Yekutieli (2001) and Fan *et al.* (2012).

In this paper, we propose a new procedure to estimate the sparse structure of the cross correlation matrix via a multiple testing procedure. We start from the sample correlation coefficient r_{jk} and suggest to threshold r_{jk}s universally at the level t as

$$\hat{\rho}_{jk} = r_{jk} \cdot \mathrm{I}\left(\left|r_{jk}\right| \geq t\right), \quad j = 1, 2, \ldots, p,\ k = 1, 2, \ldots, q, \tag{1.2}$$

where $\mathrm{I}(A)$ is an indicator function for the event A. For the decision of thresholding level t, we consider the multiple testing problem (1.1) with the test statistic z_{jk}, the Fisher's z-transformation of r_{jk} for an individual null hypothesis H_{0jk}. The level t is set to control the local false discovery rate at a prespecified level. To evaluate and compare the performance, we apply both new procedure and an alternative one by Cai and Liu (2016) to paired proteomic data ($\mathbf{X}$) and transcriptomic data ($\mathbf{Y}$) in TCGA breast cancer cohort. Then, we identify the significantly correlated pairs for both procedures and show that the estimated cross correlation matrix by new 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 adaptive thresholding procedure developed by Cai and Liu (2016) and propose new procedure to estimate the sparse structure of high-dimensional cross correlation matrix. In Section 3, a comparison between the methods, new procedure and that of Cai and Liu (2016), using breast cancer data is evaluated. Finally, we conclude the paper with a brief summary in Section 4.

2. Methods

To begin with, let present the problem with data structure we are interested in this paper and review the existing method proposed by Cai and Liu (2016). Lastly, we describe the new procedure for estimating the sparse structure of cross correlation matrix.

2.1. Cross correlation matrix

Suppose we observed a vector pair $(\mathbf{X}_i, \mathbf{Y}_i)$ for subject $i = 1, \ldots, n$, 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. We assume that the data $\mathbf{Z}_i = (\mathbf{X}_i^\top, \mathbf{Y}_i^\top)^\top$ for each subject follows the multivariate normal distribution with mean vector and correlation matrix as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} \mathbf{R}_{XX} & \mathbf{R}_{XY} \\ \mathbf{R}_{YX} & \mathbf{R}_{YY} \end{pmatrix},$$

where $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$ are mean vectors with length p and q and $\mathbf{R}_{XX}$, $\mathbf{R}_{XY}$ and $\mathbf{R}_{YY}$ are submatrices of covariance matrix $\mathbf{R}$ with size $p \times p$, $p \times q$ and $q \times q$, respectively. Further, we have data $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$ 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

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

The estimator of sparse cross correlation matrix $\mathbf{R}_{XY}$ is the thresholding estimator given in (1.2). To decide the thresholding level t, we are interested in the simultaneous correlation tests between variables X_j and Y_k for $1 \leqslant j \leqslant p$ and $1 \leqslant k \leqslant q$. That is, we will apply multiple testing procedure to find the non-zero correlation pairs while controlling the FDR at a given level α.

2.2. Adaptive thresholding procedure

Cai and Liu (2011, 2016) propose an adaptive thresholding method for sparse cross correlation matrix estimation and a large-scale multiple testing procedure for correlations in one sample case. With the two random vectors $\mathbf{X}$ and $\mathbf{Y}$ with dimension p and q respectively, the procedure simultaneously tests the hypotheses

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

which is equivalent to the hypothesis (1.1) for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, q$. They suggest using the test statistic

$$T_{jk} = \frac{\sum_{i=1}^n \left(X_{ij} - \bar{X}_j\right)\left(Y_{ik} - \bar{Y}_k\right)}{\sqrt{n\hat{\theta}_{jk}}},$$

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

$$\hat{\theta}_{jk} = \frac{1}{n} \sum_{i=1}^{n} \left[\left(X_{ij} - \bar{X}_j \right) \left(Y_{ik} - \bar{Y}_k \right) - \hat{\sigma}_{jk} \right]^2$$

and

$$\hat{\sigma}_{jk} = \frac{1}{n} \sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j \right) \left(Y_{ik} - \bar{Y}_k \right).$$

With a prespecified level $0 < \alpha < 1$, let t be the threshold level which is defined as

$$\hat{t}^{c} = \inf \left\{ 0 \leqslant t \leqslant b_{p,q} : \frac{G(t)\, pq}{\max \left\{ \sum_{j=1}^{p} \sum_{k=1}^{q} I\left(\left| T_{jk} \right| \geqslant t \right), 1 \right\}} \leqslant \alpha \right\}, \tag{2.1}$$

where $b_{p,q} = \sqrt{4 \log(p+q) - 2 \log \log(p+q)}$ and $G(t) = 2 - 2\Phi(t)$ when the sample size is large and defines a sparse structure of the cross correlation matrix. If the sample size is small, the null distribution of T_{jk}, or $G(t)$, can be approximated by bootstrap method (refer in (Cai and Liu, 2016)). If $\hat{t}^{c}$ does not exist, they set $\hat{t}^{c} = \sqrt{2 \log(pq)}$. Then the null hypothesis H_{0jk} is rejected whenever $|T_{jk}| \geqslant \hat{t}^{c}$, equivalently,

$$\left| r_{jk} \right| \geq \frac{\sqrt{\hat{\theta}_{jk}}}{\sqrt{n \hat{\sigma}_{jj} \hat{\sigma}_{kk}}} \cdot \hat{t}^{c},$$

where r_{jk} is the sample correlation coefficient and we have a cross correlation matrix estimator $\hat{\mathbf{R}}^{c}_{XY}$ with entries

$$\hat{\rho}^{c}_{jk} = r_{jk} \cdot \mathrm{I} \left(\left| r_{jk} \right| \geq \frac{\sqrt{\hat{\theta}_{jk}}}{\sqrt{n \hat{\sigma}_{jj} \hat{\sigma}_{kk}}} \cdot \hat{t}^{c} \right), \quad j = 1, 2, \ldots, p,\ k = 1, 2, \ldots, q. \tag{2.2}$$

2.3. New procedure

In this section, we propose a large-scale multiple testing procedure for estimating sparse cross correlation matrix. We first construct a test statistic for testing no correlation between each pair as in (1.1). The test statistic has asymptotically standard normal distribution under the null hypothesis H_{0jk}. Then we use the local FDR procedure to control the overall FDR at the prespecified level α.

The traditional statistic for association is the sample correlation coefficient, r_{jk} between variables X_j and Y_k, which is defined as

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

for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, q$.

Since the variances of sample correlation coefficient are unstable, we use the variance stabilization method, Fisher's z-transformation which is defined as

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

When data $(\mathbf{X}_i^\top, \mathbf{Y}_i^\top)^\top$ follows a multivariate normal distribution, it has been shown that $F(r_{jk})$ is approximately distributed by normal distribution with mean $\ln((1+\rho_{jk})/(1-\rho_{jk}))/2$ and standard deviation $1/\sqrt{n-3}$ with sample size n. Under the null hypothesis, we obtain a standardized normal test statistic

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

We then apply local FDR procedure to these test statistics z_{jk} for $1 \le j \le p$ and $1 \le k \le q$.

Among various FDR procedures, we consider the local FDR proposed by Efron (2004) which is a Bayes version of Benjamini and Hochberg (1995) procedure focusing on density rather than tail areas. The local FDR at the observation z_{jk} is defined as

$$\mathrm{lfdr}\left(z_{jk}\right) = \frac{\pi_0 f_0\left(z_{jk}\right)}{\pi_0 f_0\left(z_{jk}\right) + \pi_1 f_1\left(z_{jk}\right)},$$

where f_0 and f_1 are null and alternative densities with corresponding prior probabilities π_0 and $\pi_1 (= 1-\pi_0)$. Then, the null hypothesis H_{0jk} is rejected whenever $\mathrm{lfdr}(z_{jk}) \le \alpha$ so that we control FDR at the prespecified level α. In view of thresholding estimator (1.2), the thresholding level t is estimated as

$$\hat{t}^{\mathrm{lfdr}} = \inf_{r_{jk}}\left\{\left|r_{jk}\right| \,\middle|\, \mathrm{lfdr}\left(z_{jk}\right) \le \alpha,\ z_{jk} = \sqrt{n-3}F\left(r_{jk}\right)\right\},$$

which defines a sparse structure and so gives a cross correlation matrix estimator $\hat{\mathbf{R}}_{XY}^{\mathrm{lfdr}}$ with elements

$$\hat{\rho}_{jk}^{\mathrm{lfdr}} = r_{jk}\cdot \mathrm{I}\left(\left|r_{jk}\right| \ge \hat{t}^{\mathrm{lfdr}}\right), \quad j = 1,2,\ldots,p,\ k = 1,2,\ldots,q. \tag{2.3}$$

In this paper, we estimate local FDR by the R package *locfdr* by Efron (2004).

3. Data example

In this section, we apply 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 aim 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.

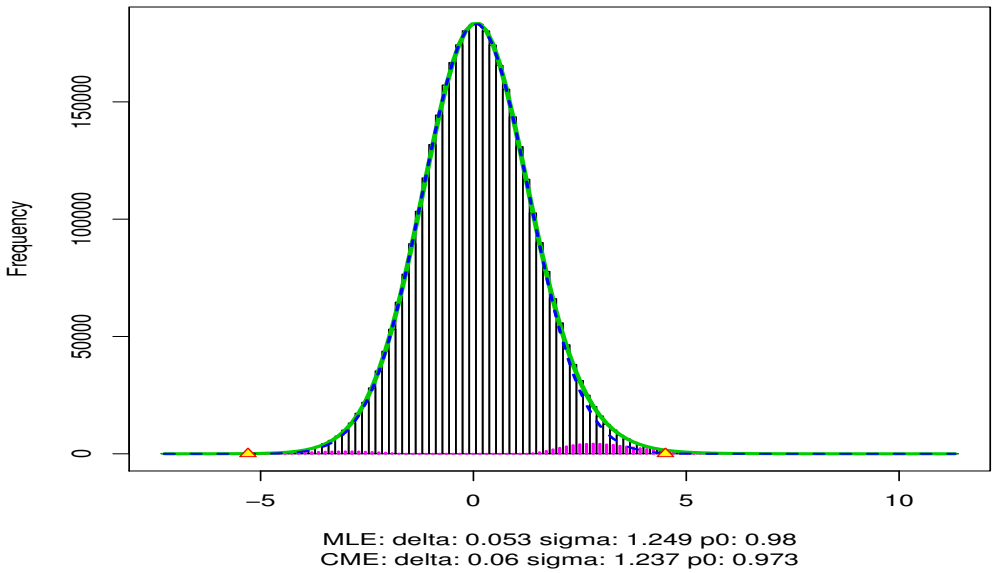

Figure 1: *Histogram of z-values for the protein and mRNA expression pairs data, N = 3,721,230. Green curve is estimate of mixture density $f(z)$. Blue curve is estimate of null density $\pi_0 f_0(z)$ with delta, sigma, and p0 which are the estimates of mean, standard deviation, and mixing probability of null distribution $f_0(z)$ for each method, MLE and CME.*

We capitalize on the fact that the TFAs are assembled into protein complexes while in action, and thus hypothesize 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 (1,195 proteins), which have been known to regulate as many as 3,114 target genes according to the TF and regulatory element databases such as ENCODE, and TTRUST (Han *et al.*, 2015), ITFP (Zheng *et al.*, 2008), TRED (Zhao *et al.*, 2005).

Figure 1 shows the histogram of the $1{,}195 \times 3{,}114 = 3{,}721{,}230$ z-values. The green solid line is the fitted mixture density $f(z)$. The 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 line is the null density $\pi_0 f_0(z)$ estimated by maximum likelihood (MLE). Both MLE and central matching method (CME) estimate the empirical null distribution similarly. The newly proposed procedure of estimating the cross correlation matrix uses local FDR (lfdr) cutoff value 0.1. Using the empirical null distribution estimation by MLE, more than 99.9% of the entries are penalized to zero by (1.2), resulting in a sparse cross correlation matrix estimation. A total of 3,860 (TFA, mRNA) pairs (0.1%) have non-zero correlation, with all the pairs having strong correlation values larger than |0.5|.

Further, we estimate the sparse cross correlation matrix using the adaptive thresholding procedure proposed by Cai and Liu (2016). We also control the FDR at the level $\alpha = 0.1$ by (2.2) and a total of 57,941 pairs (1.56%) are found with non-zero correlation. Among the non-zero values, only 10.6% (6,138 pairs) have strong correlation values larger than |0.5|, indicating the problem of under-penalization of cross correlation matrix One reason for this would be the FDR in Cai and Liu (2016) is that by (Benjamini and Hochberg, 1995), not local FDR by Efron (2004), and the latter is more

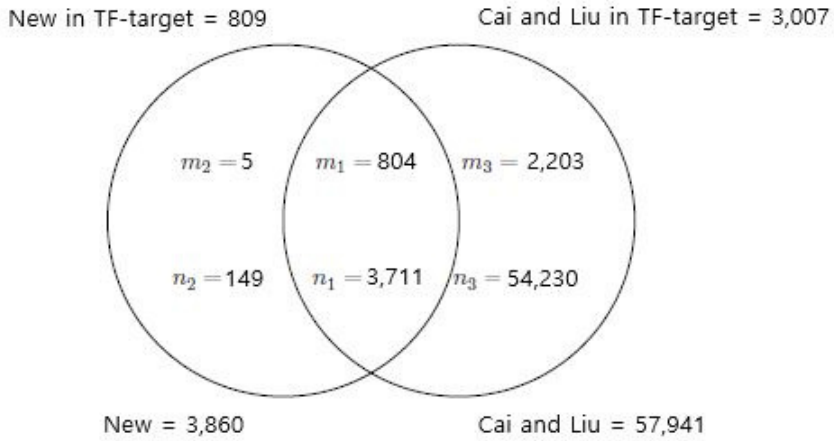

Figure 2: *Venn diagram of non-zero correlation pairs detected by new procedure (New) and adaptive thresholding procedure (Cai and Liu). n_1 denotes the number of non-zero correlation pairs in both procedure, n_2 and n_3 denote the number of unique non-zero correlation pairs in each procedure; m_1, m_2 and m_3 are the number of benchmark pairs in the corresponding n_i pairs for $i = 1, 2, 3$.*

Table 1: Estimation of cross correlation matrix at the different FDR level α by new procedure (New) and adaptive thresholding procedure (Cai and Liu)

Method	$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$	
	New	Cai and Liu	New	Cai and Liu	New	Cai and Liu
Ture discovery rate	20.96%	5.17%	25.35%	8.46%	28.93%	17.42%
False non-discovery rate	3.22%	3.21%	3.23%	3.20%	3.24%	3.22%

True discovery rate is the proportion of benchmarked pairs among the non-zero correlation pairs detected by each method.
False non-discovery rate is the proportion of zero correlation pairs thresholded by each method among benchmark pairs.

conservative than the former which is shown in Appendix.

Further, to evaluate the accuracy of procedures, we benchmark (TFA, mRNA) pairs with non-zero correlation against the known transcription regulatory networks 101,272 pairs and compare the coverage rates. In Figure 2, the Venn diagram shows the number of non-zero correlation (TFA, mRNA) pairs with and without benchmark for both procedures. The TF-target pairs are benchmark pairs used. For the new procedure, a total of 3,860 (TFA, mRNA) pairs have non-zero correlation and among these 809 pairs are known as TF-target pairs. On the other hand, the adaptive thresholding procedure by Cai and Liu (2016) produces a substantial amount of non-zero correlation pairs (57,941) compared to the pairs found by new procedure (3,860). However, the proportion of benchmarked pairs among all non-zero correlation pairs ($m_1 + m_2/n_1 + n_2 = 20.96\%$) by new procedure is substantially higher than that ($m_1 + m_3/n_1 + n_3 = 5.17\%$) by adaptive thresholding procedure. Moreover, the proportion of unique non-zero correlation pairs under benchmark among all unique non-zero correlation pairs ($m_2/n_2 = 3.4\%, m_3/n_3 = 4.1\%$) are nearly the same and both methods show almost the same false negative rate of 3.2%. These findings suggest that the new procedure is more efficient in finding the significantly correlated pairs or estimating the sparse structure of cross correlation matrix than the adaptive thresholding procedure. Table 1 supports this result by showing the similar pattern at the different FDR level α.

4. Conclusion

In this paper, we propose a new method to estimate the sparsity structure of cross correlation matrix $\mathbf{R}_{XY}$ 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 thresholds the traditional sample correlations by controlling local FDR and so FDR at a predetermined level α. In doing so, we adapt the Efron's local false discovery rate procedure (Efron, 2004) to test the individual hypotheses on ρ_{jk}s simultaneously. Lastly, applying the method to breast cancer data in TCGA, we show the procedure more efficiently estimate and define the sparse structure of high-dimensional cross correlation matrix than adaptive thresholding procedure by Cai and Liu (2016). As a future work, with the recent advances in multiple testing literature, we may be able to refine the procedure.

Appendix:

We claim that local FDR by (Efron, 2004) (lfdr(y)) is more conservative than the FDR in Cai and Liu (2016) by showing 'lfdr $< \alpha$' approximately gurantees 'FDR $< \alpha$'. Here, the FDR in Cai and Liu (2016) is the expected number of false discovery proportion in Benjamini and Hochberg (1995).

Let us start with a few notations to be used in this section. Let N be the number of hypotheses, $\mathcal{R}$ be the rejection region, FDR($\mathcal{R}$) be the FDR by Benjamini and Hochberg (1995). In addition, let Fdr($\mathcal{R}$) be the Bayesian FDR defined as

$$\mathrm{Fdr}\left(\mathcal{R}\right) = \frac{\pi_0 F_0\left(\mathcal{R}\right)}{F\left(\mathcal{R}\right)},$$

where π_0 is the proportion of the null hypotheses among N total hypotheses, $F_0(\mathcal{R})$ and $F_1(\mathcal{R})$ are the probabilities of the rejection $\mathcal{R}$ under the null and alternative hypothesis, respectively, and $F(\mathcal{R}) = \pi_0 F_0(\mathcal{R}) + (1-\pi_0)F_1(\mathcal{R})$. The local FDR is defined as

$$\mathrm{lfdr}\left(z\right) = \frac{\pi_0 f_0\left(z\right)}{f\left(z\right)} = \frac{\pi_0 f_0\left(z\right)}{\pi_0 f_0\left(z\right) + \left(1-\pi_0\right) f_1\left(z\right)}, \tag{A.1}$$

where $f_0(z)$ and $f_1(z)$ are probability distribution functions of z under the null and alternative hypothesis as we define earlier.

Before proving the claim, we recall two results from Efron and Tibshirani (2002), the averaging theorem and the conservative bias theorem. The averaging theorem tells that

$$\mathbb{E}_f\left(\mathrm{lfdr}(z)\middle| z \in \mathcal{R}\right) = \mathrm{Fdr}\left(\mathcal{R}\right) \tag{A.2}$$

and the conservative bias theorem is that

$$\mathrm{FDR}\left(\mathcal{R}\right) \leq \mathbb{E}\left(\widehat{\mathrm{Fdr}}\left(\mathcal{R}\right)\right), \tag{A.3}$$

where $\widehat{\mathrm{Fdr}}(\mathcal{R})$ is the empirical estimate of the Baysian FDR that is

$$\widehat{\mathrm{Fdr}}\left(\mathcal{R}\right) = \frac{\pi_0 F_0\left(\mathcal{R}\right)}{\hat{F}\left(\mathcal{R}\right)} \text{ with } \widehat{F}\left(\mathcal{R}\right) = \frac{\#\left(z_i \in \mathcal{R}\right)}{N}.$$

Now, let us show our main claim. We set the rejection region $\mathcal{R}$ as

$$\mathcal{R} = \left\{z \,\middle|\, \mathrm{lfdr}(z) \leq \alpha\right\}.$$

Suppose we for a moment assume $\mathbb{E}(\widehat{\text{Fdr}}(\mathcal{R}))$ is close to $\text{Fdr}(\mathcal{R})$. Then,

$$\text{FDR}(\mathcal{R}) \leq \mathbb{E}\left(\widehat{\text{Fdr}}(\mathcal{R})\right) \tag{A.4}$$

$$\approx \text{Fdr}(\mathcal{R}) \leq \alpha, \tag{A.5}$$

where (A.4) is from the conservative bias theorem (A.3) and (A.5) is from the averaging theorem (A.2). Thus, this completes the proof of the claim.

Finally, let us discuss the assumption that '$\mathbb{E}(\widehat{\text{Fdr}}(\mathcal{R}))$ is close to $\text{Fdr}(\mathcal{R})$'. Suppose Z_is are independently from $F(z)$ and $F(\mathcal{R}) > 0$, and thus $\widehat{F}(\mathcal{R}) > 0$ almost surely. Recall that $\widehat{F}(\mathcal{R}) = \#\{Z_i \in \mathcal{R}\}/N$ and $W \equiv \#\{Z_i \in \mathcal{R}\}$ follows the binomial distribution $B(N, F(\mathcal{R}))$. Therefore, the law of large numbers and bounded convergence theorem shows that $\widehat{\text{Fdr}}(\mathcal{R})$ converges to $\text{Fdr}(\mathcal{R})$ almost surely, and thus $\mathbb{E}(\widehat{\text{Fdr}}(\mathcal{R}))$ should be close to $\text{Fdr}(\mathcal{R})$, when N is large enough.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF-2019R1F1A1056779).

References

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

Bennett CM, Wolford GL, and Miller MB (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.

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

Dubois PC, Trynka G, Franke L *et al.* (2010). Multiple common variants for celiac disease influencing immune gene expression, *Nature Genetics*, **42**, 295–302.

Efron B and Tibshirani R (2002). Empirical Bayes methods and false discovery rates for microarrays, *Genetic Epidemiology*, **23**, 70–86.

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

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, *et al.* (2015). TRRUST: A reference database of human transcriptional regulatory interactions, *Scientific Reports*, **5**, 11432.

Huttlin EL, Ting L, Bruckner RJ, *et al.* (2015). The bioplex network: A systematic exploration of the human interactome, *Cell*, **162**, 425–440.

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

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

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

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

Shaw P, Greenstein D, Lerch J, *et al.* (2006). Intellectual ability and cortical development in children and adolescents, *Nature*, **440**, 676–679.

Shedden K and Taylor J (2005). Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas, *Methods of Microarray Data Analysis*, (pp. 121–131), Springer, Boston.

Storey JD (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 TT (2015). Testing differential networks with applications to detecting gene-by-gene interactions, *Biometrika*, **102**, 247–266.

Yu D, Lee SH, Lim J, Xiao G, Craddock RC, and Biswal BB (2018). Fused lasso regression for identifying differential correlations in brain connectome graphs, *Statistical Analysis and Data Mining*, **11**, 203–226.

Zhao F, Xuan Z, Liu L, and Zhang MQ (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.

Received April 01, 2022; Revised July 12, 2022; Accepted July 28, 2022