View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by ScholarlyCommons@Penn

Penn Libraries University of Pennsylvania

University of Pennsylvania
ScholarlyCommons

Statistics Papers

Wharton Faculty Research

9-2016

Minimax Rate-Optimal Estimation of High-Dimensional Covariance Matrices with Incomplete Data

T. Tony Cai
University of Pennsylvania

Anru Zhang

Follow this and additional works at: https://repository.upenn.edu/statistics_papers

Part of the Business Analytics Commons, Management Sciences and Quantitative Methods Commons, Mathematics Commons, and the Statistics and Probability Commons

Recommended Citation

Cai, T., & Zhang, A. (2016). Minimax Rate-Optimal Estimation of High-Dimensional Covariance Matrices with Incomplete Data. *Journal of Multivariate Analysis, 150* 55-74. http://dx.doi.org/10.1016/j.jmva.2016.05.002

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/statistics_papers/603
For more information, please contact repository@pobox.upenn.edu.

Minimax Rate-Optimal Estimation of High-Dimensional Covariance Matrices with Incomplete Data

Abstract

Missing data occur frequently in a wide range of applications. In this paper, we consider estimation of high-dimensional covariance matrices in the presence of missing observations under a general missing completely at random model in the sense that the missingness is not dependent on the values of the data. Based on incomplete data, estimators for bandable and sparse covariance matrices are proposed and their theoretical and numerical properties are investigated.

Minimax rates of convergence are established under the spectral norm loss and the proposed estimators are shown to be rate-optimal under mild regularity conditions. Simulation studies demonstrate that the estimators perform well numerically. The methods are also illustrated through an application to data from four ovarian cancer studies. The key technical tools developed in this paper are of independent interest and potentially useful for a range of related problems in high-dimensional statistical inference with missing data.

Keywords

adaptive thresholding, bandable covariance matrix, generalized sample covariance matrix, missing data, optimal rate of convergence, sparse convergence matrix, thresholding

Disciplines

Business | Business Analytics | Management Sciences and Quantitative Methods | Mathematics | Statistics and Probability

This technical report is available at ScholarlyCommons: https://repository.upenn.edu/statistics_papers/603

Minimax Rate-optimal Estimation of High-dimensional Covariance Matrices with Incomplete Data [*]

T. Tony Cai and Anru Zhang

Abstract

Missing data occur frequently in a wide range of applications. In this paper, we consider estimation of high-dimensional covariance matrices in the presence of missing observations under a general missing completely at random model in the sense that the missingness is not dependent on the values of the data. Based on incomplete data, estimators for bandable and sparse covariance matrices are proposed and their theoretical and numerical properties are investigated.

Minimax rates of convergence are established under the spectral norm loss and the proposed estimators are shown to be rate-optimal under mild regularity conditions. Simulation studies demonstrate that the estimators perform well numerically. The methods are also illustrated through an application to data from four ovarian cancer studies. The key technical tools developed in this paper are of independent interest and potentially useful for a range of related problems in high-dimensional statistical inference with missing data.

[*] T. Tony Cai is Dorothy Silberberg Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA (E-mail: tcai@wharton.upenn.edu); Anru Zhang is Assistant Professor, University of Wisconsin-Madison, Madison, WI (E-mail: anruzhang@stat.wisc.edu). The research of Tony Cai and Anru Zhang was supported in part by NSF Grant DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

Keywords: Adaptive thresholding, bandable covariance matrix, generalized sample covariance matrix, missing data, optimal rate of convergence, sparse covariance matrix, thresholding.

1 Introduction

The problem of missing data arises frequently in a wide range of fields, including biomedical studies, social science, engineering, economics, and computer science. Statistical inference in the presence of missing observations has been well studied in classical statistics. See, e.g., Ibrahim and Molenberghs [18] for a review of missing data methods in longitudinal studies and Schafer [26] for literature on handling multivariate data with missing observations. See Little and Rubin [20] and the references therein for a comprehensive treatment of missing data problems.

Missing data also occurs in contemporary high-dimensional inference problems, whose dimension p can be comparable to or even much larger than the sample size n. For example, in large-scale genome-wide association studies (GWAS), it is common for many subjects to have missing values on some genetic markers due to various reasons, including insufficient resolution, image corruption, and experimental error during the laboratory process. Also, different studies may have different volumes of genomic data available by design. For instance, the four genomic ovarian cancer studies discussed in Section 4 have throughput measurements of mRNA gene expression levels, but only one of these also has microRNA measurements (Cancer Genome Atlas Research Network [11], Bonome et al. [4], Tothill et al. [27] and Dressman et al. [15]). Discarding samples with any missingness is highly inefficient and could induce bias due to non-random missingness. It is of significant interest to integrate multiple high-throughput studies of the same disease, not only to boost statistical power but also to improve the biological interpretability. However, considerable challenges arise when integrating such studies due to missing data.

Although there have been significant recent efforts to develop methodologies and theories for high dimensional data analysis, there is a paucity of methods with theoretical guarantees for statistical inference with missing data in the high-dimensional setting. Under the assumption that the components are missing uniformly and completely at random (*MUCR*), Loh and Wainwright [21] proposed a non-convex optimization approach to high-dimensional linear regression, Lounici [23] introduced a method for estimating a low-rank covariance matrix and Lounici [22] considered sparse principal component analysis. In these papers, theoretical properties of the procedures were analyzed. These methods and theoretical results critically depend on the MUCR assumption.

Covariance structures play a fundamental role in high-dimensional statistics. It is of direct interest in a wide range of applications including genomic data analysis, particularly for hypothesis generation. Knowledge of the covariance structure is critical to many statistical methods, including discriminant analysis, principal component analysis, clustering analysis, and regression analysis. In the high-dimensional setting with complete data, inference on the covariance structure has been actively studied in recent years. See Cai, Ren and Zhou [7] for a survey of recent results on minimax and adaptive estimation of high-dimensional covariance and precision matrices under various structural assumptions. Estimation of high-dimensional covariance matrices in the presence of missing data also has wide applications in biomedical studies, particularly in integrative genomic analysis which holds great potential in providing a global view of genome function (see Hawkins et al. [17]).

In this paper, we consider estimation of high-dimensional covariance matrices in the presence of missing observations under a general missing completely at random (*MCR*) model in the sense that the missingness is not dependent on the values of the data. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be n independent copies of a p dimensional random vector $\mathbf{X}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. Instead of observing the complete sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, one observes the sample with missing values, where the observed coordinates of $\mathbf{X}_k$ are indicated by a vector $\mathbf{S}_k \in \{0,1\}^p$, $k = 1, ..., n$. That is,

$$X_{jk} \text{ is observed if } S_{jk} = 1 \text{ and } X_{jk} \text{ is missing if } S_{jk} = 0. \tag{1}$$

Here X_{jk} and S_{jk} are respectively the jth coordinate of the vectors $\mathbf{X}_k$ and $\mathbf{S}_k$. We denote the incomplete sample with missing values by $\mathbf{X}^* = \{\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*\}$. The major goal of the present paper is to estimate $\boldsymbol{\Sigma}$, the covariance matrix of $\mathbf{X}$, with theoretical guarantees based on the incomplete data $\mathbf{X}^*$ in the high-dimensional setting where p can be much larger than n.

This paper focuses on estimation of high-dimensional bandable covariance matrices and sparse covariance matrices in the presence of missing data. These two classes of covariance matrices arise frequently in many applications, including genomics, econometrics, signal processing, temporal and spatial data analyses, and chemometrics. Estimation of these high-dimensional structured covariance matrices have been well studied in the setting of complete data in a number of recent papers, e.g., Bickel and Levina [2, 3], Karoui [16], Rothman et al. [24], Cai and Zhou [10], Cai and Liu [5], Cai et al. [6, 9] and Cai and Yuan [8]. Given an incomplete sample $\mathbf{X}^*$ with missing values, we introduced a "generalized" sample covariance matrix, which can be viewed as an analog of the usual sample covariance matrix in the case of complete data. For estimation of bandable covariance matrices, where the entries of the matrix decay as they move away from the diagonal, a blockwise tridiagonal estimator is introduced and is shown to be rate-optimal. We then consider estimation of sparse covariance matrices. An adaptive thresholding estimator based on the generalized sample covariance matrix is proposed. The estimator is shown to achieve the optimal rate of convergence over a large class of approximately sparse covariance matrices under mild conditions.

The technical analysis for the case of missing data is much more challenging than that for the complete data, although some of the basic ideas are similar. To facilitate the theoretical analysis of the proposed estimators, we establish two key technical results, first, a large deviation result for a sub-matrix of the generalized sample covariance matrix and second, a large deviation bound for the self-normalized entries of the generalized sample covariance matrix. These technical tools are not only important for the present paper, but also useful for other related problems in high-dimensional statistical inference with missing data.

A simulation study is carried out to examine the numerical performance of the proposed estimation procedures. The results show that the proposed estimators perform well numerically. Even in the MUCR setting, our proposed procedures for estimating bandable, sparse covariance matrices, which do not rely on the information of the missingness mechanism, outperform the ones specifically designed for MUCR. The advantages are more significant under the setting of missing completely at random but not uniformly. We also illustrate our procedure with an application to data from four ovarian cancer studies that have different volumes of genomic data by design. The proposed estimators enable us to estimate the covariance matrix by integrating the data from all four studies and lead to a more accurate estimator. Such high-dimensional covariance matrix estimation with missing data is also useful for other types of data integration. See further discussions in Section 4.4.

The rest of the paper is organized as follows. Section 2 considers estimation of bandable covariance matrices with incomplete data. The minimax rate of convergence is established for the spectral norm loss under regularity conditions. Section 3 focuses on estimation of high-dimensional sparse covariance matrices and introduces an adaptive thresholding estimator in the presence of missing observations. Asymptotic properties of the estimator under the spectral norm loss is also studied. Numerical performance of the proposed methods is investigated in Section 4 through both simulation studies and an analysis of an ovarian cancer dataset. Section 5 discusses a few related problems. Finally the proofs of the main results are given in Section 6 and the Supplement.

2 Estimation of Bandable Covariance Matrices

In this section, we consider estimation of bandable covariance matrices with incomplete data. Bandable covariance matrices, whose entries decay as they move away from the diagonal, arise frequently in temporal and spatial data analysis. See, e.g., Bickel and Levina [2] and Cai et al. [7] and the references therein. The procedure relies on a "generalized" sample covariance matrix. We begin with basic notation and definitions that will be used throughout the rest of the paper.

2.1 Notation and Definitions

Matrices and vectors are denoted by boldface letters. For a vector $\boldsymbol{\beta} \in \mathbb{R}^p$, we denote the Euclidean q-norm by $\|\boldsymbol{\beta}\|_q$, i.e., $\|\boldsymbol{\beta}\|_q = \sqrt[q]{\sum_{i=1}^p |\beta_i|^q}$. Let $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = \sum_i \lambda_i(\mathbf{A})\mathbf{u}_i\mathbf{v}_i^\top$ be the singular value decomposition of a matrix $\mathbf{A} \in \mathbb{R}^{p_1 \times p_2}$, where $\mathbf{D} = \text{diag}\{\lambda_1(\mathbf{A}), \ldots\}$ with $\lambda_1(\mathbf{A}) \geq \cdots \geq 0$ being the singular values. For $1 \leq q \leq \infty$, the Schatten-q norm $\|\mathbf{A}\|_q$ is defined by $\|\mathbf{A}\|_q = \{\sum_i \lambda_i^q(\mathbf{A})\}^{1/q}$. In particular, $\|\mathbf{A}\|_2 = \sqrt{\sum_i \lambda_i^2(\mathbf{A})}$ is the Frobenius norm of $\mathbf{A}$ and will be denoted as $\|\mathbf{A}\|_F$; $\|\mathbf{A}\|_\infty = \lambda_1(\mathbf{A})$ is the spectral norm of $\mathbf{A}$ and will be simply denoted as $\|\mathbf{A}\|$. For $1 \leq q \leq \infty$ and $A \in \mathbb{R}^{p_1 \times p_2}$, we denote the operator ℓ_q norm of $\mathbf{A}$ by $\|\mathbf{A}\|_{\ell_q}$ which is defined as $\|\mathbf{A}\|_{\ell_q} = \max_{x \in \mathbb{R}^{p_2}} \|\mathbf{A}\mathbf{x}\|_q / \|\mathbf{x}\|_q$. The following are well known facts about the various norms of a matrix $A = (a_{ij})$,

$$\|\mathbf{A}\|_{\ell_1} = \max_j \sum_{i=1}^{p_1} |a_{ij}|, \quad \|\mathbf{A}\|_{\ell_2} = \|\mathbf{A}\| = \lambda_1(\mathbf{A}), \quad \|\mathbf{A}\|_{\ell_\infty} = \max_i \sum_{j=1}^{p_2} |a_{ij}|, \tag{2}$$

and, if $\mathbf{A}$ is symmetric, $\|\mathbf{A}\|_{\ell_1} = \|\mathbf{A}\|_{\ell_\infty} \geq \|\mathbf{A}\|_{\ell_2}$. When R_1, R_2 are two subsets of $\{1, \ldots, p_1\}$, $\{1, \ldots, p_2\}$ respectively, we note $\mathbf{A}_{R_1 \times R_2} = (a_{ij})_{i \in R_1, j \in R_2}$ as the sub-matrix of $\mathbf{A}$ with indices R_1 and R_2. In addition, we simply write $\mathbf{A}_{R_1 \times R_1}$ as $\mathbf{A}_{R_1}$.

We denote by $\mathbf{X}_1, \ldots, \mathbf{X}_n$ a complete random sample (without missing observations) from a p-dimensional distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The sample mean and sample covariance matrix are defined as

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^n \mathbf{X}_k, \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^n \left(\mathbf{X}_k - \bar{\mathbf{X}}\right)\left(\mathbf{X}_k - \bar{\mathbf{X}}\right)^\top. \tag{3}$$

Now we introduce the notation related to the incomplete data with missing observations. Generally, we use the superscript "$*$" to denote objects related to missing values. Let $\mathbf{S}_1, ..., \mathbf{S}_n$ be the indicator vectors for the observed values (see (1)) and let $\mathbf{X}^* = \{\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*\}$ be the observed incomplete data where the observed entries are indexed by the vectors $\mathbf{S}_1, ..., \mathbf{S}_n \in$

$\{0,1\}^p$. In addition, we define

$$n_{ij}^* = \sum_{k=1}^{n} S_{ik} S_{jk}, \quad 1 \le i, j \le p. \tag{4}$$

Here n_{ij}^* is the number of vectors $\mathbf{X}_k^*$ in which the i^{th} and j^{th} entries are both observed. For convenience, we also denote

$$n_i^* = n_{ii}^*, \quad n_{\min}^* = \min_{i,j} n_{ij}^*. \tag{5}$$

Given a sample $\mathbf{X}^* = \{\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*\}$ with missing values, the sample mean and sample covariance matrix can no longer be calculated in the usual way. Instead, we propose the "generalized sample mean" $\bar{\mathbf{X}}^*$ defined by

$$\bar{\mathbf{X}}^* = (\bar{X}_i^*)_{1 \le i \le p} \quad \text{with} \quad \bar{X}_i^* = \frac{1}{n_i^*} \sum_{k=1}^{n} X_{ik} S_{ik}, \quad 1 \le i \le p, \tag{6}$$

where X_{ik} is the ith entry of $\mathbf{X}_k$, and the "generalized sample covariance matrix" $\hat{\mathbf{\Sigma}}^*$ defined by

$$\hat{\mathbf{\Sigma}}^* = (\hat{\sigma}_{ij}^*)_{1 \le i,j \le p} \quad \text{with} \quad \hat{\sigma}_{ij}^* = \frac{1}{n_{ij}^*} \sum_{k=1}^{n} (X_{ik} - \bar{X}_i^*)(X_{jk} - \bar{X}_j^*) S_{ik} S_{jk}. \tag{7}$$

As will be seen later, the generalized sample mean $\bar{\mathbf{X}}^*$ and the generalized sample covariance matrix $\hat{\mathbf{\Sigma}}^*$ play similar roles as those of the conventional sample mean and sample covariance matrix in inference problems, but the technical analysis can be much more involved. Some distinctions between the generalized sample covariance matrix $\hat{\mathbf{\Sigma}}^*$ and the usual sample covariance matrix $\hat{\mathbf{\Sigma}}$ are that $\hat{\mathbf{\Sigma}}^*$ is in general not non-negative definite, and each entry $\hat{\sigma}_{ij}^*$ is the average of a varying number (n_{ij}^*) of samples, which create additional difficulties in the technical analysis.

Regarding the mechanism of missingness, the assumption we use for the theoretical analysis is *missing completely at random*. This is a more general setting than the one considered previously by Loh and Wainwright [21] and Lounici [22].

Assumption 2.1 (Missing Completely at Random (MCR)) $\mathbf{S} = \{\mathbf{S}_1, \ldots, \mathbf{S}_n\}$ is not dependent on the values of $\mathbf{X}$. Here $\mathbf{S}$ can be either deterministic or random, but independent of $\mathbf{X}$.

We adopt Assumption 1 in Chen et al. [13] and assume that the random vector $\mathbf{X}$ is sub-Gaussian satisfying the following assumption.

Assumption 2.2 (Sub-Gaussian Assumption) $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$. Here the columns $\mathbf{X}_k$ are i.i.d. and can be expressed as

$$\mathbf{X}_k = \mathbf{\Gamma}\mathbf{Z}_k + \boldsymbol{\mu}, \quad k = 1, \ldots, n, \tag{8}$$

where $\boldsymbol{\mu}$ is a fixed p-dimensional mean vector, $\mathbf{\Gamma} \in \mathbb{R}^{p\times q}$ is a fixed matrix with $q \geq p$ so that $\mathbf{\Gamma}\mathbf{\Gamma}^\top = \mathbf{\Sigma}$, $\mathbf{Z}_k = (Z_{1k}, \ldots, Z_{mk})^\top$ is an m-dimensional random vector with the components mean 0, variance 1, and i.i.d. sub-Gaussian, with the exception of i.i.d. Rademacher. More specifically, each Z_{ik} satisfies that $\mathrm{E}Z_{ik} = 0, \mathrm{var}(Z_{ik}) = 1$, $0 < \mathrm{var}(Z_{ik}^2) < \infty$, and there exists $\tau > 0$ such that $\mathrm{E}e^{tZ_{ik}} \leq \exp(\tau t^2/2)$ for all $t > 0$.

Note that the exclusion of the Rademacher distribution in Assumption 2.2 is only required for estimation of sparse covariance matrices. See Remark 3.3 for further discussions.

2.2 Rate-optimal Blockwise Tridiagonal Estimator

We follow Bickel [2] and Cai et al. [9] and consider estimating the covariance matrix $\mathbf{\Sigma}$ over the parameter space $\mathcal{U}_\alpha = \mathcal{U}_\alpha(M_0, M)$ where

$$\mathcal{U}_\alpha(M_0, M) = \left\{ \mathbf{\Sigma} : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \leq Mk^{-\alpha} \text{ for all } k, \|\mathbf{\Sigma}\| \leq M_0 \right\}. \tag{9}$$

Suppose we have n i.i.d. samples with missing values $\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*$ with covariance matrix $\mathbf{\Sigma} \in \mathcal{U}_\alpha(M_0, M)$. We propose a blockwise tridiagonal estimator $\hat{\mathbf{\Sigma}}^{\mathrm{bt}}$ to estimate $\mathbf{\Sigma}$. We begin by dividing the generalized sample covariance matrix $\hat{\mathbf{\Sigma}}^*$ given by (7) into blocks of size $k \times k$ for some k. More specifically, pick an integer k and let $N = \lceil p/k \rceil$. Set $I_j = \{(j-1)k+1, \ldots, jk\}$ for $1 \leq j \leq N-1$, and $I_N = \{(N-1)k+1, \ldots, p\}$. For $1 \leq j, j' \leq N$ and $\mathbf{A} = (a_{i_1,i_2})_{p\times p}$, define

$$\mathbf{A}_{I_j \times I_{j'}} = (a_{i_1,i_2})_{i_1 \in I_j, i_2 \in I_{j'}}$$

and define the blockwise tridiagonal estimator $\hat{\mathbf{\Sigma}}^{\mathrm{bt}}$ by

$$\hat{\mathbf{\Sigma}}_{I_j \times I_{j'}} = \begin{cases} \hat{\mathbf{\Sigma}}^*_{I_j \times I_{j'}}, & \text{if } |j - j'| \leq 1; \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

That is, $\hat{\mathbf{\Sigma}}_{I_j \times I_{j'}}$ is estimated by its sample counterpart if and only if j and j' differ by at most 1. The weight matrix of the blockwise tridiagonal estimator $\hat{\mathbf{\Sigma}}^{\mathrm{bt}}$ is illustrated in Figure 1.

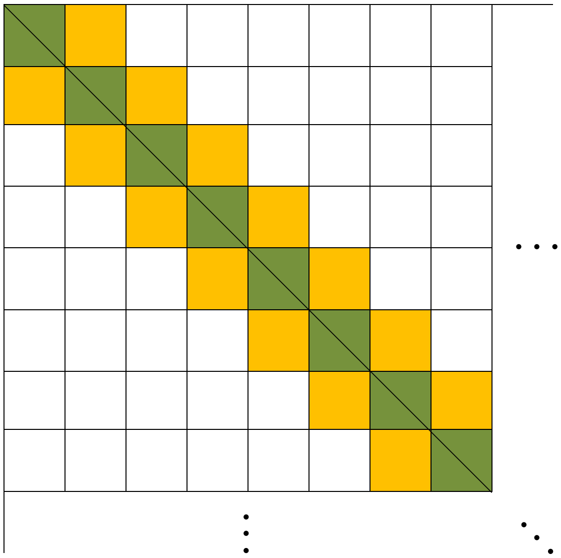


Figure 1: Weight matrix for the blockwise tridiagonal estimator.

Theorem 2.1 *Suppose Assumptions 2.1 and 2.2 hold. Then, conditioning on* $\mathbf{S}$*, the blockwise tridiagonal* $\hat{\mathbf{\Sigma}}^{\mathrm{bt}}$ *with* $k = (n^*_{\min})^{1/(2\alpha+1)}$ *satisfies*

$$\sup_{\mathbf{\Sigma} \in \mathcal{U}_\alpha(M, M_0)} \mathrm{E}\|\hat{\mathbf{\Sigma}}^{\mathrm{bt}} - \mathbf{\Sigma}\|^2 \leq C(n^*_{\min})^{-2\alpha/(2\alpha+1)} + C\frac{\ln p}{n^*_{\min}}, \tag{11}$$

where C *is a constant depending only on* M*,* M_0*, and* τ *from Assumption 2.2.*

The optimal choice of block size k depends on the unknown "smoothness parameter" α. In practice, k can be chosen by cross-validation. See Section 4.1 for further discussions. Moreover, the convergence rate in (11) is optimal as we also have the following lower bound result.

Proposition 2.1 *For any $n_0 \geq 1$ such that $p \leq \exp(\gamma n_0)$ for some constant $\gamma > 0$, conditioning on* $\mathbf{S}$ *we have*

$$\inf_{\hat{\mathbf{\Sigma}}} \sup_{\substack{\mathbf{\Sigma}\in\mathcal{U}_\alpha(M,M_0)\\ \mathbf{S}:n^*_{\min}\geq n_0}} \mathrm{E}\left(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2\right) \geq C(n_0)^{-2\alpha/(2\alpha+1)} + C\frac{\ln p}{n_0}.$$

Remark 2.1 (Tapering and banding estimators) It should be noted that the same rate of convergence can also be attained by tapering and banding estimators with suitable choices of tapering and banding parameters. Specifically, let $\hat{\mathbf{\Sigma}}^{\mathrm{tp}}$ and $\hat{\mathbf{\Sigma}}^{\mathrm{bd}}$ be respectively the tapering and banded estimators proposed in Cai et al. [9] and Bickel and Levina [2] with

$$\hat{\mathbf{\Sigma}}^{\mathrm{tp}} = \hat{\mathbf{\Sigma}}^{\mathrm{tp}}_k = (w^{\mathrm{tp}}_{ij}\hat{\sigma}^*_{ij})_{1\leq i,j\leq p} \quad \text{and} \quad \hat{\mathbf{\Sigma}}^{\mathrm{bd}} = \hat{\mathbf{\Sigma}}^{\mathrm{bd}}_k = (w^{\mathrm{bd}}_{ij}\hat{\sigma}^*_{ij})_{1\leq i,j\leq p}, \tag{12}$$

where w^{tp}_{ij} and w^{bd}_{ij} are the weights defined as

$$w^{\mathrm{tp}}_{ij} = \begin{cases} 1, & \text{when } |i-j| \leq k/2, \\ 2 - \frac{|i-j|}{k_h}, & \text{when } k/2 < |i-j| < k \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad w^{\mathrm{bd}}_{ij} = \begin{cases} 1, & \text{when } |i-j| \leq k, \\ 0, & \text{otherwise} \end{cases}. \tag{13}$$

Then the estimators $\hat{\mathbf{\Sigma}}^{\mathrm{tp}}$ and $\hat{\mathbf{\Sigma}}^{\mathrm{bd}}$ with $k = (n^*_{\min})^{1/(2\alpha+1)}$ attains the rate given in (11).

The proof of Theorem 2.1 shares some basic ideas with that for the complete data case (See, e.g. Theorem 2 in Cai et al. [9]). However, it relies on a new key technical tool which is a large deviation result for a sub-matrix of the generalized sample covariance matrix under the spectral norm. This random matrix result for the case of missing data, stated in the following lemma, can be potentially useful for other, related high-dimensional missing data problems. The proof of Lemma 2.1, given in Section 6, is more involved than the complete data case, as in the generalized sample covariance matrix each entry, $\hat{\sigma}^*_{ij}$, is the average of a varying number of samples.

Lemma 2.1 *Suppose Assumptions 2.1 and 2.2 hold. Let* $\hat{\mathbf{\Sigma}}^*$ *be the generalized sample covariance matrix defined in* (7) *and let A and B be two subsets of* $\{1, \ldots, p\}$. *Then, conditioning* *on* $\mathbf{S}$*, the submatrix* $\hat{\boldsymbol{\Sigma}}^*_{A\times B}$ *satisfies*

$$\begin{aligned}&\Pr\left(\|\hat{\boldsymbol{\Sigma}}^*_{A\times B}-\boldsymbol{\Sigma}_{A\times B}\|\le x\right)\\ \ge&1-C\cdot(49)^{|A\cup B|}\exp\left\{-cn^*_{\min}\min\left(\frac{x^2}{\tau^4\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|},\frac{x}{\tau^2\left(\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|\right)^{1/2}}\right)\right\}\end{aligned} \tag{14}$$

for all $x>0$*. Here* $C>0$ *and* $c>0$ *are two absolute constants.*

3 Estimation of Sparse Covariance Matrices

In this section, we consider estimation of high-dimensional sparse covariance matrices in the presence of missing data. We introduce an adaptive thresholding estimator based on incomplete data and investigate its asymptotic properties.

3.1 Adaptive Thresholding Procedure

Sparse covariance matrices arise naturally in a range of applications including genomics. Estimation of sparse covariance matrices has been considered in several recent papers in the setting of complete data (see, e.g., Bickel and Levina [3], El Karoui [16], Rothman et al. [24], Cai and Zhou [10] and Cai and Liu [5]). Estimation of a sparse covariance matrix is intrinsically a heteroscedastic problem in the sense that the variances of the entries of the sample covariance matrix can vary over a wide range. To treat the heteroscedasticity of the sample covariances, Cai and Liu [5] introduced an adaptive thresholding procedure which adapts to the variability of the individual entries of the sample covariance matrix and outperforms the universal thresholding method. The estimator is shown to be simultaneously rate optimal over collections of sparse covariance matrices.

In the present setting of missing data, the usual sample covariance matrix is not available. Instead we apply the idea of adaptive thresholding to the generalized sample covariance matrix $\hat{\boldsymbol{\Sigma}}^*$. The procedure can be described as follows. Note that $\hat{\boldsymbol{\Sigma}}^*$ defined in (7) is a nearly unbiased estimate of $\mathbf{\Sigma}$, we may write it element-wise as

$$\hat{\sigma}_{ij}^{*} \approx \sigma_{ij} + \sqrt{\frac{\theta_{ij}}{n_{ij}^{*}}} z_{ij}, \quad 1 \le i, j \le p,$$

where z_i is approximately normal with mean 0 and variance 1, and θ_{ij} describes the uncertainty of estimator σ_{ij}^{*} to σ_{ij} such that

$$\theta_{ij} = \operatorname{var}\left\{(X_i - \mu_i)(X_j - \mu_j) - \sigma_{ij}\right\}.$$

We can estimate θ_{ij} by

$$\hat{\theta}_{ij}^{*} = \frac{1}{n_{ij}^{*}} \sum_{k=1}^{n} \left\{(X_{ik} - \bar{X}_i^{*})(X_{jk} - \bar{X}_j^{*}) - \hat{\sigma}_{ij}^{*}\right\}^2 S_{ik} S_{jk}. \tag{15}$$

Lemma 3.1 given at the end of this section shows that $\hat{\theta}_{ij}^{*}$ is a good estimate of θ_{ij}.

Since the covariance matrix $\mathbf{\Sigma}$ is assumed to be sparse, it is natural to estimate $\mathbf{\Sigma}$ by individually thresholding $\hat{\theta}_{ij}^{*}$ according to its own variability as measured by $\hat{\theta}_{ij}^{*}$. Define the thresholding level λ_{ij} by

$$\lambda_{ij} = \delta \sqrt{\frac{\hat{\theta}_{ij}^{*} \ln p}{n_{ij}^{*}}}, \quad 1 \le i, j \le p,$$

where δ is a thresholding constant which can be taken as 2.

Let T_λ be a thresholding function satisfying the following conditions,

(1). $|T_\lambda(z)| \le c_T |y|$ for all z, y such that $|z - y| \le \lambda$;

(2). $T_\lambda(z) = 0$ for $|z| \le \lambda$;

(3). $|T_\lambda(z) - z| \le \lambda$, for all $z \in \mathbb{R}$.

These conditions are met by many well-used thresholding functions, including the soft thresholding rule $T_\lambda(z) = \operatorname{sgn}(z)(z - \lambda)_+$, where $\operatorname{sgn}(z)$ is the sign function such that $\operatorname{sgn}(z) = 1$ if $z > 0$, $\operatorname{sgn}(z) = 0$ if $z = 0$, and $\operatorname{sgn}(z) = -1$ if $z < 0$, and the adaptive lasso rule $T_\lambda(z) = z(1 - |\lambda/z|^\eta)_+$ with $\eta \ge 1$ (see Rothman et al. [24]). The hard thresholding function does not satisfy Condition (1), but our analysis also applies to hard thresholding under similar conditions.

The covariance matrix $\mathbf{\Sigma}$ is estimated by $\hat{\mathbf{\Sigma}}^{\mathrm{at}} = (\hat{\sigma}_{ij}^{\mathrm{at}})_{1\le i,j\le p}$ where $\hat{\sigma}_{ij}^{\mathrm{at}}$ is the thresholding estimator defined by

$$\hat{\sigma}_{ij}^{\mathrm{at}} = T_{\lambda_{ij}}(\hat{\sigma}_{ij}^{*}). \tag{16}$$

Note that here each entry $\hat{\sigma}_{ij}^{*}$ is thresholded according to its own variability.

3.2 Asymptotic Properties

We now investigate the properties of the thresholding estimator $\hat{\mathbf{\Sigma}}^{\mathrm{at}}$ over the following parameter space for sparse covariance matrices,

$$\mathcal{H}(c_{n,p}) = \left\{ \mathbf{\Sigma} = (\sigma_{ij}) : \max_{1\le i\le p} \sum_{j=1}^{p} \min\left\{ (\sigma_{ii}\sigma_{jj})^{1/2}, \frac{|\sigma_{ij}|}{\sqrt{(\ln p)/n}} \right\} \le c_{n,p} \right\}. \tag{17}$$

The parameter space $\mathcal{H}(c_{n,p})$ contains a large collection of sparse covariance matrices and does not impose any constraint on the variances σ_{ii}, $i = 1, ..., p$. The collection $\mathcal{H}(c_{n,p})$ contains some other commonly used classes of sparse covariance matrices in the literature, including an ℓ_q ball assumption $\max_i \sum_{j=1}^{p} |\sigma_{ij}|^q \le s_{n,p}$ in Bickel and Levina [3], and a weak ℓ_q ball assumption $\max_{1\le j\le p} \left\{ \left|\sigma_{j[k]}\right|^q \right\} \le s_{n,p}/k$ for each integer k in Cai and Zhou [10] where $\left|\sigma_{j[k]}\right|$ is the kth largest entry in magnitude of the jth row $(\sigma_{ij})_{1\le i\le p}$. See Cai et al. [7] for more discussions.

We have the following result on the performance of $\hat{\mathbf{\Sigma}}^{\mathrm{at}}$ over the parameter space $\mathcal{H}(c_{n,p})$.

Theorem 3.1 *Suppose that $\delta \ge 2$, $\ln p = o((n_{\min}^{*})^{1/3})$ and Assumptions 2.1 and 2.2 hold. Then, conditioning on $\mathbf{S}$, there exists a constant C not depending on p, $n_{\min}^{*}$ or n such that for any $\mathbf{\Sigma} \in \mathcal{H}(c_{n,p})$,*

$$\Pr\left(\left\| \hat{\mathbf{\Sigma}}^{\mathrm{at}} - \mathbf{\Sigma} \right\| \le C c_{n,p} \sqrt{\frac{\ln p}{n_{\min}^{*}}} \right) \ge 1 - O\left\{ (\ln p)^{-1/2} p^{-\delta+2} \right\}. \tag{18}$$

Moreover, if we further assume that $p \ge (n_{\min}^{*})^{\xi}$ and $\delta \ge 4 + 1/\xi$, we in addition have

$$\mathrm{E}\left(\|\hat{\mathbf{\Sigma}}^{\mathrm{at}} - \mathbf{\Sigma}\|^2 \right) \le C c_{n,p}^2 \frac{\ln p}{n_{\min}^{*}}. \tag{19}$$

Moreover, the lower bound result below shows that the rate in (19) is optimal.

Proposition 3.1 *For any $n_0 \geq 1$ and $c_{n,p} > 0$ such that $c_{n,p} \leq M n_0^{1/2} (\ln p)^{-3/2}$ for some constant $M > 0$, conditioning on $\mathbf{S}$ we have*

$$\inf_{\hat{\boldsymbol{\Sigma}}} \sup_{\substack{\boldsymbol{\Sigma} \in \mathcal{H}(c_{n,p}) \\ \mathbf{S}: n^*_{\min} \geq n_0}} \mathrm{E}\left(\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|^2 \right) \geq C c_{n,p}^2 \frac{\ln p}{n_0}.$$

Remark 3.1 (ℓ_q norm loss) We focus in this paper on estimation under the spectral norm loss. The results given in Theorem 3.1 can be easily generalized to the general matrix ℓ_q norm for $1 \leq q \leq \infty$. The results given in Equations (18) and (19) remain valid when the spectral norm is replaced by the matrix ℓ_q norm for $1 \leq q \leq \infty$.

Remark 3.2 (Positive definiteness) Under mild conditions on $\boldsymbol{\Sigma}$, the estimator $\hat{\boldsymbol{\Sigma}}^{\mathrm{at}}$ is positive definite with high probability. However, $\hat{\boldsymbol{\Sigma}}^{\mathrm{at}}$ is not guaranteed to be positive definite for a given data set. Whenever $\hat{\boldsymbol{\Sigma}}^{\mathrm{at}}$ is not positive semi-definite, a simple extra step can make the final estimator $\hat{\boldsymbol{\Sigma}}^{\mathrm{at}}_{+}$ positive definite and also rate-optimal.

Write the eigen-decomposition of $\hat{\boldsymbol{\Sigma}}^{\mathrm{at}}$ as $\hat{\boldsymbol{\Sigma}}^{\mathrm{at}} = \sum_{i=1}^{p} \hat{\lambda}_i \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^{\top}$, where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ are the eigenvalues and $\hat{\mathbf{v}}_i$ are the corresponding eigenvectors. Define the final estimator

$$\hat{\boldsymbol{\Sigma}}^{\mathrm{at}}_{+} = \hat{\boldsymbol{\Sigma}}^{\mathrm{at}} + \left(|\hat{\lambda}_p| + \frac{\ln p}{n^*_{\min}} \right) I\{\hat{\lambda}_p < 0\} \cdot \mathbf{I}_{p \times p},$$

where $\mathbf{I}_{p \times p}$ is the $p \times p$ identity matrix. Then $\hat{\boldsymbol{\Sigma}}^{\mathrm{at}}_{+}$ is a positive definite matrix with the same structure as that of $\hat{\boldsymbol{\Sigma}}^{\mathrm{at}}$. It is easy to show that $\hat{\boldsymbol{\Sigma}}^{\mathrm{at}}_{+}$ and $\hat{\boldsymbol{\Sigma}}^{\mathrm{at}}$ attains the same rate of convergence over $\mathcal{H}(c_{n,p})$. See Cai, Ren and Zhou [7] for further discussions.

Remark 3.3 (Exclusion of the Rademacher Distribution) To guarantee that $\hat{\theta}^*_{ij}$ is a good estimate of θ_{ij}, one important condition needed in the theoretical analysis is that $\theta_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$ is bounded from below by a positive constant. However when the components of $\mathbf{Z}_k$ in (8) are i.i.d. Rademacher, it is possible that $\theta_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}} = 0$. For example, If Z_1 and Z_2 are i.i.d. Rademacher and $X_i = Z_1 + Z_2$ and $X_j = Z_1 - Z_2$, then $\mathrm{var}(X_i X_j) = \mathrm{var}(Z_1^2 - Z_2^2) = 0$, and this implies $\theta_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}} = 0$.

A key technical tool in the analysis of the adaptive thresholding estimator is a large deviation result for the self-normalized entries of the generalized sample covariance matrix. The following lemma, proved in Section 6, plays a critical role in the proof of Theorem 3.1 and can be useful for other high-dimensional inference problems with missing data.

Lemma 3.1 *Suppose* $\ln p = o((n^*_{\min})^{1/3})$ *and Assumptions 2.1 and 2.2 hold. For any constants* $\delta \geq 2$, $\varepsilon > 0$, $M > 0$, *conditioning on* $\mathbf{S}$, *we have*

$$\Pr\left(\frac{|\hat{\sigma}^*_{ij} - \sigma_{ij}|}{(\hat{\theta}^*_{ij})^{1/2}} \geq \delta\sqrt{\frac{\ln p}{n^*_{ij}}}, \forall 1 \leq i, j \leq p\right) = O\left\{(\ln p)^{-1/2} p^{-\delta+2}\right\}, \tag{20}$$

$$\Pr\left(\max_{ij} \frac{|\hat{\theta}^*_{ij} - \theta_{ij}|}{\sigma_{ii}\sigma_{jj}} \geq \varepsilon\right) = O(p^{-M}). \tag{21}$$

In addition to optimal estimation of a sparse covariance matrix $\mathbf{\Sigma}$ under the spectral norm loss, it is also of significant interest to recover the support of $\mathbf{\Sigma}$, i.e., the locations of the nonzero entries of $\mathbf{\Sigma}$. The problem has been studied in the case of complete data in, e.g., Cai and Liu [5] and Rothman et al. [24]. With incomplete data, the support can be similarly recovered through adaptive thresholding. Specifically, define the support of $\mathbf{\Sigma} = (\sigma_{ij})_{1 \leq i,j \leq p}$ by $\mathrm{supp}(\mathbf{\Sigma}) = \{(i,j) : \sigma_{ij} \neq 0\}$. Under the condition that the non-zero entries of $\mathbf{\Sigma}$ are sufficiently bounded away from zero, the adaptive thresholding estimator $\hat{\mathbf{\Sigma}}^{\mathrm{at}}$ recovers the support $\mathrm{supp}(\mathbf{\Sigma})$ consistently. It is noteworthy that in the support recovery analysis, the sparsity assumption is not directly needed.

Theorem 3.2 (Support Recovery) *Suppose* $\ln p = o((n^*_{\min})^{1/3})$ *and Assumptions 2.1 and 2.2 hold. Let* γ *be any positive constant. Suppose* $\mathbf{\Sigma}$ *satisfies*

$$|\sigma_{ij}| > (4 + \gamma)\sqrt{\frac{\theta_{ij}\ln p}{n^*_{ij}}}, \quad \textit{for all } (i,j) \in \mathrm{supp}(\mathbf{\Sigma}). \tag{22}$$

Let $\hat{\mathbf{\Sigma}}^{\mathrm{at}}$ *be the adaptive thresholding estimator with* $\delta = 2$, *then, conditioning on* $\mathbf{S}$, *we have*

$$\Pr\left\{\mathrm{supp}(\hat{\mathbf{\Sigma}}^{\mathrm{at}}) = \mathrm{supp}(\mathbf{\Sigma})\right\} \to 1 \quad \textit{as} \quad n,\ p \to \infty. \tag{23}$$

4 Numerical Results

We investigate in this section the numerical performance of the proposed estimators through simulations. The proposed adaptive thresholding procedure is also illustrated with an estimation of the covariance matrix based on data from four ovarian cancer studies.

The estimators $\hat{\mathbf{\Sigma}}^{\mathrm{bt}}$ and $\hat{\mathbf{\Sigma}}^{\mathrm{at}}$ introduced in the previous sections all require specification of the tuning parameters (k or δ). Cross-validation is a simple and practical data-driven method for the selection of these tuning parameters. Numerical results indicate that the proposed estimators with the tuning parameter selected by cross-validation perform well empirically. We begin by introducing the following K-fold cross-validation method for the empirical selection of the tuning parameters.

4.1 Cross-validation

For a pre-specified positive integer N, we construct a grid T of non-negative numbers. For bandable covariance matrix estimation, we set $T = \left\{1, \lceil p^{1/N} \rceil, \ldots, \lceil p^{N/N} \rceil\right\}$, and for sparse covariance matrix estimation, we let $T = \{0, 1/N, \ldots, 4N/N\}$.

Given n samples $\mathbf{X}^* \in \mathbb{R}^{p\times n}$ with missing values, for a given positive integer K, we randomly divide them into two groups of size $n_1 \approx n(K-1)/K$, $n_2 \approx n/K$ for H times. For $h = 1, \ldots, H$, we denote by J_1^h and $J_2^h \subseteq \{1, \ldots, n\}$ the index sets of the two groups for the h-th split. The proposed estimator, $\hat{\mathbf{\Sigma}}^{\mathrm{bt}}$ for bandable covariance matrices, or $\hat{\mathbf{\Sigma}}^{\mathrm{at}}$ for sparse covariance matrices, is then applied to the first group of data $\mathbf{X}^*_{J_1^h}$ with each value of the tuning parameter $t \in T$ and denote the result by $\hat{\mathbf{\Sigma}}^{\mathrm{bt}}_h(t)$ or $\hat{\mathbf{\Sigma}}^{\mathrm{at}}_h(t)$ respectively. Denote the generalized sample covariance matrix of the second group of data $\mathbf{X}^*_{J_2^h}$ by $\hat{\mathbf{\Sigma}}^*_h$ and set

$$\hat{R}(t) = \frac{1}{H}\sum_{h=1}^{H} \|\hat{\mathbf{\Sigma}}_h(t) - \hat{\mathbf{\Sigma}}^*_h\|_F^2, \tag{24}$$

where $\hat{\mathbf{\Sigma}}_h(t)$ is either $\hat{\mathbf{\Sigma}}^{\mathrm{bt}}(t)$ for bandable covariance matrices, or $\hat{\mathbf{\Sigma}}^{\mathrm{at}}(t)$ for sparse covariance matrices.

The final tuning parameter is chosen to be

$$t_* = \arg\min_T \hat{R}(t)$$

and the final estimator $\hat{\mathbf{\Sigma}}^{\mathrm{bt}}$ (or $\hat{\mathbf{\Sigma}}^{\mathrm{at}}$) is calculated using this choice of the tuning parameter t_*. In the following numerical studies, we will use 5-fold cross-validation (i.e., $K = 5$) to select the tuning parameters.

Remark 4.1 The Frobenius norm used in (24) can be replaced by other losses such as the spectral norm. Our simulation results indicate that using the Frobenius norm in (24) works well, even when the true loss is the spectral norm loss.

4.2 Simulation Studies

In the simulation studies, we consider the following two settings for the missingness. The first is MUCR where each entry X_{ik} is observed with probability $0 < \rho \le 1$, and the second is missing not uniformly but completely at random (MCR) where the complete data matrix $\mathbf{X}$ is divided into four equal-size blocks,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{(11)} & \mathbf{X}_{(12)} \\ \mathbf{X}_{(21)} & \mathbf{X}_{(22)} \end{bmatrix}, \quad \mathbf{X}_{(11)}, \mathbf{X}_{(12)}, \mathbf{X}_{(21)}, \mathbf{X}_{(22)} \in \mathbb{R}^{\frac{p}{2} \times \frac{n}{2}},$$

and each entry of $\mathbf{X}_{(11)}$ and $\mathbf{X}_{(22)}$ is observed with probability $\rho^{(1)}$ and each entry of $\mathbf{X}_{(12)}$ and $\mathbf{X}_{(21)}$ is observed with probability $\rho^{(2)}$, for some $0 < \rho^{(1)}, \rho^{(2)} \le 1$.

As mentioned in the introduction, high-dimensional inference for missing data has been studied in the case of MUCR and we would like to compare our estimators with the corresponding estimators based on a different sample covariance matrix designed for the MUCR case. Under the assumption that $\mathrm{E}\mathbf{X} = 0$ and each entry of $\mathbf{X}$ is observed independently with probability ρ, Wainwright [21] and Lounici [23] introduced the following substitute of the usual sample covariance matrix

$$\hat{\mathbf{\Sigma}}^{\bullet} = (\sigma^{\bullet}_{ij})_{1 \le i,j \le p} \quad \text{with} \quad \hat{\sigma}^{\bullet}_{ij} = \begin{cases} \frac{1}{n(1-\rho)^2} \sum_{k=1}^n X^*_{ik} X^*_{jk}, & i \ne j \\ \frac{1}{n(1-\rho)} \sum_{k=1}^n X^*_{ik} X^*_{jk}, & i = j \end{cases} \tag{25}$$

where the missing entries of $\mathbf{X}^*$ are replaced by 0's. It is easy to show that $\hat{\mathbf{\Sigma}}^{\bullet}$ is a consistent estimator of $\mathbf{\Sigma}$ under MUCR and could be used similarly as the sample covariance matrix in the complete data setting.

For more general settings where $\mathrm{E}\mathbf{X} \neq 0$ and the coordinates $X_1, X_2, ..., X_p$ are observed with different probabilities $\rho_1, \ldots, \rho_p$, $\hat{\mathbf{\Sigma}}^{\bullet}$ can be generalized as

$$\hat{\mathbf{\Sigma}}^{\bullet} = (\hat{\sigma}_{ij}^{\bullet})_{1 \le i,j \le p} \quad \text{with} \quad \hat{\sigma}_{ij}^{\bullet} = \begin{cases} \frac{1}{n(1-\hat{\rho}_i)(1-\hat{\rho}_j)} \sum_{k=1}^{n} X_{ik,c}^* X_{jk,c}^*, & i \neq j \\ \\ \frac{1}{n(1-\hat{\rho}_i)} \sum_{k=1}^{n} X_{ik,c}^* X_{jk,c}^*, & i = j \end{cases} \tag{26}$$

where for $i = 1, \ldots, p$ and $k = 1, \ldots, n$, $\hat{\rho}_i = \frac{1}{n} \sum_{k=1}^{n} S_{ik}$ and $X_{ik,c}^* = X_{ik}^* - \bar{X}_i^*$.

Based on $\hat{\mathbf{\Sigma}}^{\bullet}$, we can analogously define the corresponding blockwise tridiagonal estimator $\hat{\mathbf{\Sigma}}^{\mathrm{bt}\bullet}$ for bandable covariance matrices, and adaptive thresholding estimator $\hat{\mathbf{\Sigma}}^{\mathrm{at}\bullet}$ for sparse covariance matrices.

We first consider estimation of bandable covariance matrices and compare the proposed blockwise tridiagonal estimator $\hat{\mathbf{\Sigma}}^{\mathrm{bt}}$ with the corresponding estimator $\hat{\mathbf{\Sigma}}^{\mathrm{bt}\bullet}$. For both methods, the tuning parameter k is selected by 5-fold cross-validation with N varying from 20 to 50. The following bandable covariance matrices are considered:

1. (Linear decaying bandable model) $\mathbf{\Sigma} = (\sigma_{ij})_{1 \le i,j \le p}$ with $\sigma_{ij} = \max\{0, 1 - |i-j|/5\}$.

2. (Squared decaying bandable model) $\mathbf{\Sigma} = (\sigma_{ij})_{1 \le i,j \le p}$ with $\sigma_{ij} = (|i-j|+1)^{-2}$.

For missingness, both MUCR and MCR are considered and (25) and (26) are used to calculate $\hat{\mathbf{\Sigma}}^{\bullet}$ respectively. The proposed procedure $\hat{\mathbf{\Sigma}}^{\mathrm{bt}}$ is compared with the estimator $\hat{\mathbf{\Sigma}}^{\mathrm{bt}\bullet}$, which is based on $\hat{\mathbf{\Sigma}}^{\bullet}$. The results for the spectral norm, ℓ_1 norm and Frobenius norm losses are reported in Table 1. It is easy to see from Table 1 that the proposed estimator $\hat{\mathbf{\Sigma}}^{\mathrm{bt}}$ generally outperforms $\hat{\mathbf{\Sigma}}^{\mathrm{bt}\bullet}$, especially in the fast decaying setting.

Now we consider estimation of sparse covariance matrices with missing values under the following two models.

1. (Permutation Bandable Model) $\mathbf{\Sigma} = (\sigma_{ij})_{1 \le i,j \le p}$, where $\sigma_{ij} = \max(0, 1 - 0.2 \cdot |s(i) - s(j)|)$ and $s(i), i = 1, \ldots, p$ is a random permutation of $\{1, \ldots, p\}$.

	Spectral norm		ℓ_1 norm		Frobenius norm	
(p, n)	$\hat{\boldsymbol{\Sigma}}^{\mathrm{bt}}$	$\hat{\boldsymbol{\Sigma}}^{\mathrm{bt}\bullet}$	$\hat{\boldsymbol{\Sigma}}^{\mathrm{bt}}$	$\hat{\boldsymbol{\Sigma}}^{\mathrm{bt}\bullet}$	$\hat{\boldsymbol{\Sigma}}^{\mathrm{bt}}$	$\hat{\boldsymbol{\Sigma}}^{\mathrm{bt}\bullet}$
Linear Decay Bandable Model, MUCR $\rho = .5$						
$(50, 50)$	2.78(0.17)	2.88(0.18)	4.37(0.57)	4.57(0.76)	7.73(0.85)	7.85(0.80)
$(50, 200)$	1.44(0.06)	1.56(0.07)	2.52(0.17)	2.71(0.19)	3.91(0.18)	4.16(0.16)
$(200, 100)$	2.25(0.13)	2.44(0.16)	3.83(0.32)	4.22(0.46)	10.27(0.29)	10.89(0.29)
$(200, 200)$	1.67(0.07)	1.82(0.08)	2.81(0.19)	3.08(0.22)	7.19(0.19)	7.68(0.14)
$(500, 200)$	2.00(0.07)	2.18(0.10)	3.45(0.16)	3.74(0.27)	12.10(0.36)	12.87(0.42)
Squared Decay Bandable Model, MUCR $\rho = .5$						
$(50, 50)$	1.34(0.08)	1.40(0.11)	2.28(0.16)	2.37(0.21)	3.78(0.19)	3.91(0.18)
$(50, 200)$	0.82(0.01)	0.84(0.01)	1.47(0.03)	1.49(0.02)	2.24(0.02)	2.30(0.02)
$(200, 100)$	1.13(0.01)	1.17(0.02)	2.12(0.05)	2.18(0.07)	5.74(0.04)	5.91(0.05)
$(200, 200)$	0.92(0.00)	0.94(0.00)	1.66(0.02)	1.72(0.03)	4.49(0.02)	4.61(0.01)
$(500, 200)$	0.97(0.00)	0.98(0.00)	1.80(0.02)	1.86(0.02)	7.15(0.01)	7.35(0.01)
Linear Decay Bandable Model, MCR $\rho^{(1)} = .8, \rho^{(2)} = .2$						
$(50, 50)$	2.76(0.26)	3.46(1.43)	4.24(0.73)	5.87(2.91)	7.03(1.25)	8.47(1.29)
$(50, 200)$	1.51(0.11)	2.64(0.40)	2.52(0.30)	4.29(0.99)	3.62(0.30)	5.77(0.45)
$(200, 100)$	2.32(0.22)	3.93(0.67)	3.73(0.47)	6.21(1.11)	9.04(0.48)	13.47(0.84)
$(200, 200)$	1.67(0.10)	3.23(0.27)	2.71(0.26)	4.91(0.49)	6.32(0.11)	11.32(0.49)
$(500, 200)$	1.98(0.09)	3.78(0.20)	3.19(0.20)	5.70(0.42)	10.39(0.12)	18.48(0.49)
Squared Decay Bandable Model, MCR $\rho^{(1)} = .8, \rho^{(2)} = .2$						
$(50, 50)$	1.26(0.08)	1.49(0.13)	2.21(0.23)	2.60(0.28)	3.48(0.14)	4.18(0.23)
$(50, 200)$	0.82(0.01)	0.88(0.04)	1.47(0.05)	1.77(0.11)	2.18(0.04)	2.68(0.11)
$(200, 100)$	1.06(0.01)	1.30(0.04)	1.96(0.04)	2.44(0.07)	5.32(0.02)	6.51(0.06)
$(200, 200)$	0.90(0.00)	0.96(0.03)	1.60(0.02)	1.99(0.06)	4.27(0.02)	5.26(0.15)
$(500, 200)$	0.93(0.00)	1.03(0.01)	1.69(0.01)	2.11(0.03)	6.73(0.01)	8.25(0.04)

Table 1: Comparsion between $\hat{\boldsymbol{\Sigma}}^{\mathrm{bt}}$ and $\hat{\boldsymbol{\Sigma}}^{\mathrm{bt}\bullet}$ in different settings of bandable covariance matrix estimation.

2. (Randomly Sparse Model) $\boldsymbol{\Sigma} = \mathbf{I}_p + (\mathbf{D}+\mathbf{D}^\top)/(\|\mathbf{D}+\mathbf{D}^\top\|+0.01)$, where $\mathbf{D}$ is randomly generated as

$$\mathbf{D} = (d_{ij})_{1\le i,j\le p}, \quad d_{ij} = \begin{cases} 1 & \text{w.p. } 0.1 \\ 0 & \text{w.p. } 0.8 \\ -1 & \text{w.p. } 0.1 \end{cases} \quad \text{for } i \neq j; \quad d_{ii} = 0.$$

Similar to the sparse covariance matrix estimation, for missingness, we consider both MUCR and MCR. The results for the spectral norm, matrix ℓ_1 norm and Frobenius norm losses are summarized in Table 2. It can be seen from Table 2 that, even under the MUCR setting, the proposed estimator $\hat{\boldsymbol{\Sigma}}^{\text{at}}$ based on the generalized sample covariance matrix is uniformly better than the one based on $\hat{\boldsymbol{\Sigma}}^{\bullet}$. In the more general MCR setting, the difference in the performance between the two estimators is even more significant.

4.3 Comparison with Complete Samples

For covariance matrix estimation with missing data, an interesting question is: what is the "effective sample size"? That is, for samples with missing values, we would like to know the equivalent size of complete samples such that the accuracy for covariance matrix estimation is approximately the same. We now compare the performance of the proposed estimator based on the incomplete data with the corresponding estimator based on the complete data for various sample sizes. We fix the dimension $p = 100$. For the incomplete data, we consider $n = 1000$ and MUCR with $\rho = .5$. The covariance matrix $\boldsymbol{\Sigma}$ is chosen as

- Linear Decaying Bandable Model (in Bandable Covariance Matrix Estimation);
- Permutation Bandable Model (in Sparse Covariance Matrix Estimation);

Correspondingly, we consider the similar settings for the complete data with the same $\boldsymbol{\Sigma}$ and p, but different sample size n_c, where n_c can be one of the following three values,

1. $\overline{n^*_{\text{pair}}} = \sum_{i,j=1}^n n^*_{ij}/p^2$: the average number of pairs of (x_i, x_j)'s that can be observed within the same sample;

	Spectral norm		ℓ_1 norm		Frobenius norm	
(p, n)	$\hat{\boldsymbol{\Sigma}}^{\text{at}}$	$\hat{\boldsymbol{\Sigma}}^{\text{at}\bullet}$	$\hat{\boldsymbol{\Sigma}}^{\text{at}}$	$\hat{\boldsymbol{\Sigma}}^{\text{at}\bullet}$	$\hat{\boldsymbol{\Sigma}}^{\text{at}}$	$\hat{\boldsymbol{\Sigma}}^{\text{at}\bullet}$
Permutation Bandable Model, MUCR $\rho = .5$						
$(50, 50)$	4.26(0.24)	4.45(0.41)	5.58(0.58)	6.19(7.54)	11.34(0.79)	11.73(1.08)
$(50, 200)$	1.70(0.05)	1.74(0.06)	3.31(0.32)	3.42(0.38)	4.93(0.09)	5.07(0.16)
$(200, 100)$	3.48(0.07)	3.66(0.58)	5.80(0.39)	6.23(14.89)	18.34(0.81)	19.37(5.50)
$(200, 200)$	2.12(0.04)	2.20(0.03)	4.17(0.29)	4.44(0.32)	11.46(0.14)	11.94(0.13)
$(500, 200)$	2.28(0.03)	3.51(0.17)	4.17(0.15)	6.55(0.72)	16.85(0.10)	21.96(0.49)
Randomly Sparse Model, MUCR $\rho = .5$						
$(50, 50)$	1.76(0.07)	1.96(0.62)	3.69(0.24)	4.20(5.89)	5.75(0.51)	6.27(2.95)
$(50, 200)$	1.05(0.00)	1.06(0.00)	2.73(0.04)	2.74(0.05)	3.75(0.03)	3.77(0.04)
$(200, 100)$	1.40(0.01)	1.45(0.01)	4.88(0.08)	4.94(0.09)	8.34(0.07)	8.50(0.07)
$(200, 200)$	1.07(0.00)	1.09(0.01)	4.44(0.03)	4.46(0.03)	7.42(0.02)	7.43(0.02)
$(500, 200)$	1.14(0.01)	1.31(0.01)	6.39(0.04)	6.65(0.08)	11.73(0.01)	12.23(0.05)
Permutation Bandable Model, MCR $\rho^{(1)} = .8, \rho^{(2)} = .2$						
$(50, 50)$	4.23(0.38)	4.71(1.17)	6.67(2.30)	7.46(8.92)	11.22(1.34)	11.71(2.01)
$(50, 200)$	1.64(0.05)	2.79(0.39)	2.94(0.21)	4.52(0.95)	4.41(0.13)	6.29(0.46)
$(200, 100)$	3.17(0.06)	4.16(0.57)	5.73(0.66)	8.11(1.87)	15.93(0.53)	18.03(0.77)
$(200, 200)$	2.00(0.03)	3.22(0.18)	3.65(0.16)	5.70(0.60)	9.83(0.11)	13.29(0.55)
$(500, 200)$	2.22(0.03)	3.45(0.17)	4.09(0.17)	6.44(0.96)	16.80(0.14)	21.93(0.45)
Randomly Sparse Model, MCR $\rho^{(1)} = .8, \rho^{(2)} = .2$						
$(50, 50)$	2.15(0.46)	2.19(0.49)	4.21(0.94)	4.47(4.65)	6.36(0.96)	7.25(1.57)
$(50, 200)$	1.09(0.02)	1.16(0.04)	2.82(0.19)	2.99(0.32)	3.83(0.10)	4.00(0.20)
$(200, 100)$	1.46(0.02)	1.82(0.03)	4.96(0.12)	5.61(0.21)	8.45(0.07)	10.10(0.14)
$(200, 200)$	1.08(0.00)	1.20(0.01)	4.46(0.04)	4.57(0.05)	7.43(0.02)	7.66(0.04)
$(500, 200)$	1.12(0.01)	1.33(0.01)	6.35(0.04)	6.60(0.07)	11.71(0.02)	12.20(0.06)

Table 2: Comparsion between $\hat{\boldsymbol{\Sigma}}^{\text{at}}$ and $\hat{\boldsymbol{\Sigma}}^{\text{at}\bullet}$ in different settings of sparse covariance matrix estimation.

Setting	sample size	Spectral norm	ℓ_1 norm	Frobenius norm
		Bandable Covariance Matrix Estimation		
Missing Data	$n = 1000$	0.72(0.01)	1.25(0.03)	2.40(0.01)
Complete Data	$n_c = \overline{n^*_{\text{pair}}}$	0.97(0.03)	1.49(0.05)	2.48(0.04)
Complete Data	$n_c = \overline{n^*_{\text{s}}}$	0.65(0.01)	1.01(0.03)	1.69(0.03)
Complete Data	$n_c = n$	0.48(0.01)	0.73(0.01)	1.22(0.01)
		Sparse Covariance Matrix Estimation		
Missing Data	$n = 1000$	0.75(0.01)	1.37(0.04)	2.90(0.02)
Complete Data	$n_c = \overline{n^*_{\text{pair}}}$	0.83(0.02)	1.31(0.05)	2.94(0.04)
Complete Data	$n_c = \overline{n^*_{\text{s}}}$	0.65(0.01)	1.01(0.03)	1.86(0.04)
Complete Data	$n_c = n$	0.45(0.01)	0.64(0.01)	1.12(0.01)

Table 3: Comparison between incomplete samples and complete samples.

2. $\overline{n^*_{\text{s}}} = \sum_{i=1}^n n^*_i/p$: the average number of single x_i's can be observed;

3. n: the same number of samples with the missing values.

The results for all the settings are summarized in Table 3. It can be seen that the equivalent sample size depends on the loss function and in general it is between $\overline{n^*_{\text{pair}}}$ and $\overline{n^*_{\text{s}}}$. Overall, the average risk under the missing data setting is most comparable to that under the complete data setting for the sample size of $n_c = \overline{n^*_{\text{pair}}}$, the average number of observed pairs.

4.4 Analysis of Ovarian Cancer Data

In this section, we illustrate the proposed adaptive thresholding procedure with an application to data from four ovarian cancer genomic studies, Cancer Genome Atlas Research Network [11] (TCGA), Bonome et al. [4] (BONO), Dressman et al. [15] (DRES) and Tothill et al. [27] (TOTH). The method introduced in Sections 3 enables us to estimate the covariance matrix by integrating data from all four studies and thus yields a more accurate estimator. The data structure is illustrated in Figure 2. The gene expression markers (the first 426 rows) are observed in all four studies without any missingness (the top black block in Figure 2). The miRNA expression markers are observed in 552 samples from the TCGA study (the bottom left block in Figure 2) and completely missing in the 881 samples from the TOTH, DRES, BONO and part of TCGA studies (the white block in Figure 2).

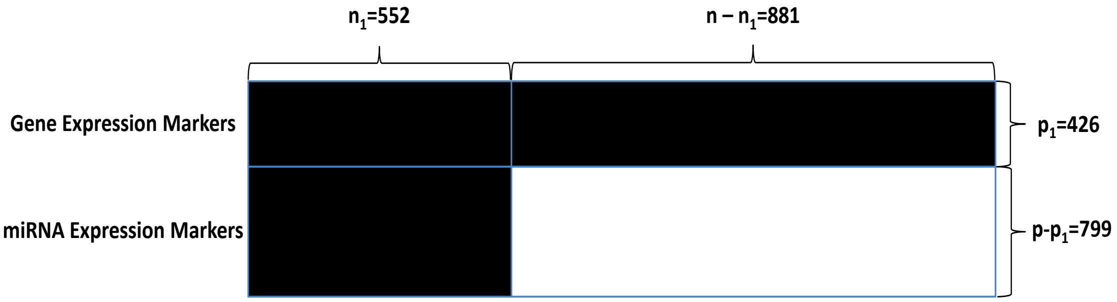
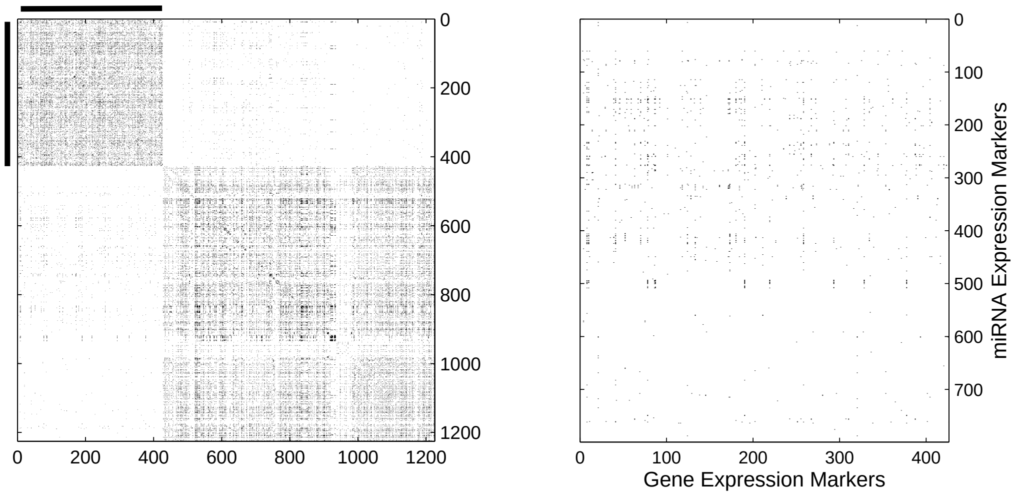


Figure 2: Illustration of the ovarian cancer dataset. Black block = completely observed; White block = completely missing.

Our goal is to estimate the covariance matrix $\mathbf{\Sigma}$ of the 1225 variables with the particular interest in the cross-covariances between the gene and miRNA expression markers. It is clear that the missingness here is not uniformly at random. On the other hand, it is reasonable to assume the missingness does not depend on the value of the data and thus missing completely at random (Assumption 2.1) can be assumed. We apply the adaptive thresholding procedure with $\delta = 2$ to estimate the covariance matrix and recover its support based on all the observations. The support of the estimate is shown in a heatmap in Figure 3. The left panel is for the whole covariance matrix and the right panel zooms into the cross-covariances between the gene and miRNA expression markers.

It can be seen from Figure 3 that the two diagonal blocks, with 12.24% and 8.39% nonzero off-diagonal entries respectively, are relatively dense, indicating that the relationships among



(a) Covariance matrix of the gene and miRNA expression markers. The gene expression markers are marked with lines.

(b) Cross-covariances between the gene and miRNA expression markers. 1294 (.38%) gene-miRNA pairs were detected.

Figure 3: Heatmaps of the covariance matrix estimate with all the observed data.

the gene expression markers and those among the miRNA expression markers, as measured by their covariances, are relatively close. In contrast, the cross-covariances between gene and miRNA expression markers are very sparse with only 0.38% of significant gene-miRNA pairs. The gene and miRNA expression markers affect each other through different mechanisms, the cross-covariances between the gene and miRNA markers are of significant interest (see Ko et al. [19]). It is worthwhile to take a closer look at the cross-covariance matrix displayed on the right panel in Figure 3. For each given gene, we count the number of miRNAs whose covariances with this gene are significant, and then rank all the genes by the counts. Similarly, we rank all the miRNAs. The top 5 genes and the top 5 miRNA expression markers are shown in Table 4.4.

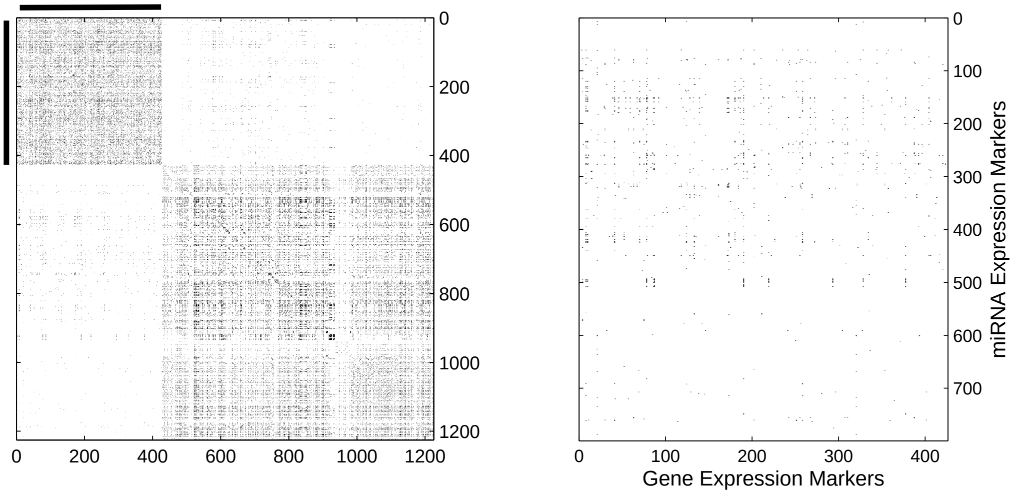
Many of these gene and miRNA expression markers have been studied before in the literature. For example, the miRNA expression markers hsa-miR-142-5p and hsa-miR-142-3p have been demonstrated in Andreopoulos and Anastassiou [1] as standing out among the miRNA

Gene Expression Marker	Counts	miRNA Expression Marker	Counts
ACTA2	61	hsa-miR-142-5p	31
INHBA	57	hsa-miR-142-3p	29
COL10A1	53	hsa-miR-22	26
BGN	46	hsa-miR-21*	24
NID1	41	hsa-miR-146a	21

Table 4: Genes and miRNA's with most selected pairs

markers as having higher correlations with more genes, as well as methylation sites. Carraro et al. [12] finds that inhibition of miR-142-3p leads to ectopic expression of the gene marker ACTA2. This indicates strong interaction between miR-142-3p and ACTA2.

To further demonstrate the robustness of our proposed procedure against missingness, we consider a setting with additional missing observations. We first randomly select half of the 552 complete samples (where both gene and miRNA expression markers are observed) and half of the 881 incomplete samples (where only gene expression markers are observed), and then independently mask each entry of the selected samples with probability 0.05. The proposed adaptive thresholding procedure is then applied to the data with these additional missing values. The estimated covariance matrix is shown in heatmaps in Figure 4. These additional missing observations do not significantly affect the estimation accuracy. Figure 4 is visually very similar to Figure 3. To quantify the similarity between the two estimates, we calculate the Matthews correlation coefficient (MCC) between them. The value of MCC is equal to 0.9441, which indicates that the estimate based on the data with the additional missingness is very close to the estimate based on the original samples. We also pay close attention to the cross-covariance matrix displayed on the right panel in Figure 4 and rank the gene and miRNA expression markers in the same way as before. The top 5 genes and the top 5 miRNA expression markers, listed in Table 5, are nearly identical to those given in Table 4.4, which are based on the original samples. These results indicate that the proposed method is robust against additional missingness.



(a) Covariance matrix of the gene and miRNA expression markers. The gene expression markers are marked with lines.

(b) Cross-covariances between the gene and miRNA expression markers. 1176 (.35%) gene-miRNA pairs were detected.

Figure 4: Heatmaps of the covariance matrix estimate with additional missing values.

5 Discussions

We considered in the present paper estimation of bandable and sparse covariance matrices in the presence of missing observations. The pivotal quantity is the generalized sample covariance matrix defined in (7). The technical analysis is more challenging due to the missing data. We have mainly focused on the spectral norm loss in the theoretical analysis. Performance under other losses such as the Frobenius norm can also be analyzed.

To illustrate the proposed methods, we integrated four ovarian cancer studies. These methods for high-dimensional covariance matrix estimation with missing data are also useful for other types of data integration. For example, linking multiple data sources such as electronic data records, medicare data, registry data and patient reported outcomes could greatly

Gene Expression Marker	Counts	miRNA Expression Marker	Counts
ACTA2	60	hsa-miR-142-3p	31
INHBA	56	hsa-miR-142-5p	30
COL10A1	50	hsa-miR-146a	21
BGN	43	hsa-miR-150	21
NID1	40	hsa-miR-21*	21

Table 5: Genes and miRNA's with most selected pairs after masking

increase the power of exploratory studies such as phenome-wide association studies (Denny et al. [14]). However, missing data inevitably arises and may hinder the potential of integrative analysis. In addition to random missingness due to unavailable information on a small fraction of patients, many variables such as the genetic measurements may only exist in one or two data sources and are hence structurally missing for other data sources. Our proposed methods could potentially provide accurate recovery of the covariance matrix in the presence of missingness.

In this paper, we allowed the proportion of missing values to be non-negligible as long as the minimum number of occurrences of any pair of variables $n^*_{\min}$ is of order n. An interesting question is what happens when the number of observed values is large but $n^*_{\min}$ is small (or even zero). We believe that the covariance matrix $\mathbf{\Sigma}$ can still be well estimated under certain global structural assumptions. This is out of the scope of the present paper and is an interesting problem for future research.

The key ideas and techniques developed in this paper can be used for a range of other related problems in high-dimensional statistical inference with missing data. For example, the same techniques can also be applied to estimation of other structured covariance matrices such as Toeplitz matrices, which have been studied in the literature in the case of complete data. When there are missing data, we can construct similar estimators using the generalized sample covariance matrix. The large deviation bounds for a sub-matrix and self-normalized entries of the generalized sample covariance matrix developed in Lemmas 3.1 and 2.1 would be helpful for analyzing the properties of the estimators.

The techniques can also be used on two-sample problems such as estimation of differential correlation matrices and hypothesis testing on the covariance structures. The generalized sample covariance matrix can be standardized to form the generalized sample correlation matrix which can then be used to estimate the differential correlation matrix in the two-sample case. It is also of significant interest in some applications to test the covariance structures in both one- and two-sample settings based on incomplete data. In the one-sample case, it is of interest to test the hypothesis $\{H_0 : \mathbf{\Sigma} = \mathbf{I}\}$ or $\{H_0 : \mathbf{R} = \mathbf{I}\}$, where $\mathbf{R}$ is the correlation matrix. In the two-sample case, one wishes to test the equality of two covariance matrices $\{H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2\}$. These are interesting problems for further exploration in the future.

6 Proofs

We prove Theorem 2.1 and the key technical result Lemma 6.1 for the bandable covariance matrix estimation in this section.

6.1 Proof of Lemma 2.1

To prove this lemma, we first introduce the following technical tool for the spectral norm of the sub-matrices.

Lemma 6.1 *Suppose $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ is any positive semi-definite matrix, $A, B \in \{1, \ldots, p\}$, then*

$$\|\mathbf{\Sigma}_{A\times B}\| \leq (\|\mathbf{\Sigma}_A\| \|\mathbf{\Sigma}_B\|)^{1/2}. \tag{27}$$

The proof of Lemma 6.1 is provided later and now we move back to the proof of Lemma 2.1. Without loss of generality, we assume that $\boldsymbol{\mu} = \mathrm{E}\mathbf{X} = 0$. We further define

$$\breve{\mathbf{\Sigma}}^* = (\breve{\sigma}^*_{ij})_{1\leq i,j\leq p}, \quad \breve{\sigma}^*_{ij} = \frac{1}{n^*_{ij}} \sum_{k=1}^{n} X_{ik}X_{jk}S_{ik}S_{jk}. \tag{28}$$

Also for convenience of presentation, we use $C, C_1, c, \ldots$ to denote uniform constants, whose exact values may vary in different senarios. The lemma is now proved in the following steps:

1. We first consider for fixed unit vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ with $\mathrm{supp}(\mathbf{a}) \subseteq A, \mathrm{supp}(\mathbf{b}) \subseteq B$, the tail bound of $\mathbf{a}^\top(\hat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma})\mathbf{b}$. We would like to show that there exist uniform constants $C_1, c > 0$ such that for all $x > 0$,

   $$\begin{aligned}&\Pr\left\{\left|\mathbf{a}^\top\left(\hat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}\right)\mathbf{b}\right| \geq x\right\}\\ \leq & C_1 \exp\left\{-cn^*_{\min}\min\left(\frac{x^2}{\tau^4\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|}, \frac{x}{\tau^2(\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|)^{1/2}}\right)\right\}.\end{aligned} \tag{29}$$

   Specifically, we will bound $\mathbf{a}^\top(\breve{\boldsymbol{\Sigma}}^* - \hat{\boldsymbol{\Sigma}})\mathbf{b}$ and $\mathbf{a}^\top(\breve{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma})\mathbf{b}$ separately in the next two steps.

2. We consider $\mathbf{a}^\top(\breve{\boldsymbol{\Sigma}}^* - \hat{\boldsymbol{\Sigma}})\mathbf{b}$ first. Since

   $$\breve{\sigma}^*_{ij} - \hat{\sigma}^*_{ij} = \frac{1}{n^*_{ij}}\sum_{k=1}^n (X_{jk}\bar{X}^*_i + X_{ik}\bar{X}^*_j)S_{ik}S_{jk} - \bar{X}^*_i\bar{X}^*_j,$$

   $\mathbf{a}^\top(\breve{\boldsymbol{\Sigma}}^* - \hat{\boldsymbol{\Sigma}}^*)\mathbf{b}$ can be written as

   $$\begin{aligned}&\mathbf{a}^\top(\breve{\boldsymbol{\Sigma}}^* - \hat{\boldsymbol{\Sigma}}^*)\mathbf{b} = \sum_{i,j=1}^p a_i b_j(\breve{\sigma}^*_{ij} - \hat{\sigma}^*_{ij})\\ =&\sum_{i,j=1}^p a_i b_j\Bigg(\frac{\sum_{k=1}^n X_{ik}S_{ik}}{n^*_i}\cdot\frac{\sum_{l=1}^n X_{jl}S_{il}S_{jl}}{n^*_{ij}}\\ &+\frac{\sum_{k=1}^n X_{ik}S_{ik}S_{jk}}{n^*_{ij}}\cdot\frac{\sum_{l=1}^n X_{jl}S_{jl}}{n^*_j} - \frac{\sum_{k=1}^n X_{ik}S_{ik}}{n^*_i}\cdot\frac{\sum_{l=1}^n X_{jl}S_{jl}}{n^*_j}\Bigg)\\ =&\sum_{i,j=1}^p\sum_{k,l=1}^n X_{ik}X_{jl}a_ib_j\left(\frac{S_{ik}S_{il}S_{jl}}{n^*_in^*_{ij}} + \frac{S_{ik}S_{jk}S_{jl}}{n^*_{ij}n^*_j} - \frac{S_{ik}S_{jl}}{n^*_in^*_j}\right).\end{aligned} \tag{30}$$

We can calculate from (30) that

$$\begin{aligned}
&\left|\mathrm{E}\mathbf{a}^\top(\breve{\boldsymbol{\Sigma}}^* - \hat{\boldsymbol{\Sigma}}^*)\mathbf{b}\right| \\
=&\left|\sum_{i,j=1}^{p}\sum_{k=1}^{n}\sigma_{ij}a_ib_j\left(\frac{S_{ik}S_{ik}S_{jk}}{n_i^*n_{ij}^*} + \frac{S_{ik}S_{jk}S_{jk}}{n_{ij}^*n_j^*} - \frac{S_{ik}S_{jk}}{n_i^*n_j^*}\right)\right| \\
=&\left|\sum_{i,j=1}^{p}\sigma_{ij}\frac{a_i}{n_i^*}b_j + \sum_{i,j=1}^{p}a_i\frac{b_j}{n_j^*}\sigma_{ij} - \sum_{k=1}^{n}\sum_{i,j=1}^{p}\frac{S_{ik}a_i}{n_i^*}\frac{S_{jk}b_j}{n_j^*}\sigma_{ij}\right| \\
\leq&\left|\left(\frac{a_1}{n_1^*},\dots,\frac{a_p}{n_p^*}\right)\boldsymbol{\Sigma}\mathbf{b}\right| + \left|\mathbf{a}^\top\boldsymbol{\Sigma}\left(\frac{b_1}{n_1^*},\dots,\frac{b_p}{n_p^*}\right)^\top\right| \qquad (31)\\
&+\sum_{k=1}^{n}\left|\left(\frac{S_{1k}a_1}{n_1^*},\dots,\frac{S_{pk}a_p}{n_p^*}\right)\boldsymbol{\Sigma}\left(\frac{S_{1k}b_1}{n_1^*},\dots,\frac{S_{pk}b_p}{n_p^*}\right)^\top\right| \\
\leq&\|\boldsymbol{\Sigma}_{A\times B}\|\frac{\|\mathbf{a}\|_2\|\mathbf{b}\|_2}{n_{\min}^*} + \|\boldsymbol{\Sigma}_{A\times B}\|\frac{\|\mathbf{a}\|_2\|\mathbf{b}\|_2}{n_{\min}^*} \\
&+\sum_{k=1}^{n}\|\boldsymbol{\Sigma}_{A\times B}\|\cdot\frac{1}{2}\left\{\left\|\left(\frac{S_{1k}a_1}{n_1^*},\dots,\frac{S_{pk}a_p}{n_p^*}\right)\right\|_2^2 + \left\|\left(\frac{S_{1k}b_1}{n_1^*},\dots,\frac{S_{pk}b_p}{n_p^*}\right)\right\|_2^2\right\}.
\end{aligned}$$

For the last term in (31), we have the following bound,

$$\begin{aligned}
&\sum_{k=1}^{n}\|\boldsymbol{\Sigma}_{A\times B}\|\cdot\frac{1}{2}\left\{\left\|\left(\frac{S_{1k}a_1}{n_1^*},\dots,\frac{S_{pk}a_p}{n_p^*}\right)\right\|_2^2 + \left\|\left(\frac{S_{1k}b_1}{n_1^*},\dots,\frac{S_{pk}b_p}{n_p^*}\right)\right\|_2^2\right\} \\
=&\|\boldsymbol{\Sigma}_{A\times B}\|\sum_{k=1}^{n}\sum_{i=1}^{p}\frac{1}{2}\left(\frac{S_{ik}a_i^2}{n_i^{*2}} + \frac{S_{ik}b_i^2}{n_i^{*2}}\right) \\
=&\|\boldsymbol{\Sigma}_{A\times B}\|\sum_{i=1}^{p}\frac{1}{2}\left(\frac{a_i^2+b_i^2}{n_i^*}\right) \\
\leq&\|\boldsymbol{\Sigma}_{A\times B}\|\sum_{i=1}^{p}\frac{a_i^2+b_i^2}{2n_{\min}^*} \leq \frac{(\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|)^{1/2}}{n_{\min}^*}.
\end{aligned}$$

Thus, by (31) and the inequality above, we have

$$\left|\mathrm{E}a^\top(\breve{\boldsymbol{\Sigma}}^* - \hat{\boldsymbol{\Sigma}}^*)b\right| \leq \frac{3(\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|)^{1/2}}{n_{\min}^*}. \qquad (32)$$

The last term of (30) can be treated as a quadratic form of the vectorization of $\mathbf{X}$: $\mathrm{vec}(\mathbf{X}) \in \mathbb{R}^{pn}$. We note the last term as $\mathrm{vec}(\mathbf{X})^\top\mathbf{Q}\mathrm{vec}(\mathbf{X})$, where $\mathbf{Q} \in \mathbb{R}^{pn\times pn}$ and

$$\mathbf{Q}_{(i,k),(j,l)} = a_ib_j\left(\frac{S_{ik}S_{il}S_{jl}}{n_i^*n_{ij}^*} + \frac{S_{ik}S_{jk}S_{jl}}{n_{ij}^*n_j^*} - \frac{S_{ik}S_{jl}}{n_i^*n_j^*}\right), \quad 1 \leq i,j \leq p, 1 \leq k,l \leq n.$$

$\mathbf{Q}$ has the following properties,

$$\begin{aligned}\|\mathbf{Q}\|_F^2 =& \sum_{i,j=1}^{p}\sum_{k,l=1}^{n} a_i^2 b_j^2 \left(\frac{S_{ik}S_{il}S_{jl}}{n_i^* n_{ij}^*} + \frac{S_{ik}S_{jk}S_{jl}}{n_{ij}^* n_j^*} - \frac{S_{ik}S_{jl}}{n_i^* n_j^*}\right)^2 \\ \leq & \sum_{i,j=1}^{p} a_i^2 b_j^2 \sum_{k,l=1}^{n}\left(2\frac{S_{ik}S_{il}S_{jl}}{n_i^{*2} n_{ij}^{*2}} + 2\frac{S_{ik}S_{jk}S_{jl}}{n_{ij}^{*2} n_j^{*2}} + \frac{S_{ik}S_{jl}}{n_i^{*2} n_j^{*2}}\right), \quad \text{since } S_{ik} \in \{0,1\}; \\ \leq & \sum_{i,j=1}^{p} a_i^2 b_j^2 \frac{5}{n_{\min}^{*2}} = \frac{5\|\mathbf{a}\|_2^2\|\mathbf{b}\|_2^2}{n_{\min}^{*2}} = \frac{5}{n_{\min}^{*2}};\end{aligned} \tag{33}$$

$$\|\mathbf{Q}\| \leq \|\mathbf{Q}\|_F \leq \frac{\sqrt{5}\|\mathbf{a}\|_2\|\mathbf{b}\|_2}{n_{\min}^*} \leq \frac{\sqrt{5}}{n_{\min}^*}. \tag{34}$$

For $\text{vec}(\mathbf{X}) \in \mathbb{R}^{pn}$, since its segments $\{\mathbf{X}_k, k = 1, \ldots, p\}$ are independent and $\mathbf{X}_k = \mathbf{\Gamma}\mathbf{Z}_k$, we can further write $\text{vec}(\mathbf{X}) = \mathbf{D}_{\mathbf{\Gamma}}\text{vec}(\mathbf{Z})$, where $\mathbf{D}_{\mathbf{\Gamma}} \in \mathbb{R}^{pn\times qn}$ is with n diagonal blocks of $\mathbf{\Gamma}$, $\text{vec}(\mathbf{Z})$ is a (qn)-dimensional i.i.d. sub-Gaussian random vector. Based on Hanson-Wright's inequality (Theorem 1.1 in Rudelson and Vershynin [25]),

$$\begin{aligned} & \Pr\left\{\left|\mathbf{a}^\top\left(\breve{\mathbf{\Sigma}}^* - \hat{\mathbf{\Sigma}}^*\right)\mathbf{b} - \text{E}\mathbf{a}^\top\left(\breve{\mathbf{\Sigma}}^* - \hat{\mathbf{\Sigma}}^*\right)\mathbf{b}\right| \geq x\right\} \\ = & \Pr\left\{\left|\text{vec}(\mathbf{X})^\top\mathbf{Q}\text{vec}(\mathbf{X}) - \text{E}\text{vec}(\mathbf{X})^\top\mathbf{Q}\text{vec}(\mathbf{X})\right| \geq x\right\} \\ = & \Pr\left[\left|\text{vec}(\mathbf{Z})^\top\mathbf{D}_{\mathbf{\Gamma}}^\top\mathbf{Q}\mathbf{D}_{\mathbf{\Gamma}}\text{vec}(\mathbf{Z}) - \text{E}\left\{\text{vec}(\mathbf{Z})^\top\mathbf{D}_{\mathbf{\Gamma}}^\top\mathbf{Q}\mathbf{D}_{\mathbf{\Gamma}}\text{vec}(\mathbf{Z})\right\}\right| \geq x\right] \\ \leq & 2\exp\left\{-c\min\left(\frac{x^2}{\tau^4\|\mathbf{D}_{\mathbf{\Gamma}}^\top\mathbf{Q}\mathbf{D}_{\mathbf{\Gamma}}\|_F^2}, \frac{x}{\tau^2\|\mathbf{D}_{\mathbf{\Gamma}}\mathbf{Q}\mathbf{D}_{\mathbf{\Gamma}}\|}\right)\right\}.\end{aligned} \tag{35}$$

Here $c > 0$ is a uniform constant. Since $\mathbf{Q}$ is supported on $\{(i,k),(j,l) : i \in A, j \in B\}$, we have $\mathbf{D}_{\mathbf{\Gamma}}^\top\mathbf{Q}\mathbf{D}_{\mathbf{\Gamma}} = \mathbf{D}_{\mathbf{\Gamma}_A}^\top\mathbf{Q}_{A\times B}\mathbf{D}_{\mathbf{\Gamma}_B}$. Here $\mathbf{D}_{\mathbf{\Gamma}_A} \in \mathbb{R}^{|A|n\times qn}, \mathbf{D}_{\mathbf{\Gamma}_B} \in \mathbb{R}^{|B|n\times qn}$ are with n diagonal block $\mathbf{\Gamma}_{A\times[q]}$ and $\mathbf{\Gamma}_{B\times[q]}$, respectively, where $[q] = \{1, \ldots, q\}$. Since $\mathbf{\Gamma}_{A\times[q]}\mathbf{\Gamma}_{A\times[q]}^\top = \mathbf{\Sigma}_A$, $\mathbf{\Gamma}_{B\times[q]}\mathbf{\Gamma}_{B\times[q]}^\top = \mathbf{\Sigma}_B$, we know

$$\|\mathbf{D}_{\mathbf{\Gamma}_A}\| = \|\mathbf{\Gamma}_{A\times[q]}\| \leq \|\mathbf{\Sigma}_A\|^{1/2}, \quad \|\mathbf{\Gamma}_{B\times[q]}\| \leq \|\mathbf{D}_{\mathbf{\Gamma}_B}\| \leq \|\mathbf{\Sigma}_B\|^{1/2}.$$

Then we further have

$$\begin{aligned}
&\Pr\left\{\left|\mathbf{a}^\top\left(\breve{\boldsymbol{\Sigma}}^*-\hat{\boldsymbol{\Sigma}}^*\right)\mathbf{b}-\mathrm{E}\mathbf{a}^\top\left(\breve{\boldsymbol{\Sigma}}^*-\hat{\boldsymbol{\Sigma}}^*\right)\mathbf{b}\right|\geq x\right\}\\
\leq&2\exp\left\{-c\min\left(\frac{x^2}{\tau^4\|\mathbf{D}_{\boldsymbol{\Gamma}_A}^\top\mathbf{Q}_{A\times B}\mathbf{D}_{\boldsymbol{\Gamma}_B}\|_F^2},\frac{x}{\tau^2\|\mathbf{D}_{\boldsymbol{\Gamma}_A}^\top\mathbf{Q}_{A\times B}\mathbf{D}_{\boldsymbol{\Gamma}_B}\|}\right)\right\}\\
\leq&2\exp\left\{-c\min\left(\frac{x^2}{\tau^4\|\mathbf{D}_{\boldsymbol{\Gamma}_B}\|^2\|\mathbf{D}_{\boldsymbol{\Gamma}_A}^\top\|^2\|\mathbf{Q}\|_F^2},\frac{x}{\tau^2\|\mathbf{D}_{\boldsymbol{\Gamma}_B}\|\|\mathbf{D}_{\boldsymbol{\Gamma}_A}^\top\|\|\mathbf{Q}\|}\right)\right\}\\
\leq&2\exp\left[-c\min\left\{\frac{x^2}{\tau^4\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|\|\mathbf{Q}\|_F^2},\frac{x}{\tau^2(\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|)^{1/2}\|\mathbf{Q}\|}\right\}\right]\\
\leq&2\exp\left[-c\min\left\{\frac{x^2n_{\min}^{*2}}{\tau^4\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|},\frac{xn_{\min}^*}{\tau^2(\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|)^{1/2}}\right\}\right].
\end{aligned}\tag{36}$$

We define $x'=\max\left\{x-3(\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|)^{1/2}/n_{\min}^*,0\right\}$, combining the inequality above and (32), we have

$$\begin{aligned}
&\Pr\left\{\left|\mathbf{a}^\top\left(\breve{\boldsymbol{\Sigma}}^*-\hat{\boldsymbol{\Sigma}}^*\right)\mathbf{b}\right|\geq x\right\}\leq\Pr\left\{\left|\mathbf{a}^\top\left(\breve{\boldsymbol{\Sigma}}^*-\hat{\boldsymbol{\Sigma}}^*\right)\mathbf{b}-\mathrm{E}\mathbf{a}^\top\hat{\boldsymbol{\Sigma}}^*\mathbf{b}\right|\geq x'\right\}\\
\leq&2\exp\left[-c\min\left\{\frac{(x')^2n_{\min}^{*2}}{\tau^4\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|},\frac{x'n_{\min}^*}{\tau^2(\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|)^{1/2}}\right\}\right]\\
\leq&2\exp\left[-c'\min\left\{\frac{x^2n_{\min}^{*2}}{\tau^4\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|},\frac{xn_{\min}^*}{\tau^2(\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|)^{1/2}}\right\}+C\max\left(\frac{1}{\tau^4},\frac{1}{\tau^2}\right)\right]\\
\leq&C\exp\left[-c'\min\left\{\frac{x^2n_{\min}^{*2}}{\tau^4\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|},\frac{xn_{\min}^*}{\tau^2(\|\boldsymbol{\Sigma}_A\|\|\boldsymbol{\Sigma}_B\|)^{1/2}}\right\}\right].
\end{aligned}\tag{37}$$

In the last inequality above, we used a fact that τ is lower bounded by a uniform constant. This is due to Assumption 2.2 that $\mathrm{E}(Z)=0$, $\mathrm{var}(Z)=1$, $\mathrm{E}\exp(tZ)\leq\exp(t^2\tau^2/2)$. Then,

$$\exp(4\tau^2/2)\geq\frac{1}{2}\left\{\mathrm{E}\exp(2Z)+\mathrm{E}\exp(-2Z)\right\}=\sum_{k=0}^{\infty}\frac{2^{2k}\mathrm{E}Z^{2k}}{(2k)!}\geq2\mathrm{E}Z^2=2,$$

which implies $\tau^2\geq\frac{1}{2}\ln(2)$.

3. It is easy to see that $\mathrm{E}\breve{\mathbf{\Sigma}}^* = \mathbf{\Sigma}$, so $\mathrm{E}\mathbf{a}^\top(\breve{\mathbf{\Sigma}}^* - \mathbf{\Sigma})\mathbf{b} = 0$. Then

$$\begin{aligned}
& a^\top(\breve{\mathbf{\Sigma}}^* - \mathbf{\Sigma})b \\
=& \sum_{i,j=1}^{p} a_i b_j \left(\frac{1}{n_{ij}^*}\sum_{k=1}^{n} X_{ik}X_{jk}S_{ik}S_{jk}\right) - \mathrm{E}\sum_{i,j=1}^{p} a_i b_j \left(\frac{1}{n_{ij}^*}\sum_{k=1}^{n} X_{ik}X_{jk}S_{ik}S_{jk}\right) \\
=& \sum_{k=1}^{n}\sum_{i,j=1}^{p}\left(\frac{a_i b_j S_{ik}S_{jk}}{n_{ij}^*}X_{ik}X_{jk} - \mathrm{E}\frac{a_i b_j S_{ik}S_{jk}}{n_{ij}^*}X_{ik}X_{jk}\right) \qquad (38)\\
\triangleq& \sum_{k=1}^{n}\left(\mathbf{X}_k^\top \mathbf{C}^k \mathbf{X}_k - \mathrm{E}\mathbf{X}_k^\top \mathbf{C}^k \mathbf{X}_k\right) \\
=& \sum_{k=1}^{n}\left(\mathbf{Z}_k^\top \mathbf{\Gamma}^\top \mathbf{C}^k \mathbf{\Gamma}\mathbf{Z}_k - \mathrm{E}\mathbf{Z}_k^\top \mathbf{\Gamma}^\top \mathbf{C}^k \mathbf{\Gamma}\mathbf{Z}_k\right).
\end{aligned}$$

Here $\mathbf{C}^k \in \mathbb{R}^{p\times p}$ is a matrix such that $C_{ij}^k = a_i b_j S_{ik}S_{jk}/n_{ij}^*$. Note that $\mathbf{C}^k$ is supported on $A \times B$, we can prove the following properties of $\mathbf{C}^k$.

$$\begin{aligned}
\|\mathbf{\Gamma}^\top \mathbf{C}^k \mathbf{\Gamma}\|_F =& \sqrt{\mathrm{tr}\left(\mathbf{C}^k \mathbf{\Gamma}\mathbf{\Gamma}^\top \mathbf{C}^{k\top}\mathbf{\Gamma}\mathbf{\Gamma}^\top\right)} \\
=& \sqrt{\mathrm{tr}\left(\mathbf{C}^k_{A\times B}\mathbf{\Gamma}_{B\times[q]}\mathbf{\Gamma}^\top_{B\times[q]}\mathbf{C}^{k\top}_{A\times B}\mathbf{\Gamma}_{A\times[q]}\mathbf{\Gamma}^\top_{A\times[q]}\right)} \qquad (39)\\
\le& \|\mathbf{\Gamma}_{B\times[q]}\|\|\mathbf{\Gamma}_{A\times[q]}\|\sqrt{\mathrm{tr}(\mathbf{C}^k\mathbf{C}^{k\top})} \\
\le& (\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|)^{1/2}\sqrt{\mathrm{tr}(\mathbf{C}^k\mathbf{C}^{k\top})};
\end{aligned}$$

$$\begin{aligned}
&\sum_{k=1}^{n}\|\mathbf{\Gamma}^\top \mathbf{C}^k \mathbf{\Gamma}\|_F^2 \le \|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|\|\mathbf{C}^k\|_F^2 = \|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|\sum_{k=1}^{n}\sum_{i,j=1}^{p}\left(\frac{a_i b_j S_{ik}S_{jk}}{n_{ij}^*}\right)^2 \\
=& \|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|\sum_{k=1}^{n}\sum_{i,j=1}^{p}\frac{S_{ik}S_{jk}a_i^2 b_j^2}{n_{ij}^{*2}} = \|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|\sum_{i,j=1}^{p}\frac{a_i^2 b_j^2}{n_{ij}^*} \qquad (40)\\
\le& \|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|\frac{\|\mathbf{a}\|_2^2\|\mathbf{b}\|_2^2}{n_{\min}^*} = \frac{\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|}{n_{\min}^*};
\end{aligned}$$

$$\begin{aligned}
\|\mathbf{\Gamma}^\top \mathbf{C}^k \mathbf{\Gamma}\| \le& \|\mathbf{\Gamma}^\top \mathbf{C}^k \mathbf{\Gamma}\|_F \le (\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|)^{1/2}\sqrt{\mathrm{tr}(\mathbf{C}^k\mathbf{C}^{k\top})} \\
\le& (\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|)^{1/2}\sqrt{\sum_{i,j=1}^{p}\left(\frac{a_i b_j S_{ik}S_{jk}}{n_{ij}^*}\right)^2} \qquad (41)\\
\le& (\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|)^{1/2}\sqrt{\sum_{i,j=1}^{p}\frac{a_i^2 b_j^2}{n_{\min}^{*2}}} \le \frac{\sqrt{\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|}}{n_{\min}^*}.
\end{aligned}$$

Now, note that the last line of (38) can be also equivalently written as

$$\mathrm{vec}(\mathbf{Z})^\top \mathbf{C}^{con} \mathrm{vec}(\mathbf{Z})^\top - \mathrm{E}\mathrm{vec}(\mathbf{Z})^\top \mathbf{C}^{con} \mathrm{vec}(\mathbf{Z})^\top,$$

$$\mathbf{C}^{con} = \begin{bmatrix} \mathbf{\Gamma}^\top \mathbf{C}^1 \mathbf{\Gamma} & & \\ & \ddots & \\ & & \mathbf{\Gamma}^\top \mathbf{C}^n \mathbf{\Gamma} \end{bmatrix} \in \mathbb{R}^{(qn)\times(qn)},$$

where $\mathrm{vec}(\mathbf{Z})$ is the vectorization of $\mathbf{Z}$, which is an qn-dimensional i.i.d. sub-Gaussian vector. Based on the properties of $\mathbf{C}^k$ above, we have

$$\|\mathbf{C}^{con}\|_F^2 = \sum_{k=1}^n \|\mathbf{\Gamma}^\top \mathbf{C}^k \mathbf{\Gamma}\|_F^2 \overset{(40)}{\leq} \frac{\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|}{n^*_{\min}},$$

$$\|\mathbf{C}^{con}\| \leq \max_{1\leq k\leq n} \|\mathbf{\Gamma}^\top \mathbf{C}^k \mathbf{\Gamma}\| \overset{(41)}{\leq} \frac{(\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|)^{1/2}}{n^*_{\min}}.$$

Now applying Hanson-Wright's inequality (Theorem 1.1 in Rudelson and Vershynin [25]), we have

$$\begin{aligned} &\Pr\left\{\left|\mathrm{vec}(\mathbf{Z}_k)^\top \mathbf{C}^{con} \mathrm{vec}(\mathbf{Z}_k) - \mathrm{E}\mathrm{vec}(\mathbf{Z}_k)^\top \mathbf{C}^{con} \mathrm{vec}(\mathbf{Z}_k)\right| \geq x\right\} \\ \leq& 2\exp\left\{-c\min\left(\frac{x^2}{\tau^4\|\mathbf{C}^{con}\|_F^2}, \frac{x}{\tau^2\|\mathbf{C}^{con}\|}\right)\right\} \\ \leq& 2\exp\left\{-cn^*_{\min}\min\left(\frac{x^2}{\tau^4\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|}, \frac{x}{\tau^2(\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|)^{1/2}}\right)\right\}. \end{aligned} \tag{42}$$

Thus,

$$\begin{aligned} &\Pr\left\{\left|\mathbf{a}^\top(\breve{\mathbf{\Sigma}}^* - \mathbf{\Sigma})\mathbf{b}\right| \geq x\right\} \\ \leq& 2\exp\left\{-cn^*_{\min}\min\left(\frac{x^2}{\tau^4\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|}, \frac{x}{\tau^2(\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|)^{1/2}}\right)\right\}. \end{aligned} \tag{43}$$

Here c is a uniform constant. Combining (43) and (37), we have (29).

4. Next, we use the ε-net technique to give the bound on $\|\hat{\mathbf{\Sigma}}^*_{A\times B} - \mathbf{\Sigma}_{A\times B}\|$. Denote $\mathbf{D}^* = \hat{\mathbf{\Sigma}}^*_{A\times B} - \mathbf{\Sigma}_{A\times B}$. Suppose $S^A_{1/3}$ is the (1/3)-net for all unit vectors in $\mathbb{R}^{|A|}$; similarly $S^B_{1/3}$ is the (1/3)-net for all unit vectors in $\mathbb{R}^{|B|}$. Based on the proof of Lemma 3 in Cai et al.

[9], we can let $\mathrm{Card}(S^A_{1/3}) \le 7^k$, $\mathrm{Card}(S^B_{1/3}) \le 7^k$. Since for all $\mathbf{a}, \mathbf{a}_0 \in \mathbb{R}^{|A|}, \mathbf{b}, \mathbf{b}_0 \in \mathbb{R}^{|B|}$,

$$\begin{aligned} \left|\mathbf{a}^\top \mathbf{D}^* \mathbf{b}\right| - \left|\mathbf{a}_0^\top \mathbf{D}^* \mathbf{b}_0\right| \le \left|\mathbf{a}^\top \mathbf{D}^* \mathbf{b} - \mathbf{a}_0^\top \mathbf{D}^* \mathbf{b}_0\right| \le \left|(\mathbf{a}-\mathbf{a}_0)^\top \mathbf{D}^* \mathbf{b}\right| + \left|\mathbf{a}_0^\top \mathbf{D}^* (\mathbf{b}-\mathbf{b}_0)\right| \\ \le (\|\mathbf{a}-\mathbf{a}_0\|_2 + \|\mathbf{b}-\mathbf{b}_0\|_2)\, \|\mathbf{D}^*\|, \end{aligned} \tag{44}$$

we have for all $\mathbf{a} \in \mathbb{R}^{|A|}, \mathbf{b} \in \mathbb{R}^{|B|}, \|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = 1$, we can find $\mathbf{a}_0 \in S^A_{1/3}, \mathbf{b}_0 \in S^B_{1/3}$ such that $\|\mathbf{a}_0 - \mathbf{a}\|_2 \le 1/3, \|\mathbf{b}_0 - \mathbf{b}\|_2 \le 1/3$, then

$$|\mathbf{a}^\top \mathbf{D}^* \mathbf{b}| \le |\mathbf{a}_0^\top \mathbf{D}^* \mathbf{b}_0| + \frac{2}{3}\|\mathbf{D}^*\| \le \sup_{\mathbf{a}_0 \in S^A_{1/3}, \mathbf{b}_0 \in S^B_{1/3}} |\mathbf{a}_0^\top \mathbf{D}^* \mathbf{b}_0| + \frac{2}{3}\|\mathbf{D}^*\|,$$

$$\|\mathbf{D}^*\| = \sup_{\substack{\mathbf{a}\in\mathbb{R}^{|A|}, \mathbf{b}\in\mathbb{R}^{|B|}, \\ \|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = 1}} |\mathbf{a}^\top \mathbf{D}^* \mathbf{b}| \le \sup_{\mathbf{a}_0 \in S^A_{1/3}, \mathbf{b}_0 \in S^B_{1/3}} |\mathbf{a}_0^\top \mathbf{D}^* \mathbf{b}_0| + \frac{2}{3}\|\mathbf{D}^*\|,$$

which yields

$$\|\hat{\mathbf{\Sigma}}^*_{A\times B} - \mathbf{\Sigma}_{A\times B}\| = \|\mathbf{D}^*\| \le 3 \sup_{\mathbf{a}_0 \in S^A_{1/3}, \mathbf{b}_0 \in S^B_{1/3}} |\mathbf{a}_0^\top \mathbf{D}^* \mathbf{b}_0|. \tag{45}$$

Finally, by combining (29) and the inequality above, we know there exist uniform constants $C_1, c > 0$ such that for all $t > 0$,

$$\begin{aligned} &\Pr\left(\|\hat{\mathbf{\Sigma}}^*_{A\times B} - \mathbf{\Sigma}_{A\times B}\| \ge x\right) \le \Pr\left(\sup_{\mathbf{a}_0 \in S^A_{1/3}, \mathbf{b}_0 \in S^B_{1/3}} |\mathbf{a}_0^\top \mathbf{D}^* \mathbf{b}_0| \ge \frac{x}{3}\right) \\ \le & C_1 (7)^{|A|+|B|} \exp\left[-c n^*_{\min} \min\left\{\frac{x^2}{\tau^4 \|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|}, \frac{x}{\tau^2 (\|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|)^{1/2}}\right\}\right]. \end{aligned} \tag{46}$$

Since $|A| + |B| \le 2|A \cup B|$, we have finished the proof of Lemma 2.1. □

Proof of Lemma 6.1. Since $\mathbf{\Sigma}$ is positive semi-definite, we can find the Cholesky decomposition such that $\mathbf{\Sigma} = \mathbf{V}\mathbf{V}^\top$. Then $\mathbf{\Sigma}_{A\times B} = \mathbf{V}_{A\times[p]}\mathbf{V}^\top_{B\times[p]}$ and

$$\begin{aligned} \|\mathbf{\Sigma}_{A\times B}\| &= \max_{\substack{\mathbf{x}\in\mathbb{R}^{|A|}, \mathbf{y}\in\mathbb{R}^{|B|} \\ \|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1}} \mathbf{x}^\top \mathbf{V}_{A\times[p]}\mathbf{V}^\top_{B\times[p]}\mathbf{y} \\ &\le \max_{\substack{\mathbf{x}\in\mathbb{R}^{|A|}, \mathbf{y}\in\mathbb{R}^{|B|} \\ \|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1}} \left(\mathbf{x}^\top \mathbf{V}_{A\times[p]}\mathbf{V}^\top_{A\times[p]}\mathbf{x}\right)^{1/2} \left(\mathbf{y}^\top \mathbf{V}_{B\times[p]}\mathbf{V}^\top_{B\times[p]}\mathbf{y}\right)^{1/2} \\ &= \max_{\substack{\mathbf{x}\in\mathbb{R}^{|A|}, \mathbf{y}\in\mathbb{R}^{|B|} \\ \|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1}} \left(\mathbf{x}^\top \mathbf{\Sigma}_A \mathbf{x}\right)^{1/2} \left(\mathbf{y}^\top \mathbf{\Sigma}_B \mathbf{y}\right)^{1/2} = \|\mathbf{\Sigma}_A\|\|\mathbf{\Sigma}_B\|. \end{aligned}$$

Here we have used the Cauchy-Schwarz inequality. □

6.2 Proof of Theorem 2.1

Define $\mathbf{B} = (b_{ij})_{1\le i,j\le p}$ such that $b_{ij} = \sigma_{ij}$ if $i \in I_s$, $j \in I_{s'}$ and $|s - s'| \le 1$, and 0 otherwise. Let $\mathbf{\Delta} = \mathbf{\Sigma} - \mathbf{B}$. Then

$$\|\hat{\mathbf{\Sigma}}^{\mathrm{bt}} - \mathbf{\Sigma}\| \le \|\hat{\mathbf{\Sigma}}^{\mathrm{bt}} - \mathbf{B}\| + \|\mathbf{\Delta}\|.$$

It is easy to see that

$$\|\mathbf{\Delta}\| \le \|\mathbf{\Delta}\|_{\ell_1} \le \max_i \sum_{j:|i-j|\ge k} |\sigma_{ij}| \le Mk^{-\alpha}.$$

To bound $\|\hat{\mathbf{\Sigma}}^{\mathrm{bt}} - \mathbf{B}\|$, note that

$$\|\hat{\mathbf{\Sigma}}^{\mathrm{bt}} - \mathbf{B}\| = \sup_{u\in\mathbb{R}^p:\|u\|_2=1} \left|\langle u, (\hat{\mathbf{\Sigma}}^{\mathrm{bt}} - \mathbf{B})u\rangle\right|.$$

For any $u \in \mathbb{R}^p$, $\|u\|_2 = 1$, we have

$$\begin{aligned}\left|\langle u, (\hat{\mathbf{\Sigma}}^{\mathrm{bt}} - \mathbf{B})u\rangle\right| &\le \sum_{s,s':|s-s'|\le 1} \left|\left\langle u_{I_s}, (\hat{\mathbf{\Sigma}}^*_{I_s\times I_{s'}} - \mathbf{\Sigma}_{I_s\times I_{s'}})u_{I_{s'}}\right\rangle\right| \\ &\le \sum_{s,s':|s-s'|\le 1} \|u_{I_s}\|_2\|u_{I_{s'}}\|_2\|\hat{\mathbf{\Sigma}}^*_{I_s\times I_{s'}} - \mathbf{\Sigma}_{I_s\times I_{s'}}\| \\ &\le \left(\sum_{s,s':|s-s'|\le 1} \|u_{I_s}\|_2\|u_{I_{s'}}\|_2\right)\left(\max_{|s-s'|\le 1} \|\hat{\mathbf{\Sigma}}^*_{I_s\times I_{s'}} - \mathbf{\Sigma}_{I_s\times I_{s'}}\|\right).\end{aligned}$$

The Cauchy-Schwarz inequality yields

$$\sum_{s,s':|s-s'|\le 1} \|u_{I_s}\|_2\|u_{I_{s'}}\|_2 \le \frac{1}{2}\sum_{s,s':|s-s'|\le 1}\left(\|u_{I_s}\|_2^2 + \|u_{I_{s'}}\|_2^2\right) \le 3\sum_{s=1}^N \|u_{I_s}\|_2^2 = 3. \tag{47}$$

Therefore,

$$\|\hat{\mathbf{\Sigma}}^{\mathrm{bt}} - \mathbf{\Sigma}\| \le \|\hat{\mathbf{\Sigma}}^* - \mathbf{B}\| + \|\mathbf{\Delta}\| \le 3\max_{|s-s'|\le 1}\left\|\hat{\mathbf{\Sigma}}^*_{I_s\times I_{s'}} - \mathbf{\Sigma}_{I_s\times I_{s'}}\right\| + Mk^{-\alpha},$$

which yields

$$\mathrm{E}\|\hat{\mathbf{\Sigma}}^{\mathrm{bt}} - \mathbf{\Sigma}\|^2 \le 18\mathrm{E}\left(\max_{|s-s'|\le 1}\left\|\hat{\mathbf{\Sigma}}^*_{I_s\times I_{s'}} - \mathbf{\Sigma}_{I_s\times I_{s'}}\right\|\right)^2 + 2M^2k^{-2\alpha}.$$

According to lemma 2.1, there exists constant $C, c > 0$ which only depend on τ such that for all $x > 0$,

$$\Pr\left(\max_{|s-s'|\le 1}\|\hat{\mathbf{\Sigma}}_{I_s\times I_{s'}} - \mathbf{\Sigma}_{I_s\times I_{s'}}\| \ge x\right) \le C\lceil\frac{p}{k}\rceil(49)^k \exp\left\{-cn^*_{\min}\min\left(\frac{x^2}{\|\mathbf{\Sigma}\|^2}, \frac{x}{\|\mathbf{\Sigma}\|}\right)\right\}. \quad (48)$$

Now we set $t = C'(k + \ln p)/n^*_{\min}$ for C' large enough. The spectral norm risk satisfies

$$\begin{aligned}
\mathrm{E}\|\hat{\mathbf{\Sigma}}^{\mathrm{bt}} - \mathbf{\Sigma}\|^2 \le & 18\mathrm{E}\max_{|s-s'|\le 1}\left\|\hat{\mathbf{\Sigma}}^*_{I_s\times I_{s'}}\right\| + 2M^2k^{-2\alpha} \\
\le & 18\int_0^\infty \Pr\left(\max_{|s-s'|\le 1}\|\hat{\mathbf{\Sigma}}_{I_s\times I_{s'}} - \mathbf{\Sigma}_{I_s\times I_{s'}}\|^2 \ge x\right)dx + 2M^2k^{-2\alpha} \\
\le & 18t + 18\int_t^\infty \Pr\left(\max_{|s-s'|\le 1}\|\hat{\mathbf{\Sigma}}_{I_s\times I_{s'}} - \mathbf{\Sigma}_{I_s\times I_{s'}}\|^2 \ge x\right)dx + 2M^2k^{-2\alpha} \quad (49) \\
\le & 18t + C\lceil\frac{p}{k}\rceil(49)^k\int_t^\infty \exp\left\{-c'n^*_{\min}\min\left(x, x^{\frac{1}{2}}\right)\right\}dx + 2M^2k^{-2\alpha} \\
\le & 18t + C\lceil\frac{p}{k}\rceil(49)^k\frac{1}{n^*_{\min}}\exp\left(-c'n^*_{\min}t\right) + 2M^2k^{-2\alpha},
\end{aligned}$$

then (49) yields

$$\mathrm{E}\|\hat{\mathbf{\Sigma}}^{\mathrm{bt}} - \mathbf{\Sigma}\|^2 \le C\left(\frac{k + \ln p}{n^*_{\min}} + k^{-2\alpha}\right), \quad (50)$$

where C only depends on τ, M, M_0. We can finally finish the proof of Theorem 2.1 by taking $k = (n^*_{\min})^{1/(2\alpha+1)}$. □

Acknowledgments

We thank Tianxi Cai for the ovarian cancer data set and for helpful discussions. We also thank the Editor, the Associate editor, one referee and Zoe Russek for useful comments which have helped to improve the presentation of the paper.

References

[1] B. Andreopoulos and D. Anastassiou. Integrated analysis reveals hsa-mir-142 as a representative of a lymphocyte-specific gene expression and methylation signature. *Cancer Informatics*, 11:61–75, 2012.

[2] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36:199–227, 2008.

[3] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36:2577–2604, 2008.

[4] T. Bonome, J.-Y. Lee, D.-C. Park, M. Radonovich, C. Pise-Masison, J. Brady, G. J. Gardner, K. Hao, W. H. Wong, J. C. Barrett, and et al. Expression profiling of serous low malignant potential, low-grade, and high-grade tumors of the ovary. *Cancer Research*, 65:10602–10612, 2005.

[5] T. T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.*, 106:672–684, 2011.

[6] T. T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Rel.*, 161:781–815, 2015.

[7] T. T. Cai, Z. Ren, and H. H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.*, 10:1–59, 2016.

[8] T. T. Cai and M. Yuan. Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.*, 40:2014–2042, 2012.

[9] T. T. Cai, C.-H. Zhang, and H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38:2118–2144, 2010.

[10] T. T. Cai and H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.*, 40:2389–2420, 2012.

[11] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474:609–615, 2011.

[12] G. Carraro, A. Shrestha, J. Rostkovius, A. Contreras, C.-M. Chao, E. El Agha, B. MacKenzie, S. Dilai, D. Guidolin, M. M. Taketo, et al. mir-142-3p balances proliferation and differentiation of mesenchymal cells during lung development. *Development*, 141(6):1272–1281, 2014.

[13] S. X. Chen, L. X. Zhang, and P. S. Zhong. Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.*, 105:810–819, 2010.

[14] J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, 26:1205–1210, 2010.

[15] H. K. Dressman, A. Berchuck, G. Chan, J. Zhai, A. Bild, R. Sayer, J. Cragun, J. Clarke, R. S. Whitaker, and e. a. Li, L. An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *J. Clin. Oncol.*, 25:517–525, 2007.

[16] N. El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.*, 36:2717–2756, 2008.

[17] R. D. Hawkins, G. C. Hon, and B. Ren. Next-generation genomics: an integrative approach. *Nat. Rev. Genet.*, 11:476–486, 2010.

[18] J. G. Ibrahim and G. Molenberghs. Missing data methods in longitudinal studies: a review. *Test*, 18:1–43, 2009.

[19] S. Y. Ko, N. Barengo, A. Ladanyi, J. S. Lee, F. Marini, E. Lengyel, and H. Naora. Hoxa9 promotes ovarian cancer growth by stimulating cancer-associated fibroblasts. *J. Clin. Invest.*, 122:3603–3617, 2012.

[20] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* 2nd Edition. John Wiley & Sons, New York, 2002.

[21] P.-L. Loh and M. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Ann. Statist.*, 40:1637–1664, 2012.

[22] K. Lounici. Sparse principal component analysis with missing observations. *High dimensional probability VI*, 66 of Prog. Proba., Institute of Mathematical Statistics (IMS) Collections:327–356, 2013.

[23] K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, to appear, 2014.

[24] A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.*, 104:177–186, 2009.

[25] M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18:1–9, 2013.

[26] J. L. Schafer. *Analysis of Incomplete Multivariate Data.* CRC press, 2010.

[27] R. W. Tothill, A. V. Tinker, J. George, R. Brown, S. B. Fox, S. Lade, D. S. Johnson, M. K. Trivett, D. Etemadmoghadam, and B. e. a. Locandro. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.*, 14:5198–5208, 2008.

Appendix: Proofs

In this appendix we collect the proofs for the main results of the sparse covariance matrix estimation and Propositions 2.1 and 3.1.

Proof of Lemma 3.1

The main strategy for the proof of this lemma is similar to that for Lemma 2 in Cai and Liu [5]. Without loss of generality, we can translate X and assume that $\mathrm{E}\mathbf{X} = \boldsymbol{\mu} = 0$. First, we show the following property on θ_{ij}

$$c\sigma_{ii}\sigma_{jj} \leq \theta_{ij} \leq C\sigma_{ii}\sigma_{jj}. \tag{51}$$

Here $c, C > 0$ only depend on the distribution of $\mathbf{Z}$. Denote $\mathbf{a}, \mathbf{b}$ as the i-th and j-th row vector of $\boldsymbol{\Gamma}$, then $\|\mathbf{a}\|_2^2 = \mathrm{var}(X_i) = \sigma_{ii}$, $\|b\|_2^2 = \sigma_{jj}$. Recall that

$$\theta_{ij} = \mathrm{var}(X_iX_j - \sigma_{ij}) = \mathrm{var}(\mathbf{a}^\top\mathbf{Z}\mathbf{b}^\top\mathbf{Z} - \mathrm{E}\mathbf{a}^\top\mathbf{Z}\mathbf{b}^\top\mathbf{Z}),$$

thus

$$\theta_{ij} = \mathrm{var}\left(\mathbf{Z}^\top\mathbf{a}\mathbf{b}^\top\mathbf{Z}\right) \leq \mathrm{E}\left(\mathbf{Z}^\top\mathbf{a}\mathbf{b}^\top\mathbf{Z}\right)^2 \leq \sqrt{\mathrm{E}(\mathbf{Z}^\top\mathbf{a})^4\mathrm{E}(\mathbf{Z}^\top\mathbf{b})^4}. \tag{52}$$

Since

$$\mathrm{E}\left(\mathbf{Z}^\top\mathbf{a}\right)^4 \leq \left(\sum_{s=1}^q a_s^4\mathrm{E}Z_s^4 + 6\sum_{1\leq s<t\leq q} a_s^2a_t^2\mathrm{E}Z_s^2\mathrm{E}Z_t^2\right) \leq 3\|\mathbf{a}\|_2^4\mathrm{E}Z^4 \leq C_\tau\sigma_{ii}^2,$$

similarly $\mathrm{E}\left(\mathbf{Z}^\top\mathbf{b}\right)^4 \leq C_\tau\sigma_{jj}^2$, we know $\theta_{ij} \leq C_\tau\sigma_{ii}\sigma_{jj}$ for some constant C_τ only depending on τ.

On the other hand, since the entries of $\mathbf{Z}$ are i.i.d. with mean 0 and variance 1 and

$\text{var}(Z^2) > 0$, we know $\text{E}Z^4 > (\text{E}Z^2)^2 = 1$. We can calculate that

$$\begin{aligned}\theta_{ij} =&\text{var}(X_iX_j) = \text{E}(X_iX_j)^2 - (\text{E}X_iX_j)^2\\ =&\text{E}\left\{\sum_{s=1}^{q} a_sb_sZ_s^2 + \sum_{1\le s<t\le q}(a_sb_t + a_tb_s)Z_sZ_t\right\}^2 - \left(\sum_{s=1}^{q} a_sb_s\right)^2\\ =&\sum_{s=1}^{q} a_s^2b_s^2\text{E}Z_s^4 + \sum_{1\le s<t\le q}\left(a_s^2b_t^2 + a_t^2b_s^2 + 4a_sb_sa_tb_t\right)\text{E}Z_s^2Z_t^2 - \left(\sum_{s=1}^{q} a_sb_s\right)^2\\ =&\sum_{s=1}^{q} a_s^2b_s^2(\text{E}Z^4 - 3) + \left(\sum_{s=1}^{q} a_s^2\right)\left(\sum_{t=1}^{q} b_t^2\right) + \left(\sum_{s=1}^{q} a_sb_s\right)^2.\end{aligned}$$

When $\text{E}Z^4 \ge 3$, it is clear that

$$\theta_{ij} \ge \left(\sum_{s=1}^{q} a_s^2\right)\left(\sum_{s=1}^{q} b_s^2\right) = \|\mathbf{a}\|_2^2\|\mathbf{b}\|_2^2 = \sigma_{ii}\sigma_{jj};$$

when $\text{E}Z^4 < 3$, note $\xi = \text{E}Z^4$, $x = \sum_{s:a_sb_s\ge 0} a_sb_s$, $y = -\sum_{s:a_sb_s<0} a_sb_s$, then $x, y \ge 0$ and

$$\begin{aligned}\theta_{ij} \ge& - (3-\xi)(x^2+y^2) + (x-y)^2 + \frac{3-\xi}{2}\left(\sum_{s=1}^{q} a_s^2\right)\left(\sum_{s=1}^{q} b_s^2\right) + \frac{\xi-1}{2}\left(\sum_{s=1}^{q} a_s^2\right)\left(\sum_{s=1}^{q} b_s^2\right)\\ \ge& - (3-\xi)(x^2+y^2) + (x-y)^2 + \frac{3-\xi}{2}\left(\sum_{s=1}^{q} |a_sb_s|\right)^2 + \frac{\xi-1}{2}\|\mathbf{a}\|_2^2\|\mathbf{b}\|_2^2\\ =& - (3-\xi)(x^2+y^2) + (x-y)^2 + \frac{3-\xi}{2}(x+y)^2 + \frac{\xi-1}{2}\|\mathbf{a}\|_2^2\|\mathbf{b}\|_2^2\\ \ge&\frac{\xi-1}{2}(x-y)^2 + \frac{\xi-1}{2}\sigma_{ii}\sigma_{jj} \ge c\sigma_{ii}\sigma_{jj}.\end{aligned}$$

Here $c = (\xi - 1)/2$ only depends on the distribution of Z.

Now we further normalize each row of $\mathbf{\Gamma}$ such that $\|\mathbf{\Gamma}_{i\cdot}\|_2 = 1$ $\text{var}(X_i) = \text{var}(\mathbf{\Gamma}_i\mathbf{Z}) = 1$ for $1 \le i \le p$. The rest of the proof is essentially the same as Lemma 2 in Cai and Liu [5] thus we will not go into details. Let

$$\tilde{\theta}_{ij}^* = \frac{1}{n_{ij}^*}\sum_{k=1}^{n}\left(X_{ik}X_{jk} - \tilde{\sigma}_{ij}^*\right)^2 S_{ik}S_{jk}, \quad \tilde{\sigma}_{ij}^* = \frac{1}{n_{ij}^*}\sum_{k=1}^{n} X_{ik}X_{jk}S_{ik}S_{jk}. \tag{53}$$

We would like to show

$$\Pr\left(\max_{ij}|\hat{\theta}_{ij}^* - \tilde{\theta}_{ij}^*| \ge C_1\sqrt{\ln p/n_{ij}^*}\right) = O(p^{-M}). \tag{54}$$

Denote $X_i^{(j)*}$ as the average of X_i's for those samples X_i, X_k are both observed, i.e. $X_i^{(j)*} = \sum_{k=1}^n S_{ik}S_{jk}X_{ik}/n_{ij}^*$. Then,

$$\begin{aligned}
\hat{\theta}_{ij}^* =& \frac{1}{n_{ij}^*}\sum_{k=1}^n S_{ik}S_{jk}\left(X_{ik}X_{jk} - \bar{X}_i^* X_{jk} - \bar{X}_j^* X_{ki} - \tilde{\sigma}_{ij}^* + \bar{X}_i^{(j)*}\bar{X}_j + \bar{X}_j^{(i)*}\bar{X}_i\right)^2 \\
=& \tilde{\theta}_{ij}^* + \frac{2}{n_{ij}^*}\sum_{k=1}^n S_{ik}S_{jk}\left(X_{ik}X_{jk} - \tilde{\sigma}_{ij}^*\right)\left(\bar{X}_i^{(j)*}\bar{X}_j^* + \bar{X}_j^{(i)*}\bar{X}_i^* - \bar{X}_i^* X_{jk} - \bar{X}_j^* X_{ik}\right) \qquad (55) \\
&+ \frac{1}{n_{ij}^*}\sum_{k=1}^n S_{ik}S_{jk}\left(\bar{X}_i^{(j)*}\bar{X}_i^* + \bar{X}_j^{(i)*}\bar{X}_i^* - \bar{X}_i^* X_{jk} - \bar{X}_j^* X_{ik}\right)^2.
\end{aligned}$$

Similarly to Lemma 2 in Cai and Liu [5], we could have

$$\Pr\left(\max_{i,j}|\bar{X}_i^{(j)*}| \geq C_2\sqrt{\frac{\ln p}{n_{ij}^*}}\right) = O\left(p^{-M}\right), \tag{56}$$

$$\Pr\left(\max_{ij}\left|\frac{1}{n_{ij}^*}\sum_{k=1}^n S_{ik}S_{jk}X_{ik}^2X_{jk}\bar{X}_j^*\right| \geq C_5\sqrt{\frac{\ln p}{n_{ij}^*}}\right). \tag{57}$$

and similar bounds for the other terms in the right hand side of (55). Hence we have proved (54). By (51), we can directly get

$$\Pr\left(\left|\tilde{\theta}_{ij}^* - \theta_{ij}\right| \geq \varepsilon\right) = O(p^{-M}), \tag{58}$$

by applying the result in Lemma 2 in Cai and Liu [5] on the samples $\mathbf{X}_k, k \in \{k : S_{ik} = S_{jk} = 1\}$. Combining (54) and (58), we can proved (21).

The proof of (20) is omitted here because it is essentially the same as that of Lemma 2 in [5]. □

Proof of Theorem 3.1

First without loss of generality, we can assume that $\boldsymbol{\mu} = \mathrm{E}\mathbf{X}_k = 0$. Based on Assumption 2.2, we have for each k, $\mathbf{X}_k = \boldsymbol{\Gamma}\mathbf{Z}_k$, where $\mathbf{Z}_k$ is an i.i.d. sub-Gaussian random vector. Based on the proof of Lemma 3.1, we know $c\sigma_{ii}\sigma_{jj} \leq \theta_{ij} \leq C\sigma_{ii}\sigma_{jj}$, where c, C are constants which only depend on the distribution of $\mathbf{Z}$. We will prove Theorem 3.1 in several steps.

1. For $\varepsilon > 0$, we first consider the loss under the event that

$$Q = \left\{ |\hat{\sigma}_{ij}^* - \sigma| / \hat{\theta}_{ij}^* \leq \delta \sqrt{\ln p / n_{ij}^*}, \forall 1 \leq i, j \leq p, \quad \text{and} \quad \max_{ij} |\hat{\theta}_{ij}^* - \theta_{ij}| / (\sigma_{ii}\sigma_{jj}) \leq \varepsilon. \right\}. \tag{59}$$

Since $|\hat{\sigma}_{ij}^* - \sigma_{ij}| \leq \delta \sqrt{\hat{\theta}_{ij}^* \ln p / n_{ij}^*} = \lambda_{ij}$, by Condition (1) of $T_{\lambda_{ij}}$, we have

$$|T_{\lambda_{ij}}(\hat{\sigma}_{ij}^*) - \sigma_{ij}| \leq c_T |\sigma_{ij}|.$$

Besides, by condition (3) of $T_{\lambda_{ij}}$,

$$|T_{\lambda_{ij}}(\hat{\sigma}_{ij}^*) - \sigma_{ij}| \leq |T_{\lambda_{ij}}(\hat{\sigma}_{ij}^*) - \hat{\sigma}_{ij}^*| + |\hat{\sigma}_{ij}^* - \sigma_{ij}| \overset{(59)}{\leq} \lambda_{ij} + \delta \sqrt{\frac{\hat{\theta}_{ij}^* \ln p}{n_{ij}^*}}$$
$$\leq 2\delta \sqrt{\frac{\hat{\theta}_{ij}^* \ln p}{n_{ij}^*}} \leq 2\delta \sqrt{\frac{(\theta_{ij} + \varepsilon \sigma_{ii}\sigma_{jj}) \ln p}{n_{ij}^*}} \leq C \sqrt{\frac{\sigma_{ii}\sigma_{jj} \ln p}{n_{ij}^*}}.$$

Since $n_{\min}^* \leq n_{ij}^* \leq n$, thus

$$\|\hat{\mathbf{\Sigma}}^{\text{at}} - \mathbf{\Sigma}\|_{\ell_1} \leq \max_i \sum_{j=1}^p \left| T_{\lambda_{ij}}(\hat{\mathbf{\Sigma}}_{ij}^*) - \sigma_{ij} \right| \leq \max_i \sum_{j=1}^p C \min \left\{ |\sigma_{ij}|, \sqrt{\frac{\sigma_{ii}\sigma_{jj} \ln p}{n_{ij}^*}} \right\}$$
$$\overset{(17)}{\leq} C c_{n,p} \sqrt{\frac{\ln p}{n_{\min}^*}}.$$

Since $\hat{\mathbf{\Sigma}}^{\text{at}} - \mathbf{\Sigma}$ is a symmetric matrix, we have

$$\|\hat{\mathbf{\Sigma}}^{\text{at}} - \mathbf{\Sigma}\|_{\ell_q} \leq \|\hat{\mathbf{\Sigma}}^{\text{at}} - \mathbf{\Sigma}\|_{\ell_1} \leq C c_{n,p} \sqrt{\frac{\ln p}{n_{\min}^*}}$$

for all $1 \leq q \leq \infty$.

By Lemma 3.1, we know (59) happens with probability at least $1 - O\left\{(\ln p)^{-1/2} p^{-\delta+2}\right\}$, which implies (18).

2. Next we consider (19). We apply Lemma 2.1 by restricting $\mathbf{\Sigma}$ on $\{i,j\} \times \{i,j\}$ and set $A = \{i\}$, $B = \{j\}$, then there exists $C_1, c_1 > 0$ such that

$$\Pr\left(|\hat{\sigma}_{ij}^* - \sigma_{ij}| \leq x\right) \geq 1 - C_1 \exp\left[-c_1 n_{ij}^* \min\left\{\frac{x^2}{\sigma_{ii}\sigma_{jj}}, \frac{x}{(\sigma_{ii}\sigma_{jj})^{1/2}}\right\}\right] \tag{60}$$

holds for all $x > 0$. Therefore,

$$\Pr\left\{|\hat{\sigma}^*_{ij} - \sigma_{ij}| \leq x(\sigma_{ii}\sigma_{jj})^{1/2}, \forall 1 \leq i, j \leq p\right\} \geq 1 - C_1 p^2 \exp\left\{-c_1 n^*_{\min} \min(x, x^2)\right\}.$$

We also have

$$|T_{\lambda_{ij}}(\hat{\sigma}^*_{ij}) - \sigma_{ij}| \leq c_T|\hat{\sigma}^*_{ij}| + |\sigma_{ij}| \leq (1 + c_T)|\sigma_{ij}| + c_T|\hat{\sigma}^*_{ij} - \sigma_{ij}|.$$

Thus,

$$\begin{aligned}
&\mathrm{E}\|\hat{\boldsymbol{\Sigma}}^{\mathrm{at}} - \boldsymbol{\Sigma}\|^2_{\ell_1} = \int_Q \|\hat{\boldsymbol{\Sigma}}^{\mathrm{at}} - \boldsymbol{\Sigma}\|^2_{\ell_1} dP + \int_{Q^c} \|\hat{\boldsymbol{\Sigma}}^{\mathrm{at}} - \boldsymbol{\Sigma}\|^2_{\ell_1} dP \\
\leq & C c^2_{n,p} \frac{\ln p}{n^*_{\min}} + \int_{Q^c} \left(\max_i \sum_{j=1}^p |T_{\lambda_{ij}}(\hat{\sigma}_{ij}) - \sigma_{ij}|\right)^2 dP \qquad (61) \\
\leq & C c^2_{n,p} \frac{\ln p}{n^*_{\min}} + C\int_{Q^c} \left(\max_i \sum_{j=1}^p |\sigma_{ij}|\right)^2 dP + C\int_{Q^c} \left(\max_i \sum_{j=1}^p |\hat{\sigma}^*_{ij} - \sigma_{ij}|\right)^2 dP.
\end{aligned}$$

For the second term above, we have

$$\begin{aligned}
C\int_{Q^c} \left(\max_i \sum_{j=1}^p |\sigma_{ij}|\right)^2 dP \leq & C\int_{Q^c} \left[\max_i \sum_{j=1}^p \min\left\{(\sigma_{ii}\sigma_{jj})^{1/2}, \frac{|\sigma_{ij}|}{\sqrt{\ln p/n^*_{\min}}}\right\}\right]^2 dP \\
\leq & C\Pr(Q^c) c^2_{n,p} \leq C c^2_{n,p} p^{-\delta+2}(\ln p)^{-1/2}.
\end{aligned}$$

Based on the assumption that $p \geq (n^*_{\min})^{\xi}$, $\delta \geq 4 + 1/\xi$, we have

$$C\int_{Q^c} \left(\max_i \sum_{j=1}^p |\sigma_{ij}|\right)^2 \leq C c^2_{n,p} \frac{\ln p}{n^*_{\min}}. \qquad (62)$$

We denote $K = \max_i \sigma_{ii}$, then $K \le c_{n,p}$. For the third term above in (61), we have

$$\begin{aligned}
& C\int_{Q^c}\left(\max_i \sum_{j=1}^p |\hat{\sigma}^*_{ij} - \sigma_{ij}|\right)^2 dP \le Cp^2 \int_{Q^c}\left(\max_{ij} |\hat{\sigma}^*_{ij} - \sigma_{ij}|\right)^2 dP \\
\le & Cp^2 \int_0^\infty x \Pr\left(\{\max_{ij} |\hat{\sigma}^*_{ij} - \sigma_{ij}| \ge x\} \cap Q^c\right) dx \\
= & Cp^2 \int_0^K x \Pr\left(\{\max_{ij} |\hat{\sigma}^*_{ij} - \sigma_{ij}| \ge x\} \cap Q^c\right) dx \\
& + Cp^2 \int_K^\infty x \Pr\left(\{\max_{ij} |\hat{\sigma}^*_{ij} - \sigma_{ij}| \ge x\} \cap Q^c\right) dx \\
\le & Cp^2 K^2 \Pr(Q^c) + Cp^2 \int_K^\infty x \exp\left\{-c_1 n^*_{\min} \min\left(\frac{x^2}{\max_i \sigma^2_{ii}}, \frac{x}{\max_i \sigma_{ii}}\right)\right\} dx \\
\le & Cp^2 c^2_{n,p} p^{-\delta+2} (\ln p)^{-1/2} + Cp^2 \int_K^\infty x \exp(-c_1 n^*_{\min} x/K) dx \\
\le & Cp^{-\delta+4} (\ln p)^{-1/2} c^2_{n,p} + Cp^2 c^2_{n,p} \exp(-c n^*_{\min}).
\end{aligned}$$

Based on the assumption $\ln p = o((n^*_{\min})^{1/3})$ and $p \ge (n^*_{\min})^\xi$, $\delta \ge 4 + 1/\xi$, we have

$$C\int_{Q^c}\left(\max_i \sum_{j=1}^p |\hat{\sigma}^*_{ij} - \sigma_{ij}|\right)^2 dP \le C c^2_{n,p} \frac{\ln p}{n^*_{\min}}. \tag{63}$$

Combining (62), (63) and (61), we have finished the proof of (19) under the additional assumption that $p \ge (n^*_{\min})^\xi$, $\delta \ge 4 + 1/\xi$. □

Proof of Theorem 3.2.

By Lemma 3.1, we know

$$\Pr\left(|\hat{\sigma}^*_{ij} - \sigma_{ij}| \ge 2\sqrt{\frac{\ln p \hat{\theta}^*_{ij}}{n^*_{ij}}}, \exists 1 \le i, j \le p\right) = O\left\{(\ln p)^{-1/2}\right\}; \tag{64}$$

for all $\varepsilon > 0$,

$$\Pr\left(|\hat{\theta}^*_{ij} - \theta_{ij}| / \sqrt{\sigma_{ii}\sigma_{jj}} \ge \varepsilon, \exists 1 \le i, j \le p\right) = O(p^{-M}). \tag{65}$$

Also by the proof of Lemma 3.1, there exists $c > 0$ such that

$$\theta_{ij} / \sqrt{\sigma_{ij}\sigma_{ij}} > c. \tag{66}$$

When $\delta = 2$, the thresholding level is

$$\lambda_{ij} = 2\sqrt{\frac{\hat{\theta}^*_{ij} \ln p}{n^*_{ij}}}. \tag{67}$$

Therefore,

$$\begin{aligned}
&\Pr\left\{\operatorname{supp}(\hat{\mathbf{\Sigma}}^{\text{at}}) \neq \operatorname{supp}(\mathbf{\Sigma})\right\} \\
=&\Pr\left\{T_{\lambda_{ij}}(\hat{\sigma}^*_{ij}) = 0, \exists (i,j) \in \operatorname{supp}(\mathbf{\Sigma})\right\} + \Pr\left\{T_{\lambda_{ij}}(\hat{\sigma}^*_{ij}) \neq 0, \exists (i,j) \notin \operatorname{supp}(\mathbf{\Sigma})\right\} \\
\leq&\Pr\left\{|\hat{\sigma}^*_{ij}| \leq \lambda_{ij}, \exists (i,j) \in \operatorname{supp}(\mathbf{\Sigma})\right\} + \Pr\left\{|\hat{\sigma}^*_{ij}| > \lambda_{ij}, \exists (i,j) \notin \operatorname{supp}(\mathbf{\Sigma})\right\} \\
\overset{(67)}{\leq}& \Pr\left(|\hat{\sigma}^*_{ij} - \sigma_{ij}| \geq 2\sqrt{\frac{\hat{\theta}^*_{ij} \ln p}{n^*_{ij}}}, \exists 1 \leq i,j \leq p\right) \\
&+ \Pr\left\{|\sigma_{ij}| \leq 4\sqrt{\frac{\hat{\theta}^*_{ij} \ln p}{n^*_{ij}}}, \exists (i,j) \in \operatorname{supp}(\mathbf{\Sigma})\right\} \\
\overset{(64)}{\leq}& O\left\{(\ln p)^{-1/2}\right\} + \Pr\left\{(4+\gamma)\sqrt{\frac{\theta_{ij} \ln p}{n^*_{ij}}} \leq 4\sqrt{\frac{\hat{\theta}^*_{ij}}{n^*_{ij}}}, \exists (i,j) \in \operatorname{supp}(\mathbf{\Sigma})\right\} \\
\leq& O\left\{(\ln p)^{-1/2}\right\} + \Pr\left\{\left(\frac{4+\gamma}{4}\right)^2 - 1 \leq \frac{\hat{\theta}^*_{ij} - \theta_{ij}}{\theta_{ij}}, \exists (i,j) \in \operatorname{supp}(\mathbf{\Sigma})\right\} \\
\overset{(65)(66)}{=}& o(1),
\end{aligned}$$

which means $\Pr\left\{\operatorname{supp}(\hat{\mathbf{\Sigma}}^{\text{at}}) \neq \operatorname{supp}(\mathbf{\Sigma})\right\} = o(1)$. □

Proof of Propositions 2.1 and 3.1.

For given $n_0 \geq 1$, we again consider a special pattern of missingness $\mathbf{S}_0$:

$$(\mathbf{S}_0)_{ij} = \begin{cases} 1, & 1 \leq i \leq n_0, 1 \leq j \leq p \\ 0, & n_0 + 1 \leq i \leq n, 1 \leq j \leq p. \end{cases}$$

Under this missingness pattern, $n^*_{\min} = n_0$, and the problem essentially becomes complete data problem with n_0 samples. Now, Propositions 2.1, 3.1 directly follow Theorem 3 of Cai et al. [9] and Theorem 2 of Cai and Zhou [10], respectively.