Support Recovery in Sparse PCA with Incomplete Data

Hanbyul Lee
Department of Statistics
Purdue University
West Lafayette, IN 47906
lee3078@purdue.edu

Qifan Song
Department of Statistics
Purdue University
West Lafayette, IN 47906
qfsong@purdue.edu

Jean Honorio
Department of Computer Science
Purdue University
West Lafayette, IN 47906
jhonorio@purdue.edu

Abstract

We study a practical algorithm for sparse principal component analysis (PCA) of incomplete and noisy data. Our algorithm is based on the semidefinite program (SDP) relaxation of the non-convex l_1-regularized PCA problem. We provide theoretical and experimental evidence that SDP enables us to exactly recover the true support of the sparse leading eigenvector of the unknown true matrix, despite only observing an incomplete (missing uniformly at random) and noisy version of it. We derive sufficient conditions for exact recovery, which involve matrix incoherence, the spectral gap between the largest and second-largest eigenvalues, the observation probability and the noise variance. We validate our theoretical results with incomplete synthetic data, and show encouraging and meaningful results on a gene expression dataset.

1 Introduction

Principal component analysis (PCA) is one of the most popular methods to reduce data dimension which is widely used in various applications including genetics, image processing, engineering, and many others. However, standard PCA is usually not preferred when principal components depend on only a small number of variables, because it provides dense vectors as a solution which degrades interpretability of the result. This can be worse especially in the high-dimensional setting where the solution of standard PCA is inconsistent as addressed in several works [Paul, 2007, Nadler, 2008, Johnstone and Lu, 2009]. To solve the inconsistency issue and improve interpretability, *sparse PCA* has been proposed, which enforces sparsity in the PCA solution so that dimension reduction and variable selection can be simultaneously performed. Theoretical and algorithmic researches on sparse PCA have been actively conducted over the past few years [Zou et al., 2006, Amini and Wainwright, 2008, Journée et al., 2010, Ma, 2013, Lei and Vu, 2015, Berk and Bertsimas, 2019, Richtárik et al., 2021].

In this paper, we consider a special situation where the data to which sparse PCA is applied are not completely observed, but partially missing. Missing data frequently occurs in a wide range of machine learning problems, where sparse PCA is no exception. There are various reasons and situations where data becomes incomplete, such as failures of hardware, high expenses of sampling, and preserving privacy. One concrete example is the analysis of single-cell RNA sequence (scRNA-seq) data [Park and Zhao, 2019], where the cells are divided into several distinct types which can be characterized with only a small number of genes among tens of thousands of genes. Sparse PCA can be effectively utilized here to reduce the dimension (from numerous cells to a few cell types) and to select a small number of genes that affect the reduced data. However, since scRNA-seq data usually have many missing values due to technical and sampling issues, the existing sparse PCA theory and method designed for fully observed data cannot be directly applied, and new methodology and theory are in demand.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Despite the need for theoretical research and algorithmic development of sparse PCA for incomplete data, there have not been many studies yet. Lounici [2013] and Kundu et al. [2015] considered two different optimization objectives for sparse PCA on incomplete data, which impose l_1 regularization and l_0 constraint on the classic PCA loss function using a (bias-corrected) incomplete matrix, respectively. It was shown that the solution of each problem has a non-trivial error bound under certain conditions, but the optimization problems they considered are either nonconvex or NP-hard, and thus theoretical studies of computational feasible algorithms are still lacking. More recently, Park and Zhao [2019] proposed a computationally tractable two-step algorithm based on matrix factorization and completion, but its first step is an iterative algorithm that requires singular value decomposition in every iteration, which incurs a lot of cost in memory and time under a high-dimensional setting.

With this motivation, we suggest a computational friendly convex optimization problem via a semidefinite relaxation of the l_1 regularized PCA, to solve the sparse PCA on incomplete data. We note that very efficient scalable SDP solvers exist in practice [Yurtsever et al., 2021]. We assume that the unknown true matrix $\boldsymbol{M}^* \in \mathbb{R}^{d\times d}$ is symmetric and has a sparse leading eigenvector $\boldsymbol{u}_1$. Our goal is to exactly recover the support of this sparse leading eigenvector, i.e., to find the set J correctly where $J = supp(\boldsymbol{u}_1) = \{i : u_{1,i} \neq 0\}$. Given a noisy observation $\boldsymbol{M}$ for the unknown true matrix $\boldsymbol{M}^*$, it is intuitive to consider imposing a regularization term on the PCA quadratic loss that aims to find the first principal component. When using the l_1 regularizer, the optimization problem can be written as:

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}^\top \boldsymbol{x}=1}{\arg\max}\, \boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} - \rho\|\boldsymbol{x}\|_1^2.$$

Hence, J is estimated with $supp(\hat{\boldsymbol{x}})$. However, this intuitively appealing objective is nonconvex and very difficult to solve, so the following semidefinite relaxation can be considered as an alternative:

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{X}\succeq 0 \text{ and } tr(\boldsymbol{X})=1}{\arg\max} \langle \boldsymbol{M}, \boldsymbol{X}\rangle - \rho\|\boldsymbol{X}\|_{1,1}.$$

By letting $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$, the equivalence of the above two objective functions can be easily justified. Since $supp(\boldsymbol{x}) = supp(diag(\boldsymbol{x}\boldsymbol{x}^\top))$, we estimate the support J by $\hat{J} = supp(diag(\hat{\boldsymbol{X}}))$ in the semidefinite problem. This kind of relaxation has been studied by d'Aspremont et al. [2004] and Lei and Vu [2015], but their works were limited to complete data. Surprisingly, without any additional modifications on the relaxation problem such as using matrix factorization or matrix completion, we show that it is possible to exactly recover true support J with the above semidefinite program itself when $\boldsymbol{M}$ is an incomplete observation. Our main contribution is to prove this claim theoretically and experimentally.

In Section 3, we provide theoretical justification (i.e., Theorem 1) that we can exactly recover the true support J with high probability by obtaining a unique solution of the semidefinite problem, under proper conditions. The conditions involve matrix coherence parameters, the spectral gap between the largest and second-largest eigenvalues of the true matrix, the observation probability and the noise variance, which are discussed in detail in Corollaries 1 and 2. Specifically, we show that the sample complexity is related to the matrix coherence parameters as well as the matrix dimension d and the support size s. We prove that the observation probability p has the bound of $p = \omega\left(\frac{1}{d^{-1}+1}\right)$ in the worst scenario in terms of the matrix coherence, while it has a smaller lower bound $p = \omega\left(\frac{1}{(\log s)^{-1}+1}\right)$ in the best scenario. In Section 4, we provide experimental results on incomplete synthetic datasets and a gene expression dataset. The experiment on the synthetic datasets validate our theoretical results, and the experiment on the gene expression dataset gives us a consistent result with prior studies. We also show that our SDP algorithm outperforms several other sparse PCA approaches in the synthetic dataset.

2 Preliminaries

2.1 Notation

We first introduce the notations used throughout the paper. Matrices are bold capital, vectors are bold lowercase and scalars or entries are not bold. For any positive integer n, we denote $[n] := \{1, \ldots, n\}$. For any vector $\boldsymbol{a} \in \mathbb{R}^d$ and index set $J \subseteq [d]$, $\boldsymbol{a}_J$ denotes the $|J|$-dimensional vector consisting of

the entries of $\boldsymbol{a}$ in J. For any matrix $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$ and index sets $J_1 \subseteq [d_1]$ and $J_2 \subseteq [d_2]$, $\boldsymbol{A}_{J_1,J_2}$ and $\boldsymbol{A}_{J_1,:}(\boldsymbol{A}_{:,J_2})$ denote the $|J_1| \times |J_2|$ sub-matrix of $\boldsymbol{A}$ consisting of rows in J_1 and columns in J_2, and the $|J_1| \times d_2$ ($d_1 \times |J_2|$) sub-matrix of $\boldsymbol{A}$ consisting of rows in J_1 (columns in J_2), respectively. $\|\boldsymbol{a}\|_1$, $\|\boldsymbol{a}\|_2$ and $\|\boldsymbol{a}\|_\infty$ represent the l_1 norm, l_2 norm and maximum norm of a vector $\boldsymbol{a}$, respectively. $\{\boldsymbol{e}_i \ : \ i \in [d]\}$ indicates the standard basis of $\mathbb{R}^d$.

A variety of norms on matrices will be used: we denote by $\|\boldsymbol{A}\|_2$ the spectral norm and by $\|\boldsymbol{A}\|_F$ the Frobenius norm of a matrix $\boldsymbol{A}$. We let $\|\boldsymbol{A}\|_{1,1} = \sum_{i\in[d_1],j\in[d_2]} |A_{i,j}|$, $\|\boldsymbol{A}\|_{\max} = \|\boldsymbol{A}\|_{\infty,\infty} = \max_{i\in[d_1],j\in[d_2]} |A_{i,j}|$, $\|\boldsymbol{A}\|_{2,\infty} = \max_{j\in[d_2]} \|\boldsymbol{A}_{:,j}\|_2$ and $\|\boldsymbol{A}\|_{1,\infty} = \max_{j\in[d_2]} \|\boldsymbol{A}_{:,j}\|_1$ represent the $l_{1,1}$ norm, the entrywise l_∞ norm, the $l_{2,\infty}$ norm and the $l_{1,\infty}$ norm of a matrix $\boldsymbol{A}$, respectively. The trace of $\boldsymbol{A}$ is denoted $tr(\boldsymbol{A})$, and the matrix inner product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is denoted $\langle \boldsymbol{A}, \boldsymbol{B} \rangle$. Also, $\sigma_i(\boldsymbol{A})$ and $\lambda_i(\boldsymbol{A})$ represent the ith largest singular value and the ith largest eigenvalue of $\boldsymbol{A}$, respectively.

The notation $C, C_1, \ldots, c, c_1, \ldots$ denote positive constants whose values may change from line to line. The notation $f(x) = o(g(x))$ or $f(x) \ll g(x)$ means $\lim_{x\to\infty} f(x)/g(x) = 0$; $f(x) = \omega(g(x))$ or $f(x) \gg g(x)$ means $\lim_{x\to\infty} f(x)/g(x) = \infty$; $f(x) = O(g(x))$ or $f(x) \lesssim g(x)$ means that there exists a constant C such that $f(x) \leq Cg(x)$ asymptotically; $f(x) = \Omega(g(x))$ or $f(x) \gtrsim g(x)$ means that there exists a constant C such that $f(x) \geq Cg(x)$ asymptotically; $f(x) = \Theta(g(x))$ or $f(x) \simeq g(x)$ means that there exists constants C and C' such that $Cg(x) \leq f(x) \leq C'g(x)$ asymptotically.

2.2 Model

We now introduce our model assumption. Suppose that an unknown matrix $\boldsymbol{M}^* \in \mathbb{R}^{d\times d}$ is symmetric. The spectral decomposition of $\boldsymbol{M}^*$ is given by

$$\boldsymbol{M}^* = \sum_{k\in[d]} \lambda_k(\boldsymbol{M}^*)\boldsymbol{u}_k\boldsymbol{u}_k^\top,$$

where $\lambda_1(\boldsymbol{M}^*) \geq \cdots \geq \lambda_d(\boldsymbol{M}^*)$ are its eigenvalues and $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d \in \mathbb{R}^d$ are the corresponding eigenvectors. We assume that $\lambda_1(\boldsymbol{M}^*) > \lambda_2(\boldsymbol{M}^*)$ and the leading eigenvector $\boldsymbol{u}_1$ of $\boldsymbol{M}^*$ is sparse, that is, for some set $J \in [d]$,

$$\begin{cases} u_{1,i} \neq 0 & \text{if } i \in J \\ u_{1,i} = 0 & \text{otherwise.} \end{cases}$$

With a notation $supp(\boldsymbol{a}) := \{i \in [d] : a_i \neq 0\}$ for any vector $\boldsymbol{a} \in \mathbb{R}^d$, we can write $J = supp(\boldsymbol{u}_1)$. Also, we denote the size of J by s.

Incomplete and noisy observation Suppose that we have only noisy observations of the entries of $\boldsymbol{M}^*$ over a sampling set $\Omega \subseteq [d] \times [d]$. Specifically, we observe a symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ such that

$$M_{i,j} = M_{j,i} = \delta_{i,j} \cdot (M^*_{i,j} + \epsilon_{i,j})$$

for $1 \leq i \leq j \leq d$, where $\delta_{i,j} = 1$ if $(i,j) \in \Omega$ and $\delta_{i,j} = 0$ otherwise, and $\epsilon_{i,j}$ is the noise at location (i,j). In this paper, we consider the following assumptions on random sampling and random noise: for $1 \leq i \leq j \leq d$,

- Each (i,j) is included in the sampling set Ω independently with probability p (that is, $\delta_{i,j} \overset{i.i.d.}{\sim} Ber(p)$.)
- $\delta_{i,j}$'s and $\epsilon_{i,j}$'s are mutually independent.
- $\mathbb{E}[\epsilon_{i,j}] = 0$ and $\mathsf{Var}[\epsilon_{i,j}] = \sigma^2$.
- $|\epsilon_{i,j}| \leq B$ almost surely.

Finally, we define the coherence parameters of the sub-matrices $\boldsymbol{M}^*_{J,J}$, $\boldsymbol{M}^*_{J^c,J}$ and $\boldsymbol{M}^*_{J^c,J^c}$.

Definition 1 (Coherence parameters)**.** *We define the coherence parameters $\mu_0(\boldsymbol{M}^*_{J,J})$, $\mu_1(\boldsymbol{M}^*_{J,J})$, $\mu_2(\boldsymbol{M}^*_{J^c,J})$ and $\mu_3(\boldsymbol{M}^*_{J^c,J^c})$ as follows:*

$$\mu_0(\boldsymbol{M}^*_{J,J}) := \frac{\|\boldsymbol{M}^*_{J,J}\|_{\max}}{\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J})}, \quad \mu_1(\boldsymbol{M}^*_{J,J}) := \frac{\|\boldsymbol{M}^*_{J,J}\|_{\max}}{\|\boldsymbol{M}^*_{J,J}\|_{2,\infty}},$$

$$\mu_2(\boldsymbol{M}^*_{J^c,J}) := \frac{\|\boldsymbol{M}^*_{J^c,J}\|_{\max}}{\|\boldsymbol{M}^*_{J^c,J}\|_F}, \qquad \mu_3(\boldsymbol{M}^*_{J^c,J^c}) := \frac{\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max}}{\max\left\{\|\boldsymbol{M}^*_{J^c,J^c}\|_2, \|\boldsymbol{M}^*_{J^c,J^c}\|_{2,\infty}\right\}}.$$

We use μ_0, μ_1, μ_2 and μ_3 as shorthand for $\mu_0(\boldsymbol{M}^*_{J,J})$, $\mu_1(\boldsymbol{M}^*_{J,J})$, $\mu_2(\boldsymbol{M}^*_{J^c,J})$ and $\mu_3(\boldsymbol{M}^*_{J^c,J^c})$, respectively.

Intuitively, when each coherence parameter is small, all the entries of the corresponding matrix have comparable magnitudes. Note that $\frac{1}{s} \le \mu_0 \le 1$, $\frac{1}{\sqrt{s}} \le \mu_1 \le 1$, $\frac{1}{\sqrt{s(d-s)}} \le \mu_2 \le 1$, $\frac{1}{d-s} \le \mu_3 \le 1$.

Remark. *We note that our setting is different from the case in which a data matrix $\boldsymbol{Y}$ is observed with missing entries and the covariance matrix $\boldsymbol{M} = \boldsymbol{Y}^\top\boldsymbol{Y}$ is analyzed for sparse PCA. Instead, we directly observe a general symmetric incomplete matrix $\boldsymbol{M}$, which is not necessarily positive semidefinite. Our model can be therefore applied to various data types other than covariance matrices (e.g., undirected random graphs with missing edges.) When a noisy and incomplete data matrix $\boldsymbol{Y}$ is given and its covariance matrix is considered, our current analysis cannot be directly applied and one may need some bias correction technique in the algorithm, since $\mathbb{E}[\boldsymbol{M}] = \mathbb{E}[\boldsymbol{Y}^\top\boldsymbol{Y}]$ is not the true matrix $\boldsymbol{Y}^{*\top}\boldsymbol{Y}^*$ where $\boldsymbol{Y}^* = \mathbb{E}[\boldsymbol{Y}]$. Furthermore, the independent missingness mechanism over $\boldsymbol{Y}$ does not imply independent missingness over $\boldsymbol{M}$. We leave this for future work.*

3 Main Results

As mentioned in the introduction, we consider the following semidefinite programming (SDP) in order to recover the true support J:

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{X}\succeq 0 \text{ and } tr(\boldsymbol{X})=1}{\arg\max} \langle \boldsymbol{M}, \boldsymbol{X}\rangle - \rho\|\boldsymbol{X}\|_{1,1}, \tag{1}$$

where we estimate J by $\hat{J} = supp(diag(\hat{\boldsymbol{X}}))$. We recall that (1) is a convex relaxation of the following nonconvex problem:

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}^\top\boldsymbol{x}=1}{\arg\max}\, \boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x} - \rho\|\boldsymbol{x}\|_1^2. \tag{2}$$

In Theorem 1, we will show that under appropriate conditions, the solution of (1) attains $\hat{J} = J$ with high probability. Our main technical tool used in the proof is the primal-dual witness argument [Wainwright, 2009]. We start with deriving the sufficient conditions for the primal-dual solutions of (1) to be uniquely determined and satisfy $supp(diag(\hat{\boldsymbol{X}})) = J$. We then establish a proper candidate solution which meets the derived sufficient conditions, where we make use of the Karush-Kuhn-Tucker (KKT) conditions of (2) to set up a reasonable candidate. We finally develop the conditions under which the established candidate solution satisfies the sufficient conditions from the primal-dual witness argument of (1) with high probability. Detailed proof is given in Appendix B.

Theorem 1. *Under the model defined in Section 2.2, assume that the following conditions hold:*

$$2\sqrt{2}\cdot\frac{K_1+\rho s}{p(\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J}))} \le \min_{i\in J}|u_{1,i}|,$$

$$\rho > 2\sqrt{ps^c\cdot\left\{(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_F^2 + (d-s)s\sigma^2\right\}} + p\cdot\|\boldsymbol{M}^*_{J^c,J}\|_{\max}$$

$$(K_2 + p\cdot\|\boldsymbol{M}^*_{J^c,J}\|_2)^2\cdot(1+\sqrt{s})^2 \le \left\{p\cdot(\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J})) - 2\cdot K_1 - 2\rho s\right\}$$
$$\times\left\{p\cdot(\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_1(\boldsymbol{M}^*_{J^c,J^c})) - K_1 - K_3 - \rho d\right\},$$

where $c > 0$, and K_1, K_2 and K_3 are defined as follows:

$$K_1 := (c+1)\cdot R_1\log(2s) + \sqrt{2(c+1)}\cdot R_2\sqrt{\log(2s)}$$
$$K_2 := (c+1)\cdot R_3\log d + \sqrt{2(c+1)}\cdot R_4\sqrt{\log d}$$
$$K_3 := (c+1)\cdot R_5\log(2(d-s)) + \sqrt{2(c+1)}\cdot R_6\sqrt{\log(2(d-s))}$$

and

$$R_1 := \max\{(1-p)\|\boldsymbol{M}^*_{J,J}\|_{\max} + B,\ \ p\|\boldsymbol{M}^*_{J,J}\|_{\max}\},$$
$$R_2 := \sqrt{p(1-p)}\|\boldsymbol{M}^*_{J,J}\|_{2,\infty} + \sqrt{ps\sigma^2},$$
$$R_3 := \max\{(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_{\max} + B,\ \ p\|\boldsymbol{M}^*_{J^c,J}\|_{\max}\},$$
$$R_4 := \max\{\sqrt{p(1-p)}\|\boldsymbol{M}^*_{J^c,J}\|_{2,\infty} + \sqrt{p(d-s)\sigma^2}, \sqrt{p(1-p)}\|\boldsymbol{M}^*_{J,J^c}\|_{2,\infty} + \sqrt{ps\sigma^2}\},$$
$$R_5 := \max\{(1-p)\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max} + B,\ \ p\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max}\},$$
$$R_6 := \sqrt{p(1-p)}\|\boldsymbol{M}^*_{J^c,J^c}\|_{2,\infty} + \sqrt{p(d-s)\sigma^2}.$$

Then the optimal solution $\hat{\boldsymbol{X}}$ to the problem (1) is unique and satisfies $supp(diag(\hat{\boldsymbol{X}})) = J$ with probability at least $1 - s^{-c} - d^{-c} - (2s)^{-c} - (2(d-s))^{-c}$.

Remark 1. *Our proposed method actually guarantees more than support recovery. Under the sufficient conditions, the optimal solution $\hat{\boldsymbol{X}}$ can be expressed as $\hat{\boldsymbol{X}} = \begin{pmatrix} \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top & 0 \\ 0 & 0 \end{pmatrix}$, where $\hat{\boldsymbol{x}}$ satisfies $\left\|\boldsymbol{u}_{1,J} - \hat{\boldsymbol{x}}\right\|_2 \le \min_{j\in J}|u_{1,j}|$ which is shown in our proof. This implies consistency of our solution in eigenvector estimation.*

Remark 2. *To deal with missingness, we utilized concentration inequalities such as matrix Bernstein inequality and Chebychev's inequality, which is one key technical difference from prior works on sparse PCA with complete data [Lei and Vu, 2015]. Beyond this, under the primal-dual witness framework, while Lei and Vu [2015] employed implicit constraints in the optimization problem, we made all the constraints explicit and directly derived the sufficient conditions to avoid strong conditions. We note that simply applying concentration inequalities to the approach of Lei and Vu [2015] results in stricter requirements on $\boldsymbol{M}^*_{J^c,J^c}$ and p.*

We now consider the following two particular scenarios:

(s1) $B = \sigma_2 = 0$, that is, the observation $\boldsymbol{M}$ is noiseless (but still incomplete).

(s2) The rank of $\boldsymbol{M}^*$ is 1.

to better interpret the conditions of $\boldsymbol{M}^*$ and p listed in Theorem 1 and understand under what circumstance these conditions hold. For both cases, we set $p \ge 0.5$ for simplicity. Under the first setting, we can re-express the conditions on $\boldsymbol{M}^*$ for exact sparse recovery of J in a more interpretable way (specifically, in terms of coherence parameters and spectral gap) as well as the conditions on p. In the second setting, we aim to investigate that the maximum level of noise that is allowed by Theorem 1. Corollaries 1 and 2 include the results of the two settings (s1) and (s2), respectively.

Corollary 1. *Assume that $B = \sigma_2 = 0$, $p \ge 0.5$ and $\min_{i\in J}|u_{1,i}| = \Omega(\frac{1}{\sqrt{s}})$. Denote $\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J})$ by $\bar{\lambda}(\boldsymbol{M}^*_{J,J})$. If the following conditions hold:*

$$\mu_0 = o\left(\frac{1}{\sqrt{s}\log s}\right), \tag{3}$$

$$\|\boldsymbol{M}^*_{J^c,J}\|_{\max} = o\left(\frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{s} \cdot \min\left\{\mu_2,\ \frac{1}{s},\ \frac{\sqrt{s}}{\log d}\right\}\right), \tag{4}$$

$$\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max} = o\left(\bar{\lambda}(\boldsymbol{M}^*_{J,J}) \cdot \min\left\{\mu_3,\ \frac{1}{\log(d-s)}\right\}\right), \tag{5}$$

$$\begin{aligned}\sqrt{\frac{1-p}{p}} = o\Bigg(\min\Bigg\{&\mu_1\sqrt{\log s}, \\ &\frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})\mu_2}{\|\boldsymbol{M}^*_{J^c,J}\|_{\max}} \cdot \min\left\{\frac{1}{s^2\sqrt{s}},\ \frac{1}{s\sqrt{s(d-s)}}\right\}, \\ &\frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})\mu_3}{\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max}} \cdot \frac{1}{\sqrt{\log(d-s)}}\Bigg\}\Bigg),^{1}\end{aligned} \tag{6}$$

$$\rho = \Theta\left(\frac{p\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{s^2}\right), \tag{7}$$

then the conditions in Theorem 1 hold asymptotically, that is, when s and d are sufficiently large, the optimal solution $\hat{\boldsymbol{X}}$ to the problem (1) is unique and satisfies $supp(diag(\hat{\boldsymbol{X}})) = J$ with probability at least $1 - s^{-1} - d^{-1} - (2s)^{-1} - (2(d-s))^{-1}$.

Conditions on true matrix $\boldsymbol{M}^*$ From the conditions in Corollary 1, we can find desirable properties on the matrix $\boldsymbol{M}^*$ as follows:

- *Incoherence of $\boldsymbol{M}^*_{J,J}$, and coherence of $\boldsymbol{M}^*_{J^c,J}$ and $\boldsymbol{M}^*_{J^c,J^c}$*: From the coherence parameter in (3) and those in (4), (5) and (6), we see that the sub-matrix $\boldsymbol{M}^*_{J,J}$ and the sub-matrices $\boldsymbol{M}^*_{J^c,J}$ and $\boldsymbol{M}^*_{J^c,J^c}$ are expected to be incoherent and coherent, respectively. This is different from other problems involving incomplete matrices, such as matrix completion [Candès and Recht, 2009] and standard PCA on incomplete data [Cai et al., 2021], where the entire matrix, not a sub-matrix, is required to be incoherent.

  We can easily check the need of incoherence of $\boldsymbol{M}^*_{J^c,J}$ with an example that the sub-matrix has only one entry with a large magnitude while the other entries have relatively small values. Even if the true leading eigenvector of the sub-matrix is not sparse, the sparse PCA algorithm may produce a solution $\hat{J}$ which has a smaller size than that of the true support J. However, for $\boldsymbol{M}^*_{J^c,J}$ and $\boldsymbol{M}^*_{J^c,J^c}$, coherence is preferable: intuitively speaking, when $\boldsymbol{M}^*_{J^c,J}$ and $\boldsymbol{M}^*_{J^c,J^c}$ are the most coherent, that is, only one entry is nonzero in each sub-matrix, and all other entries are zero, missing the entries in $\boldsymbol{M}^*_{J^c,J}$ and $\boldsymbol{M}^*_{J^c,J^c}$ does not change the leading eigenvector of $\boldsymbol{M}^*$. On the other hand, when $\boldsymbol{M}^*_{J^c,J}$ and $\boldsymbol{M}^*_{J^c,J^c}$ are incoherent, that is, all the entries have comparable magnitudes, missing only a few entries changes the leading eigenvector and its sparsitency, so that sparse PCA is likely to fail to recover J. A simple illustration can be found in the Appendix A.

- *Large spectral gap $\bar{\lambda}(\boldsymbol{M}^*_{J,J})$ $(= \lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J}))$*: This can be found in (4), (5) and (6). A sufficiently large spectral gap requirement has been also discussed in the work on sparse PCA on the complete matrix [Lei and Vu, 2015]. It ensures the uniqueness and identifiability of the orthogonal projection matrix with respect to the principal subspace. If the spectral gap of eigenvalues is nearly zero, then the top two eigenvectors are indistinguishable given the observational noise, leading to failure to recover the sparsity of the leading eigenvector.
  We also note that $\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J}) \geq \lambda_1(\boldsymbol{M}^*) - \lambda_2(\boldsymbol{M}^*)$ since $\lambda_1(\boldsymbol{M}^*_{J,J}) = \lambda_1(\boldsymbol{M}^*)$ and $\lambda_2(\boldsymbol{M}^*_{J,J}) \leq \lambda_2(\boldsymbol{M}^*)$. Hence, a large $\lambda_1(\boldsymbol{M}^*) - \lambda_2(\boldsymbol{M}^*)$ implies a large $\bar{\lambda}(\boldsymbol{M}^*_{J,J})$.

- *Small magnitudes of $\boldsymbol{M}^*_{J^c,J}$ and $\boldsymbol{M}^*_{J^c,J^c}$*: This can also be found in (4), (5) and (6). This condition is also natural: if the magnitudes are relatively small, missing the entries will not make a big impact to the result.

Conditions on p (ratio of missing data) For simplicity, suppose that $\bar{\lambda}(\boldsymbol{M}^*_{J,J}) = O(s)$ and $s = O(\log d)$. We consider two extreme cases where the coherence parameters are maximized and minimized. We discuss the bound of the sample complexity in each case.

- *The best scenario where the bound of the sample complexity is the lowest*: Suppose that $\mu_1 = o(\frac{1}{\log s})$ and $\mu_2 = \mu_3 = 1$ (note that when $\mu_0 = o\big(\frac{1}{\sqrt{s}\log s}\big)$, μ_1 is upper bounded by $o\big(\frac{1}{\log s}\big)$.) Then the condition (6) can be written as:

$$p = \frac{1}{1+\epsilon_1}, \quad \epsilon_1 = o\Bigg(\min\Big\{\frac{1}{\log s}, \quad \frac{1}{sd\|\boldsymbol{M}^*_{J^c,J}\|^2_{\max}}\Big\}\Bigg).$$

  Furthermore, if $\|\boldsymbol{M}^*_{J^c,J}\|_{\max} = O(\sqrt{\frac{\log s}{sd}})$, then $p = \omega\Big(\frac{1}{(\log s)^{-1}+1}\Big)$.

[1]This can be expressed equivalently as $p = \frac{1}{1+\epsilon^2}$ where $\epsilon = o\Big(\min\Big\{\mu_1\sqrt{\log s}, \ \frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})\mu_2}{\|\boldsymbol{M}^*_{J^c,J}\|_{\max}} \cdot \min\big\{\frac{1}{s^2\sqrt{s}}, \ \frac{1}{s\sqrt{s(d-s)}}\big\}, \ \frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})\mu_3}{\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max}} \cdot \frac{1}{\sqrt{\log(d-s)}}\Big\}\Big)$.

- *The worst scenario where the bound of the sample complexity is the highest*: Suppose that $\mu_1 = \frac{1}{\sqrt{s}}$, $\mu_2 = \frac{1}{\sqrt{s(d-s)}}$ and $\mu_3 = \frac{1}{d-s}$. In this case, the condition (6) can be written as:

$$p = \frac{1}{1+\epsilon_2}, \ \ \epsilon_2 = o\Bigg(\min\Big\{ \frac{\log s}{s}, \ \frac{1}{(sd\|\boldsymbol{M}^*_{J^c,J}\|_{\max})^2}, \ \frac{1}{(d\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max})^2} \Big\} \Bigg).$$

  Suppose that $\|\boldsymbol{M}^*_{J^c,J}\|_{\max} = \Theta(\frac{1}{s\sqrt{d}})$ and $\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max} = \Theta(\frac{1}{\sqrt{d}})$ for instance, then $p = \omega\Big(\frac{1}{d^{-1}+1}\Big)$.

Next, we consider the second setting (s2) where the rank of $\boldsymbol{M}^*$ is assumed to be 1, that is, $\boldsymbol{M}^* = \lambda_1(\boldsymbol{M}^*)\boldsymbol{u}_1\boldsymbol{u}_1^\top$ (without loss of generality, we assume $\lambda_1(\boldsymbol{M}^*) > 0$). Trivially, $\boldsymbol{M}^*_{J^c,J} = \boldsymbol{M}^*_{J,J^c} = \boldsymbol{M}^*_{J^c,J^c} = 0$ and Theorem 1 can be greatly simplified. Here, we focus on analyzing how much noise (parameters B and σ^2) is allowed.

Corollary 2. *Assume that $p \geq 0.5$ and the rank of $\boldsymbol{M}^*$ is 1, that is, $\boldsymbol{M}^* = \lambda_1(\boldsymbol{M}^*)\boldsymbol{u}_1\boldsymbol{u}_1^\top$. Let $\lambda_1(\boldsymbol{M}^*) > 0$. Suppose that s and d satisfy $\frac{1}{\sqrt{s}} \leq \frac{12+\frac{d-s}{s}+8\sqrt{2}a_2-\sqrt{(4-\frac{d-s}{s}-8\sqrt{2}a_2)^2+512a_1^2(1+\sqrt{s})^2}}{4\sqrt{2}+\sqrt{2}\cdot\frac{d-s}{s}+16a_2-16\sqrt{2}a_1^2(1+\sqrt{s})^2}$ where $a_1 = (2-\frac{1}{p})\cdot\frac{\log d}{8\sqrt{2}\log(2s)} + \frac{\sqrt{\max\{d-s,s\}}\cdot\sqrt{\log d}}{16s^2\sqrt{d-s}}$ and $a_2 = (2-\frac{1}{p})\cdot\frac{\log(2(d-s))}{8\sqrt{2}\log(2s)} + \frac{\sqrt{\log(2(d-s))}}{16s^2}$. If the following conditions hold:*

$$\frac{\max_{i,j\in J}|u_{1,i}u_{1,j}|}{\min_{i\in J}|u_{1,i}|} \leq \frac{1}{16\sqrt{2}\log(2s)},$$

$$\frac{\max_{i\in J}|u_{1,i}|}{\min_{i\in J}|u_{1,i}|} \leq \frac{1}{16\sqrt{2}\sqrt{\log(2s)}}\cdot\sqrt{\frac{p}{1-p}},$$

$$B \leq (2p-1)\lambda_1(\boldsymbol{M}^*)\cdot\max_{i,j\in J}|u_{1,i}u_{1,j}|,$$

$$2\sqrt{2}\cdot\sqrt{p\sigma^2 s^2(d-s)} < \rho \leq \frac{1}{8\sqrt{2}s}\cdot p\lambda_1(\boldsymbol{M}^*)\cdot\min_{i\in J}|u_{1,i}|,$$

then the optimal solution $\hat{\boldsymbol{X}}$ to the problem (1) is unique and satisfies $supp(diag(\hat{\boldsymbol{X}})) = J$ with probability at least $1 - s^{-1} - d^{-1} - (2s)^{-1} - (2(d-s))^{-1}$.

Conditions on noise parameters B and σ^2 For simplicity, let $\lambda_1(\boldsymbol{M}^*) = O(s)$ and $\forall|u_{1,i}| = \Theta(\frac{1}{\sqrt{s}})$. Then the above conditions in Corollary 2 imply that

$$B \lesssim p \ \ \text{and} \ \ \sigma^2 \lesssim \frac{p}{s^3(d-s)}.$$

The condition for B is relatively moderate while σ^2 needs to be extremely small to satisfy the condition in Corollary 2. We comment this is only a sufficient condition, and the experimental results show that (1) can succeed even with σ^2 larger than the aforementioned bound.

4 Numerical Results

We perform the SDP algorithm of (1) on synthetic and real data to validate our theoretic results and show how well the true support of the sparse principal component is exactly recovered. We also compare our SDP algorithm with other sparse PCA methods and show that our method performs better. Our experiments were executed on `MATLAB` and standard `CVX` code was used, although very efficient scalable SDP solvers exist in practice [Yurtsever et al., 2021].

4.1 Synthetic Data

We perform two lines of experiments:

1. With the spectral gap $\lambda_1(\boldsymbol{M}^*) - \lambda_2(\boldsymbol{M}^*)$ and the noise parameters B and σ^2 fixed, we compare the results for different s and d.

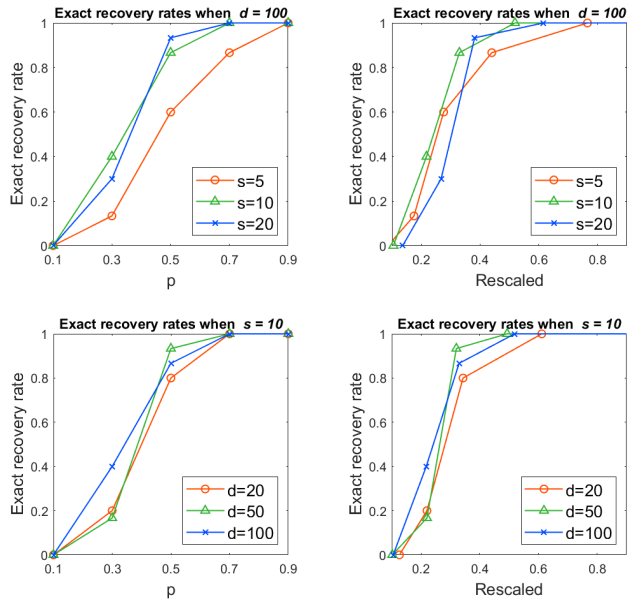

Figure 1: Results of experiment 1 on synthetic data.

2. With s and d fixed, we compare the results for different spectral gaps and noise parameters.

In each experiment, we generate the true matrix $\boldsymbol{M}^*$ as follows: the leading eigenvector $\boldsymbol{u}_1$ is set to have s number of non-zero entries. $\lambda_2(\boldsymbol{M}^*), \ldots, \lambda_d(\boldsymbol{M}^*)$ are randomly selected from a normal distribution with mean 0 and standard deviation 1, and $\lambda_1(\boldsymbol{M}^*)$ is set to $\lambda_2(\boldsymbol{M}^*)$ plus the spectral gap. The orthogonal eigenvectors are randomly selected, while the non-zero entries of the leading eigenvector $\boldsymbol{u}_1$ are made to have a value of at least $\frac{1}{2\sqrt{s}}$.

When generating the observation $\boldsymbol{M}$, we first add to $\boldsymbol{M}^*$ the entry-wise noise which is randomly selected from a truncated normal distribution with support $[-B, B]$. The normal distribution to be truncated is set to have mean 0 and standard deviation σ_{normal}. After adding the entry-wise noise, we generate an incomplete matrix $\boldsymbol{M}$ by selecting the observed entries uniformly at random with probability $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

In each setting, we run the algorithm (1) and verify if the solution exactly recovers the true support. We repeat each experiment 30 times with different random seeds, and calculate the rate of exact recovery in each setting.

Experiment 1 In this experiment, we fix the spectral gap $\lambda_1(\boldsymbol{M}^*) - \lambda_2(\boldsymbol{M}^*)$ as 20 and the noise parameters B and σ^2 as 5 and 0.01. We use the tuning parameter $\rho = 0.1$. We try three different matrix dimensions $d \in \{20, 50, 100\}$ and three different support sizes $s \in \{5, 10, 20\}$.

To check whether the bound of the sample complexity obtained in Corollary 1 is tight, we calculate the coherence parameters and the maximum magnitudes of the sub-matrices at each setting, and calculate the following rescaled parameter:

$$\sqrt{\frac{p}{1-p}} \cdot \min\left\{\mu_1\sqrt{\log s}, \frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})\mu_2}{\|\boldsymbol{M}^*_{J^c,J}\|_{\max}} \cdot \min\left\{\frac{1}{s^2\sqrt{s}}, \frac{1}{s\sqrt{s(d-s)}}\right\}, \frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})\mu_3}{\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max}} \cdot \frac{1}{\sqrt{\log(d-s)}}\right\},$$

which is derived from (6). If the exact recovery rate versus this rescaled parameter is the same across different settings, then we empirically justify that the bound of the sample complexity we derive is "tight"[2] in the sense that the exact recovery rate is solely determined by this rescaled parameter.

Figure 1 shows the experimental results. The two plots above are the experimental results for different values of s when $d = 100$, and the two plots below are for different values of d when $s = 10$.

[2]We note that "tight" does not mean an agreement between necessary and sufficient conditions here.

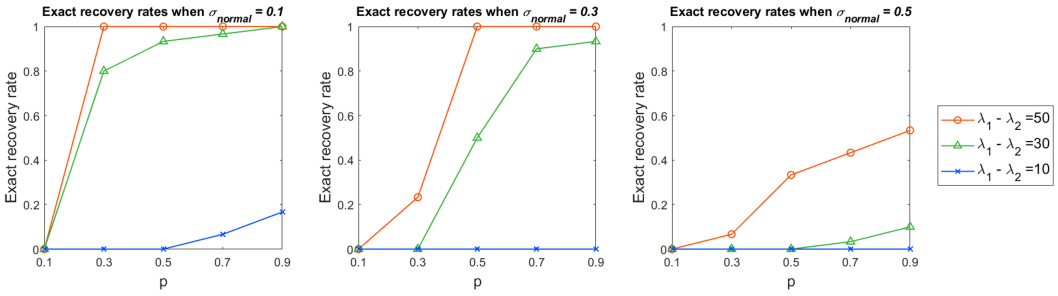

Figure 2: Results of experiment 2 on synthetic data.

The x-axis of the left graphs represents p, and the x-axis of the right graphs indicates the rescaled parameter.

We can see from the two graphs on the right that the exact recovery rate versus the rescaled parameter is the same in different settings of d and s. This means that our bound of the sample complexity is tight.

Another observation we can make is that the exact recovery rate is not necessarily increasing or decreasing as s or d increases or decreases. This is probably because coherences and maximum magnitudes of sub-matrices are involved in the sample complexity as well.

Experiment 2 Here, we fix the matrix dimension d as 100 and the support size s as 50. We set $B = 5$. We try three different spectral gaps $\lambda_1(\boldsymbol{M}^*) - \lambda_2(\boldsymbol{M}^*) \in \{10, 30, 50\}$ and three different standard deviations of the normal distribution, $\sigma_{normal} \in \{0.1, 0.3, 0.5\}$. We try two different tuning parameters $\rho \in \{0.1, 0.01\}$ and report the best result.

Figure 2 demonstrates the experimental results. The three plots show the results when σ_{normal} is 0.1, 0.3 and 0.5, respectively. The red, green and blue lines indicate the cases where the spectral gap $\lambda_1(\boldsymbol{M}^*) - \lambda_2(\boldsymbol{M}^*)$ is 50, 30 and 10, respectively. From the plots, we can observe that the exact recovery rate increases as σ^2 is small and $\lambda_1(\boldsymbol{M}^*) - \lambda_2(\boldsymbol{M}^*)$ is large, which is consistent with the conditions we have checked in Corollaries 1 and 2.

4.2 Gene Expression Data

We analyze a gene expression dataset (GSE21385) from the Gene Expression Omnibus website (`https://www.ncbi.nlm.nih.gov/geo/`.) The dataset examines rheumatoid arthritis synovial fibroblasts, which together with synovial macrophages, are the two leading cell types that invade and degrade cartilage and bone.

The original data set contains 56 subjects and 112 genes. We compute its incomplete covariance matrix, where 87% of the matrix entries are observed since some subject/gene pairs are unobserved. With this incomplete covariance matrix, we solve the semidefinite program in (1) for sparse PCA with $\rho = 2$.

By solving (1), we find that the support of the solution contains 3 genes: beta-1 catenin (CTNNB), hypoxanthine-guanine phosphoribosyltransferase 1 (HPRT1) and semaphorin III/F (SEMA3F). Our result is consistent with prior studies on rheumatoid arthritis since CTNNB has been found to be upregulated [Iwamoto et al., 2018], SEMA3F has been found to be downregulated [Tang et al., 2018], and HPRT1 is known to be a housekeeping gene [Mesko et al., 2013]. Additional illustration of this application can be found in Appendix G.

4.3 Comparison with Other Methods

We compare our SDP algorithm with three different methods. First, we consider two sparse PCA algorithms where missing cells are treated as zero: the diagonal thresholding sparse PCA (DTSPCA) by Johnstone and Lu [2009] and the iterative thresholding sparse PCA (ITSPCA) by Ma [2013]. Second, we consider the combination of imputation and our SDP method: we first estimate the

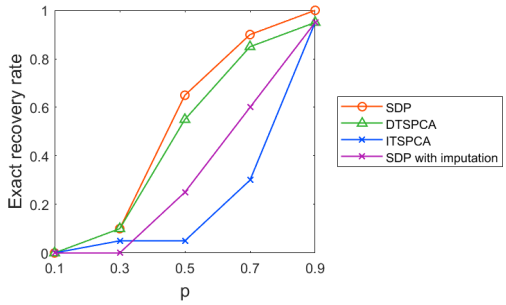

Figure 3: Comparison with other methods.

missing elements of $\boldsymbol{M}$ by using the matrix completion method based on the following Frobenius norm miminization with nuclear norm penalty $\|\boldsymbol{Y}\|_*$:

$$\tilde{\boldsymbol{M}} = \underset{\boldsymbol{Y}:symmetric}{\arg\min} \|P_\Omega(\boldsymbol{Y}) - \boldsymbol{M}\|_F^2 + \tau\|\boldsymbol{Y}\|_*$$

and then implemented the SDP method with the completed matrix $\tilde{\boldsymbol{M}}$. Details of the simulation settings and algorithm implementations are provided in Appendix F.

Figure 3 demonstrates that the exact recovery rate of our SDP method is higher than those of the three other methods. The comparison shows that the imputation-based strategy fails to work under the small-p regime, and we conjecture the following rationale: it is known that the matrix completion algorithm can be successful under the low-rank assumption. However, we do not assume that the true matrix is low-rank, but just that its spectral gap $\lambda_1(\boldsymbol{M}^*) - \lambda_2(\boldsymbol{M}^*)$ is large enough, which probably leads to the failure of matrix completion. Over the 20 simulation replications, we found that the average of $\max_{i,j} |\tilde{M}_{ij} - M^*_{ij}|$ was 1.0260 while the averages of $\max_{i,j} |M^*_{ij}|$ and $\min_{i,j} |M^*_{ij}|$ are 2.3200 and 0.0002, respectively. This shows that the matrix completion has not been done very precisely. Given such an unsuccessful matrix completion, the imputed cells introduce more noise into the inference and the result of sparse PCA can be even worse than that of simply using zero for the missing entries.

5 Concluding Remarks

We have presented the sufficient conditions to exactly recover the true support of the sparse leading eigenvector by solving a simple semidefinite programming on an incomplete and noisy observation. We have shown that the conditions involve matrix coherence, spectral gap, matrix magnitudes, sample complexity and variance of noise, and provided empirical evidence to justify our theoretical results. To the best of our knowledge, we provide the first theoretical guarantee for exact support recovery with sparse PCA on incomplete data. While we currently focus on a uniformly missing at random setup, an interesting open question is whether it is possible to provide guarantees for a deterministic pattern of missing entries.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2134209-DMS.

References

Arash A Amini and Martin J Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *2008 IEEE international symposium on information theory*, pages 2454–2458. IEEE, 2008.

Lauren Berk and Dimitris Bertsimas. Certifiably optimal sparse principal component analysis. *Mathematical Programming Computation*, 11(3):381–420, 2019.

Changxiao Cai, Gen Li, Yuejie Chi, H Vincent Poor, and Yuxin Chen. Subspace estimation from unbalanced and incomplete data matrices: $\ell 2,\infty$ statistical guarantees. *The Annals of Statistics*, 49 (2):944–967, 2021.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Alexandre d'Aspremont, Laurent Ghaoui, Michael Jordan, and Gert Lanckriet. A direct formulation for sparse pca using semidefinite programming. *Advances in neural information processing systems*, 17, 2004.

Naoki Iwamoto, Shoichi Fukui, Ayuko Takatani, Toshimasa Shimizu, Masataka Umeda, Ayako Nishino, Takashi Igawa, Tomohiro Koga, Shin-ya Kawashiri, Kunihiro Ichinose, et al. Osteogenic differentiation of fibroblast-like synovial cells in rheumatoid arthritis is induced by microrna-218 through a robo/slit pathway. *Arthritis research & therapy*, 20(1):1–10, 2018.

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(2), 2010.

Abhisek Kundu, Petros Drineas, and Malik Magdon-Ismail. Approximating sparse pca from incomplete data. *Advances in Neural Information Processing Systems*, 28, 2015.

Jing Lei and Vincent Q Vu. Sparsistency and agnostic inference in sparse pca. *The Annals of Statistics*, 43(1):299–322, 2015.

Karim Lounici. Sparse principal component analysis with missing observations. In *High dimensional probability VI*, pages 327–356. Springer, 2013.

Zongming Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

Bertalan Mesko, Szilard Poliska, Andrea Váncsa, Zoltan Szekanecz, Karoly Palatka, Zsolt Hollo, Attila Horvath, Laszlo Steiner, Gabor Zahuczky, Janos Podani, et al. Peripheral blood derived gene panels predict response to infliximab in rheumatoid arthritis and crohn's disease. *Genome medicine*, 5(6):1–10, 2013.

Boaz Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817, 2008.

Seyoung Park and Hongyu Zhao. Sparse principal component analysis with missing observations. *The Annals of Applied Statistics*, 13(2):1016–1042, 2019.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

Peter Richtárik, Majid Jahani, Selin Damla Ahipaşaoğlu, and Martin Takáč. Alternating maximization: unifying framework for 8 sparse pca formulations and efficient parallel codes. *Optimization and Engineering*, 22(3):1493–1519, 2021.

Man Wai Tang, Beatriz Malvar Fernández, Simon P Newsom, Jaap D van Buul, Timothy RDJ Radstake, Dominique L Baeten, Paul P Tak, Kris A Reedquist, and Samuel García. Class 3 semaphorins modulate the invasive capacity of rheumatoid arthritis fibroblast-like synoviocytes. *Rheumatology*, 57(5):909–920, 2018.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5): 2183–2202, 2009.

Alp Yurtsever, Joel A Tropp, Olivier Fercoq, Madeleine Udell, and Volkan Cevher. Scalable semidefinite programming. *SIAM Journal on Mathematics of Data Science*, 3(1):171–200, 2021.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See the supplemental material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide the code in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.
   (b) Did you mention the license of the assets? [Yes] See Section 4.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Examples of coherent and incoherent sub-matrices $\boldsymbol{M}^*_{J^c,J}$ and $\boldsymbol{M}^*_{J^c,J^c}$

For illustration, here we present a noiseless case (that is, we only focus on the change of eigen-structure caused by missing values,) and set $J = \{1, \ldots, s\}$, i.e., only the first s entries of the true leading eigenvector are nonzero. We let $d = 20$ and $s = 10$ in the below examples.

In the following four examples, we show a complete or incomplete matrix, followed by its leading eigenvector. We separate each matrix and its leading eigenvector by an arrow. The entries in $J^c \times J$, $J \times J^c$ and $J^c \times J^c$ sub-matrices are marked in bold. The missing entries are marked in red.

$$\begin{pmatrix} 1 & \cdots & 1 & 1 & 1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cdots & 1 & 1 & 1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ 1 & \cdots & 1 & 2 & 0 & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ 1 & \cdots & 1 & 0 & 2 & \mathbf{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1} & \mathbf{-1} & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} \Rightarrow \begin{pmatrix} 0.3162 \\ \vdots \\ 0.3162 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & \cdots & 1 & 1 & 1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cdots & 1 & 1 & 1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ 1 & \cdots & 1 & 2 & 0 & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ 1 & \cdots & 1 & 0 & 2 & \mathbf{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1} & \mathbf{-1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} \Rightarrow \begin{pmatrix} 0.3162 \\ \vdots \\ 0.3162 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

The example on the left is a complete matrix having coherent sub-matrices $\boldsymbol{M}^*_{J^c,J}$ and $\boldsymbol{M}^*_{J^c,J^c}$, and the example on the right is its incomplete counterpart. We can observe that missing some entries does not change the leading eigenvector in this case.

$$\begin{pmatrix} 4 & 2 & \cdots & 4 & 2 & \mathbf{1} & \mathbf{-1} & \cdots & \mathbf{-1} \\ 2 & 4 & \cdots & 2 & 4 & \mathbf{-1} & \mathbf{1} & \cdots & \mathbf{1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 4 & 2 & \cdots & 4 & 2 & \mathbf{1} & \mathbf{-1} & \cdots & \mathbf{-1} \\ 2 & 4 & \cdots & 2 & 4 & \mathbf{-1} & \mathbf{1} & \cdots & \mathbf{1} \\ \mathbf{1} & \mathbf{-1} & \cdots & \mathbf{1} & \mathbf{-1} & \mathbf{1} & \mathbf{-1} & \cdots & \mathbf{-1} \\ \mathbf{-1} & \mathbf{1} & \cdots & \mathbf{-1} & \mathbf{1} & \mathbf{-1} & \mathbf{1} & \cdots & \mathbf{1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{-1} & \mathbf{1} & \cdots & \mathbf{-1} & \mathbf{1} & \mathbf{-1} & \mathbf{1} & \cdots & \mathbf{1} \end{pmatrix} \Rightarrow \begin{pmatrix} 0.3162 \\ \vdots \\ 0.3162 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 4 & 2 & \cdots & 4 & 2 & \mathbf{1} & \mathbf{-1} & \cdots & \mathbf{-1} \\ 2 & 4 & \cdots & 2 & 4 & \mathbf{-1} & \mathbf{1} & \cdots & \mathbf{1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 4 & 2 & \cdots & 4 & 2 & \mathbf{1} & \mathbf{-1} & \cdots & \mathbf{-1} \\ 2 & 4 & \cdots & 2 & 4 & \mathbf{0} & \mathbf{1} & \cdots & \mathbf{1} \\ \mathbf{1} & \mathbf{-1} & \cdots & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{-1} & \cdots & \mathbf{-1} \\ \mathbf{-1} & \mathbf{1} & \cdots & \mathbf{-1} & \mathbf{1} & \mathbf{-1} & \mathbf{1} & \cdots & \mathbf{1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{-1} & \mathbf{1} & \cdots & \mathbf{-1} & \mathbf{1} & \mathbf{-1} & \mathbf{1} & \cdots & \mathbf{1} \end{pmatrix} \Rightarrow \begin{pmatrix} 0.3172 \\ \vdots \\ 0.3155 \\ 0.0115 \\ -0.001 \\ \vdots \\ -0.001 \end{pmatrix}$$

However, as shown above, if the sub-matrices $\boldsymbol{M}^*_{J^c,J}$ and $\boldsymbol{M}^*_{J^c,J^c}$ are highly incoherent, then missing only one entry in the sub-matrix $\boldsymbol{M}^*_{J^c,J}$ changes the leading eigenvector significantly. In this case, the support of the leading eigenvector of the incomplete matrix is $\{1, \ldots, d\}$, so that it is more difficult to exactly recover the true support J.

B Proof of Theorem 1

We use the primal-dual witness construction [Wainwright, 2009] to obtain the sufficient conditions for support recovery. The following proposition indicates the sufficient conditions to recover the support without false positives by using the optimization problem (1).

Proposition 1. *If $\hat{\boldsymbol{X}} \in \mathbb{R}^{d\times d}$, $\hat{\boldsymbol{Z}} \in \mathbb{R}^{d\times d}$ and $\hat{\mu} \in \mathbb{R}$ satisfy the following:*

$$\begin{aligned}
&\hat{\boldsymbol{X}}_{J,J} \succeq 0, tr(\hat{\boldsymbol{X}}_{J,J}) = 1 \\
&\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{Z}}_{J,J} \preceq \hat{\mu}\boldsymbol{I} \\
&\boldsymbol{M} - \rho\hat{\boldsymbol{Z}} \preceq \hat{\mu}\boldsymbol{I} \\
&\hat{Z}_{ij} \in \partial|\hat{X}_{ij}| \quad \textit{for each } (i,j) \in J \times J \\
&\hat{Z}_{ij} \in (-1,1) \quad \textit{for each } (i,j) \notin J \times J \\
&(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{Z}}_{J,J})\hat{\boldsymbol{X}}_{J,J} = \hat{\mu} \cdot \hat{\boldsymbol{X}}_{J,J} \\
&(\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{Z}}_{J^c,J})\hat{\boldsymbol{X}}_{J,J} = 0,
\end{aligned}$$

then $\hat{\boldsymbol{X}}$ is an optimal solution to the problem (1), and it satisfies $supp(diag(\hat{\boldsymbol{X}})) \subseteq J$.

The proof is provided in E.1. Recall that our goal is to exactly recover the support J where $s = |J|$, and the conditions in Proposition 1 do not guarantee $supp(diag(\hat{\boldsymbol{X}})) = J$ yet. We now want to construct a solution $\hat{\boldsymbol{X}}$ satisfying not only the above optimality conditions but also $supp(diag(\hat{\boldsymbol{X}})) = J$. To

find a reasonable candidate, we look at the KKT conditions of the nonconvex problem (2) and apply the primal-dual witness argument. We note that the problem (2) is only used for getting some initial intuition in order to later construct a desirable solution to the problem (1), and solving the problem (2) is not our interest. Proposition 2 represents the sufficient conditions for the desirable solution to be uniquely obtained. The proof is deferred to E.2.

Proposition 2. *Consider a 3-tuple of* $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}, \hat{\boldsymbol{w}}) \in \mathbb{R}^s \times \mathbb{R}^s \times \mathbb{R}^{d-s}$ *such that*

$$\hat{z}_i = \mathit{sign}(u_{1,i}) \quad \text{for all } i \in J,$$
$$\hat{\boldsymbol{x}} \text{ is the leading eigenvector of } \boldsymbol{M}_{J,J} - \rho \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top,$$
$$\hat{\boldsymbol{w}} = \frac{1}{\rho \|\hat{\boldsymbol{x}}\|_1} \boldsymbol{M}_{J^c,J} \hat{\boldsymbol{x}}.$$

If the 3-tuple $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}, \hat{\boldsymbol{w}}) \in \mathbb{R}^s \times \mathbb{R}^s \times \mathbb{R}^{d-s}$ *satisfies the following conditions:*

$$\mathit{sign}(\hat{x}_i) = \mathit{sign}(u_{1,i}) \quad \text{for all } i \in J \tag{8}$$
$$\|\hat{\boldsymbol{w}}\|_\infty < 1 \tag{9}$$
$$\lambda_1(\boldsymbol{M}_{J,J} - \rho \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) = \lambda_1(\boldsymbol{M} - \rho(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)^\top (\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)) \tag{10}$$
$$\lambda_1(\boldsymbol{M}_{J,J} - \rho \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) > \lambda_2(\boldsymbol{M}_{J,J} - \rho \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) \tag{11}$$

then for $\hat{\boldsymbol{X}} := \begin{pmatrix} \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top & 0 \\ 0 & 0 \end{pmatrix}$, $\hat{\boldsymbol{X}}$ *is a unique optimal solution to the problem (1) and satisfies* $\mathit{supp}(\mathit{diag}(\hat{\boldsymbol{X}})) = J$.

For clarity of exposition, our abuse of notation seemingly assumes $J = [s]$ when we join vectors and matrices, for instance when we join $(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)$ and $\left(\begin{smallmatrix} \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top & 0 \\ 0 & 0 \end{smallmatrix}\right)$. It should be clear that for $J \neq [s]$, one will need to properly interleave vector entries or matrix rows/columns.

What remains is to derive the sufficient conditions for (8) - (11). Lemmas 1 to 4 below presents the sufficient conditions for (8) - (11), respectively. We provide the proofs of the lemmas in E.4 to E.7.

Lemma 1 (Sufficient Conditions for (8))**.** *If the following inequality holds:*

$$2\sqrt{2} \cdot \frac{K_1 + \rho s}{p(\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J}))} \leq \min_{i \in J} |u_{1,i}|,$$

then the condition (8) holds with probability at least $1 - (2s)^{-c}$ *for any* $c > 0$*. Note that* K_1 *is defined in Lemma 6.*

Lemma 2 (Sufficient Conditions for (9))**.** *If the following inequality holds:*

$$\rho > 2\sqrt{ps^c \cdot \left\{(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_F^2 + (d-s)s\sigma^2\right\}} + p \cdot \|\boldsymbol{M}^*_{J^c,J}\|_{\max},$$

then the condition (9) holds with probability at least $1 - s^{-c}$ *for any* $c > 0$.

Lemma 3 (Sufficient Conditions for (10))**.** *In addition to the conditions in Lemma 2, if the following inequality holds:*

$$(K_2 + p \cdot \|\boldsymbol{M}^*_{J^c,J}\|_2)^2 \cdot (1 + \sqrt{s})^2 \leq \left\{p \cdot (\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J})) - 2 \cdot K_1 - 2\rho s\right\}$$
$$\times \left\{p \cdot (\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_1(\boldsymbol{M}^*_{J^c,J^c})) - K_1 - K_3 - \rho d\right\},$$

then the condition (10) holds with probability at least $1 - s^{-c} - d^{-c} - (2s)^{-c} - (2(d-s))^{-c}$ *for any* $c > 0$*. Note that* K_1*,* K_2 *and* K_3 *are defined in Lemma 6.*

Lemma 4 (Sufficient Conditions for (11))**.** *In addition to the conditions in Lemma 2, if the following inequality holds:*

$$p \cdot (\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J})) \geq 2 \cdot K_1 + 2\rho s,$$

then the condition (11) holds with probability at least $1 - (2s)^{-c}$ *for any* $c > 0$*. Note that* K_1 *is defined in Lemma 6.*

Since $\min_{i \in J} |u_{1,i}| < \sqrt{2}$ always holds, the sufficient condition in Lemma 1 implies the sufficient condition in Lemma 4. Hence, we do not need the condition in Lemma 4. Combining the sufficient condtions in Lemma 1 to Lemma 3 gives the result of Theorem 1.

C Proof of Corollary 1

When $B = \sigma^2 = 0$ and $p \geq 0.5$, we can write

$$\begin{aligned} K_1 &\lesssim p\|\boldsymbol{M}^*_{J,J}\|_{\max}\log s + \sqrt{p(1-p)}\|\boldsymbol{M}^*_{J,J}\|_{2,\infty}\sqrt{\log s} \\ &= p\log s\|\boldsymbol{M}^*_{J,J}\|_{\max}\left(1 + \sqrt{\frac{1-p}{p}}(\mu_1\sqrt{\log s})^{-1}\right). \end{aligned}$$

Under the conditions (3), (6) and (7),

$$\begin{aligned} K_1 + \rho s &\lesssim p\log s\|\boldsymbol{M}^*_{J,J}\|_{\max} + \frac{p\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{s} = p\log s\cdot\bar{\lambda}(\boldsymbol{M}^*_{J,J})\mu_0 + \frac{p\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{s} \\ &\ll p\bar{\lambda}(\boldsymbol{M}^*_{J,J})\log s\frac{1}{\sqrt{s}\log s} + \frac{p\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{\sqrt{s}} = \frac{2p\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{\sqrt{s}} \lesssim \frac{p\bar{\lambda}(\boldsymbol{M}^*_{J,J})\cdot\min_{i\in J}|u_{1,i}|}{2\sqrt{2}}. \end{aligned}$$

Hence, $2\sqrt{2}\cdot\frac{K_1+\rho s}{p\bar{\lambda}(\boldsymbol{M}^*_{J,J})} \leq \min_{i\in J}|u_{1,i}|$ holds asymptotically.

Next, under the conditions (4), (6) and (7),

$$\begin{aligned} &2\sqrt{sp(1-p)}\|\boldsymbol{M}^*_{J^c,J}\|_F + p\|\boldsymbol{M}^*_{J^c,J}\|_{\max} \leq p\|\boldsymbol{M}^*_{J^c,J}\|_{\max}\left(1 + 2\sqrt{s}\sqrt{\frac{1-p}{p}}\mu_2^{-1}\right) \\ &\ll p\|\boldsymbol{M}^*_{J^c,J}\|_{\max}\left(\frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{s^2\|\boldsymbol{M}^*_{J^c,J}\|_{\max}} + \frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{s^2\|\boldsymbol{M}^*_{J^c,J}\|_{\max}}\right) \lesssim \frac{p\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{s^2} \lesssim \rho. \end{aligned}$$

Therefore, $\rho > 2\sqrt{sp(1-p)}\|\boldsymbol{M}^*_{J^c,J}\|_F + p\|\boldsymbol{M}^*_{J^c,J}\|_{\max}$ holds asymptotically.

Lastly, we will show that under the conditions (4), (5), (6) and (7),

$$\begin{aligned} (K_2 + p\cdot\|\boldsymbol{M}^*_{J^c,J}\|_2)^2\cdot(1+\sqrt{s})^2 \leq &\left\{p\cdot\bar{\lambda}(\boldsymbol{M}^*_{J,J}) - 2\cdot K_1 - 2\rho s\right\} \\ &\times\left\{p\cdot\lambda_1(\boldsymbol{M}^*_{J,J}) - K_1 - \rho s - K_3 - \frac{1}{\rho}\cdot(p\|\boldsymbol{M}^*_{J,J^c}\|_{\infty,2} + c_0\sqrt{d-s})^2\right\}, \end{aligned}$$

where $c_0 = \sqrt{sp(1-p)}\|\boldsymbol{M}^*_{J^c,J}\|_F$. Here, the bound (27) is used instead of (26). Under the conditions (4) and (6),

$$\begin{aligned} &(K_2 + p\cdot\|\boldsymbol{M}^*_{J^c,J}\|_2)\cdot(1+\sqrt{s}) \\ &\lesssim \left(p\|\boldsymbol{M}^*_{J^c,J}\|_{\max}\log d + \sqrt{p(1-p)}\max\left\{\|\boldsymbol{M}^*_{J^c,J}\|_{2,\infty}, \|\boldsymbol{M}^{*\top}_{J^c,J}\|_{2,\infty}\right\}\sqrt{\log d} + p\|\boldsymbol{M}^*_{J^c,J}\|_2\right)\cdot\sqrt{s} \\ &\leq p\|\boldsymbol{M}^*_{J^c,J}\|_{\max}\cdot\left(\log d + \sqrt{\log d}\sqrt{\frac{1-p}{p}}\mu_2^{-1} + \mu_2^{-1}\right)\cdot\sqrt{s} \\ &\ll p\|\boldsymbol{M}^*_{J^c,J}\|_{\max}\cdot\left(\frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})\sqrt{s}}{s\|\boldsymbol{M}^*_{J^c,J}\|_{\max}} + \frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})\sqrt{\log d}}{s\sqrt{s(d-s)}\|\boldsymbol{M}^*_{J^c,J}\|_{\max}} + \frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{s\|\boldsymbol{M}^*_{J^c,J}\|_{\max}}\right)\cdot\sqrt{s} \\ &\lesssim p\|\boldsymbol{M}^*_{J^c,J}\|_{\max}\cdot\frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{\sqrt{s}\|\boldsymbol{M}^*_{J^c,J}\|_{\max}}\cdot\sqrt{s} = p\bar{\lambda}(\boldsymbol{M}^*_{J,J}). \end{aligned}$$

Also, since $2\sqrt{2}\cdot\frac{K_1+\rho s}{p\bar{\lambda}(\boldsymbol{M}^*_{J,J})} \leq \min_{i\in J}|u_{1,i}| \leq \frac{1}{\sqrt{s}}$,

$$p\bar{\lambda}(\boldsymbol{M}^*_{J,J}) - 2K_1 - 2\rho s \geq p\bar{\lambda}(\boldsymbol{M}^*_{J,J}) - \frac{p\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{\sqrt{2s}} \simeq p\bar{\lambda}(\boldsymbol{M}^*_{J,J}).$$

Moreover, under the conditions (5) and (6),

$$\begin{aligned} K_3 &\lesssim p\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max}\log(d-s) + \sqrt{p(1-p)}\|\boldsymbol{M}^*_{J^c,J^c}\|_{2,\infty}\sqrt{\log(d-s)} \\ &= p\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max}\log(d-s)\cdot\left\{1 + \sqrt{\frac{1-p}{p}}\mu_3^{-1}\frac{1}{\sqrt{\log(d-s)}}\right\} \\ &\ll p\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max}\log(d-s)\cdot\left\{\frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max}\log(d-s)} + \frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max}\log(d-s)}\right\} \simeq p\bar{\lambda}(\boldsymbol{M}^*_{J,J}), \end{aligned}$$

and under the conditions (4), (6) and (7),

$$\begin{aligned}
&\frac{1}{\rho}\cdot(p\|\boldsymbol{M}^*_{J,J^c}\|_{\infty,2}+\sqrt{s(d-s)p(1-p)}\|\boldsymbol{M}^*_{J,J^c}\|_F)^2\\
&\le \frac{1}{\rho}\cdot\left\{p\|\boldsymbol{M}^*_{J,J^c}\|_{\max}\Big(\mu_2^{-1}+\sqrt{s(d-s)}\sqrt{\frac{1-p}{p}}\mu_2^{-1}\Big)\right\}^2\\
&\ll \frac{1}{\rho}\cdot\left\{p\|\boldsymbol{M}^*_{J,J^c}\|_{\max}\Big(\frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{s\|\boldsymbol{M}^*_{J,J^c}\|_{\max}}+\frac{\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{s\|\boldsymbol{M}^*_{J,J^c}\|_{\max}}\Big)\right\}^2 \simeq \frac{1}{\rho}\cdot\frac{(p\bar{\lambda}(\boldsymbol{M}^*_{J,J}))^2}{s^2}\\
&\simeq \frac{s^2}{p\bar{\lambda}(\boldsymbol{M}^*_{J,J})}\cdot\frac{(p\bar{\lambda}(\boldsymbol{M}^*_{J,J}))^2}{s^2}=p\bar{\lambda}(\boldsymbol{M}^*_{J,J}).
\end{aligned}$$

Hence,

$$\begin{aligned}
&p\cdot\lambda_1(\boldsymbol{M}^*_{J,J})-K_1-\rho s-K_3-\frac{1}{\rho}\cdot(p\|\boldsymbol{M}^*_{J,J^c}\|_{\infty,2}+c_0\sqrt{d-s})^2\\
&\ge p\bar{\lambda}(\boldsymbol{M}^*_{J,J})\cdot(1-o(1))-\frac{p\bar{\lambda}(\boldsymbol{M}^*_{J,J})}{2\sqrt{2s}}\simeq p\bar{\lambda}(\boldsymbol{M}^*_{J,J}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&(K_2+p\cdot\|\boldsymbol{M}^*_{J^c,J}\|_2)^2\cdot(1+\sqrt{s})^2\ll\big(p\bar{\lambda}(\boldsymbol{M}^*_{J,J})\big)^2\\
&\lesssim\Big\{p\cdot\bar{\lambda}(\boldsymbol{M}^*_{J,J})-2\cdot K_1-2\rho s\Big\}\times\Big\{p\cdot\lambda_1(\boldsymbol{M}^*_{J,J})-K_1-\rho s-K_3-\frac{1}{\rho}\cdot(p\|\boldsymbol{M}^*_{J,J^c}\|_{\infty,2}+c_0\sqrt{d-s})^2\Big\},
\end{aligned}$$

that is, the desired result holds asymptotically.

D Proof of Corollary 2

Since $\boldsymbol{M}^*_{J^c,J}=\boldsymbol{M}^*_{J,J^c}=\boldsymbol{M}^*_{J^c,J^c}=0$ and $\boldsymbol{M}^*_{J,J}=\lambda_1(\boldsymbol{M}^*)\boldsymbol{u}_{1,J}\boldsymbol{u}_{1,J}^\top$ for $\lambda_1(\boldsymbol{M}^*)>0$, the conditions in Theorem 1 can be written as follows:

$$2\sqrt{2}\cdot\frac{K_1+\rho s}{p\lambda_1(\boldsymbol{M}^*)}\le\min_{i\in J}|u_{1,i}|, \tag{12}$$

$$\rho>2\sqrt{2}\cdot\sqrt{p\sigma^2s^2(d-s)}, \tag{13}$$

$$K_2^2\cdot(1+\sqrt{s})^2\le\Big\{p\lambda_1(\boldsymbol{M}^*)-2\cdot K_1-2\rho s\Big\}\times\Big\{p\lambda_1(\boldsymbol{M}^*)-K_1-K_3-\rho d\Big\}, \tag{14}$$

where $c=1$.

First, when $B\le(2p-1)\lambda_1(\boldsymbol{M}^*)\cdot\max_{i,j\in J}|u_{1,i}u_{1,j}|=(2p-1)\|\boldsymbol{M}^*_{J,J}\|_{\max}$ and $p\ge0.5$, K_1 is expressed as

$$K_1=2p\|\boldsymbol{M}^*_{J,J}\|_{\max}\log(2s)+2\big\{\sqrt{p(1-p)}\|\boldsymbol{M}^*_{J,J}\|_{2,\infty}+\sqrt{ps\sigma^2}\big\}\sqrt{\log(2s)}.$$

Hence, if the following inequalities hold:

$$\begin{aligned}
2p\|\boldsymbol{M}^*_{J,J}\|_{\max}\log(2s)&\le\frac{1}{8\sqrt{2}}\cdot p\lambda_1(\boldsymbol{M}^*)\cdot\min_{i\in J}|u_{1,i}|,\\
2\sqrt{p(1-p)}\|\boldsymbol{M}^*_{J,J}\|_{2,\infty}\cdot\sqrt{\log(2s)}&\le\frac{1}{8\sqrt{2}}\cdot p\lambda_1(\boldsymbol{M}^*)\cdot\min_{i\in J}|u_{1,i}|,\\
2\sqrt{ps\sigma^2}\cdot\sqrt{\log(2s)}&\le\frac{1}{8\sqrt{2}}\cdot p\lambda_1(\boldsymbol{M}^*)\cdot\min_{i\in J}|u_{1,i}|,\\
\rho s&\le\frac{1}{8\sqrt{2}}\cdot p\lambda_1(\boldsymbol{M}^*)\cdot\min_{i\in J}|u_{1,i}|,
\end{aligned}$$

that is, if the following inequalities hold:

$$\|\boldsymbol{M}^*_{J,J}\|_{\max} \le \frac{1}{16\sqrt{2}\log(2s)} \cdot \lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}|, \tag{15}$$

$$\|\boldsymbol{M}^*_{J,J}\|_{2,\infty} \le \frac{1}{16\sqrt{2}\sqrt{\log(2s)}} \cdot \sqrt{\frac{p}{1-p}} \cdot \lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}|, \tag{16}$$

$$2\sqrt{p\sigma^2 s\log(2s)} \le \frac{1}{8\sqrt{2}} \cdot p\lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}|, \tag{17}$$

$$\rho \le \frac{1}{8\sqrt{2}s} \cdot p\lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}|, \tag{18}$$

then (12) holds. Note that (13) and (18) imply that

$$2\sqrt{2} \cdot \sqrt{p\sigma^2 s^2(d-s)} < \frac{1}{8\sqrt{2}s} \cdot p\lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}|.$$

Since $2s^3(d-s) \ge \log(2s)$ for any $s \ge 1$ and $d > s$, (13) and (18) are sufficient for (17).

Now, we will derive the sufficient conditions for (14). First, by the conditions (12) and (18), we have that

$$2K_1 + 2\rho s \le \frac{1}{\sqrt{2}} \cdot p\lambda_1(\boldsymbol{M}^*) \min_{i\in J} |u_{1,i}|,$$

$$K_1 + \rho d = K_1 + \rho s + \rho(d-s) \le \frac{1}{2\sqrt{2}} \cdot p\lambda_1(\boldsymbol{M}^*) \min_{i\in J} |u_{1,i}| + \frac{d-s}{8\sqrt{2}s} \cdot p\lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}|.$$

Also, since

$$2B \le 2(2p-1)\|\boldsymbol{M}^*_{J,J}\|_{\max} \le (2p-1)\frac{1}{8\sqrt{2}\log(2s)} \cdot \lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}| = (2-\frac{1}{p}) \cdot \frac{p\lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}|}{8\sqrt{2}\log(2s)}$$

$$2\sqrt{p\sigma^2} < \frac{1}{16s^2\sqrt{d-s}} \cdot p\lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}|,$$

we can state

$$\begin{aligned} K_2 &= 2B\log d + 2\sqrt{p\sigma^2 \max\{d-s, s\}} \cdot \sqrt{\log d} \\ &\le p\lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}| \cdot \left\{ \underbrace{(2-\frac{1}{p}) \cdot \frac{\log d}{8\sqrt{2}\log(2s)} + \frac{\sqrt{\max\{d-s,s\}} \cdot \sqrt{\log d}}{16s^2\sqrt{d-s}}}_{=:a_1} \right\} \end{aligned}$$

and

$$\begin{aligned} K_3 &= 2B\log(2(d-s)) + 2\sqrt{p\sigma^2(d-s)}\sqrt{\log(2(d-s))} \\ &\le p\lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}| \cdot \left\{ \underbrace{(2-\frac{1}{p}) \cdot \frac{\log(2(d-s))}{8\sqrt{2}\log(2s)} + \frac{\sqrt{\log(2(d-s))}}{16s^2}}_{=:a_2} \right\}. \end{aligned}$$

Therefore, the following inequality is sufficient for (14):

$$\left(p\lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}| \cdot a_1 \cdot (1+\sqrt{s})\right)^2 \le \left\{p\lambda_1(\boldsymbol{M}^*) - \frac{1}{\sqrt{2}} \cdot p\lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}|\right\}$$
$$\times \left\{p\lambda_1(\boldsymbol{M}^*) - \frac{1}{2\sqrt{2}} \cdot p\lambda_1(\boldsymbol{M}^*) \min_{i\in J} |u_{1,i}| - \frac{d-s}{8\sqrt{2}s} \cdot p\lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}| - a_2 \cdot p\lambda_1(\boldsymbol{M}^*) \cdot \min_{i\in J} |u_{1,i}|\right\}$$

$$\Leftrightarrow \left(a_1 \cdot (1+\sqrt{s}) \cdot \min_{i\in J} |u_{1,i}|\right)^2 \le \left\{1 - \frac{1}{\sqrt{2}} \cdot \min_{i\in J} |u_{1,i}|\right\} \times \left\{1 - \left(\frac{1}{2\sqrt{2}} + \frac{d-s}{8\sqrt{2}s} + a_2\right) \cdot \min_{i\in J} |u_{1,i}|\right\}.$$

Note that the quadratic inequality $(ax)^2 \leq (1-bx)(1-cx)$ holds if $a^2 \neq bc$ and $0 \leq x \leq \frac{b+c-\sqrt{(b-c)^2+4a^2}}{2(bc-a^2)}$. By using this fact, we have that if the following inequality holds:

$$\min_{i\in J} |u_{1,i}| \leq \frac{12+\frac{d-s}{s}+8\sqrt{2}a_2-\sqrt{(4-\frac{d-s}{s}-8\sqrt{2}a_2)^2+512a_1^2(1+\sqrt{s})^2}}{4\sqrt{2}+\sqrt{2}\cdot\frac{d-s}{s}+16a_2-16\sqrt{2}a_1^2(1+\sqrt{s})^2},$$

then (14) holds. Since $\min_{i\in J}|u_{1,i}| \leq \frac{1}{\sqrt{s}}$, the following inequality is sufficient for (14):

$$\frac{1}{\sqrt{s}} \leq \frac{12+\frac{d-s}{s}+8\sqrt{2}a_2-\sqrt{(4-\frac{d-s}{s}-8\sqrt{2}a_2)^2+512a_1^2(1+\sqrt{s})^2}}{4\sqrt{2}+\sqrt{2}\cdot\frac{d-s}{s}+16a_2-16\sqrt{2}a_1^2(1+\sqrt{s})^2}.$$

In sum, if the following inequalities hold:

$$\begin{aligned}
\|\boldsymbol{M}^*_{J,J}\|_{\max} &\leq \frac{1}{16\sqrt{2}\log(2s)}\cdot\lambda_1(\boldsymbol{M}^*)\cdot\min_{i\in J}|u_{1,i}|,\\
\|\boldsymbol{M}^*_{J,J}\|_{2,\infty} &\leq \frac{1}{16\sqrt{2}\sqrt{\log(2s)}}\cdot\sqrt{\frac{p}{1-p}}\cdot\lambda_1(\boldsymbol{M}^*)\cdot\min_{i\in J}|u_{1,i}|,\\
B &\leq (2p-1)\lambda_1(\boldsymbol{M}^*)\cdot\max_{i,j\in J}|u_{1,i}u_{1,j}|,\\
\rho &\leq \frac{1}{8\sqrt{2}s}\cdot p\lambda_1(\boldsymbol{M}^*)\cdot\min_{i\in J}|u_{1,i}|,\\
\rho &> 2\sqrt{2}\cdot\sqrt{p\sigma^2 s^2(d-s)},\\
\frac{1}{\sqrt{s}} &\leq \frac{12+\frac{d-s}{s}+8\sqrt{2}a_2-\sqrt{(4-\frac{d-s}{s}-8\sqrt{2}a_2)^2+512a_1^2(1+\sqrt{s})^2}}{4\sqrt{2}+\sqrt{2}\cdot\frac{d-s}{s}+16a_2-16\sqrt{2}a_1^2(1+\sqrt{s})^2},
\end{aligned}$$

then the desired result holds, where $a_1 = (2-\frac{1}{p})\cdot\frac{\log d}{8\sqrt{2}\log(2s)}+\frac{\sqrt{\max\{d-s,s\}}\cdot\sqrt{\log d}}{16s^2\sqrt{d-s}}$ and $a_2 = (2-\frac{1}{p})\cdot\frac{\log(2(d-s))}{8\sqrt{2}\log(2s)}+\frac{\sqrt{\log(2(d-s))}}{16s^2}$. Since $\|\boldsymbol{M}^*\|_{\max} = \lambda_1(\boldsymbol{M}^*)\max_{i,j\in J}|u_{1,i}u_{1,j}|$ and $\|\boldsymbol{M}^*\|_{2,\infty} = \lambda_1(\boldsymbol{M}^*)\max_{i\in J}\sqrt{\sum_{j\in J}u_{1,i}^2u_{1,j}^2} = \lambda_1(\boldsymbol{M}^*)\max_{i\in J}|u_{1,i}|$, the first two conditions can be written as $\frac{\max_{i,j\in J}|u_{1,i}u_{1,j}|}{\min_{i\in J}|u_{1,i}|} \leq \frac{1}{16\sqrt{2}\log(2s)}$ and $\frac{\max_{i\in J}|u_{1,i}|}{\min_{i\in J}|u_{1,i}|} \leq \frac{1}{16\sqrt{2}\sqrt{\log(2s)}}\cdot\sqrt{\frac{p}{1-p}}$.

E Other Proofs

E.1 Proof of Proposition 1

With the primal variable $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ and the dual variables $\boldsymbol{Z} \in \mathbb{R}^{d\times d}$, $\boldsymbol{\Lambda} \in \mathbb{R}^{d\times d}$ and $\mu \in \mathbb{R}$, the Lagrangian of the problem (1) is written as

$$L(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Lambda},\mu) = -\langle\boldsymbol{M},\boldsymbol{X}\rangle + \rho\langle\boldsymbol{X},\boldsymbol{Z}\rangle - \langle\boldsymbol{\Lambda},\boldsymbol{X}\rangle + \mu\cdot(tr(\boldsymbol{X})-1)$$

where $Z_{ij} \in \partial|X_{ij}|$ for each $i,j \in [d]$. According to the standard KKT condition, we can derive that $(\hat{\boldsymbol{X}},\hat{\boldsymbol{Z}},\hat{\boldsymbol{\Lambda}},\hat{\mu})$ is optimal if and only if the followings hold:

- Primal feasibility: $\hat{\boldsymbol{X}} \succeq 0$, $tr(\hat{\boldsymbol{X}}) = 1$
- Dual feasibility: $\hat{\boldsymbol{\Lambda}} \succeq 0$, $\hat{Z}_{ij} \in \partial|\hat{X}_{ij}|$ for each $i,j \in [d]$
- Complementary slackness: $\langle\hat{\boldsymbol{\Lambda}},\hat{\boldsymbol{X}}\rangle = 0$ ($\Leftrightarrow \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{X}} = 0$ if $\hat{\boldsymbol{X}} \succeq 0$ and $\hat{\boldsymbol{\Lambda}} \succeq 0$)
- Stationarity: $\hat{\boldsymbol{\Lambda}} = -\boldsymbol{M} + \rho\hat{\boldsymbol{Z}} + \hat{\mu}\cdot\boldsymbol{I}$.

By substituting $\hat{\boldsymbol{\Lambda}}$ with $-\boldsymbol{M} + \rho\hat{\boldsymbol{Z}} + \hat{\mu} \cdot \boldsymbol{I}$, it can be shown that the above conditions are equivalent to

$$\begin{aligned}
&\hat{\boldsymbol{X}} \succeq 0, tr(\hat{\boldsymbol{X}}) = 1 \\
&\boldsymbol{M} - \rho\hat{\boldsymbol{Z}} \preceq \hat{\mu}\boldsymbol{I} \\
&\hat{Z}_{ij} \in \partial|\hat{X}_{ij}| \quad \text{for each } i, j \in [d] \\
&(\boldsymbol{M} - \rho\hat{\boldsymbol{Z}})\hat{\boldsymbol{X}} = \hat{\mu} \cdot \hat{\boldsymbol{X}}.
\end{aligned}$$

To use the primal-dual witness construction, we now consider the following restricted problem:

$$\max_{\boldsymbol{X} \succeq 0, tr(\boldsymbol{X})=1 \text{ and } supp(\boldsymbol{X}) \subseteq J \times J} \langle \boldsymbol{M}, \boldsymbol{X} \rangle - \rho \|\boldsymbol{X}\|_{1,1}. \tag{19}$$

Similarly to the above, we can derive that $\hat{\boldsymbol{X}} = \begin{pmatrix} \hat{\boldsymbol{X}}_{J,J} & 0 \\ 0 & 0 \end{pmatrix}$ is optimal to the problem (19) if and only if

$$\begin{aligned}
&\hat{\boldsymbol{X}}_{J,J} \succeq 0, tr(\hat{\boldsymbol{X}}_{J,J}) = 1 \\
&\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{Z}}_{J,J} \preceq \hat{\mu}\boldsymbol{I} \\
&\hat{Z}_{ij} \in \partial|\hat{X}_{ij}| \quad \text{for each } i, j \in J \\
&(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{Z}}_{J,J})\hat{\boldsymbol{X}}_{J,J} = \hat{\mu} \cdot \hat{\boldsymbol{X}}_{J,J}.
\end{aligned}$$

Now, we want for the above solution $\hat{\boldsymbol{X}} = \begin{pmatrix} \hat{\boldsymbol{X}}_{J,J} & 0 \\ 0 & 0 \end{pmatrix}$ to satisfy the optimality conditions of the original problem (1). Furthermore, by assuming the strict dual feasibility, we want to guarantee $supp(diag(\hat{\boldsymbol{X}})) \subseteq J$. We can easily derive their sufficient conditions listed below:

$$\begin{aligned}
&\hat{\boldsymbol{X}}_{J,J} \succeq 0, tr(\hat{\boldsymbol{X}}_{J,J}) = 1 \\
&\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{Z}}_{J,J} \preceq \hat{\mu}\boldsymbol{I} \\
&\boldsymbol{M} - \rho\hat{\boldsymbol{Z}} \preceq \hat{\mu}\boldsymbol{I} \\
&\hat{Z}_{ij} \in \partial|\hat{X}_{ij}| \quad \text{for each } (i, j) \in J \times J \\
&\hat{Z}_{ij} \in (-1, 1) \quad \text{for each } (i, j) \notin J \times J \\
&(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{Z}}_{J,J})\hat{\boldsymbol{X}}_{J,J} = \hat{\mu} \cdot \hat{\boldsymbol{X}}_{J,J} \\
&(\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{Z}}_{J^c,J})\hat{\boldsymbol{X}}_{J,J} = 0.
\end{aligned}$$

If the above conditions hold, then $\hat{\boldsymbol{X}} = \begin{pmatrix} \hat{\boldsymbol{X}}_{J,J} & 0 \\ 0 & 0 \end{pmatrix}$ is optimal to the problem (1) and satisfies $supp(diag(\hat{\boldsymbol{X}})) \subseteq J$.

E.2 Proof of Proposition 2

With the primal variable $\boldsymbol{x} \in \mathbb{R}^d$ and the dual variables $\boldsymbol{z} \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$, the Lagrangian of the problem (2) is written as

$$L(\boldsymbol{x}, \boldsymbol{z}, \lambda) = -\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} + \rho\langle \boldsymbol{x}, \boldsymbol{z} \rangle^2 + \lambda(\boldsymbol{x}^\top \boldsymbol{x} - 1) = \boldsymbol{x}^\top(-\boldsymbol{M} + \rho\boldsymbol{z}\boldsymbol{z}^\top + \lambda\boldsymbol{I})\boldsymbol{x} - \lambda$$

where $z_i \in \partial|x_i|$ for $i \in [d]$. By denoting the primal solution by $\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{x}}_1^\top, \tilde{\boldsymbol{x}}_2^\top)^\top \in \mathbb{R}^s \times \mathbb{R}^{d-s}$ and the dual solutions by $(\tilde{\boldsymbol{z}}^\top, \tilde{\boldsymbol{w}}^\top)^\top \in \mathbb{R}^s \times \mathbb{R}^{d-s}$ and $\tilde{\lambda} \in \mathbb{R}$, the KKT conditions of (2) are given as follows:

- Primal feasibility: $\tilde{\boldsymbol{x}}_1^\top \tilde{\boldsymbol{x}}_1 + \tilde{\boldsymbol{x}}_2^\top \tilde{\boldsymbol{x}}_2 = 1$
- Dual feasibility: $\begin{cases} \tilde{z}_i = \text{sign}(\tilde{x}_{1,i}) & \text{if } \tilde{x}_{1,i} \neq 0 \\ \tilde{z}_i \in [-1, 1] & \text{if } \tilde{x}_{1,i} = 0, \end{cases} \begin{cases} \tilde{w}_i = \text{sign}(\tilde{x}_{2,i}) & \text{if } \tilde{x}_{2,i} \neq 0 \\ \tilde{w}_i \in [-1, 1] & \text{if } \tilde{x}_{2,i} = 0 \end{cases}$

- Stationarity: $\left[\boldsymbol{M} - \rho(\tilde{\boldsymbol{z}}^\top, \tilde{\boldsymbol{w}}^\top)^\top(\tilde{\boldsymbol{z}}^\top, \tilde{\boldsymbol{w}}^\top)\right]\begin{pmatrix}\tilde{\boldsymbol{x}}_1\\ \tilde{\boldsymbol{x}}_2\end{pmatrix} = \tilde{\lambda}\begin{pmatrix}\tilde{\boldsymbol{x}}_1\\ \tilde{\boldsymbol{x}}_2\end{pmatrix}$
  $\tilde{\lambda}\boldsymbol{I} \succeq \boldsymbol{M} - \rho(\tilde{\boldsymbol{z}}^\top, \tilde{\boldsymbol{w}}^\top)^\top(\tilde{\boldsymbol{z}}^\top, \tilde{\boldsymbol{w}}^\top)$.

Here, it can be easily checked that the primal feasibility and the stationarity conditions are equivalent to the following:

$$\begin{pmatrix}\tilde{\boldsymbol{x}}_1\\ \tilde{\boldsymbol{x}}_2\end{pmatrix} \text{ is the leading eigenvector of } \boldsymbol{M} - \rho(\tilde{\boldsymbol{z}}^\top, \tilde{\boldsymbol{w}}^\top)^\top(\tilde{\boldsymbol{z}}^\top, \tilde{\boldsymbol{w}}^\top).$$

To proceed with primal-dual witness argument, we now consider the KKT conditions for the problem (2) with an additional constraint $supp((\boldsymbol{x}^\top, \boldsymbol{y}^\top)^\top) \subseteq J$, that is, $\boldsymbol{y} = \boldsymbol{0}$. With the primal solution $\hat{\boldsymbol{x}} \in \mathbb{R}^s$ and the dual solution $\hat{\boldsymbol{z}} \in \mathbb{R}^s$, the KKT conditions are given by

$$\begin{aligned}&\hat{\boldsymbol{x}} \text{ is the leading eigenvector of } \boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top\\ &\hat{z}_i = \text{sign}(\hat{x}_i) \quad \text{if } \hat{x}_i \neq 0, \;\; \hat{z}_i \in [-1, 1] \quad \text{if } \hat{x}_i = 0\end{aligned}$$

Now we will show that if the following conditions hold, the solution $(\hat{\boldsymbol{x}}^\top, \boldsymbol{0}^\top)^\top$ satisfies the KKT conditions of the original problem (2) and the strict dual feasibility:

$$\hat{\boldsymbol{x}} \text{ is the leading eigenvector of } \boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top \tag{20}$$

$$\hat{z}_i = \text{sign}(\hat{x}_i) \quad \text{if } \hat{x}_i \neq 0, \;\; \hat{z}_i \in [-1, 1] \quad \text{if } \hat{x}_i = 0$$

$$\hat{\boldsymbol{w}} = \frac{1}{\rho\|\hat{\boldsymbol{x}}\|_1}\boldsymbol{M}_{J^c,J}\hat{\boldsymbol{x}} \tag{21}$$

$$\|\hat{\boldsymbol{w}}\|_\infty < 1.$$

$$\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) = \lambda_1(\boldsymbol{M} - \rho(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)^\top(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)).$$

Let $\hat{\lambda} = \lambda_1\big(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top\big) = \lambda_1\big(\boldsymbol{M} - \rho(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)^\top(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)\big)$. For $(\hat{\boldsymbol{x}}^\top, \boldsymbol{0}^\top)^\top$ to be the leading eigenvector of $\boldsymbol{M} - \rho(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)^\top(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)$,

$$\left[\boldsymbol{M} - \rho(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)^\top(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)\right]\begin{pmatrix}\hat{\boldsymbol{x}}\\ \boldsymbol{0}\end{pmatrix} = \hat{\lambda}\begin{pmatrix}\hat{\boldsymbol{x}}\\ \boldsymbol{0}\end{pmatrix} \tag{22}$$

needs to be satisfied. It can be easily checked that (22) is equivalent to (20) and (21).

Now, let $\hat{\boldsymbol{X}} = \begin{pmatrix}\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top & 0\\ 0 & 0\end{pmatrix}$, $\hat{\boldsymbol{Z}} = (\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)^\top(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)$ and $\hat{\mu} = \lambda_1\big(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top\big) = \lambda_1\big(\boldsymbol{M} - \rho(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)^\top(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)\big)$. Then it can be easily shown that $(\hat{\boldsymbol{X}}, \hat{\boldsymbol{Z}}, \hat{\mu})$ satisfies the sufficient conditions in Proposition 1. That is, $\hat{\boldsymbol{X}}$ constructed above is an optimal solution to the problem (1) and satisfies $supp(diag(\hat{\boldsymbol{X}})) \subseteq J$. To ensure that there is no false positive, we consider the additional condition that $\text{sign}(\hat{x}_i) = \text{sign}(u_{1,i})$ for all $i \in J$.

Lastly, for the uniqueness, we need an additional condition presented in the following lemma.

Lemma 5. *For* $\hat{\boldsymbol{X}} = \begin{pmatrix}\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top & 0\\ 0 & 0\end{pmatrix}$ *and* $\hat{\boldsymbol{Z}} = (\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)^\top(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)$ *constructed above, if the following condition holds:*

$$\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) > \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)$$

then the solution $\hat{\boldsymbol{X}}$ *is a unique optimal solution to the problem (1).*

Proof. According to the standard primal-dual witness construction, we only need to show that under the condition, $\hat{\boldsymbol{X}}_{J,J} = \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top$ is a unique optimal solution to the restricted problem (19).

Assume that there exists another optimal solution to the problem (19), say $\tilde{\boldsymbol{X}}_{J,J}$. Also, denote its dual optimal solution by $\tilde{\boldsymbol{Z}}_{J,J}$. Then, we can write

$$\begin{aligned}&\langle \boldsymbol{M}_{J,J}, \hat{\boldsymbol{X}}_{J,J}\rangle - \rho\|\hat{\boldsymbol{X}}_{J,J}\|_{1,1} = \langle \boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top\rangle = \hat{\boldsymbol{x}}^\top(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)\hat{\boldsymbol{x}}\\ &= \langle \boldsymbol{M}_{J,J}, \tilde{\boldsymbol{X}}_{J,J}\rangle - \rho\|\tilde{\boldsymbol{X}}_{J,J}\|_{1,1} = \langle \boldsymbol{M}_{J,J} - \rho\tilde{\boldsymbol{Z}}_{J,J}, \tilde{\boldsymbol{X}}_{J,J}\rangle.\end{aligned}$$

Recall that $\hat{\boldsymbol{x}}$ is the leading eigenvector of $\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top$, that is, $\hat{\boldsymbol{x}}^\top(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)\hat{\boldsymbol{x}} = \lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)$. Now, we will show that $\langle \boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top, \tilde{\boldsymbol{X}}_{J,J}\rangle < \lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)$ for any matrix $\tilde{\boldsymbol{X}}_{J,J} \neq \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top$ such that $\tilde{\boldsymbol{X}}_{J,J} \succeq 0$ and $tr(\tilde{\boldsymbol{X}}_{J,J}) = 1$. Let $\tilde{\boldsymbol{X}}_{J,J} = \sum_{i\in J}\theta_i \boldsymbol{v}_i\boldsymbol{v}_i^\top$, which is the spectral decomposition of $\tilde{\boldsymbol{X}}_{J,J}$. We can derive that

$$\langle \boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top, \tilde{\boldsymbol{X}}_{J,J}\rangle = \langle \boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top, \sum_{i\in J}\theta_i\boldsymbol{v}_i\boldsymbol{v}_i^\top\rangle = \sum_{i\in J}\theta_i\boldsymbol{v}_i^\top(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)\boldsymbol{v}_i \leq \lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)$$

where the last inequality holds since $\sum_{i\in J}\theta_i = tr(\tilde{\boldsymbol{X}}_{J,J}) = 1$ and $\boldsymbol{v}_i^\top(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)\boldsymbol{v}_i \leq \lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)$. Here, the equality holds only if $\theta_1 = 1$, $\theta_i = 0$ for $i \neq 1$ and $\boldsymbol{v}_1 = \hat{\boldsymbol{x}}$, that is, $\tilde{\boldsymbol{X}}_{J,J} = \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top$. Therefore, $\langle \boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top, \tilde{\boldsymbol{X}}_{J,J}\rangle < \lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)$ for any matrix $\tilde{\boldsymbol{X}}_{J,J} \neq \hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\top$ such that $\tilde{\boldsymbol{X}}_{J,J} \succeq 0$ and $tr(\tilde{\boldsymbol{X}}_{J,J}) = 1$.

With this fact, we can derive that

$$\begin{aligned}
\langle \boldsymbol{M}_{J,J}, \hat{\boldsymbol{X}}_{J,J}\rangle - \rho\|\hat{\boldsymbol{X}}_{J,J}\|_{1,1} &= \hat{\boldsymbol{x}}^\top(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)\hat{\boldsymbol{x}} = \lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) \\
&> \langle \boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top, \tilde{\boldsymbol{X}}_{J,J}\rangle = \langle \boldsymbol{M}_{J,J} - \rho\tilde{\boldsymbol{Z}}_{J,J}, \tilde{\boldsymbol{X}}_{J,J}\rangle + \rho\langle \tilde{\boldsymbol{Z}}_{J,J} - \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top, \tilde{\boldsymbol{X}}_{J,J}\rangle \\
&= \langle \boldsymbol{M}_{J,J}, \tilde{\boldsymbol{X}}_{J,J}\rangle - \rho\|\tilde{\boldsymbol{X}}_{J,J}\|_{1,1} + \rho\langle \tilde{\boldsymbol{Z}}_{J,J} - \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top, \tilde{\boldsymbol{X}}_{J,J}\rangle.
\end{aligned}$$

Since $\langle \boldsymbol{M}_{J,J}, \hat{\boldsymbol{X}}_{J,J}\rangle - \rho\|\hat{\boldsymbol{X}}_{J,J}\|_{1,1} = \langle \boldsymbol{M}_{J,J}, \tilde{\boldsymbol{X}}_{J,J}\rangle - \rho\|\tilde{\boldsymbol{X}}_{J,J}\|_{1,1}$ by assumption, the above inequality implies $\langle \tilde{\boldsymbol{Z}}_{J,J} - \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top, \tilde{\boldsymbol{X}}_{J,J}\rangle < 0$, that is, $\langle \tilde{\boldsymbol{Z}}_{J,J}, \tilde{\boldsymbol{X}}_{J,J}\rangle < \langle \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top, \tilde{\boldsymbol{X}}_{J,J}\rangle$. This contradicts the fact that $\langle \tilde{\boldsymbol{Z}}_{J,J}, \tilde{\boldsymbol{X}}_{J,J}\rangle = \sup_{\|\boldsymbol{Z}_{J,J}\|_{\max}\leq 1}\langle \boldsymbol{Z}_{J,J}, \tilde{\boldsymbol{X}}_{J,J}\rangle$, and thus the desired result holds.

□

E.3 Lemma 6

The following lemma is frequently used in the proofs.

Lemma 6. *For any $c > 0$,*

$$\begin{aligned}
\|\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J}\|_2 &\leq (c+1)\cdot R_1\log(2s) + \sqrt{2(c+1)}\cdot R_2\sqrt{\log(2s)} =: K_1 \\
&\qquad \textit{with probability at least } 1-(2s)^{-c}, \\
\|\mathbb{E}[\boldsymbol{M}_{J^c,J}] - \boldsymbol{M}_{J^c,J}\|_2 &\leq (c+1)\cdot R_3\log d + \sqrt{2(c+1)}\cdot R_4\sqrt{\log d} =: K_2 \\
&\qquad \textit{with probability at least } 1-d^{-c}, \\
\|\mathbb{E}[\boldsymbol{M}_{J^c,J^c}] - \boldsymbol{M}_{J^c,J^c}\|_2 &\leq (c+1)\cdot R_5\log(2(d-s)) + \sqrt{2(c+1)}\cdot R_6\sqrt{\log(2(d-s))} =: K_3 \\
&\qquad \textit{with probability at least } 1-(2(d-s))^{-c}
\end{aligned}$$

where

$$\begin{aligned}
R_1 &:= \max\{(1-p)\|\boldsymbol{M}^*_{J,J}\|_{\max} + B,\ \ p\|\boldsymbol{M}^*_{J,J}\|_{\max}\}, \\
R_2 &:= \sqrt{p(1-p)}\|\boldsymbol{M}^*_{J,J}\|_{2,\infty} + \sqrt{ps\sigma^2}, \\
R_3 &:= \max\{(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_{\max} + B,\ \ p\|\boldsymbol{M}^*_{J^c,J}\|_{\max}\}, \\
R_4 &:= \max\{\sqrt{p(1-p)}\|\boldsymbol{M}^*_{J^c,J}\|_{2,\infty} + \sqrt{p(d-s)\sigma^2}, \sqrt{p(1-p)}\|\boldsymbol{M}^*_{J,J^c}\|_{2,\infty} + \sqrt{ps\sigma^2}\}, \\
R_5 &:= \max\{(1-p)\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max} + B,\ \ p\|\boldsymbol{M}^*_{J^c,J^c}\|_{\max}\}, \\
R_6 &:= \sqrt{p(1-p)}\|\boldsymbol{M}^*_{J^c,J^c}\|_{2,\infty} + \sqrt{p(d-s)\sigma^2}.
\end{aligned}$$

Proof. We use the matrix Bernstein inequality presented in Theorem 2. Here, we only show the upper bound of $\|\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J}\|_2$, since the others can be derived similarly. Note that

$$\begin{aligned}\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J} &= \sum_{i,j\in J} (\mathbb{E}[M_{i,j}] - M_{i,j})\boldsymbol{e}_i\boldsymbol{e}_j^\top = \sum_{i,j\in J} (p\cdot M^*_{i,j} - \delta_{i,j}(M^*_{i,j} + \epsilon_{i,j}))\boldsymbol{e}_i\boldsymbol{e}_j^\top \\ &= \sum_{i,j\in J} \underbrace{((p-\delta_{i,j})\cdot M^*_{i,j} - \delta_{i,j}\epsilon_{i,j})}_{=:a_{i,j}}\boldsymbol{e}_i\boldsymbol{e}_j^\top \\ &= \sum_{i,j\in J, i<j} a_{i,j}(\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top) + \sum_{i\in J} a_{i,i}\boldsymbol{e}_i\boldsymbol{e}_i^\top,\end{aligned}$$

which can be viewed as a sum of independent zero-mean matrices. It is straightforward to conclude that

$$\begin{aligned}&\max\{\|a_{i,j}(\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top)\|_2,\ \|a_{i,i}\boldsymbol{e}_i\boldsymbol{e}_i^\top\|_2\ :\ i,j\in J,\ i<j\} \\ &\qquad\qquad \le \max\{(1-p)\|\boldsymbol{M}^*_{J,J}\|_{\max} + B,\ \ p\|\boldsymbol{M}^*_{J,J}\|_{\max}\} = R_1\ \ a.s.,\end{aligned}$$

$$\left\|\sum_{i,j\in J, i<j} \mathbb{E}[a_{i,j}(\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top)]^2 + \sum_{i\in J}\mathbb{E}[a_{i,i}\boldsymbol{e}_i\boldsymbol{e}_i^\top]^2\right\|_2 \le p(1-p)\|\boldsymbol{M}^*_{J,J}\|^2_{2,\infty} + ps\sigma^2 \le R_2^2.$$

By the matrix Bernstein inequality, one has that with probability at least $1-(2s)^{-c}$,

$$\|\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J}\|_2 \le (c+1)\cdot R_1\log(2s) + \sqrt{2(c+1)}\cdot R_2\sqrt{\log(2s)}.$$

□

Theorem 2 (Matrix Bernstein inequality (e.g., Theorem 1.6 in Tropp [2012])). *Consider a finite sequence $\{\boldsymbol{Z}_k\}$ of independent, random matrices with dimensions $d_1 \times d_2$. Assume that each random matrix satisfies*

$$\mathbb{E}[\boldsymbol{Z}_k] = \boldsymbol{0} \quad \textit{and} \quad \|\boldsymbol{Z}_k\|_2 \le R_1 \textit{ almost surely.}$$

Also, suppose that

$$\max\left\{\left\|\sum_k \mathbb{E}[\boldsymbol{Z}_k\boldsymbol{Z}_k^\top]\right\|_2, \left\|\sum_k \mathbb{E}[\boldsymbol{Z}_k^\top\boldsymbol{Z}_k]\right\|_2\right\} \le R_2^2.$$

Then, for all $t \ge 0$,

$$\mathbb{P}\left\{\left\|\sum_k \boldsymbol{Z}_k\right\|_2 \ge t\right\} \le (d_1+d_2)\cdot\exp\left(\frac{-t^2/2}{R_1 t/3 + R_2^2}\right).$$

The above inequality implies that

$$\begin{aligned}\left\|\sum_k \boldsymbol{Z}_k\right\|_2 &\le \frac{(c+1)R_1}{3}\log(d_1+d_2) + \sqrt{\left\{\frac{(c+1)R_1}{3}\log(d_1+d_2)\right\}^2 + 2(c+1)R_2^2\log(d_1+d_2)} \\ &\le (c+1)R_1\log(d_1+d_2) + \sqrt{2(c+1)}R_2\sqrt{\log(d_1+d_2)}\end{aligned}$$

with probability at least $1-(d_1+d_2)^{-c}$ for any $c > 0$.

E.4 Proof of Lemma 1

First, the following lemma can be easily shown.

Lemma 7. *For any unit vectors $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathbb{R}^d$ such that $y_i \ne 0$ for $\forall i \in [d]$, if $\|\boldsymbol{x}-\boldsymbol{y}\|_2 \le \min_{i\in[d]}|y_i|$, then $sign(x_i) = sign(y_i)$ for $\forall i \in [d]$.*

Proof. If $\boldsymbol{x} = \boldsymbol{y}$, then it is trivial that $sign(x_i) = sign(y_i)$ for $\forall i \in [d]$. If $\boldsymbol{x} \ne \boldsymbol{y}$, then for any $i \in [d]$,

$$|x_i - y_i| < \|\boldsymbol{x}-\boldsymbol{y}\|_2 \le \min_{i\in[d]}|y_i| \le |y_i|,$$

where the first inequality is strict since both $\boldsymbol{x}$ and $\boldsymbol{y}$ are unit vectors. The above inequality implies that

$$y_i - |y_i| < x_i < y_i + |y_i|,$$

that is, $0 < x_i < 2y_i$ if $y_i > 0$, and $2y_i < x_i < 0$ if $y_i < 0$. Therefore, $sign(x_i) = sign(y_i)$ holds for any $i \in [d]$. □

Now, let $\hat{z}_i = \text{sign}(u_{1,i})$ for all $i \in J$ and $\hat{\boldsymbol{x}}$ be the leading eigenvector of $\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top$. We will derive the upper bound of $\|\boldsymbol{u}_1 - \hat{\boldsymbol{x}}\|_2$. By applying the Davis-Kahan sinΘ theorem, we obtain

$$\|\boldsymbol{u}_1 - \hat{\boldsymbol{x}}\|_2 \leq \frac{2\sqrt{2}}{p(\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J}))}\|\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J} + \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top\|_2$$

where $\mathbb{E}[\boldsymbol{M}_{J,J}] = p \cdot \boldsymbol{M}^*_{J,J}$. By the triangle inequality, we can upper bound

$$\|\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J} + \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top\|_2 \leq \|\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J}\|_2 + \|\rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top\|_2 = \|\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J}\|_2 + \rho s.$$

From Lemma 6, one has that with probability at least $1 - (2s)^{-c}$,

$$\|\boldsymbol{u}_1 - \hat{\boldsymbol{x}}\|_2 \leq 2\sqrt{2} \cdot \frac{K_1 + \rho s}{p(\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J}))}.$$

By Lemma 7, if

$$2\sqrt{2} \cdot \frac{K_1 + \rho s}{p(\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J}))} \leq \min_{i \in J} |u_{1,i}|,$$

then $sign(\hat{x}_i) = sign(u_{1,i})$ for all $i \in J$ with probability at least $1 - (2s)^{-c}$.

E.5 Proof of Lemma 2

First, we can derive the upper bound of $\|\hat{\boldsymbol{w}}\|_\infty$ as follows:

$$\begin{aligned}\|\hat{\boldsymbol{w}}\|_\infty &= \frac{1}{\rho\|\hat{\boldsymbol{x}}\|_1}\|\boldsymbol{M}_{J^c,J}\hat{\boldsymbol{x}}\|_\infty = \frac{1}{\rho\|\hat{\boldsymbol{x}}\|_1} \cdot \max_{i \in J^c}\left|\sum_{j \in J} M_{i,j}\hat{x}_j\right| \\ &\leq \frac{1}{\rho\|\hat{\boldsymbol{x}}\|_1} \cdot \left(\max_{i \in J^c}\max_{j \in J}|M_{i,j}|\right) \cdot \sum_{j \in J}|\hat{x}_j| = \frac{1}{\rho} \cdot \|\boldsymbol{M}_{J^c,J}\|_{\max}.\end{aligned}$$

For each $M_{i,j}$, $i \in J^c$ and $j \in J$, we now apply Chebyshev's inequality as follows:

$$\mathbb{P}\big(|M_{i,j}| \geq |\mathbb{E}[M_{i,j}]| + c\big) \leq \mathbb{P}\big(|M_{i,j} - \mathbb{E}[M_{i,j}]| \geq c\big) \leq \frac{\mathsf{Var}[M_{i,j}]}{c^2} \tag{23}$$

for any $c > 0$. Note that for each $i \in J^c$ and $j \in J$,

$$\mathbb{E}[M_{i,j}] = p \cdot M^*_{i,j} \text{ and } \mathsf{Var}[M_{i,j}] = p(1-p)(M^*_{i,j})^2 + p\sigma^2.$$

With the assumption that $p \cdot \|\boldsymbol{M}^*_{J^c,J}\|_{\max} < \rho$, letting $\gamma := \frac{p}{\rho}\|\boldsymbol{M}^*_{J^c,J}\|_{\max}$ and $c := \frac{1+\gamma}{2} \cdot \rho - p \cdot |M^*_{i,j}|$ yields that $\gamma < 1$ and $c > 0$. By plugging c into (23), we have that

$$\begin{aligned}\mathbb{P}\left(|M_{i,j}| \geq \frac{1+\gamma}{2} \cdot \rho\right) &\leq \frac{p(1-p)(M^*_{i,j})^2 + p\sigma^2}{\left(\frac{1+\gamma}{2} \cdot \rho - p \cdot |M^*_{i,j}|\right)^2} \leq \frac{p(1-p)(M^*_{i,j})^2 + p\sigma^2}{\left(\frac{1+\gamma}{2} \cdot \rho - p \cdot \|\boldsymbol{M}^*_{J^c,J}\|_{\max}\right)^2} \\ &= \frac{p(1-p)(M^*_{i,j})^2 + p\sigma^2}{\left(\frac{1+\gamma}{2} \cdot \rho - \gamma\rho\right)^2} = \frac{p(1-p)(M^*_{i,j})^2 + p\sigma^2}{\left(\frac{1-\gamma}{2} \cdot \rho\right)^2}\end{aligned}$$

for each $i \in J^c$ and $j \in J$. Hence,

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{M}_{J^c,J}\|_{\max} \geq \frac{1+\gamma}{2} \cdot \rho\right) &\leq \sum_{i \in J^c}\sum_{j \in J}\frac{p(1-p)(M^*_{i,j})^2 + p\sigma^2}{\left(\frac{1-\gamma}{2} \cdot \rho\right)^2} \\ &= \frac{4}{(1-\gamma)^2\rho^2} \cdot \left\{p(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_F^2 + p(d-s)s\sigma^2\right\} \\ &= \frac{4}{\left(\rho - p \cdot \|\boldsymbol{M}^*_{J^c,J}\|_{\max}\right)^2} \cdot \left\{p(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_F^2 + p(d-s)s\sigma^2\right\}.\end{aligned}$$

Since $\gamma < 1$, we have that

$$\mathbb{P}\left(\|\boldsymbol{M}_{J^c,J}\|_{\max} < \rho\right) \geq 1 - \frac{4}{\left(\rho - p \cdot \|\boldsymbol{M}^*_{J^c,J}\|_{\max}\right)^2} \cdot \left\{p(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_F^2 + p(d-s)s\sigma^2\right\}.$$

If the following inequality holds:

$$\sqrt{4ps^c \cdot \left\{(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_F^2 + (d-s)s\sigma^2\right\}} < \rho - p \cdot \|\boldsymbol{M}^*_{J^c,J}\|_{\max},$$

then $\frac{4}{\left(\rho - p\cdot\|\boldsymbol{M}^*_{J^c,J}\|_{\max}\right)^2} \cdot \left\{p(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_F^2 + p(d-s)s\sigma^2\right\} \leq s^{-c}$, that is, $\|\boldsymbol{M}_{J^c,J}\|_{\max} < \rho$ holds with probability at least $1 - s^{-c}$. $\|\hat{\boldsymbol{w}}\|_\infty \leq \frac{1}{\rho}\|\boldsymbol{M}_{J^c,J}\|_{\max}$, and thus the desired result holds.

E.6 Proof of Lemma 3

Lemma 8 shows that if the following inequality holds:

$$\|\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top\|_2^2 \leq \left\{\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)\right\} \cdot \left\{\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \lambda_1(\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top)\right\},$$

then $\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) = \lambda_1(\boldsymbol{M} - \rho(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)^\top(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top))$.

Now, we derive the upper or lower bounds of $\|\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top\|_2$, $\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)$ and $\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \lambda_1(\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top)$. First,

$$\begin{aligned}
\|\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top\|_2 &= \left\|\boldsymbol{M}_{J^c,J} - \rho \cdot \frac{1}{\rho\|\hat{\boldsymbol{x}}\|_1}\boldsymbol{M}_{J^c,J}\hat{\boldsymbol{x}}\hat{\boldsymbol{z}}^\top\right\|_2 = \left\|\boldsymbol{M}_{J^c,J} \cdot \left(I - \frac{\hat{\boldsymbol{x}}\hat{\boldsymbol{z}}^\top}{\|\hat{\boldsymbol{x}}\|_1}\right)\right\|_2 \\
&\leq \|\boldsymbol{M}_{J^c,J}\|_2 \cdot \left\|I - \frac{\hat{\boldsymbol{x}}\hat{\boldsymbol{z}}^\top}{\|\hat{\boldsymbol{x}}\|_1}\right\|_2 \leq \|\boldsymbol{M}_{J^c,J}\|_2 \cdot \left(1 + \frac{\|\hat{\boldsymbol{x}}\|_2\|\hat{\boldsymbol{z}}\|_2}{\|\hat{\boldsymbol{x}}\|_1}\right) \\
&\leq \|\boldsymbol{M}_{J^c,J}\|_2 \cdot (1 + \sqrt{s}).
\end{aligned}$$

Since $\|\boldsymbol{M}_{J^c,J}\|_2 \leq \|\mathbb{E}[\boldsymbol{M}_{J^c,J}] - \boldsymbol{M}_{J^c,J}\|_2 + \|\mathbb{E}[\boldsymbol{M}_{J^c,J}]\|_2 \leq K_2 + p \cdot \|\boldsymbol{M}^*_{J^c,J}\|_2$ with probability at least $1 - d^{-c}$ by Lemma 6, we have that for any $c > 0$,

$$\|\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top\|_2 \leq (K_2 + p \cdot \|\boldsymbol{M}^*_{J^c,J}\|_2) \cdot (1 + \sqrt{s}) \tag{24}$$

with probability at least $1 - d^{-c}$.

Next, by Weyl's inequality,

$$\begin{aligned}
\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) &\geq \lambda_1(\mathbb{E}[\boldsymbol{M}_{J,J}]) - \lambda_2(\mathbb{E}[\boldsymbol{M}_{J,J}]) - 2 \cdot \|\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J} + \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top\|_2 \\
&\geq p \cdot (\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J})) - 2 \cdot \|\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J}\|_2 - 2\rho s.
\end{aligned}$$

By Lemma 6, we have that for any $c > 0$,

$$\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) \geq p \cdot (\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J})) - 2 \cdot K_1 - 2\rho s \tag{25}$$

with probability at least $1 - (2s)^{-c}$.

Finally, by applying Weyl's inequality and the triangle inequality, we have that

$$\begin{aligned}
\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \lambda_1(\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top) &\geq p \cdot \lambda_1(\boldsymbol{M}^*_{J,J}) - \|\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J}\|_2 - \rho s \\
&\quad - \|\mathbb{E}[\boldsymbol{M}_{J^c,J^c}] - \boldsymbol{M}_{J^c,J^c}\|_2 - p \cdot \lambda_1(\boldsymbol{M}^*_{J^c,J^c}) - \rho \cdot \|\hat{\boldsymbol{w}}\|_2^2.
\end{aligned}$$

Note that under the conditions in Lemma 2, $\|\hat{\boldsymbol{w}}\|_\infty < 1$, that is, $\|\hat{\boldsymbol{w}}\|_2 < \sqrt{d-s}$ holds with probability at least $1 - s^{-c}$. We can also use the upper bound of $\|\hat{\boldsymbol{w}}\|_2$ derived in Lemma 9. By applying Lemma 6, we have that for any $c > 0$,

$$\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \|\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top\|_2 \geq p \cdot (\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_1(\boldsymbol{M}^*_{J^c,J^c})) - K_1 - K_3 - \rho d \tag{26}$$

or

$$\begin{aligned}
\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \|\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top\|_2 &\geq p \cdot (\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_1(\boldsymbol{M}^*_{J^c,J^c})) - K_1 - \rho s \\
&\quad - K_3 - \frac{1}{\rho} \cdot (p\|\boldsymbol{M}^*_{J,J^c}\|_{\infty,2} + c_0\sqrt{d-s})^2
\end{aligned} \tag{27}$$

with probability at least $1 - s^{-c} - (2s)^{-c} - (2(d-s))^{-c}$, where $c_0 = \sqrt{s^c \cdot \{p(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_F^2 + ps(d-s)\sigma^2\}}$.

From (24)-(26), we can derive that if the following inequality holds:

$$(K_2 + p \cdot \|\boldsymbol{M}^*_{J^c,J}\|_2)^2 \cdot (1+\sqrt{s})^2 \leq \Big\{p \cdot (\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J})) - 2 \cdot K_1 - 2\rho s\Big\} \times \Big\{p \cdot (\lambda_1(\boldsymbol{M}^*_{J,J}) - \|\boldsymbol{M}^*_{J^c,J^c}\|_2) - K_1 - K_3 - \rho d\Big\},$$

then for any $c > 0$, the desired result holds with probability at least $1 - s^{-c} - d^{-c} - (2s)^{-c} - (2(d-s))^{-c}$.

Lemma 8. *If the following inequality holds:*

$$\|\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top\|_2^2 \leq \big\{\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)\big\} \cdot \big\{\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \|\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top\|_2\big\},$$

then $\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) = \lambda_1(\boldsymbol{M} - \rho(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top)^\top(\hat{\boldsymbol{z}}^\top, \hat{\boldsymbol{w}}^\top))$.

Proof. Let $\hat{\boldsymbol{Z}} = \begin{pmatrix} \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top & \hat{\boldsymbol{z}}\hat{\boldsymbol{w}}^\top \\ \hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top & \hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top \end{pmatrix}$. First, we can show that $\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)$ is an eigenvalue of $\boldsymbol{M} - \rho\hat{\boldsymbol{Z}}$ where its corresponding eigenvector is $(\hat{\boldsymbol{x}}^\top, 0^\top)^\top \in \mathbb{R}^d$. This is because

$$(\boldsymbol{M} - \rho\hat{\boldsymbol{Z}})\begin{pmatrix} \hat{\boldsymbol{x}} \\ 0 \end{pmatrix} = \begin{pmatrix} (\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)\hat{\boldsymbol{x}} \\ (\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top)\hat{\boldsymbol{x}} \end{pmatrix} = \lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) \cdot \begin{pmatrix} \hat{\boldsymbol{x}} \\ 0 \end{pmatrix}$$

where the last equality holds since $\hat{\boldsymbol{x}}$ is the leading eigenvector of $\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top$ and

$$(\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top)\hat{\boldsymbol{x}} = \boldsymbol{M}_{J^c,J}\hat{\boldsymbol{x}} - \rho \cdot \frac{1}{\rho\|\hat{\boldsymbol{x}}\|_1}\boldsymbol{M}_{J^c,J}\hat{\boldsymbol{x}} \cdot \|\hat{\boldsymbol{x}}\|_1 = 0.$$

Now, it is sufficient to show that for all $\boldsymbol{y} = (\boldsymbol{y}_1^\top,\ \boldsymbol{y}_2^\top)^\top$ such that $\boldsymbol{y}_1 \in \mathbb{R}^s$, $\boldsymbol{y}_2 \in \mathbb{R}^{d-s}$, $\|\boldsymbol{y}_1\|_2^2 + \|\boldsymbol{y}_2\|_2^2 = 1$ and $\hat{\boldsymbol{x}}^\top\boldsymbol{y}_1 = 0$,

$$\boldsymbol{y}^\top(\boldsymbol{M} - \rho\hat{\boldsymbol{Z}})\boldsymbol{y} \leq \lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top),$$

which implies that $\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)$ is the largest eigenvalue of $\boldsymbol{M} - \rho\hat{\boldsymbol{Z}}$. Note that

$$\begin{aligned}
&\boldsymbol{y}^\top(\boldsymbol{M} - \rho\hat{\boldsymbol{Z}})\boldsymbol{y} = \boldsymbol{y}_1^\top(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)\boldsymbol{y}_1 + 2\boldsymbol{y}_2^\top(\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top)\boldsymbol{y}_1 + \boldsymbol{y}_2^\top(\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top)\boldsymbol{y}_2 \\
&\leq \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) \cdot \|\boldsymbol{y}_1\|_2^2 + 2\|\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top\|_2 \cdot \|\boldsymbol{y}_1\|_2 \cdot \|\boldsymbol{y}_2\|_2 + \lambda_1(\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top) \cdot \|\boldsymbol{y}_2\|_2^2 \\
&= \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) \cdot (1 - \|\boldsymbol{y}_2\|_2^2) + 2\|\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top\|_2 \cdot \sqrt{1 - \|\boldsymbol{y}_2\|_2^2} \cdot \|\boldsymbol{y}_2\|_2 + \lambda_1(\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top) \cdot \|\boldsymbol{y}_2\|_2^2 \\
&= \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) + (\lambda_1(\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top) - \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)) \cdot \|\boldsymbol{y}_2\|_2^2 \\
&\quad + 2\|\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top\|_2 \cdot \sqrt{\|\boldsymbol{y}_2\|_2^2 \cdot (1 - \|\boldsymbol{y}_2\|_2^2)} \\
&= \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) + (\lambda_1(\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top) - \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)) \cdot t + 2\|\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top\|_2 \cdot \sqrt{t \cdot (1-t)}
\end{aligned}$$

where $0 \leq t := \|\boldsymbol{y}_2\|_2^2 \leq 1$. The first inequality holds since $\boldsymbol{y}_1/\|\boldsymbol{y}_1\|_2$ is orthogonal to $\hat{\boldsymbol{x}}$, the leading eigenvector of $\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top$. The above upper bound of $\boldsymbol{y}^\top(\boldsymbol{M} - \rho\hat{\boldsymbol{Z}})\boldsymbol{y}$ implies that if the following inequality holds for any $t \in [0,1]$:

$$\lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) + (\lambda_1(\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top) - \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)) \cdot t + 2\|\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top\|_2 \cdot \sqrt{t \cdot (1-t)} \leq \lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top),$$

then $\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)$ is the largest eigenvalue of $\boldsymbol{M} - \rho\hat{Z}$. From Lemma 10, we have that if the following inequality holds:

$$\|\boldsymbol{M}_{J^c,J} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{z}}^\top\|_2^2 \leq \big\{\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top)\big\} \cdot \big\{\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \lambda_1(\boldsymbol{M}_{J^c,J^c} - \rho\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top)\big\},$$

then $\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) = \lambda_1(\boldsymbol{M} - \rho\hat{\boldsymbol{Z}})$. □

Lemma 9. *Let* $c_0 = \sqrt{s^c \cdot \{p(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_F^2 + ps(d-s)\sigma^2\}}$ *for any* $c > 0$*. Then,*

$$\|\hat{\boldsymbol{w}}\|_2 \le \frac{p}{\rho} \cdot \|\boldsymbol{M}^*_{J,J^c}\|_{\infty,2} + \frac{c_0\sqrt{d-s}}{\rho}$$

with probability at least $1 - s^{-c}$*.*

Proof. First, we can derive the upper bound of $\|\hat{\boldsymbol{w}}\|_2$ as follows:

$$\begin{aligned}\|\hat{\boldsymbol{w}}\|_2 &= \frac{1}{\rho\|\hat{\boldsymbol{x}}\|_1}\|\boldsymbol{M}_{J^c,J}\hat{\boldsymbol{x}}\|_2 = \frac{1}{\rho\|\hat{\boldsymbol{x}}\|_1} \cdot \sqrt{\sum_{i\in J^c}\left(\sum_{j\in J} M_{i,j}\hat{x}_j\right)^2} \\ &\le \frac{1}{\rho\|\hat{\boldsymbol{x}}\|_1} \cdot \sqrt{\sum_{i\in J^c}\left(\max_{j\in J}|M_{i,j}| \cdot \|\hat{\boldsymbol{x}}\|_1\right)^2} = \frac{1}{\rho} \cdot \sqrt{\sum_{i\in J^c}\left(\max_{j\in J}|M_{i,j}|\right)^2}.\end{aligned}$$

By Chebyshev's inequality, for each $i \in J^c$ and $j \in J$,

$$\mathbb{P}\big(|M_{i,j}| \ge |\mathbb{E}[M_{i,j}]| + c_0\big) \le \frac{\mathsf{Var}[M_{i,j}]}{c_0^2} \tag{28}$$

for any $c_0 > 0$, where $\mathbb{E}[M_{i,j}] = p \cdot M^*_{i,j}$ and $\mathsf{Var}[M_{i,j}] = p(1-p)(M^*_{i,j})^2 + p\sigma^2$. That is,

$$\begin{aligned}&\mathbb{P}\big(|M_{i,j}| \le pM^*_{i,j} + c_0 \text{ for } \forall i \in J^c \text{ and } j \in J\big) \\ &\ge 1 - \sum_{i\in J^c, j\in J}\mathbb{P}\big(|M_{i,j}| \ge |\mathbb{E}[M_{i,j}]| + c_0\big) \\ &\ge 1 - \frac{p(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_F^2 + ps(d-s)\sigma^2}{c_0^2}.\end{aligned}$$

Let $c_0 = \sqrt{s^c \cdot \{p(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_F^2 + ps(d-s)\sigma^2\}}$ for any $c > 0$. Then $|M_{i,j}| \le pM^*_{i,j} + c_0$ for any $i \in J^c$ and $j \in J$ with probability at least $1 - s^{-c}$. This also means that $\max_{j\in J}|M_{i,j}| \le p\max_{j\in J}|M^*_{i,j}| + c_0$ and

$$\sqrt{\sum_{i\in J^c}\left(\max_{j\in J}|M_{i,j}|\right)^2} \le \sqrt{\sum_{i\in J^c}\left(p\max_{j\in J}|M^*_{i,j}|\right)^2} + \sqrt{\sum_{i\in J^c} c_0^2}$$

with probability at least $1 - s^{-c}$. Therefore, for $c_0 = \sqrt{s^c \cdot \{p(1-p)\|\boldsymbol{M}^*_{J^c,J}\|_F^2 + ps(d-s)\sigma^2\}}$,

$$\|\hat{\boldsymbol{w}}\|_2 \le \frac{p}{\rho} \cdot \|\boldsymbol{M}^*_{J,J^c}\|_{\infty,2} + \frac{c_0\sqrt{d-s}}{\rho}$$

with probability at least $1 - s^{-c}$ for any $c > 0$, where $\|\boldsymbol{M}^*_{J,J^c}\|_{\infty,2} = \sqrt{\sum_{i\in J^c}\left(\max_{j\in J}|M^*_{i,j}|\right)^2}$.

□

Lemma 10. *Assume* $a \ne 0$*. If* $a^2 \le c(b+c)$ *holds, then* $2a\sqrt{t(1-t)} \le bt + c$ *for all* $t \in [0,1]$*.*

Proof.

$$\begin{aligned}&2a\sqrt{t(1-t)} \le bt + c \quad \text{for all } t \in [0,1] \\ &\Leftarrow 4a^2t(1-t) \le (bt+c)^2,\ bt + c \ge 0 \quad \text{for all } t \in [0,1] \\ &\Leftrightarrow (4a^2+b^2)\left(t - \frac{2a^2-bc}{4a^2+b^2}\right)^2 + c^2 - \frac{(2a^2-bc)^2}{4a^2+b^2} \ge 0,\ bt + c \ge 0 \quad \text{for all } t \in [0,1] \\ &\Leftarrow c^2 - \frac{(2a^2-bc)^2}{4a^2+b^2} \ge 0,\ c \ge 0,\ b + c \ge 0 \\ &\Leftrightarrow a^2 \le c(b+c).\end{aligned}$$

□

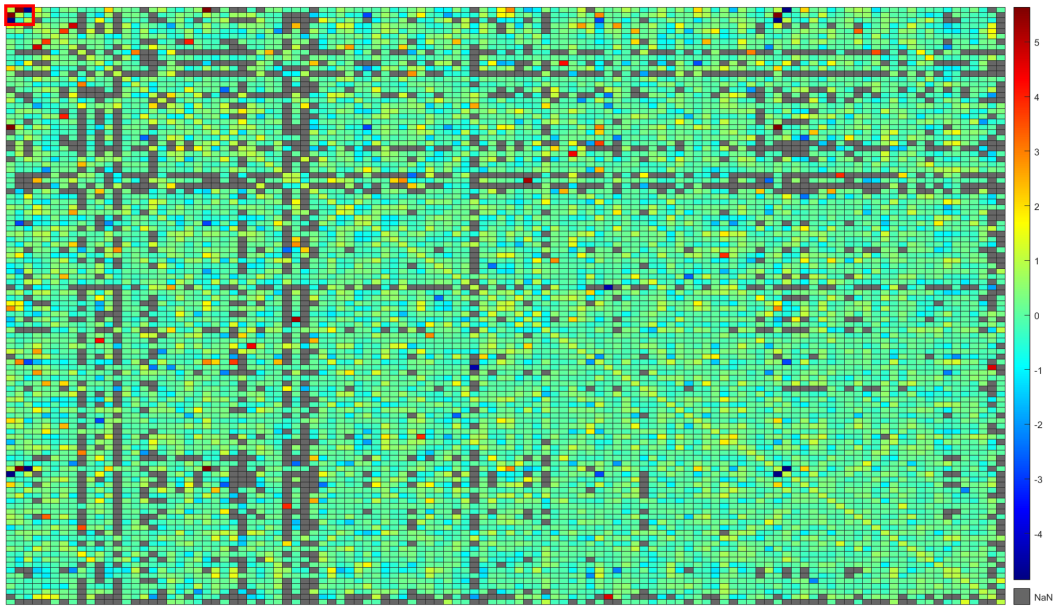

Figure 4: Heatmap of the Incomplete Covariance Matrix of the Gene Expression Data. We permuted all the genes, so that that the 3 genes in the support (selected by SDP optimization) appear in the top/left corner of this visualization.

E.7 Proof of Lemma 4

By Weyl's inequality and the triangle inequality,

$$\begin{aligned}
&\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) \\
&\geq \lambda_1(\mathbb{E}[\boldsymbol{M}_{J,J}]) - \lambda_2(\mathbb{E}[\boldsymbol{M}_{J,J}]) - 2 \cdot \|\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J}\|_2 - 2\rho \cdot \|\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top\|_2 \\
&= p \cdot (\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J})) - 2 \cdot \|\mathbb{E}[\boldsymbol{M}_{J,J}] - \boldsymbol{M}_{J,J}\|_2 - 2\rho \cdot \|\hat{\boldsymbol{z}}\|_2^2.
\end{aligned}$$

Note that $\|\hat{\boldsymbol{z}}\|_2 = \sqrt{s}$ and under the conditions in Lemma 2. Also, by applying Lemma 6, we have that

$$\lambda_1(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) - \lambda_2(\boldsymbol{M}_{J,J} - \rho\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\top) > p \cdot (\lambda_1(\boldsymbol{M}^*_{J,J}) - \lambda_2(\boldsymbol{M}^*_{J,J})) - 2 \cdot K_1 - 2\rho s$$

with probability at least $1 - (2s)^{-c}$ for any $c > 0$. Therefore, the desired result holds.

F Comparison with Other Methods for Sparse PCA

For the simulation in Section 4.3, we set $d = 50$, $s = 10$, $\lambda_1(\boldsymbol{M}^*) - \lambda_2(\boldsymbol{M}^*) = 10$, $B = 5$ and $\sigma_{normal} = 0.1$. We repeat each experiment 20 times with different random seeds, and calculate the rate of exact recovery.

We first consider two different algorithms where missing cells are treated as zero: the diagonal thresholding sparse PCA (DTSPCA) by Johnstone and Lu [2009] and the iterative thresholding sparse PCA (ITSPCA) by Ma [2013]. We use the tuning parameters $\rho = 0.1$, $\alpha = 5$ and $\gamma = 0.01$ for our SDP algorithm, DTSPCA and ITSPCA, respectively.

Next, we consider the combination of imputation and our SDP method. We first estimate the missing elements of $\boldsymbol{M}$ by using the matrix completion method based on the following Frobenius norm miminization with nuclear norm penalty $\|\boldsymbol{Y}\|_*$:

$$\tilde{\boldsymbol{M}} = \underset{\boldsymbol{Y}:symmetric}{\arg\min} \|P_\Omega(\boldsymbol{Y}) - \boldsymbol{M}\|_F^2 + \tau\|\boldsymbol{Y}\|_*$$

and then implemented the SDP method with the completed matrix $\tilde{\boldsymbol{M}}$. We use the tuning parameters $\rho = 0.1$ and $\tau = 0.1$.

G Additional Illustration: Heatmap of the Incomplete Covariance Matrix of Gene Expression Data in Section 4.2

A sparse eigenvector with a large eigenvalue should lead to a block in the covariance matrix with large values. To check this, we illustrate the heatmap of the incomplete covariance matrix of the gene expression data in Section 4.2. In Figure 4, the red box at the top left corner indicates the submatrix whose rows and columns correspond to the 3 genes (beta-1 catenin (CTNNB), hypoxanthine-guanine phosphoribosyltransferase 1 (HPRT1) and semaphorin III/F (SEMA3F)) selected from the sparse PCA method. We can observe that two of the three off-diagonals of this submatrix have significantly larger values than the other elements of the covariance matrix.