STATISTICAL INFERENCES AND APPLICATIONS FOR A LOW-RANK MATRIX

by

Juhee Cho

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2016

Date of final oral examination: 05/02/16

The dissertation is approved by the following members of the Final Oral Committee:
Karl Rohe, Assistant Professor, Statistics
Sunduz Keles, Professor, Statistics
Ming Yuan, Professor, Statistics
Anru Zhang, Assistant Professor, Statistics
Eva DuGoff, Assistant Professor, Population Health Sciences

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisor, Professor Karl Rohe, for his supervision, encouragement and continuous support of my Ph.D. study, research and career planning.

I would also like to thank Professor Sushmita Roy and Professor Eva H. Dugoff for introducing me to the fields of regulatory networks inference and physician referral networks, respectively. I thank Professor Sunduz Keles for her guidance, especially as my course advisor early in my graduate career. I also thank the other two members of my committee, Professor Ming Yuan and Professor Anru Zhang, for providing insightful comments, suggestions, and future directions on my work. I deeply appreciate the time and effort they spent on my behalf.

My sincere thanks also give to the Department of Statistics at University of Wisconsin-Madison. Thanks for providing me a wonderful Ph.D. program, as well as the financial support and all other resources.

Thank the weekly group meeting members, Tai Qin, Norbert Binkiewicz, Mohammad Khabbazian, Donggyu Kim, Thu Le, Song Wang, Yilin Zhang, Tianjie wang, Xiaoyi Yang, and Alan Sayler, for the discussions and inspiring ideas from various prospectives. Thank Emma Krauska and Zoe Russek for their help editing my manuscipts. Thank my friends, Jeein, and Mihee, for their support which helped me through times of struggle and stress.

Last but not least, I would like to thank my parents and my younger brother for their kind love and support. The existence of my family is the greatest blessing in my life.

ABSTRACT

Abstract

Low-rank matrix is a very reasonable assumption to make in many of real-world applications. Recently, as technology advances and various data types appear, researchers actively study its theoretical properties and empirical applicability. This work particularly investigates statistical properties and finds an application of a low-rank matrix in matrix completion and network analysis literature.

Matrix completion algorithms recover a low rank matrix from a small fraction of noisy entries. In practice, the singular vectors and singular values of the low rank matrix play a pivotal role for statistical analyses and inferences. Chapter 1 proposes estimators of those quantities and studies their convergence rate and asymptotic distribution. Then, the proposed estimators for the singular vectors and singular values combine to form a consistent estimator of the full low rank matrix which is computed with a non-iterative algorithm. In the cases studied in this chapter, this estimator achieves the minimax lower bound. The numerical experiments corroborate our theoretical results.

Various matrix completion algorithms have been developed in the past decade. Thresholded singular value decomposition (SVD) was a popular technique in implementing many of them. A sizable number of studies have shown its theoretical and empirical excellence, but choosing the right threshold level still remains as a key empirical difficulty. Chapter 2 proposes a novel matrix completion algorithm which iterates thresholded SVD with theoretically-justified and data-dependent values of thresholding parameters. The estimate of the proposed algorithm enjoys the minimax error rate and shows outstanding empirical performances. The thresholding scheme that the proposed algorithm uses can be viewed as a solution to a non-convex optimization problem, understanding of whose theoretical convergence guarantee is known to be limited. Chapter 2 investigates this problem by introducing a simpler algorithm, generalized-*softImpute*, analyzing its convergence behavior, and connecting it to the proposed algorithm.

Network is a vibrant area in statistics, biology, and computer science. Recently,

an emerging type of data in these fields is samples of labeled networks. The “labels" of networks imply that the nodes are labeled and that the same set of nodes reappears in all of the networks. Also, they mean that values (e.g. age, gender, or healthy v.s. sick) or vectors of values characterizing the associated network are also observed. Chapter 3 develops methods to estimate an induced subgraph which varies across the networks as the associated characteristics vary. Then, it applies the method and analyzes transcriptional regulatory network data observed from 41 diverse human cell types.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 ASYMPTOTIC THEORY FOR ESTIMATING THE SINGULAR VECTORS AND VALUES OF A PARTIALLY-OBSERVED LOW RANK MATRIX WITH NOISE

1.1 Introduction

The matrix completion problem arises in several different machine learning and engineering applications, ranging from collaborative filtering (Rennie and Srebro (2005)), to computer vision (Weinberger and Saul (2006)), to positioning (Montanari and Oh (2010)), and to recommender systems (Bennett and Lanning (2007)). The literature has established a sizable body of algorithmic research (Rennie and Srebro (2005); Keshavan et al. (2009); Cai et al. (2010); Mazumder et al. (2010); Hastie et al. (2014); Cho et al. (2015b)) and theoretical results (Fazel (2002); Srebro et al. (2004); Candès and Recht (2009); Candès and Plan (2010); Keshavan et al. (2010); Recht (2011); Gross (2011); Negahban et al. (2011); Koltchinskii et al. (2011a); Rohde et al. (2011); Koltchinskii et al. (2011b); Candès and Plan (2011); Negahban and Wainwright (2012); Cai and Zhou (2013); Davenport et al. (2014); Chatterjee (2014)). This extant literature is primarily focused on estimating the unobserved entries of the matrix. In several of these previous estimation techniques, the algorithms first estimate the singular vectors and singular values of the low rank matrix. Also, based upon classical multivariate statistics, these singular vectors and singular values can serve various types of statistical analyses and inferences. For example, the overarching aim in the Netflix problem was to predict the unobserved film ratings and the previous algorithms and theories served this purpose. However, if one wishes to interpret the resulting model predictions, then the estimated singular vectors and singular values can provide insights on (i) the main latent factors of film preferences and (ii) their relative strengths, respectively. In the Netflix example,

> "The first factor has on one side lowbrow comedies and horror movies, aimed at a male or adolescent audience (Half Baked, Freddy vs. Jason),

> while the other side contains drama or comedy with serious undertones and strong female leads (Sophie's Choice, Moonstruck). The second factor has independent, critically acclaimed, quirky films (Punch-Drunk Love, I Heart Huckabees) on one side, and mainstream formulaic films (Armageddon, Runaway Bride) on the other side." (Koren et al. (2009))

This inference is based upon the leading singular vectors of the estimated matrix. To the best of our knowledge, no previous research has studied the statistical properties of the estimated singular vectors and singular values.

This paper proposes estimators of the singular vectors and singular values of the low rank matrix as well as an estimator of the low rank matrix itself. First, Lemma 1.3.1 studies the singular vectors and singular values of a partially observed matrix that simply substitutes zeros for the unobserved entries; the resulting estimators are biased. The proposed estimators adjust for this bias. Theorem 1.4.1 finds the convergence rate for the bias-adjusted singular vector estimators and Theorem 1.4.3 gives a multivariate central limit theorem for the bias-adjusted singular value estimators. Despite the fact that the proposed estimators are built upon a partially observed matrix, they converge at the same rate as the standard estimators built from a fully observed matrix up to a constant factor which depends on the probability of observing each entry. Combining the proposed singular vector and value estimators, Section 1.4 gives a one-step consistent estimator of the low rank matrix which does not iterate over several singular value decompositions or eigenvalue decompositions. The mean squared error of this estimator achieves the minimax lower bound in Theorems 5-7 (Koltchinskii et al. (2011a)).

The rest of this paper is organized as follows. Section 3.5 describes the model setup. Section 1.3 shows that the singular vectors and singular values of a partially observed matrix are biased and suggests a bias-adjusted alternative. Section 1.4 finds (1) the convergence rates of the estimated singular vectors and (2) the asymptotic distribution of the estimated singular values. Section 1.4 proposes and studies a one-step consistent estimator of the full matrix. Section 1.5 corroborates the theoretical findings with numerical experiments. Finally, Section 1.6 provides the

proofs of our main theoretical results. The proofs of the other results are collected in the Appendix.

1.2 Model setup

The underlying matrix that we wish to estimate is an $n \times d$ matrix M_0 with rank r. By singular value decomposition (SVD),

$$M_0 = U \Lambda V^T, \tag{1.1}$$

for orthonormal matrices $U = (U_1, \ldots, U_r) \in \mathbb{R}^{n \times r}$ and $V = (V_1, \ldots, V_r) \in \mathbb{R}^{d \times r}$ containing the left and right singular vectors, and a diagonal matrix $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r) \in \mathbb{R}^{r \times r}$ containing the singular values. M_0 is corrupted by noise $\epsilon \in \mathbb{R}^{n \times d}$, where the entries of ϵ are i.i.d. sub-Gaussian random variables with mean zero and variance σ^2. Let $y \in \{0,1\}^{n \times d}$ be such that $y_{kh} = 1$ if the (k,h)-th entry of $M_0 + \epsilon$ is observed and $y_{kh} = 0$ otherwise. The entries of y are i.i.d. Bernoulli(p) and independent of the entries of ϵ. Thus, the total number of observed entries in $M_0 + \epsilon$ is a Binomial(nd, p) random variable. We observe y and the partially observed matrix $M \in \mathbb{R}^{n \times d}$, where

$$M_{kh} = \left[y \cdot (M_0 + \epsilon) \right]_{kh} = \begin{cases} M_{0kh} + \epsilon_{kh} & \text{if observed } (y_{kh} = 1) \\ 0 & \text{otherwise } (y_{kh} = 0) \end{cases}$$

for $1 \leqslant k \leqslant n$ and $1 \leqslant h \leqslant d$. Throughout the paper, it is presumed that $r \ll d \leqslant n$. Moreover, the entries of M_0 are bounded in absolute value by a constant $L > 0$.

Remark 1. *Depending on the case, the noise ϵ can be related to the measurement system so that assuming that there exist errors for unobserved entries does not make sense. Hence, assume a hierarchical model as follows;*

$$\begin{aligned} &\epsilon_{ij} | y_{ij} = 0 = 0 \ \textit{a.s.}, \\ &\epsilon_{ij} | y_{ij} = 1 \sim \textit{subgaussian, and} \end{aligned}$$

$$y_{ij} \quad \sim \textit{ i.i.d. Bernoulli}(p).$$

In this setting, the results obtained in this paper would still hold although it may require more techniques or minor changes in the proof. For simplicity of the paper, we only focus on the original setting.

1.3 Estimation of singular values and vectors of M_0

The vast majority of previous estimators of M_0 have been initialized with M, in effect imputing the missing values with zero. In this section, we study the properties of singular vectors and values of M. This suggests alternative estimators of the singular vectors and values of M_0.

Properties of singular values and vectors of M

Define

$$\hat{\Sigma} := M^\top M \quad \text{and} \quad \hat{\Sigma}_t := MM^\top.$$

Then, the eigenvectors of $\hat{\Sigma}$ and $\hat{\Sigma}_t$ are the same as the right and left singular vectors of M, respectively, and the squared root of eigenvalues of $\hat{\Sigma}$ are the same as the singular values of M. The following lemma shows that $\hat{\Sigma}$ and $\hat{\Sigma}_t$ are biased estimators of $M_0^\top M_0$ and $M_0 M_0^\top$, respectively.

Lemma 1.3.1. *Under the model setup in Section 3.5, we have*

$$\mathbb{E}\,\hat{\Sigma} = p^2 M_0^\top M_0 + p(1-p)\,\textit{diag}(M_0^\top M_0) + npd\sigma^2 I_d, \tag{1.2}$$

and similarly,

$$\mathbb{E}\,\hat{\Sigma}_t = p^2 M_0 M_0^\top + p(1-p)\,\textit{diag}(M_0 M_0^\top) + dpd\sigma^2 I_n, \tag{1.3}$$

where I_d *and* I_n *are* $d \times d$ *and* $n \times n$ *identity matrices, respectively.*

The proof of this lemma is in Appendix A. The right-hand side of (1.2) contains terms beyond $p^2 M_0^T M_0$ and they make the singular vectors and singular values of M biased estimators of the singular vectors and values of M_0. While the bias coming from $np\sigma^2 I_d$ is manageable[1], the bias coming from $p(1-p)\operatorname{diag}(M_0^T M_0)$ is not. The same applies to $\hat{\Sigma}_t$ in (1.3).

To get rid of the terms producing unmanageable biases, we define $\hat{\Sigma}_p$ and $\hat{\Sigma}_{pt}$ and their eigenvectors and eigenvalues as follows,

$$\begin{aligned}\hat{\Sigma}_p :=& \hat{\Sigma} - (1-p)\operatorname{diag}(\hat{\Sigma}) \\ =& (V_p, V_{pc})\operatorname{diag}(\lambda_{p1}^2, \ldots, \lambda_{pd}^2)(V_p, V_{pc})^T, \text{ and} \\ \hat{\Sigma}_{pt} :=& \hat{\Sigma}_t - (1-p)\operatorname{diag}(\hat{\Sigma}_t) \\ =& (U_p, U_{pc})\operatorname{diag}(\lambda_{pt1}^2, \ldots, \lambda_{ptn}^2)(U_p, U_{pc})^T,\end{aligned} \tag{1.4}$$

where

$$\begin{aligned} V_p &= (V_{p1}, \ldots, V_{pr}) \in \mathbb{R}^{d\times r}, \qquad V_{pc} = (V_{pr+1}, \ldots, V_{pd}) \in \mathbb{R}^{d\times(d-r)}, \\ U_p &= (U_{p1}, \ldots, U_{pr}) \in \mathbb{R}^{n\times r}, \qquad U_{pc} = (U_{pr+1}, \ldots, U_{pn}) \in \mathbb{R}^{n\times(n-r)}. \end{aligned}$$

The following proposition shows that $\hat{\Sigma}_p$ and $\hat{\Sigma}_{pt}$ adjust the bias.

Proposition 1.3.1. *Under the model setup in Section 3.5, we have by eigendecomposition,*

$$\begin{aligned}\mathbb{E}\,\hat{\Sigma}_p &= p^2 M_0^T M_0 + np^2\sigma^2 I_d = (V, V_c)\ddot{\Lambda}_p^2 (V, V_c)^T \text{ and} \\ \mathbb{E}\,\hat{\Sigma}_{pt} &= p^2 M_0 M_0^T + dp^2\sigma^2 I_n = (U, U_c)\ddot{\Lambda}_{pt}^2 (U, U_c)^T,\end{aligned}$$

where V *and* U *are as defined in* (2.1), $V_c \in \mathbb{R}^{d\times(d-r)}$, $U_c \in \mathbb{R}^{n\times(n-r)}$,

$$\begin{aligned}\ddot{\Lambda}_p^2 &= \mathit{diag}(\ddot{\lambda}_{p1}^2, \ldots, \ddot{\lambda}_{pd}^2) \\ &= \mathit{diag}(p^2[\lambda_1^2 + n\sigma^2], \ldots, p^2[\lambda_r^2 + n\sigma^2], p^2 n\sigma^2, \ldots, p^2 n\sigma^2) \in \mathbb{R}^{d\times d}, \text{ and} \\ \ddot{\Lambda}_{pt}^2 &= \mathit{diag}(p^2[\lambda_1^2 + d\sigma^2], \ldots, p^2[\lambda_r^2 + d\sigma^2], p^2 d\sigma^2, \ldots, p^2 d\sigma^2) \in \mathbb{R}^{n\times n}.\end{aligned}$$

[1] This term does not change the singular vectors of $\mathbb{E}\,\hat{\Sigma}$; it merely increases each singular value by $np\sigma^2$.

The proof of this proposition easily follows from Lemma 1.3.1 and (1.4).

Proposition 1.3.1 shows that the top r eigenvectors of $\mathbb{E}\,\hat{\Sigma}_p$ and $\mathbb{E}\,\hat{\Sigma}_{pt}$ are the same as the right and left singular vectors of M_0, respectively. Also, the top r eigenvalues of $\mathbb{E}\,\hat{\Sigma}_p$ are easily adjusted to match the singular values of M_0 as follows,

$$\lambda_i^2 = \frac{1}{p^2}\ddot{\lambda}_{p_i}^2 - n\sigma^2, \quad \text{for } i = 1, \dots, r.$$

Estimators of singular values and vectors of M_0

The results in Proposition 1.3.1 suggest plug-in estimators using the leading eigenvectors and eigenvalues of $\hat{\Sigma}_p$ and the leading eigenvectors of $\hat{\Sigma}_{pt}$ as estimators of V, Λ, and U, respectively. However, since p is an unknown parameter in practice, the proposed estimators use instead of p the proportion of observed entries in M, $\hat{p}$, which is defined as

$$\hat{p} = \frac{\sum_{k=1}^{n}\sum_{h=1}^{d} y_{kh}}{nd}. \tag{1.5}$$

Using $\hat{p}$, define $\hat{\Sigma}_{\hat{p}}$ and $\hat{\Sigma}_{\hat{p}t}$ as

$$\hat{\Sigma}_{\hat{p}} := \hat{\Sigma} - (1-\hat{p})\,\mathrm{diag}(\hat{\Sigma}) \quad \text{and} \quad \hat{\Sigma}_{\hat{p}t} := \hat{\Sigma}_t - (1-\hat{p})\,\mathrm{diag}(\hat{\Sigma}_t). \tag{1.6}$$

By eigendecomposition,

$$\hat{\Sigma}_{\hat{p}} = (\hat{V}, \hat{V}_c)\,\Lambda_{\hat{p}}^2\,(\hat{V}, \hat{V}_c)^T \quad \text{and} \quad \hat{\Sigma}_{\hat{p}t} = (\hat{U}, \hat{U}_c)\,\Lambda_{\hat{p}t}^2\,(\hat{U}, \hat{U}_c)^T, \tag{1.7}$$

where $\hat{V} \in \mathbb{R}^{d\times r}$, $\hat{V}_c \in \mathbb{R}^{d\times(d-r)}$, $\Lambda_{\hat{p}}^2 = \mathrm{diag}(\lambda_{\hat{p}1}^2, \dots, \lambda_{\hat{p}d}^2) \in \mathbb{R}^{d\times d}$, $\hat{U} \in \mathbb{R}^{n\times r}$, $\hat{U}_c \in \mathbb{R}^{n\times(n-r)}$, and $\Lambda_{\hat{p}t}^2 = \mathrm{diag}(\lambda_{\hat{p}t1}^2, \dots, \lambda_{\hat{p}tn}^2) \in \mathbb{R}^{n\times n}$. Then, estimate the left and right singular vectors, U and V, of M_0 by $\hat{U}$ and $\hat{V}$, respectively. Also, estimate the singular values, λ_i, $i = 1, \dots, r$, of M_0 by

$$\hat{\lambda}_i = \sqrt{\frac{1}{\hat{p}^2}\left(\lambda_{\hat{p}_i}^2 - \hat{\tau}_{\hat{p}}\right)} \quad \text{for } i = 1, \dots, r, \tag{1.8}$$

where $\hat{\tau}_{\hat{p}} = \frac{1}{d-r}\mathrm{tr}\left(\hat{V}_c^\top \hat{\Sigma}_{\hat{p}} \hat{V}_c\right)$.

For any $A \in \mathbb{R}^{n\times d}$, let the i-th left singular vector of A be denoted by $\mathbf{u}_i(A)$, the i-th right singular vector of A by $\mathbf{v}_i(A)$, and the top i-th singular value of A by $\boldsymbol{\lambda}_i(A)$ for $i = 1, \ldots, d$. Then, Algorithm 1 summarizes the steps to compute the proposed estimators of the singular values and vectors of M_0.

Algorithm 1 Estimators of U_i, V_i, and λ_i for $i = 1, \ldots, r$

Require: M, y, and r

$\hat{p} \leftarrow \frac{1}{nd}\sum_{k=1}^{n}\sum_{h=1}^{d} y_{kh}$

$\hat{\Sigma}_{\hat{p}} \leftarrow M^\top M - (1-\hat{p})\mathrm{diag}(M^\top M)$

$\hat{\Sigma}_{t\hat{p}} \leftarrow MM^\top - (1-\hat{p})\mathrm{diag}(MM^\top)$

$\hat{V}_i \leftarrow \mathbf{v}_i(\hat{\Sigma}_{\hat{p}}), \quad \forall i \in \{1, \ldots, r\}$

$\hat{U}_i \leftarrow \mathbf{u}_i(\hat{\Sigma}_{\hat{p}t}), \quad \forall i \in \{1, \ldots, r\}$

$\hat{\tau}_{\hat{p}} \leftarrow \frac{1}{d-r}\sum_{i=r+1}^{d} \boldsymbol{\lambda}_i(\hat{\Sigma}_{\hat{p}})$

$\hat{\lambda}_i \leftarrow \frac{1}{\hat{p}}\sqrt{\boldsymbol{\lambda}_i(\hat{\Sigma}_{\hat{p}}) - \hat{\tau}_{\hat{p}}}, \quad \forall i \in \{1, \ldots, r\}$

return $\hat{V}_i$, $\hat{U}_i$, and $\hat{\lambda}_i$ for $i = 1, \ldots, r$

1.4 Asymptotic theory

This section investigates the statistical properties of the estimators proposed in (1.7) and (1.8).

Convergence rate of the estimated singular vectors and asymptotic distribution of the estimated singular values

Let $x = (x_1, \ldots, x_n)^\top$ be a n-dimensional vector and $A = (A_{kh})$ a $n \times d$ matrix. Then, the ℓ_p-norm is defined as follows,

$$\|x\|_p = \left(\sum_{i=1}^{p} |x_i|^p\right)^{1/p}, \quad \text{and} \quad \|A\|_p = \sup\{\|Ax\|_p, \|x\|_p = 1\}, \ \ p = 1, 2, \infty.$$

The spectral norm $\|A\|_2$ is a square root of the largest eigenvalue of $AA^\top$,

$$\|A\|_1 = \max_{1\leqslant h\leqslant d}\sum_{k=1}^{n}|A_{kh}|, \text{ and } \|A\|_\infty = \max_{1\leqslant k\leqslant n}\sum_{h=1}^{d}|A_{kh}|.$$

The squared Frobenius norm is defined by $\|A\|_F^2 = \operatorname{tr}\left(A^\top A\right)$, the trace of $A^\top A$. We denote by $c > 0$ and $C > 0$ generic constants that are free of n, d, and p, and different from appearance to appearance.

To measure how close the proposed estimator $\hat{V}$ is to V (or, $\hat{U}$ to U), we introduce a classical notion of distance between subspaces. Let $\mathcal{R}(Z_1)$ denote a column space spanned by $Z_1 \in \mathbb{R}^{d\times r}$ and $\mathcal{R}(Z_2)$ by $Z_2 \in \mathbb{R}^{d\times r}$. Then, to measure the dissimilarity between $\mathcal{R}(Z_1)$ and $\mathcal{R}(Z_2)$, consider the following loss function

$$\|\sin(Z_1, Z_2)\|_F^2 = \|\sin\Theta(\mathcal{R}(Z_1), \mathcal{R}(Z_2))\|_F^2,$$

where $\sin\Theta(\mathcal{R}(Z_1), \mathcal{R}(Z_2))$ is a diagonal matrix of singular values (canonical angles) of $P_1P_2^\perp$ with orthogonal projections P_1 and P_2 of Z_1 and Z_2, respectively. Here $P^\perp = I - P$. The canonical angles generalize the notion of angles between lines and are often used to define the distance between subspaces. If the columns of Z_1 and Z_2 are singular vectors, $\mathcal{R}(Z_1)$ and $\mathcal{R}(Z_2)$ have projections $P_1 = Z_1Z_1^\top$ and $P_2 = Z_2Z_2^\top$, respectively, and $\|\sin(Z_1, \hat{Z}_2)\|_F^2 = \|Z_1Z_1^\top(Z_2Z_2^\top)^\perp\|_F^2 = \frac{1}{2}\|Z_1Z_1^\top - Z_2Z_2^\top\|_F^2$. Proposition 2.2 in Vu and Lei (2013) relates this subspace distance to the Frobenius distance

$$\frac{1}{2}\inf_{O\in\mathbb{V}_{r,r}}\|Z_1 - Z_2O\|_F^2 \leqslant \|\sin(Z_1, Z_2)\|_F^2 \leqslant \inf_{O\in\mathbb{V}_{r,r}}\|Z_1 - Z_2O\|_F^2, \tag{1.9}$$

where $\mathbb{V}_{r,r} = \{O \in \mathbb{R}^{r\times r} : O^\top O = I_r \text{ and } OO^\top = I_r\}$ denotes the Stiefel manifold of $r \times r$ orthonormal matrices. In other words, the distance between two subspaces corresponds to the minimal distance between their orthonormal bases.

Assumption 1.

(1) $\lambda_i = b_i\sqrt{nd}, i = 1, \ldots, r$, *where* $\frac{1}{c} \leqslant b_i \leqslant c$ *for a constant* $c > 0$;

(2) there exists a constant $m \in \{1, \dots, r\}$ *such that* $b_m > b_{m+1}$*, where* $b_{r+1} = 0$*;*

(3) $d \leqslant n \leqslant e^{d^\alpha}$ *for a constant* $\alpha < 1$ *free of* n*,* d*, and* p*.*

Remark 2. *To motivate Assumption 3 (1), suppose that a non-vanishing proportion of entries of* M_0 *contains non-vanishing signals (i.e.* $M_{0kh}^2 \geqslant c_0$ *for some constant* $c_0 > 0$*) and that the rank of* M_0 *is fixed. Then,*

$$\sum_{k=1}^{n} \sum_{h=1}^{d} M_{0kh}^2 = \|M_0\|_F^2 \geqslant cnd$$

for some constant $c > 0$*. Because the squared Frobenius norm is also the sum of the squared singular values of* M_0*, the order of the singular values of* M_0 *should be* $\sqrt{nd}$ *(see also Fan et al. (2013)). Assumption 3(1) may seem uncommon in the matrix completion literature, but consider the widely-used assumption (II.2) in Candès and Plan (2010),*

$$\max_{1 \leqslant k \leqslant n} |U_{ik}| \leqslant \sqrt{C/n} \quad \text{and} \quad \max_{1 \leqslant h \leqslant d} |V_{ih}| \leqslant \sqrt{C/d}$$

for $i = 1, \dots, r$ *and a constant* $C \geqslant 1$*, which prevents spiky singular vectors. Under the model setup in Section 3.5 where the entries of* M_0 *are bounded in absolute value by a constant* $L > 0$*, this implies Assumption 3(1).*

The following theorem shows the convergence of $\hat{V}$ to V and $\hat{U}$ to U.

Theorem 1.4.1. *Under the model setup in Section 3.5 and Assumption 3, let* $\hat{V}^{(m)}$ *and* $\hat{U}^{(m)}$ *be the first* m *columns of* $\hat{V}$ *and* $\hat{U}$ *defined in* (1.7)*, respectively, and let* $V^{(m)}$ *and* $U^{(m)}$ *be the first* m *columns of* V *and* U *defined in* (2.1)*, respectively. Then, for large* n *and* d*,*

$$\mathbb{E} \left\| \sin\left(\hat{V}^{(m)}, V^{(m)}\right) \right\|_F^2 \leqslant \frac{C_1 n^{-1}}{p\,(b_m^2 - b_{m+1}^2)^2} \tag{1.10}$$

and

$$\mathbb{E} \left\| \sin\left(\hat{U}^{(m)}, U^{(m)}\right) \right\|_F^2 \leqslant \frac{C_2 d^{-1}}{p\,(b_m^2 - b_{m+1}^2)^2}, \tag{1.11}$$

where C_1 *and* C_2 *are generic constants free of* n, d, *and* p.

The proof of this theorem is in Section 1.6.

Remark 3. *As long as* $\frac{p\,d}{\log n} \to \infty$, *the convergence rates in Theorem 1.4.1 will hold. Hence, under the setting where* p *goes to zero, if* $d/\log n$ *diverges fast enough that* $\frac{p\,d}{\log n} \to \infty$, *we can still obtain the results in Theorem 1.4.1.*

Remark 4. *Despite the fact that* $\hat{V}^{(m)}$ *is built on a partially observed matrix* M, *Theorem 1.4.1 gives the convergence rate* $\frac{n^{-1/2}}{(b_m^2 - b_{m+1}^2)}$ *which is the standard convergence rate for eigenvectors (Anderson et al. (1958)). The effect of the partial observations appears in the denominator of the right-hand side of* (1.10) *as* p. *A similar discussion applies to* $\hat{U}^{(m)}$ *in* (1.11).

The next theorem shows the asymptotic distribution of $\hat{\lambda}_i^2$ centered around λ_i^2.

Theorem 1.4.2. *Suppose* $nd^{-1} \to \infty$. *Then, under the model setup in Section 3.5 and Assumption 3, we have*

$$\frac{\sum_{i=1}^{m} \hat{\lambda}_i^2 - \sum_{i=1}^{m} \lambda_i^2}{\sqrt{nd}\sigma_\lambda} \to \mathcal{N}(0,1) \text{ in distribution}, \quad \text{as } n \text{ and } d \to \infty.$$

where

$$\sigma_\lambda^2 = \frac{4(1-p)}{p}\left\{ \sum_{k=1}^{n}\sum_{h=1}^{d} M_{0kh}^2 \left(\sum_{i=1}^{m} b_i U_{ik} V_{ih} \right)^2 - \left(\sum_{i=1}^{m} b_i^2 \right)^2 \right\} + \frac{4\sigma^2}{p} \sum_{i=1}^{m} b_i^2,$$

U_{ik} *is the* k*-th entry of* U_i, *and* V_{ih} *is the* h*-th entry of* V_i.

The proof of this theorem is in Section 1.6.

Remark 5. *As long as* $\frac{p\,d}{\log n} \to \infty$ *and* $pnd^{-1} \to \infty$, *the asymptotic normality result in Theorem 1.4.2 will hold. Hence, under the setting where* p *goes to zero, if* $d/\log n$ *and* n/d *diverge fast enough that* $\frac{p\,d}{\log n} \to \infty$ *and* $pn/d \to \infty$, *we can still obtain the results in Theorem 1.4.2.*

Remark 6. *Theorem 1.4.2 shows that the convergence rate of* $\sum_{i=1}^{m} \hat{\lambda}_i^2$ *is* $\sqrt{nd}$*. Considering Assumption 3(1), it is an optimal rate. However, since the results are based on partially observed entries, the asymptotic variance,* σ_λ^2*, increases with the rate* p^{-1}*. For example, when we have a fully-observed matrix,* σ_λ^2 *simply becomes* $4\sigma^2 \sum_{i=1}^{m} b_i^2$ *which is a lower bound for* σ_λ^2*.*

One of the main purposes of this paper is to investigate asymptotic behaviors of the estimators of the singular values of M_0. An application of the proof of Theorem 1.4.2 and the delta method provides a multivariate central limit theorem for $\hat{\lambda}_1, \ldots, \hat{\lambda}_r$.

Theorem 1.4.3. *Suppose that*

$$b_i > b_{i+1} \text{ for all } i \in \{1, \ldots, r\} \quad \text{and} \quad nd^{-1} \to \infty.$$

Then, under the model setup in Section 3.5 and Assumption 3, we have

$$\Upsilon^{-1/2} \begin{pmatrix} \hat{\lambda}_1 - \lambda_1 \\ \vdots \\ \hat{\lambda}_r - \lambda_r \end{pmatrix} \to \mathcal{N}(0, I_r) \text{ in distribution, as } n \text{ and } d \to \infty,$$

where $\Upsilon = \Upsilon^T \in \mathbb{R}^{r \times r}$ *consists of*

$$\Upsilon_{ij} = \begin{cases} \frac{(1-p)}{p} \left(\sum_{k=1}^{n} \sum_{h=1}^{d} M_{0kh}^2 U_{ik}^2 V_{ih}^2 - b_i^2 \right) + \frac{\sigma^2}{p} & \text{if } i = j \\ \frac{(1-p)}{p} \left(\sum_{k=1}^{n} \sum_{h=1}^{d} M_{0kh}^2 U_{ik} V_{ih} U_{jk} V_{jh} - b_i b_j \right) & \text{if } i \neq j. \end{cases} \tag{1.12}$$

Thus, $|\hat{\lambda}_i - \lambda_i| = O_p\left(\frac{1}{\sqrt{p}}\right)$.

Remark 7. *As in case of Theorem 1.4.2 (see Remark 5), as long as* $\frac{p\,d}{\log n} \to \infty$ *and* $pnd^{-1} \to \infty$*, the asymptotic normality result in Theorem 1.4.3 will hold. Note that Theorems 1.4.2 and 1.4.3 require an additional condition,* $pnd^{-1} \to \infty$*, to the condition required for Theorem 1.4.1,* $\frac{p\,d}{\log n} \to \infty$*. Under the setting where* p *is a constant, this additional condition implies that* d/n *has to go to zero. The rationale behind this is as*

follows. In Theorems 1.4.2 and 1.4.3, we find the limiting distribution on the singular values of M_0 *from a* $d \times d$ *matrix* $\hat{\Sigma}_{\hat{p}}$*, while the total number of observations is* nd*. That is, the size of our parameter space is* d^2 *and the total amount of information we can use to find asymptotic properties on the parameters is* nd*. Since our observations are even noisy, we need an enough number of observations to achieve our goal. When* $d/n \to 0$*, we can make the approximation errors in the singular values of* $\hat{\Sigma}_{\hat{p}}$ *negligible and find the limiting distribution on the singular values of* M_0*.*

Remark 8. *The results of Theorems 1.4.2 and 1.4.3 help us to make statistical inference on the singular values of* M_0*. For example, they open up possibilities for us to evaluate how many factors are significant or how influential each factor is, by providing the distribution of the singular values.*

Theorems 1.4.1-1.4.3 show that the proposed estimators for U, V, and λ_i's are asymptotically unbiased and have optimal convergence rates. With these well-developed estimators for the singular values and vectors of M_0, the following section proposes a consistent estimator of M_0.

A consistent estimator of M_0

Suppose that $b_i > b_{i+1}$ for all $i = 1, \dots, r$. Theorem 1.4.1 and (1.9) imply that $\hat{V}_i$ and $\hat{U}_i$ can estimate V_i and U_i up to constant factors $\operatorname{sign}(\langle \hat{V}_i, V_i \rangle)$ and $\operatorname{sign}(\langle \hat{U}_i, U_i \rangle)$, respectively. Let $s_0 = (s_{01}, \dots, s_{0r}) \in \{-1, 1\}^r$ be

$$s_{0i} = \operatorname{sign}(\langle \hat{V}_i, V_i \rangle) \operatorname{sign}(\langle \hat{U}_i, U_i \rangle) \quad \text{for} \quad i \in \{1, \dots, r\}. \tag{1.13}$$

Then, $\hat{M}(s_0) = \sum_{i=1}^{r} s_{0i} \hat{\lambda}_i \hat{U}_i \hat{V}_i^T$ becomes a consistent estimator of M_0. However, since s_0 is an unknown parameter in practice, we employ $\hat{s} = (\hat{s}_1, \dots, \hat{s}_r) \in \{-1, 1\}^r$ as an estimator of s_0;

$$\hat{s} = \underset{s \in \{-1,1\}^r}{\arg\min} \, \| \mathcal{P}_\Omega\big(\hat{M}(s)\big) - \mathcal{P}_\Omega\big(M\big) \|_F^2, \tag{1.14}$$

where Ω contains indices of the observed entries, $y_{kh} = 1 \Leftrightarrow (k,h) \in \Omega$, and $\mathcal{P}_\Omega(A)$ for any $A \in \mathbb{R}^{n\times d}$ denotes the projection of A onto Ω,

$$\mathcal{P}_\Omega(A)_{kh} = \begin{cases} A_{kh} & \text{if } (k,h) \in \Omega \\ 0 & \text{if } (k,h) \notin \Omega \end{cases} \quad \text{for } 1 \leqslant k \leqslant n \text{ and } 1 \leqslant h \leqslant d.$$

Hence, the proposed estimator of M_0 is

$$\hat{M}(\hat{s}) = \sum_{i=1}^{r} \hat{s}_i \hat{\lambda}_i \hat{U}_i \hat{V}_i^T. \tag{1.15}$$

Remark 9. *Finding* $\hat{s}$ *as in* (1.14) *requires* 2^r *computations. Hence, it can be a computational bottleneck or even impossible for a large* r*. In such cases, we suggest an alternate way to find* $\hat{s}$ *as follows;*

$$\hat{s}_i^{alternate} = sign(\langle \hat{V}_i, \mathbf{v}_i(M)\rangle)\ sign(\langle \hat{U}_i, \mathbf{u}_i(M)\rangle) \quad for\ \ i = 1, \dots, r.$$

Note that if we use V_i *and* U_i *instead of* $\mathbf{v}_i(M)$ *and* $\mathbf{u}_i(M)$*, this gives us the true sign* s_0 *in* (1.13).

In the following we show that $\hat{M}(\hat{s})$ is a consistent estimator of M_0 under certain conditions. The steps to compute $\hat{M}(\hat{s})$ using $\{\hat{V}_i, \hat{U}_i, \hat{\lambda}_i\}_{i=1}^r$ from Algorithm 1 are summarized in Algorithm 2.

Algorithm 2 Estimator of M_0

Require: $\hat{V}_i$, $\hat{U}_i$, and $\hat{\lambda}_i$ for $i = 1, \dots, r$
$\hat{s} \leftarrow \arg\min_{s\in\{-1,1\}^r} \left\| \mathcal{P}_\Omega\left(\sum_{i=1}^r s_i \hat{\lambda}_i \hat{U}_i \hat{V}_i^T \right) - \mathcal{P}_\Omega(M) \right\|_F^2$
$\hat{M}(\hat{s}) \leftarrow \sum_{i=1}^r \hat{s}_i \hat{\lambda}_i \hat{U}_i \hat{V}_i^T$
return $\hat{M}(\hat{s})$

Assumption 2.

(1) $\lim_{n\to\infty, d\to\infty} \mathbb{P}\Big(\min_{s\in\{-1,1\}^r} \|\mathcal{P}_\Omega(\hat{M}(s)) - \mathcal{P}_\Omega(M)\|_F^2 < \|\mathcal{P}_\Omega(\hat{M}(s_0)) - \mathcal{P}_\Omega(M)\|_F^2 \Big) = 0;$

(2) $b_i > b_{i+1}$ *for all* $i = 1, \dots, r$.

Remark 10. *When the rank* r *is 1, it is more straightforward to understand Assumption 5(1). Assuming that* $s_0 = 1$*, it means that*

$$\lim_{n\to\infty, d\to\infty} \mathbb{P}\Big(\|\mathcal{P}_\Omega(-\hat{\lambda}\hat{u}\hat{V}^T) - \mathcal{P}_\Omega(M)\|_F^2 < \|\mathcal{P}_\Omega(\hat{\lambda}\hat{u}\hat{V}^T) - \mathcal{P}_\Omega(M)\|_F^2 \Big) = 0.$$

That is, the probability that $\hat{s}$ *picks a different sign than the true sign* $s_0 = 1$ *goes to zero with the dimensionality. Given the asymptotic properties of our estimators* $\hat{\lambda}$, $\hat{u}$*, and* $\hat{V}$*, this is not an unreasonable assumption to make.*

Theorem 1.4.4. *Under the model setup in Section 3.5 and Assumptions 3-5, for any given* $\eta > 0$*, there exists a constant* $C_\eta > 0$ *such that for sufficiently large* n*,*

$$\mathbb{P}\left(\frac{p\, b_r^4}{n} \left\| \hat{M}(\hat{s}) - M_0 \right\|_F^2 \geqslant C_\eta \right) \leqslant \eta.$$

Or alternatively,

$$\|\hat{M}(\hat{s}) - M_0\|_F^2 = \frac{1}{p\, b_r^4}\, o_p\left(h_n n\right),$$

where h_n *can be anything that diverges very slowly with the dimensionality, for example,* $\log(\log d)$.

The proof of this theorem is in Section 1.6.

Remark 11. *As in case of Theorem 1.4.1 (see Remark 3), as long as* $\frac{p\,d}{\log n} \to \infty$*, the convergence rates in Theorem 1.4.4 will hold. If we let* $p = \frac{N}{nd}$ *so that* N *represents the number of observed entries in the population sense, this condition implies that* $\frac{N}{n \log n} \to \infty$*. Therefore, for* $\hat{M}(\hat{s})$ *to be consistent, the number of observed entries should increase at a faster rate than* $n \log n$*. This is a comparable result to Theorem 1 in Candès and Plan (2010).*

Remark 12. *The additional condition,* $p n d^{-1} \to \infty$*, required for Theorems 1.4.2 and 1.4.3 (see Remarks 5 and 7), is not needed for Theorems 1.4.1 and 1.4.4. It means that if* p *is a constant, even though* $d/n \to c$ *for some* $0 < c \leqslant 1$ *or* $d \leqslant n$*, the results in Theorems 1.4.1 and 1.4.4 will still hold, but the results in Theorems 1.4.2 and 1.4.3 will not.*

Remark 13. *Theorem 1.4.4 shows that* $\frac{1}{nd}\|\hat{M}(\hat{s}) - M_0\|_F^2$ *is bounded by* $C p^{-1} d^{-1}$ *for some constant* $C > 0$*. Under the setting where the rank of* M_0 *is fixed as in this paper, this is matched to the minimax lower bound in Theorems 5-7 (Koltchinskii et al. (2011a)). The previous estimators that obtain the minimax rate are computed via semidefinite programs that require iterating over several SVDs. However, the proposed estimator is a non-iterative algorithm.*

Remark 14. *Chatterjee (2014) established the minimax error rate for estimators of a general class of noisy incomplete matrices which extend beyond low rank matrix completion. In the regime studied herein, the convergence rate of our estimator of* M_0 *is faster than the convergence rate in Theorem 2.1 (Chatterjee (2014)). This is likely because we consider a smaller class of matrices, where the singular values of a low rank matrix have the divergence rate* $\sqrt{nd}$ *(Assumption 3(1)). Remark 2 justifies this assumption in the setting of low rank matrix completion.*

Throughout this paper, we have assumed that the rank, r, of M_0 is known. However, it is an unknown parameter and needs to be estimated. The following lemma proposes an estimator of r and shows its consistency.

Lemma 1.4.1. *Let* $C_d > 0$ *such that* $C_d/d \to 0$ *and* $C_d \to \infty$*, for example,* $C_d = c \log d$ *for any* $c > 0$*. Also, let* $\hat{r} = \left|\{i \in \{1, \ldots, d\} \mid \lambda_{\hat{p}i}^2 \geqslant p^2 n C_d\}\right|$ *where* $\lambda_{\hat{p}i}^2$ *is defined in* (1.7)*. Then, for any given* $\delta > 0$*, we have*

$$\mathbb{P}(\hat{r} = r) = 1 - O(n^{-\delta}).$$

The proof of this lemma is in Appendix A.

Remark 15. *Empirically to find* C_d *and* $\hat{r}$ *in Lemma 1.4.1, we suggest using a scree plot of the singular values of* $\hat{\Sigma}_{\hat{p}}$ *in* (1.6)*.*

Remark 16. *As long as* C_d *satisfies* $\sigma^2 p^2 n < p^2 n\, C_d \leqslant (\sigma^2 + b_r^2 d)\, p^2 n$, *consistency of* $\hat{r}$ *in Lemma 1.4.1 will hold. However, in the finite sample case, if the noise level* σ^2 *is larger than* $b_r^2 d$, *it can be difficult to observe a singular-value gap and determine* $\hat{r}$ *using the scree plot of the singular values of* $\hat{\Sigma}_{\hat{p}}$.

1.5 Numerical experiments

Simulations

This section studies the performance of the proposed estimators using several values of the dimension n and the probability p.

To simulate M_0, generate $A \in [-5,5]^{n\times 2}$, $B \in [-5,5]^{d\times 2}$ to contain i.i.d. Uniform$[-5,5]$ random variables and define

$$M_0 = AB^T \in \mathbb{R}^{n\times d}.$$

Each entry of M_0 is observed with probability p and unobserved with probability $1-p$. The observed entries of M_0 are corrupted by noise ϵ as defined in Section 3.5, where ϵ_{kh} are i.i.d. $\mathcal{N}(0,1)$. The dimension n varies from 100 to 1000 and p from 0.1 to 1, while $d = 2\sqrt{n}$. Each simulation was repeated 500 times and the errors were averaged.

Figures 1.1 and 1.2 summarize the resulting mean squared errors calculated by $\frac{1}{nd}\|\hat{M}(\hat{s}) - M_0\|_F^2$, $\|\mathrm{diag}(\hat{\lambda}_1, \hat{\lambda}_2) - \Lambda\|_F^2$, $\|\hat{V} - V\|_F^2$, and $\|\hat{U} - U\|_F^2$, when n and p increase along the x-axis, respectively. The MSE for $\hat{V}$ decreases more rapidly than the MSE for $\hat{U}$ and both MSEs decrease when p increases; this is consistent with the results in Theorem 1.4.1. The MSE of $\hat{M}$ decreases with the increase of n and p. The MSE of $\hat{\lambda}$ stays stable over the changes of n since it is measured on $\hat{\lambda}_i$ instead of $\hat{\lambda}_i^2$ (see Theorem 1.4.3), but decreases with the increase of p.

We further studied the asymptotic normality of $\sum_{i=1}^2 \hat{\lambda}_i$ in Theorem 1.4.3. Figure 1.3 graphs the QQ plot of $\sum_{i=1}^2 \hat{\lambda}_i - \sum_{i=1}^2 \lambda_i$, where the dimension n is fixed at 1000 and p varies from 0.1 to 1. This shows that the asymptotic normality holds across

various values of p.

A data example

To illustrate the proposed estimation methods, this section analyzes the MovieLens 100k data (GroupLens (2015)). The data set consists of 100,000 ratings from 943 users and 1682 movies and each user has rated at least 20 movies. Taking this partially observed data matrix as M, we computed $\hat{\Sigma}_{\hat{p}}$ as in (1.6) and plotted the scree plot of the singular values of $\hat{\Sigma}_{\hat{p}}$ to determine $\hat{r}$. Figure 1.4 shows the result. Since there exists a singular value gap between the 3rd and 4th singular values, we chose $\hat{r} = 3$. Then, we computed the estimators of the singular vectors and values and the estimator of the full low rank matrix as illustrated in Algorithms 1 and 2.

The estimated singular vectors help us understand what the main factors of movie preferences are. Table 1.1 shows lists of movies that characterize the top 3 singular vectors (factors of movie preferences). Particularly, it presents 5 movies that correspond to the largest values in each singular vector and 5 movies that correspond to the smallest values. The 1st factor has well-known and top-rated

Table 1.1: Lists of movies that characterize each of the top 3 singular vectors

1st singular vector	One side (well-known, top-rated)	Silence of the Lambs, Fargo, Star Wars, Return of the Jedi, Raiders of the Lost Ark
	The other side (unknown, pooly-rated)	A Further Gesture, Mat i syn, A Very Natural Thing, Hush, Office Killer
2nd singular vector	One side (box-office hit in 90's)	Scream, Air Force One, The Rock, Contact, Liar Liar
	The other side (classic in 40's-60's)	Citizen Kane, The Graduate, Casablanca, The African Queen, Dr. Strangelove
3rd singular vector	One side (action, thriller)	Jurassic Park, Top Gun, Speed, True Lies, Batman
	The other side (drama)	Il Postino, Secrets & Lies, English Patient, Full Monty, L.A. Confidential

movies on one side and unknown and poorly-rated movies on the other side. The 2nd factor has box-office hit movies in 1990's on one side and memorable classic

movies in 1940's-1960's on the other side. The 3rd factor has action and thriller movies on one side and quieter and drama movies on the other side.

The estimated singular values help us see how influential the main factors of movie preferences are. Particularly, Figure 1.5 shows the estimated singular values and their 95% confidence intervals. For the standard deviation used in the confidence intervals, we used $\Upsilon_{ii}^{-1/2}$ from (1.12) in Theorem 1.4.4. Computing $\Upsilon_{ii}^{-1/2}$ requires information on the values of the parameters M_0, U, V, λ_i, p, and σ^2, but we replaced these with the estimated values $\hat{M}(\hat{s}), \hat{U}, \hat{V}, \hat{\lambda}_i, \hat{p}$, and $\hat{\tau}_{\hat{p}}/n\hat{p}^2$. From Figure 1.5, we observe that all 3 factors of movie preferences are significant.

To find the RMSE of our estimator of the full low rank matrix, $\hat{M}(\hat{s})$, we used 5 training and 5 test data sets from 5-fold cross validation which is publicly provided in GroupLens (2015). The RMSE was computed by

$$\sqrt{\frac{\|\mathcal{P}_{\Omega_{test}}(\hat{M}(\hat{s})) - \mathcal{P}_{\Omega_{test}}(M)\|_F^2}{|\Omega_{test}|}},$$

where Ω_{test} contains indices of observed entries that belong to the test set, $\mathcal{P}_{\Omega_{test}}$ for a matrix $A \in \mathbb{R}^{n\times d}$ denotes the projection of A onto Ω_{test}, and $|\Omega_{test}|$ denotes the cardinality of Ω_{test}. The average of the resulting RMSEs was 1.656.

1.6 Proofs

Proofs for Theorem 1.4.1

The proof of the following proposition and lemmas are in Appendix A.

Proposition 1.6.1. *Under the model setup in Section 3.5 and Assumption 3, we have for large* n *and* d*,*

$$\mathbb{E}\left\|\sin\left(V_p^{(m)}, V^{(m)}\right)\right\|_F^2 \leqslant \frac{C_1\, n^{-1}}{p\,(b_m^2 - b_{m+1}^2)^2}, \text{ and} \tag{1.16}$$

$$\mathbb{E}\left\|\sin\left(U_p^{(m)}, U^{(m)}\right)\right\|_F^2 \leqslant \frac{C_2\, d^{-1}}{p\,(b_m^2 - b_{m+1}^2)^2},$$

where V_p *and* U_p *are defined in* (1.4) *and* C_1 *and* C_2 *are generic constants free of* n, d, *and* p.

Lemma 1.6.1. *Under the model setup in Section 3.5 and Assumption 3, for any given* $\mu_1 > 0$*, there exists a large constant* $C_{\mu_1} > 0$ *such that*

$$\frac{1}{nd}\left\|\hat{\Sigma}_p - \mathbb{E}\hat{\Sigma}_p\right\|_2 \leqslant C_{\mu_1} \max\left\{p\frac{\log n}{d},\, p^{3/2}\sqrt{\frac{\log n}{n}}\right\} \tag{1.17}$$

with probability at least $1 - O\left(n^{-\mu_1}\right)$*, where* $\hat{\Sigma}_p$ *is defined in* (1.4)*. Similarly, for any given* $\mu_2 > 0$*, there exists a large constant* $C_{\mu_2} > 0$ *such that*

$$\frac{1}{nd}\left\|\hat{\Sigma}_{pt} - \mathbb{E}\left(\hat{\Sigma}_{pt}\right)\right\|_2 \leqslant C_{\mu_2} \max\left\{p\frac{\log n}{d},\, p^{3/2}\sqrt{\frac{\log n}{d}}\right\}$$

with probability at least $1 - O\left(n^{-\mu_2}\right)$*, where* $\hat{\Sigma}_{pt}$ *is defined in* (1.4).

Lemma 1.6.2. *Under the model setup in Section 3.5 and Assumption 3, for any given* $\nu_1 > 0$*, there exists a large constant* $C_{\nu_1} > 0$ *such that*

$$\frac{1}{nd} \left\| \hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p \right\|_2 \leqslant C_{\nu_1} p^{3/2} \sqrt{\frac{\log n}{nd}} \frac{1}{d} \tag{1.18}$$

with probability at least $1 - O(n^{-\nu_1})$*, where* $\hat{\Sigma}_{\hat{p}}$ *and* $\hat{\Sigma}_p$ *are defined in* (1.6) *and* (1.4)*, respectively. Similarly, for any given* $\nu_2 > 0$*, there exists a large constant* $C_{\nu_2} > 0$ *such that*

$$\frac{1}{nd} \left\| \hat{\Sigma}_{\hat{p}t} - \hat{\Sigma}_{pt} \right\|_2 \leqslant C_{\nu_2} p^{3/2} \sqrt{\frac{\log n}{nd}} \frac{1}{n}$$

with probability at least $1 - O(n^{-\nu_2})$*, where* $\hat{\Sigma}_{\hat{p}t}$ *and* $\hat{\Sigma}_{pt}$ *are defined in* (1.6) *and* (1.4)*, respectively.*

Lemma 1.6.3. *Under the model setup in Section 3.5 and Assumption 3, we have for large* n *and* d*,*

$$\mathbb{E} \left\| \frac{1}{nd} \left(\hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p \right) V_p^{(m)} \right\|_F^2 \leqslant C_1 \max \left\{ \frac{p^3(1-p)}{nd^3}, \frac{p^2(1-p)}{n^2 d^{5/2}} \right\} \tag{1.19}$$

and

$$\mathbb{E} \left\| \frac{1}{nd} \left(\hat{\Sigma}_{\hat{p}t} - \hat{\Sigma}_{pt} \right) U_p^{(m)} \right\|_F^2 \leqslant C_2 \max \left\{ \frac{p^3(1-p)}{dn^3}, \frac{p^2(1-p)}{d^2 n^{5/2}} \right\},$$

where $\hat{\Sigma}_{\hat{p}}$ *and* $\hat{\Sigma}_{\hat{p}t}$ *are defined in* (1.6)*,* $\hat{\Sigma}_p$*,* $\hat{\Sigma}_{pt}$*,* V_p*, and* U_p *are defined in* (1.4)*, and* C_1 *and* C_2 *are generic constants free of* n*,* d*, and* p*.*

Proof of Theorem 1.4.1. We only prove (1.10) because (1.11) can be proved similarly.

By triangle inequality and Proposition 1.6.1, we have

$$\begin{aligned} &\mathbb{E} \| \sin \left(\hat{V}^{(m)}, V^{(m)} \right) \|_F^2 \\ &\leqslant 4 \, \mathbb{E} \| \sin \left(\hat{V}^{(m)}, V_p^{(m)} \right) \|_F^2 + 4 \mathbb{E} \| \sin \left(V_p^{(m)}, V^{(m)} \right) \|_F^2 \end{aligned}$$

$$\leqslant 4\,\mathbb{E}\|\sin\left(\hat{V}^{(m)}, V_p^{(m)}\right)\|_F^2 + \frac{C\,n^{-1}}{p\,(b_m^2 - b_{m+1}^2)^2}. \tag{1.20}$$

Now, consider $\mathbb{E}\|\sin\left(\hat{V}^{(m)}, V_p^{(m)}\right)\|_F^2$. Let

$$E_1 = \left\{ \max_{1\leqslant i\leqslant d} \frac{1}{nd}|\lambda^2_{p\,i} - \ddot{\lambda}^{\,2}_{p\,i}| < t_1 \right\},$$

where $t_1 = C_1' \, p\frac{\log n}{d} + C_1'' \, p^{3/2}\sqrt{\frac{\log n}{n}}$, and

$$E_2 = \left\{ \frac{1}{nd}|\lambda^2_{p\,m+1} - \lambda^2_{\hat{p}\,m+1}| < t_2 \right\}.$$

where $t_2 = C_2\, p^{3/2}\sqrt{\frac{\log n}{nd}}\frac{1}{d}$. Then, by Weyl's theorem (Li (1998a)), Lemma 1.6.1, and Lemma 1.6.2, we have for large constants C_1', C_1'', and C_2,

$$\mathbb{P}(E_1^c) \leqslant \mathbb{P}\left(\frac{1}{nd}\left\|\hat{\Sigma}_p - \mathbb{E}\hat{\Sigma}_p\right\|_2 \geqslant t_1\right) = O\left(n^{-4}\right) \text{ and}$$

$$\mathbb{P}(E_2^c) \leqslant \mathbb{P}\left(\frac{1}{nd}\left\|\hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p\right\|_2 \geqslant t_2\right) = O\left(n^{-4}\right).$$

Thus, for large n and d,

$$\begin{aligned}
&\mathbb{E}\|\sin\left(\hat{V}^{(m)}, V_p^{(m)}\right)\|_F^2 \\
&= \mathbb{E}\left\{ \|\sin\left(\hat{V}^{(m)}, V_p^{(m)}\right)\|_F^2 \; \mathbb{1}_{(E_1\cap E_2)^c} \right\} \\
&\qquad + \mathbb{E}\left\{ \|\sin\left(\hat{V}^{(m)}, V_p^{(m)}\right)\|_F^2 \; \mathbb{1}_{E_1\cap E_2} \right\} \\
&\leqslant m\left\{ \mathbb{E}\left(\mathbb{1}_{E_2^c}\right) + \mathbb{E}\left(\mathbb{1}_{E_1^c}\right) \right\} + \mathbb{E}\left\{ \frac{\left\| \frac{1}{nd}\left(\hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p\right) V_p^{(m)} \right\|_F^2}{\left(\frac{1}{nd}|\lambda^2_{p\,m} - \lambda^2_{\hat{p}\,m+1}|\right)^2} \; \mathbb{1}_{E_1\cap E_2} \right\} \\
&\leqslant cn^{-4} + \mathbb{E}\left\{ \frac{\left\| \frac{1}{nd}\left(\hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p\right) V_p^{(m)} \right\|_F^2 \; \mathbb{1}_{E_1\cap E_2}}{\left(\frac{1}{nd}|\ddot{\lambda}^{\,2}_{p\,m} - \ddot{\lambda}^{\,2}_{p\,m+1}| - t_2 - 2t_1\right)^2} \right\}
\end{aligned}$$

$$\leqslant cn^{-4} + \mathbb{E}\left\{ \frac{\left\| \frac{1}{nd} \left(\hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p \right) V_p^{(m)} \right\|_F^2}{\left(\frac{1}{2nd} |\ddot{\lambda}_{p\,m}^{\,2} - \ddot{\lambda}_{p\,m+1}^{\,2}| \right)^2} \right\}$$

$$\leqslant cn^{-4} + \frac{C(1-p)}{(b_m^2 - b_{m+1}^2)^2} \max\left\{ \frac{1}{pnd^3}, \frac{1}{p^2n^2d^{5/2}} \right\}, \tag{1.21}$$

where $\mathbb{1}_E$ is an indicator function of an event E, the first inequality holds by the fact that $\| \sin(\hat{V}^{(m)}, V_p^{(m)}) \|_F^2 \leqslant m$ and Davis-Kahan $\sin\theta$ theorem (Theorem 3.1 in Li (1998b)), and the last inequality is due to Lemma 1.6.3.

By (1.20) and (1.21), the result (1.10) follows. □

Proofs for Theorem 1.4.2

The proof of the following propositions are in Appendix A.

Proposition 1.6.2. *Under the assumptions in Theorem 1.4.2, we have*

$$\sqrt{nd}\, \Gamma_{nd}^{-1/2} \left[\begin{pmatrix} \frac{1}{nd\,p^2} \sum_{i=1}^m \lambda_{pi}^2 \\ \frac{p^2}{nd} \sum_{i=1}^m (\lambda_i^2 + n\sigma^2)\, \hat{p} \end{pmatrix} - \begin{pmatrix} \frac{1}{nd} \sum_{i=1}^m \left[\lambda_i^2 + n\sigma^2 \right] \\ \frac{p^3}{nd} \sum_{i=1}^m (\lambda_i^2 + n\sigma^2) \end{pmatrix} \right]$$

$$\to \mathcal{N}(0, I_2) \ \textit{in distribution}, \quad \textit{as } n, d \to \infty,$$

where λ_{pi}, λ_i, *and* $\hat{p}$ *are defined in* (1.4), (2.1), *and* (1.5), *respectively, and* $\Gamma_{nd} = \Gamma_{nd}^T \in \mathbb{R}^{2\times 2}$ *consists of*

$$(\Gamma_{nd})_{11} = \frac{4(1-p)}{p} \sum_{k=1}^n \sum_{h=1}^d M_{0kh}^2 \left\{ \sum_{i=1}^m b_i U_{ik} V_{ih} \right\}^2 + \frac{4\sigma^2}{p} \sum_{i=1}^m b_i^2,$$

$$(\Gamma_{nd})_{12} = 2p^2(1-p) \left(\sum_{i=1}^m b_i^2 \right)^2, \textit{ and } (\Gamma_{nd})_{22} = p^5(1-p) \left(\sum_{i=1}^m b_i^2 \right)^2.$$

Proposition 1.6.3. *Under the model setup in Section 3.5 and Assumption 3, let*

$$\hat{\tau}_p = \frac{1}{d-r} tr\left(V_{pc}^T \hat{\Sigma}_p V_{pc} \right),$$

where $\hat{\Sigma}_p$ *and* V_{pc} *are defined in* (1.4)*. Then, we have* $\hat{\tau}_p - np^2\sigma^2 = O_p\left(p\sqrt{n}\right)$.

Proof of Theorem 1.4.2. We have

$$\begin{aligned}
&\frac{1}{\sqrt{nd}}\left\{\sum_{i=1}^{m}\hat{\lambda}_i^2 - \sum_{i=1}^{m}\lambda_i^2\right\} \\
&= \frac{1}{\sqrt{nd}}\left\{\left(\hat{p}^{-2}\sum_{i=1}^{m}\lambda_{\hat{p}\,i}^2 - \sum_{i=1}^{m}\left[\lambda_i^2 + n\sigma^2\right]\right) + m\left(n\sigma^2 - \frac{1}{\hat{p}^2}\hat{\tau}_{\hat{p}}\right)\right\} \\
&= \frac{1}{\sqrt{nd}}\{(a) + m\,(b)\}.
\end{aligned}$$

First, consider the term (a). We have

$$\begin{aligned}
(a) &= \frac{1}{\hat{p}^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_{\hat{p}}\hat{V}^{(m)}\right) - \sum_{i=1}^{m}\left[\lambda_i^2 + n\sigma^2\right] \\
&= \left\{\frac{1}{p^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_p\hat{V}^{(m)}\right) - \sum_{i=1}^{m}\left[\lambda_i^2 + n\sigma^2\right]\right\} \\
&\quad + \left\{\frac{1}{p^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_{\hat{p}}\hat{V}^{(m)}\right) - \frac{1}{p^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_p\hat{V}^{(m)}\right)\right\} \\
&\quad + \left\{\frac{1}{\hat{p}^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_{\hat{p}}\hat{V}^{(m)}\right) - \frac{1}{p^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_{\hat{p}}\hat{V}^{(m)}\right)\right\} \\
&= (i) + (ii) + (iii). \qquad (1.22)
\end{aligned}$$

By (1.9), there is $O \in \mathbb{V}_{m,m}$ such that

$$\|\hat{V}^{(m)} - V_p^{(m)}O\|_F^2 \leqslant 2\|\sin(\hat{V}^{(m)}, V_p^{(m)})\|_F^2 \quad \text{and} \quad O_i^T V_p^{(m)T}\hat{\Sigma}_p V_p^{(m)}O_i = \lambda_{pi}^2,$$

where O_i is the i-th column of O. Then, the term (i) is

$$\begin{aligned}
(i) &= \frac{1}{p^2}\operatorname{tr}\left(O^T V_p^{(m)T}\hat{\Sigma}_p V_p^{(m)}O\right) - \sum_{i=1}^{m}\left[\lambda_i^2 + n\sigma^2\right] \\
&\qquad + \frac{1}{p^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_p\hat{V}^{(m)} - O^T V_p^{(m)T}\hat{\Sigma}_p V_p^{(m)}O\right) \\
&= \frac{1}{p^2}\operatorname{tr}\left(V_p^{(m)T}\hat{\Sigma}_p V_p^{(m)}\right) - \sum_{i=1}^{m}\left[\lambda_i^2 + n\sigma^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{p^2}\sum_{i=1}^{m}\left(\hat{V}_i^T\hat{\Sigma}_p\hat{V}_i - \mathcal{O}_i^T V_p{}^T\hat{\Sigma}_p V_p\mathcal{O}_i\right)\\
= \;& \frac{1}{p^2}\sum_{i=1}^{m}\lambda_{pi}^2 - \sum_{i=1}^{m}\left[\lambda_i^2+n\sigma^2\right] + O_p\left(\frac{1}{pd^2}\right), \qquad (1.23)
\end{aligned}
$$

where the last equality holds by the fact that

$$
\begin{aligned}
&\left|\sum_{i=1}^{m}\left(\hat{V}_i^T\hat{\Sigma}_p\hat{V}_i - \mathcal{O}_i^T V_p^{(m)T}\hat{\Sigma}_p V_p^{(m)}\mathcal{O}_i\right)\right|\\
&= \left|\sum_{i=1}^{m}\left[(\hat{V}_i - V_p^{(m)}\mathcal{O}_i)^T\hat{\Sigma}_p(\hat{V}_i - V_p^{(m)}\mathcal{O}_i) + 2\lambda_{pi}^2\mathcal{O}_i^T V_p^{(m)T}\hat{V}_i - 2\lambda_{pi}^2\right]\right|\\
&= \left|\sum_{i=1}^{m}\left[(\hat{V}_i - V_p^{(m)}\mathcal{O}_i)^T\hat{\Sigma}_p(\hat{V}_i - V_p^{(m)}\mathcal{O}_i) - \lambda_{pi}^2\left\|\hat{V}_i - V_p^{(m)}\mathcal{O}_i\right\|_2^2\right]\right|\\
&\leqslant 2\lambda_{p1}^2\sum_{i=1}^{m}\left\|\hat{V}_i - V_p^{(m)}\mathcal{O}_i\right\|_2^2\\
&= 2\lambda_{p1}^2\left\|\hat{V}^{(m)} - V_p^{(m)}\mathcal{O}\right\|_F^2\\
&= O_p\left(\frac{p}{d^2}\right), \qquad (1.24)
\end{aligned}
$$

where the last equality is due to (1.9), (1.21), and (1.25) below; by the application of Weyl's theorem (Li (1998a)) and Lemma 1.6.1, we can show

$$
\lambda_{p1}^2 = O_p(p^2nd). \qquad (1.25)
$$

The term (ii) is

$$
\begin{aligned}
\mathbb{E}\,|(ii)| &= \mathbb{E}\left|\frac{1}{p^2}(\hat{p}-p)\,\mathrm{tr}\left(\hat{V}^{(m)T}\mathrm{diag}(\hat{\Sigma})\hat{V}^{(m)}\right)\right|\\
&\leqslant \frac{m}{p^2}\,\mathbb{E}\left|(\hat{p}-p)\max_{1\leqslant i\leqslant m}\hat{V}_i^T\mathrm{diag}(\hat{\Sigma})\hat{V}_i\right|\\
&\leqslant \frac{m}{p^2}\left\{\mathbb{E}(\hat{p}-p)^2\right\}^{1/2}\left\{\mathbb{E}\Big[\max_{1\leqslant i\leqslant m}\hat{V}_i^T\mathrm{diag}(\hat{\Sigma})\hat{V}_i\Big]^2\right\}^{1/2}\\
&\leqslant \frac{m}{p^2}\left\{\mathbb{E}(\hat{p}-p)^2\right\}^{1/2}\left\{\mathbb{E}\Big[\left\|\mathrm{diag}(\hat{\Sigma})\right\|_2^2\Big]\right\}^{1/2}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{m}{p^2}\sqrt{\frac{p(1-p)}{nd}}\left\{\mathbb{E}\Big[\,\|\mathrm{diag}(\hat{\Sigma})\|_2^2\,\Big]\right\}^{1/2} \\
&= O\left(\max\left\{\frac{1}{p}, \sqrt{\frac{n}{pd}}\right\}\right), \qquad (1.26)
\end{aligned}
$$

where the second inequality is due to Hölder's inequality and the last equality holds by the fact that

$$
\begin{aligned}
&\mathbb{E}\Big[\,\|\mathrm{diag}(\hat{\Sigma})\|_2^2\Big] \\
&\leqslant 4\,\mathbb{E}\Big[\,\|\mathrm{diag}(\hat{\Sigma}) - p\,\mathrm{diag}(M_0^T M_0) - np\sigma^2 I_d\|_2^2 \\
&\qquad\qquad + \|p\,\mathrm{diag}(M_0^T M_0) + np\sigma^2 I_d\|_2^2\Big] \\
&= 4\,\mathbb{E}\left[\max_{1\leqslant h\leqslant d}\Big|\sum_{k=1}^{n}\left(M_{kh}^2 - pM_{0kh}^2 - p\sigma^2\right)\Big|^2\right] \\
&\qquad\qquad + 4\left\{\max_{1\leqslant h\leqslant d} p\sum_{k=1}^{n} M_{0kh}^2 + np\sigma^2\right\}^2 \\
&\leqslant 4\sum_{h=1}^{d}\mathbb{E}\left\{\Big|\sum_{k=1}^{n}\left[M_{kh}^2 - p(M_{0kh}^2 + \sigma^2)\right]\Big|^2\right\} + 4\left\{np(L^2+\sigma^2)\right\}^2 \\
&= 4\sum_{h=1}^{d}\sum_{k=1}^{n}\mathbb{E}\left[M_{kh}^2 - p(M_{0kh}^2 + \sigma^2)\right]^2 + 4\left\{np(L^2+\sigma^2)\right\}^2 \\
&= O\left(\max\{pnd, p^2n^2\}\right).
\end{aligned}
$$

The term (iii) in (1.22) is

$$
\begin{aligned}
(iii) &= \left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right)\left[\mathrm{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_{\hat{p}}\hat{V}^{(m)}\right) - p^2\sum_{i=1}^{m}\left(\lambda_i^2 + n\sigma^2\right)\right] \\
&\quad + \left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right)p^2\sum_{i=1}^{m}\left(\lambda_i^2 + n\sigma^2\right) \\
&= \left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right)\Bigg[\mathrm{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_{\hat{p}}\hat{V}^{(m)}\right) - \mathrm{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_p\hat{V}^{(m)}\right) \\
&\qquad\qquad + \mathrm{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_p\hat{V}^{(m)}\right) - \mathrm{tr}\left(\mathcal{O}^T V_p^{(m)T}\hat{\Sigma}_p V_p^{(m)}\mathcal{O}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \operatorname{tr}\left(\mathcal{O}^{\mathsf{T}} V_p^{(m)\mathsf{T}} \hat{\Sigma}_p V_p^{(m)} \mathcal{O}\right) - p^2 \sum_{i=1}^{m} \left(\lambda_i^2 + n\sigma^2\right) \Bigg] \\
&\quad + \left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right) p^2 \sum_{i=1}^{m} \left(\lambda_i^2 + n\sigma^2\right) \\
&= O_p\left(\frac{1}{\sqrt{p^5 nd}}\right) \left[O_p\left(\max\left\{p, \sqrt{\frac{p^3 n}{d}}\right\}\right) + O_p\left(\frac{p}{d^2}\right) + O_p\left(\sqrt{p^3 nd}\right) \right] \\
&\quad + \left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right) p^2 \sum_{i=1}^{m} \left(\lambda_i^2 + n\sigma^2\right) \\
&= O_p\left(\frac{1}{p}\right) + \left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right) p^2 \sum_{i=1}^{m} \left(\lambda_i^2 + n\sigma^2\right), \qquad (1.27)
\end{aligned}
$$

where the third equality is due to (1.26), (1.24), Proposition 1.6.2, and the fact that

$$
\sqrt{nd}\left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right) \to \mathcal{N}\left(0, \frac{4(1-p)}{p^5}\right) \text{ in distribution, as } n, d \to \infty, \qquad (1.28)
$$

by CLT and Delta method. From (1.23), (1.26), and (1.27), we have

$$
\begin{aligned}
(a) &= \frac{1}{p^2} \sum_{i=1}^{m} \lambda_{p\,i}^2 - \sum_{i=1}^{m} \left[\lambda_i^2 + n\sigma^2\right] \\
&\quad + \left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right) p^2 \sum_{i=1}^{m} \left(\lambda_i^2 + n\sigma^2\right) + o_p\left(\sqrt{\frac{nd}{p}}\right). \qquad (1.29)
\end{aligned}
$$

Second, the term (b) is

$$
\begin{aligned}
(b) &= n\sigma^2 - \frac{1}{\hat{p}^2}\hat{\tau}_{\hat{p}} \\
&= \left(n\sigma^2 - \frac{1}{p^2}\hat{\tau}_p\right) + \left(\frac{1}{p^2} - \frac{1}{\hat{p}^2}\right)\hat{\tau}_{\hat{p}} + \frac{1}{p^2}\left(\hat{\tau}_p - \hat{\tau}_{\hat{p}}\right) \\
&= O_p\left(\frac{\sqrt{n}}{p}\right) + O_p\left(\sqrt{\frac{n}{pd}}\right) + \frac{1}{p^2}\left(\hat{\tau}_p - \hat{\tau}_{\hat{p}}\right) \\
&= o_p\left(\sqrt{\frac{nd}{p}}\right), \qquad (1.30)
\end{aligned}
$$

where the third equality is due to Proposition 1.6.3 and (1.28), and the last equality holds by the fact that there is $\tilde{\mathcal{O}} \in \mathbb{V}_{d-r,d-r}$ by (1.9) such that

$$\|\hat{V}_c^{(m)} - V_{pc}^{(m)}\tilde{\mathcal{O}}\|_F^2 \leqslant 2\|\sin(\hat{V}_c^{(m)}, V_{pc}^{(m)})\|_F^2 \quad \text{and} \quad \tilde{\mathcal{O}}_i^T V_{pc}^T \hat{\Sigma}_p V_{pc} \tilde{\mathcal{O}}_i = \lambda_{p\,r+i}^2,$$

where $\tilde{\mathcal{O}}_i$ is the i-th column of $\tilde{\mathcal{O}}$, and that

$$\begin{aligned}
&|\hat{\tau}_p - \hat{\tau}_{\hat{p}}| \\
&= \frac{1}{(d-r)}\Big|\mathrm{tr}\left(\tilde{\mathcal{O}}^T V_{pc}^T \hat{\Sigma}_p V_{pc}\tilde{\mathcal{O}}\right) - \mathrm{tr}\left(\hat{V}_c^T \hat{\Sigma}_p \hat{V}_c\right) \\
&\qquad\qquad + \mathrm{tr}\left(\hat{V}_c^T \hat{\Sigma}_p \hat{V}_c\right) - \mathrm{tr}\left(\hat{V}_c^T \hat{\Sigma}_{\hat{p}} \hat{V}_c\right)\Big| \\
&\leqslant \frac{1}{(d-r)}\left|\mathrm{tr}\left(\tilde{\mathcal{O}}^T V_{pc}^T \hat{\Sigma}_p V_{pc}\tilde{\mathcal{O}}\right) - \mathrm{tr}\left(\hat{V}_c^T \hat{\Sigma}_p \hat{V}_c\right)\right| \\
&\qquad\qquad + \frac{1}{(d-r)}\left|\mathrm{tr}\left(\hat{V}_c^T \hat{\Sigma}_p \hat{V}_c\right) - \mathrm{tr}\left(\hat{V}_c^T \hat{\Sigma}_{\hat{p}} \hat{V}_c\right)\right| \\
&\leqslant \frac{1}{(d-r)} 4\lambda_{p1}^2 \left\|\sin(V_{pc}, \hat{V}_c)\right\|_F^2 + \frac{1}{(d-r)}\left|(\hat{p}-p)\mathrm{tr}\left(\hat{V}_c^T \mathrm{diag}(\hat{\Sigma})\hat{V}_c\right)\right| \\
&= \frac{1}{(d-r)} 4\lambda_{p1}^2 \left\|\sin(V_p, \hat{V})\right\|_F^2 + O_p\left(\max\left\{p,\, p^{3/2}\sqrt{\frac{n}{d}}\right\}\right) \\
&= O_p\left(\frac{p}{d^3}\right) + O_p\left(\max\left\{p,\, p^{3/2}\sqrt{\frac{n}{d}}\right\}\right),
\end{aligned}$$

where the second inequality can be derived similarly to (1.24), the second equality holds similarly to (1.26), and the last equality is due to (1.21) and (1.25).

Combining the results in (1.29) and (1.30), we have

$$\begin{aligned}
&\frac{1}{\sqrt{nd}}\left\{\sum_{i=1}^m \hat{\lambda}_i^2 - \sum_{i=1}^m \lambda_i^2\right\} \\
&= \frac{1}{\sqrt{nd}}\{(a) + m\,(b)\} \\
&= \frac{1}{\sqrt{nd}}\left\{\frac{1}{p^2}\sum_{i=1}^m \lambda_{p\,i}^2 - \sum_{i=1}^m\left[\lambda_i^2 + n\sigma^2\right] + \left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right) p^2 \sum_{i=1}^m \left(\lambda_i^2 + n\sigma^2\right)\right\} \\
&\qquad + o_p(1).
\end{aligned}$$

Thus, by Proposition 1.6.2, Delta method and Slutsky's theorem, we have

$$\frac{1}{\sqrt{ndp}\sigma_\lambda}\left\{\sum_{i=1}^{m}\hat{\lambda}_i^2-\sum_{i=1}^{m}\lambda_i^2\right\}\ \to\ \mathcal{N}(0,1)\ \text{ in distribution,}\quad \text{as } n,d\to\infty,$$

where $\sigma_\lambda^2=\begin{pmatrix}1 & -2p^{-3}\end{pmatrix}\Gamma_{nd}\begin{pmatrix}1\\ -2p^{-3}\end{pmatrix}$. □

Proofs for Theorem 1.4.4

The proof of the following Proposition is in Appendix A.

Proposition 1.6.4. *Under the model setup in Section 3.5, Assumption 3, and Assumption 5(2), we have*

$$\left\|\hat{M}(s_0)-M_0\right\|_F^2=\frac{1}{p\,b_r^4}\,O_p\left(n\right),$$

where $\hat{M}(s_0)$ *are defined in* (1.13) *and* (1.15) *and* M_0 *is defined in* (2.1).

Proof of Theorem 1.4.4. For any given $\eta>0$, we have for a large n,

$$\mathbb{P}\left(\min_{s\in\{-1,1\}^r}\|\mathcal{P}_\Omega(\hat{M}(s))-\mathcal{P}_\Omega(M)\|_F^2<\|\mathcal{P}_\Omega(\hat{M}(s_0))-\mathcal{P}_\Omega(M)\|_F^2\right)\leqslant\eta/2$$

by Assumption 5(1). Also, for any given $\eta>0$, we can find $C_\eta>0$, free of n, d, and p, such that for large n,

$$\mathbb{P}\left(\frac{p\,b_r^4}{n}\left\|\hat{M}(s_0)-M_0\right\|_F^2\geqslant C_\eta\right)\leqslant\eta/2$$

by Proposition 1.6.4. Therefore, for any given $\eta>0$, we can find $C_\eta>0$ such that

$$\begin{aligned}&\mathbb{P}\left(\frac{p\,b_r^4}{n}\left\|\hat{M}(\hat{s})-M_0\right\|_F^2\geqslant C_\eta\right)\\&=\mathbb{P}\left(\frac{p\,b_r^4}{n}\left\|\hat{M}(s_0)-M_0\right\|_F^2\geqslant C_\eta, s_0=\hat{s}\right)\end{aligned}$$

$$
\begin{aligned}
&\quad + \mathbb{P}\left(\frac{p\, b_r^4}{n} \left\| \hat{M}(\hat{s}) - M_0 \right\|_F^2 \geqslant C_\eta, s_0 \neq \hat{s} \right) \\
&\leqslant \mathbb{P}\left(\frac{p\, b_r^4}{n} \left\| \hat{M}(s_0) - M_0 \right\|_F^2 \geqslant C_\eta \right) \\
&\quad + \mathbb{P}\left(\min_{s \in \{-1,1\}^r} \| \mathcal{P}_\Omega(\hat{M}(s)) - \mathcal{P}_\Omega(M) \|_F^2 < \| \mathcal{P}_\Omega(\hat{M}(s_0)) - \mathcal{P}_\Omega(M) \|_F^2 \right) \\
&\leqslant \eta/2 + \eta/2 \\
&= \eta.
\end{aligned}
$$

Or, for any given $\eta > 0$ and $\zeta > 0$, there exists $N_\zeta > 0$ such that for all $n \geqslant N_\zeta$,

$$
\begin{aligned}
&\mathbb{P}\left(\frac{p\, b_r^4}{h_n n} \left\| \hat{M}(\hat{s}) - M_0 \right\|_F^2 > \eta \right) \\
&= \mathbb{P}\left(\frac{p\, b_r^4}{h_n n} \left\| \hat{M}(s_0) - M_0 \right\|_F^2 > \eta, s_0 = \hat{s} \right) \\
&\quad + \mathbb{P}\left(\frac{p\, b_r^4}{h_n n} \left\| \hat{M}(\hat{s}) - M_0 \right\|_F^2 > \eta, s_0 \neq \hat{s} \right) \\
&\leqslant \mathbb{P}\left(\frac{p\, b_r^4}{h_n n} \left\| \hat{M}(s_0) - M_0 \right\|_F^2 \geqslant \eta \right) \\
&\quad + \mathbb{P}\left(\min_{s \in \{-1,1\}^r} \| \mathcal{P}_\Omega(\hat{M}(s)) - \mathcal{P}_\Omega(M) \|_F^2 < \| \mathcal{P}_\Omega(\hat{M}(s_0)) - \mathcal{P}_\Omega(M) \|_F^2 \right) \\
&\leqslant \zeta/2 + \zeta/2 \\
&= \zeta,
\end{aligned}
$$

where the second inequality holds due to Assumption 5(1) and Proposition 1.6.4. □

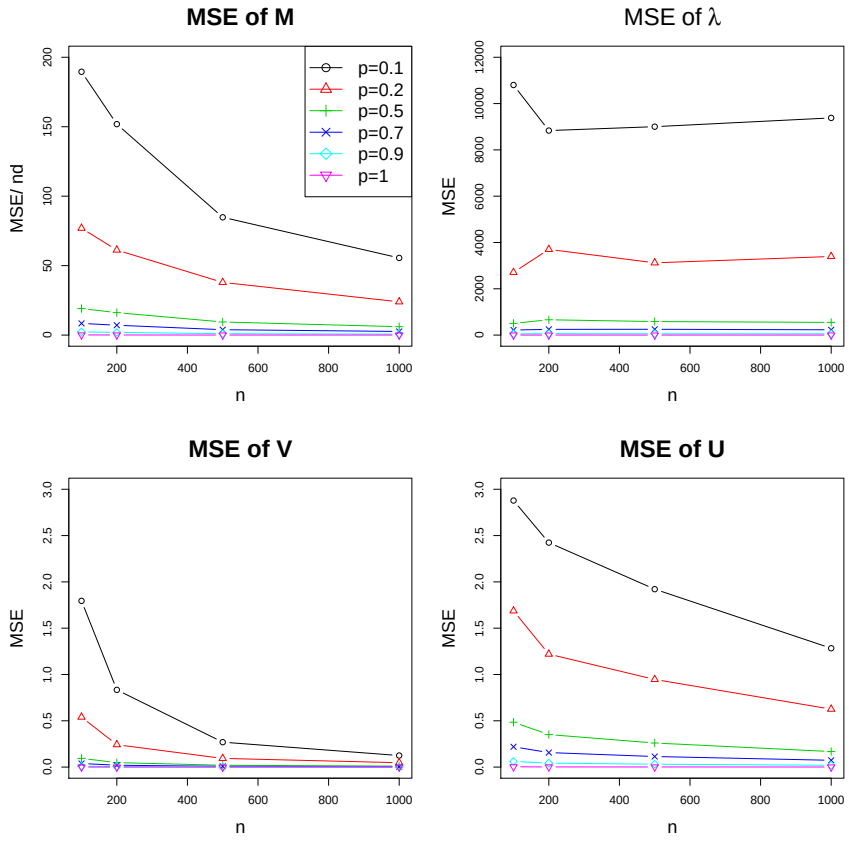

Figure 1.1: The mean squared errors for six different values of p when n increases. Each point on the plots correspond to an average over 500 replicates.

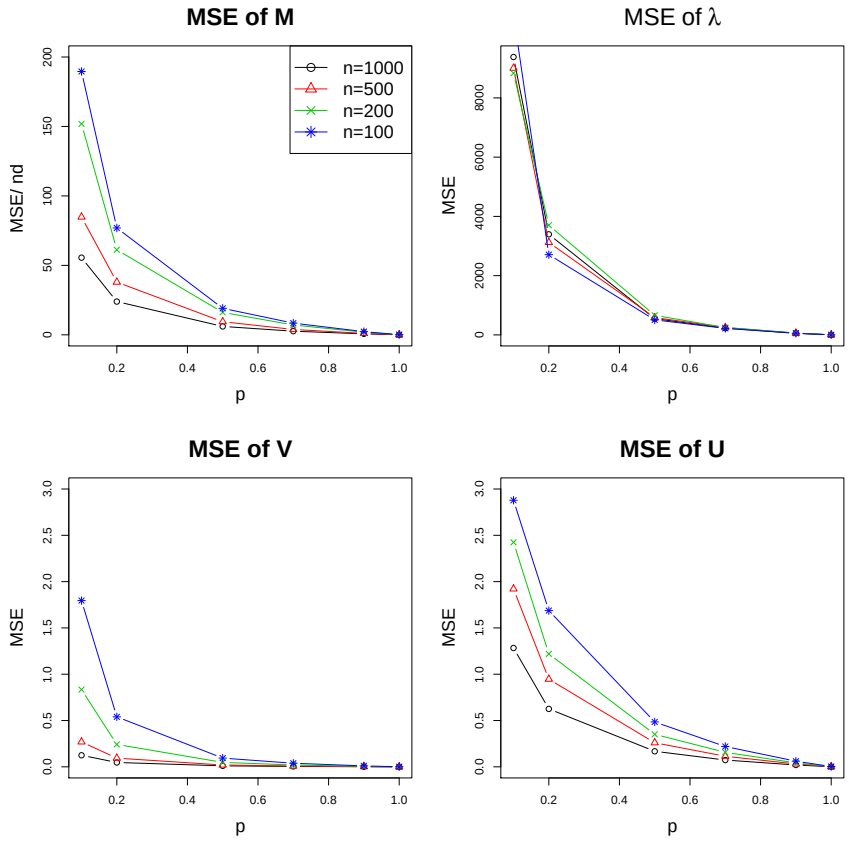

Figure 1.2: The same mean squared errors as the ones in Figure 1 plotted for four different values of n when p increases. Each point on the plots correspond to an average over 500 replicates.

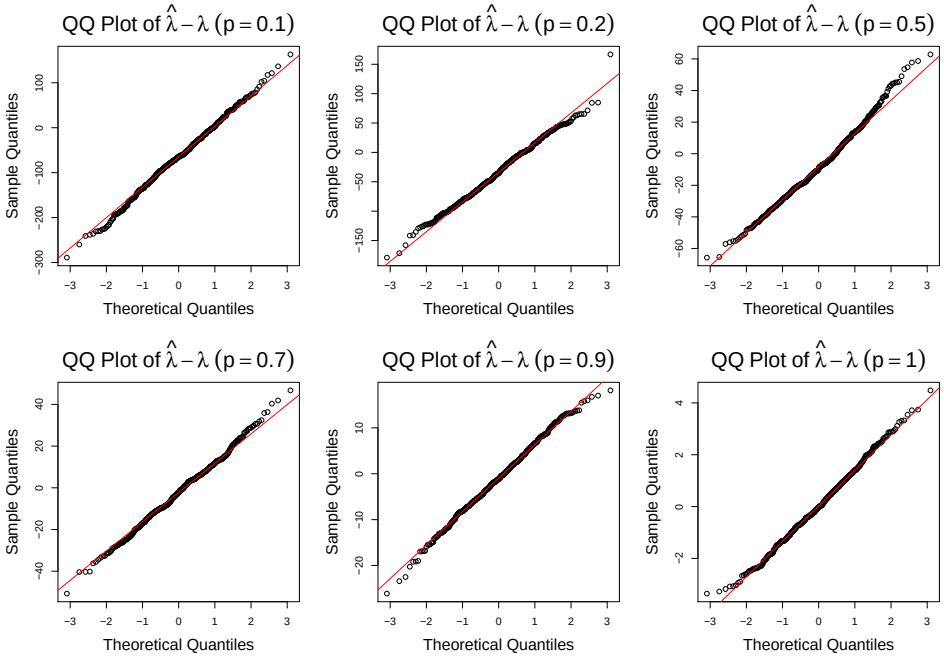

Figure 1.3: Asymptotic normality of $\sum_{i=1}^{2} \hat{\lambda}_i - \sum_{i=1}^{2} \lambda_i$ as p varies from 0.1 to 1. Across the plots, we fixed n to be 1000.

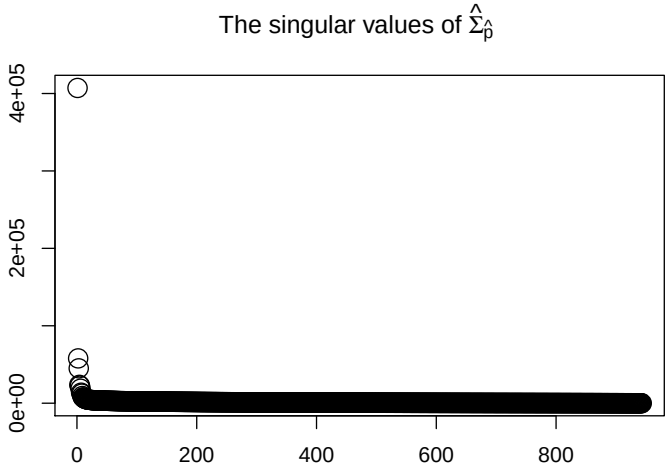

Figure 1.4: The singular values of $\hat{\Sigma}_{\hat{p}}$ computed by taking the MovieLens 100k data matrix as M. From this scree plot, we choose $\hat{r}$ to be 3.

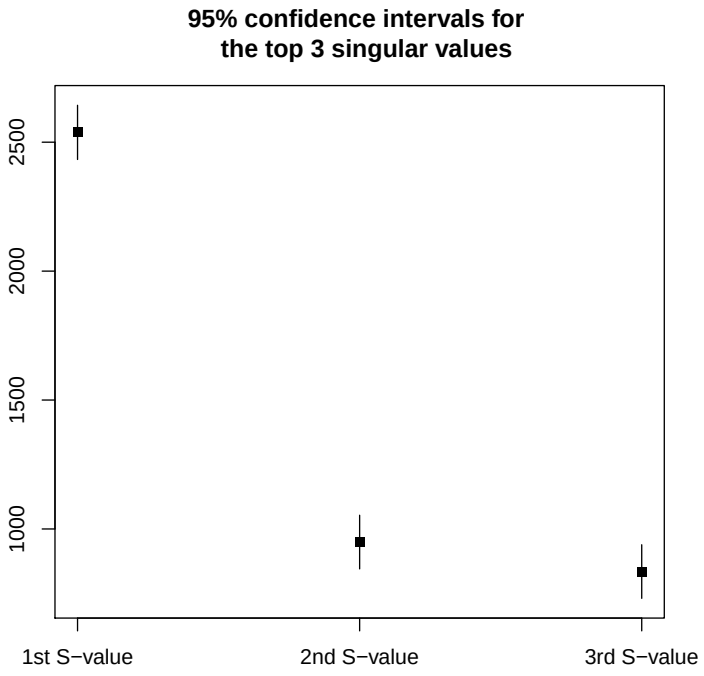

Figure 1.5: The 3 estimated singular values and their 95% confidence intervals.

2 INTELLIGENT INITIALIZATION AND ADAPTIVE THRESHOLDING FOR ITERATIVE MATRIX COMPLETION; SOME STATISTICAL AND ALGORITHMIC THEORY FOR *adaptive-impute*

2.1 Introduction

Matrix completion appears in a variety of areas where it recovers a low-rank or approximately low-rank matrix from a small fraction of observed entries such as collaborative filtering (Rennie and Srebro (2005)), computer vision (Weinberger and Saul (2006)), positioning (Montanari and Oh (2010)), and recommender systems (Bennett and Lanning (2007)). Early work in this field was done by Achlioptas and McSherry (2001), Azar et al. (2001), Fazel (2002), Srebro et al. (2004), and Rennie and Srebro (2005). Later, Candès and Recht (2009) introduced the technique of matrix completion by minimizing the nuclear norm under convex constraints. This opened up a significant overlap with compressed sensing (Candès et al. (2006), Donoho (2006)) and led to accelerated research in matrix completion. They and others (Candès and Recht (2009), Candès and Tao (2010), Keshavan et al. (2010), Gross (2011), Recht (2011)) showed that the technique can exactly recover a low-rank matrix in the noiseless case. Many of the following works showed the approximate recovery of the low-rank matrix with the presence of noise (Candès and Plan (2010), Negahban et al. (2011), Koltchinskii et al. (2011a), Rohde et al. (2011)). Several other papers studied matrix completion in various settings (e.g. Davenport et al. (2014), Negahban and Wainwright (2012)) and proposed different estimation procedures of matrix completion (Srebro et al. (2004), Keshavan et al. (2009), Koltchinskii et al. (2011b), Cai and Zhou (2013), Chatterjee (2014)) than the ones by Candès and Recht (2009). In addition to the theoretical advances, a large number of algorithms have emerged (e.g. Rennie and Srebro (2005), Cai et al. (2010), Keshavan et al. (2009), Mazumder et al. (2010), Hastie et al. (2014)). An overview is well summarized in Mazumder et al. (2010) and Hastie et al. (2014).

Many of matrix completion algorithms employ thresholded singular value decomposition (SVD) which soft- or hard- thresholds the singular values. The statistical literature has responded by investigating its theoretical optimality and strong empirical performances. However, a key empirical difficulty of employing thresholded SVD for matrix completion is to find the right way and level of threshold. Depending on the choice of the thresholding scheme, the rank of the estimated low-rank matrix and predicted values for unobserved entries can widely change. Despite its importance, we lack understanding on how to choose the threshold level and what bias or error we eliminate by thresholding.

We propose a novel iterative matrix completion algorithm, *Adaptive-Impute*, which recovers the underlying low-rank matrix from a few noisy entries via differentially and adaptively thresholded SVD. Specifically, the proposed *Adaptive-Impute* algorithm differentially thresholds the singular values and adaptively updates the threshold levels on every iteration. As was the case with adaptive Lasso (Zou (2006)) and adaptive thresholding for sparse covariance matrix estimation (Cai and Liu (2011)), the proposed thresholding scheme gives *Adaptive-Impute* stronger empirical performances than the thresholding scheme that uses a single thresholding parameter for all singular values throughout the iterations (e.g. *softImpute* (Mazumder et al. (2010))). Although *Adaptive-Impute* employs multiple thresholding parameters changing over iterations, we suggest specified values for the thresholding parameters that are theoretically-justified and data-dependent. Hence, *Adaptive-Impute* is free of the tuning problems associated with the choice of threshold levels. Its single tuning parameter is the rank of the resulting estimator. We suggest a way to choose the rank based on singular value gaps (for details, see Section 2.5). This novel threshold scheme of *Adaptive-Impute* makes it estimation via non-convex optimization, understanding of whose theoretical guarantees is known to be limited. However, to solve this problem and help understand the convergence behavior of *Adaptive-Impute*, we introduce a simpler algorithm than *Adaptive-Impute*, generalized-*softImpute*, and derive a sufficient condition under which it converges. Then, we prove that *Adaptive-Impute* behaves almost the same as generalized-*softImpute*. Numerical experiments and a real data analysis in Section

2.5 suggest superior performances of *Adaptive-Impute* over the existing *softImpute*-type algorithms.

The rest of this paper is organized as follows. Section 3.5 describes the model setup. Section 2.3 introduces the proposed algorithm *Adaptive-Impute*. Section 2.4 introduces a generalized-*softImpute*, a simpler algorithm than *Adaptive-Impute*. Section 2.5 presents numerical experiment results. Section 2.6 concludes the paper with discussion. All proofs are collected in Section B.

2.2 The model setup

Suppose that we have an $n \times d$ matrix of rank r,

$$M_0 = U\Lambda V^T, \tag{2.1}$$

where by SVD, $U = (U_1, \ldots, U_r) \in \mathbb{R}^{n\times r}$, $V = (V_1, \ldots, V_r) \in \mathbb{R}^{d\times r}$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r) \in \mathbb{R}^{r\times r}$, and $\lambda_1 \geqslant \ldots \geqslant \lambda_r \geqslant 0$. The entries of M_0 are corrupted by noise $\epsilon \in \mathbb{R}^{n\times d}$ whose entries are i.i.d. sub-Gaussian random variables with mean zero and variance σ^2. Hence, we can only observe $M_F = M_0 + \epsilon$. However, oftentimes in real world applications, not all entries of M_F are observable. So, define $y \in \mathbb{R}^{n\times d}$ such that $y_{ij} = 1$ if the (i, j)-th entry of M_F is observed and $y_{ij} = 0$ if it is not observed. The entries of y are assumed to be i.i.d. Bernoulli(p) and independent of the entries of ϵ. Then, the partially-observed noisy low-rank matrix $M \in \mathbb{R}^{n\times d}$ is written as

$$M_{ij} = y_{ij}M_{Fij} = \begin{cases} M_{0ij} + \epsilon_{ij} & \text{if observed } (y_{ij} = 1) \\ 0 & \text{otherwise } (y_{ij} = 0). \end{cases}$$

Throughout the paper, we assume that $r \ll d \leqslant n$ and the entries of M_0 are bounded by a positive constant L in absolute value. In this paper, we develop an iterative algorithm to recover M_0 from M and investigate its theoretical properties and empirical performances.

2.3 *Adaptive-Impute* algorithm

Initialization

We first introduce some notation. Let a set Ω contain indices of the observed entries, $y_{ij} = 1 \Leftrightarrow (i,j) \in \Omega$. Then, for any matrix $A \in \mathbb{R}^{n\times d}$, denote by $\mathcal{P}_\Omega(A)$ the projection of A onto Ω and by $\mathcal{P}_\Omega^\perp(A)$ the projection of A onto the complement of Ω;

$$[\mathcal{P}_\Omega(A)]_{ij} = \begin{cases} A_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{if } (i,j) \notin \Omega \end{cases} \quad \text{and} \quad \left[\mathcal{P}_\Omega^\perp(A)\right]_{ij} = \begin{cases} 0 & \text{if } (i,j) \in \Omega \\ A_{ij} & \text{if } (i,j) \notin \Omega. \end{cases}$$

That is, $\mathcal{P}_\Omega(A) + \mathcal{P}_\Omega^\perp(A) = A$. We let $\mathbf{u}_i(A)$ denote the i-th left singular vector of A, $\mathbf{v}_i(A)$ the i-th right singular vector of A, and $\boldsymbol{\lambda}_i(A)$ the i-th singular value of A such that $\boldsymbol{\lambda}_1(A) \geqslant \ldots \geqslant \boldsymbol{\lambda}_d(A)$. The squared Frobenius norm is defined by $\|A\|_F^2 = \text{tr}\left(A^\top A\right)$, the trace of $A^\top A$, and the nuclear norm by $\|A\|_* = \sum_{i=1}^d \boldsymbol{\lambda}_i(A)$, the sum of the singular values of A. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, diag(A) represents a matrix with diagonal elements of A on the diagonal and zeros elsewhere.

Many of the iterative matrix completion algorithms (e.g. Cai et al. (2010), Mazumder et al. (2010), Keshavan et al. (2009), Chatterjee (2014)) in the current literature initialize with M, where the unobserved entries begin at zero. This initialization works well with algorithms that are based on convex optimization or that are robust to the initial. However, for algorithms that are based on non-convex optimization or that are sensitive to the initial, filling the unobserved entries with zeros may not be a good choice. Cho et al. (2015a) proposed a one-step consistent estimator, $\hat{M}$, that attains the minimax error rate (Koltchinskii et al. (2011a)), r/pd, and requires only two eigendecompositions. *Adaptive-Impute* employs the entries of this one-step consistent estimator instead of zeros as initial values of the unobserved entries. Algorithm 3 describes how to compute the initial $\hat{M}$ of *Adaptive-Impute*. The following theorem shows that $\hat{M}$ achieves the minimax error rate.

Assumption 3.

Algorithm 3 Initialization (Cho et al. (2015a))

Require: M, y, and r

$\hat{p} \leftarrow \frac{1}{nd} \sum_{i=1}^{n} \sum_{j=1}^{d} y_{ij}$

$\Sigma_{\hat{p}} \leftarrow M^T M - (1-\hat{p})\mathrm{diag}(M^T M)$

$\Sigma_{t\hat{p}} \leftarrow MM^T - (1-\hat{p})\mathrm{diag}(MM^T)$

$\hat{V}_i \leftarrow \mathbf{v}_i(\Sigma_{\hat{p}}), \quad \forall i \in \{1, \ldots, r\}$

$\hat{U}_i \leftarrow \mathbf{u}_i(\Sigma_{t\hat{p}}), \quad \forall i \in \{1, \ldots, r\}$

$\widetilde{\alpha} \leftarrow \frac{1}{d-r} \sum_{i=r+1}^{d} \lambda_i(\Sigma_{\hat{p}})$

$\hat{\tau}_i \leftarrow \lambda_i(\Sigma_{\hat{p}}) - \frac{1}{\hat{p}} \sqrt{\lambda_i(\Sigma_{\hat{p}}) - \widetilde{\alpha}}, \quad \forall i \in \{1, \ldots, r\}$

$\hat{\lambda}_i \leftarrow \lambda_i(\Sigma_{\hat{p}}) - \hat{\tau}_i, \quad \forall i \in \{1, \ldots, r\}$

$\hat{s} = (\hat{s}_1, \ldots, \hat{s}_r) \leftarrow \arg\min_{s \in \{-1,1\}^r} \left\| \mathcal{P}_\Omega \left(\sum_{i=1}^{r} s_i \hat{\lambda}_i \hat{U}_i \hat{V}_i^T - M \right) \right\|_F^2$

$\hat{M} \leftarrow \sum_{i=1}^{r} \hat{s}_i \hat{\lambda}_i \hat{U}_i \hat{V}_i^T$

return $\hat{M}$

(1) $pd/\log n \to \infty$ *and* $n, d \to \infty$ *with* $d \leqslant n \leqslant e^{d^\beta}$, *where* $\beta < 1$ *free of* n, d, *and* p;

(2) $\lambda_i = b_i \sqrt{nd}$ *for all* $i = 1, \ldots, r$, *where* $\{b_i\}_{i=1,\ldots,r}$ *are positive bounded values;*

(3) $b_i > b_{i+1}$ *for all* $i = 1, \ldots, r$, *where* $b_{r+1} = 0$;

(4) $\lim_{n,d \to \infty} \mathbb{P}\Big(\min_{s \in \{-1,1\}^r} \left\| \mathcal{P}_\Omega \big(\sum_{i=1}^{r} s_i \hat{\lambda}_i \hat{U}_i \hat{V}_i^T - M \big) \right\|_F^2 < \left\| \mathcal{P}_\Omega \big(\sum_{i=1}^{r} s_{0i} \hat{\lambda}_i \hat{U}_i \hat{V}_i^T - M \big) \right\|_F^2 \Big) = 0,$

where $s = (s_1, \ldots, s_r)$ *and* $s_{0i} = sign(\langle \hat{V}_i, V_i \rangle)\, sign(\langle \hat{U}_i, U_i \rangle)$ *for* $i = 1, \ldots, r$.

Remark 17. *Under the setting where the rank* r *is fixed as in this paper, Assumption 3(2) implies that the underlying low-rank matrix* M_0 *is dense. More specifically, note that the squared Frobenius norm indicates both the sum of all squared entries of a matrix and the sum of its singular values squared. Also, note that* $\|M_0\|_F^2 = \sum_{i=1}^{r} \lambda_i^2(M_0) = cnd$ *for some constant* $c > 0$ *by Assumption 3(2). Thus, the sum of all squared entries of* M_0 *has an order* nd. *This means that a non-vanishing proportion of entries of* M_0 *contains non-vanishing signals with dimensionality (see Fan et al. (2013)). For more discussion, see Remark 2 in Cho et al. (2015a).*

Remark 18. *The singular vectors,* $\{\hat{U}_i\}_{i=1}^r$ *and* $\{\hat{V}_i\}_{i=1}^r$*, that compose* $\hat{M}$ *are consistent estimators of* U *and* V *up to signs (for details, see Cho et al. (2015a)). Hence, when combining them with* $\{\hat{\lambda}_i\}_{i=1}^r$ *to reconstruct* $\hat{M}$*, a sign problem happens. Assumption 3(4) assures that as* n *and* d *increase, the probability of choosing different signs than the true signs,* $\{s_{0i}\}_{i=1}^r$*, goes to zero. Given the asymptotic consistency of* $\{\hat{U}_i\}_{i=1}^r$*,* $\{\hat{V}_i\}_{i=1}^r$*, and* $\{\hat{\lambda}_i\}_{i=1}^r$*, this is not an unreasonable assumption to make.*

Proposition 2.3.1. *(Theorem 4.4 in Cho et al. (2015a)) Under Assumption 3 and the model setup in Section 3.5,* $\hat{M}$ *is a consistent estimator of* M_0*. In particular,*

$$\frac{1}{nd}\|\hat{M} - M_0\|_F^2 = o_p\left(\frac{h_n}{pd}\right),$$

where h_n *diverges very slowly with the dimensionality, for example,* $\log(\log d)$*.*

Remark 19. *Since* h_n *in Proposition 2.3.1 can be any quantity that diverges slowly with the dimensionality, the convergence rate of* $\hat{M}$ *can be thought of as* $1/pd$*. Under the setting where the rank of* M_0 *is fixed as in this paper, it is matched to the minimax error rate,* r/pd*, found in Koltchinskii et al. (2011a).*

Using $\hat{M}$ to initialize *Adaptive-Impute* has two major advantages. First, since $\hat{M}$ is already a consistent estimator of M_0 achieving the minimax error rate, it allows a series of the iterates of *Adaptive-Impute* coming after $\hat{M}$ to be also consistent estimators of M_0 achieving the minimax error rate (see Theorem 2.3.1). Second, because *Adaptive-Impute* is based on a non-convex optimization problem (see Section 2.4), its convergence may depend on initial values. $\hat{M}$ provides *Adaptive-Impute* a suitable initializer.

Adaptive thresholds

To motivate the novel thresholding scheme of *Adaptive-Impute*, we first consider the case where a fully-observed noisy low-rank matrix is available. Specifically, suppose that the probability of observing each entry, p, is 1 and thus $M_F = M_0 + \epsilon$

is observed. Under the model setup in Section 3.5 we can easily show that

$$\mathbb{E}(M_F^T M_F) = M_0^T M_0 + n\sigma^2 I_d \quad \text{and} \quad \mathbb{E}(M_F M_F^T) = M_0 M_0^T + d\sigma^2 I_n, \tag{2.2}$$

where I_d and I_n are identity matrices of size d and n, respectively. This shows that the eigenvectors of $\mathbb{E}(M_F^T M_F)$ and $\mathbb{E}(M_F M_F^T)$ are the same as the right and left singular vectors of M_0. Also, the top r eigenvalues of $\mathbb{E}(M_F^T M_F)$ consist of the squared singular values of M_0 and a noise, $n\sigma^2$, the latter of which is the same as the average of the bottom $d - r$ eigenvalues of $\mathbb{E}(M_F^T M_F)$. In light of this, we want the estimator of M_0 based on M_F to keep the first r singular vectors of M_F as they are, but adjust the bias occuring in the singular values of M_F. Thus, the resulting estimator is

$$\hat{M}^F = \sum_{i=1}^{r} \sqrt{\lambda_i^2(M_F) - \alpha}\, \mathbf{u}_i(M_F)\mathbf{v}_i(M_F)^T, \quad \text{where } \alpha = \frac{1}{d-r} \sum_{i=r+1}^{d} \lambda_i^2(M_F). \tag{2.3}$$

A simple extension of Proposition 2.3.1 shows that $\hat{M}^F$ achieves the best possible minimax error rate of convergence, $1/d$, since $p = 1$.

Now consider the cases where a partially-observed noisy low-rank matrix M is available. For each iteration $t \geqslant 1$, we fill out the unobserved entries of M with the corresponding entries of the previous iterate Z_t, treat the completed matrix $\widetilde{M}_t = \mathcal{P}_\Omega(M) + \mathcal{P}_\Omega^\perp(Z_t)$ as if it is a fully-observed matrix M_F, and find the next iterate Z_{t+1} in the same way that we found $\hat{M}^F$ from M_F in (2.3);

$$Z_{t+1} = \sum_{i=1}^{r} \sqrt{\lambda_i^2(\widetilde{M}_t) - \widetilde{\alpha}_t}\, \mathbf{u}_i(\widetilde{M}_t)\mathbf{v}_i(\widetilde{M}_t)^T, \quad \text{where } \widetilde{\alpha}_t = \frac{1}{d-r} \sum_{i=r+1}^{d} \lambda_i^2(\widetilde{M}_t). \tag{2.4}$$

Note that the difference in (2.4) from (2.3) is in the usage of $\widetilde{M}_t$ instead of M_F. Hence, the performance of *Adaptive-Impute* may depend on how close $\mathcal{P}_\Omega(Z_t)$ is to $\mathcal{P}_\Omega(M_0)$. Algorithm 4 summarizes these computing steps of *Adaptive-Impute* continued from Algorithm 3.

The following theorem illustrates that the iterates of *Adaptive-Impute* retain the

Algorithm 4 *Adaptive-Impute*

Require: M, y, r, and $\varepsilon > 0$

$Z_1 \leftarrow \hat{M}$ # from Algorithm 3

repeat for $t = 1, 2, \ldots$

$\widetilde{M}_t \leftarrow \mathcal{P}_\Omega(M) + \mathcal{P}_\Omega^\perp(Z_t)$

$V_i^{(t)} \leftarrow \mathbf{v}_i(\widetilde{M}_t), \quad \forall i \in \{1, \ldots, r\}$

$U_i^{(t)} \leftarrow \mathbf{u}_i(\widetilde{M}_t), \quad \forall i \in \{1, \ldots, r\}$

$\widetilde{\boldsymbol{\alpha}}_t \leftarrow \frac{1}{d-r} \sum_{i=r+1}^{d} \boldsymbol{\lambda}_i^2(\widetilde{M}_t)$

$\tau_{t,i} \leftarrow \boldsymbol{\lambda}_i(\widetilde{M}_t) - \sqrt{\boldsymbol{\lambda}_i^2(\widetilde{M}_t) - \widetilde{\boldsymbol{\alpha}}_t}, \qquad \forall i \in \{1, \ldots, r\}$ #

Adaptive thresholds

$\lambda_i^{(t)} \leftarrow \boldsymbol{\lambda}_i(\widetilde{M}_t) - \tau_{t,i} \left(= \sqrt{\boldsymbol{\lambda}_i^2(\widetilde{M}_t) - \widetilde{\boldsymbol{\alpha}}_t} \right), \quad \forall i \in \{1, \ldots, r\}$

$Z_{t+1} \leftarrow \sum_{i=1}^{r} \lambda_i^{(t)} U_i^{(t)} V_i^{(t)T}$

$t \leftarrow t + 1$

until $\|Z_{t+1} - Z_t\|_F^2 / \|Z_t\|_F^2 \leqslant \varepsilon$

return Z_{t+1}

statistical performance of the initializer $\hat{M}$.

Assumption 4. *For all* $i = 1, \ldots, r$, *sign*$(\langle \mathbf{u}_i(\widetilde{M}_t), U_i \rangle) =$ *sign*$(\langle \mathbf{v}_i(\widetilde{M}_t), V_i \rangle)$.

Theorem 2.3.1. *Under Assumptions 3-4 and the model setup in Section 3.5, we have for any fixed value of* t,

$$\frac{1}{nd} \|Z_t - M_0\|_F^2 = o_p\left(\frac{h_n}{pd}\right), \text{ as } n, d \to \infty \text{ with any } h_n \to \infty$$

where h_n *diverges very slowly with the dimensionality, for example,* $\log(\log d)$.

Remark 20. *Similarly as in Remark 19, since* h_n *is a quantity diverging very slowly, the convergence rate of* Z_t *can be thought of as* $1/pd$ *which is matched to the minimax error rate,* r/pd *(Koltchinskii et al. (2011a)).*

Non-convexity of *Adaptive-Impute*

We can view *Adaptive-Impute* as an estimation method via non-convex optimization.

For $t \geqslant 1$, define

$$\tau_{t,i} = \begin{cases} \lambda_i(\widetilde{M}_t) - \sqrt{\lambda_i^2(\widetilde{M}_t) - \widetilde{\alpha}_t}, & i \leqslant r \\ \lambda_{r+1}(\widetilde{M}_t), & i > r \end{cases}, \tag{2.5}$$

where $\widetilde{\alpha}_t = \frac{1}{d-r}\sum_{i=r+1}^{d} \lambda_i^2(\widetilde{M}_t)$ and $\widetilde{M}_t = \mathcal{P}_\Omega(M) + \mathcal{P}_\Omega^\perp(Z_t)$. Then, in each iteration *Adaptive-Impute* provides a solution to the problem

$$\min_{Z \in \mathbb{R}^{n\times d}} \frac{1}{2nd}\|\widetilde{M}_t - Z\|_F^2 + \sum_{i=1}^{d} \frac{\tau_{t,i}}{\sqrt{nd}}\frac{\lambda_i(Z)}{\sqrt{nd}}. \tag{2.6}$$

Note that the threshold parameters, $\tau_{t,i}$, have dependence on both the i-th singular value and the t-th iteration. The following theorem provides an explicit solution to (2.6).

Theorem 2.3.2. *Let X be an $n \times d$ matrix and let $n \geqslant d$. The optimization problem*

$$\min_{Z} \frac{1}{2nd}\|X - Z\|_F^2 + \sum_{i=1}^{d} \frac{\tau_i}{\sqrt{nd}}\frac{\lambda_i(Z)}{\sqrt{nd}} \tag{2.7}$$

has a solution which is given by

$$\hat{Z} = \Phi(\Delta - \boldsymbol{\tau})_+\Psi^T, \tag{2.8}$$

where $\Phi\Delta\Psi^T$ is the SVD of X, $\boldsymbol{\tau} = diag(\tau_1, \ldots, \tau_d) \in \mathbb{R}^{d\times d}$, $(\Delta - \boldsymbol{\tau})_+ = diag\big((\lambda_1(X) - \tau_1)_+, \ldots, (\lambda_d(X) - \tau_d)_+\big) \in \mathbb{R}^{d\times d}$, and $c_+ = \max(c, 0)$ for any $c \in \mathbb{R}$.

Remark 21. *To see how Theorem 2.3.2 provides a solution to* (2.6)*, let $X = \widetilde{M}_t$ and $\tau_i = \tau_{t,i}$ as specified in* (2.5)*. Then,* (2.6) *and* (2.7) *become the same and $\hat{Z}$ in* (2.8) *gives the explicit form of the $(t+1)$-th iterate, Z_{t+1}, in Algorithm 4.*

If all of the thresholding parameters in (2.6) are equal such that $\tau = \tau_{t,1} = \ldots = \tau_{t,d}$ for all $1 \leqslant i \leqslant d$ and $t \geqslant 1$, the optimization problem (2.6) becomes equivalent to that of *softImpute* (Mazumder et al. (2010)) and Theorem 2.3.2 provides

an iterative solution to it. While *softImpute* requires finding the right value of a thresholding parameter τ by using a cross validation (CV) technique which is time-consuming and often does not have a straightforward validation criteria, *Adaptive-Impute* suggests specific values of the thresholding levels as in (2.5). The novel thresholding scheme of *Adaptive-Impute* together with the rank constraint results in superior empirical performances over the existing *softImpute*-type algorithms (see Section 2.5).

The thresholding scheme of *Adaptive-Impute* can be viewed as a solution to a non-convex optimization problem since at every iteration it differentially and adaptively thresholds the singular values. As Hastie and others alluded to a similar issue for matrix completion methods via non-convex optimization in Hastie et al. (2014), it is hard to provide a direct convergence guarantee of *Adaptive-Impute*. So, in the following section we introduce a generalized-*softImpute* algorithm, simpler than *Adaptive-Impute* and yet still non-convex, and investigate its asymptotic convergence. It hints at the convergent behavior of *Adaptive-Impute* in the asymptotic sense.

2.4 Generalized *softImpute*

Generalized-*softImpute* is an algorithm which iteratively solves the problem,

$$\min_{Z\in\mathbb{R}^{n\times d}} Q_\tau(Z|Z_t^g) := \frac{1}{2nd}\left\|\mathcal{P}_\Omega(M)+\mathcal{P}_\Omega^\perp(Z_t^g)-Z\right\|_F^2+\sum_{i=1}^{d}\frac{\tau_i}{\sqrt{nd}}\frac{\lambda_i(Z)}{\sqrt{nd}}, \tag{2.9}$$

to ultimately solve the optimization problem,

$$\min_{Z\in\mathbb{R}^{n\times d}} f_\tau(Z) := \frac{1}{2nd}\|\mathcal{P}_\Omega(M)-\mathcal{P}_\Omega(Z)\|_F^2+\sum_{i=1}^{d}\frac{\tau_i}{\sqrt{nd}}\frac{\lambda_i(Z)}{\sqrt{nd}}. \tag{2.10}$$

Note that generalized-*softImpute* differentially penalizes the singular values, but the thresholding parameters do not change over iterations. The iterative solutions of generalized-*softImpute* are denoted by $Z_{t+1}^g := \arg\min_{Z\in\mathbb{R}^{n\times d}} Q_\tau(Z|Z_t^g)$ for $t \geqslant 1$ and Theorem 2.3.2 provides a closed form of Z_{t+1}^g. If $\tau_i = \tau$ for all $1 \leqslant i \leqslant d$,

generalized-*softImpute* will be equivalent to *softImpute* and both (2.9) and (2.10) become convex problems. However, by differentially penalizing the singular values, generalized-*softImpute* ends up solving a non-convex optimization problem. Theorem 2.4.1 below shows that despite the non-convexity of generalized-*softImpute*, the iterates of generalized-*softImpute*, $\{Z_t^g\}_{t\geqslant 1}$, converge to a solution of problem (2.10) under certain conditions.

Assumption 5. *Let* $\widetilde{M}_t^g = \mathcal{P}_\Omega(M) + \mathcal{P}_\Omega^\perp(Z_t^g)$ *and* $D_t^g := \widetilde{M}_t^g - Z_{t+1}^g$. *Then,*

$$\frac{1}{nd}\left\|D_t^g - D_{t+1}^g\right\|_F^2 + \frac{2}{nd}\langle D_t^g - D_{t+1}^g, Z_{t+1}^g - Z_{t+2}^g\rangle \geqslant 0 \quad \textit{for all } \ t \geqslant 1.$$

Theorem 2.4.1. *Let* Z_∞ *be a limit point of the sequence* Z_t^g. *Under Assumption 5, if the minimizer* Z^s *of* (2.10) *satisfies*

$$Z^s \in \left\{ Z \in \mathbb{R}^{n\times d} : \sum_{i=1}^{d} \tau_i\lambda_i(Z) \geqslant \sum_{i=1}^{d} \tau_i\lambda_i(Z_\infty) + \langle (Z - Z_\infty), D_\infty\rangle \right\}, \tag{2.11}$$

we have $f_\tau(Z_\infty) = f_\tau(Z^s)$ *and* $\lim_{t\to\infty} f_\tau(Z_t^g) = f(Z^s)$.

Remark 22. *If* $\tau_i = \tau$ *for all* i *as in case of* softImpute, *Assumption 5 and* (2.11) *are always satisfied because* $\frac{1}{\tau}D_t^g$ *belongs to the sub-gradient of* $\left\|Z_{t+1}^g\right\|_*$.

Remark 23. *If* Z^s *is unique, then generalized-*softImpute *finds the global minimum point of* (2.10) *by Theorem 2.4.1.*

Generalized-*softImpute* resembles *Adaptive-Impute* in a sense that both of them employ different thresholding parameters on $\lambda_i(Z)$'s. However, *Adaptive-Impute* updates these tuning parameters every iteration while generalized-*softImpute* does not. The following lemmas show that despite this difference, the convergent behavior of *Adaptive-Impute* is asymptotically close to that of generalized-*softImpute*.

Lemma 2.4.1. *Under Assumptions 3-4 and the model setup in Section 3.5, we have*

$$\left|\frac{\tau_{t,i}}{\sqrt{nd}} - \frac{\tau_{t+1,i}}{\sqrt{nd}}\right| = o_p\left(\sqrt{\frac{h_n}{pd}}\right) \quad \textit{for } \ i = 1,\ldots,d,$$

where $\tau_{t,i}$ *is defined in* (2.5).

Lemma 2.4.2. *Let* $D_t := \widetilde{M}_t - Z_{t+1}$, *where* $\widetilde{M}_t$ *and* Z_t *are as defined in Algorithm 4. Then, under Assumptions 3-4 and the model setup in Section 3.5, we have*

$$\frac{1}{nd}\|D_t - D_{t+1}\|_F^2 + \frac{2}{nd}\langle D_t - D_{t+1}, Z_{t+1} - Z_{t+2}\rangle + o_p\left(\frac{h_n}{pd}\right) \geqslant 0.$$

Lemma 2.4.1 shows that for large n and d, thresholding parameters of *Adaptive-Impute* are stable between iterations so that *Adaptive-Impute* behaves similarly to generalized-*softImpute*. Lemma 2.4.2 shows how Assumption 5 is adapted in *Adaptive-Impute*. It implies a possibility of *Adaptive-Impute* satisfying Assumption 5 asymptotically. Although this still does not provide a guarantee of convergence of *Adaptive-Impute*, numerical results below support this possibility.

2.5 Numerical results

In this section, we conducted simulations and a real-data analysis to compare *Adaptive-Impute* for estimating M_0 with the four different versions of *softImpute*:

1. *Adaptive-Impute*: the proposed algorithm, as summarized in Algorithm 4;
2. *softImpute*: the original *softImpute* algorithm (Mazumder et al. (2010));
3. *softImpute-Rank*: *softImpute* with rank restriction (Hastie et al. (2014));
4. *softImpute-ALS*: *Maximum-Margin Matrix Factorization* (Hastie et al. (2014));
5. *softImpute-ALS-Rank*: *rank-restricted Maximum-Margin Matrix Factorization* in Algorithm 3.1 (Hastie et al. (2014)).

SoftImpute algorithms were implemented with the `R` package, `softImpute` (Hastie and Mazumder (2015)). The R code for *Adaptive-Impute* is available at `https://github.com/chojuhee/hello-world/blob/master/adaptiveImpute_Rfunction`. In

this R code, we made two adjustments from Algorithms 3 and 4 for technical reasons. First, in almost all real world applications that needed matrix completion, the entries of M_0 are bounded below and above by constants L_1 and L_2 such that

$$L_1 \leqslant M_{0ij} \leqslant L_2$$

and smaller or larger values than the constants do not make sense. So, after each iteration of *Adaptive-Impute*, $t \geqslant 1$, we replace the values of Z_t that are smaller than L_1 with L_1 and the values of Z_t that are greater than L_2 with L_2. Second, the cardinality of the set, $\{-1, 1\}^r$, that we search over to find $\hat{s}$ in Algorithm 3 increases exponentially. Hence, finding $\hat{s}$ easily becomes a computational bottleneck of *Adaptive-Impute* or is even impossible for large r. We suggest two possible solutions to this problem. One solution is to find $\hat{s}$ by computing $\hat{s}_i = \text{sign}(\langle \hat{V}_i, \mathbf{v}_i(M) \rangle) \, \text{sign}(\langle \hat{U}_i, \mathbf{u}_i(M) \rangle)$ for $i = 1, \ldots, r$. Note that if we use V_i and U_i instead of $\mathbf{v}_i(M)$ and $\mathbf{u}_i(M)$, this gives us the true sign s_0 under Assumption 3. The other solution is to use a linear regression. Let a vector of the observed entries of M be the dependent variable and let a vector of the corresponding entries of $\hat{\lambda}_i \hat{U}_i \hat{V}_i^T$ be the i-th column of the design matrix for $i = 1, \ldots, r$. Then, we set $\hat{s}$ to be the coefficients of the regression line whose intercept is forced to be 0. The difference in the results of these two methods are negligible. In the following experiment, we only reported the results of the former solution for simplicity, while the R code provided in `https://github.com/chojuhee/hello-world/blob/master/adaptiveImpute_Rfunction` are written for both solutions.

Simulation study

To create $M_0 = AB^T \in \mathbb{R}^{n \times d}$, we sampled $A \in \mathbb{R}^{n \times r}$ and $B \in \mathbb{R}^{d \times r}$ to contain i.i.d. uniform$[-5, 5]$ random variables and a noise matrix $\epsilon \in \mathbb{R}^{n \times d}$ to contain i.i.d. $\mathcal{N}(0, \sigma^2)$. Then, each entry of $M_0 + \epsilon$ was observed independently with probability p. Across simulations, $n = 1700$, $d = 1000$, $r \in \{5, 10, 20, 50\}$, σ varies from 0.1 to 50, and p varies from 0.1 to 0.9. For each simulation setting, the data was sampled 100 times and the errors were averaged.

To evaluate performance of the algorithms, we measured three different types of errors; test, training, and total errors; the test error, $\texttt{Test}(\hat{M}) = \|\mathcal{P}_{\Omega}^{\perp}(\hat{M} - M_0)\|_F^2 / \|\mathcal{P}_{\Omega}^{\perp}(M_0)\|_F^2$, represents the distance between the estimate $\hat{M}$ and the parameter M_0 measured on the unobserved entries, the training error, $\texttt{Training}(\hat{M}) = \|\mathcal{P}_{\Omega}(\hat{M} - M_0)\|_F^2 / \|\mathcal{P}_{\Omega}(M_0)\|_F^2$, the distance measured on the observed entries, and the total error, $\texttt{Total}(\hat{M}) = \|\hat{M} - M_0\|_F^2 / \|M_0\|_F^2$, the distance measured on all entries. For ease of comparison, Figure 2.1 and 2.3 plot the relative efficiencies with respect to *softImpute*-Rank. For example, the relative test efficiency of *Adaptive-Impute* with respect to *softImpute*-Rank is defined as $\texttt{Test}(\hat{M}_{\texttt{rank}})/\texttt{Test}(\hat{M}_{\texttt{adapt}})$, where $\hat{M}_{\texttt{adapt}}$ is an estimate of *Adaptive-Impute* and $\hat{M}_{\texttt{rank}}$ is an estimate of *softImpute-Rank*. The relative total and training efficiencies with respect to *softImpute*-Rank are defined similarly.

We used the best tuning parameter for the algorithms in comparison. Specifically, for algorithms with rank restriction (including *Adaptive-Impute*), we provided the true rank (i.e. 5, 10, 20, or 50). For *softImpute*-type algorithms, an oracle tuning parameter was chosen to minimize the total error.

Figure 2.1 shows the change of the relative efficiencies as the probability of observing each entry, p, increases with $\sigma = 1$. Three columns of plots in Figure 2.1 correspond to three different types of errors and four rows of plots to four different values of the rank. In all cases, *Adaptive-Impute* outperforms the competitors and works especially better when p is small. Among *softImpute*-type algorithms, the algorithms with rank constraint (i.e. *softImpute*-Rank and *softImpute*-ALS-Rank) perform better than the ones without (i.e. *softImpute* and *softImpute*-ALS). Figure 2.2 shows the change of the absolute errors that are used to compute relative efficiencies in Figure 2.1 as the probability of observing each entry, p, increases.

Figure 2.3 shows the change of the log relative efficiencies as the standard deviation (SD) of each entry of ϵ, σ, increases with $p = 0.1$. When the noise level is under 15, *Adaptive-Impute* outperforms the competitors, but when the noise level is over 15, *softImpute*-type algorithms start to outperform *Adaptive-Impute*. Hence, *softImpute*-type algorithms are more robust to large noises than *Adaptive-Impute*. It may be because when there exist large noises dominating the signals, the conditions

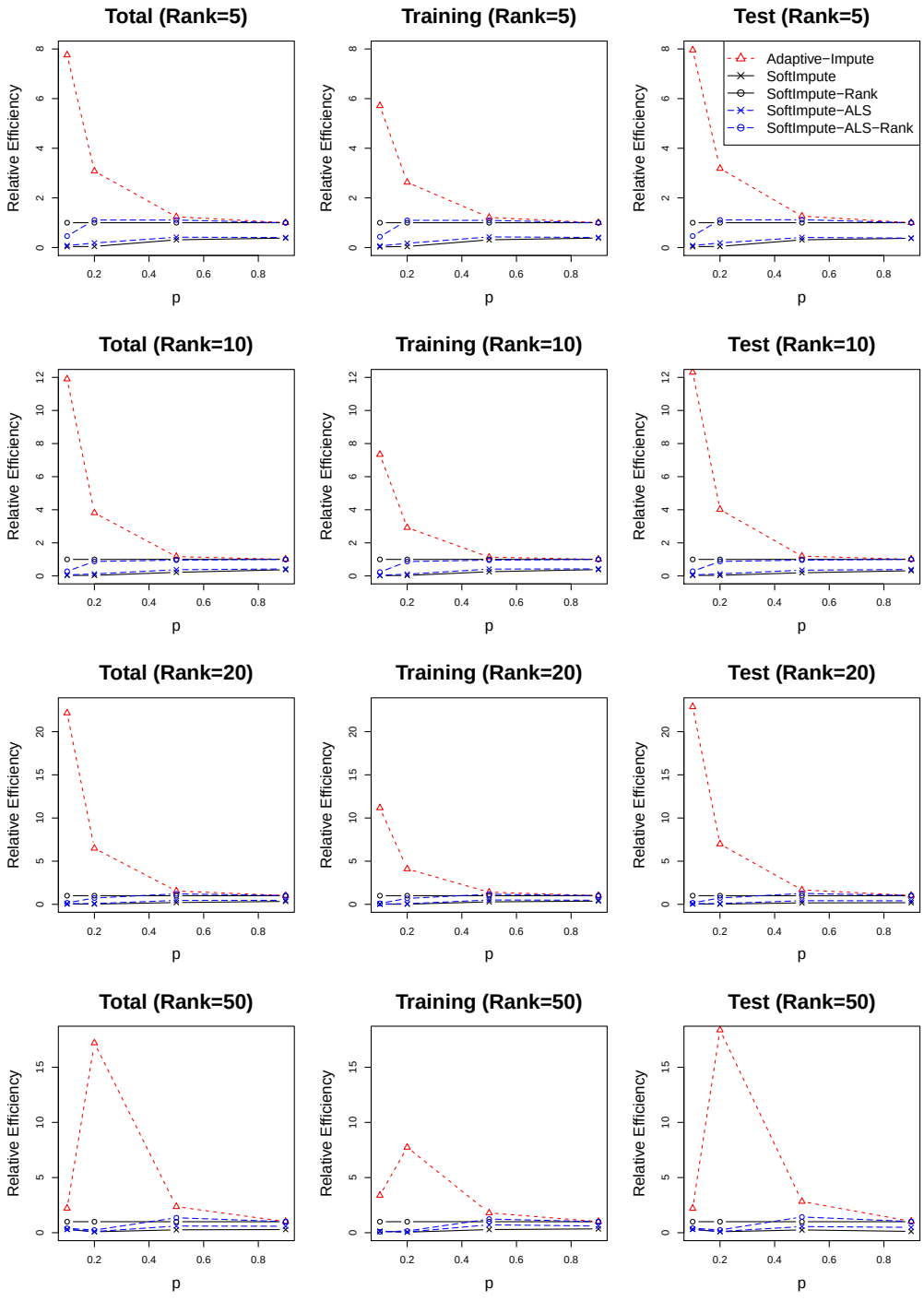

Figure 2.1: The relative efficiency plotted against the probability of observing each entry, p, when $\sigma = 1$. Training errors are measured over the observed entries, test errors over the unobserved entries, and total errors over all entries.

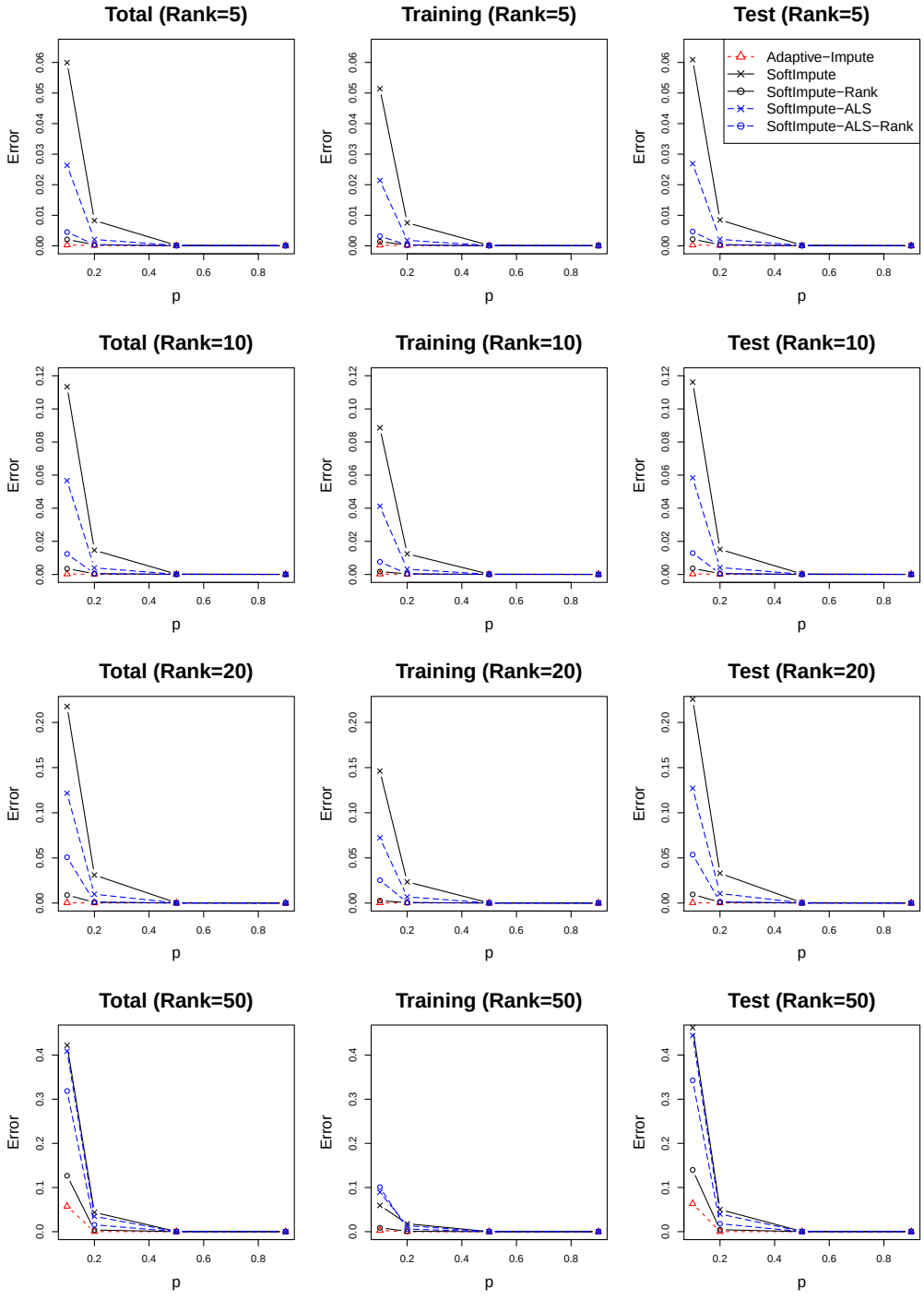

Figure 2.2: Change of the absolute errors when the probability of observing each entry, p increases and $\sigma = 1$.

for convergence presented in Section 2.4 are not satisfied. In real life applications, however, it is not common to observe such large noises that dominate the signals. Figure 2.4 shows the change of the absolute errors that are used to compute relative efficiencies in Figure 2.3.

Figure 2.5 shows convergence of the iterates of *Adaptive-Impute* to the underlying low-rank matrix over iterations; that is, the change of log $\texttt{Total}(Z_t)$, $\texttt{Training}(Z_t)$, and $\texttt{Test}(Z_t)$ errors as t increases. Across all plots, $n = 1700$, $d = 1000$, $p = 0.1$, and the errors were averaged over 100 replicates. In all cases, we observe that *Adaptive-Impute* converges well. Particularly, the smaller value of noise and/or rank is, the faster *Adaptive-Impute* converges.

A real data example

We applied *Adaptive-Impute* and the competing methods to a real data, MovieLens 100k (GroupLens (2015)). We used 5 training and 5 test data sets from 5-fold CV which are publicly available in GroupLens (2015). For the rank used in *Adaptive-Impute* and *softImpute*-type algorithms with rank constraint, we chose 3 based on a scree plot (Figure 2.6). Lemma 2 in Cho et al. (2015a) provides justification of using the scree plot and the singular value gap to choose the rank. For the thresholding parameters for *softImpute*-type algorithms, we chose the optimal values which result in the smallest test errors. The test errors were measured by normalized mean absolute error (NMAE) (Herlocker et al. (2004)),

$$\frac{1}{(M_{max} - M_{min})|\Omega_{test}|} \sum_{(i,j)\in\Omega_{test}} |\hat{M}_{ij} - M_{ij}|,$$

where the set Ω_{test} contains indices of the entries in test data, $|\Omega_{test}|$ is the cardinality of Ω_{test}, $M_{max} = \max\{\{M_{i,j}\} \setminus 0\}$ is the largest entry of M, and $M_{min} = \min\{\{M_{i,j}\} \setminus 0\}$ is the smallest entry of M.

Figure 2.7 summarizes the resulting NMAEs. Five points in the x-axis correspond to the 5-fold CV test data, the y-axis represents the values of NMAE, and the five different lines on the plane correspond to the 5 different algorithms in

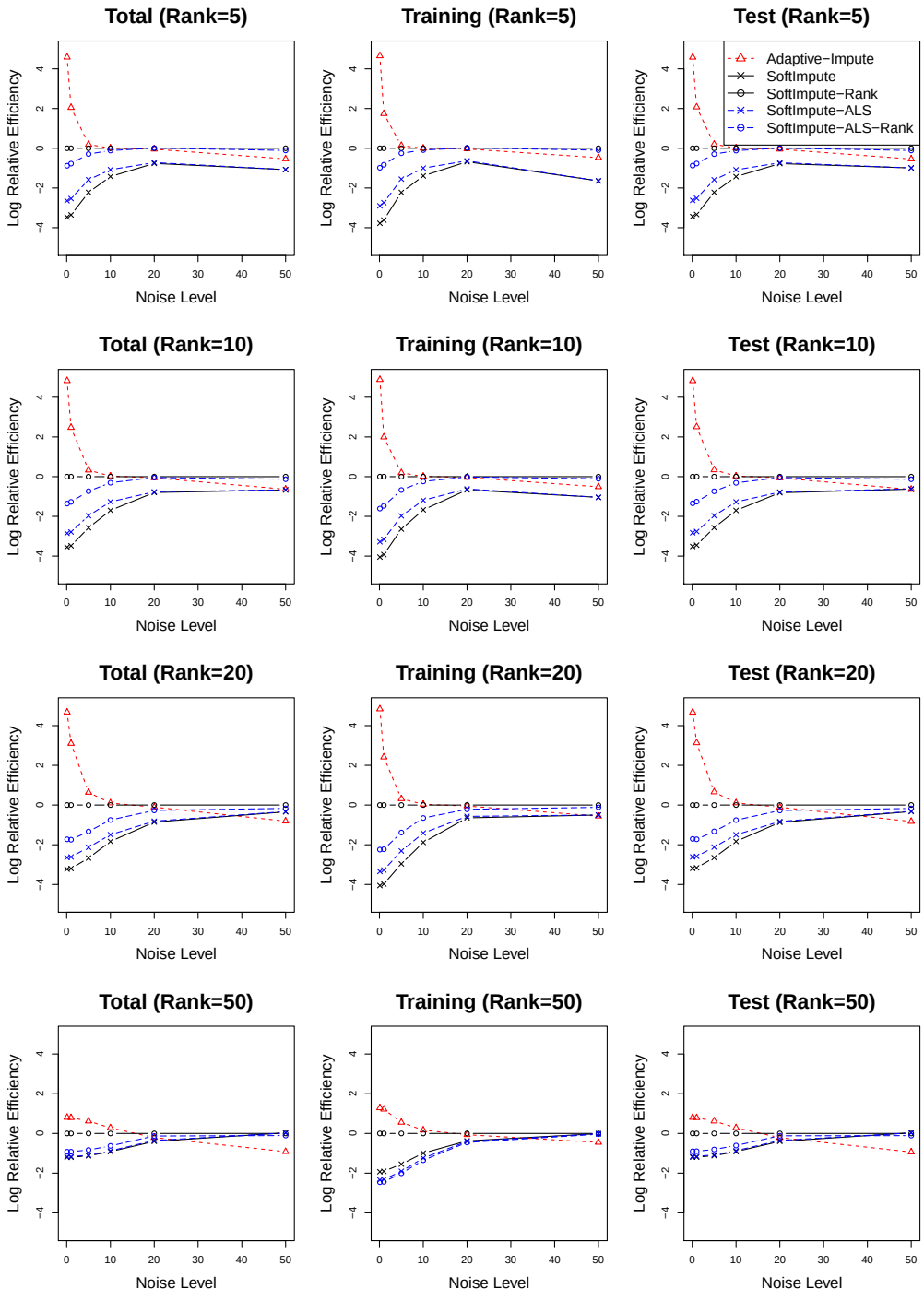

Figure 2.3: The log relative efficiency plotted against the SD of each entry of ϵ when $p = 0.1$. Training errors are measured over the observed entries, test errors over the unobserved entries, and total errors over all entries.

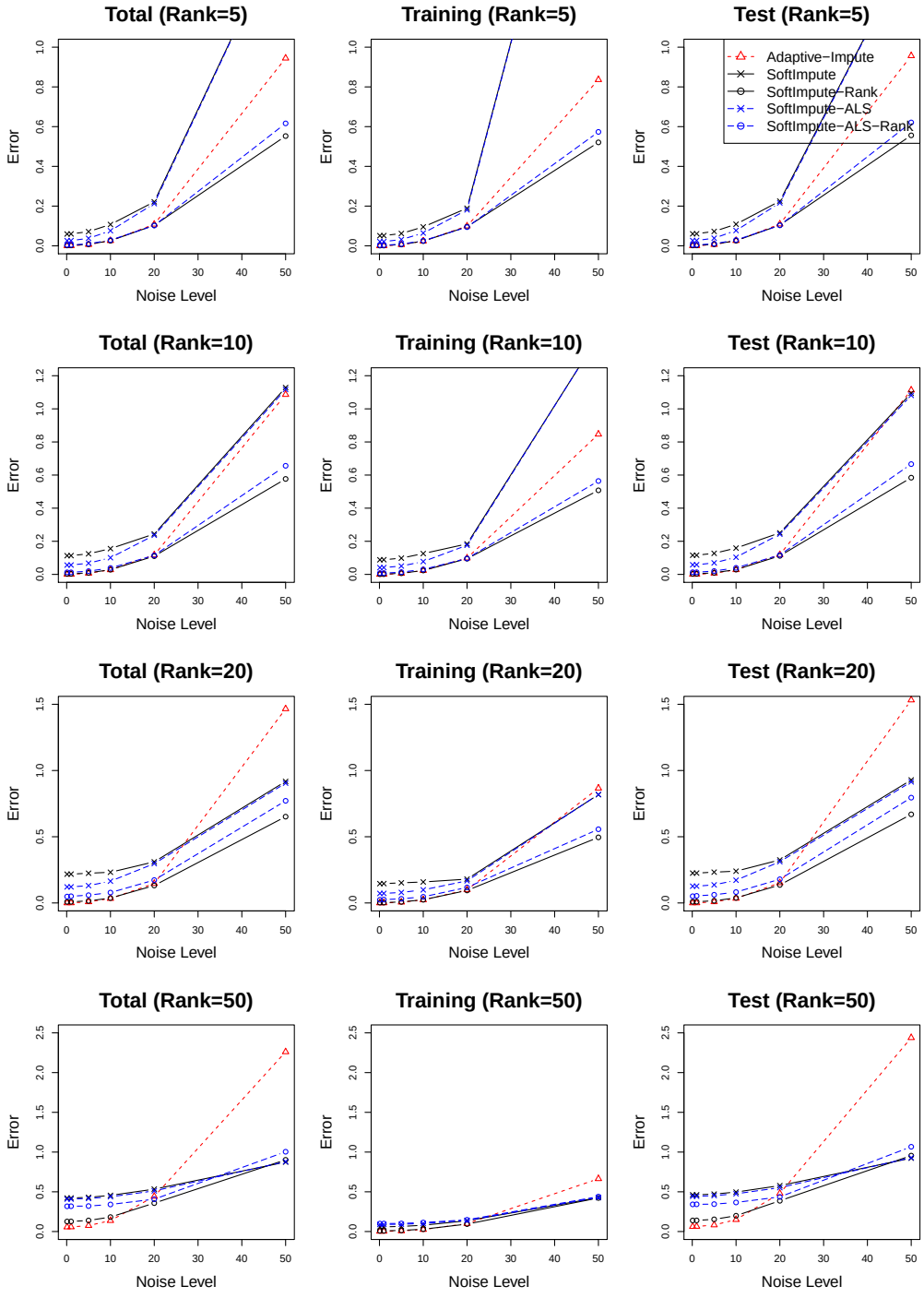

Figure 2.4: Change of the absolute errors when the SD of each entry of ϵ increases and $p = 0.1$.

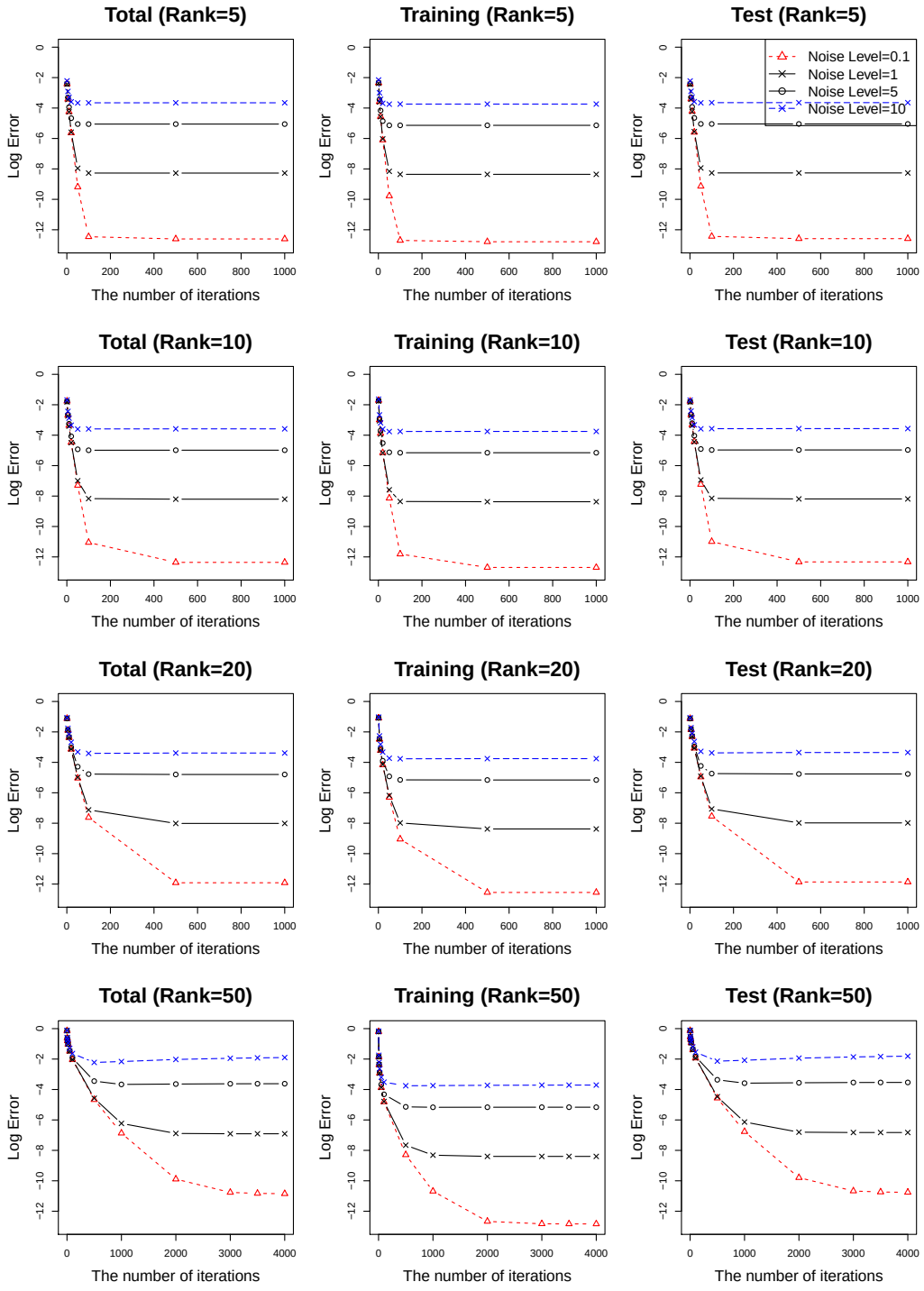

Figure 2.5: Convergence of the iterates of *Adaptive-Impute* to the underlying low-rank matrix. In all plots, $n = 1700$, $d = 1000$, $p = 0.1$, and all points were averaged over 100 replicates.

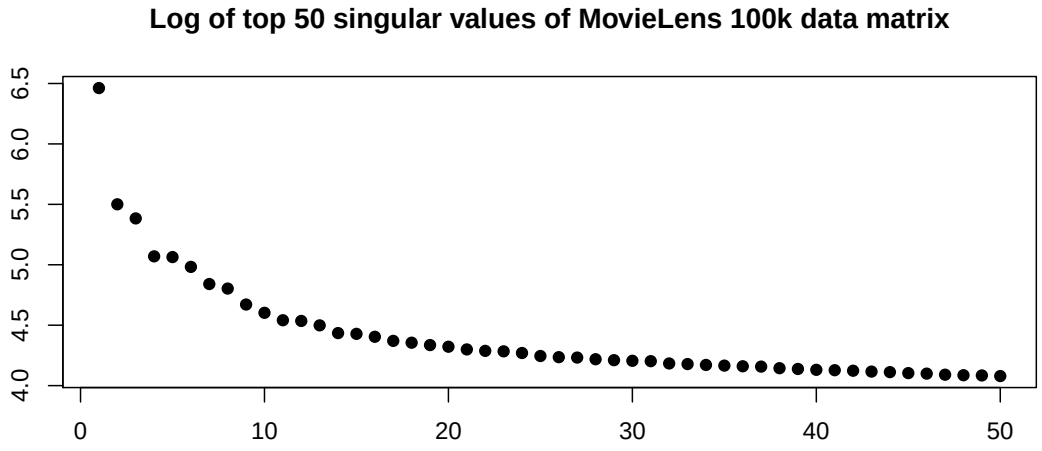

Figure 2.6: Log of the top 50 singular values of the MovieLens 100k data matrix (GroupLens (2015)).

comparison. We observe that *Adaptive-Impute* outperforms all of the other algorithms. Specifically, the test errors of *Adaptive-Impute* reduce those of *softImpute*-type algorithms by 6%-16%. Among *softImpute*-type algorithms, the ones with rank constraint (i.e. *softImpute*-Rank and *softImpute*-ALS-Rank) performs better than the ones without (i.e. *softImpute*and *softImpute*-ALS). This is the same result to the simulation results.

2.6 Discussion

Choosing the right thresholding parameter for matrix completion algorithms using thresholded SVD often poses empirical challenges. This paper proposed a novel thresholded SVD algorithm for matrix completion, *Adaptive-Impute*, which employs a theoretically-justified and data-dependent set of thresholding parameters. We established its theoretical guarantees on statistical performance and showed its

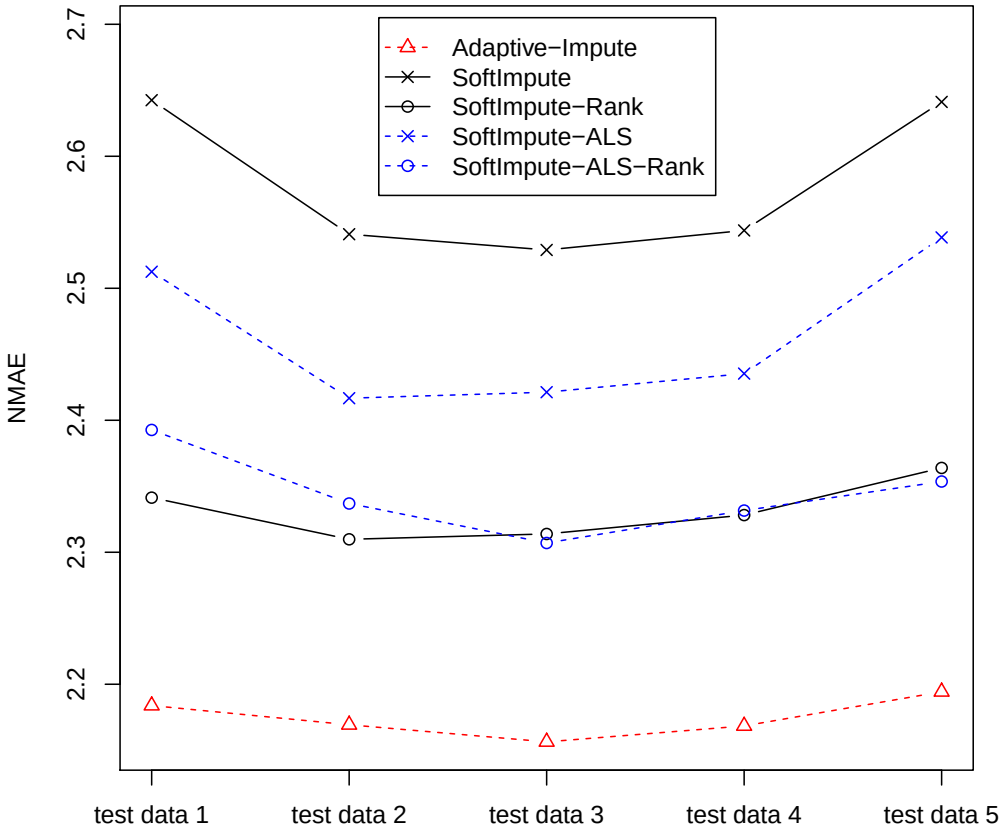

Figure 2.7: The NMAEs of *Adaptive-Impute* and its competitors measured in 5-fold CV test data from MovieLens 100k (GroupLens (2015)).

strong performances in both simulated and real data. It provides understanding on the effects of thresholding and the right threshold level. Yet, there is a newly open problem. Although we proposed a reasonable remedy in the paper, the choice of the rank of the underlying low-rank matrix is of another great practical interest. To estimate the rank and completely automate the entire procedure of *Adaptive-Impute*

would be a potential direction for future research.

3 ESTIMATING INDUCED SUBGRAPHS IN SAMPLES OF LABELED GRAPHS

3.1 Introduction

Network is a vibrant area in statistics, biology, and computer science. Networks describe connections among nodes, such as people's friendships (represented online by Facebook, Twitter, etc.) or functional relationships between proteins in a living cell. Recently, an emerging type of data in these fields is samples of labeled networks (or graphs) and the associated characteristics. The "labels" of networks imply that the nodes are labeled and that the same set of nodes reappear in all of the networks. Also, "labels" have a dual meaning that there are values (e.g. age, gender, or healthy v.s. sick) or vectors of values associated with each network.

The statistical literature has produced a number of relevant new models, methodologies, and theoretical results for network data (e.g. Hoff et al. (2002), Airoldi et al. (2009), Bickel and Chen (2009)). These results have focused on studying a network in isolation. However, in many cases, we encounter multiple networks. In this chapter, we suggest statistical methods to simultaneously analyze multiple networks or multiple networks together with the associated characteristics. Particularly, we propose estimation methods for induced subgraphs from sampled networks which vary with the associated characteristics. The proposed methods deploy three basic statistical tools to this novel data regime; principal component analysis (PCA), linear regression, and canonical correlation analysis (CCA). In case where only multiple networks are observed, PCA can find us induced subgraphs that have the most variability across all of the sampled networks If in addition to multiple networks, a characterizing value (e.g. age, gender, or healthy v.s. sick) to each network is also observed, we can study more direct relationship between networks and the associated characteristics using linear(trace) regression. If several different values are observed as an associated characteristic to each network, CCA can be a good way to analyze them. The most similar to our work is a series of

methods developed for graph classification (Vogelstein et al. (2013)). We can make use of the results from our suggested method for graph classification, but our interest is rather to understand the general variability across the networks and find induced subgraphs that are closely related to it than to focus on classification of graphs.

In the following sections, we introduce an interesting data that motivated our research (section 3.2), give a definition of induced subgraphs (section 3.3), develop PCA, linear(trace) regression, and CCA penalized by sparsity and low-rank constraints using the Alternating Direction Method of Multipliers (ADMM) in Boyd et al. (2011) (section 3.4), show the results of data analysis (section 3.5), and conclude with discussion (section 3.6).

Preliminaries The following notations will be used throughout the chapter. $\|\cdot\|_1$ denotes the ℓ_1 norm, $\|\cdot\|_{1,1}$ denotes a matrix version of the ℓ_1 norm, $\|\cdot\|_2$ denotes the ℓ_2 norm, $\|\cdot\|_*$ denotes the nuclear norm, and $\|\cdot\|_F$ denotes the Frobenius norm. For two matrices A and B of the same size, $\langle A, B\rangle = \mathrm{tr}(A^T B)$ denotes a matrix version of the inner product where $\mathrm{tr}(Z)$ for a square matrix Z is the sum of diagonals on Z. Let z be a vector in $\mathbb{R}^{p^2}$ and Z be a matrix in $\mathbb{R}^{p\times p}$. Then, $\mathcal{V}(Z) \in \mathbb{R}^{p^2}$ denotes a stack of columns of Z and $\mathcal{M}(z)$ denotes a matrix in $\mathbb{R}^{p\times p}$ whose first column is the first m elements of z, second column is the next m elements of z, and so on. The i-th row of a matrix Z is denoted by $Z_{i\cdot}$, the j-th column of Z is by $Z_{\cdot j}$, and the element in i-th row and j-th column of Z is by Z_{ij}.

Throughout, we study n number of samples of labeled graphs (or networks) and the associated characteristics. Define the k-th sampled graph as $G^{(k)}(V, E^{(k)})$ and the associated characteristic $y^{(k)} \in \mathbb{R}^q$, where $k \in \{1, \ldots, n\}$, $q \geqslant 1$, $V = \{v_1, v_2, \ldots, v_p\}$ is the node set that is common in all $G^{(k)}$, and $E^{(k)}$ is the edge set that contains a pair (i, j) if there is an edge from node v_i to node v_j in $G^{(k)}$. $E^{(k)}$ can be represented by the adjacency matrix $A^{(k)} \in \mathbb{R}^{p\times p}$. $A^{(k)}_{ij} > 0$ if (i, j) is in the edge set $E^{(k)}$ and $A^{(k)}_{ij} = 0$ otherwise. $E = \cup_k E^{(k)}$ is a collection of all edge sets. Define the diagonal matrices $R^{(k)} \in \mathcal{R}^{p\times p}$ and $O^{(k)} \in \mathcal{R}^{p\times p}$ and the normalized Graph

Laplacian $L^{(k)} \in \mathcal{R}^{p \times p}$ in the following way:

$$R_{ii}^{(k)} = \sum_j A_{ij}^{(k)}, \quad O_{jj}^{(k)} = \sum_i A_{ij}^{(k)}, \quad \text{and} \quad L^{(k)} = (R^{(k)})^{-1/2} A^{(k)} (O^{(k)})^{-1/2}.$$

The design matrices $X^{(A)}$ and $X^{(L)}$ in $\mathbb{R}^{n \times p^2}$ denote $\left(\mathcal{V}(A^{(1)}), \ldots, \mathcal{V}(A^{(n)}) \right)^T$ and $\left(\mathcal{V}(L^{(1)}), \ldots, \mathcal{V}(L^{(n)}) \right)^T$, respectively, so that the k-th row of $X^{(A)}$ is a vectorized adjacency matrix $A^{(k)}$ and the k-th row of $X^{(L)}$ is a vectorized normalized Graph Laplacian $L^{(k)}$. Sometimes, we simply use X to indicate a design matrix and in such cases, X can be interpreted as either $X^{(A)}$ or $X^{(L)}$. Regarding the associated characteristic $y^{(k)} \in \mathbb{R}^q$, if it is a scalar so that $q = 1$, $y = (y^{(1)}, \ldots, y^{(n)})^T$ denotes a vector in $\mathbb{R}^n$, and if it is a vector so that $q > 1$, $Y = (y^{(1)}, \ldots, y^{(n)})^T$ denotes a matrix in $\mathbb{R}^{n \times q}$.

3.2 Human transcriptional regulatory network data

The data used in this analysis consist of 41 human regulatory networks and their associated group labels. Specifically, Neph et al. (2012) assembled 41 human regulatory networks, each of which represent interactions between 475 transcription factors (TFs) in a human cell type, using genome-wide maps of in vivo DNaseI footprints. Then, they classified the 41 human cell types into 8 groups based on the developmental and functional properties. One of their goals was to understand *variability between* networks of different functional types and *similarity within* the same type and to find the TFs which play an pivotal role in it. So, they applied Ward clustering to 41 cell-type networks, each of which represented by the NND vector, a vector that encapsulates the relative number of interactions observed in a cell type for each of the 475 TFs (Alon (2006)). The resulting clusters strikingly parallel the functional groups (Figure 3.1) and the functionally-related cell types share similar core transcriptional regulatory networks. However, their analysis resulted from summarizing the networks in isolation and comparing the summaries rather than simultaneously analyzing all networks, the latter of which is a goal of this research.

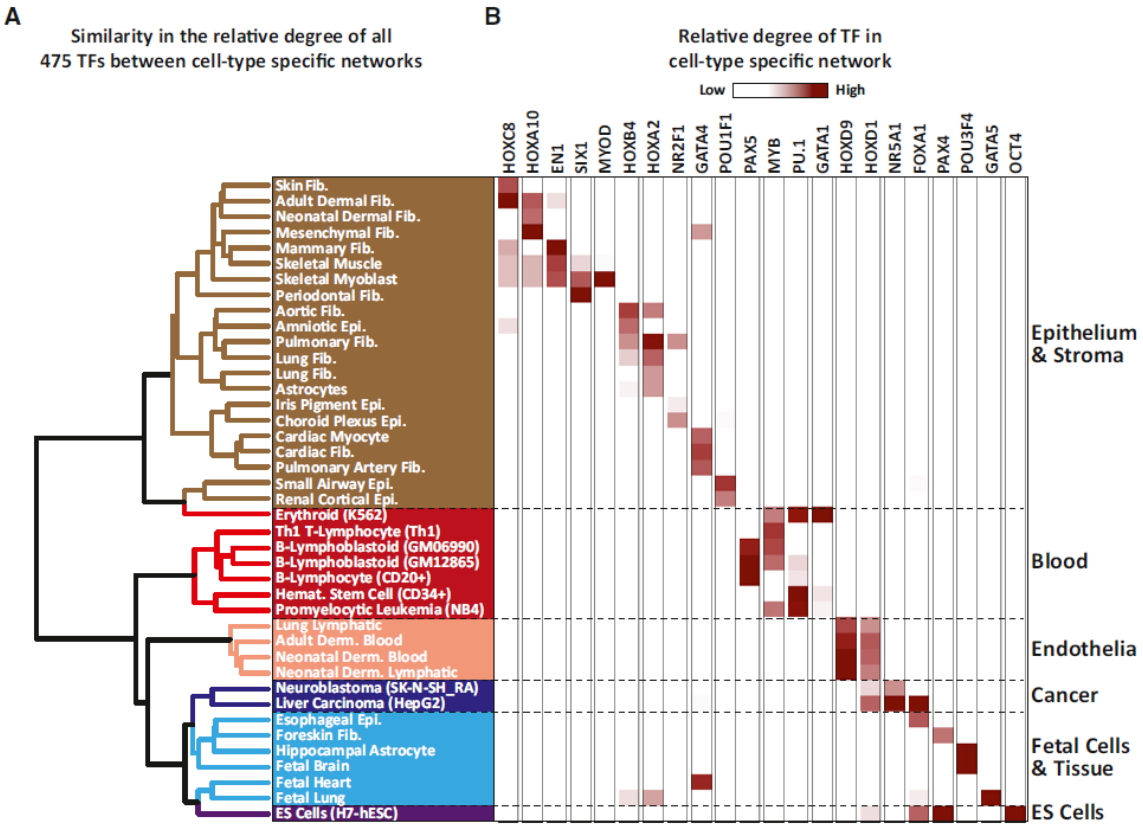

Figure 3.1: Figure 4 in Neph et al. (2012). Part A is obtained by taking an NND vector as a location of each cell type in $\mathbb{R}^{475}$ and performing Ward clustering. We can see that functionally related groups are almost perfectly clustered together. Part B shows how the master regulatory TFs contribute to the clustering of functionally related cell-type networks

In the following sections, we develop methods to estimate node-induced subgraphs whose interaction patterns differ depending on the associated characteristics of the regulatory networks. As a first step, we provide the exact definition of *induced subgraphs*

3.3 Induced subgraphs

Among all possible subgraphs, we are particularly interested in induced subgraphs of two types: A *node-induced subgraph* is a subset of the nodes in sampled graphs together with all edges whose endpoints are both in the subset. A formal definition of this is

$$G'(V', E') \quad \text{where} \quad V' \subset V \quad \text{and} \quad E' = \{(i, j) \in E \,|\, \forall i \in V', \forall j \in V'\}.$$

When the nodes (or TFs), edge density or graph pattern within which changes between the networks, are main interests, we would look for a node-induced subgraph.

An *edge-induced subgraph* is a subset of the edges in sampled graphs together with all nodes that are the edges' endpoints. A formal definition of this is

$$G''(V'', E'') \quad \text{where} \quad E'' \subset E \quad \text{and} \quad V'' = \{v_i, v_j \in V \,|\, \forall (i, j) \in E''\}.$$

When specific edges (or signals between brain spots) driving different characteristics between subjects are main interests, we would look for an edge-induced subgraph.

Since our interest is in finding master TFs, interaction patterns among which are different between networks of different functional types and similar within the same type, we focus on developing methods for estimating node-induced subgraphs. Hereafter, we simply call a node-induced subgraph as an induced subgraph.

3.4 Methods

Motivation

Suppose that we have paired samples $(A^{(k)}, y^{(k)}), k = 1, \ldots, n$ where $A^{(k)} \in \mathbb{R}^{p \times p}$, $y^{(k)} \in \mathbb{R}$, and

$$y^{(k)} = \langle A^{(k)}, B \rangle + \epsilon_k, \quad k \in \{1, \ldots, n\}, \tag{3.1}$$

for $B \in \mathbb{R}^{p\times p}$ and i.i.d. $\epsilon_k \in \mathbb{R}$ with mean zero and variance σ^2. Note that $\langle A^{(k)}, B\rangle$ is the sum of element-by-element multiplications between $A^{(k)}$ and B. So, B picks up all edges in $A^{(k)}$ that correspond to nonzero elements of B and drops off all edges in $A^{(k)}$ that correspond to zeros of B. In light of this, to find a node-induced subgraph which associates with the change of $y^{(k)}$, the coefficient matrix B should be blockmized. To estimate a blockmized B, we will apply sparsity and low-rank constraints, $\|\cdot\|_{1,1}$ and $\|\cdot\|_*$, on B.

Estimation

In this section, we introduce estimation methods to find induced subgraphs based on the adjacency matrices, $A^{(k)}, k = 1,\dots,n$, and the design matrix $X^{(A)}$ abbreviated to X for ease of notation. However, the same discussion can be applied to the normalized Graph Laplacians, $L^{(k)}, k = 1,\dots,n$, and the corresponding design matrix $X^{(L)}$. We also assume that y, Y, and X are centered throughout this section.

To find a blockmized B in (3.1), we solve the following problem

$$\min_{B\in\mathbb{R}^{p\times p}} \frac{1}{2}\sum_{k=1}^{n}\left(y^{(k)} - \langle A^{(k)}, B\rangle\right)^2 + \lambda\|B\|_{1,1} \quad \text{s.t.} \quad \|B\|_* \leqslant c \tag{3.2}$$

where $B = \mathcal{M}(\beta)$. Note that the latter equation of (3.2) is in fact a Trace regression with sparse and low-rank constraints.

The explicit form of solution to (3.2) does not exist, but using the ADMM, the problem can be solved. Let $\mathcal{N}_c = \{M \in \mathbb{R}^{p\times p} : \|M\|_* \leqslant c\}$ and $\mathbb{I}_{\mathcal{N}_c}$ be the convex indicator function which takes the values 0 and $+\infty$ on and off $\mathcal{N}_c$. Then, the problem (3.2) can be written into the ADMM problem

$$\min_{B,F\in\mathbb{R}^{p\times p}} \mathbb{I}_{\mathcal{N}_c}(B) + \frac{1}{2}\sum_{k=1}^{n}\left(y^{(k)} - \langle A^{(k)}, B\rangle\right)^2 + \lambda\|F\|_{1,1} \quad \text{s.t. } B - F = 0 \tag{3.3}$$

and its solution given by Boyd et al. (2011) is

$$
\begin{aligned}
B^{(new)} &= \arg\min_{B} \;\; \mathbb{I}_{\mathcal{N}_c}(B) + \frac{1}{2}\sum_{k=1}^{n}\left(y^{(k)} - \langle A^{(k)}, B\rangle\right)^2 + \frac{\rho}{2}\left\|B - F + U\right\|^2 \\
&= \arg\min_{B\in\mathcal{N}_c} \;\; \frac{1}{2}\sum_{k=1}^{n}\left(y^{(k)} - \langle A^{(k)}, B\rangle\right)^2 + \frac{\rho}{2}\left\|B - F + U\right\|^2 \\
&= \arg\min_{\mathcal{M}(\beta)\in\mathcal{N}_c} \;\; \frac{1}{2}\left\|y - X\beta\right\|^2 + \frac{\rho}{2}\left\|\beta - f + u\right\|^2 \\
&=: \mathcal{P}_{\mathcal{N}_c}\left\{\mathcal{M}\left[\left(I + \frac{1}{\rho}X^TX\right)^{-1}\left(\frac{1}{\rho}X^Ty + f - u\right)\right]\right\} \\
F^{(new)} &= \arg\min_{F} \;\; \lambda\left\|F\right\|_{1,1} + \frac{\rho}{2}\left\|F - B^{(new)} - U\right\|^2 \\
&=: \mathcal{S}_{\lambda/\rho}(B^{(new)} + U) \\
U^{(new)} &= U + B^{(new)} - F^{(new)}
\end{aligned}
$$

where $\beta = \mathcal{V}(B), f = \mathcal{V}(F), u = \mathcal{V}(U)$, and $\mathcal{P}_{\mathcal{N}_c}$ is the Euclidean projection onto $\mathcal{N}_c$. The entire steps are summarized in the algorithm 5.

Algorithm 5 Trace regression with sparsity and low-rank constraints

Require: $A, X, y, \lambda \geqslant 0, c > 0, \rho > 0, \epsilon > 0$

$F^{(0)} \leftarrow 0, U^{(0)} \leftarrow 0$ # Initialization

repeat $t = 1, 2, \ldots$

$B^{(t)} \leftarrow \mathcal{P}_{\mathcal{N}_c}\left\{\mathcal{M}\left[\left(I + \frac{1}{\rho}X^TX\right)^{-1}\left(\frac{1}{\rho}X^Ty + f^{(t-1)} - u^{(t-1)}\right)\right]\right\}$ # Projection

$F^{(t)} \leftarrow \mathcal{S}_{\lambda/\rho}\left(B^{(t)} + U^{(t-1)}\right)$ # Elementwise soft thresholding

$U^{(t)} \leftarrow U^{(t-1)} + B^{(t)} - F^{(t)}$ # Dual variable update

until $\|B^{(t)} - F^{(t)}\|_2^2 \vee \rho^2\|F^{(t)} - F^{(t-1)}\|_2^2 \leqslant \epsilon^2$ # Stopping criterion

return $F^{(t)}$

Extension

We can extend the idea of estimating induced subgraphs to the case where only sampled networks are available and the case where both sampled networks and associated vector-valued characteristics are available.

PCA with sparse and low-rank constraints

Suppose that we have multiple networks. To find induced subgraphs that vary across the networks, we propose the following sparse and low-rank PCA problem

$$\min_{u\in\mathbb{R}^n,\|u\|=1,v\in\mathbb{R}^{p^2}} \frac{1}{2}\left\|X-uv^T\right\|_F^2+\gamma\|v\|_1+\tau\|\mathcal{M}(v)\|_* \tag{3.4}$$

for a design matrix $X\in\mathbb{R}^{n\times p^2}$. Note that without the low-rank penalization term, $\tau\|\mathcal{M}(v)\|_*$, this problem setting is the same as sPCA-rSVD in Shen and Huang (2008). The solution for (3.4) is given by an iterative update as in Shen and Huang (2008);

$$v^{(new)}=\arg\min_{v\in\mathbb{R}^{p^2}} \frac{1}{2}\left\|X-uv^T\right\|_F^2+\gamma\|v\|_1+\tau\|\mathcal{M}(v)\|_* \tag{3.5}$$

$$\begin{aligned} u^{(new)} &= \arg\min_{u:\|u\|=1} \frac{1}{2}\left\|X-uv^{(new)^T}\right\|_F^2 \\ &= Xv^{(new)}/\left\|Xv^{(new)}\right\|_2 . \end{aligned}$$

Unlike the update $u^{(new)}$ where an explicit form of solution exists, the update $v^{(new)}$ does not have an explicit form, so we use the Alternating Direction Method of Multipliers (ADMM) from Boyd et al. (2011) to solve (3.5). The ADMM problem we solve is

$$\min_{q,w\in\mathbb{R}^{p^2}} \frac{1}{2}\left\|X-uq^T\right\|_F^2+\gamma\|w\|_1+\tau\|\mathcal{M}(q)\|_* \quad \text{s.t. } q-w=0\,.$$

The solution for this is (Boyd et al. (2011))

$$
\begin{aligned}
q^{(new)} &= \arg\min_{q} \ \frac{1}{2}\left\|X - uq^{T}\right\|_{F}^{2} + \tau\left\|\mathcal{M}(q)\right\|_{*} + \frac{\rho}{2}\left\|q - w + s\right\|_{2}^{2} \\
&= \sum_{j}\left(\sigma_{j}(Z) - \frac{\tau}{1+\rho}\right)_{+} u_{j}(Z)v_{j}(Z)^{T}, \quad \text{where } \ Z = \mathcal{M}\left(\frac{X^{T}u + \rho(w-s)}{1+\rho}\right) \\
w^{(new)} &= \arg\min_{w} \ \gamma\left\|w\right\|_{1} + \frac{\rho}{2}\left\|w - q^{(new)} - s\right\|_{2}^{2} \\
&=: \mathcal{S}_{\gamma/\rho}(q^{(new)} + s) \\
s^{(new)} &= s + q^{(new)} - w^{(new)}.
\end{aligned}
$$

where $(z)_{+} := \max(z, 0)$ for a $z \in \mathbb{R}$, $\sigma_{j}(Z)$ denotes the j-th singular value of Z, $u_{j}(Z)$ and $v_{j}(Z)$ denote the j-th left and right singular vectors of Z, respectively, and $\mathcal{S}_{\gamma/\rho}$ is the elementwise soft thresholding operator. The entire steps are summarized in the algorithm 6.

CCA with sparse and low-rank constraints

Suppose that data contain paired samples $(y^{(k)}, A^{(k)}), k = 1, \ldots, n$, where $y^{(k)} \in \mathbb{R}^{q}, q > 1$ is a vector. This creates two design matrices $X = \left(\mathcal{V}(A^{(1)}), \ldots, \mathcal{V}(A^{(n)})\right)^{T} \in \mathbb{R}^{n \times p^{2}}$ and $Y = (y^{(1)}, \ldots, y^{(n)})^{T} \in \mathbb{R}^{n \times q}$ and to find induced subgraphs, we propose sparse and low-rank CCA method.

Canonical correlation analysis (CCA) is a multivariate statistical model that facilitates the study of interrelationship among sets of multiple dependent variables and multiple independent variables (Lattin et al. (2003), Green (2014)). Whereas multiple regression predicts a single dependent variable from a set of multiple independent variables, canonical correlation simultaneously predicts multiple dependent variables from multiple independent variables. The CCA problem is defined as

$$
\max_{g \in \mathbb{R}^{q_1}, h \in \mathbb{R}^{q_2}} g^{T}X^{T}Yh \quad \text{s.t.} \quad g^{T}X^{T}Xg \leqslant 1 \quad \text{and} \quad h^{T}Y^{T}Yh \leqslant 1 \tag{3.6}
$$

Algorithm 6 PCA with sparsity and low-rank constraints

Require: $X, \gamma \geqslant 0, \tau \geqslant 0, \rho > 0, \epsilon_1 > 0, \epsilon_2 > 0$

Let $\sigma u^* v^{*\mathsf{T}}$ be the best rank-one approximation of X. Then,

$u^{(0)} \leftarrow u^*, v^{(0)} \leftarrow \sigma v^*$ # Initialization

repeat $t = 1, 2, \ldots$

repeat $\ell = 1, 2, \ldots$

$$q^{(\ell)} \leftarrow \sum_j \left(\sigma_j(Z^{(\ell-1)}) - \tfrac{\tau}{1+\rho}\right)_+ u_j(Z^{(\ell-1)}) v_j(Z^{(\ell-1)})^{\mathsf{T}}$$

$$\text{where } Z^{(\ell-1)} = \mathcal{M}\left(\tfrac{X^{\mathsf{T}} u^{(\ell-1)} + \rho(w^{(\ell-1)} - s^{(\ell-1)})}{1+\rho}\right)$$

$$w^{(\ell)} \leftarrow \mathcal{S}_{\lambda/\rho}(q^{(\ell)} + s^{(\ell-1)})$$

$$s^{(\ell)} \leftarrow s^{(\ell-1)} + q^{(\ell)} - w^{(\ell)}$$

until $\left\|q^{(\ell)} - w^{(\ell)}\right\|_2^2 \vee \rho^2 \left\|w^{(\ell)} - w^{(\ell-1)}\right\|_2^2 \leqslant \epsilon_1^2$

$v^{(t)} \leftarrow w^{(\ell)}$ # Update v

$u^{(t)} \leftarrow Xv^{(t)} / \left\|Xv^{(t)}\right\|_2$ # Update u

until $\left\|v^{(t)} - v^{(t-1)}\right\|_2^2 \vee \left\|u^{(t)} - u^{(t-1)}\right\|_2^2 \leqslant \epsilon_2^2$ # Stopping criterion

return $u^{(t)}, v^{(t)}$

where the two design matrices $X \in \mathbb{R}^{n \times q_1}$ and $Y \in \mathbb{R}^{n \times q_2}$ are centered and $q1$ and q_2 are at least two. It finds a set of weights g and h, for X and Y, respectively, so that the correlation between the linear combinations of Xg and Yh is maximized. Now, to find induced subgraphs, we impose sparsity and low-rank constrains to g so that the proposed sparse and low-rank CCA problem becomes

$$\max_{g \in \mathbb{R}^{p^2}, h \in \mathbb{R}^q} g^{\mathsf{T}} X^{\mathsf{T}} Y h - \omega \|g\|_1 - \mu \|\mathcal{M}(g)\|_* \quad \text{s.t. } g^{\mathsf{T}} g \leqslant 1 \text{ and } h^{\mathsf{T}} h \leqslant 1\,. \tag{3.7}$$

Note that this is a special case of the penalized CCA suggested in Witten et al. (2009). The solution to (3.7) is given by

$$g^{(new)} = \underset{g \in \mathbb{R}^{p^2}}{\arg\max} \quad g^{\mathsf{T}} X^{\mathsf{T}} Y h - \omega \|g\|_1 - \mu \|\mathcal{M}(g)\|_* \quad \text{s.t.} \quad g^{\mathsf{T}} g \leqslant 1 \tag{3.8}$$

$$\begin{aligned} h^{(new)} &= \underset{h \in \mathbb{R}^q}{\arg\max} \quad g^T X^T Y h \quad \text{s.t.} \quad h^T h \leqslant 1 \\ &= Y^T X g^{(new)} / \left\| Y^T X g^{(new)} \right\|_2 . \end{aligned}$$

Since there is no explicit form of solution for the update $g^{(new)}$, we solve the problem (3.8) using ADMM. The ADMM problem to solve (3.8) is

$$\min_{a, r \in \mathbb{R}^{p^2}} \quad -a^T X^T Y h + \omega \left\| r \right\|_1 + \mu \left\| \mathcal{M}(a) \right\|_* \quad \text{s.t.} \quad a^T a \leqslant 1$$

and the solution to this is given by Boyd et al. (2011) as follows;

$$\begin{aligned} a^{(new)} &= \underset{a : \|a\|_2 \leqslant 1}{\arg\min} \quad -a^T X^T Y h + \mu \left\| \mathcal{M}(a) \right\|_* + \frac{\rho}{2} \left\| a - r + e \right\|^2 \\ &= \frac{1}{\kappa} \sum_j \left(\sigma_j(M) - \frac{\mu}{\rho} \right)_+ u_j(M) v_j(M)^T, \text{ where } M = \mathcal{M}\left(\frac{X^T Y h + \rho(r - e)}{\rho} \right) \\ r^{(new)} &= \underset{r}{\arg\min} \quad \omega \left\| r \right\|_1 + \frac{\rho}{2} \left\| r - a^{(new)} - e \right\|^2 \\ &=: \mathcal{S}_{\omega/\rho}(a^{(new)} + e) \\ e^{(new)} &= e + a^{(new)} - r^{(new)} . \end{aligned}$$

where $\kappa = \left\| \sum_j \left(\sigma_j(M) - \frac{\mu}{\rho} \right)_+ u_j(M) v_j(M)^T \right\|_2$. The entire steps are summarized in the algorithm 7.

3.5 Results

Setup

Let $A^{(k)} \in \{0,1\}^{475 \times 475}, k = 1, \ldots, 41$ be the adjacency matrices representing the regulatory networks of 475 TFs from 41 tissue cells so that $A^{(k)}_{i,j} = 1$, if the i-th TF regulates the j-th TF in the k-th cell, and $= 0$, otherwise. In the analysis, to rule out the worries about the degree imbalance between the TFs, we used the normalized Graph Laplacians, $L^{(k)} \in \{0,1\}^{475 \times 475}, k = 1, \ldots, 41$, derived from

Algorithm 7 CCA with sparsity and low-rank constraints

Require: $X, Y, \omega \geqslant 0, \mu \geqslant 0, \rho > 0, \epsilon_1 > 0, \epsilon_2 > 0$

Let g^* and h^* be the solutions to the CCA problem (3.6). Then,

$g^{(0)} \leftarrow g^* \in \mathbb{R}^{p^2}$, $h^{(0)} \leftarrow h^* \in \mathbb{R}^q$ # Initialization

repeat $t = 1, 2, \ldots$

  repeat $\ell = 1, 2, \ldots$

    $a^{(\ell)} \leftarrow \frac{1}{\kappa^{(\ell)}} \sum_j \left(\sigma_j(M^{(\ell-1)}) - \frac{\mu}{1+\rho}\right)_+ u_j(M^{(\ell-1)}) v_j(M^{(\ell-1)})^T$

      where $M^{(\ell-1)} = \mathcal{M}\left(\frac{X^T Y h^{(\ell-1)} + \rho(r^{(\ell-1)} - e^{(\ell-1)})}{\rho}\right)$

    $r^{(\ell)} \leftarrow \mathcal{S}_{\omega/\rho}(a^{(\ell)} + e^{(\ell-1)})$

    $e^{(\ell)} \leftarrow a^{(\ell-1)} + r^{(\ell)} - e^{(\ell)}$

  until $\left\|a^{(\ell)} - r^{(\ell)}\right\|_2^2 \vee \rho^2 \left\|r^{(\ell)} - r^{(\ell-1)}\right\|_2^2 \leqslant \epsilon_1^2$

  $g^{(t)} \leftarrow r^{(\ell)}$ # Update g

  $h^{(t)} \leftarrow Y^T X g^{(t)} / \left\|Y^T X g^{(t)}\right\|_2$ # Update h

until $\left\|g^{(t)} - g^{(t-1)}\right\|_2^2 \vee \left\|h^{(t)} - h^{(t-1)}\right\|_2^2 \leqslant \epsilon_2^2$ # Stopping criterion

return $g^{(t)}, h^{(t)}$

$A^{(k)} \in \{0,1\}^{475\times 475}, k = 1, \ldots, 41$.

For the associated characteristic $y^{(k)}$ to the adjacency matrix $A^{(k)}$, we could employ 8 functional group labels given by Neph et al. (2012), but instead of using them, we found and used real-valued associated characteristics that remarkably parallel the funcional grouping. Specifically, we first computed the similarities between $L^{(k)}, k = 1, \ldots, 41$ by computing $S_{ij} = \langle L^{(i)}, L^{(j)} \rangle$ and created a similarity matrix $S = \{\langle L^{(i)}, L^{(j)} \rangle\}_{i,j=1}^{41} \in \mathbb{R}^{41\times 41}$. If we had computed the similarity S_{ij} with $\langle A^{(i)}, A^{(j)} \rangle$ instead of $\langle L^{(i)}, L^{(j)} \rangle$, S_{ij} would have indicated the number of common ones (edges) between $A^{(i)}$ and $A^{(j)}$, that is, the number of the same TF interaction patterns observed in both i-th and j-th cell types. Since $L^{(k)}$ are the normalized version of $A^{(k)}$, the similarities $S_{ij} = \langle L^{(i)}, L^{(j)} \rangle$ can be considered as a normalized version of $\langle A^{(i)}, A^{(j)} \rangle$. Now, Eigen-decompose S and let

$$S = \Gamma D \Gamma^T = (\Gamma D^{1/2})(\Gamma D^{1/2})^T = \Lambda \Lambda^T \quad \text{where} \quad \Lambda = \Gamma D^{1/2}, \tag{3.9}$$

where $\Gamma \in \mathbb{R}^{41\times 41}$ contains the eigenvectors of S and $D \in \mathbb{R}^{41\times 41}$ contains the eigenvalues of S. Take and plot the 1st column of Λ, and we obtain Figure 3.2. It shows that the 8 functional groups of cell types are well-clustered, and they

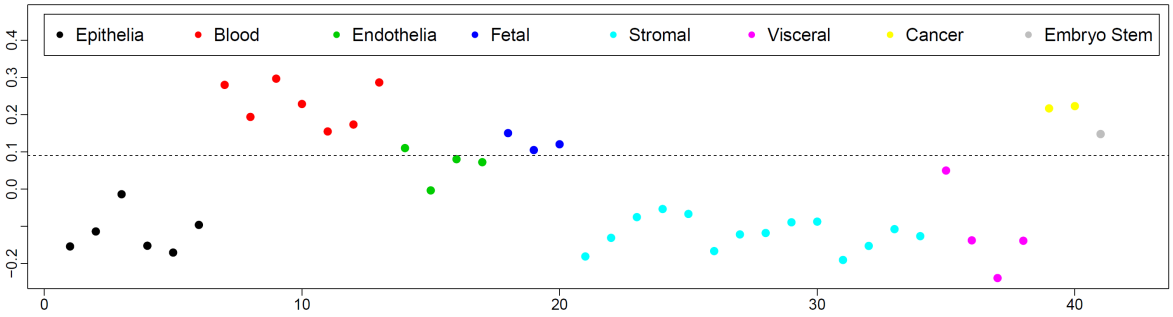

Figure 3.2: Plotting the first column of Λ.

can be more divided into two larger groups than 8 functional groups, Epithelia, Endothelia, Stromal, and Visceral cells under the dotted line versus the rest cell types above the dotted line. We set the 1st column of Λ as a response variable y and found the induced subgraph which drives such patterns of y using (3.2).

Under this setup, the estimation problem (3.2) can be written similarly as a sparse and low-rank PCA. Specifically, let $X^{(L)} = \left(\mathcal{V}(L^{(1)}), \ldots, \mathcal{V}(L^{(n)})\right)^T = U\Sigma V^T$ by singular value decomposition. Then, by construction,

$$S = \{\langle L^{(i)}, L^{(j)}\rangle\}_{i,j=1}^{41} = X^{(L)}{X^{(L)}}^T = (U\Sigma)(U\Sigma)^T = (X^{(L)}V)(X^{(L)}V)^T.$$

Since $S = \Lambda\Lambda^T$ from (3.9), by taking y as the 1st column of Λ, the model (3.1) can be viewed as

$$y = X^{(L)}\beta + \epsilon = X^{(L)}V_{\cdot 1} + \epsilon = \begin{pmatrix} \langle L^{(1)}, B\rangle \\ \langle L^{(2)}, B\rangle \\ \vdots \\ \langle L^{(n)}, B\rangle \end{pmatrix} + \epsilon$$

and estimation problem (3.2) as a procedure of approximating $V_{\cdot 1}$ with a sparse and low-rank $B = \mathcal{M}(\beta)$. Note that the sparse and low-rank PCA in eqrefeqn:sl-pca finds loading vectors that maximize the variance of the corresponding principal component and that is sparse and has low-rank when transformed into a matrix. In this case, we do not know in advance what the resulting principal component would be. Contrarily, in the case above where $y = \Lambda_{\cdot 1}$ and $X = X^{(L)}$, we first find the principal component by the ordinary PCA method and then make use of it to estimate the sparse and low-rank loading vector. That is, in the latter case, we try to recover the loading vector that gives the given principal component and that satisfies the sparse and low-rank assumptions.

Below, we show the analysis results.

Results

To choose the best tuning parameters, we performed 5-fold cross validation with the loss function, $\|y^{(test)} - X^{(test)}\hat{\beta}^{(train)}\|^2$ and chose the values which minimized the loss. The resulting estimate, $\hat{B} = \mathcal{M}(\hat{\beta})$, had 69 nonzero elements and was nearly of rank 4, meaning that the singular values except for the first 4 were much smaller than 10^{-6}. Figure 3.3 shows the selected 69 TF interactions across 41 cell types that correspond to the nonzero elements of $\hat{B}$ so that the i-th column of Figure 3.3 shows which edges among 69 are active and which edges are not active in each cell type. We can observe that Epithelia, Endothelia, Stromal, and Visceral cells share similar graph pattern around part B, and so do the rest of cell types. Edges around part A are only active in Blood cells.

In particular, we choose two subgraphs from Figure 3.3; the first subgraph, denoted by A, distinguishes blood cells from the rest and the second subgraph, denoted by B, distinguishes a group of Epithelia, Stromal, and Visceral cells from the rest. These two subgraphs are depicted across 41 cells in Figure 3.4 and 3.5.

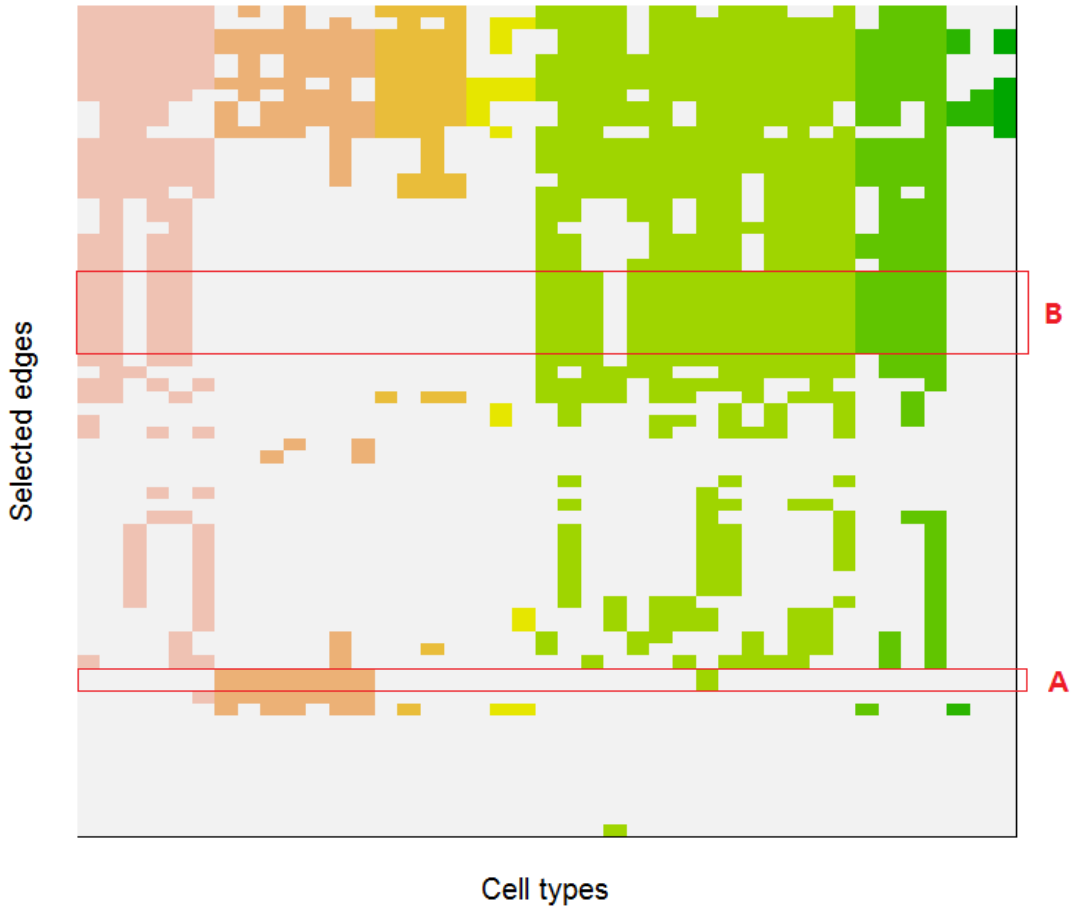

Figure 3.3: (sparse and low-rank linear regression) The graph presents how the 69 edges selected by the estimate $\hat{B}$ behave in each of 41 cell types. From the left, Epithelia, Blood, Endothelia, Fetal, Stromal, Visceral, Cancer, Embryonic Stem cells are presented. The edges belonging to A and B distinguish Blood cells from the rest and a group of Epithelia, Stromal, and Visceral cells from the rest, respectively.

3.6 Discussion

We proposed and implemented methods to estimate induced subgraphs from samples of labeled graphs. Specifically, we developed sparse and low-rank Linear(trace) Regression, PCA, and CCA using the ADMM. Applying these methods to a real

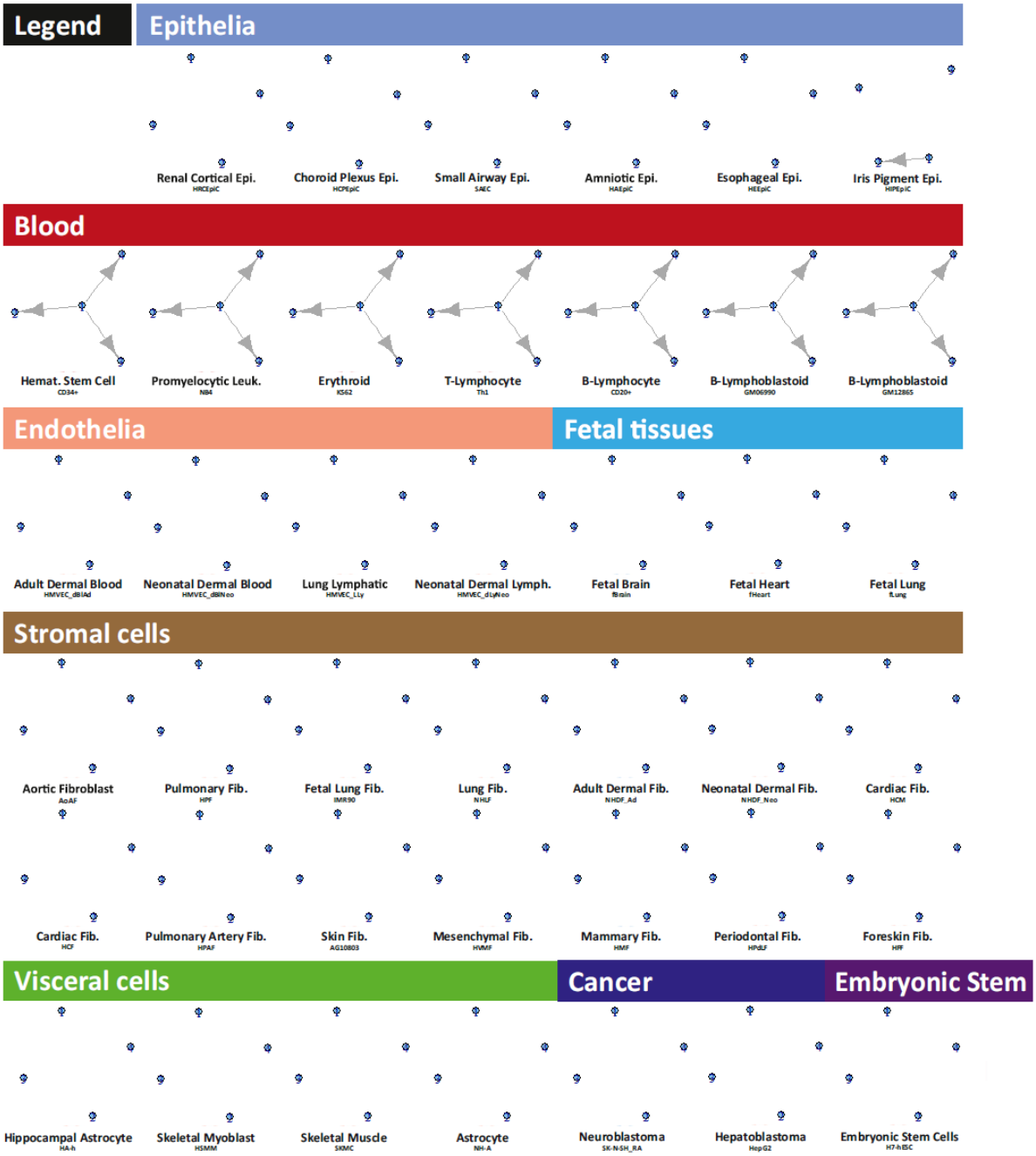

Figure 3.4: A signature subgraph A which distinguish blood cells from the rest.

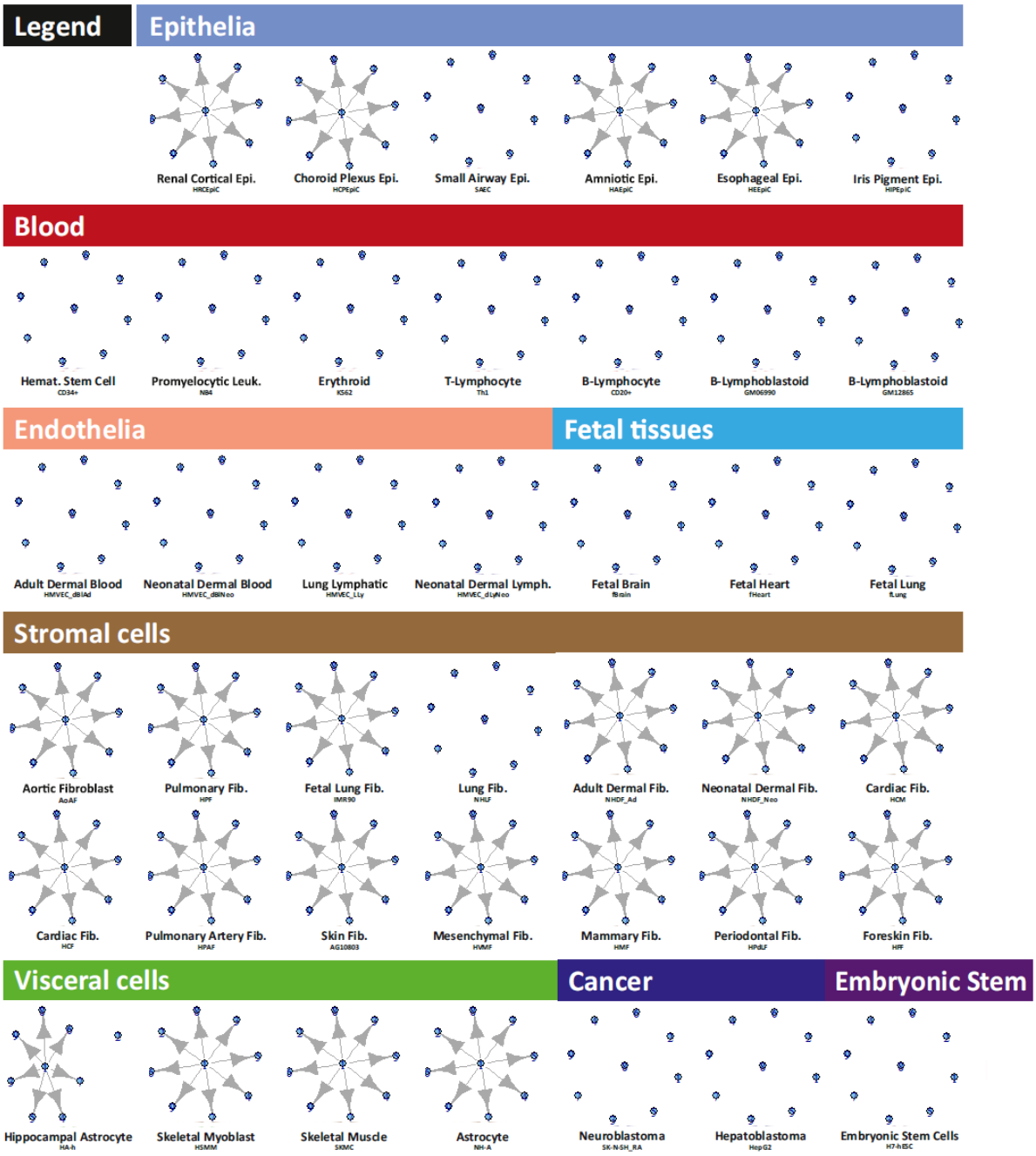

Figure 3.5: A signature subgraph B which distinguish Epithelia, Stromal, Visceral cells from the rest.

data, 41 human transcriptional factor regulatory networks, gave us interesting results. Still, for theoretical guarantees of the proposed estimation methods, a further investigation should follow. We leave this as a future research.

A APPENDIX FOR CHAPTER 1

Proofs for Lemma 1.3.1

Proof of Lemma 1.3.1. Let

$$M_y = [(y_{kh} - p)M_{0kh}]_{1\leqslant k\leqslant n, 1\leqslant h\leqslant d} \quad \text{and} \quad \epsilon_y = [y_{kh}\epsilon_{kh}]_{1\leqslant k\leqslant n, 1\leqslant h\leqslant d},$$

both in $\mathbb{R}^{n\times d}$. Then, $M = pM_0 + M_y + \epsilon_y$ and

$$\begin{aligned}\hat{\Sigma} &= p^2 M_0^T M_0 + M_y^T M_y + \epsilon_y^T \epsilon_y \\ &\quad + pM_0^T M_y + pM_y^T M_0 + pM_0^T \epsilon_y + p\epsilon_y^T M_0 + M_y^T \epsilon_y + \epsilon_y^T M_y. \quad (A.1)\end{aligned}$$

The result (1.2) follows since under the model setup in Section 3.5,

$$\mathbb{E}M_y = 0, \ \ \mathbb{E}\epsilon_y = 0, \ \ \mathbb{E}(M_y^T M_y) = p(1-p)\,\mathrm{diag}(M_0^T M_0), \ \ \mathbb{E}(\epsilon_y^T \epsilon_y) = np\sigma^2 I_d,$$

$$\mathbb{E}(M_0^T M_y) = 0, \quad \mathbb{E}(M_0^T \epsilon_y) = 0, \quad \text{and} \quad \mathbb{E}(M_y^T \epsilon_y) = 0.$$

We can similarly show the result (1.3). □

Proofs for Section 1.6

Proof of Proposition 1.6.1. We only show the result (1.16), since the other result can be shown similarly.

Let

$$E = \left\{ \frac{1}{nd} |\ddot{\lambda}_{p\,m+1}^2 - \lambda_{p\,m+1}^2| < t \right\},$$

where $t = C_1 p\frac{\log n}{d} + C_2 p^{3/2}\sqrt{\frac{\log n}{n}}$. Note that $\frac{t}{p^2} \to 0$. By Weyl's theorem (Li (1998a)) and Lemma 1.6.1, we have for large constants $C_1, C_2 > 0$,

$$\mathbb{P}(E^c) \leqslant \mathbb{P}\left(\frac{1}{nd} \left\| \hat{\Sigma}_p - \mathbb{E}\hat{\Sigma}_p \right\|_2 \geqslant t \right) = O(n^{-2}).$$

Thus, for large n,

$$
\begin{aligned}
&\mathbb{E}\left\|\sin\left(V_p^{(m)}, V^{(m)}\right)\right\|_F^2 \\
&= \mathbb{E}\left\{ \left\|\sin\left(V_p^{(m)}, V^{(m)}\right)\right\|_F^2 \ \mathbb{1}_{E^c} \right\} + \mathbb{E}\left\{ \left\|\sin\left(V_p^{(m)}, V^{(m)}\right)\right\|_F^2 \ \mathbb{1}_{E} \right\} \\
&\leqslant m\,\mathbb{P}\left(E^c\right) + \mathbb{E}\left\{ \frac{\left\| \frac{1}{nd}\frac{1}{p^2}\left(\hat{\Sigma}_p - \mathbb{E}\,\hat{\Sigma}_p\right) V^{(m)} \right\|_F^2 \ \mathbb{1}_E}{\left(\frac{1}{nd}\frac{1}{p^2}|\ddot{\lambda}_{p\,m}^{\,2} - \lambda_{p\,m+1}^2|\right)^2} \right\} \\
&\leqslant m\,\mathbb{P}\left(E^c\right) + \mathbb{E}\left\{ \frac{\left\| \frac{1}{nd}\frac{1}{p^2}\left(\hat{\Sigma}_p - \mathbb{E}\,\hat{\Sigma}_p\right) V^{(m)} \right\|_F^2 \ \mathbb{1}_E}{\left(\frac{1}{nd}\frac{1}{p^2}\left|\ddot{\lambda}_{p\,m}^{\,2} - \ddot{\lambda}_{p\,m+1}^{\,2}\right| - \frac{t}{p^2}\right)^2} \right\} \\
&\leqslant m\,\mathbb{P}\left(E^c\right) + \mathbb{E}\left\{ \frac{\left\| \frac{1}{nd}\frac{1}{p^2}\left(\hat{\Sigma}_p - \mathbb{E}\,\hat{\Sigma}_p\right) V^{(m)} \right\|_F^2 \ \mathbb{1}_E}{\left(\frac{1}{2nd}\frac{1}{p^2}\left|\ddot{\lambda}_{p\,m}^{\,2} - \ddot{\lambda}_{p\,m+1}^{\,2}\right|\right)^2} \right\} \\
&\leqslant cn^{-2} + \frac{Cn^{-1}}{p\,(b_m^2 - b_{m+1}^2)^2}, \qquad \text{(A.2)}
\end{aligned}
$$

where $\mathbb{1}_E$ is an indicator function of an event E, the first inequality is due to the fact that $\|\sin(\hat{V}^{(m)}, V_p^{(m)})\|_F^2 \leqslant m$ and Davis-Kahan $\sin\theta$ theorem (Theorem 3.1 in Li (1998b)), and the last inequality holds by Lemma A.0.1 below. □

Lemma A.0.1. *Under the model setup in Section 3.5 and Assumption 3, we have for large* n *and* d,

$$\mathbb{E}\left\| \frac{1}{nd}\,\frac{1}{p^2}\left(\hat{\Sigma}_p - \mathbb{E}\hat{\Sigma}_p\right) V^{(m)} \right\|_F^2 \leqslant \frac{C_1}{p\,n} \qquad \text{(A.3)}$$

and

$$\mathbb{E}\left\| \frac{1}{nd}\,\frac{1}{p^2}\left(\hat{\Sigma}_{pt} - \mathbb{E}\hat{\Sigma}_{pt}\right) U^{(m)} \right\|_F^2 \leqslant \frac{C_2}{p\,d},$$

where $\hat{\Sigma}_p$ *and* $\hat{\Sigma}_{pt}$ *are defined in* (1.4) *and* C_1 *and* C_2 *are generic constants free of* n, d, *and* p.

Proof of Lemma A.0.1. We only show the result (A.3) because the other result holds similarly.

From (A.1), (1.4), Proposition 1.3.1, and triangle inequality, we have

$$
\begin{aligned}
&\left\| \left(\hat{\Sigma}_p - \mathbb{E}\hat{\Sigma}_p\right) V^{(m)} \right\|_F \\
&\leqslant \left\| \left[M_y^T M_y - (1-p)\mathrm{diag}(M_y^T M_y) - p^2(1-p)\,\mathrm{diag}(M_0^T M_0)\right] V^{(m)} \right\|_F \\
&+ \left\| \left[\epsilon_y^T \epsilon_y - (1-p)\mathrm{diag}(\epsilon_y^T \epsilon_y) - np^2\sigma^2 I_d\right] V^{(m)} \right\|_F \\
&+p \left\| \left[M_y^T M_0 - (1-p)\mathrm{diag}(M_y^T M_0)\right] V^{(m)} \right\|_F \\
&+p \left\| \left[M_0^T M_y - (1-p)\mathrm{diag}(M_0^T M_y)\right] V^{(m)} \right\|_F \\
&+p \left\| \left[\epsilon_y^T M_0 - (1-p)\mathrm{diag}(\epsilon_y^T M_0)\right] V^{(m)} \right\|_F \\
&+p \left\| \left[M_0^T \epsilon_y - (1-p)\mathrm{diag}(M_0^T \epsilon_y)\right] V^{(m)} \right\|_F \\
&+ \left\| \left[M_y^T \epsilon_y - (1-p)\mathrm{diag}(M_y^T \epsilon_y)\right] V^{(m)} \right\|_F \\
&+ \left\| \left[\epsilon_y^T M_y - (1-p)\mathrm{diag}(\epsilon_y^T M_y)\right] V^{(m)} \right\|_F \\
&= (A) + (B) + p\,(C) + p\,(D) + p\,(E) + p\,(F) + (G) + (H). \qquad \text{(A.4)}
\end{aligned}
$$

We examine the convergence rates of the above terms, (A)-(H).

First, consider the term (A) in (A.4). Then, we have

$$
\begin{aligned}
&\mathbb{E} \left\| \left[M_y^T M_y - (1-p)\mathrm{diag}(M_y^T M_y) - p^2(1-p)\,\mathrm{diag}(M_0^T M_0)\right] V^{(m)} \right\|_F^2 \\
&= \sum_{i=1}^{d}\sum_{j=1}^{m} \mathbb{E}\Bigg\{ \sum_{k=1}^{n}\sum_{h=1}^{d} \Big[p\left((y_{ki}-p)^2 - p(1-p)\right) M_{0ki}^2 V_{ji}\, \mathbb{1}_{(h=i)} \\
&\qquad\qquad + (y_{ki}-p)(y_{kh}-p) M_{0ki} M_{0kh} V_{jh}\, \mathbb{1}_{(h\neq i)} \Big] \Bigg\}^2 \\
&= \sum_{i=1}^{d}\sum_{j=1}^{m} \Bigg\{ \sum_{k=1}^{n}\sum_{h=1}^{d} \Big[p^2\, \mathbb{E}\left((y_{ki}-p)^2 - p(1-p)\right)^2 M_{0ki}^4 V_{ji}^2\, \mathbb{1}_{(h=i)} \\
&\qquad\qquad + \mathbb{E}\left((y_{ki}-p)^2 (y_{kh}-p)^2\right) M_{0ki}^2 M_{0kh}^2 V_{jh}^2\, \mathbb{1}_{(h\neq i)} \Big] \Bigg\} \\
&= \sum_{i=1}^{d}\sum_{j=1}^{m} \Bigg\{ \sum_{k=1}^{n}\sum_{h=1}^{d} \Big[p^3(1-p)(2p-1)^2 M_{0ki}^4 V_{ji}^2\, \mathbb{1}_{(h=i)} \\
&\qquad\qquad + p^2(1-p)^2 M_{0ki}^2 M_{0kh}^2 V_{jh}^2\, \mathbb{1}_{(h\neq i)} \Big] \Bigg\} \\
&\leqslant p^2(1-p) L^4 \sum_{i=1}^{d}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{h=1}^{d} V_{jh}^2
\end{aligned}
$$

$$
\begin{aligned}
&= p^2(1-p)L^4 \sum_{i=1}^{d}\sum_{j=1}^{m}\sum_{k=1}^{n} 1 \\
&\leqslant Cp^2(1-p)ndm. \qquad (A.5)
\end{aligned}
$$

Similarly to (A.5), we can show that the expected values of the terms (B), (D), (F), (G), and (H) squared are bounded by Cp^2ndm.

Second, consider the term (C) in (A.4). Then, we have

$$
\begin{aligned}
&\mathbb{E}\left\| \left[M_y^T M_0 - (1-p)\mathrm{diag}(M_y^T M_0) \right] V^{(m)} \right\|_F^2 \\
&= \sum_{i=1}^{d}\sum_{j=1}^{m} \mathbb{E}\Bigg\{ \sum_{k=1}^{n} (y_{ki} - p) \sum_{h=1}^{d} \Big[p\, M_{0ki}^2 V_{ji}\, \mathbb{1}_{(h=i)} \\
&\qquad\qquad + M_{0ki} M_{0kh} V_{jh}\, \mathbb{1}_{(h\neq i)} \Big] \Bigg\}^2 \\
&= \sum_{i=1}^{d}\sum_{j=1}^{m}\sum_{k=1}^{n} \mathbb{E}(y_{ki} - p)^2 \left\{ \sum_{h=1}^{d} M_{0ki} M_{0kh} V_{jh} \Big[1 - (1-p)\mathbb{1}_{(h=i)} \Big] \right\}^2 \\
&= p(1-p) \sum_{i=1}^{d}\sum_{j=1}^{m}\sum_{k=1}^{n} \left\{ \sum_{h=1}^{d} M_{0ki} M_{0kh} V_{jh} \Big[1 - (1-p)\mathbb{1}_{(h=i)} \Big] \right\}^2 \\
&\leqslant p(1-p)L^4 \sum_{i=1}^{d}\sum_{j=1}^{m}\sum_{k=1}^{n} \left\{ \sum_{h=1}^{d} |V_{jh}| \right\}^2 \\
&\leqslant Cp(1-p)ndm^2, \qquad (A.6)
\end{aligned}
$$

where the last inequality holds due to Cauchy-Schwarz inequality.

Lastly, for the term (E) in (A.4),

$$
\begin{aligned}
&\mathbb{E}\left\| \left[\epsilon_y^T M_0 - (1-p)\mathrm{diag}(\epsilon_y^T M_0) \right] V^{(m)} \right\|_F^2 \\
&= \sum_{i=1}^{d}\sum_{j=1}^{m} \mathbb{E}\Bigg\{ \sum_{k=1}^{n} y_{ki}\epsilon_{ki} \sum_{h=1}^{d} M_{0kh} V_{jh} \Big[1 - (1-p)\mathbb{1}_{(h=i)} \Big] \Bigg\}^2 \\
&= \sum_{i=1}^{d}\sum_{j=1}^{m}\sum_{k=1}^{n} \mathbb{E}\big(y_{ki}^2 \epsilon_{ki}^2\big) \left\{ \sum_{h=1}^{d} M_{0kh} V_{jh} \Big[1 - (1-p)\mathbb{1}_{(h=i)} \Big] \right\}^2
\end{aligned}
$$

$$
\begin{aligned}
&= p\sigma^2 \sum_{i=1}^{d}\sum_{j=1}^{m}\sum_{k=1}^{n}\left\{\sum_{h=1}^{d} M_{0kh}V_{jh}\left[1-(1-p)\mathbb{1}_{(h=i)}\right]\right\}^2 \\
&\leqslant p\sigma^2 L^2 \sum_{i=1}^{d}\sum_{j=1}^{m}\sum_{k=1}^{n}\left\{\sum_{h=1}^{d}|V_{jh}|\right\}^2 \\
&\leqslant Cpnd^2, \qquad \text{(A.7)}
\end{aligned}
$$

where last inequality holds due to Cauchy-Schwarz inequality.

The result follows from (A.5)-(A.7). □

Lemma A.0.2. *Under the model setup in Section 3.5 and Assumption 3, we have for any given* $\xi_1 > 0$,

$$\|M_y\|_2 \leqslant C_{\xi_1}\sqrt{p\,n\log n}$$

with probability $1 - O(n^{-\xi_1})$. *Similarly, we have for any given* $\xi_2 > 0$,

$$\|\epsilon_y\|_2 \leqslant C_{\xi_2}\sqrt{p\,n\log n}$$

with probability $1 - O(n^{-\xi_2})$.

Proof of Lemma A.0.2. Let $M_y^{(i,j)} \in \mathbb{R}^{n\times d}$ be such that

$$M_{y\,kh}^{(i,j)} = \begin{cases} (y_{kh}-p)M_{0kh}, & (k,h)=(i,j) \\ 0, & (k,h)\neq(i,j) \end{cases} \quad \text{for } 1\leqslant k\leqslant n \text{ and } 1\leqslant h\leqslant d.$$

Then,

$$\frac{1}{nd}M_y = \frac{1}{nd}\sum_{i=1}^{n}\sum_{j=1}^{d} M_y^{(i,j)},$$

$\mathbb{E}(M_y^{(i,j)}) = 0$, and $\|M_y^{(i,j)}\|_2 \leqslant L$ for all $1\leqslant k\leqslant n$ and $1\leqslant h\leqslant d$. Also, we have

$$
\begin{aligned}
\left\|\frac{1}{nd}\mathbb{E}\left(M_y^{(i,j)}M_y^{(i,j)\mathsf{T}}\right)\right\|_2 &= \left\|\frac{p(1-p)}{nd}\mathrm{diag}\left(M_0M_0^{\mathsf{T}}\right)\right\|_2 \leqslant \frac{p\,L^2}{n} \text{ and} \\
\left\|\frac{1}{nd}\mathbb{E}\left(M_y^{(i,j)\mathsf{T}}M_y^{(i,j)}\right)\right\|_2 &= \left\|\frac{p(1-p)}{nd}\mathrm{diag}\left(M_0^{\mathsf{T}}M_0\right)\right\|_2 \leqslant \frac{p\,L^2}{d}. \qquad \text{(A.8)}
\end{aligned}
$$

Thus, by Proposition 1 in Koltchinskii et al. (2011a), we have

$$\left\| \frac{1}{nd} M_y \right\|_2 \leqslant C \max\left(\sqrt{\frac{p\, L^2}{d}} \sqrt{\frac{\log n}{nd}}, L \frac{\log n}{nd} \right) \leqslant C \sqrt{\frac{p\, \log n}{nd^2}}$$

with probability at least $1 - n^{-\xi_1}$.

In a similar way together with Proposition 2 in Koltchinskii et al. (2011a), we can show that $\left\| \frac{1}{nd} \epsilon_y \right\|_2 \leqslant C \sqrt{\frac{p\, \log n}{nd^2}}$ with probability at least $1 - n^{-\xi_2}$.

□

Proof of Lemma 1.6.1. We only show the result (1.17) because the other result holds similarly.

From (A.1), Proposition 1.3.1 and triangle inequality, we have

$$\begin{aligned}
&\frac{1}{nd} \left\| \hat{\Sigma}_p - \mathbb{E}\hat{\Sigma}_p \right\|_2 \\
&\leqslant \frac{1}{nd} \left\| M_y^T M_y - (1-p) \mathrm{diag}(M_y^T M_y) - p^2(1-p)\, \mathrm{diag}(M_0^T M_0) \right\|_2 \\
&+ \frac{1}{nd} \left\| \epsilon_y^T \epsilon_y - (1-p) \mathrm{diag}(\epsilon_y^T \epsilon_y) - np^2\sigma^2 I_d \right\|_2 \\
&+ 2\, \frac{1}{nd} \left\| pM_y^T M_0 - (1-p) p \mathrm{diag}(M_y^T M_0) \right\|_2 \\
&+ 2\, \frac{1}{nd} \left\| p\epsilon_y^T M_0 - (1-p) p \mathrm{diag}(\epsilon_y^T M_0) \right\|_2 \\
&+ 2\, \frac{1}{nd} \left\| M_y^T \epsilon_y - (1-p) \mathrm{diag}(M_y^T \epsilon_y) \right\|_2 \\
&= (I) + (II) + 2\,(III) + 2\,(IV) + 2\,(V). \qquad \text{(A.9)}
\end{aligned}$$

Because of similarity, we provide arguments only for (I) and (IV).

Consider the term (I) in (A.9). First, we have by Lemma A.0.2

$$\frac{1}{nd} \left\| M_y^T M_y \right\|_2 = nd \left\| \frac{1}{nd} M_y \right\|_2^2 \leqslant Cp \frac{\log n}{d} \qquad \text{(A.10)}$$

with probability at least $1 - O(n^{-\mu_1})$. Also, we have with probability at least

$1 - O(n^{-\mu_1})$,

$$\begin{aligned}
&\frac{1-p}{nd} \left\| \operatorname{diag}(M_y^T M_y) + p^2 \operatorname{diag}(M_0^T M_0) \right\|_2 \\
&\leqslant \frac{1-p}{nd} \left\| \operatorname{diag}(M_y^T M_y) - p(1-p) \operatorname{diag}(M_0^T M_0) \right\|_2 \\
&\quad + \frac{p(1-p)}{nd} \left\| \operatorname{diag}(M_0^T M_0) \right\|_2 \\
&= (1-p) \max_{1 \leqslant h \leqslant d} \left| \sum_{k=1}^{n} \frac{\left[(y_{kh} - p)^2 - p(1-p) \right] M_{0kh}^2}{nd} \right| \\
&\quad + \frac{p(1-p)}{nd} \max_{1 \leqslant h \leqslant d} \left| \sum_{k=1}^{n} M_{0kh}^2 \right| \\
&\leqslant C \sqrt{\frac{p \log n}{n}} \frac{1}{d} + \frac{p(1-p) L^2}{d} \\
&\leqslant C p d^{-1}, \qquad \text{(A.11)}
\end{aligned}$$

where the second inequality holds by (A.12) below. Take $t^2 = c \frac{\log n}{n d^2} p(1-p)(3p^2 - 3p + 1)$ for some large constant $c > 0$. Then, by Bernstein's inequality,

$$\begin{aligned}
&\mathbb{P}\left(\max_{1 \leqslant h \leqslant d} \left| \sum_{k=1}^{n} \frac{\left[(y_{kh} - p)^2 - p(1-p) \right] M_{0kh}^2}{nd} \right| \geqslant t \right) \\
&\leqslant \sum_{h=1}^{d} \mathbb{P}\left(\left| \sum_{k=1}^{n} \left[(y_{kh} - p)^2 - p(1-p) \right] M_{0kh}^2 \right| \geqslant ndt \right) \\
&\leqslant 2d \, \exp\left\{ -\frac{nd^2 t^2}{2L^4 p(1-p)(3p^2 - 3p + 1)} \right\} \\
&= C n^{-\mu_1}. \qquad \text{(A.12)}
\end{aligned}$$

By (A.10) and (A.11), we have

$$(I) \leqslant C p \frac{\log n}{d} \qquad \text{(A.13)}$$

with probability at least $1 - O(n^{-\mu_1})$. Similarly, we can show that (II) and (V) are bounded by $C p \frac{\log n}{d}$ with probability at least $1 - O(n^{-\mu_1})$.

Consider the term (IV) in (A.9). We have

$$(IV)^2 \leqslant \left\{ \max_{1\leqslant j\leqslant d} \sum_{i=1}^{d} \left| \sum_{k=1}^{n} X_{kij}^{(IV)} \right| \right\} \left\{ \max_{1\leqslant i\leqslant d} \sum_{j=1}^{d} \left| \sum_{k=1}^{n} X_{kij}^{(IV)} \right| \right\},$$

where $nd\, X_{kij}^{(IV)} = p\, y_{ki}\epsilon_{ki}M_{0kj}\mathbb{1}_{(i\neq j)} + p^2 y_{ki}\epsilon_{ki}M_{0kj}\mathbb{1}_{(i=j)}$ and hence $X_{kij}^{(IV)}$ are centered sub-Gaussian random variables under the model setup in Section 3.5. Then, we have for any $\rho \in \mathbb{R}$ and for all $1 \leqslant k \leqslant n$, $1 \leqslant i \leqslant d$, and $1 \leqslant j \leqslant d$,

$$\mathbb{E} \exp\left\{\rho X_{kij}^{(IV)}\right\} \leqslant \exp\left\{ \frac{\rho^2 \frac{p^3\beta}{n^2d^2}}{2} \right\} \text{ for some constant } \beta > 0.$$

Take $t^2 = cp^3 \frac{\log n}{n}$ for some large constant $c > 0$ and $\rho = \frac{t/d}{n\frac{p^3\beta}{n^2d^2}}$. Then, by Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left(\max_{1\leqslant j\leqslant d} \sum_{i=1}^{d} \left| \sum_{k=1}^{n} X_{kij}^{(IV)} \right| > t \right) &\leqslant \sum_{j=1}^{d}\sum_{i=1}^{d} \mathbb{P}\left(\left| \sum_{k=1}^{n} X_{kij}^{(IV)} \right| > t/d \right) \\
&\leqslant 2\sum_{j=1}^{d}\sum_{i=1}^{d} \frac{\mathbb{E}\left(\exp\left\{ \rho \sum_{k=1}^{n} X_{kij}^{(IV)} \right\} \right)}{\exp\{\rho(t/d)\}} \\
&\leqslant 2d^2 \exp\left\{ -\rho\frac{t}{d} + \frac{\rho^2}{2}\frac{p^3\beta}{nd^2} \right\} \\
&= 2d^2 \exp\left\{ -\frac{nt^2}{2p^3\beta} \right\} \\
&= Cn^{-\mu_1}. \qquad \text{(A.14)}
\end{aligned}$$

Similarly,

$$\mathbb{P}\left(\max_{1\leqslant i\leqslant d} \sum_{j=1}^{d} \left| \sum_{k=1}^{n} X_{kij}^{(IV)} \right| > t \right) \leqslant Cn^{-\mu_1}. \qquad \text{(A.15)}$$

By (A.14) and (A.15), with probability at least $1 - O(n^{-\mu_1})$,

$$|(IV)| \leqslant Cp^{3/2}\sqrt{\frac{\log n}{n}}. \tag{A.16}$$

Similarly, we can show that (III) is bounded by $Cp^{3/2}\sqrt{\frac{\log n}{n}}$ with probability at least $1 - O(n^{-\mu_1})$.

The statement is showed by (A.13) and (A.16). □

Proof of Lemma 1.6.2. We only show the result (1.18) because the other result holds similarly.

By triangle inequality, we have

$$\begin{aligned}
&\frac{1}{nd}\left\|\hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p\right\|_2 = \frac{1}{nd}\left\|(\hat{p} - p)\,\mathrm{diag}(\hat{\Sigma})\right\|_2 \\
&\leqslant \frac{|\hat{p} - p|}{nd}\Big\{\|\mathrm{diag}(\hat{\Sigma}) - \mathrm{diag}(pM_0^T M_0 + np\sigma^2 I_d)\|_2 \\
&\qquad\qquad + \|\mathrm{diag}(pM_0^T M_0 + np\sigma^2 I_d)\|_2\Big\}.
\end{aligned} \tag{A.17}$$

We will look at the terms in (A.17) one by one.

By Bernstein's inequality, we have for large constant $C > 0$,

$$\begin{aligned}
&\mathbb{P}\left(|\hat{p} - p| \geqslant C\sqrt{\frac{p(1-p)\log n}{nd}}\right) \\
&= \mathbb{P}\left(\left|\sum_{k=1}^{n}\sum_{h=1}^{d}(y_{kh} - p)\right| \geqslant C\sqrt{p(1-p)nd\log n}\right) \\
&\leqslant 2\,\exp\{-\nu_1 \log n\} \\
&= 2n^{-\nu_1}.
\end{aligned} \tag{A.18}$$

Take $t^2 = c\frac{p\log n}{nd^2}$ for some large constant $c > 0$. Then, since $y_{ki}^2(M_{0ki} + \epsilon_{ki})^2 - p(M_{0ki}^2 + \sigma^2), k = 1, \ldots, n$, are independent centered sub-exponential random

variables, we have by Proposition 5.16 in Vershynin (2010),

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{nd}\left\|\mathrm{diag}(\hat{\Sigma})-\mathrm{diag}(pM_0^TM_0+np\sigma^2I_d)\right\|_2\geqslant t\right)\\
&=\mathbb{P}\left(\frac{1}{nd}\max_{1\leqslant i\leqslant d}\left|\sum_{k=1}^{n}\left[y_{ki}^2(M_{0ki}+\epsilon_{ki})^2-p(M_{0ki}^2+\sigma^2)\right]\right|\geqslant t\right)\\
&\leqslant\sum_{i=1}^{d}\mathbb{P}\left(\left|\sum_{k=1}^{n}\left[y_{ki}^2(M_{0ki}+\epsilon_{ki})^2-p(M_{0ki}^2+\sigma^2)\right]\right|\geqslant ndt\right)\\
&\leqslant 2d\ \exp\left\{-\frac{n^2d^2t^2}{c_1np}\right\}\\
&\leqslant Cn^{-\nu_1}.
\end{aligned}\tag{A.19}$$

Also, note that

$$\begin{aligned}
\left\|\frac{1}{nd}\mathrm{diag}(pM_0^TM_0+np\sigma^2I_d)\right\|_2 &= \frac{1}{nd}\max_{1\leqslant i\leqslant d}\ p\sum_{k=1}^{n}M_{0ki}^2+np\sigma^2\\
&\leqslant \frac{p(L^2+\sigma^2)}{d}.
\end{aligned}\tag{A.20}$$

Combining the results in (A.17)-(A.20), we have

$$\frac{1}{nd}\left\|\hat{\Sigma}_{\hat{p}}-\hat{\Sigma}_p\right\|_2\leqslant Cp^{3/2}\sqrt{\frac{\log n}{nd}}\ \frac{1}{d}\tag{A.21}$$

with probability at least $1-O\left(n^{-\nu_1}\right)$. □

Proof of Lemma 1.6.3. We only show the result (1.19) because the other result holds similarly.

We have

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{nd}\left(\hat{\Sigma}_{\hat{p}}-\hat{\Sigma}_p\right)V^{(m)}\right\|_F^2\\
&\leqslant m\,\mathbb{E}\left\|\frac{1}{nd}\left(\hat{\Sigma}_{\hat{p}}-\hat{\Sigma}_p\right)\right\|_2^2
\end{aligned}$$

$$
\begin{aligned}
&\leqslant m\,\mathbb{E}\left\{(\hat{p}-p)^2\left\|\frac{1}{nd}\mathrm{diag}(\hat{\Sigma})\right\|_2^2\right\}\\
&\leqslant 4m\,\mathbb{E}\left\{(\hat{p}-p)^2\left\|\frac{1}{nd}\mathrm{diag}(\hat{\Sigma})-\frac{1}{nd}\mathrm{diag}(pM_0^TM_0+np\sigma^2I_d)\right\|_2^2\right\}\\
&\qquad+4m\left\|\frac{1}{nd}\mathrm{diag}(pM_0^TM_0+np\sigma^2I_d)\right\|_2^2\ \mathbb{E}\,(\hat{p}-p)^2\\
&\leqslant 4m\sqrt{\mathbb{E}\,(\hat{p}-p)^4\ \mathbb{E}\left\|\frac{1}{nd}\mathrm{diag}(\hat{\Sigma})-\frac{1}{nd}\mathrm{diag}(pM_0^TM_0+np\sigma^2I_d)\right\|_2^4}\\
&\qquad+4m\,\frac{p^2(L^2+\sigma^2)^2}{d^2}\frac{p(1-p)}{nd}\\
&\leqslant C_1\frac{p^2(1-p)}{n^2d^{5/2}}+C_2\frac{p^3(1-p)}{nd^3},
\end{aligned}
\tag{A.22}
$$

where the fourth inequality holds by Hölder's inequality and the fifth inequality is due to the fact that

$$
\begin{aligned}
&\mathbb{E}\,(\hat{p}-p)^4\ \mathbb{E}\left\|\frac{1}{nd}\mathrm{diag}(\hat{\Sigma})-\frac{1}{nd}\mathrm{diag}(pM_0^TM_0+np\sigma^2I_d)\right\|_2^4\\
&\leqslant\frac{\mathbb{E}\big\{\sum_{k=1}^n\sum_{h=1}^d(y_{kh}-p)\big\}^4}{n^4d^4}\\
&\qquad\frac{\mathbb{E}\left\{\max_{1\leqslant i\leqslant d}\left|\sum_{k=1}^n\left[y_{ki}^2(M_{0ki}+\epsilon_{ki})^2-p(M_{0ki}^2+\sigma^2)\right]\right|^4\right\}}{n^4d^4}\\
&\leqslant\frac{\mathbb{E}\big\{\sum_{k=1}^n\sum_{h=1}^d(y_{kh}-p)\big\}^4}{n^4d^4}\\
&\qquad\frac{d\,\mathbb{E}\left[\sum_{k=1}^n\left(y_{ki}^2(M_{0ki}+\epsilon_{ki})^2-p(M_{0ki}^2+\sigma^2)\right)\right]^4}{n^4d^4}\\
&=\frac{O\left(p^2(1-p)^2n^2d^2\right)\ O\left(p^2n^2d\right)}{n^8d^8}.
\end{aligned}
\tag{A.23}
$$

□

Proofs for Section 1.6

Lemma A.0.3. *Under the model setup in Section 3.5 and Assumption 3, we have*

$$\begin{aligned}
&\sum_{i=1}^{m} \lambda_{pi}^2 - p^2 \left[\sum_{i=1}^{m} \lambda_i^2 + n\sigma^2 \right] \\
&= 2p \sum_{k=1}^{n} \sum_{h=1}^{d} (y_{kh} - p) M_{0kh} \sum_{i=1}^{m} V_{ih} \left[\left(\sum_{h'=1}^{d} M_{0kh'} V_{ih'} \right) - (1-p) M_{0kh} V_{ih} \right] \\
&\quad + 2p \sum_{k=1}^{n} \sum_{h=1}^{d} y_{kh} \epsilon_{kh} \sum_{i=1}^{m} V_{ih} \left[\left(\sum_{h'=1}^{d} M_{0kh'} V_{ih'} \right) - (1-p) M_{0kh} V_{ih} \right] \\
&\quad + o_p\left(\sqrt{nd}\right) \\
&= (i) + (ii) + O_p\left(p\sqrt{n} + pd\right)
\end{aligned}$$

and $(i)+(ii) = O_p\left(\sqrt{p^3 nd}\right)$*, where* λ_{pi} *and* λ_i *are defined in* (1.4) *and* (2.1)*, respectively.*

Proof of Lemma A.0.3. We have

$$\begin{aligned}
&\sum_{i=1}^{m} \lambda_{pi}^2 - p^2 \left[\sum_{i=1}^{m} \lambda_i^2 + n\sigma^2 \right] \\
&= \mathrm{tr}(V_p{}^{(m)T} \hat{\Sigma}_p V_p{}^{(m)}) - \mathrm{tr}\left(V^{(m)T} (p^2 M_0^T M_0 + np^2\sigma^2 I_d) V^{(m)} \right) \\
&= \mathrm{tr}(\mathcal{O}^T V^{(m)T} \hat{\Sigma}_p V^{(m)} \mathcal{O}) \\
&\qquad + \mathrm{tr}(V_p{}^{(m)T} \hat{\Sigma}_p V_p{}^{(m)} - \mathcal{O}^T V^{(m)T} \hat{\Sigma}_p V^{(m)} \mathcal{O}) \\
&\qquad - \mathrm{tr}\left(V^{(m)T} (p^2 M_0^T M_0 + np^2\sigma^2 I_d) V^{(m)} \right) \\
&= \mathrm{tr}(V^{(m)T} \hat{\Sigma}_p V^{(m)}) + \mathrm{tr}(V_p{}^{(m)T} \hat{\Sigma}_p V_p{}^{(m)} - \mathcal{O}^T V^{(m)T} \hat{\Sigma}_p V^{(m)} \mathcal{O}) \\
&\qquad - \mathrm{tr}\left(V^{(m)T} (p^2 M_0^T M_0 + np^2\sigma^2 I_d) V^{(m)} \right) \\
&= \mathrm{tr}\Big(V^{(m)T} (M_y^T M_y - (1-p)\mathrm{diag}(M_y^T M_y) \\
&\qquad\qquad - p^2(1-p)\,\mathrm{diag}(M_0^T M_0)) V^{(m)} \Big) \\
&\qquad + \mathrm{tr}\left(V^{(m)T} (\epsilon_y^T \epsilon_y - (1-p)\mathrm{diag}(\epsilon_y^T \epsilon_y) - np^2\sigma^2 I_d) V^{(m)} \right) \\
&\qquad + \mathrm{tr}\Big(V^{(m)T} (pM_0^T M_y + pM_y^T M_0 \\
&\qquad\qquad - (1-p)p\mathrm{diag}(M_0^T M_y + M_y^T M_0)) V^{(m)} \Big) \\
&\qquad + \mathrm{tr}\Big(V^{(m)T} (pM_0^T \epsilon_y + p\epsilon_y^T M_0
\end{aligned}$$

$$
\begin{aligned}
&\qquad -(1-p)p\mathrm{diag}(M_0^T\epsilon_y + \epsilon_y^T M_0))V^{(m)}\Big) \\
&+\mathrm{tr}\Big(V^{(m)T}(M_y^T\epsilon_y + \epsilon_y^T M_y \\
&\qquad -(1-p)\mathrm{diag}(M_y^T\epsilon_y + \epsilon_y^T M_y))V^{(m)}\Big) \\
&+\mathrm{tr}({V_p}^{(m)T}\hat{\Sigma}_p {V_p}^{(m)} - \mathcal{O}^T V^{(m)T}\hat{\Sigma}_p V^{(m)}\mathcal{O}) \\
&= (a)+(b)+(c)+(d)+(e)+(f), \qquad \text{(A.24)}
\end{aligned}
$$

where $\mathcal{O} \in \mathbb{V}_{m,m}$ is a solution to $\inf_{Q\in\mathbb{V}_{m,m}}\|V_p^{(m)} - V^{(m)}Q\|_F^2$ and the fourth equality holds by (1.4) and (A.1). Below, we examine the six terms (a)-(f) one by one.

The term (a) in (A.24) is

$$
\begin{aligned}
(a) &= \sum_{i=1}^{m} V_i^T\left(M_y^T M_y - (1-p)\mathrm{diag}(M_y^T M_y) - p^2(1-p)\,\mathrm{diag}(M_0^T M_0)\right)V_i \\
&= \sum_{i=1}^{m}\Bigg\{\sum_{k=1}^{n}\left(\sum_{h=1}^{d}(y_{kh}-p)M_{0kh}V_{ih}\right)^2 \\
&\qquad -(1-p)\sum_{k=1}^{n}\sum_{h=1}^{d}(y_{kh}-p)^2 M_{0kh}^2 V_{ih}^2 \\
&\qquad -p^2(1-p)\sum_{k=1}^{n}\sum_{h=1}^{d}M_{0kh}^2 V_{ih}^2\Bigg\} \\
&= \sum_{k=1}^{n}\sum_{h=1}^{d}p\Big[(y_{kh}-p)^2 - p(1-p)\Big]M_{0kh}^2\sum_{i=1}^{m}V_{ih}^2 \\
&\qquad +2\sum_{k=1}^{n}\sum_{h<h'}^{1\sim d}(y_{kh}-p)(y_{kh'}-p)M_{0kh}M_{0kh'}\sum_{i=1}^{m}V_{ih}V_{ih'} \qquad \text{(A.25)}
\end{aligned}
$$

Note that the two terms in (A.25) are centered and uncorrelated with each other. So, the variance is

$$
\begin{aligned}
\mathrm{var}(a) &= \left\{\sum_{k=1}^{n}\sum_{h=1}^{d}p^3(1-p)(2p-1)^2 M_{0kh}^4\left(\sum_{i=1}^{m}V_{ih}^2\right)^2\right\} \\
&\qquad +\left\{4\sum_{k=1}^{n}\sum_{h<h'}^{1\sim d}p^2(1-p)^2 M_{0kh}^2 M_{0kh'}^2\left(\sum_{i=1}^{m}V_{ih}V_{ih'}\right)^2\right\}
\end{aligned}
$$

$$
\begin{aligned}
&\leqslant m\sum_{i=1}^{m}\sum_{k=1}^{n}\sum_{h=1}^{d} p^3(1-p)(2p-1)^2 M_{0kh}^4 V_{ih}^4 \\
&\quad +4m\sum_{i=1}^{m}\sum_{k=1}^{n}\sum_{h<h'}^{1\sim d} p^2(1-p)^2 M_{0kh}^2 M_{0kh'}^2 V_{ih}^2 V_{ih'}^2 \\
&\leqslant mL^4p^3(1-p)(2p-1)^2\sum_{i=1}^{m}\sum_{k=1}^{n}\sum_{h=1}^{d} V_{ih}^4 \\
&\quad +4mL^4p^2(1-p)^2\sum_{i=1}^{m}\sum_{k=1}^{n}\sum_{h,h'}^{1\sim d} V_{ih}^2 V_{ih'}^2 \\
&\leqslant Cp^2(1-p)n, \qquad \text{(A.26)}
\end{aligned}
$$

where the first inequality is due to Jensen's inequality. This shows that the term (a) is $O_p(p\sqrt{n})$. Similarly, we can show that the terms (b) and (e) are $O_p(p\sqrt{n})$.

The term (c) in (A.24) is

$$
\begin{aligned}
&\frac{1}{2p}(c) \\
&= \sum_{i=1}^{m} V_i^T \left(M_0^T M_y - (1-p)\mathrm{diag}(M_0^T M_y)\right) V_i \\
&= \sum_{i=1}^{m}\left\{\sum_{k=1}^{n}\left(\sum_{h=1}^{d} M_{0kh}V_{ih}\right)\left(\sum_{h'=1}^{d}(y_{kh'}-p)M_{0kh'}V_{ih'}\right)\right. \\
&\qquad\qquad \left. -(1-p)\sum_{k=1}^{n}\sum_{h=1}^{d}(y_{kh}-p)M_{0kh}^2V_{ih}^2\right\} \\
&= \sum_{k=1}^{n}\sum_{h=1}^{d}(y_{kh}-p)M_{0kh}\sum_{i=1}^{m} V_{ih}\left[\left(\sum_{h'=1}^{d} M_{0kh'}V_{ih'}\right) - (1-p)M_{0kh}V_{ih}\right].
\end{aligned}
$$

Then, its variance is

$$
\begin{aligned}
&\left(\frac{1}{2p}\right)^2 \mathrm{var}(c) \\
&= \sum_{k=1}^{n}\sum_{h=1}^{d} p(1-p)M_{0kh}^2\left\{\sum_{i=1}^{m} V_{ih}\left[\left(\sum_{h'=1}^{d} M_{0kh'}V_{ih'}\right) - (1-p)M_{0kh}V_{ih}\right]\right\}^2 \\
&\leqslant Cp(1-p)nd,
\end{aligned}
$$

where the last inequality is due to Assumption 3(1) and the fact that

$$
\begin{aligned}
&\sum_{k=1}^{n}\sum_{h=1}^{d} M_{0kh}^2\left\{\sum_{i=1}^{m} V_{ih}\left[\left(\sum_{h'=1}^{d} M_{0kh'}V_{ih'}\right)-(1-p)M_{0kh}V_{ih}\right]\right\}^2\\
&=\sum_{k=1}^{n}\sum_{h=1}^{d} M_{0kh}^2\left\{\sum_{i=1}^{m}\lambda_i U_{ik}V_{ih}-(1-p)\sum_{i=1}^{m} M_{0kh}V_{ih}^2\right\}^2\\
&=\sum_{k=1}^{n}\sum_{h=1}^{d} M_{0kh}^2\left\{\sum_{i=1}^{m}\lambda_i U_{ik}V_{ih}\right\}^2\\
&\qquad+(1-p)^2\sum_{k=1}^{n}\sum_{h=1}^{d}\left\{\sum_{i=1}^{m} M_{0kh}^2V_{ih}^2\right\}^2\\
&\qquad-2(1-p)\sum_{k=1}^{n}\sum_{h=1}^{d} M_{0kh}\left\{\sum_{i=1}^{m}\lambda_i U_{ik}V_{ih}\right\}\left\{\sum_{i=1}^{m} M_{0kh}^2V_{ih}^2\right\}\\
&=\sum_{k=1}^{n}\sum_{h=1}^{d} M_{0kh}^2\left\{\sum_{i=1}^{m}\lambda_i U_{ik}V_{ih}\right\}^2+O(nd)\\
&=O(nd). \qquad\qquad (A.27)
\end{aligned}
$$

The term (d) in (A.24) is

$$
\begin{aligned}
\frac{1}{2p}(d) &= \sum_{i=1}^{m} V_i^T\left(M_0^T\epsilon_y-(1-p)\mathrm{diag}(M_0^T\epsilon_y)\right)V_i\\
&= \sum_{i=1}^{m}\left\{\sum_{k=1}^{n}\left(\sum_{h=1}^{d} M_{0kh}V_{ih}\right)\left(\sum_{h'=1}^{d} y_{kh'}\epsilon_{kh'}V_{ih'}\right)\right.\\
&\qquad\left.-(1-p)\sum_{k=1}^{n}\sum_{h=1}^{d} y_{kh}\epsilon_{kh}M_{0kh}V_{ih}^2\right\}\\
&= \sum_{k=1}^{n}\sum_{h=1}^{d} y_{kh}\epsilon_{kh}\left\{\sum_{i=1}^{m} V_{ih}\left[\left(\sum_{h'=1}^{d} M_{0kh'}V_{ih'}\right)-(1-p)M_{0kh}V_{ih}\right]\right\}.
\end{aligned}
$$

Then, its variance is

$$\left(\frac{1}{2p}\right)^2 \mathrm{var}(d) = \sum_{k=1}^{n}\sum_{h=1}^{d} p\sigma^2 \left\{ \sum_{i=1}^{m} V_{ih}\left[\left(\sum_{h'=1}^{d} M_{0kh'}V_{ih'}\right) - (1-p)M_{0kh}V_{ih}\right]\right\}^2$$
$$\leqslant Cpnd,$$

where the last inequality is due to Assumption 3(1) and the fact that

$$\begin{aligned}
&\sum_{k=1}^{n}\sum_{h=1}^{d}\left\{ \sum_{i=1}^{m} V_{ih}\left[\left(\sum_{h'=1}^{d} M_{0kh'}V_{ih'}\right) - (1-p)M_{0kh}V_{ih}\right]\right\}^2 \\
&= \sum_{k=1}^{n}\sum_{h=1}^{d}\left\{ \sum_{i=1}^{m} \lambda_i U_{ik}V_{ih} - (1-p)\sum_{i=1}^{m} M_{0kh}V_{ih}^2\right\}^2 \\
&= \sum_{i=1}^{m}\lambda_i^2 + (1-p)^2\sum_{k=1}^{n}\sum_{h=1}^{d}\left(\sum_{i=1}^{m} M_{0kh}V_{ih}^2\right)^2 \\
&\qquad -2(1-p)\sum_{i=1}^{m}\lambda_i^2\sum_{h=1}^{d}V_{ih}^2\sum_{i'=1}^{m}V_{i'h}^2 \\
&= \sum_{i=1}^{m}\lambda_i^2 + O(n). \qquad \text{(A.28)}
\end{aligned}$$

The term (f) in (A.24) is

$$\begin{aligned}
|(f)| &= \left|\mathrm{tr}(V_p{}^{(m)T}\hat{\Sigma}_p V_p{}^{(m)} - \mathcal{O}^T V^{(m)T}\hat{\Sigma}_p V^{(m)}\mathcal{O})\right| \\
&\leqslant \sum_{i=1}^{m}\left|\mathcal{O}_i^T V^T\hat{\Sigma}_p V\mathcal{O}_i - V_{pi}{}^T\hat{\Sigma}_p V_{pi}\right| \\
&= \sum_{i=1}^{m}\left\{\left|(V\mathcal{O}_i - V_{pi})^T\hat{\Sigma}_p(V\mathcal{O}_i - V_{pi}) + 2\lambda_{pi}^2 V_{pi}{}^T(V\mathcal{O}_i - V_{pi})\right|\right\} \\
&\leqslant \sum_{i=1}^{m}\lambda_{p1}^2\left(\left\|V\mathcal{O}_i - V_{pi}\right\|_2^2 + 2\left|V_{pi}{}^T(V\mathcal{O}_i - V_{pi})\right|\right) \\
&= \sum_{i=1}^{m}\lambda_{p1}^2\Big(\left\|V\mathcal{O}_i - V_{pi}\right\|_2^2
\end{aligned}$$

$$
\begin{aligned}
&\qquad + \left|\mathcal{O}_i^T V^T V \mathcal{O}_i - \mathcal{O}_i^T V^T V_{p_i} - V_{p_i}^T V \mathcal{O}_i + V_{p_i}^T V_{p_i})\right| \Big) \\
&= \sum_{i=1}^{m} 2\lambda_{p1}^2 \left\| V\mathcal{O}_i - V_{p_i} \right\|_2^2 \\
&= 2\lambda_{p1}^2 \left\| V^{(m)}\mathcal{O} - V_p{}^{(m)} \right\|_F^2 \\
&= O_p(pd), \qquad \text{(A.29)}
\end{aligned}
$$

where $\mathcal{O}_i$ is the i-th column of $\mathcal{O}$ and the last equality holds by Proposition 1.6.1, (1.9), and (1.25).

Therefore, the result follows from (A.24)-(A.29). □

Proof of Proposition 1.6.2. By Cramèr-Wold device, it is enough to show that for any given $(c_1, c_2)^T \in \mathbb{R}^2 \setminus (0,0)^T$,

$$
\begin{aligned}
&\frac{1}{\sqrt{nd}\gamma_{c_1,c_2}} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}^T \left[\begin{pmatrix} p^{-2} \sum_{i=1}^m \lambda_{p_i}^2 \\ p^2 \sum_{i=1}^m (\lambda_i^2 + n\sigma^2)\, \hat{p} \end{pmatrix} - \begin{pmatrix} \sum_{i=1}^m \left[\lambda_i^2 + n\sigma^2\right] \\ p^3 \sum_{i=1}^m (\lambda_i^2 + n\sigma^2) \end{pmatrix} \right] \\
&\to \mathcal{N}(0,1) \text{ in distribution}, \quad \text{as } n, d \to \infty,
\end{aligned}
$$

where $\gamma_{c_1,c_2}^2 = \begin{pmatrix} c_1 & c_2 \end{pmatrix} \Gamma_{nd} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$. When $c_1 = 0$, this can be directly showed by CLT. Thus, we only consider the case where $c_1 \neq 0$.

We have

$$
\begin{aligned}
&\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}^T \left[\begin{pmatrix} p^{-2} \sum_{i=1}^m \lambda_{p_i}^2 \\ p^2 \sum_{i=1}^m (\lambda_i^2 + n\sigma^2)\, \hat{p} \end{pmatrix} - \begin{pmatrix} \sum_{i=1}^m \left[\lambda_i^2 + n\sigma^2\right] \\ p^3 \sum_{i=1}^m (\lambda_i^2 + n\sigma^2) \end{pmatrix} \right] \\
&= c_1 \frac{1}{p^2} \sum_{i=1}^{m} \left[\lambda_{p_i}^2 - p^2\left(\lambda_i^2 + n\sigma^2\right)\right] + c_2\, p^2 \sum_{i=1}^{m} (\lambda_i^2 + n\sigma^2)\,(\hat{p} - p) \\
&= \frac{2c_1}{p} \sum_{k=1}^{n} \sum_{h=1}^{d} (y_{kh} - p) M_{0kh} \sum_{i=1}^{m} V_{ih} \left[\left(\sum_{h'=1}^{d} M_{0kh'} V_{ih'} \right) \right. \\
&\qquad\qquad \left. -(1-p) M_{0kh} V_{ih} \right]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{2c_1}{p}\sum_{k=1}^{n}\sum_{h=1}^{d} y_{kh}\epsilon_{kh}\sum_{i=1}^{m} V_{ih}\left[\left(\sum_{h'=1}^{d} M_{0kh'}V_{ih'}\right)-(1-p)M_{0kh}V_{ih}\right]\\
&+o_p\left(\sqrt{\frac{nd}{p}}\right)+\frac{c_2\,p^2}{nd}\sum_{k=1}^{n}\sum_{h=1}^{d}(y_{kh}-p)\sum_{i=1}^{m}(\lambda_i^2+n\sigma^2)\\
=&\sum_{k=1}^{n}\sum_{h=1}^{d}(y_{kh}-p)\left\{\frac{2c_1}{p}M_{0kh}\sum_{i=1}^{m}V_{ih}\left[\left(\sum_{h'=1}^{d}M_{0kh'}V_{ih'}\right)\right.\right.\\
&\qquad\left.\left.-(1-p)M_{0kh}V_{ih}\right]+\frac{c_2\,p^2}{nd}\sum_{i=1}^{m}(\lambda_i^2+n\sigma^2)\right\}\\
&+\frac{2c_1}{p}\sum_{k=1}^{n}\sum_{h=1}^{d} y_{kh}\epsilon_{kh}\sum_{i=1}^{m} V_{ih}\left[\left(\sum_{h'=1}^{d} M_{0kh'}V_{ih'}\right)-(1-p)M_{0kh}V_{ih}\right]\\
&+o_p\left(\sqrt{\frac{nd}{p}}\right)\\
=&\,(a)+(b)+o_p\left(\sqrt{\frac{nd}{p}}\right),
\end{aligned}
\tag{A.30}
$$

where the second equality holds by Lemma A.0.3. Since the terms (a) and (b) are centered and not correlated with each other under the model setup in Section 3.5, we have

$$
\begin{aligned}
&\mathrm{var}\left[(a)+(b)\right]=\mathrm{var}\left[(a)\right]+\mathrm{var}\left[(b)\right]\\
&=\sum_{k=1}^{n}\sum_{h=1}^{d}\mathbb{E}(y_{kh}-p)^2\left\{\frac{2c_1}{p}M_{0kh}\sum_{i=1}^{m}V_{ih}\left[\left(\sum_{h'=1}^{d}M_{0kh'}V_{ih'}\right)\right.\right.\\
&\qquad\left.\left.-(1-p)M_{0kh}V_{ih}\right]+\frac{c_2\,p^2}{nd}\sum_{i=1}^{m}(\lambda_i^2+n\sigma^2)\right\}^2\\
&\quad+\frac{4c_1^2}{p^2}\sum_{k=1}^{n}\sum_{h=1}^{d}\mathbb{E}\left(y_{kh}^2\epsilon_{kh}^2\right)\left\{\sum_{i=1}^{m}V_{ih}\left[\left(\sum_{h'=1}^{d}M_{0kh'}V_{ih'}\right)\right.\right.\\
&\qquad\left.\left.-(1-p)M_{0kh}V_{ih}\right]\right\}^2\\
&=p(1-p)\sum_{k=1}^{n}\sum_{h=1}^{d}\left\{\frac{2c_1M_{0kh}}{p}\sum_{i=1}^{m}\lambda_i U_{ik}V_{ih}+c_2p^2\sum_{i=1}^{m}b_i^2\right\}^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{4\sigma^2 c_1^2}{p}\sum_{i=1}^{m}\lambda_i^2+O\left(\frac{n}{p}\right)\\
&=\frac{4c_1^2(1-p)}{p}\sum_{k=1}^{n}\sum_{h=1}^{d}M_{0kh}^2\left\{\sum_{i=1}^{m}\lambda_i U_{ik}V_{ih}\right\}^2+\frac{4\sigma^2 c_1^2}{p}\sum_{i=1}^{m}\lambda_i^2\\
&\quad+4c_1c_2\,nd\,p^2(1-p)\left(\sum_{i=1}^{m}b_i^2\right)^2+c_2^2\,nd\,p^5(1-p)\left(\sum_{i=1}^{m}b_i^2\right)^2\\
&\quad+O\left(\frac{n}{p}\right)\\
&=nd\,(c_1\,c_2)\Gamma_{nd}\begin{pmatrix}c_1\\c_2\end{pmatrix}+O\left(\frac{n}{p}\right),
\end{aligned}
\tag{A.31}
$$

where the third equality is due to (A.27), (A.28) and Assumption 3(1). Note that

$$
nd\,(c_1\,c_2)\Gamma_{nd}\begin{pmatrix}c_1\\c_2\end{pmatrix}\geqslant\frac{4c_1^2\sigma^2}{p}\sum_{i=1}^{m}\lambda_i^2\geqslant\frac{c\,nd}{p}.
\tag{A.32}
$$

Thus, Liapunov's condition is satisfied with $(a)+(b)$ because we have

$$
\begin{aligned}
&\sum_{k=1}^{n}\sum_{h=1}^{d}\mathbb{E}\Bigg|(y_{kh}-p)\Bigg\{\frac{2c_1}{p}M_{0kh}\sum_{i=1}^{m}V_{ih}\Bigg[\Bigg(\sum_{h'=1}^{d}M_{0kh'}V_{ih'}\Bigg)\\
&\qquad\qquad-(1-p)M_{0kh}V_{ih}\Bigg]+\frac{c_2\,p^2}{nd}\sum_{i=1}^{m}(\lambda_i^2+n\sigma^2)\Bigg\}\\
&\quad+y_{kh}\epsilon_{kh}\Bigg\{\frac{2c_1}{p}\sum_{i=1}^{m}V_{ih}\Bigg[\Bigg(\sum_{h'=1}^{d}M_{0kh'}V_{ih'}\Bigg)-(1-p)M_{0kh}V_{ih}\Bigg]\Bigg\}\Bigg|^3\\
&\leqslant 8\sum_{k=1}^{n}\sum_{h=1}^{d}\Bigg\{\mathbb{E}|y_{kh}-p|^3\Bigg|\frac{2c_1}{p}M_{0kh}\sum_{i=1}^{m}V_{ih}\Bigg[\Bigg(\sum_{h'=1}^{d}M_{0kh'}V_{ih'}\Bigg)\\
&\qquad\qquad-(1-p)M_{0kh}V_{ih}\Bigg]+O(1)\Bigg|^3\\
&\qquad+\mathbb{E}|y_{kh}\epsilon_{kh}|^3\Bigg|\frac{2c_1}{p}\sum_{i=1}^{m}V_{ih}\Bigg[\Bigg(\sum_{h'=1}^{d}M_{0kh'}V_{ih'}\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&\qquad\qquad -(1-p)M_{0kh}V_{ih}\Big]\Big|^3\Bigg\} \\
&\leqslant \frac{C}{p^2}\sum_{k=1}^{n}\sum_{h=1}^{d}\Bigg\{\Big|\sum_{i=1}^{m}V_{ih}\Big[\Big(\sum_{h'=1}^{d}M_{0kh'}V_{ih'}\Big) \\
&\qquad\qquad -(1-p)M_{0kh}V_{ih}\Big]\Big|^3+O(1)\Bigg\} \\
&\leqslant \frac{C}{p^2}\sum_{k=1}^{n}\sum_{h=1}^{d}\Bigg\{\sum_{i=1}^{m}|V_{ih}|^3\left(\Big|\sum_{h'=1}^{d}M_{0kh'}V_{ih'}\Big|^3+|V_{ih}|^3\right)+O(1)\Bigg\} \\
&\leqslant \frac{C}{p^2}\sum_{i=1}^{m}\sum_{k=1}^{n}\Big|\sum_{h'=1}^{d}M_{0kh'}V_{ih'}\Big|^3+O(nd) \\
&= O\left(\frac{nd^{3/2}}{p^2}\right), \qquad\qquad (A.33)
\end{aligned}
$$

where the first inequality holds by Assumption 3(1), and the last two lines are due to Cauchy-Schwarz inequality.

By (A.30)-(A.33), Liapunov CLT and Slutsky theorem, we have

$$
\begin{aligned}
&\frac{1}{\sqrt{nd}\gamma_{c_1,c_2}}\begin{pmatrix}c_1\\c_2\end{pmatrix}^T\left[\begin{pmatrix}p^{-2}\sum_{i=1}^{m}\lambda_{pi}^2\\ p^2\sum_{i=1}^{m}(\lambda_i^2+n\sigma^2)\,\hat{p}\end{pmatrix}-\begin{pmatrix}\sum_{i=1}^{m}\left[\lambda_i^2+n\sigma^2\right]\\ p^3\sum_{i=1}^{m}(\lambda_i^2+n\sigma^2)\end{pmatrix}\right] \\
&\to \mathcal{N}(0,1) \text{ in distribution}, \quad \text{as } n,d\to\infty.
\end{aligned}
$$

□

Proof of Proposition 1.6.3. Similarly to the proof of (A.24), we have

$$
\begin{aligned}
&\hat{\tau}_p - np^2\sigma^2 \\
&= \frac{1}{d-r}\mathrm{tr}\left(V_{pc}^T\hat{\Sigma}_pV_{pc}\right)-\frac{1}{d-r}\mathrm{tr}\left(V_c^T\mathbb{E}\,\hat{\Sigma}_pV_c\right) \\
&= \frac{1}{d-r}\mathrm{tr}\left(V_c^T\hat{\Sigma}_pV_c\right) \\
&\quad +\frac{1}{d-r}\mathrm{tr}\left(V_{pc}^T\hat{\Sigma}_pV_{pc}-\mathcal{O}^TV_c^T\hat{\Sigma}_pV_c\mathcal{O}^T\right)-\frac{1}{d-r}\mathrm{tr}\left(V_c^T\mathbb{E}\,\hat{\Sigma}_pV_c\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{d-r}\mathrm{tr}\Big(V_c^T\big(M_y^T M_y - (1-p)\mathrm{diag}(M_y^T M_y) \\
&\qquad\qquad -p^2(1-p)\,\mathrm{diag}(M_0^T M_0)\big)V_c\Big) \\
&\quad + \frac{1}{d-r}\mathrm{tr}\Big(V_c^T\left(\epsilon_y^T\epsilon_y - (1-p)\mathrm{diag}(\epsilon_y^T\epsilon_y)\right)V_c - np^2\sigma^2 I_{d-r}\Big) \\
&\quad -2p(1-p)\frac{1}{d-r}\mathrm{tr}\Big(V_c^T\left(\mathrm{diag}(M_y^T M_0)\right)V_c\Big) \\
&\quad -2p(1-p)\frac{1}{d-r}\mathrm{tr}\Big(V_c^T\left(\mathrm{diag}(\epsilon_y^T M_0)\right)V_c\Big) \\
&\quad +2\frac{1}{d-r}\mathrm{tr}\Big(V_c^T\left(M_y^T\epsilon_y - (1-p)\mathrm{diag}(M_y^T\epsilon_y)\right)V_c\Big) \\
&\quad +\frac{1}{d-r}\mathrm{tr}\left(V_{pc}^T\hat{\Sigma}_p V_{pc} - \mathcal{O}^T V_c^T\hat{\Sigma}_p V_c\mathcal{O}\right) \\
&= (A)+(B)-2p(1-p)\cdot(C)-2p(1-p)\cdot(D)+2\cdot(E)+(F),
\end{aligned}
$$

where $\mathcal{O}\in\mathbb{V}_{d-r,d-r}$ is a solution to $\inf_{Q\in\mathbb{V}_{d-r,d-r}}\|V_{pc}-V_c Q\|_F^2$, and the third equality is due to the fact that $M_0 V_c = U\Lambda V^T V_c = 0$. We will show that (A)-(F) are $O_p\left(p\sqrt{n}\right)$.

Since the first five terms, (A)-(E), are centered, we only need to check their variances to find their rates. The variances of the terms (A), (B), and (E) are $O\left(p^2 n\right)$, which can be shown similarly to the proof of (A.26). The variance of the term (C) is

$$
\begin{aligned}
\mathrm{var}(C) &\leqslant \frac{1}{d-r}\sum_{i=1}^{d-r}\mathbb{E}\left[V_{ci}^T\left(\mathrm{diag}(M_y^T M_0)\right)V_{ci}\right]^2 \\
&= \frac{1}{d-r}\sum_{i=1}^{d-r}\mathrm{var}\left[\sum_{k=1}^{n}\sum_{h=1}^{d}(y_{kh}-p)M_{0kh}^2 V_{cih}^2\right] \\
&= \frac{1}{d-r}\sum_{i=1}^{d-r}\left[L^4\sum_{k=1}^{n}O(p(1-p))\right] \\
&= O(pn),
\end{aligned}
$$

where the inequality is due to Jensen's inequality. Similarly, the variance of the term (D) is $O(pn)$.

Now, consider the term (F). Similarly to the proof of (A.29),

$$\begin{aligned}
|(F)| &\leqslant \frac{1}{d-r}\left|\operatorname{tr}\left(V_{pc}^T\hat{\Sigma}_p V_{pc} - \mathcal{O}^T V_c^T \hat{\Sigma}_p V_c \mathcal{O}\right)\right| \\
&\leqslant \frac{1}{d-r}\sum_{i=1}^{d-r}\left|{V_{pc}}_i^T \hat{\Sigma}_p V_{pc_i} - \mathcal{O}_i^T V_c^T \hat{\Sigma}_p V_c \mathcal{O}_i\right| \\
&\leqslant \frac{1}{d-r}\cdot 2\lambda_{p1}^2 \left\|V_{pc} - V_c\mathcal{O}\right\|_F^2 \\
&\leqslant \frac{1}{d-r}\cdot 4\lambda_{p1}^2 \left\|\sin\left(V_{pc}, V_c\right)\right\|_F^2 \\
&= \frac{1}{d-r}\cdot 2\lambda_{p1}^2 \left\|V_{pc}V_{pc}^T - V_c V_c^T\right\|_F^2 \\
&= \frac{1}{d-r}\cdot 2\lambda_{p1}^2 \left\|(I_d - V_p V_p^T) - (I_d - VV^T)\right\|_F^2 \\
&= \frac{1}{d-r}\cdot 2\lambda_{p1}^2 \left\|V_p V_p^T - VV^T\right\|_F^2 \\
&= \frac{1}{d-r}\cdot 4\lambda_{p1}^2 \left\|\sin(V_p, V)\right\|_F^2 \\
&= O_p(p),
\end{aligned}$$

where $\mathcal{O}_i$ is the i-th column of $\mathcal{O}$, the third inequality can be derived similarly to the proof of (1.24), and the last equality holds by Proposition 1.6.1 and (1.25). □

Proofs for Section 1.6

Proof of Proposition 1.6.4. Let $\Delta_{\lambda_i} = \hat{\lambda}_i - \lambda_i$, $\Delta_{U_i} = \operatorname{sign}(\langle \hat{U}_i, U_i\rangle)\hat{U}_i - U_i$, and $\Delta_{V_i} = \operatorname{sign}(\langle \hat{V}_i, V_i\rangle)\hat{V}_i - V_i$ for all $i \in \{1, \ldots, r\}$. Similarly to the proof of Theorem 1.4.2, we can show that for all $i = 1, \ldots, r$,

$$|\Delta_{\lambda_i}| = O_p\left(\frac{1}{\sqrt{p}} + \frac{1}{p}\sqrt{\frac{d}{n}}\right). \tag{A.34}$$

Then,

$$\left\|\hat{M}(s_0) - M_0\right\|_F^2$$

$$
\begin{aligned}
&= \left\| \sum_{i=1}^{r} s_{0i}\, \hat{\lambda}_i \hat{U}_i \hat{V}_i^T - \sum_{i=1}^{r} \lambda_i U_i V_i^T \right\|_F^2 \\
&\leqslant r^2 \sum_{i=1}^{r} \left\| s_{0i}\, \hat{\lambda}_i \hat{U}_i \hat{V}_i^T - \lambda_i U_i V_i^T \right\|_F^2 \\
&= r^2 \sum_{i=1}^{r} \left\| (\lambda_i + \Delta_{\lambda_i}) (U_i + \Delta_{U_i}) (V_i + \Delta_{V_i})^T - \lambda_i U_i V_i^T \right\|_F^2 \\
&\leqslant C r^2 \sum_{i=1}^{r} \left\{ \left\| \Delta_{\lambda_i} U_i V_i^T \right\|_F^2 + \left\| \lambda_i \Delta_{U_i} V_i^T \right\|_F^2 + \left\| \lambda_i U_i \Delta_{V_i}^T \right\|_F^2 \right\} \\
&= C r^2 \sum_{i=1}^{r} \left\{ O_p \left(\frac{1}{\sqrt{p}} + \frac{1}{p} \sqrt{\frac{d}{n}} \right) + O(nd) \frac{1}{p\, b_r^4} O_p \left(\frac{1}{d} \right) + O(nd) \frac{1}{p\, b_r^4} O_p \left(\frac{1}{n} \right) \right\} \\
&= \frac{1}{p\, b_r^4} O_p(n),
\end{aligned}
$$

where the third equality holds due to (A.34) and Theorem 1.4.1. □

Proofs for Lemma 1.4.1

Proof of Lemma 1.4.1. By Weyl's theorem (Li (1998a)), Lemma 1.6.1, and Lemma 1.6.2, for any given $\delta > 0$, there exists a large constant $C_\delta > 0$ such that

$$
\begin{aligned}
\max \left\{ \left| \lambda_{\hat{p}\, r}^2 - p^2(\lambda_r^2 + n\sigma^2) \right|, \left| \lambda_{\hat{p}\, r+1}^2 - p^2 n \sigma^2 \right| \right\} &\leqslant \left\| \hat{\Sigma}_{\hat{p}} - \mathbb{E}(\hat{\Sigma}_p) \right\|_2 \\
&\leqslant \left\| \hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p \right\|_2 + \left\| \hat{\Sigma}_p - \mathbb{E}(\hat{\Sigma}_p) \right\|_2 \\
&\leqslant C_\delta\, p^{3/2} \sqrt{\frac{n \log n}{d}} \qquad \text{(A.35)}
\end{aligned}
$$

with probability at least $1 - O(n^{-\delta})$. Also, by definition of $\hat{r}$, we have

$$
\begin{aligned}
\left\{ \hat{r} = r \right\} &= \left\{ \lambda_{\hat{p}\, r}^2 \geqslant p^2 n\, C_d,\ \lambda_{\hat{p}\, r+1}^2 < p^2 n\, C_d \right\} \\
&= \Big\{ \left[\lambda_{\hat{p}\, r}^2 - p^2(\lambda_r^2 + n\sigma^2) \right] + p^2(\lambda_r^2 + n\sigma^2) \geqslant p^2 n\, C_d, \\
&\qquad \left[\lambda_{\hat{p}\, r+1}^2 - p^2 n \sigma^2 \right] + p^2 n \sigma^2 < p^2 n\, C_d \Big\}, \qquad \text{(A.36)}
\end{aligned}
$$

where $\lambda_r^2 = b_r^2\, nd$ by Assumption 3(1). The result follows by (A.35) and (A.36).

□

B APPENDIX FOR CHAPTER 2

Denote by C and C_1 generic constants whose values are free of n and p and may change from appearance to appearance. Also, denote by $\|v\|_2$ the ℓ_2-norm for any vector $v \in \mathbb{R}^d$ and by $\|A\|_2$ the spectral norm, the largest singular value of A, for any matrix $A \in \mathbb{R}^{n\times d}$.

Proof of Theorem 2.3.1

Proof of Theorem 2.3.1. We have

$$\begin{aligned}\widetilde{M}_t &= \mathcal{P}_\Omega(M) + \mathcal{P}_\Omega^\perp(Z_t)\\ &= y\cdot(M_0+\epsilon) + (1_n 1_d^T - y)\cdot Z_t\\ &= M_0 + y\cdot\epsilon + (1_n 1_d^T - y)\cdot\eta_t,\end{aligned}$$

where 1_n and 1_d are vectors of length n and d, respectively, filled with ones and $\eta_t = Z_t - M_0$, and, $A\cdot B = (A_{ij}B_{ij})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant d}$ for any A and $B \in \mathbb{R}^{n\times d}$. Assume that

$$\frac{1}{\sqrt{nd}}\|\eta_t\|_F = o_p\left(\sqrt{\frac{h_n}{p\,d}}\right). \tag{B.1}$$

Then, simple algebraic manipulations show for large n

$$\begin{aligned}\frac{1}{\sqrt{nd}}\|\eta_{t+1}\|_F &= \frac{1}{\sqrt{nd}}\|Z_{t+1} - M_0\|_F\\ &= \frac{1}{\sqrt{nd}}\left\|\sum_{i=1}^r \sqrt{\lambda_i^2(\widetilde{M}_t) - \widetilde{\alpha}_t}\,\mathbf{u}_i(\widetilde{M}_t)\mathbf{v}_i(\widetilde{M}_t)^T - \sum_{i=1}^r \lambda_i U_i V_i^T\right\|_F\\ &\leqslant \frac{C}{\sqrt{nd}}\sum_{i=1}^r\Bigg\{\left|\sqrt{\lambda_i^2(\widetilde{M}_t) - \widetilde{\alpha}_t} - \lambda_i\right|\|U_i V_i^T\|_F\\ &\qquad + \lambda_i\left\|\left(\mathbf{u}_i(\widetilde{M}_t) - U_i\mathcal{O}_i\right)V_i^T\right\|_F + \lambda_i\left\|U_i\left(\mathbf{v}_i(\widetilde{M}_t) - V_i\Omega_i\right)^T\right\|_F\Bigg\}\end{aligned}$$

$$\leqslant \quad C\sum_{i=1}^{r}\left\{\frac{1}{\sqrt{nd}}\left|\sqrt{\lambda_i^2(\widetilde{M}_t)-\widetilde{\alpha}_t}-\lambda_i\right| \right.$$
$$\left. +\frac{\lambda_i}{\sqrt{nd}}\left\|\mathbf{u}_i(\widetilde{M}_t)-U_i\mathcal{O}_i\right\|_F+\frac{\lambda_i}{\sqrt{nd}}\left\|\mathbf{v}_i(\widetilde{M}_t)-V_i\mathcal{Q}_i\right\|_F\right\}, \quad \text{(B.2)}$$

where $\mathcal{O}_i$ and $\mathcal{Q}_i$ are in $\{-1,1\}$ and minimize $\left\|\mathbf{u}_i(\widetilde{M}_t)-U_i\mathcal{O}_i\right\|_F$ and $\left\|\mathbf{v}_i(\widetilde{M}_t)-V_i\mathcal{Q}_i\right\|_F$, respectively.

To find the order of (B.2), first consider the term $\left\|\mathbf{v}_i(\widetilde{M}_t)-V_i\mathcal{O}_i\right\|_F$. By Davis-Kahan Theorem (Theorem 3.1 in Li (1998b)) and Proposition 2.2 in Vu and Lei (2013),

$$\left\|\mathbf{v}_i(\widetilde{M}_t)-V_i\mathcal{O}_i\right\|_F \leqslant \frac{\frac{1}{nd}\left\|\left(\widetilde{M}_t^T\widetilde{M}_t-\left[M_0^TM_0+np\sigma^2I\right]\right)V_i\right\|_F}{\left|\frac{1}{nd}\left(\lambda_i^2+np\sigma^2-\boldsymbol{\lambda}_{i+1}^2(\widetilde{M}_t)\right)\right|}. \quad \text{(B.3)}$$

Consider the numerator of (B.3). We have

$$\begin{aligned}
&\frac{1}{nd}\left\|\left(\widetilde{M}_t^T\widetilde{M}_t-\left[M_0^TM_0+np\sigma I\right]\right)V_i\right\|_F \\
&\leqslant \frac{1}{nd}\Big\{\left\|(y\cdot\epsilon)^T(y\cdot\epsilon)V_i-np\sigma^2V_i\right\|_F+\left\|\left[(1_n1_d^T-y)\cdot\eta_t\right]^T\left[(1_n1_d^T-y)\cdot\eta_t\right]V_i\right\|_F \\
&\qquad +\left\|M_0^T(y\cdot\epsilon)V_i\right\|_F+\left\|(y\cdot\epsilon)^TM_0V_i\right\|_F \\
&\qquad +\left\|M_0^T\left[(1_n1_d^T-y)\cdot\eta_t\right]V_i\right\|_F+\left\|\left[(1_n1_d^T-y)\cdot\eta_t\right]^TM_0V_i\right\|_F \\
&\qquad +\left\|\left[(1_n1_d^T-y)\cdot\eta_t\right]^T(y\cdot\epsilon)V_i\right\|_F+\left\|(y\cdot\epsilon)^T\left[(1_n1_d^T-y)\cdot\eta_t\right]V_i\right\|_F\Big\} \\
&=\frac{1}{nd}\Big\{O_p\left(p\sqrt{nd}\right)+o_p\left(\frac{nh_n}{p}\right)+O_p\left(\sqrt{pdn^2}\right)+O_p\left(\sqrt{pd^2n}\right) \\
&\qquad +o_p\left(\sqrt{\frac{h_ndn^2}{p}}\right)+o_p\left(\sqrt{\frac{h_ndn^2}{p}}\right)+o_p\left(\sqrt{h_nn^2}\right)+o_p\left(\sqrt{h_ndn^2}\right)\Big\} \\
&=o_p\left(\sqrt{\frac{h_n}{pd}}\right), \qquad \text{(B.4)}
\end{aligned}$$

where the first equality holds due to (2.1), Assumption 3(2), (B.1), and (B.5) and

(B.6) below. We have

$$\begin{aligned}
&\mathbb{E}\left\|(y\cdot\epsilon)^{T}(y\cdot\epsilon)V_{i}-np\sigma^{2}V_{i}\right\|_{F}^{2}\\
&=\mathbb{E}\left\{\sum_{h=1}^{d}\left[\sum_{k=1}^{n}\sum_{j=1}^{d}\left(y_{kh}y_{kj}\epsilon_{kh}\epsilon_{kj}V_{ij}-p\sigma^{2}V_{ih}\mathbb{1}_{(j=h)}\right)\right]^{2}\right\}\\
&=\sum_{h=1}^{d}\sum_{k=1}^{n}\sum_{j=1}^{d}\mathbb{E}\left(y_{kh}y_{kj}\epsilon_{kh}\epsilon_{kj}V_{ij}-p\sigma^{2}V_{ih}\mathbb{1}_{(j=h)}\right)^{2}\\
&=\sum_{k=1}^{n}\sum_{j\neq h}^{d}V_{ij}^{2}\,\mathbb{E}\left(y_{kh}^{2}y_{kj}^{2}\epsilon_{kh}^{2}\epsilon_{kj}^{2}\right)+\sum_{k=1}^{n}\sum_{j=h}^{d}V_{ij}^{2}\,\mathbb{E}\left(y_{kj}^{2}\epsilon_{kj}^{2}-p\sigma^{2}\right)^{2}\\
&=O\left(p^{2}nd\right),
\end{aligned}\tag{B.5}$$

where V_{ij} is the j-th element of V_i. Similarly, we have

$$\mathbb{E}\|(y\cdot\epsilon)V_{i}\|_{F}^{2}=O\left(pn\right),\mathbb{E}\|U_{i}^{T}(y\cdot\epsilon)\|_{F}^{2}=O\left(pd\right),\text{ and }\mathbb{E}\|y\cdot\epsilon\|_{F}^{2}=O\left(pnd\right).\tag{B.6}$$

Consider the denominator of (B.3). By Weyl's theorem (Theorem 4.3 in **?**), we have

$$\begin{aligned}
&\max_{1\leqslant i\leqslant d}\frac{1}{nd}|\lambda_{i}^{2}+np\sigma^{2}-\lambda_{i}^{2}(\widetilde{M}_{t})|\\
&\leqslant\frac{1}{nd}\left\|\widetilde{M}_{t}^{T}\widetilde{M}_{t}-\left[M_{0}^{T}M_{0}+np\sigma^{2}I\right]\right\|_{2}\\
&\leqslant\frac{1}{nd}\left\{\left\|(y\cdot\epsilon)^{T}(y\cdot\epsilon)-np\sigma^{2}I\right\|_{2}+\left\|\left[(1_{n}1_{d}^{T}-y)\cdot\eta_{t}\right]^{T}\left[(1_{n}1_{d}^{T}-y)\cdot\eta_{t}\right]\right\|_{2}\right.\\
&\qquad\left.+2\left\|M_{0}^{T}(y\cdot\epsilon)\right\|_{2}+2\left\|M_{0}^{T}\left[(1-y)\cdot\eta_{t}\right]\right\|_{2}+2\left\|(y\cdot\epsilon)^{T}\left[(1_{n}1_{d}^{T}-y)\cdot\eta_{t}\right]\right\|_{2}\right\}\\
&=\frac{1}{nd}\left\{O_{p}\left(p\sqrt{nd^{2}}\right)+o_{p}\left(\frac{nh_{n}}{p}\right)+O_{p}\left(\sqrt{p\,dn^{2}}\right)+o_{p}\left(\sqrt{\frac{dn^{2}h_{n}}{p}}\right)+o_{p}\left(\sqrt{dn^{2}h_{n}}\right)\right\}\\
&=o_{p}(1),
\end{aligned}\tag{B.7}$$

where the last two lines holds similarly to (B.4).

Thus, by (B.7) and (B.4),

$$\|\mathbf{v}_i(\widetilde{M}_t) - V_i\mathcal{O}_i\|_F = o_p\left(\sqrt{\frac{h_n}{pd}}\right). \tag{B.8}$$

Secondly, similar to the proof of (B.8), we can show $\|\mathbf{u}_i(\widetilde{M}_t) - U_i\mathcal{O}_i\|_F = o_p\left(\sqrt{h_n/pd}\right)$.

Lastly, consider the term $\frac{1}{\sqrt{nd}}\left|\sqrt{\boldsymbol{\lambda}_i^2(\widetilde{M}_t) - \widetilde{\boldsymbol{\alpha}}_t} - \lambda_i\right|$. By Taylor's expansion, there is λ_*^2 between $\boldsymbol{\lambda}_i^2(\widetilde{M}_t) - \widetilde{\boldsymbol{\alpha}}_t$ and λ_i^2 such that

$$\begin{aligned}
&\frac{1}{\sqrt{nd}}\left|\sqrt{\boldsymbol{\lambda}_i^2(\widetilde{M}_t) - \widetilde{\boldsymbol{\alpha}}_t} - \lambda_i\right| \\
&= \frac{1}{\sqrt{nd}}\left|\frac{1}{2\lambda_*}\left(\boldsymbol{\lambda}_i^2(\widetilde{M}_t) - \widetilde{\boldsymbol{\alpha}}_t - \lambda_i^2\right)\right| \\
&\leqslant \frac{1}{2\lambda_*\sqrt{nd}}\left|\boldsymbol{\lambda}_i^2(\widetilde{M}_t) - (\lambda_i^2 + np\sigma^2)\right| + \frac{1}{2\lambda_*\sqrt{nd}}\left|\widetilde{\boldsymbol{\alpha}}_t - np\sigma^2\right|.
\end{aligned} \tag{B.9}$$

We need to find the convergence rates of $\frac{1}{nd}\left|\boldsymbol{\lambda}_i^2(\widetilde{M}_t) - (\lambda_i^2 + np\sigma^2)\right|$ and $\frac{1}{nd}\left|\widetilde{\boldsymbol{\alpha}}_t - np\sigma^2\right|$. Let $V_c = (V_{r+1}, \ldots, V_d) \in \mathbb{R}^{d\times(d-r)}$ be a matrix such that $V_c^T V_c = I_{d-r}$ and $V^T V_c = 0_{r\times(d-r)}$ and let $\widetilde{V}_t = \left(\mathbf{v}_1(\widetilde{M}_t), \ldots, \mathbf{v}_d(\widetilde{M}_t)\right) \in \mathbb{R}^{d\times r}$ and $\widetilde{V}_{tc} = \left(\mathbf{v}_{r+1}(\widetilde{M}_t), \ldots, \mathbf{v}_d(\widetilde{M}_t)\right) \in \mathbb{R}^{d\times(d-r)}$ so that $\widetilde{V}_t^T\widetilde{V}_{tc} = 0_{r\times(d-r)}$. Also, let $\mathcal{O} = \operatorname{diag}(\mathcal{O}_1, \ldots, \mathcal{O}_r)$ and $\mathcal{O}_c = \operatorname{diag}(\mathcal{O}_{r+1}, \ldots, \mathcal{O}_d)$, where

$$\mathcal{O}_i := \underset{o\in\{-1,1\}}{\arg\min}\left\|V_i\, o - \mathbf{v}_i(\widetilde{M}_t)\right\|_2^2 \quad \text{for } i = 1, \ldots, d.$$

Then, we have

$$\begin{aligned}
&\frac{1}{nd}\left|\boldsymbol{\lambda}_i^2(\widetilde{M}_t) - (\lambda_i^2 + np\sigma^2)\right| \\
&= \frac{1}{nd}\left|\mathbf{v}_i(\widetilde{M}_t)^T\widetilde{M}_t^T\widetilde{M}_t\mathbf{v}_i(\widetilde{M}_t) - V_i^T\left(M_0{}^TM_0 + np\sigma^2 I\right)V_i\right| \\
&\leqslant \frac{1}{nd}\left|V_i^T\left[\widetilde{M}_t^T\widetilde{M}_t - \left(M_0{}^TM_0 + np\sigma^2 I\right)\right]V_i\right| + \frac{1}{nd}\left|\mathbf{v}_i(\widetilde{M}_t)^T\widetilde{M}_t^T\widetilde{M}_t\mathbf{v}_i(\widetilde{M}_t) - V_i^T\widetilde{M}_t^T\widetilde{M}_tV_i\right|
\end{aligned}$$

$$
\begin{aligned}
&\leqslant o_p\left(\sqrt{\frac{h_n}{pd}}\right)+\frac{1}{nd}\left|\mathbf{v}_i(\widetilde{M}_t)^T\widetilde{M}_t^T\widetilde{M}_t\mathbf{v}_i(\widetilde{M}_t)-V_i^T\widetilde{M}_t^T\widetilde{M}_tV_i\right|\\
&=o_p\left(\sqrt{\frac{h_n}{pd}}\right),
\end{aligned}
\tag{B.10}
$$

where the second inequality can be derived by the similar way to the proof of (B.4), and the last equality is due to (B.11) below. Simple algebraic manipulations show

$$
\begin{aligned}
&\frac{1}{nd}\left|\mathbf{v}_i(\widetilde{M}_t)^T\widetilde{M}_t^T\widetilde{M}_t\mathbf{v}_i(\widetilde{M}_t)-V_i^T\widetilde{M}_t^T\widetilde{M}_tV_i\right|\\
&=\frac{1}{nd}\left|\left[V_i\mathcal{O}_i-\mathbf{v}_i(\widetilde{M}_t)\right]^T\widetilde{M}_t^T\widetilde{M}_t\left[V_i\mathcal{O}_i-\mathbf{v}_i(\widetilde{M}_t)\right]+2\lambda_i^2(\widetilde{M}_t)\left[V_i\mathcal{O}_i-\mathbf{v}_i(\widetilde{M}_t)\right]^T\mathbf{v}_i(\widetilde{M}_t)\right|\\
&\leqslant\frac{2\lambda_1^2(\widetilde{M}_t)}{nd}\left\|V_i\mathcal{O}_i-\mathbf{v}_i(\widetilde{M}_t)\right\|_2^2\\
&=o_p\left(\frac{h_n}{pd}\right),
\end{aligned}
\tag{B.11}
$$

where the last equality is due to (B.7) and (B.8). Also,

$$
\begin{aligned}
&\frac{1}{nd}\left|\widetilde{\alpha}_t-np\sigma^2\right|\\
&=\frac{1}{nd}\left|\frac{1}{d-r}\sum_{j=r+1}^{d}\mathbf{v}_j(\widetilde{M}_t)^T\widetilde{M}_t^T\widetilde{M}_t\mathbf{v}_j(\widetilde{M}_t)-np\sigma^2\right|\\
&\leqslant\frac{1}{nd}\left|\frac{1}{d-r}\sum_{j=r+1}^{d}V_j^T\left[\widetilde{M}_t^T\widetilde{M}_t-\left(M_0^TM_0+np\sigma^2I\right)\right]V_j\right|\\
&\qquad+\frac{1}{nd}\left|\frac{1}{d-r}\sum_{j=r+1}^{d}\left[\mathbf{v}_j(\widetilde{M}_t)^T\widetilde{M}_t^T\widetilde{M}_t\mathbf{v}_j(\widetilde{M}_t)-V_j^T\widetilde{M}_t^T\widetilde{M}_tV_j\right]\right|\\
&=o_p\left(\sqrt{\frac{h_n}{pd}}\right)+\frac{1}{nd}\left|\frac{1}{d-r}\sum_{j=r+1}^{d}\left[\mathbf{v}_j(\widetilde{M}_t)^T\widetilde{M}_t^T\widetilde{M}_t\mathbf{v}_j(\widetilde{M}_t)-V_j^T\widetilde{M}_t^T\widetilde{M}_tV_j\right]\right|\\
&=o_p\left(\sqrt{\frac{h_n}{pd}}\right),
\end{aligned}
\tag{B.12}
$$

where the second equality can be derived by the similar way to the proof of (B.4),

and the last equality is due to (B.13) below. Similar to the proof of (B.11), we have

$$
\begin{aligned}
&\frac{1}{nd(d-r)} \sum_{j=r+1}^{d} \left| \mathbf{v}_j(\widetilde{M}_t)^T \widetilde{M}_t^T \widetilde{M}_t \mathbf{v}_j(\widetilde{M}_t) - V_j^T \widetilde{M}_t^T \widetilde{M}_t V_j \right| \\
&\leqslant \frac{1}{nd(d-r)} \sum_{j=r+1}^{d} 2\lambda_1^2(\widetilde{M}_t) \left\| V_j \mathcal{O}_j - \mathbf{v}_j(\widetilde{M}_t) \right\|_2^2 \\
&\leqslant \frac{2\lambda_1^2(\widetilde{M}_t)}{nd(d-r)} \left\| V_c \mathcal{O}_c - \widetilde{V}_{tc} \right\|_F^2 \\
&\leqslant \frac{4\lambda_1^2(\widetilde{M}_t)}{nd(d-r)} \left\| V_c V_c^T - \widetilde{V}_{tc} \widetilde{V}_{tc}^T \right\|_F^2 \\
&= \frac{4\lambda_1^2(\widetilde{M}_t)}{nd(d-r)} \left\| VV^T - \widetilde{V}_t \widetilde{V}_t^T \right\|_F^2 \\
&\leqslant \frac{4\lambda_1^2(\widetilde{M}_t)}{nd(d-r)} \left\| V\mathcal{O} - \widetilde{V}_t \right\|_F^2 \\
&= \frac{4\lambda_1^2(\widetilde{M}_t)}{nd(d-r)} \sum_{i=1}^{r} \left\| V_i \mathcal{O}_i - \mathbf{v}_i(\widetilde{M}_t) \right\|_2^2 \\
&= o_p\left(\frac{h_n}{pd^2} \right), \qquad \text{(B.13)}
\end{aligned}
$$

where the fourth and sixth lines are due to Proposition 2.2 in Vu and Lei (2013), and the last line holds from (B.8).

The three results above (B.9), (B.10), and (B.12) give $\frac{1}{\sqrt{nd}} \left| \sqrt{\lambda_i^2(\widetilde{M}_t) - \widetilde{\alpha}_t} - \lambda_i \right| = o_p\left(\sqrt{\frac{h_n}{pd}} \right)$.

Therefore, combining the results above, we have that $\frac{1}{\sqrt{nd}} \|\eta_{t+1}\|_F$ in (B.2) is $o_p\left(\sqrt{\frac{h_n}{pd}} \right)$. Since $\frac{1}{\sqrt{nd}} \|\eta_1\|_F = o_p\left(\sqrt{\frac{h_n}{pd}} \right)$ by Proposition 2.3.1, we have $\frac{1}{\sqrt{nd}} \|\eta_t\|_F = o_p\left(\sqrt{\frac{h_n}{pd}} \right)$ for any fixed t by mathematical induction. □

Proof of Theorem 2.3.2

Proof of Theorem 2.3.2. We have

$$\begin{aligned}
&\min_{Z} \ \frac{1}{2nd} \|X - Z\|_F^2 + \sum_{i=1}^{d} \frac{\tau_i}{\sqrt{nd}} \frac{\lambda_i(Z)}{\sqrt{nd}} \\
&= \min_{Z} \ \frac{1}{2nd} \left\{ \|X\|_F^2 - 2 \sum_{i=1}^{d} \tilde{\lambda}_i \cdot \tilde{u}_i^T X \tilde{v}_i + \sum_{i=1}^{d} \tilde{\lambda}_i^2 \right\} + \frac{1}{nd} \sum_{i=1}^{d} \tau_i \tilde{\lambda}_i, \quad \text{(B.14)}
\end{aligned}$$

where $\tilde{\lambda}_i = \lambda_i(Z)$, $\tilde{u}_i = \mathbf{u}_i(Z)$, and $\tilde{v}_i = \mathbf{v}_i(Z)$. Minimizing (B.14) is equivalent to minimizing

$$-2 \sum_{i=1}^{d} \tilde{\lambda}_i \cdot \tilde{u}_i^T X \tilde{v}_i + \sum_{i=1}^{d} \tilde{\lambda}_i^2 + \sum_{i=1}^{d} 2\tau_i \tilde{\lambda}_i,$$

with respect to $\tilde{\lambda}_i, \tilde{u}_i$, and $\tilde{v}_i$, $i = 1, \ldots, d$, under the conditions that $(\tilde{u}_1, \ldots, \tilde{u}_d)^T (\tilde{u}_1, \ldots, \tilde{u}_d) = I_d$, $(\tilde{v}_1, \ldots, \tilde{v}_d)^T (\tilde{v}_1, \ldots, \tilde{v}_d) = I_d$, and $\tilde{\lambda}_1 \geqslant \tilde{\lambda}_2 \geqslant \ldots \geqslant \tilde{\lambda}_d \geqslant 0$. Thus, we have

$$\begin{aligned}
&\min_{\tilde{\lambda}_i \geqslant 0, \tilde{u}_i, \tilde{v}_i, \ i=1,\ldots,d} \ -2 \sum_{i=1}^{d} \tilde{\lambda}_i \cdot \tilde{u}_i^T X \tilde{v}_i + \sum_{i=1}^{d} \tilde{\lambda}_i^2 + \sum_{i=1}^{d} 2\tau_i \tilde{\lambda}_i \\
&= \min_{\tilde{\lambda}_i \geqslant 0, \ i=1,\ldots,d} \ -2 \sum_{i=1}^{d} \tilde{\lambda}_i \cdot \lambda_i(X) + \sum_{i=1}^{d} \tilde{\lambda}_i^2 + \sum_{i=1}^{d} 2\tau_i \tilde{\lambda}_i \\
&= \min_{\tilde{\lambda}_i \geqslant 0, \ i=1,\ldots,d} \ \sum_{i=1}^{d} \left\{ \tilde{\lambda}_i^2 - 2\tilde{\lambda}_i \left[\lambda_i(X) - \tau_i \right] \right\}, \quad \text{(B.15)}
\end{aligned}$$

where the first equality is due to the facts that $\tilde{\lambda}_1 \geqslant \ldots \geqslant \tilde{\lambda}_d \geqslant 0$, and for every i, the problem

$$\max_{\|u_i\|_2^2 \leqslant 1, \|v_i\|_2^2 \leqslant 1} u_i^T X v_i \quad \text{such that} \quad u_i \perp \{\tilde{u}_1^*, \ldots, \tilde{u}_{i-1}^*\}, v_i \perp \{\tilde{v}_1^*, \ldots, \tilde{v}_{i-1}^*\}$$

is solved by $\tilde{u}_i^*, \tilde{v}_i^*$, the left and right singular vectors of X corresponding to the i-th largest singular value of X. Note that $\tilde{u}_i = \tilde{u}_i^*$. Since (B.15) is a quadratic function

of $\tilde{\lambda}_i$, the solution to the problem (B.15) is then $\tilde{\lambda}_i = (\boldsymbol{\lambda}_i(X) - \tau_i)_+$. □

Proof of Theorem 2.4.1

To ease the notation, we drop the superscript 'g' in Z_t^g, $\widetilde{M}_t^g$, and D_t^g in this section.

Lemma B.0.1. *Let* $Z_{t+1} := \arg\min_{Z \in \mathbb{R}^{n\times d}} Q_\tau(Z|Z_t)$ *in* (2.9). *Then, under Assumption 5, we have*

$$\|Z_{t+1} - Z_t\|_F^2 \to 0 \quad \textit{as } t \to \infty.$$

Proof of Lemma B.0.1. By the construction of D_t,

$$(\widetilde{M}_{t-1} - \widetilde{M}_t) - (Z_t - Z_{t+1}) - (D_{t-1} - D_t) = 0.$$

Thus, we have

$$\langle \widetilde{M}_{t-1} - \widetilde{M}_t, Z_t - Z_{t+1}\rangle - \langle Z_t - Z_{t+1}, Z_t - Z_{t+1}\rangle - \langle D_{t-1} - D_t, Z_t - Z_{t+1}\rangle = 0 \tag{B.16}$$

and

$$\langle \widetilde{M}_{t-1} - \widetilde{M}_t, \widetilde{M}_{t-1} - \widetilde{M}_t\rangle - \langle Z_t - Z_{t+1}, \widetilde{M}_{t-1} - \widetilde{M}_t\rangle - \langle D_{t-1} - D_t, \widetilde{M}_{t-1} - \widetilde{M}_t\rangle = 0. \tag{B.17}$$

Add (B.17) and (B.16), and

$$\begin{aligned} 0 &= \|\widetilde{M}_{t-1} - \widetilde{M}_t\|_F^2 - \|Z_t - Z_{t+1}\|_F^2 - \langle D_{t-1} - D_t, Z_t + \widetilde{M}_{t-1} - (Z_{t+1} + \widetilde{M}_t)\rangle \\ &= \|\widetilde{M}_{t-1} - \widetilde{M}_t\|_F^2 - \|Z_t - Z_{t+1}\|_F^2 - \|D_{t-1} - D_t\|_F^2 - 2\langle D_{t-1} - D_t, Z_t - Z_{t+1}\rangle. \end{aligned} \tag{B.18}$$

Under Assumption 5, (B.18) gives

$$\|Z_t - Z_{t+1}\|_F^2 \leqslant \left\|\widetilde{M}_{t-1} - \widetilde{M}_t\right\|_F^2,$$

and thus

$$\begin{aligned} \|Z_{t+1} - Z_t\|_F^2 &\leqslant \left\|\widetilde{M}_{t-1} - \widetilde{M}_t\right\|_F^2 \\ &\leqslant \left\|\mathcal{P}_\Omega^\perp (Z_{t-1} - Z_t)\right\|_F^2 \\ &\leqslant \|Z_t - Z_{t-1}\|_F^2 \end{aligned} \tag{B.19}$$

for all $t \geqslant 1$. This proves that the sequence $\{\|Z_{t+1} - Z_t\|_F^2\}$ converges (since it is decreasing and bounded below).

The convergence of $\{\|Z_{t+1} - Z_t\|_F^2\}$ gives

$$\|Z_{t+1} - Z_t\|_F^2 - \|Z_t - Z_{t-1}\|_F^2 \to 0 \text{ as } t \to \infty.$$

Then, by (B.19),

$$\begin{aligned} 0 &\geqslant \left\|\mathcal{P}_\Omega^\perp (Z_t - Z_{t-1})\right\|_F^2 - \|Z_t - Z_{t-1}\|_F^2 \\ &\geqslant \|Z_{t+1} - Z_t\|_F^2 - \|Z_t - Z_{t-1}\|_F^2 \\ &\to 0 \quad \text{as } t \to \infty, \end{aligned}$$

which implies

$$\left\|\mathcal{P}_\Omega^\perp (Z_t - Z_{t-1})\right\|_F^2 - \|Z_t - Z_{t-1}\|_F^2 \to 0 \Rightarrow \|\mathcal{P}_\Omega (Z_t - Z_{t-1})\|_F^2 \to 0. \tag{B.20}$$

Furthermore, similarly to the proof of Lemma 2 in Mazumder et al. (2010), we can show

$$f_\tau(Z_t) \geqslant Q_\tau(Z_{t+1}|Z_t) \geqslant Q_\tau(Z_{t+1}|Z_{t+1}) = f_\tau(Z_{t+1}) \geqslant 0 \tag{B.21}$$

for every fixed $\tau_1, \ldots, \tau_d > 0$ and $t \geqslant 1$. Thus, we have

$$Q_\tau(Z_{t+1}|Z_t) - Q_\tau(Z_{t+1}|Z_{t+1}) \to 0 \quad \text{as } t \to \infty,$$

which implies

$$\left\|\mathcal{P}_\Omega^\perp (Z_t - Z_{t+1})\right\|_F^2 \to 0 \quad \text{as } t \to \infty.$$

The above along with (B.20) gives

$$\|Z_{t+1} - Z_t\|_F^2 \to 0 \quad \text{as } t \to \infty.$$

□

Proof of Theorem 2.4.1. By the construction of D_t, we have

$$0 = \left(\widetilde{M}_t - Z_{t+1}\right) - D_t \quad \text{for all } t \geqslant 1.$$

Since Z_∞ is a limit point of the sequence Z_t, there exists a subsequence $\{n_t\} \subset \{1, 2, \ldots\}$ such that $Z_{n_t} \to Z_\infty$ as $t \to \infty$. By Lemma B.0.1, this subsequence Z_{n_t} satisfies

$$Z_{n_t} - Z_{n_t+1} \to 0$$

which implies

$$\mathcal{P}_\Omega^\perp(Z_{n_t}) - Z_{n_t+1} \to \mathcal{P}_\Omega^\perp(Z_\infty) - Z_\infty = -\mathcal{P}_\Omega(Z_\infty).$$

Hence,

$$D_{n_t} = \left(\mathcal{P}_\Omega(M) + \mathcal{P}_\Omega^\perp(Z_{n_t})\right) - Z_{n_t+1} \to \mathcal{P}_\Omega(M) - \mathcal{P}_\Omega(Z_\infty) = D_\infty. \tag{B.22}$$

Due to (2.11) and (B.22), we have

$$\begin{aligned} f_\tau(Z^s) &\geqslant f_\tau(Z_\infty) - \frac{1}{nd}\langle Z^s - Z_\infty, \mathcal{P}_\Omega(M) - \mathcal{P}_\Omega(Z_\infty) - D_\infty\rangle \\ &= f_\tau(Z_\infty). \end{aligned}$$

Since $f_\tau(Z^s) \leqslant f_\tau(Z_\infty)$ by definition of Z^s, we have $f_\tau(Z^s) = f_\tau(Z_\infty)$. Lastly, by (B.21), we have $\lim_{t\to\infty} f_\tau(Z_t) = f(Z^s)$. □

Proofs of Lemmas 2.4.1-2.4.2

Proof of Lemma 2.4.1. For $i = 1, \ldots, r$, we have

$$\begin{aligned}
&\left|\frac{\tau_{t,i}}{\sqrt{nd}} - \frac{\tau_{t+1,i}}{\sqrt{nd}}\right| \\
&= \frac{1}{\sqrt{nd}}\left|\lambda_i(\widetilde{M}_t) - \sqrt{\lambda_i^2(\widetilde{M}_t) - \widetilde{\alpha}_t} - \lambda_i(\widetilde{M}_{t+1}) + \sqrt{\lambda_i^2(\widetilde{M}_{t+1}) - \widetilde{\alpha}_{t+1}}\right| \\
&\leqslant \frac{1}{\sqrt{nd}}\left|\lambda_i(\widetilde{M}_t) - \left(\sqrt{\lambda_i^2 - np\sigma^2}\right)\right| + \frac{1}{\sqrt{nd}}\left|\sqrt{\lambda_i^2(\widetilde{M}_t) - \widetilde{\alpha}_t} - \lambda_i^2\right| \\
&\quad + \frac{1}{\sqrt{nd}}\left|\lambda_i(\widetilde{M}_{t+1}) - \left(\sqrt{\lambda_i^2 - np\sigma^2}\right)\right| + \frac{1}{\sqrt{nd}}\left|\sqrt{\lambda_i^2(\widetilde{M}_{t+1}) - \widetilde{\alpha}_{t+1}} - \lambda_i^2\right| \\
&= (I) + (II) + (III) + (IV).
\end{aligned}$$

Then, by (B.10) and (B.12), we have

$$\begin{aligned}
(I) &= \frac{1}{\sqrt{nd}}\left|\lambda_i(\widetilde{M}_t) - \left(\sqrt{\lambda_i^2 - np\sigma^2}\right)\right| \\
&= \frac{1}{2\lambda_*\sqrt{nd}}\left|\lambda_i^2(\widetilde{M}_t) - (\lambda_i^2 - np\sigma^2)\right| \\
&\leqslant \frac{1}{2\lambda_*\sqrt{nd}}\left|\lambda_i^2(\widetilde{M}_t) - \widetilde{\alpha}_t - \lambda_i^2\right| + \frac{1}{2\lambda_*\sqrt{nd}}\left|\widetilde{\alpha}_t - np\sigma^2\right| \\
&= o_p\left(\sqrt{\frac{h_n}{pd}}\right),
\end{aligned}$$

where the second equality holds for some λ_* between $\lambda_i(\widetilde{M}_t)$ and $\sqrt{\lambda_i^2 - np\sigma^2}$ by Taylor's expansion. We can similarly show that $(III) = o_p\left(\sqrt{h_n/pd}\right)$. Both of (II) and (IV) are also $o_p\left(\sqrt{h_n/pd}\right)$ by (B.9) and (B.10). □

Proof of Lemma 2.4.2. From Theorem 2.3.1 and the construction of D_t in Assumption 5, we have

$$\begin{aligned}
&\left|\frac{1}{nd}\langle D_t - D_{t+1}, Z_{t+1} - Z_{t+2}\rangle\right| \\
&\leqslant \frac{1}{nd}\left\|D_t - D_{t+1}\right\|_F \left\|Z_{t+1} - Z_{t+2}\right\|_F
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \frac{1}{nd}\left\|\widetilde{M}_t - Z_{t+1} - \left(\widetilde{M}_{t+1} - Z_{t+2}\right)\right\|_F \|Z_{t+1} - Z_{t+2}\|_F \\
&\leqslant \frac{1}{nd}\left\{ \left\|\widetilde{M}_t - \widetilde{M}_{t+1}\right\|_F + \|Z_{t+1} - Z_{t+2}\|_F \right\} \|Z_{t+1} - Z_{t+2}\|_F \\
&= \frac{1}{nd}\left\{ \left\|\mathcal{P}_\Omega^\perp\left(Z_t - Z_{t+1}\right)\right\|_F + \|Z_{t+1} - Z_{t+2}\|_F \right\} \|Z_{t+1} - Z_{t+2}\|_F \\
&\leqslant \frac{1}{nd}\left\{ \|Z_t - Z_{t+1}\|_F + \|Z_{t+1} - Z_{t+2}\|_F \right\} \|Z_{t+1} - Z_{t+2}\|_F \\
&\leqslant \frac{1}{nd}\left\{ \|Z_t - M_0\|_F + 2\|Z_{t+1} - M_0\|_F + \|Z_{t+2} - M_0\|_F \right\} \\
&\quad \times \left\{ \|Z_{t+1} - M_0\|_F + \|Z_{t+2} - M_0\|_F \right\} \\
&= o_p\left(\frac{h_n}{pd}\right).
\end{aligned}
$$

□

REFERENCES

Achlioptas, Dimitris, and Frank McSherry. 2001. Fast computation of low rank matrix approximations. In *Proceedings of the thirty-third annual acm symposium on theory of computing*, 611–618. ACM.

Airoldi, Edo M, David M Blei, Stephen E Fienberg, and Eric P Xing. 2009. Mixed membership stochastic blockmodels. In *Advances in neural information processing systems*, 33–40.

Alon, Uri. 2006. *An introduction to systems biology: design principles of biological circuits*. CRC press.

Anderson, Theodore Wilbur, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Theodore Wilbur Anderson. 1958. *An introduction to multivariate statistical analysis*, vol. 2. Wiley New York.

Azar, Yossi, Amos Fiat, Anna Karlin, Frank McSherry, and Jared Saia. 2001. Spectral analysis of data. In *Proceedings of the thirty-third annual acm symposium on theory of computing*, 619–626. ACM.

Bennett, James, and Stan Lanning. 2007. The netflix prize. In *Proceedings of kdd cup and workshop*, vol. 2007, 35.

Bickel, Peter J, and Aiyou Chen. 2009. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences* 106(50):21068–21073.

Boyd, Stephen, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* 3(1):1–122.

Cai, Jian-Feng, Emmanuel J Candès, and Zuowei Shen. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4): 1956–1982.

Cai, T Tony, and Wen-Xin Zhou. 2013. Matrix completion via max-norm constrained optimization. *arXiv preprint arXiv:1303.0341*.

Cai, Tony, and Weidong Liu. 2011. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* 106(494):672–684.

Candès, Emmanuel J, and Yaniv Plan. 2010. Matrix completion with noise. *Proceedings of the IEEE* 98(6):925–936.

———. 2011. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on* 57(4):2342–2359.

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

Candès, Emmanuel J, Justin Romberg, and Terence Tao. 2006. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on* 52(2):489–509.

Candès, Emmanuel J, and Terence Tao. 2010. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on* 56(5): 2053–2080.

Chatterjee, Sourav. 2014. Matrix estimation by universal singular value thresholding. *The Annals of Statistics* 43(1):177–214.

Cho, Juhee, Donggyu Kim, and Karl Rohe. 2015a. Asymptotic theory for estimating the singular vectors and values of a partially-observed low rank matrix with noise.

———. 2015b. Intelligent initialization and adaptive thresholding for iterative matrix completion; some statistical and algorithmic theory for adaptive-impute.

Davenport, Mark A, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 2014. 1-bit matrix completion. *Information and Inference* 3(3):189–223.

Donoho, David L. 2006. Compressed sensing. *Information Theory, IEEE Transactions on* 52(4):1289–1306.

Fan, Jianqing, Yuan Liao, and Martina Mincheva. 2013. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(4):603–680.

Fazel, Maryam. 2002. Matrix rank minimization with applications. Ph.D. thesis, PhD thesis, Stanford University.

Green, Paul E. 2014. *Mathematical tools for applied multivariate analysis*. Academic Press.

Gross, David. 2011. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on* 57(3):1548–1566.

GroupLens. 2015. Movielens100k @MISC. `http://grouplens.org/datasets/movielens/`.

Hastie, Trevor, and Rahul Mazumder. 2015. softimpute @MISC. `https://cran.r-project.org/web/packages/softImpute/index.html`.

Hastie, Trevor, Rahul Mazumder, Jason Lee, and Reza Zadeh. 2014. Matrix completion and low-rank svd via fast alternating least squares. *arXiv preprint arXiv:1410.2596*.

Herlocker, Jonathan L, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22(1):5–53.

Hoff, Peter D, Adrian E Raftery, and Mark S Handcock. 2002. Latent space approaches to social network analysis. *Journal of the american Statistical association* 97(460):1090–1098.

Keshavan, Raghunandan, Andrea Montanari, and Sewoong Oh. 2009. Matrix completion from noisy entries. In *Advances in neural information processing systems*, 952–960.

Keshavan, Raghunandan H, Andrea Montanari, and Sewoong Oh. 2010. Matrix completion from a few entries. *Information Theory, IEEE Transactions on* 56(6): 2980–2998.

Koltchinskii, Vladimir, Karim Lounici, Alexandre B Tsybakov, et al. 2011a. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* 39(5):2302–2329.

Koltchinskii, Vladimir, et al. 2011b. Von neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics* 39(6):2936–2973.

Koren, Yehuda, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42(8):30–37.

Lattin, James M, J Douglas Carroll, and Paul E Green. 2003. *Analyzing multivariate data*. Thomson Brooks/Cole Pacific Grove, CA.

Li, Ren-Cang. 1998a. Relative perturbation theory: I. eigenvalue and singular value variations. *SIAM Journal on Matrix Analysis and Applications* 19(4):956–982.

———. 1998b. Relative perturbation theory: Ii. eigenspace and singular subspace variations. *SIAM Journal on Matrix Analysis and Applications* 20(2):471–492.

Mazumder, Rahul, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research* 11:2287–2322.

Montanari, Andrea, and Sewoong Oh. 2010. On positioning via distributed matrix completion. In *Sensor array and multichannel signal processing workshop (sam), 2010 ieee*, 197–200. IEEE.

Negahban, Sahand, and Martin J Wainwright. 2012. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research* 13(1):1665–1697.

Negahban, Sahand, Martin J Wainwright, et al. 2011. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* 39(2): 1069–1097.

Neph, Shane, Andrew B Stergachis, Alex Reynolds, Richard Sandstrom, Elhanan Borenstein, and John A Stamatoyannopoulos. 2012. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150(6):1274–1286.

Recht, Benjamin. 2011. A simpler approach to matrix completion. *The Journal of Machine Learning Research* 12:3413–3430.

Rennie, Jasson DM, and Nathan Srebro. 2005. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on machine learning*, 713–719. ACM.

Rohde, Angelika, Alexandre B Tsybakov, et al. 2011. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics* 39(2):887–930.

Shen, Haipeng, and Jianhua Z Huang. 2008. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis* 99(6):1015–1034.

Srebro, Nathan, Jason Rennie, and Tommi S Jaakkola. 2004. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, 1329–1336.

Vershynin, Roman. 2010. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vogelstein, Joshua T, William Gray Roncal, R Jacob Vogelstein, and Carey E Priebe. 2013. Graph classification using signal-subgraphs: Applications in statistical connectomics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35(7): 1539–1551.

Vu, Vincent Q, and Jing Lei. 2013. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics* 41(6):2905–2947.

Weinberger, Kilian Q, and Lawrence K Saul. 2006. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision* 70(1):77–90.

Witten, Daniela M, Robert Tibshirani, and Trevor Hastie. 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* kxp008.

Zou, Hui. 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101(476):1418–1429.