Statistica Sinica **27** (2017), 1921-1948
doi:https://doi.org/10.5705/ss.202016-0205

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

Juhee Cho, Donggyu Kim and Karl Rohe

Fred Hutchinson Cancer Research Center, Princeton University and University of Wisconsin-Madison

Abstract: Matrix completion algorithms recover a low rank matrix from a small fraction of the entries, each entry contaminated with additive errors. In practice, the singular vectors and singular values of the low rank matrix play a pivotal role for statistical analyses and inferences. This paper proposes estimators of these quantities and studies their asymptotic behavior. Under the setting where the dimensions of the matrix increase to infinity and the probability of observing each entry is identical, Theorem 1 gives the rate of convergence for the estimated singular vectors; Theorem 3 gives a multivariate central limit theorem for the estimated singular values. Even though the estimators use only a partially observed matrix, they achieve the same rates of convergence as the fully observed case. These estimators combine to form a consistent estimator of the full low rank matrix that is computed with a non-iterative algorithm. In the cases studied in this paper, this estimator achieves the minimax lower bound in Koltchinskii, Lounici and Tsybakov (2011). The numerical experiments corroborate our theoretical results.

Key words and phrases: Low rank matrices, matrix completion, matrix estimation, singular value decomposition

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 (Keshavan, Montanari and Oh (2009); Cai, Candès and Shen (2010); Mazumder, Hastie and Tibshirani (2010); Hastie et al. (2014); Rennie and Srebro (2005); Cho, Kim and Rohe (2016)) and theoretical results (Fazel (2002); Srebro, Rennie and Jaakkola (2004); Candès and Recht (2009); Candès and Plan (2010); Keshavan, Montanari and Oh (2010); Candès and Plan (2011); Gross (2011); Koltchinskii

(2011); Koltchinskii, Lounici and Tsybakov (2011); Negahban and Wainwright (2011); Recht (2011); Rohde and Tsybakov (2011); Negahban and Wainwright (2012); Cai and Zhou (2013); Chatterjee (2014); Davenport et al. (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, Bell and Volinsky (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 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 finds the convergence rate for the bias-adjusted singular vector estimators and Theorem 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 4.2 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, Lounici and Tsybakov (2011)).

The rest of this paper is organized as follows. Section 2 describes the model setup. Section 3 shows that the singular vectors and singular values of a partially observed matrix are biased and suggests a bias-adjusted alternative. Section 4.1 finds (1) the convergence rates of the estimated singular vectors and (2) the asymptotic distribution of the estimated singular values. Section 4.2 proposes and studies a one-step consistent estimator of the full matrix. Section 5 corroborates the theoretical findings with numerical experiments. The Appendix A provides the proofs of our main theoretical results. The proofs of all other results are collected in the Supplement.

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),

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

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 $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r) \in \mathbb{R}^{r\times r}$ 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

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

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

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;

$$\begin{aligned} \epsilon_{ij}|y_{ij} = 0 &= 0 \ \text{ a.s.}, \\ \epsilon_{ij}|y_{ij} = 1 &\sim \ \text{subgaussian, and} \\ y_{ij} \quad &\sim \ \text{i.i.d. Bernoulli}(p). \end{aligned}$$

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.

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.

3.1. Properties of singular values and vectors of M

Define

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

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 $M_0^T M_0$ and $M_0 M_0^T$, respectively.

Lemma 1. *Under the model setup in Section 2, we have*

$$\mathbb{E}\,\hat{\Sigma} = p^2 M_0^T M_0 + p(1-p)\, diag(M_0^T M_0) + np\sigma^2 I_d, \tag{3.1}$$

and similarly,

$$\mathbb{E}\,\hat{\Sigma}_t = p^2 M_0 M_0^T + p(1-p)\, diag(M_0 M_0^T) + dp\sigma^2 I_n, \tag{3.2}$$

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

The right-hand side of (3.1) 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 since it does not change the singular vectors of $\mathbb{E}\,\hat{\Sigma}$, the bias coming from $p(1-p)\,\text{diag}(M_0^T M_0)$ is not. The same applies to $\hat{\Sigma}_t$ in (3.2).

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,

$$\begin{aligned}\hat{\Sigma}_p :=&\hat{\Sigma} - (1-p)\operatorname{diag}(\hat{\Sigma})\\ =&(V_p, V_{pc})\operatorname{diag}(\lambda_{p_1}^2, \ldots, \lambda_{p_d}^2)(V_p, V_{pc})^T, \quad \text{and}\\ \hat{\Sigma}_{pt} :=&\hat{\Sigma}_t - (1-p)\operatorname{diag}(\hat{\Sigma}_t)\\ =&(U_p, U_{pc})\operatorname{diag}(\lambda_{pt_1}^2, \ldots, \lambda_{pt_n}^2)(U_p, U_{pc})^T,\end{aligned} \tag{3.3}$$

where

$$\begin{aligned}V_p &= (V_{p_1}, \ldots, V_{p_r}) \in \mathbb{R}^{d\times r}, \quad V_{pc} = (V_{p_{r+1}}, \ldots, V_{p_{dd}}) \in \mathbb{R}^{d\times(d-r)},\\ U_p &= (U_{p_1}, \ldots, U_{p_r}) \in \mathbb{R}^{n\times r},\ U_{pc} = (U_{p_{r+1}}, \ldots, U_{p_n}) \in \mathbb{R}^{n\times(n-r)}.\end{aligned}$$

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

Proposition 1. *Under the model setup in Section* 2, *we have by eigendecomposition,*

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

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

$$\begin{aligned}\ddot{\Lambda}_p^2 &= diag(\ddot{\lambda}_{p_1}^2, \ldots, \ddot{\lambda}_{p_d}^2)\\ &= diag(p^2[\lambda_1^2 + n\sigma^2], \ldots, p^2[\lambda_r^2 + n\sigma^2], p^2 n\sigma^2, \ldots, p^2 n\sigma^2) \in \mathbb{R}^{d\times d}, \ \textit{and}\\ \ddot{\Lambda}_{pt}^2 &= diag(p^2[\lambda_1^2 + d\sigma^2], \ldots, p^2[\lambda_r^2 + d\sigma^2], p^2 d\sigma^2, \ldots, p^2 d\sigma^2) \in \mathbb{R}^{n\times n}.\end{aligned}$$

The proof of this proposition easily follows from Lemma 1 and (3.3).

Proposition 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,

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

3.2. Estimators of singular values and vectors of M_0

The results in Proposition 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

$$\hat{p} = \frac{\sum_{k=1}^n \sum_{h=1}^d y_{kh}}{nd}. \tag{3.4}$$

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

$$\hat{\Sigma}_{\hat{p}} := \hat{\Sigma} - (1-\hat{p})\,\mathrm{diag}(\hat{\Sigma}) \quad \text{and} \quad \hat{\Sigma}_{\hat{p}t} := \hat{\Sigma}_t - (1-\hat{p})\,\mathrm{diag}(\hat{\Sigma}_t). \tag{3.5}$$

By eigendecomposition,

$$\hat{\Sigma}_{\hat{p}} = (\hat{V}, \hat{V}_c)\,\Lambda^2_{\hat{p}}\,(\hat{V}, \hat{V}_c)^T \quad \text{and} \quad \hat{\Sigma}_{\hat{p}t} = (\hat{U}, \hat{U}_c)\,\Lambda^2_{\hat{p}t}\,(\hat{U}, \hat{U}_c)^T, \tag{3.6}$$

where $\hat{V} \in \mathbb{R}^{d\times r}$, $\hat{V}_c \in \mathbb{R}^{d\times(d-r)}$, $\Lambda^2_{\hat{p}} = \mathrm{diag}(\lambda^2_{\hat{p}1}, \ldots, \lambda^2_{\hat{p}d}) \in \mathbb{R}^{d\times d}$, $\hat{U} \in \mathbb{R}^{n\times r}$, $\hat{U}_c \in \mathbb{R}^{n\times(n-r)}$, and $\Lambda^2_{\hat{p}t} = \mathrm{diag}(\lambda^2_{\hat{p}t1}, \ldots, \lambda^2_{\hat{p}tn}) \in \mathbb{R}^{n\times n}$. 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, $\lambda_i, i = 1, \ldots, r$, of M_0 by

$$\hat{\lambda}_i = \sqrt{\frac{1}{\hat{p}^2}\left(\lambda^2_{\hat{p}_i} - \hat{\tau}_{\hat{p}}\right)} \quad \text{for } i = 1, \ldots, r, \tag{3.7}$$

where $\hat{\tau}_{\hat{p}} = 1/(d-r)\mathrm{tr}(\hat{V}_c^T \hat{\Sigma}_{\hat{p}} \hat{V}_c)$.

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 $\lambda_i$ for $i = 1, \ldots, r$
Require: $M$, $y$, and $r$
  $\hat{p} \leftarrow 1/nd \sum_{k=1}^{n} \sum_{h=1}^{d} y_{kh}$
  $\hat{\Sigma}_{\hat{p}} \leftarrow M^T M - (1-\hat{p})\mathrm{diag}(M^T M)$
  $\hat{\Sigma}_{t\hat{p}} \leftarrow M M^T - (1-\hat{p})\mathrm{diag}(M M^T)$
  $\hat{V}_i \leftarrow \mathbf{v}_i(\hat{\Sigma}_{\hat{p}}), \quad \forall i \in \{1, \ldots, r\}$
  $\hat{U}_i \leftarrow \mathbf{u}_i(\hat{\Sigma}_{\hat{p}t}), \quad \forall i \in \{1, \ldots, r\}$
  $\hat{\tau}_{\hat{p}} \leftarrow 1/(d-r) \sum_{i=r+1}^{d} \boldsymbol{\lambda}_i(\hat{\Sigma}_{\hat{p}})$
  $\hat{\lambda}_i \leftarrow 1/\hat{p}\sqrt{\boldsymbol{\lambda}_i(\hat{\Sigma}_{\hat{p}}) - \hat{\tau}_{\hat{p}}}, \quad \forall i \in \{1, \ldots, r\}$
  return $\hat{V}_i$, $\hat{U}_i$, and $\hat{\lambda}_i$ for $i = 1, \ldots, r$
```

4. Asymptotic Theory

This section investigates the statistical properties of the estimators proposed in (3.6) and (3.7).

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

Let $x = (x_1, \ldots, x_n)^T$ 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^T,

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

The squared Frobenius norm is defined by $\|A\|_F^2 = \operatorname{tr}\left(A^T A\right)$, the trace of $A^T A$. 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 $P_1 P_2^\perp$ 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 $P_1 = Z_1 Z_1^T$ and $P_2 = Z_2 Z_2^T$, respectively, and $\| \sin(Z_1, \hat{Z}_2)\|_F^2 = \|Z_1 Z_1^T (Z_2 Z_2^T)^\perp\|_F^2 = 1/2\|Z_1 Z_1^T - Z_2 Z_2^T\|_F^2$. Proposition 2.2 in Vu and Lei (2013) relates this subspace distance to the Frobenius distance

$$\frac{1}{2} \inf_{\mathcal{O} \in \mathbb{V}_{r,r}} \|Z_1 - Z_2 \mathcal{O}\|_F^2 \le \|\sin(Z_1, Z_2)\|_F^2 \le \inf_{\O \in \mathbb{V}_{r,r}} \|Z_1 - Z_2 \mathcal{O}\|_F^2, \tag{4.1}$$

where $\mathbb{V}_{r,r} = \{O \in \mathbb{R}^{r \times r} : O^T O = I_r \text{ and } O O^T = I_r\}$ 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* $1/c \le b_i \le c$ *for a constant* $c > 0$*;*

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

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

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

$$\sum_{k=1}^{n}\sum_{h=1}^{d} {M_0}_{kh}^2 = \|M_0\|_F^2 \geq cnd$$

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, Liao and Mincheva (2013)). Assumption 1(1) may seem uncommon in the matrix completion literature, but consider the widely-used assumption (II.2) in Candès and Plan (2010),

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

for $i = 1, \ldots, r$ and a constant $C \geq 1$, which prevents spiky singular vectors. Under the model setup in Section 2 where the entries of M_0 are bounded in absolute value by a constant Ł > 0, this implies Assumption 1(1).

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

Theorem 1. *Under the model setup in Section* 2 *and Assumption* 1*, let* $\hat{V}^{(m)}$ *and* $\hat{U}^{(m)}$ *be the first* m *columns of* $\hat{V}$ *and* $\hat{U}$ *defined in* (3.6)*, 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*,*

$$\mathbb{E}\left\|\sin\left(\hat{V}^{(m)}, V^{(m)}\right)\right\|_F^2 \leq \frac{C_1\, n^{-1}}{p\,(b_m^2 - b_{m+1}^2)^2} \tag{4.2}$$

and

$$\mathbb{E}\left\|\sin\left(\hat{U}^{(m)}, U^{(m)}\right)\right\|_F^2 \leq \frac{C_2\, d^{-1}}{p\,(b_m^2 - b_{m+1}^2)^2}, \tag{4.3}$$

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

Remark 3. As long as $p\, d/\log n \to \infty$, the convergence rates in Theorem 1 will hold. Hence, even though p goes to zero, if $d/\log n$ diverges fast enough that $p\, d/\log n \to \infty$, we can still obtain the same results.

Remark 4. Despite the fact that $\hat{V}^{(m)}$ is built on a partially observed matrix M, Theorem 1 gives the convergence rate $n^{-1/2}/(b_m^2 - b_{m+1}^2)$ 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 (4.2) as p. A similar discussion applies to $\hat{U}^{(m)}$ in (4.3).

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

Theorem 2. *Suppose $nd^{-1} \to \infty$. Then, under the model setup in Section 2 and Assumption 1, 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

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

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

Remark 5. As long as $p\,d/\log n \to \infty$ and $pnd^{-1} \to \infty$, the asymptotic normality result in Theorem 2 will hold. Hence, even though p goes to zero, if $d/\log n$ and n/d diverge fast enough that $p\,d/\log n \to \infty$ and $pn/d \to \infty$, we can still obtain the same results.

Remark 6. Theorem 2 shows that the convergence rate of $\sum_{i=1}^{m} \hat{\lambda}_i^2$ is $\sqrt{nd}$. Considering Assumption 1(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 2 and the delta method provides a multivariate central limit theorem for $\hat{\lambda}_1, \ldots, \hat{\lambda}_r$.

Theorem 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 2 and Assumption 1, we have

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

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

$$\Upsilon_{ij} = \begin{cases} \dfrac{(1-p)}{p}\left(\displaystyle\sum_{k=1}^{n}\sum_{h=1}^{d} M_{0\,kh}^{\,2} U_{ik}^2 V_{ih}^2 - b_i^2\right) + \dfrac{\sigma^2}{p} & \text{if } i = j, \\ \dfrac{(1-p)}{p}\left(\displaystyle\sum_{k=1}^{n}\sum_{h=1}^{d} M_{0\,kh}^{\,2} U_{ik} V_{ih} U_{jk} V_{jh} - b_i b_j\right) & \text{if } i \neq j. \end{cases} \tag{4.4}$$

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

Remark 7. As long as $p\,d/\log n \to \infty$ and $pnd^{-1} \to \infty$, the asymptotic normality result in Theorem 3 will hold. Note that Theorems 2 and 3 require an additional condition, $pnd^{-1} \to \infty$, to the condition required for Theorem 1, $p\,d/\log n \to \infty$. 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 2 and 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 2 and 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 1-3 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.

4.2. A consistent estimator of M_0

Suppose that $b_i > b_{i+1}$ for all $i = 1, \ldots, r$. Theorem 1 and (4.1) 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 $s_0 = (s_{01}, \ldots, s_{0r}) \in \{-1, 1\}^r$ be

$$s_{0i} = \operatorname{sign}(\langle \hat{V}_i, V_i \rangle)\operatorname{sign}(\langle \hat{U}_i, U_i \rangle) \quad \text{for} \quad i \in \{1, \ldots, r\}. \tag{4.5}$$

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 $\hat{s} = (\hat{s}_1, \ldots, \hat{s}_r) \in \{-1, 1\}^r$ as an estimator of s_0;

$$\hat{s} = \underset{s \in \{-1,1\}^r}{\arg\min}\ \|\mathcal{P}_\Omega\big(\hat{M}(s)\big) - \mathcal{P}_\Omega\big(M\big)\|_F^2, \tag{4.6}$$

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,

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

Hence, the proposed estimator of M_0 is

$$\hat{M}(\hat{s}) = \sum_{i=1}^{r} \hat{s}_i \, \hat{\lambda}_i \hat{U}_i \hat{V}_i^T. \tag{4.7}$$

Remark 9. Finding $\hat{s}$ as in (4.6) 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;

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

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 in (4.5).

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

Assumption 2.

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

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

Remark 10. When the rank r is 1, it is more straightforward to understand Assumption 2(1). Assuming that $s_0 = 1$, it means that

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

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 4. *Under the model setup in Section* 2 *and Assumptions* 1-2*, for any given* $\eta > 0$*, there exists a constant* $C_\eta > 0$ *such that for sufficiently large* n*,*

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

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,* $\log(\log d)$*.*

Remark 11. As long as $p\, d/\log n \to \infty$, the convergence rates in Theorem 4 will hold. If we let $p = N/nd$ so that N represents the number of observed entries in the population sense, this condition implies that $N/(n \log n) \to \infty$. 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, $pnd^{-1} \to \infty$, required for Theorems 2 and 3 (see Remarks 5 and 7), is not needed for Theorems 1 and 4. It means that if p is a constant, even though $d/n \to c$ for some $0 < c \leq 1$ or $d \leq n$, the results in Theorems 1 and 4 will still hold, but the results in Theorems 2 and 3 will not.

Remark 13. Theorem 4 shows that $1/nd\|\hat{M}(\hat{s}) - M_0\|_F^2$ is bounded by $Cp^{-1}d^{-1}$ 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, Lounici and Tsybakov (2011)). 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 1(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 2. *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 $\hat{r} = \left|\{i \in \{1, \ldots, d\} \mid \lambda_{\hat{p}i}^2 \geq p^2 n\, C_d\}\right|$ where $\lambda_{\hat{p}i}^2$ is defined in* (3.6)*. Then, for any given $\delta > 0$, we have*

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

Remark 15. Empirically to find C_d and $\hat{r}$ in Lemma 2, we suggest using a scree plot of the singular values of $\hat{\Sigma}_{\hat{p}}$ in (3.5).

Remark 16. As long as C_d satisfies $\sigma^2 p^2 n < p^2 n\, C_d \leq (\sigma^2 + b_r^2 d)\, p^2 n$, consistency of $\hat{r}$ in Lemma 2 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 $\hat{\Sigma}_{\hat{p}}$.

5. Numerical Experiments

5.1. 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 2, where ϵ_{kh} are i.i.d. $\mathcal{N}(0, 1)$. The dimension n varies from 100 to 1,000 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 and 2 summarize the resulting mean squared errors calculated by $1/nd \|\hat{M}(\hat{s}) - M_0\|_F^2$, $\|\text{diag}(\hat{\lambda}_1, \hat{\lambda}_2) - \Lambda\|_F^2$, $\|\hat{V} - V\|_F^2$, 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. 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 3), but decreases with the increase of p.

We further studied the asymptotic normality of $\sum_{i=1}^2 \hat{\lambda}_i$ in Theorem 3. Figure 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

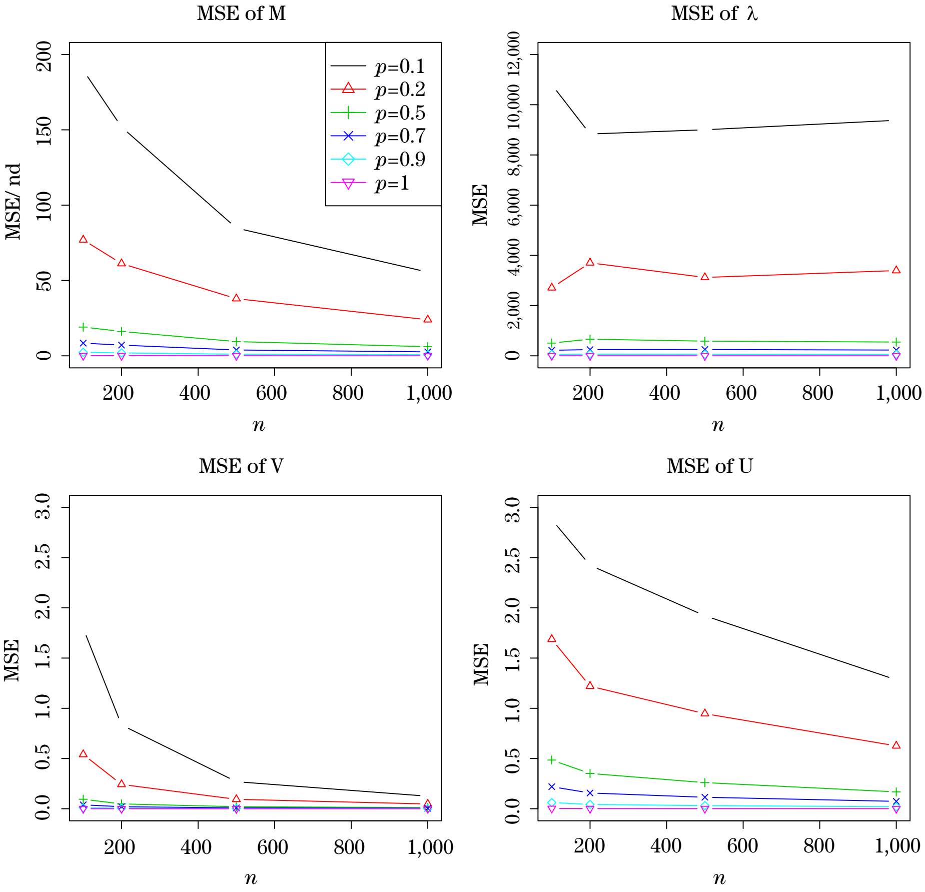

Figure 1. The mean squared errors for six different values of p when n increases. Each point on the plots correspond to an average over 500 replicates.

at 1,000 and p varies from 0.1 to 1. This shows that the asymptotic normality holds across various values of p.

5.2. 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 1,682 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 (3.5) and plotted the scree plot of the singular values of $\hat{\Sigma}_{\hat{p}}$ to determine $\hat{r}$. Figure 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

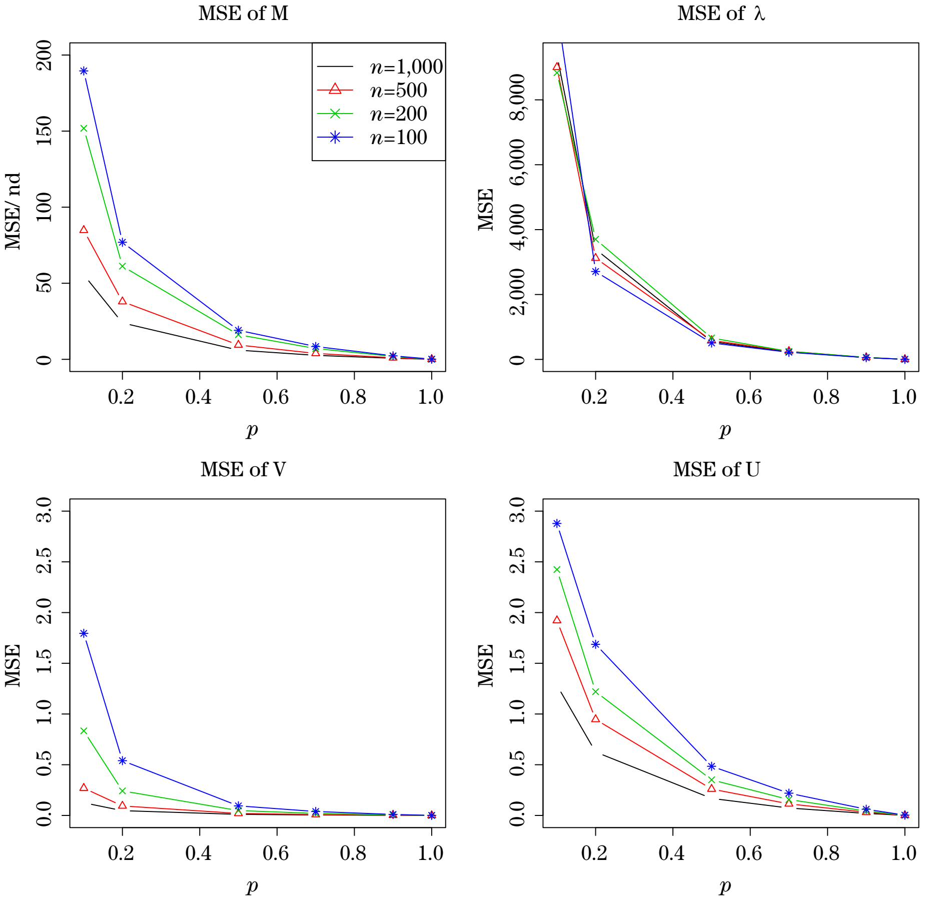

Figure 2. The same mean squared errors as the ones in Figure 1 plotted for four different values of n when p increases. Each point on the plots correspond to an average over 500 replicates.

as illustrated in Algorithms 1 and 2.

The estimated singular vectors help us understand what the main factors of movie preferences are. Table 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 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

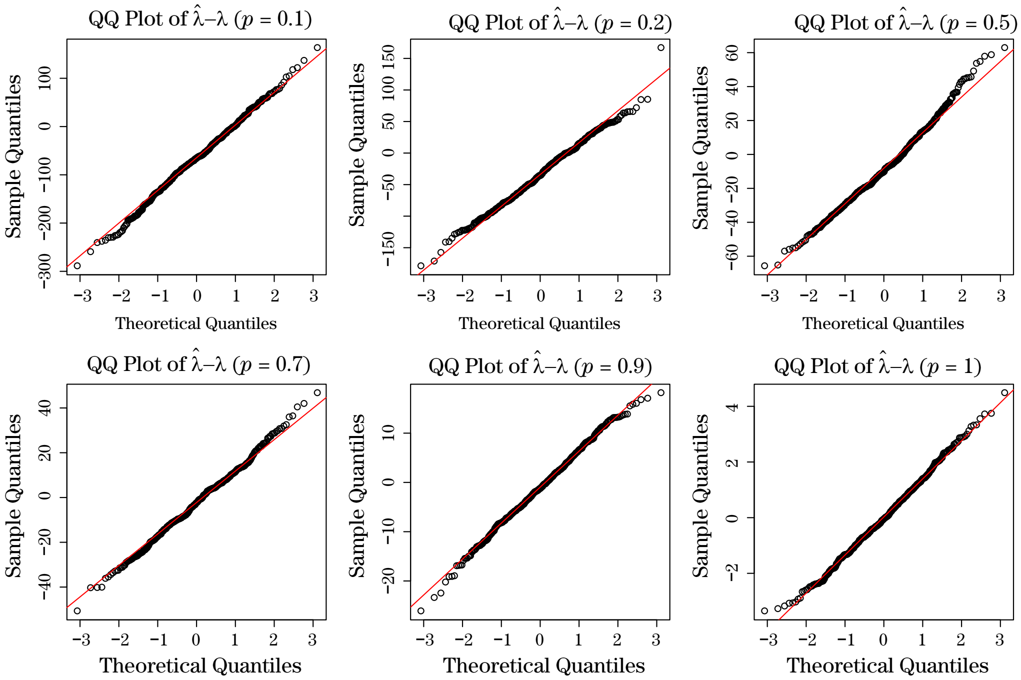

Figure 3. Asymptotic normality of $\sum_{i=1}^{2}\hat{\lambda}_i - \sum_{i=1}^{2}\lambda_i$ as p varies from 0.1 to 1. Across the plots, we fixed n to be 1,000.

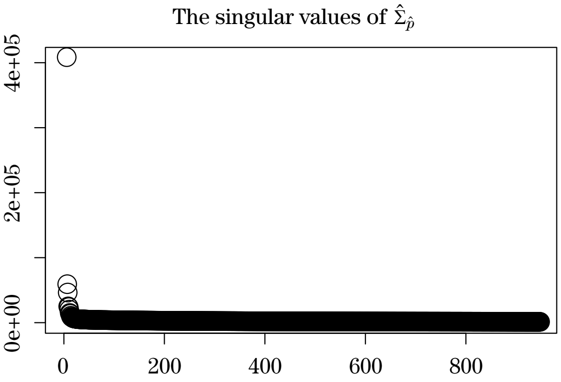

Figure 4. The singular values of $\hat{\Sigma}_{\hat{p}}$ computed by taking the MovieLens 100k data matrix as M. From this scree plot, we choose $\hat{r}$ to be 3.

of movie preferences are. Particularly, Figure 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 (4.4) in Theorem 4. Computing $\Upsilon_{ii}^{-1/2}$ requires information on the values of the parameters M_0, U, V, λ_i, p, and

Table 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

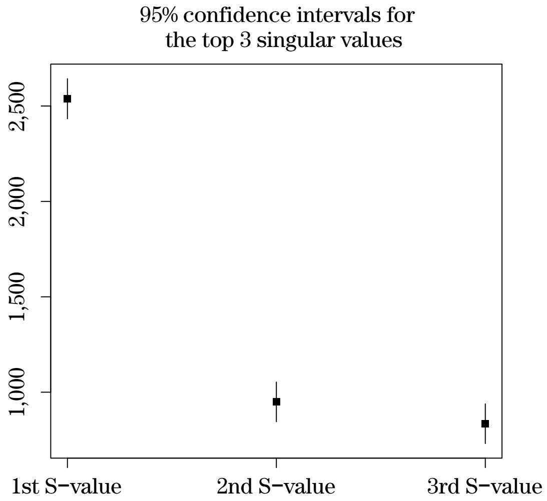

Figure 5. The 3 estimated singular values and their 95% confidence intervals.

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 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.

Supplementary Materials

The proofs for all other results than the main theoretical results are collected in the Supplement.

Acknowledgment

The authors would like to thank Anru Zhang for his helpful suggestions. The authors also appreciate the editor, associate editor and anonymous referees for careful reading of this paper and stimulating comments and suggestions. Karl Rohe's research is supported by NSF grant DMS-1309998 and ARO grant W911NF-15-1-0423.

Appendix

A.1. Proofs for Theorem 1

Proposition 2. *Under the model setup in Section* 2 *and Assumption* 1*, we have for large n and d,*

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

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

where V_p and U_p are defined in (3.3) *and C_1 and C_2 are generic constants free of $n, d,$ and p.*

Lemma 3. *Under the model setup in Section* 2 *and Assumption* 1*, for any given $\mu_1 > 0$, there exists a large constant $C_{\mu_1} > 0$ such that*

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

with probability at least $1 - O\left(n^{-\mu_1}\right)$, where $\hat{\Sigma}_p$ is defined in (3.3)*. Similarly, for any given $\mu_2 > 0$, there exists a large constant $C_{\mu_2} > 0$ such that*

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

with probability at least $1 - O\left(n^{-\mu_2}\right)$, where $\hat{\Sigma}_{pt}$ is defined in (3.3).

Lemma 4. *Under the model setup in Section* 2 *and Assumption* 1*, for any given* $\nu_1 > 0$*, there exists a large constant* $C_{\nu_1} > 0$ *such that*

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

with probability at least $1 - O\left(n^{-\nu_1}\right)$*, where* $\hat{\Sigma}_{\hat{p}}$ *and* $\hat{\Sigma}_p$ *are defined in* (3.5) *and* (3.3)*, respectively. Similarly, for any given* $\nu_2 > 0$*, there exists a large constant* $C_{\nu_2} > 0$ *such that*

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

with probability at least $1 - O\left(n^{-\nu_2}\right)$*, where* $\hat{\Sigma}_{\hat{p}t}$ *and* $\hat{\Sigma}_{pt}$ *are defined in* (3.5) *and* (3.3)*, respectively.*

Lemma 5. *Under the model setup in Section* 2 *and Assumption* 1*, we have for large* n *and* d*,*

$$\mathbb{E}\left\|\frac{1}{nd}\left(\hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p\right)V_p^{(m)}\right\|_F^2 \leq C_1 \max\left\{\frac{p^3(1-p)}{nd^3}, \frac{p^2(1-p)}{n^2 d^{5/2}}\right\} \tag{A.4}$$

and

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

where $\hat{\Sigma}_{\hat{p}}$ *and* $\hat{\Sigma}_{\hat{p}t}$ *are defined in* (3.5)*,* $\hat{\Sigma}_p$*,* $\hat{\Sigma}_{pt}$*,* V_p*, and* U_p *are defined in* (3.3)*, and* C_1 *and* C_2 *are generic constants free of* $n, d,$ *and* p*.*

Proof of Theorem 1. We only prove (4.2) because (4.3) can be proved similarly.

By triangle inequality and Proposition 2, we have

$$\begin{aligned}
&\mathbb{E}\|\sin\left(\hat{V}^{(m)}, V^{(m)}\right)\|_F^2 \\
&\leq 4\,\mathbb{E}\|\sin\left(\hat{V}^{(m)}, V_p^{(m)}\right)\|_F^2 + 4\mathbb{E}\|\sin\left(V_p^{(m)}, V^{(m)}\right)\|_F^2 \\
&\leq 4\,\mathbb{E}\|\sin\left(\hat{V}^{(m)}, V_p^{(m)}\right)\|_F^2 + \frac{C\,n^{-1}}{p\,(b_m^2 - b_{m+1}^2)^2}.
\end{aligned} \tag{A.5}$$

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

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

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

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

where $t_2 = C_2\, p^{3/2}\sqrt{\log n/nd}1/d$. Then, by Weyl's theorem (Li (1998a)), Lemma

3, and Lemma 4, we have for large constants C_1', C_1'', and C_2,

$$\mathbb{P}(E_1^c) \leq \mathbb{P}\left(\frac{1}{nd}\left\|\hat{\Sigma}_p - \mathbb{E}\hat{\Sigma}_p\right\|_2 \geq t_1\right) = O\left(n^{-4}\right) \text{ and}$$
$$\mathbb{P}(E_2^c) \leq \mathbb{P}\left(\frac{1}{nd}\left\|\hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p\right\|_2 \geq t_2\right) = O\left(n^{-4}\right).$$

Thus, for large n and d,

$$\begin{aligned}
&\mathbb{E}\|\sin\left(\hat{V}^{(m)}, V_p^{(m)}\right)\|_F^2 \\
&= \mathbb{E}\Big\{\|\sin\left(\hat{V}^{(m)}, V_p^{(m)}\right)\|_F^2 \ \mathbb{1}_{(E_1\cap E_2)^c}\Big\} + \mathbb{E}\Big\{\|\sin\left(\hat{V}^{(m)}, V_p^{(m)}\right)\|_F^2 \ \mathbb{1}_{E_1\cap E_2}\Big\} \\
&\leq m\left\{\mathbb{E}\big(\mathbb{1}_{E_2^c}\big) + \mathbb{E}\big(\mathbb{1}_{E_1^c}\big)\right\} + \mathbb{E}\left\{\frac{\left\|(1/nd)\left(\hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p\right)V_p^{(m)}\right\|_F^2}{\left((1/nd)|\lambda_{p_m}^2 - \lambda_{\hat{p}_{m+1}}^2|\right)^2}\ \mathbb{1}_{E_1\cap E_2}\right\} \\
&\leq cn^{-4} + \mathbb{E}\left\{\frac{\left\|(1/nd)\left(\hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p\right)V_p^{(m)}\right\|_F^2\ \mathbb{1}_{E_1\cap E_2}}{\left((1/nd)|\ddot{\lambda}_{p_m}^2 - \ddot{\lambda}_{p_{m+1}}^2| - t_2 - 2t_1\right)^2}\right\} \\
&\leq cn^{-4} + \mathbb{E}\left\{\frac{\left\|(1/nd)\left(\hat{\Sigma}_{\hat{p}} - \hat{\Sigma}_p\right)V_p^{(m)}\right\|_F^2}{\left((1/2nd)|\ddot{\lambda}_{p_m}^2 - \ddot{\lambda}_{p_{m+1}}^2|\right)^2}\right\} \\
&\leq cn^{-4} + \frac{C(1-p)}{(b_m^2 - b_{m+1}^2)^2}\max\left\{\frac{1}{pnd^3}, \frac{1}{p^2n^2d^{5/2}}\right\}, \qquad \text{(A.6)}
\end{aligned}$$

where $\mathbb{1}_E$ is an indicator function of an event E, the first inequality holds by the fact that $\|\sin(\hat{V}^{(m)}, V_p^{(m)})\|_F^2 \leq m$ and Davis-Kahan $\sin\theta$ theorem (Theorem 3.1 in Li (1998b)), and the last inequality is due to Lemma 5.

By (A.5) and (A.6), the result (4.2) follows.

A.2. Proofs for Theorem 2

Proposition 3. *Under the assumptions in Theorem* 2, *we have*

$$\sqrt{nd}\,\Gamma_{nd}^{-1/2}\left[\begin{pmatrix} \dfrac{1}{nd\,p^2}\displaystyle\sum_{i=1}^{m}\lambda_{p_i}^2 \\ \dfrac{p^2}{nd}\displaystyle\sum_{i=1}^{m}(\lambda_i^2 + n\sigma^2)\,\hat{p}\end{pmatrix} - \begin{pmatrix}\dfrac{1}{nd}\displaystyle\sum_{i=1}^{m}\left[\lambda_i^2 + n\sigma^2\right] \\ \dfrac{p^3}{nd}\displaystyle\sum_{i=1}^{m}(\lambda_i^2 + n\sigma^2)\end{pmatrix}\right]$$
$$\to \mathcal{N}\left(0, I_2\right) \ \textit{in distribution}, \quad \textit{as } n, d \to \infty,$$

where λ_{pi}, λ_i, *and* $\hat{p}$ *are defined in* (3.3), (2.1), *and* (3.4), *respectively, and* $\Gamma_{nd} = \Gamma_{nd}^T \in \mathbb{R}^{2\times 2}$ *consists of*

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

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

Proposition 4. *Under the model setup in Section* 2 *and Assumption* 1, *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* (3.3). *Then, we have* $\hat{\tau}_p - np^2\sigma^2 = O_p\left(p\sqrt{n}\right)$.

Proof of Theorem 2. We have

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

First, consider the term (a). We have

$$\begin{aligned}
(a) &= \frac{1}{\hat{p}^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_{\hat{p}}\hat{V}^{(m)}\right) - \sum_{i=1}^{m}\left[\lambda_i^2 + n\sigma^2\right] \\
&= \left\{\frac{1}{p^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_{p}\hat{V}^{(m)}\right) - \sum_{i=1}^{m}\left[\lambda_i^2 + n\sigma^2\right]\right\} \\
&\quad + \left\{\frac{1}{p^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_{\hat{p}}\hat{V}^{(m)}\right) - \frac{1}{p^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_{p}\hat{V}^{(m)}\right)\right\} \\
&\quad + \left\{\frac{1}{\hat{p}^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_{\hat{p}}\hat{V}^{(m)}\right) - \frac{1}{p^2}\operatorname{tr}\left(\hat{V}^{(m)T}\hat{\Sigma}_{\hat{p}}\hat{V}^{(m)}\right)\right\} \\
&= (i) + (ii) + (iii).
\end{aligned} \tag{A.7}$$

By (4.1), there is $\mathcal{O} \in \mathbb{V}_{m,m}$ such that

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

where $\mathcal{O}_i$ is the i-th column of $\mathcal{O}$. Then, the term (i) is

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

$$
\begin{aligned}
&= \frac{1}{p^2} \operatorname{tr}\left(V_p^{(m)T} \hat{\Sigma}_p V_p^{(m)} \right) - \sum_{i=1}^{m} \left[\lambda_i^2 + n\sigma^2 \right] \\
&\quad + \frac{1}{p^2} \sum_{i=1}^{m} \left(\hat{V}_i^T \hat{\Sigma}_p \hat{V}_i - \mathcal{O}_i^T {V_p}^T \hat{\Sigma}_p V_p \mathcal{O}_i \right) \\
&= \frac{1}{p^2} \sum_{i=1}^{m} \lambda_{p_i}^2 - \sum_{i=1}^{m} \left[\lambda_i^2 + n\sigma^2 \right] + O_p\left(\frac{1}{pd^2} \right), \qquad \text{(A.8)}
\end{aligned}
$$

where the last equality holds by the fact that

$$
\begin{aligned}
&\left| \sum_{i=1}^{m} \left(\hat{V}_i^T \hat{\Sigma}_p \hat{V}_i - \mathcal{O}_i^T V_p^{(m)T} \hat{\Sigma}_p V_p^{(m)} \mathcal{O}_i \right) \right| \\
&= \left| \sum_{i=1}^{m} \left[(\hat{V}_i - V_p^{(m)} \mathcal{O}_i)^T \hat{\Sigma}_p (\hat{V}_i - V_p^{(m)} \mathcal{O}_i) + 2\lambda_{pi}^2 \mathcal{O}_i^T V_p^{(m)T} \hat{V}_i - 2\lambda_{pi}^2 \right] \right| \\
&= \left| \sum_{i=1}^{m} \left[(\hat{V}_i - V_p^{(m)} \mathcal{O}_i)^T \hat{\Sigma}_p (\hat{V}_i - V_p^{(m)} \mathcal{O}_i) - \lambda_{pi}^2 \left\| \hat{V}_i - V_p^{(m)} \mathcal{O}_i \right\|_2^2 \right] \right| \\
&\le 2\lambda_{p1}^2 \sum_{i=1}^{m} \left\| \hat{V}_i - V_p^{(m)} \mathcal{O}_i \right\|_2^2 \\
&= 2\lambda_{p1}^2 \left\| \hat{V}^{(m)} - V_p^{(m)} \mathcal{O} \right\|_F^2 \\
&= O_p\left(\frac{p}{d^2} \right), \qquad \text{(A.9)}
\end{aligned}
$$

where the last equality is due to (4.1), (A.6), and (A.10) below; by the application of Weyl's theorem (Li (1998a)) and Lemma 3, we can show

$$
\lambda_{p1}^2 = O_p(p^2 nd). \qquad \text{(A.10)}
$$

The term (ii) is

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

$$= O\left(\max\left\{\frac{1}{p}, \sqrt{\frac{n}{pd}}\right\}\right), \tag{A.11}$$

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

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

The term (iii) in (A.7) is

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

$$= O_p\left(\frac{1}{p}\right) + \left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right) p^2 \sum_{i=1}^{m} \left(\lambda_i^2 + n\sigma^2\right), \tag{A.12}$$

where the third equality is due to (A.11), (A.9), Proposition 3, and the fact that

$$\sqrt{nd}\left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right) \to \mathcal{N}\left(0, \frac{4(1-p)}{p^5}\right) \text{ in distribution, as } n, d \to \infty, \tag{A.13}$$

by CLT and Delta method. From (A.8), (A.11), and (A.12), we have

$$(a) = \frac{1}{p^2}\sum_{i=1}^{m} \lambda_{p_i}^2 - \sum_{i=1}^{m}\left[\lambda_i^2 + n\sigma^2\right] + \left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right) p^2 \sum_{i=1}^{m}\left(\lambda_i^2 + n\sigma^2\right) + o_p\left(\sqrt{\frac{nd}{p}}\right). \tag{A.14}$$

Second, the term (b) is

$$\begin{aligned}
(b) &= n\sigma^2 - \frac{1}{\hat{p}^2}\hat{\tau}_{\hat{p}} \\
&= \left(n\sigma^2 - \frac{1}{p^2}\hat{\tau}_p\right) + \left(\frac{1}{p^2} - \frac{1}{\hat{p}^2}\right)\hat{\tau}_{\hat{p}} + \frac{1}{p^2}\left(\hat{\tau}_p - \hat{\tau}_{\hat{p}}\right) \\
&= O_p\left(\frac{\sqrt{n}}{p}\right) + O_p\left(\sqrt{\frac{n}{pd}}\right) + \frac{1}{p^2}\left(\hat{\tau}_p - \hat{\tau}_{\hat{p}}\right) \\
&= o_p\left(\sqrt{\frac{nd}{p}}\right),
\end{aligned} \tag{A.15}$$

where the third equality is due to Proposition 4 and (A.13), and the last equality holds by the fact that there is $\tilde{\mathcal{O}} \in \mathbb{V}_{d-r,d-r}$ by (4.1) such that

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

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

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

$$= O_p\left(\frac{p}{d^3}\right) + O_p\left(\max\left\{p,\, p^{3/2}\sqrt{\frac{n}{d}}\right\}\right),$$

where the second inequality can be derived similarly to (A.9), the second equality holds similarly to (A.11), and the last equality is due to (A.6) and (A.10).

Combining the results in (A.14) and (A.15), we have

$$\frac{1}{\sqrt{nd}}\left\{\sum_{i=1}^{m}\hat{\lambda}_i^2 - \sum_{i=1}^{m}\lambda_i^2\right\} = \frac{1}{\sqrt{nd}}\left\{(a) + m\,(b)\right\}$$

$$= \frac{1}{\sqrt{nd}}\left\{\frac{1}{p^2}\sum_{i=1}^{m}\lambda_{p_i}^2 - \sum_{i=1}^{m}\left[\lambda_i^2 + n\sigma^2\right] + \left(\frac{1}{\hat{p}^2} - \frac{1}{p^2}\right)p^2\sum_{i=1}^{m}\left(\lambda_i^2 + n\sigma^2\right)\right\} + o_p(1).$$

Thus, by Proposition 3, Delta method and Slutsky's theorem, we have

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

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

A.3. Proofs for Theorem 4

Proposition 5. *Under the model setup in Section* 2*, Assumption* 1*, and Assumption* 2(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* (4.5) *and* (4.7) *and* M_0 *is defined in* (2.1)*.*

Proof of Theorem 4. For any given $\eta > 0$, we have for a large n,

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

by Assumption 2(1). Also, for any given $\eta > 0$, we can find $C_\eta > 0$, free of n, d, and p, such that for large n,

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

by Proposition 5. Therefore, for any given $\eta > 0$, we can find $C_\eta > 0$ such that

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

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

$$+\mathbb{P}\left(\frac{p\,b_r^4}{n}\left\|\hat{M}(\hat{s}) - M_0\right\|_F^2 \ge C_\eta, s_0 \ne \hat{s}\right)$$

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

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

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

where the second inequality holds due to Assumption 2(1) and Proposition 5.

References

Anderson, T. W., Anderson, T. W., Anderson, T. W. and Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis* volume 2. Wiley New York.

Bennett, J. and Lanning, S. (2007). The netflix prize. In *Proceedings of KDD Cup and Workshop* **2007**, 35.

Cai, J.-F., Candès, E. J. and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* **20**, 1956–1982.

Cai, T. T. and Zhou, W.-X. (2013). Matrix completion via max-norm constrained optimization. *arXiv preprint arXiv:1303.0341*.

Candès, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE* **98**, 925–936.

Candès, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on* **57**, 2342–2359.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* **9**, 717–772.

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

Cho, J., Kim, D. and Rohe, K. (2016). Intelligent initialization and adaptive thresholding for iterative matrix completion; some statistical and algorithmic theory for adaptive-impute.

Davenport, M. A., Plan, Y., van den Berg, E. and Wootters, M. (2014). 1-bit matrix completion. *Information and Inference* **3**, 189–223.

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

Fazel, M. (2002). *Matrix Rank Minimization with Applications.* PhD thesis, PhD thesis, Stanford University.

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

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

Hastie, T., Mazumder, R., Lee, J. and Zadeh, R. (2014). Matrix completion and low-rank svd via fast alternating least squares. *arXiv preprint arXiv:1410.2596*.

Keshavan, R., Montanari, A. and Oh, S. (2009). Matrix completion from noisy entries. In *Advances in Neural Information Processing Systems*, pages 952–960.

Keshavan, R. H., Montanari, A. and Oh, S. (2010). Matrix completion from a few entries. *Information Theory, IEEE Transactions on* **56**, 2980–2998.

Koltchinskii, V. (2011). Von neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics* **39**, 2936–2973.

Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* **39**, 2302–2329.

Koren, Y., Bell, R. and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer* **42**, 30–37.

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

Li, R.-C. (1998b). Relative perturbation theory: Ii. eigenspace and singular subspace variations. *SIAM Journal on Matrix Analysis and Applications* **20**, 471–492.

Mazumder, R., Hastie, T. and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research* **11**, 2287–2322.

Montanari, A. and Oh, S. (2010). On positioning via distributed matrix completion. In *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2010 IEEE*, pages 197–200. IEEE.

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

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* **39**, 1069–1097.

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

Rennie, J. D. and Srebro, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine Learning*, 713–719. ACM.

Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics* **39**, 887–930.

Srebro, N., Rennie, J. and Jaakkola, T. S. (2004). Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, 1,329–1,336.

Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics* **41**, 2,905–2,947.

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

Fred Hutchinson Cancer Research Center, Public Health Sciences, 1100 Fairview Ave N, Seattle, WA 98109, USA.

E-mail: jcho23@fhcrc.org

Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ 08544, USA.

E-mail: donggyu@princeton.edu

Department of Statistics, University of Wisconsin, Madison, WI 53706, USA.

E-mail: karlrohe@stat.wisc.edu

(Received April 2016; accepted October 2016)