JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 514, 621–633, Theory and Methods
http://dx.doi.org/10.1080/01621459.2015.1021005

Taylor & Francis
Taylor & Francis Group

Structured Matrix Completion with Applications to Genomic Data Integration

Tianxi Cai, T. Tony Cai, and Anru Zhang

ABSTRACT

Matrix completion has attracted significant recent attention in many fields including statistics, applied mathematics, and electrical engineering. Current literature on matrix completion focuses primarily on independent sampling models under which the individual observed entries are sampled independently. Motivated by applications in genomic data integration, we propose a new framework of structured matrix completion (SMC) to treat structured missingness by design. Specifically, our proposed method aims at efficient matrix recovery when a subset of the rows and columns of an approximately low-rank matrix are observed. We provide theoretical justification for the proposed SMC method and derive lower bound for the estimation errors, which together establish the optimal rate of recovery over certain classes of approximately low-rank matrices. Simulation studies show that the method performs well in finite sample under a variety of configurations. The method is applied to integrate several ovarian cancer genomic studies with different extent of genomic measurements, which enables us to construct more accurate prediction rules for ovarian cancer survival. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received April 2014
Revised January 2015

KEYWORDS
Constrained minimization; Genomic data integration; Low-rank matrix; Matrix completion; Singular value decomposition; Structured matrix completion

1. Introduction

Motivated by an array of applications, matrix completion has attracted significant recent attention in different fields including statistics, applied mathematics, and electrical engineering. The central goal of matrix completion is to recover a high-dimensional low-rank matrix based on a subset of its entries. Applications include recommender systems (Koren, Bell, and Volinsky 2009), genomics (Chi et al. 2013), multi-task learning (Argyriou, Evgeniou, and Pontil 2008), sensor localization (Biswas et al. 2006; Singer and Cucuringu 2010), and computer vision (Chen and Suter 2004; Tomasi and Kanade 1992), among many others.

Matrix completion has been well studied under the uniform sampling model, where observed entries are assumed to be sampled uniformly at random. The best known approach is perhaps the constrained nuclear norm minimization (NNM), which has been shown to yield near-optimal results when the sampling distribution of the observed entries is uniform (Candès and Recht 2009; Candès and Tao 2010; Gross 2011; Recht 2011; Candes and Plan 2011). For estimating approximately low-rank matrices from uniformly sampled noisy observations, several penalized or constrained NNM estimators, which are based on the same principle as the well-known Lasso and Dantzig selector for sparse signal recovery, were proposed and analyzed (Keshavan, Montanari, and Oh 2010; Mazumder, Hastie, and Tibshirani 2010; Koltchinskii 2011; Koltchinskii et al. 2011; Rohde et al. 2011). In many applications, the entries are sampled independently but not uniformly. In such a setting, Salakhutdinov and Srebro (2010) showed that the standard NNM methods do not perform well, and proposed a weighted NNM method, which depends on the true sampling distribution. In the case of unknown sampling distribution, Foygel et al. (2011) introduced an empirically weighted NNM method. Cai and Zhou (2013) studied a max-norm constrained minimization method for the recovery of a low-rank matrix based on the noisy observations under the nonuniform sampling model. It was shown that the max-norm constrained least-square estimator is rate-optimal under the Frobenius norm loss and yields a more stable approximate recovery guarantee with respect to the sampling distributions.

The focus of matrix completion has so far been on the recovery of a low-rank matrix based on independently sampled entries. Motivated by applications in genomic data integration, we introduce in this article a new framework of matrix completion called *structured matrix completion* (SMC), where a subset of the rows and a subset of the columns of an approximately low-rank matrix are observed and the goal is to reconstruct the whole matrix based on the observed rows and columns. We first discuss the genomic data integration problem before introducing the SMC model.

1.1 Genomic Data Integration

When analyzing genome-wide studies (GWS) of association, expression profiling or methylation, ensuring adequate power

CONTACT Anru Zhang *anrzhang@wharton.upenn.edu* Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*

of the analysis is one of the most crucial goals due to the high dimensionality of the genomic markers under consideration. Because of cost constraints, GWS typically have small to moderate sample sizes and hence limited power. One approach to increase the power is to integrate information from multiple GWS of the same phenotype. However, some practical complications may hamper the feasibility of such integrative analysis. Different GWS often involve different platforms with distinct genomic coverage. For example, whole genome next generation sequencing (NGS) studies would provide mutation information on all loci while older technologies for genome-wide association studies (GWAS) would only provide information on a small subset of loci. In some settings, certain studies may provide a wider range of genomic data than others. For example, one study may provide extensive genomic measurements including gene expression, miRNA, and DNA methylation while other studies may only measure gene expression.

To perform integrative analysis of studies with different extent of genomic measurements, the naive complete observation only approach may suffer from low power. For the GWAS setting with a small fraction of loci missing, many imputation methods have been proposed in recent years to improve the power of the studies. Examples of useful methods include haplotype reconstruction, k-nearest neighbor, regression, and singular value decomposition methods (Troyanskaya et al. 2001; Kim, Golub, and Park 2005; Li and Abecasis 2006; Scheet and Stephens 2006; Wang et al. 2006; Browning and Browning 2009). Many of the haplotype phasing methods are considered to be highly effective in recovering missing genotype information (Yu and Schaid 2007). These methods, while useful, are often computationally intensive. In addition, when one study has a much denser coverage than the other, the fraction of missingness could be high and an exceedingly large number of observations would need to be imputed. It is unclear whether it is statistically or computationally feasible to extend these methods to such settings. Moreover, haplotype-based methods cannot be extended to incorporate other types of genomic data such as gene expression and miRNA data.

When integrating multiple studies with different extent of genomic measurements, the observed data can be viewed as complete rows and columns of a large matrix A and the missing components can be arranged as a submatrix of A. As such, the missingness in A is structured by design. In this article, we propose a novel SMC method for imputing the missing submatrix of A. As shown in Section 5, by imputing the missing miRNA measurements and constructing prediction rules based on the imputed data, it is possible to significantly improve the prediction performance.

1.2 Structured Matrix Completion Model

Motivated by the applications mentioned above, this article considers SMC where a subset of rows and columns are observed. Specifically, we observe $m_1 < p_1$ rows and $m_2 < p_2$ columns of a matrix $A \in \mathbb{R}^{p_1 \times p_2}$ and the goal is to recover the whole matrix. Since the singular values are invariant under row/column permutations, it can be assumed without loss of generality that we observe the first m_1 rows and m_2 columns of A, which can be written in a block form:

$$A = \begin{array}{c} \begin{matrix} m_2 & p_2 - m_2 \end{matrix} \\ \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{matrix} m_1 \\ p_1 - m_1, \end{matrix} \end{array} \tag{1}$$

where A_{11}, A_{12}, and A_{21} are observed and the goal is to recover the missing block A_{22}. See Figure 1(a) in Section 2 for a graphical display of the data. Clearly there is no way to recover A_{22} if A is an arbitrary matrix. However, in many applications such as genomic data integration discussed earlier, A is approximately low-rank, which makes it possible to recover A_{22} with accuracy. In this article, we introduce a method based on the singular value decomposition (SVD) for the recovery of A_{22} when A is approximately low rank.

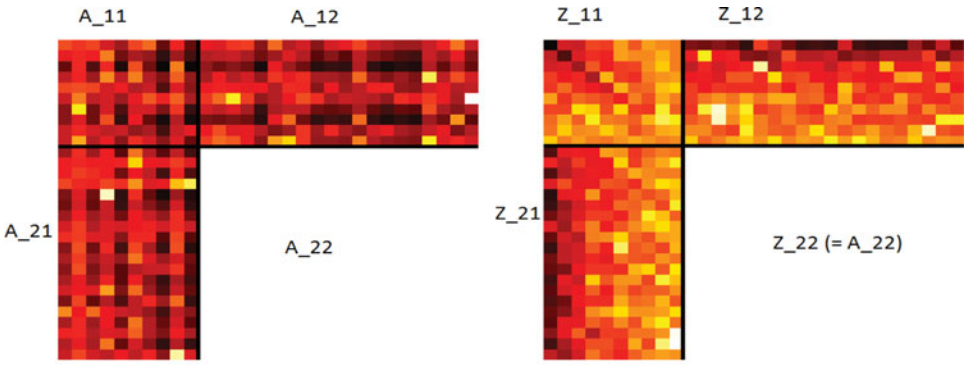
It is important to note that the observations here are much more "structured" comparing to the previous settings of matrix completion. As the observed entries are in full rows or full columns, the existing methods based on NNM are not suitable. As mentioned earlier, constrained NNM methods have been widely used in matrix completion problems based on independently observed entries. However, for the problem considered in the present article, these methods do not use the structure of the observations and do not guarantee precise recovery even for exactly low-rank matrix A (see Remark 1 in Section 2). Numerical results in Section 4 show that NNM methods do not perform well in SMC.

In this article we propose a new SMC method that can be easily implemented by a fast algorithm that only involves basic matrix operations and the SVD. The main idea of our recovery procedure is based on the Schur complement. In the ideal case when A is exactly low rank, the Schur complement of the missing block, $A_{22} - A_{21}A_{11}^{\dagger}A_{12}$, is zero and thus $A_{21}A_{11}^{\dagger}A_{12}$ can be used to recover A_{22} exactly. When A is approximately low rank, $A_{21}A_{11}^{\dagger}A_{12}$ cannot be used directly to estimate A_{22}. For this case, we transform the observed blocks using SVD; remove some unimportant rows and columns based on thresholding rules; and subsequently apply a similar procedure to recover A_{22}.

Both its theoretical and numerical properties are studied. It is shown that the estimator recovers low-rank matrices accurately and is robust against small perturbations. A lower bound result shows that the estimator is rate optimal for a class of approximately low-rank matrices. Although it is required for the theoretical analysis that there is a significant gap between the singular values of the true low-rank matrix and those of the perturbation, simulation results indicate that this gap is not really necessary in practice and the estimator recovers A accurately whenever the singular values of A decay sufficiently fast.

1.3 Organization of The Article

The rest of the article is organized as follows. In Section 2, we introduce in detail the proposed SMC methods when A is exactly or approximately low rank. The theoretical properties of the estimators are analyzed in Section 3. Both upper and lower bounds for the recovery accuracy under the Schatten-q norm loss are established. Simulation results are shown in Section 4 to investigate the numerical performance of the proposed methods. A real data application to genomic data integration



(a) heatmap of block-wise A (b) heatmap of block-wise Z after rotation

Figure 1. Illustrative example with $A \in \mathbb{R}^{30\times 30}$, $m_1 = m_2 = 10$. (A darker block corresponds to larger magnitude.)

is given in Section 5. Section 6 discusses a few practical issues related to real data applications. For reasons of space, the proofs of the main results and additional simulation results are given in the online supplement. Some key technical tools used in the proofs of the main theorems are also developed and proved in the online supplement.

2. Structured Matrix Completion: Methodology

In this section, we propose procedures to recover the submatrix A_{22} based on the observed blocks A_{11}, A_{12}, and A_{21}. We begin with basic notation and definitions that will be used in the rest of the article.

For a matrix U, we use $U_{[\Omega_1,\Omega_2]}$ to represent its submatrix with row indices Ω_1 and column indices Ω_2. We also use the Matlab syntax to represent index sets. Specifically for integers $a \le b$, "$a:b$" represents $\{a, a+1, \ldots, b\}$; and "": alone represents the entire index set. Therefore, $U_{[:,1:r]}$ stands for the first r columns of U while $U_{[(m_1+1):p_1,:]}$ stands for the $\{m_1+1, \ldots, p_1\}$th rows of U. For the matrix A given in (1), we use the notation $A_{\bullet 1}$ and $A_{1\bullet}$ to denote $[A_{11}^\intercal, A_{21}^\intercal]^\intercal$ and $[A_{11}, A_{12}]$, respectively. For a matrix $B \in \mathbb{R}^{m\times n}$, let $B = U\Sigma V^\intercal = \sum_i \sigma_i(B) u_i v_i^\intercal$ be the SVD, where $\Sigma = \mathrm{diag}\{\sigma_1(B), \sigma_2(B), \ldots\}$ with $\sigma_1(B) \ge \sigma_2(B) \ge \cdots \ge 0$ being the singular values of B in decreasing order. The smallest singular value $\sigma_{\min(m,n)}$, which will be denoted by $\sigma_{\min}(B)$, plays an important role in our analysis. We also define $B_{\max(r)} = \sum_{i=1}^{r} \sigma_i(B) u_i v_i^\intercal$ and $B_{-\max(r)} = B - B_{\max(r)} = \sum_{i\ge r+1} \sigma_i(B) u_i v_i^\intercal$. For $1 \le q \le \infty$, the Schatten-q norm $\|B\|_q$ is defined to be the vector q-norm of the singular values of B, that is, $\|B\|_q = \left(\sum_i \sigma_i^q(B)\right)^{1/q}$. Three special cases are of particular interest: when $q = 1$, $\|B\|_1 = \sum_i \sigma_i(B)$ is the nuclear (or trace) norm of B and will be denoted as $\|B\|_*$; when $q = 2$, $\|B\|_2 = \sqrt{\sum_{i,j} B_{ij}^2}$ is the Frobenius norm of B and will be denoted as $\|B\|_F$; when $q = \infty$, $\|B\|_\infty = \sigma_1(B)$ is the spectral norm of B that we simply denote as $\|B\|$. For any matrix $U \in \mathbb{R}^{p\times n}$, we use $P_U \equiv U\left(U^\intercal U\right)^\dagger U^\intercal \in \mathbb{R}^{p\times p}$ to denote the projection operator onto the column space of U. Throughout, we assume that A is *approximately rank* r in that for some integer $0 < r \le \min(m_1, m_2)$, there is a significant gap between $\sigma_r(A)$ and $\sigma_{r+1}(A)$ and the tail $\|A_{-\max(r)}\|_q = \left(\sum_{k\ge r+1} \sigma_k^q(A)\right)^{1/q}$ is small. The gap assumption enables us to provide a theoretical upper bound on the accuracy of the estimator, while it is not necessary in practice (see Section 4 for more details).

2.1 Exact Low-Rank Matrix Recovery

We begin with the relatively easy case where A is exactly of rank r. In this case, a simple analysis indicates that A can be perfectly recovered as shown in the following proposition.

Proposition 1. Suppose A is of rank r, the SVD of A_{11} is $A_{11} = U\Sigma V^\intercal$, where $U \in \mathbb{R}^{p_1\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$, and $V \in \mathbb{R}^{p_2\times r}$. If

$$\mathrm{rank}([A_{11}\ A_{12}]) = \mathrm{rank}\left(\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix}\right) = \mathrm{rank}(A) = r,$$

then $\mathrm{rank}(A_{11}) = r$ and A_{22} is exactly given by

$$A_{22} = A_{21}(A_{11})^\dagger A_{12} = A_{21} V (\Sigma)^{-1} U^\intercal A_{12}. \tag{2}$$

Remark 1. Under the same conditions as Proposition 1, the NNM

$$\hat{A}_{22} = \arg\min_B \left\| \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & B \end{bmatrix} \right\|_* \tag{3}$$

fails to guarantee the exact recovery of A_{22}. Consider the case where A is a $p_1 \times p_2$ matrix with all entries being 1. Suppose we observe arbitrary m_1 rows and m_2 columns, the NNM would yield $\hat{A}_{22} \in \mathbb{R}^{(p_1-m_1)\times(p_2-m_2)}$ with all entries being $\left(1 \wedge \sqrt{\frac{m_1 m_2}{(p_1-m_1)(p_2-m_2)}}\right)$ (see Lemma 4 in the online supplement). Hence when $m_1 m_2 < (p_1 - m_1)(p_2 - m_2)$, that is, when the size of the observed blocks are much smaller than that of A, the NNM fails to recover exactly the missing block A_{22}. See also the numerical comparison in Section 4. The NNM (3) also fails to recover A_{22} with high probability in a random matrix setting where $A = B_1 B_2^T$ with $B_1 \in \mathbb{R}^{p_1\times r}$ and $B_2 \in \mathbb{R}^{p_2\times r}$ being iid standard Gaussian matrices. See Lemma ?? in the online supplement for further details. In addition to (3), other variations of NNM have been proposed in the literature, including *penalized* NNM (Toh and Yun 2010; Mazumder, Hastie, and Tibshirani 2010),

$$\hat{A}^{\mathrm{PN}} = \arg\min_Z \left\{ \frac{1}{2} \sum_{(i_k,j_k)\in\Omega} (Z_{i_k,j_k} - A_{i_k,j_k})^2 + t\|Z\|_* \right\}; \tag{4}$$

and *constrained* NNM with relaxation (Cai, Candès, and Shen 2010),

$$\hat{A}^{\mathrm{CN}} = \arg\min_{Z} \left\{ \|Z\|_* : \ |Z_{i_k,j_k} - A_{i_k,j_k}| \leq t \text{ for } (i_k, j_k) \in \Omega \right\}, \tag{5}$$

where $\Omega = \{(i_k, j_k) : A_{i_k,j_k}$ observed, $1 \leq i_k \leq p_1, 1 \leq j_k \leq p_2\}$ and t is the tuning parameter. However, these NNM methods may not be suitable for SMC especially when only a small number of rows and columns are observed. In particular, when $m_1 \ll p_1, m_2 \ll p_2$, A is well spread in each block $A_{11}, A_{12}, A_{21}, A_{22}$, we have $\|[A_{11}A_{12}]\|_* \ll \|A\|_*$, $[A_{12}]_* \ll \|A\|_*$. Thus,

$$\left\| \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & 0 \end{bmatrix} \right\|_* \leq \left\| \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} \right\|_* + \left\| \begin{bmatrix} A_{12} \end{bmatrix} \right\|_* \ll \left\| \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right\|_*.$$

In the other words, imputing A_{22} with all zero yields a much smaller nuclear norm than imputing with the true A_{22} and hence NNM methods would generally fail to recover A_{22} under such settings.

Proposition 1 shows that, when A is exactly low-rank, A_{22} can be recovered precisely by $A_{21}(A_{11})^{\dagger}A_{12}$. Unfortunately, this result heavily relies on the exactly low-rank assumption that cannot be directly used for approximately low-rank matrices. In fact, even with a small perturbation to A, the inverse of A_{11} makes the formula $A_{21}(A_{11})^{\dagger}A_{12}$ unstable, which may lead to the failure of recovery. In practice, A is often not exactly low rank but approximately low rank. Thus, for the rest of the article, we focus on the latter setting.

2.2 Approximate Low-Rank Matrix Recovery

Let $A = U\Sigma V^{\intercal}$ be the SVD of an approximately low rank matrix A and partition $U \in \mathbb{R}^{p_1 \times p_1}$, $V \in \mathbb{R}^{p_2 \times p_2}$, and $\Sigma \in \mathbb{R}^{p_1 \times p_2}$ into blocks as

$$U = \begin{array}{c} \begin{matrix} r & p_1 - r \end{matrix} \\ \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \begin{matrix} m_1 \\ p_1 - m_1 \end{matrix} \end{array},$$

$$V = \begin{array}{c} \begin{matrix} r & p_2 - r \end{matrix} \\ \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \begin{matrix} m_2 \\ p_2 - m_2 \end{matrix} \end{array},$$

$$\sum = \begin{array}{c} \begin{matrix} r & p_2 - r \end{matrix} \\ \begin{bmatrix} \sum_1 & 0 \\ 0 & \sum_2 \end{bmatrix} \begin{matrix} r \\ p_1 - r \end{matrix} \end{array} \tag{6}$$

Then A can be decomposed as $A = A_{\max(r)} + A_{-\max(r)}$, where $A_{\max(r)}$ is of rank r with the largest r singular values of A and $A_{-\max(r)}$ is general but with small singular values. Then

$$A_{\max(r)} = U_{\bullet 1} \sum_1 V_{\bullet 1}^{\mathrm{T}} = \begin{array}{c} \begin{matrix} m_2 & p_2 - m_2 \end{matrix} \\ \begin{bmatrix} U_{11} \sum_1 V_{11}^{\mathrm{T}} & U_{11} \sum_1 V_{21}^{\mathrm{T}} \\ U_{21} \sum_1 V_{11}^{\mathrm{T}} & U_{21} \sum_1 V_{21}^{\mathrm{T}} \end{bmatrix} \begin{matrix} m_1 \\ p_1 - m_1 \end{matrix} \end{array},$$
$$\text{and} \quad A_{-\max(r)} = U_{\bullet 2} \sum\nolimits_2 V_{\bullet 2.}^{\mathrm{T}} \tag{7}$$

Here and in the sequel, we use the notation $U_{\bullet k}$ and $U_{k\bullet}$ to denote $[U_{1k}^{\intercal}, U_{2k}^{\intercal}]^{\intercal}$ and $[U_{k1}, U_{k2}]$, respectively. Thus, $A_{\max(r)}$ can be viewed as a rank-r approximation to A and obviously

$$U_{21}\Sigma_1 V_{21}^{\intercal} = \{U_{21}\Sigma_1 V_{11}^{\intercal}\}\{U_{11}\Sigma_1 V_{11}^{\intercal}\}^{-1}\{U_{11}\Sigma_1 V_{21}^{\intercal}\}.$$

We will use the observed A_{11}, A_{12}, and A_{21} to obtain estimates of $U_{\bullet 1}$, $V_{\bullet 1}$, and Σ_1 and subsequently recover A_{22} using an estimated $U_{21}\Sigma_1 V_{21}^{\intercal}$.

When r is known, that is, we know where the gap is located in the singular values of A, a simple procedure can be implemented to estimate A_{22} as described in Algorithm 1 by estimating $U_{\bullet 1}$ and $V_{\bullet 1}$ using the principal components of $A_{\bullet 1}$ and $A_{1\bullet}$.

Algorithm 1 Algorithm 1 Algorithm for Structured Matrix Completion with a given r

1: Input: $A_{11} \in \mathbb{R}^{m_1 \times m_2}$, $A_{12} \in \mathbb{R}^{(p_1 - m_1) \times m_2}$, $A_{21} \in \mathbb{R}^{m_1 \times (p_2 - m_2)}$.
2: Calculate the SVD of $A_{\bullet 1}$ and $A_{1\bullet}$ to obtain $A_{\bullet 1} = U^{(1)}\Sigma^{(1)}V^{(1)\intercal}$, $A_{1\bullet} = U^{(2)}\Sigma^{(2)}V^{(2)\intercal}$.
3: Suppose M, N are orthonormal basis of U_{11}, V_{11}. We estimate the column space of U_{11} and V_{11} by $\hat{M} = U^{(2)}_{[:,1:r]}$, $\hat{N} = V^{(1)}_{[:,1:r]}$.
4: Finally we estimate A_{22} as

$$\hat{A}_{22} = A_{21}\hat{N}(\hat{M}^{\intercal}A_{11}\hat{N})^{-1}\hat{M}^{\intercal}A_{12}. \tag{8}$$

However, Algorithm 1 has several major limitations. First, it relies on a given r, which is typically unknown in practice. Second, the algorithm need to calculate the matrix divisions, which may cause serious precision issues when the matrix is near-singular or the rank r is misspecified. To overcome these difficulties, we propose another Algorithm that essentially first estimates r with $\hat{r}$ and then apply Algorithm 1 to recover A_{22}. Before introducing the algorithm of recovery without knowing r, it is helpful to illustrate the idea with heat maps in Figures 1 and 2.

Our procedure has three steps.

1. First, we move the significant factors of $A_{\bullet 1}$ and $A_{1\bullet}$ to the front by rotating the columns of $A_{\bullet 1}$ and the rows of $A_{1\bullet}$ based on the SVD,

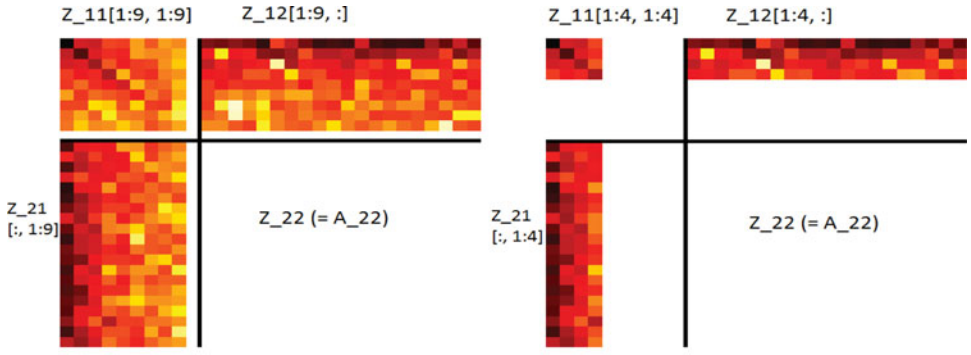
$$A_{\bullet 1} = U^{(1)}\Sigma^{(1)}V^{(1)\intercal}, \quad A_{1\bullet} = U^{(2)}\Sigma^{(2)}V^{(2)\intercal}.$$

After the transformation, we have Z_{11}, Z_{12}, Z_{21},

$$Z_{11} = U^{(2)\intercal}A_{11}V^{(1)}, \quad Z_{12} = U^{(2)\intercal}A_{12}, Z_{21} = A_{21}V^{(1)}, Z_{22} = A_{22}.$$

Clearly A and Z have the same singular values since the transformation is orthogonal. As shown in Figure 1(b), the amplitudes of the columns of $Z_{\bullet 1} = [Z_{11}^{\intercal}, Z_{21}^{\intercal}]^{\intercal}$ and the rows of $Z_{1\bullet} = [Z_{11}, Z_{12}]$ are decaying.

2. When A is exactly of rank r, the $\{r+1, \ldots, m_1\}$th rows and $\{r+1, \ldots, m_2\}$th columns of Z are zero. Due to the small perturbation term $A_{-\max(r)}$, the back columns of $Z_{\bullet 1}$ and rows of $Z_{1\bullet}$ are small but nonzero. To recover $A_{\max(r)}$, the best rank r approximation to A, a natural idea is to first delete these back rows of $Z_{1\bullet}$ and columns of $Z_{\bullet 1}$, that is, the $\{r+1, \ldots, m_1\}$th rows and $\{r+1, \ldots, m_2\}$th columns of Z.
However, since r is unknown, it is unclear how many back rows and columns should be removed. It will be



(a) Intermediate step when $\hat{r} = 9$ (b) Identify the position to truncate at $\hat{r} = 4$

Figure 2. Searching for the appropriate position to truncate from $\hat{r} = 10$ to 1.

helpful to have an estimate for r, $\hat{r}$, and then use $Z_{21,[:,1:\hat{r}]}$, $Z_{11,[1:\hat{r},1:\hat{r}]}$, and $Z_{12[1:\hat{r},:]}$ to recover A_{22}. It will be shown that a good choice of $\hat{r}$ would satisfy that $Z_{11,[1:\hat{r},1:\hat{r}]}$ is nonsingular and $\|Z_{21,[1:\hat{r},1:\hat{r}]}Z^{-1}_{11,[1:\hat{r},1:\hat{r}]}\| \le T_R$, where T_R is some constant to be specified later. Our final estimator for r would be the largest $\hat{r}$ that satisfies this condition, which can be identified recursively from $\min(m_1, m_2)$ to 1 (see Figure 2).

3. Finally, similar to (2), A_{22} can be estimated by

$$\hat{A}_{22} = Z_{21,[:,1:\hat{r}]}Z^{-1}_{11,[1:\hat{r},1:\hat{r}]}Z_{12,[1:\hat{r},:]}. \tag{9}$$

The method we propose can be summarized as the following algorithm.

Algorithm 2 Algorithm 2 Algorithm of Structured Matrix Completion with unknown r

Input: $A_{11} \in \mathbb{R}^{m_1\times m_2}$, $A_{12}^{m_1\times(p_2-m_2)}$, $A_{21}^{(p_1-m_1)\times m_2}$. Thresholding level: T_R, (or T_C).
Calculate the SVD $A_{\bullet 1} = U^{(1)}\Sigma^{(1)}V^{(1)\intercal}$, $A_{1\bullet} = U^{(2)}\Sigma^{(2)}V^{(2)\intercal}$.
Calculate $Z_{11} \in \mathbb{R}^{m_1\times m_2}$, $Z_{12} \in \mathbb{R}^{m_1\times(p_2-m_2)}$, $Z_{21} \in \mathbb{R}^{(p_1-m_1)\times m_2}$

$$Z_{11} = U^{(2)\intercal}A_{11}V^{(1)}, \quad Z_{12} = U^{(2)\intercal}A_{12}, \quad Z_{21} = A_{21}V^{(1)}.$$

for s = min(m_1, m_2) : -1: 1 **do** (Use iteration to find $\hat{r}$)
  Calculate $D_{R,s} \in \mathbb{R}^{(p_1-m_1)\times s}$ (or $D_{C,s} \in \mathbb{R}^{s\times(p_2-m_2)}$) by solving linear equation system,

$$D_{R,s} = Z_{21,[:,1:s]}Z^{-1}_{11,[1:s,1:s]} \text{ (or } D_{C,s} = Z^{-1}_{11,[1:s,1:s]}Z_{12,[1:s,:]})$$

  if $Z_{11,[1:s,1:s]}$ is not singular and $\|D_{R,s}\| \le T_R$ (or $\|D_{C,s}\| \le T_C$) **then**
    $\hat{r} = s$; **break** from the loop;
  end if
end for
if ($\hat{r}$ is not valued) **then** $\hat{r} = 0$.
end if
Finally we calculate the estimate as

$$\hat{A}_{22} = Z_{21,[:,1:\hat{r}]}Z^{-1}_{11,[1:\hat{r},1:\hat{r}]}Z_{12,[1:\hat{r},:]}$$

It can also be seen from Algorithm 2 that the estimator $\hat{r}$ is constructed based on either the row thresholding rule $\|D_{R,s}\| \le T_R$ or the column thresholding rule $\|D_{C,s}\| \le T_C$. Discussions on the choice between $D_{R,s}$ and $D_{C,s}$ are given in the next section. Let us focus for now on the row thresholding based on $D_{R,s} = Z_{21,[:,1:s]}Z^{-1}_{11,[1:s,1:s]}$. It is important to note that $Z_{21[:,1:r]}$ and $Z_{11,[1:r,1:r]}$ approximate $U_{21}\Sigma_1$ and Σ_1, respectively. The idea behind the proposed $\hat{r}$ is that when $s > r$, $Z_{21[:,1:s]}$ and $Z_{11,[1:s,1:s]}$ are nearly singular and hence $D_{R,s}$ may either be deemed singular or with unbounded norm. When $s = r$, $Z_{11,[1:s,1:s]}$ is nonsingular with $\|D_{R,s}\|$ bounded by some constant, as we show in Theorem 2. Thus, we estimate $\hat{r}$ as the largest r such that $Z_{11,[1:s,1:s]}$ is nonsingular with $\|D_{R,s}\| < T_R$.

3. Theoretical Analysis

In this section, we investigate the theoretical properties of the algorithms introduced in Section 2. Upper bounds for the estimation errors of Algorithms 1 and 2 are presented in Theorems 1 and 2, respectively, and the lower-bound results are given in Theorem 3. These bounds together establish the optimal rate of recovery over certain classes of approximately low-rank matrices. The choices of tuning parameters T_R and T_C are discussed in Corollaries 1 and 2.

Theorem 1. Suppose $\hat{A}$ is given by the procedure of Algorithm 1. Assume

$$\sigma_{r+1}(A) \le \frac{1}{2}\sigma_r(A)\cdot\sigma_{\min}(U_{11})\cdot\sigma_{\min}(V_{11}). \tag{10}$$

Then for any $1 \le q \le \infty$,

$$\left\|\hat{A}_{22} - A_{22}\right\|_q \le 3\|A_{-\max(r)}\|_q\left(1+\frac{1}{\sigma_{\min}(U_{11})}\right)\left(1+\frac{1}{\sigma_{\min}(V_{11})}\right). \tag{11}$$

Remark 2. It is helpful to explain intuitively why Condition (10) is needed. When A is approximately low-rank, the dominant low-rank component of A, $A_{\max(r)}$, serves as a good approximation to A, while the residual $A_{-\max(r)}$ is "small." The goal is to recover $A_{\max(r)}$ well. Among the three observed blocks, A_{11} is the most important and it is necessary to have $A_{\max(r)}$ dominating $A_{-\max(r)}$ in A_{11}. Note that $A_{11} = A_{\max(r),[1:m_1,1:m_2]} + A_{-\max(r),[1:m_1,1:m_2]}$,

$$\sigma_r(A_{\max(r),[1:m_1,1:m_2]}) = \sigma_r(U_{11}\Sigma_1V_{11}^{\intercal}) \ge \sigma_{\min}(U_{11})\sigma_r(A)\sigma_{\min}(V_{11}),$$

$$\|A_{-\max(r),[1:m_1,1:m_2]}\| = \|U_{12}\Sigma_2 V_{12}^{\mathsf{T}}\| \le \sigma_{r+1}(A).$$

We thus require Condition (10) in Theorem 1 for the theoretical analysis.

Theorem 1 gives an upper bound for the estimation accuracy of Algorithm 1 under the assumption that there is a significant gap between $\sigma_r(A)$ and $\sigma_{r+1}(A)$ for some known r. It is noteworthy that there are possibly multiple values of r that satisfy Condition (10). In such a case, the bound (11) applies to all such r and the largest r yields the strongest result.

We now turn to Algorithm 2, where the knowledge of r is not assumed. Theorem 2 shows that for properly chosen T_R or T_C, Algorithm 2 can lead to accurate recovery of A_{22}.

Theorem 2. Assume that there exists $r \in [1, \min(m_1, m_2)]$ such that

$$\sigma_{r+1}(A) \le \frac{1}{4}\sigma_r(A) \cdot \sigma_{\min}(U_{11})\sigma_{\min}(V_{11}). \tag{12}$$

Let T_R and T_C be two constants satisfying

$$T_R \ge \frac{1.36}{\sigma_{\min}(U_{11})} + 0.35 \quad \text{and} \quad T_C \ge \frac{1.36}{\sigma_{\min}(V_{11})} + 0.35.$$

Then for $1 \le q \le \infty$, $\hat{A}_{22}$ given by Algorithm 2 satisfies

$$\left\|\hat{A}_{22} - A_{22}\right\|_q \le 6.5T_R\left(\frac{1}{\sigma_{\min}(V_{11})} + 1\right)\|A_{-\max(r)}\|_q \tag{13}$$

$$\text{or} \quad \left\|\hat{A}_{22} - A_{22}\right\|_q \le 6.5T_C\left(\frac{1}{\sigma_{\min}(U_{11})} + 1\right)\|A_{-\max(r)}\|_q$$

when $\hat{r}$ is estimated based on the thresholding rule $\|D_{R,s}\| \le T_R$ or $\|D_{C,s}\| \le T_C$, respectively.

Besides $\sigma_r(A)$ and $\sigma_{r+1}(A)$, Theorems 1 and 2 involve $\sigma_{\min}(U_{11})$ and $\sigma_{\min}(V_{11})$, two important quantities that reflect how much the low-rank matrix $A_{\max(r)} = U_{\bullet 1}\Sigma_1 V_{\bullet 1}^{\mathsf{T}}$ is concentrated on the first m_1 rows and m_2 columns. We should note that $\sigma_{\min}(U_{11})$ and $\sigma_{\min}(V_{11})$ depend on the singular vectors of A and $\sigma_r(A)$ and $\sigma_{r+1}(A)$ are the singular values of A. The lower bound in Theorem 3 indicates that $\sigma_{\min}(U_{11})$, $\sigma_{\min}(V_{11})$, and the singular values of A together quantify the difficulty of the problem: recovery of A_{22} gets harder as $\sigma_{\min}(U_{11})$ and $\sigma_{\min}(V_{11})$ become smaller or the $\{r+1, \ldots, \min(p_1, p_2)\}$th singular values become larger. Define the class of approximately rank-r matrices $\mathcal{F}_r(M_1, M_2)$ by

$$\mathcal{F}_r(M_1, M_2) = \left\{A \in \mathbb{R}^{p_1\times p_2} : \begin{array}{l}\sigma_{\min}(U_{11}) \ge M_1, \sigma_{\min}(V_{11}) \ge M_2, \\ \sigma_{r+1}(A) \le \frac{1}{2}\sigma_r(A)\sigma_{\min}(U_{11})\sigma_{\min}(V_{11})\end{array}\right\}. \tag{14}$$

Theorem 3 (Lower Bound). Suppose $r \le \min(m_1, m_2, p_1 - m_1, p_2 - m_2)$ and $0 < M_1, M_2 < 1$, then for all $1 \le q \le \infty$,

$$\inf_{\hat{A}_{22}}\sup_{A\in\mathcal{F}_r(M_1,M_2)}\frac{\|\hat{A}_{22} - A_{22}\|_q}{\|A_{-\max(r)}\|_q} \ge \frac{1}{4}\left(\frac{1}{M_1} + 1\right)\left(\frac{1}{M_2} + 1\right). \tag{15}$$

Remark 3. Theorems 1, 2, and 3 together immediately yield the optimal rate of recovery over the class $\mathcal{F}_r(M_1M_2)$,

$$\inf_{\hat{A}_{22}}\sup_{A\in\mathcal{F}_r(M_1,M_2)}\frac{\|\hat{A}_{22} - A_{22}\|_q}{\|A_{-\max(r)}\|_q} \asymp \left(\frac{1}{M_1} + 1\right)\left(\frac{1}{M_2} + 1\right) \quad \text{for } 0 \le M_1, M_2 < 1, 1 \le q \le \infty. \tag{16}$$

Since U_{11} and V_{11} are determined by the SVD of A and $\sigma_{\min}(U_{11})$ and $\sigma_{\min}(V_{11})$ are unknown based only on A_{11}, A_{12}, and A_{21}, it is thus not straightforward to choose the tuning parameters T_R and T_C in a principled way. Theorem 2 also does not provide information on the choice between row and column thresholding. Such a choice generally depends on the problem setting. We consider below two settings where either the row/columns of A are randomly sampled or A is itself a random low-rank matrix. In such settings, when A is approximately rank r and at least $O(r\log r)$ number of rows and columns are observed, Algorithm 2 gives accurate recovery of A with fully specified tuning parameter. We first consider in Corollary 1 a fixed matrix A with the observed m_1 rows and m_2 columns selected uniformly randomly.

Corollary 1 (Random Rows/Columns). Let $A = U\Sigma V^{\mathsf{T}}$ be the SVD of $A \in \mathbb{R}^{p_1\times p_2}$. Set

$$W_r^{(1)} = \frac{p_1}{r}\max_{1\le i\le p_1}\sum_{j=1}^{r}U_{ij}^2 \quad \text{and} \quad W_r^{(2)} = \frac{p_2}{r}\max_{1\le i\le p_2}\sum_{j=1}^{r}V_{ij}^2. \tag{17}$$

Let $\Omega_1 \subset \{1, \ldots, p_1\}$ and $\Omega_2 \subset \{1, \ldots, p_2\}$ be, respectively, the index set of the observed m_1 rows and m_2 columns. Then A can be decomposed as

$$A_{11} = A_{[\Omega_1,\Omega_2]},\ A_{21} = A_{[\Omega_1^c,\Omega_2]},\quad A_{12} = A_{[\Omega_1,\Omega_2^c]},\ A_{22} = A_{[\Omega_1^c,\Omega_2^c]}. \tag{18}$$

1. Let Ω_1 and Ω_2 be independently and uniformly selected from $\{1, \ldots, p_1\}$ and $\{1, \ldots, p_2\}$ with or without replacement, respectively. Suppose there exists $r \le \min(m_1, m_2)$ such that

$$\sigma_{r+1}(A) \le \frac{1}{6}\sigma_r(A)\sqrt{\frac{m_1m_2}{p_1p_2}},$$

and the number of rows and number of columns we observed satisfy

$$m_1 \ge 12.5rW_r^{(1)}(\log(r) + c),$$
$$m_2 \ge 12.5rW_r^{(2)}(\log(r) + c), \quad \text{for some constant } c > 1.$$

Algorithm 2 with either column thresholding with the break condition $\|D_{R,s}\| \le T_R$ where $T_R = 2\sqrt{\frac{p_1}{m_1}}$ or row thresholding with the break condition $\|D_{C,s}\| \le T_C$ where $T_C = 2\sqrt{\frac{p_2}{m_2}}$ satisfies, for all $1 \le q \le \infty$,

$$\|\hat{A}_{22} - A_{22}\|_q \le 29\|A_{-\max(r)}\|_q\sqrt{\frac{p_1p_2}{m_1m_2}}$$
$$\text{with probability} \ge 1 - 4\exp(-c).$$

2. If Ω_1 is uniformly randomly selected from $\{1, \ldots, p_1\}$ with or without replacement (Ω_2 is not necessarily random), and there exists $r \le m_2$ such that

$$\sigma_{r+1}(A) \le \frac{1}{5}\sigma_r(A)\sigma_{\min}(V_{11})\sqrt{\frac{m_1}{p_1}}$$

and the number of observed rows satisfies

$$m_1 \ge 12.5 r W_r^{(1)}\left(\log(r) + c\right) \quad \text{for some constant } c > 1, \tag{19}$$

then Algorithm 2 with the break condition $\|D_{R,s}\| \le T_R$, where $T_R \ge 2\sqrt{\frac{p_1}{m_1}}$ satisfies, for all $1 \le q \le \infty$,

$$\left\|\hat{A}_{22} - A_{22}\right\|_q \le 6.5\|A_{-\max(r)}\|_q T_R\left(\frac{1}{\sigma_{\min}(V_{11})} + 1\right)$$

with probability $\ge 1 - 2\exp(-c)$.

3. Similarly, if Ω_2 is uniformly randomly selected from $\{1, \ldots, p_2\}$ with or without replacement (Ω_1 is not necessarily random) and there exists $r \le m_2$ such that

$$\sigma_{r+1}(A) \le \frac{1}{5}\sigma_r(A)\sigma_{\min}(U_{11})\sqrt{\frac{m_2}{p_2}},$$

and the number of observed columns satisfies

$$m_2 \ge 12.5 r W_r^{(2)}\left(\log(r) + c\right) \quad \text{for some constant } c > 1, \tag{20}$$

then Algorithm 2 with the break condition $\|D_{C,s}\| \le T_C$, where $T_C \ge 2\sqrt{\frac{p_2}{m_2}}$ satisfies, for all $1 \le q \le \infty$,

$$\left\|\hat{A}_{22} - A_{22}\right\|_q \le 6.5\|A_{-\max(r)}\|_q T_C\left(\frac{1}{\sigma_{\min}(U_{11})} + 1\right)$$

with probability $\ge 1 - 2\exp(-c)$.

Remark 4. The quantities $W_r^{(1)}$ and $W_r^{(2)}$ in Corollary 1 measure the variation of amplitude of each row or each column of $A_{\max(r)}$. When $W_r^{(1)}$ and $W_r^{(2)}$ become larger, a small number of rows and columns in $A_{\max(r)}$ would have larger amplitude than others, while these rows and columns would be missed with large probability in the sampling of Ω, which means the problem would become harder. Hence, more observations for the matrix with larger $W_r^{(1)}$ and $W_r^{(2)}$ are needed as shown in (19).

We now consider the case where A is a random matrix.

Corollary 2 (Random Matrix). Suppose $A \in \mathbb{R}^{p_1 \times p_2}$ is a random matrix generated by $A = U\Sigma V^\intercal$, where the singular values Σ and singular space V are fixed, and U has orthonormal columns that are randomly sampled based on the Haar measure. Suppose we observe the first m_1 rows and first m_2 columns of A. Assume there exists $r < \frac{1}{2}\min(m_1, m_2)$ such that

$$\sigma_{r+1}(A) \le \frac{1}{5}\sigma_r(A)\sigma_{\min}(V_{11})\sqrt{\frac{m_1}{p_1}}.$$

Then there exist uniform constants $c, \delta > 0$ such that if $m_1 \ge cr$, $\hat{A}_{22}$ is given by Algorithm 2 with the break condition $\|D_{R,s}\| \le T_R$, where $T_R \ge 2\sqrt{\frac{p_1}{m_1}}$, we have for all $1 \le q \le \infty$,

$$\left\|\hat{A}_{22} - A_{22}\right\|_q \le 6.5\|A_{-\max(r)}\|_q T_R\left(\frac{1}{\sigma_{\min}(V_{11})} + 1\right)$$

with probability at least $1 - e^{-\delta m_1}$.

Parallel results hold for the case when U is fixed and V has orthonormal columns that are randomly sampled based on the Haar measure, and we observe the first m_1 rows and first m_2 columns of A. Assume there exists $r < \frac{1}{2}\min(m_1, m_2)$ such that

$$\sigma_{r+1}(A) \le \frac{1}{5}\sigma_r(A)\sigma_{\min}(U_{11})\sqrt{\frac{m_2}{p_2}}.$$

Then there exist uniform constants $c, \delta > 0$ such that if $m_2 \ge cr$, $\hat{A}_{22}$ is given by Algorithm 2 with column thresholding with the break condition $\|D_{C,s}\| \le T_C$, where $T_C \ge 2\sqrt{\frac{p_2}{m_2}}$, we have for all $1 \le q \le \infty$,

$$\left\|\hat{A}_{22} - A_{22}\right\|_q \le 6.5\|A_{-\max(r)}\|_q T_C\left(\frac{1}{\sigma_{\min}(U_{11})} + 1\right)$$

with probability at least $1 - e^{-\delta m_2}$.

4. Simulation

In this section, we show results from extensive simulation studies that examine the numerical performance of Algorithm 2 on randomly generated matrices for various values of p_1, p_2, m_1, and m_2. We first consider settings where a gap between some adjacent singular values exists, as required by our theoretical analysis.

Then we investigate settings where the singular values decay smoothly with no significant gap between adjacent singular values. The results show that the proposed procedure performs well even when there is no significant gap, as long as the singular values decay at a reasonable rate.

We also examine how sensitive the proposed estimators are to the choice of the threshold and the choice between row and column thresholding. In addition, we compare the performance of the SMC method with that of the NNM method. Finally, we consider a setting similar to the real data application discussed in the next section. Results shown below are based on 200–500 replications for each configuration. Additional simulation results on the effect of m_1, m_2, and ratio p_1/m_1 are provided in the supplement. Throughout, we generate the random matrix A from $A = U\Sigma V$, where the singular values of the diagonal matrix Σ are chosen accordingly for different settings. The singular spaces U and V are drawn randomly from the Haar measure. Specifically, we generate iid standard Gaussian matrix $\tilde{U} \in \mathbb{R}^{p_1 \times \min(p_1, p_2)}$ and $\tilde{V} \in \mathbb{R}^{p_2 \times \min(p_1, p_2)}$, then apply the QR decomposition to $\tilde{U}$ and $\tilde{V}$ and assign U and V with the Q part of the result.

We first consider the performance of Algorithm 2 when a significant gap between the rth and $(r+1)$th singular values of A. We fixed $p_1 = p_2 = 1000$, $m_1 = m_2 = 50$ and choose the singular values as

$$\{\underbrace{1, \ldots, 1}_{r}, g^{-1}1^{-1}, g^{-1}2^{-1}, \ldots\},$$
$$g = 1, 2, \ldots, 10, \quad r = 4, 12, \text{ and } 20. \tag{21}$$

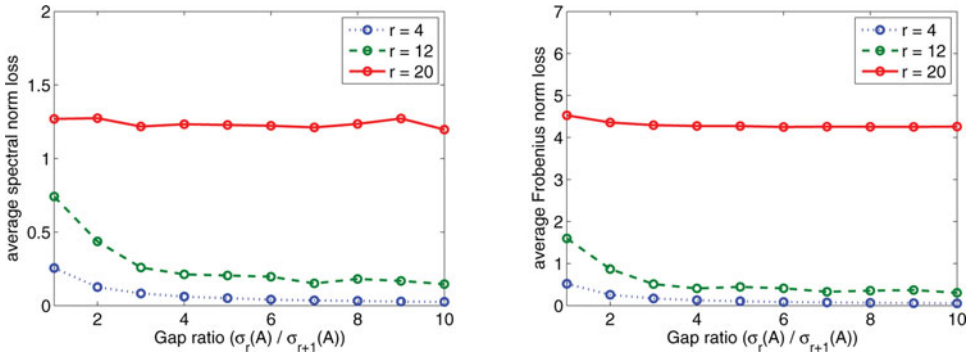


Figure 3. Spectral norm loss (left panel) and Frobenius norm loss (right panel) when there is a gap between $\sigma_r(A)$ and $\sigma_{r+1}(A)$. The singular values of A are given by (21), $p_1 = p_2 = 1000$, and $m_1 = m_2 = 50$.

Here r is the rank of the major low-rank part $A_{\max(r)}$, $g = \frac{\sigma_r(A)}{\sigma_{r+1}(A)}$ is the gap ratio between the rth and $(r+1)$th singular values of A. The average loss of $\hat{A}_{22}$ from Algorithm 2 with the row thresholding and $T_R = 2\sqrt{p_1/m_1}$ under both the spectral norm and Frobenius norm losses are given in Figure 3. The results suggest that our algorithm performs better when r gets smaller and gap ratio $g = \sigma_r(A)/\sigma_{r+1}(A)$ gets larger. Moreover, even when $g = 1$, namely, there is no significant gap between any adjacent singular values, our algorithm still works well for small r. As will be seen in the following simulation studies, this is generally the case as long as the singular values of A decay sufficiently fast.

We now turn to the settings with the singular values being $\{j^{-\alpha},\ j = 1, 2, \ldots, \min(p_1, p_2)\}$ and various choices of α, p_1, and p_2. Hence, no significant gap between adjacent singular values exists under these settings and we aim to demonstrate that our method continues to work well. We first consider $p_1 = p_2 = 1000$, $m_1 = m_2 = 50$ and let α range from 0.3 to 2. Under this setting, we also study how the choice of thresholds affects the performance of our algorithm. For simplicity, we report results only for row thresholding as results for column thresholding are similar. The average loss of $\hat{A}_{22}$ from Algorithm 2 with $T_R \in \{c\sqrt{m_1/p_1},\ c \in [1, 6]\}$ under both the spectral norm and Frobenius norm are given in Figure 4. In general, the algorithm performs well provided that α is not too small and as expected, the average loss decreases with a higher decay rate in the singular values. This indicates that the existence of a significant gap between adjacent singular values is not necessary in practice, provided that the singular values decay sufficiently fast. When comparing the results across different choices of the threshold, $c = 2$ as suggested in our theoretical analysis is indeed the optimal choice. Thus, in all subsequent numerical analysis, we fix $c = 2$.

To investigate the impact of row versus column thresholding, we let the singular value decay rate be $\alpha = 1$, $p_1 = 300$, $p_2 = 3000$, and m_1 and m_2 varying from 10 to 150. The original matrix A is generated the same way as before. We apply row and column thresholding with $T_R = 2\sqrt{p_1/m_1}$ and $T_C = 2\sqrt{p_2/m_2}$. It can be seen from Figure 5 that when the observed rows and columns are selected randomly, the results are not sensitive to the choice between row and column thresholding.

We next turn to the comparison between our proposed SMC algorithm and the penalized NNM method, which recovers A

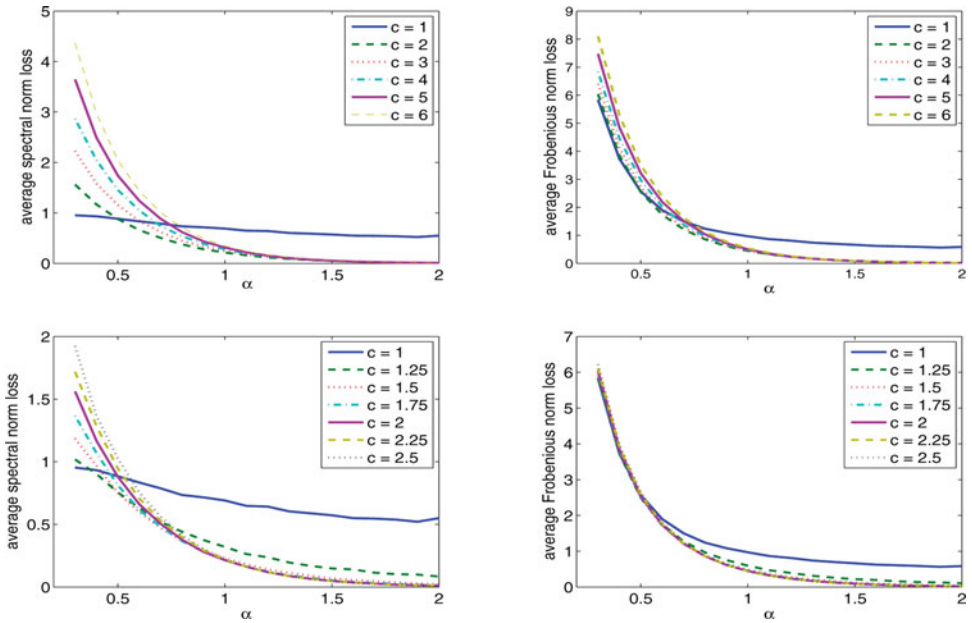
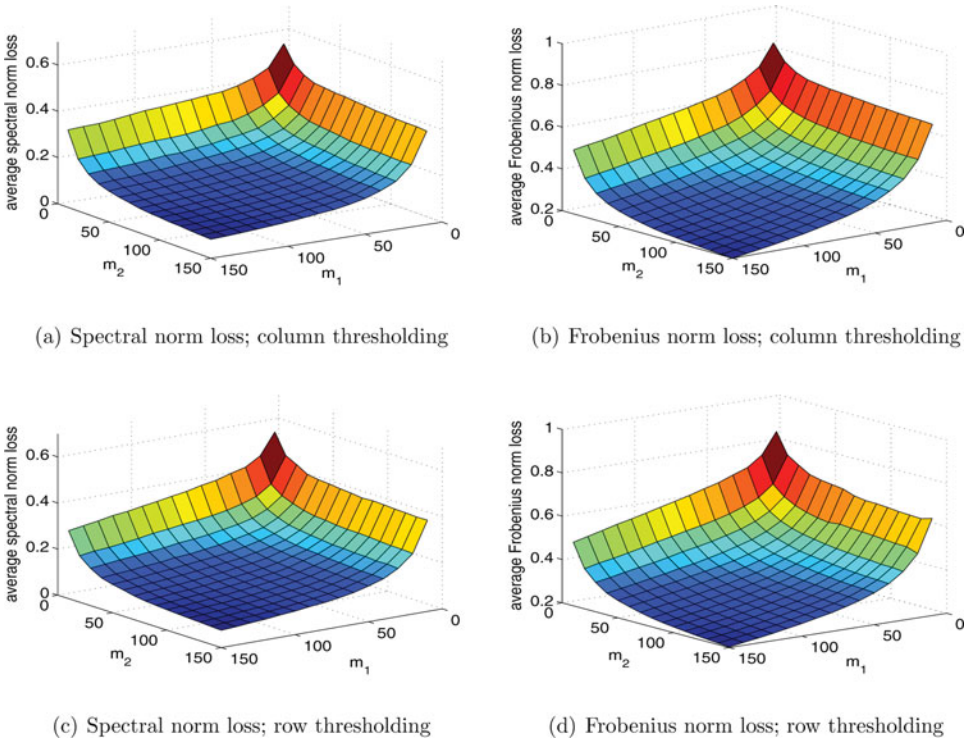


Figure 4. Spectral norm loss (left panel) and Frobenius norm loss (right panel) as the thresholding constant c varies. The singular values of A are $\{j^{-\alpha},\ j = 1, 2, \ldots\}$ with α varying from 0.3 to 2, $p_1 = p_2 = 1000$, and $m_1 = m_2 = 50$.



(a) Spectral norm loss; column thresholding (b) Frobenius norm loss; column thresholding

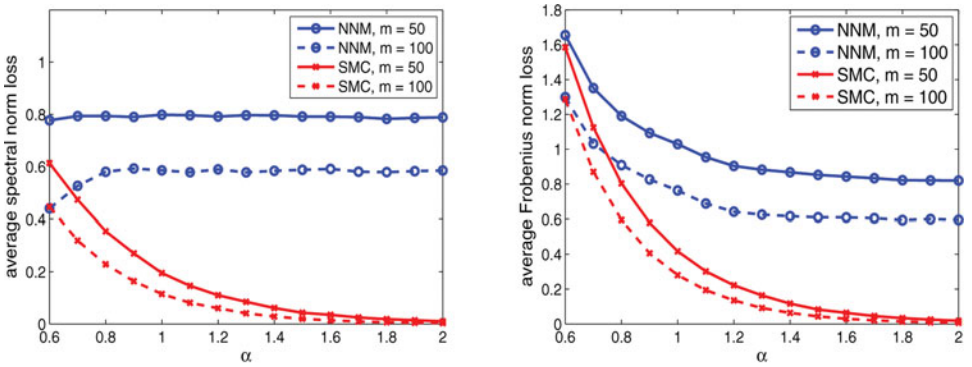
(c) Spectral norm loss; row thresholding (d) Frobenius norm loss; row thresholding

Figure 5. Spectral and Frobenius norm losses with column/row thresholding. The singular values of A are $\{j^{-1}, j = 1, 2, \ldots\}$, $p_1 = 300$, $p_2 = 3000$, and $m_1, m_2 = 10, \ldots, 150$.

by (4). The solution to (4) can be solved by the spectral regularization algorithm by Mazumder, Hastie, and Tibshirani (2010) or the accelerated proximal gradient algorithm by Toh and Yun (2010), where these two methods provide similar results. We use five-fold cross-validation to select the tuning parameter t. Details on the implementation can be found in the online supplement.

We consider the setting where $p_1 = p_2 = 500$, $m_1 = m_2 = 50, 100$ and the singular value decay rate α ranges from 0.6 to 2. As shown in Figure 6, the proposed SMC method substantially outperform the penalized NNM method with respect to both the spectral and Frobenius norm loss, especially as α increases.

Finally, we consider a simulation setting that mimics the ovarian cancer data application considered in the next section, where $p_1 = 1148$, $p_2 = 1225$, $m_1 = 230$, $m_2 = 426$ and the singular values of A decay at a polynomial rate α. Although the singular values of the full matrix are unknown, we estimate the decay rate based on the singular values of the fully observed 552 rows of the matrix from the TCGA study, denoted by $\{\sigma_j, j = 1, \ldots, 522\}$. A simple linear regression of $\{\log(\sigma_j), j = 1, \ldots, 522\}$ on $\{\log(j), j = 1, \ldots, 522\}$ estimates α as 0.8777. In the simulation, we randomly generate $A \in \mathbb{R}^{p_1 \times p_2}$ such that the singular values are fixed as $\{j^{-0.8777}, j = 1, 2, \ldots\}$. For comparison, we also obtained results for $\alpha = 1$ as well as those based on the penalized NNM method with five-cross-validation. As shown in Table 1, the relative spectral norm loss and relative Frobenius norm loss of the proposed method are reasonably small and substantially smaller than those from the penalized NNM method.



(a) Spectral norm loss (b) Frobenius norm loss

Figure 6. Comparison of the proposed SMC method with the NNM method with five-cross-validation for the settings with singular values of A being $\{j^{-\alpha}, j = 1, 2, \ldots\}$ for α ranging from 0.6 to 2, $p_1 = p_2 = 500$, and $m_1 = m_2 = 50$ or 100.

Table 1. Relative spectral norm loss ($\|\hat{A}_{22} - A_{22}\|/\|A_{22}\|$) and Frobenius norm loss ($\|\hat{A}_{22} - A_{22}\|_F/\|A_{22}\|_F$) for $p_1 = 1148$, $p_2 = 1225$, $m_1 = 230$, $m_2 = 426$ and singular values of A being $\{j^{-\alpha} : j = 1, 2, \ldots\}$.

	Relative spectral norm loss		Relative Frobenius norm loss	
	SMC	NNM	SMC	NNM
$\alpha = 0.8777$	0.1253	0.4614	0.2879	0.6122
$\alpha = 1$	0.0732	0.4543	0.1794	0.5671

5. Application in Genomic Data Integration

In this section, we apply our proposed procedures to integrate multiple genomic studies of ovarian cancer (OC). OC is the fifth leading cause of cancer mortality among women, attributing to 14,000 deaths annually (Siegel, Naishadham, and Jemal 2013). OC is a relatively heterogenous disease with 5 year survival rate varying substantially among different subgroups. The overall 5 year survival rate is near 90% for stage I cancer. But the majority of the OC patients are diagnosed as stage III/IV diseases and tend to develop resistance to chemotherapy, resulting a 5 year survival rate only about 30% (Holschneider and Berek 2000). On the other hand, a small minority of advanced cancers are sensitive to chemotherapy and do not relapse after treatment completion. Such a heterogeneity in disease progression is likely to be in part attributable to variations in underlying biological characteristics of OC (Berchuck et al. 2005). This heterogeneity and the lack of successful treatment strategies motivated multiple genomic studies of OC to identify molecular signatures that can distinguish OC subtypes, and in turn help to optimize and personalize treatment. For example, the Cancer Genome Atlas (TCGA) comprehensively measured genomic and epigenetic abnormalities on high grade OC samples (Cancer Genome Atlas Research Network 2011). A gene expression risk score based on 193 genes, $\mathcal{G}$, was trained on 230 training samples, denoted by TCGA$^{(t)}$, and shown as highly predictive of OC survival when validated on the TCGA independent validation set of size 322, denoted by TCGA$^{(v)}$, as well as on several independent OC gene expression studies including those from Bonome et al. (2005) (BONO), Dressman et al. (2007) (DRES), and Tothill et al. (2008) (TOTH).

The TCGA study also showed that clustering of miRNA levels overlaps with gene-expression-based clusters and is predictive of survival. It would be interesting to examine whether combining miRNA with $\mathcal{G}$ could improve survival prediction when compared to $\mathcal{G}$ alone. One may use TCGA$^{(v)}$ to evaluate the added value of miRNA. However, TCGA$^{(v)}$ is of limited sample size. Furthermore, since miRNA was only measured for the TCGA study, its utility in prediction cannot be directly validated using these independent studies. Here, we apply our proposed SMC method to impute the missing miRNA values and subsequently construct prediction rules based on both $\mathcal{G}$ and the imputed miRNA, denoted by $\widehat{\text{miRNA}}$, for these independent validation sets. To facilitate the comparison with the analysis based on TCGA$^{(v)}$ alone where miRNA measurements are observed, we only used the miRNA from TCGA$^{(t)}$ for imputation and reserved the miRNA data from TCGA$^{(v)}$ for validation purposes. To improve the imputation, we also included additional 300 genes that were previously used in a prognostic gene expression signature for predicting ovarian cancer survival (Denkert et al. 2009). This results in a total of $m_1 = 426$ unique gene expression variables available for imputation. Detailed information on the data used for imputation is shown in Figure 7. Prior to imputation, all gene expression and miRNA levels are log transformed and centered to have mean zero within each study to remove potential platform or batch effects. Since the observable rows (indexing subjects) can be viewed as random whereas the observable columns (indexing genes and miRNAs) are not random, we used row thresholding with threshold $T_R = 2\sqrt{p_1/m_1}$ as suggested in the theoretical and simulation results. For comparison, we also imputed data using the penalized NNM method with tuning parameter t selected via five-fold cross-validation.

We first compared $\widehat{\text{miRNA}}$ to the observed miRNA on TCGA$^{(v)}$. Our imputation yielded a rank 2 matrix for $\widehat{\text{miRNA}}$ and the correlations between the two right and left singular vectors $\widehat{\text{miRNA}}$ to that of the observed miRNA variables are 0.90, 0.71, 0.34, 0.14, substantially higher than that of those from the NNM method, with the corresponding values 0.45, 0.06, 0.10, 0.05. This suggests that the SMC imputation does a good job in recovering the leading projections of the miRNA measurements and outperforms the NNM method.

To evaluate the utility of $\widehat{\text{miRNA}}$ for predicting OC survival, we used the TCGA$^{(t)}$ to select 117 miRNA markers that are marginally associated with survival with a nominal p-value threshold of 0.05. We use the two leading principal components (PCs) of the 117 miRNA markers, $\text{miRNA}^{\text{PC}} = (\text{miRNA}_1^{\text{PC}}, \text{miRNA}_2^{\text{PC}})^T$, as predictors for the survival outcome in addition to $\mathcal{G}$. The imputation enables us to integrate information from four studies including TCGA$^{(t)}$, which could substantially improve efficiency and prediction performance. We first assessed the association between $\{\text{miRNA}^{\text{PC}}, \mathcal{G}\}$ and OC survival by fitting a stratified Cox model (Kalbfleisch and Prentice 2011) to the integrated data that combines TCGA$^{(v)}$ and the three additional studies via either the SMC or NNM methods. In addition, we fit the Cox model to (i) TCGA$^{(v)}$ set alone with miRNA^{PC} obtained from the observed miRNA; and (ii) each individual study separately with imputed miRNA^{PC}. As shown in Table 2(a), the log hazard ratio (logHR) estimates

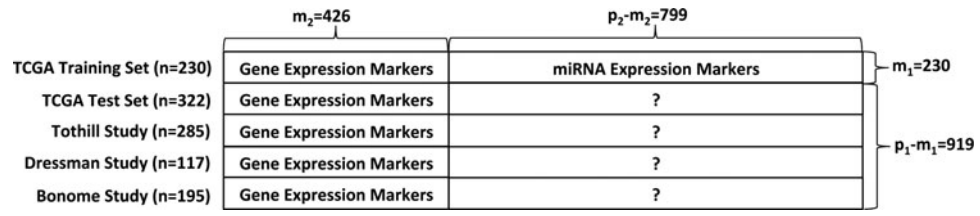


Figure 7. Imputation scheme for integrating multiple OC genomic studies.

Table 2. Shown in (a) are the estimates of the log hazard ratio (logHR) along with their corresponding standard errors (SE) and *p*-values by fitting stratified Cox model integrating information from four independent studies with imputed miRNA based on the SMC method and the nuclear norm minimization (NNM); and Cox model to the TCGA test data with original observed miRNA (Ori.). Shown also are the estimates for each individual studies by fitting separate Cox models with imputed miRNA.

(a) Integrated analysis with imputed miRNA versus single study with observed miRNA

	logHR			SE			*p*-Value		
	Ori.	SMC	NNM	Ori.	SMC	NNM	Ori.	SMC	NNM
$\mathcal{G}$	0.067	0.143	0.168	0.041	0.034	0.028	0.104	0.000	
$\text{miRNA}_1^{\text{PC}}$	− 0.012	− 0.019	− 0.013	0.009	0.006	0.012	0.218	0.001	0.283
$\text{miRNA}_2^{\text{PC}}$	0.023	0.018	− 0.005	0.014	0.009	0.014	0.092	0.039	0.725

(b) Estimates for individual studies with imputed miRNA from the SMC method

	logHR				SE				*p*-Value			
	TCGA	TOTH	DRES	BONO	TCGA	TOTH	DRES	BONO	TCGA	TOTH	DRES	BONO
$\mathcal{G}$	0.051	0.377	0.174	0.311	0.048	0.069	0.132	0.117	0.286	0.000	0.187	0.008
$\text{miRNA}_1^{\text{PC}}$	− 0.014	− 0.021	− 0.031	− 0.010	0.011	0.012	0.014	0.014	0.207	0.082	0.030	0.484
$\text{miRNA}_2^{\text{PC}}$	0.014	0.045	− 0.021	0.036	0.016	0.018	0.022	0.019	0.391	0.009	0.336	0.054

(c) Estimates for individual studies with imputed miRNA from the NNM method

	logHR				SE				*p*-Value			
	TCGA	TOTH	DRES	BONO	TCGA	TOTH	DRES	BONO	TCGA	TOTH	DRES	BONO
$\mathcal{G}$	0.082	0.405	0.361	0.258	0.037	0.066	0.114	0.088	0.028	0.000	0.002	0.003
$\text{miRNA}_1^{\text{PC}}$	− 0.045	0.016	0.055	− 0.008	0.021	0.026	0.031	0.023	0.034	0.544	0.076	0.721
$\text{miRNA}_2^{\text{PC}}$	0.008	− 0.086	− 0.043	0.019	0.026	0.027	0.034	0.029	0.758	0.002	0.201	0.496

for miRNA^{PC} from the integrated analysis, based on both SMC and NNM methods, are similar in magnitude to those obtained based on the observed miRNA values with $\text{TCGA}^{(v)}$. However, the integrated analysis has substantially smaller standard error (SE) estimates due the increased sample sizes. The estimated logHRs are also reasonably consistent across studies when separate models were fit to individual studies.

We also compared the prediction performance of the model based on $\mathcal{G}$ alone to the model that includes both $\mathcal{G}$ and the imputed miRNA^{PC}. Combining information from all four studies via standard meta-analysis, the average improvement in *C*-statistic was 0.032 (SE = 0.013) for the SMC method and 0.001 (SE = 0.009) for the NNM method, suggesting that the imputed miRNA^{PC} from the SMC method has much higher predictive value compared to those obtained from the NNM method.

In summary, the results shown above suggest that our SMC procedure accurately recovers the leading PCs of the miRNA variables. In addition, adding miRNA^{PC} obtained from imputation using the proposed SMC method could significantly improve the prediction performance, which confirms the value of our method for integrative genomic analysis. When comparing to the NNM method, the proposed SMC method produces summaries of miRNA that is more correlated with the truth and yields leading PCs that are more predictive of OC survival.

6. Discussions

The present article introduced a new framework of SMC where a subset of the rows and columns of an approximately low-rank matrix are observed. We proposed an SMC method for the recovery of the whole matrix with theoretical guarantees. The proposed procedure significantly outperforms the conventional NNM method for matrix completion, which does not take into account the special structure of the observations. As shown by our theoretical and numerical analyses, the widely adopted NNM methods for matrix completion are not suitable for the SMC setting. These NNM methods perform particularly poorly when a small number of rows and columns are observed.

The key assumption in matrix completion is the matrix being approximately low rank. This is reasonable in the ovarian cancer application since as indicated in the results from the TCGA study (Cancer Genome Atlas Research Network 2011), the patterns observed in the miRNA signature are highly correlated with the patterns observed in the gene expression signature. This suggests the high correlation among the selected gene expression and miRNA variables. Results from the imputation based on the approximate low rank assumption given in Section 5 are also encouraging with promising correlations with true signals and good prediction performance from the imputed miRNA signatures. We expect that this imputation method will also work well in genotyping and sequencing applications, particularly for regions with reasonably high linkage disequilibrium.

Another main assumption that is needed in the theoretical analysis is that there is a significant gap between the rth and $(r+1)$th singular values of A. This assumption may not be valid in real practice. In particular, the singular values of the ovarian dataset analyzed in Section 5 is decreasing smoothly without a significant gap. However, it has been shown in the simulation studies presented in Section 4 that, although there is no significant gap between any adjacent singular values of the matrix to be recovered, the proposed SMC method works well as long as the singular values decay sufficiently fast. Theoretical analysis for the proposed SMC method under more general patterns of singular value decay warrants future research.

To implement the proposed Algorithm 2, major decisions include the choice of threshold values and choosing between column thresholding and row thresholding. Based on both theoretical and numerical studies, optimal threshold values can be set as $T_C = 2\sqrt{p_2/m_2}$ for column thresholding and $T_R =$

$2\sqrt{p_1/m_1}$ for row thresholding. Simulation results in Section 4 show that when both rows and columns are randomly chosen, the results are very similar. In the real data applications, the choice between row thresholding and column thresholding depends on whether the rows or columns are more "homogenous," or closer to being randomly sampled. For example, in the ovarian cancer dataset analyzed in Section 5, the rows correspond to the patients and the columns correspond to the gene expression levels and miRNA levels. Thus, the rows are closer to random sample than the columns, consequently it is more natural to use the row thresholding in this case.

We have shown both theoretically and numerically in Sections 3 and 4 that Algorithm 2 provides a good recovery of A_{22}. However, the naive implementation of this algorithm requires $\min(m_1, m_2)$ matrix inversions and multiplication operations in the for loop that calculates $\|D_{R,s}\|$ (or $\|D_{C,s}\|$), $s \in \{\hat{r}, \hat{r}+1, \ldots, \min(m_1, m_2)\}$. Taking into account the relationship among $D_{R,s}$ (or $D_{C,s}$) for different s's, it is possible to simultaneously calculate all $\|D_{R,s}\|$ (or $\|D_{C,s}\|$) and accelerate the computations. For reasons of space, we leave optimal implementation of Algorithm 2 as future work.

7. Supplementary Materials

We provide additional simulation results and the proofs of the main theorems in the online supplement. Some key technical tools used in the proofs of the main results are also developed and proved. The proofs rely on the results in Cai and Zhang (2014), Gross and Nesme (2010), Laurent and Massart (2000), Vershynin (2010) and Vershynin (2013).

Acknowledgment

The authors thank the editor, associate editor, and referee for their detailed and constructive comments that have helped to improve the presentation of the article.

Funding

The research of Tianxi Cai was supported in part by NIH Grants R01 GM079330 and U54 H6007963; the research of Tony Cai and Anru Zhang was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

References

Argyriou, A., Evgeniou, T., and Pontil, M. (2008), "Convex Multi-Task Feature Learning," *Machine Learning*, 73, 243–272. [621]

Berchuck, A., Iversen, E. S., Lancaster, J. M., Pittman, J., Luo, J., Lee, P., Murphy, S., Dressman, H. K., Febbo, P. G., West, M., Nevins, J. R., and Marks, J. R. (2005), "Patterns of Gene Expression That Characterize Long-Term Survival in Advanced Stage Serous Ovarian Cancers," *Clinical Cancer Research*, 11, 3686–3696. [630]

Biswas, P., Lian, T.-C., Wang, T.-C., and Ye, Y. (2006), "Semidefinite Programming Based Algorithms for Sensor Network Localization," *ACM Transactions on Sensor Networks (TOSN)*, 2, 188–220. [621]

Bonome, T., Lee, J.-Y., Park, D.-C., Radonovich, M., Pise-Masison, C., Brady, J., Gardner, G. J., Hao, K., Wong, W. H., Barrett, J. C., Lu, K. H., Sood, A. K., Gershenson, D. M., Mok, S. C., and Birrer, M. J. (2005), "Expression Profiling of Serous Low Malignant Potential, Low-Grade, and High-Grade Tumors of the Ovary," *Cancer Research*, 65, 10602–10612. [630]

Browning, B. L., and Browning, S. R. (2009), "A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals," *The American Journal of Human Genetics*, 84, 210–223. [622]

Cai, J.-F., Candès, E., and Shen, Z. (2010), "A Singular Value Thresholding Algorithm for Matrix Completion," *SIAM Journal on Optimization*, 20, 1956–1982. [624]

Cai, T. T., and Zhang, A. (2014), "Perturbation Bound on Unilateral Singular Vectors," Technical Report. [632]

Cai, T. T., and Zhou, W. (2013), "Matrix Completion via Max-Norm Constrained Optimization," *arXiv preprint arXiv:1303.0341*. [621]

Cancer Genome Atlas Research Network. (2011), "Integrated Genomic Analyses of Ovarian Carcinoma," *Nature*, 474, 609–615. [630]

Candes, E. J., and Plan, Y. (2011), "Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements," *IEEE Transactions on Information Theory*, 57, 2342–2359. [621]

Candès, E. J., and Recht, B. (2009), "Exact Matrix Completion via Convex Optimization," *Foundations of Computational Mathematics*, 9, 717–772. [621]

Candès, E. J., and Tao, T. (2010), "The Power of Convex Relaxation: Near-Optimal Matrix Completion," *IEEE Transactions on Information Theory*, 56, 2053–2080. [621]

Chen, P., and Suter, D. (2004), "Recovering the Missing Components in a Large Noisy Low-Rank Matrix: Application to Sfm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 1051–1063. [621]

Chi, E. C., Zhou, H., Chen, G. K., Del Vecchyo, D. O., and Lange, K. (2013), "Genotype Imputation via Matrix Completion," *Genome Research*, 23, 509–518. [621]

Denkert, C., Budczies, J., Darb-Esfahani, S., Györffy, B., Sehouli, J., Könsgen, D., Zeillinger, R., Weichert, W., Noske, A., Buckendahl, A.-C., Müller, B. M., Dietel, M., and Lage, H. (2009), "A Prognostic Gene Expression Index in Ovarian CancerValidation Across Different Independent Data Sets," *The Journal of Pathology*, 218, 273–280. [630]

Dressman, H. K., Berchuck, A., Chan, G., Zhai, J., Bild, A., Sayer, R., Cragun, J., Clarke, J., Whitaker, R .S., Li, L., Gray, J., Marks, J., Ginsburg, G. S., Potti, A., West, M., Nevins, J. R., and Lancaster, J. M. (2007), "An Integrated Genomic-Based Approach to Individualized Treatment of Patients With Advanced-Stage Ovarian Cancer," *Journal of Clinical Oncology*, 25, 517–525. [630]

Foygel, R., Salakhutdinov, R., Shamir, O., and Srebro, N. (2011), "Learning With the Weighted Trace-Norm Under Arbitrary Sampling Distributions," *NIPS*, 2133–2141. [621]

Gross, D. (2011), "Recovering Low-Rank Matrices From Few Coefficients in Any Basis," *IEEE Transactions on Information Theory*, 57, 1548–1566. [621]

Gross, D., and Nesme, V. (2010), "Note on Sampling Without Replacing From a Finite Collection of Matrices," *arXiv preprint, arXiv:1001.2738*. [632]

Holschneider, C. H., and Berek, J. S. (2000), "Ovarian Cancer: Epidemiology, Biology, and Prognostic Factors," *Seminars in Surgical Oncology*, 19, 3–10. [630]

Kalbfleisch, J. D., and Prentice, R. L. (2011), *The Statistical Analysis of Failure Time Data* (Vol. 360), Hoboken, NJ: Wiley. [630]

Keshavan, R. H., Montanari, A., and Oh, S. (2010), "Matrix Completion From Noisy Entries," *Journal of Machine Learning Research*, 11, 2057–2078. [621]

Kim, H., Golub, G. H., and Park, H. (2005), "Missing Value Estimation for DNA Microarray Gene Expression Data: Local Least Squares Imputation," *Bioinformatics*, 21, 187–198. [622]

Koltchinskii, V. (2011), "Von Neumann Entropy Penalization and Low-Rank Matrix Estimation," *Annals of Statistics*, 39, 2936–2973. [621]

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011), "Nuclear-Norm Penalization and Optimal Rates for Noisy Low-Rank Matrix Completion," *Annals of Statistics*, 39, 2302–2329. [621]

Koren, Y., Bell, R., and Volinsky, C. (2009), "Matrix Factorization Techniques for Recommender Systems," *Computer*, 42, 30–37. [621]

Laurent, B., Massart, P. (2000), "Adaptive Estimation of a Quadratic Functional by Model Selection," *Annals of Statistics*, 28, 1302–1338. [632]

Li, Y., and Abecasis, G. R. (2006), "Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference," *The American Journal of Human Genetics*, 79, 2290. [622]

Mazumder, R., Hastie, T., and Tibshirani, R. (2010), "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research*, 11, 2287–2322. [621,623,629]

Recht, B. (2011), "A Simpler Approach to Matrix Completion," *Journal of Machine Learning Research*, 12, 3413–3430. [621]

Rohde, A., and Tsybakov, A. B. (2011), "Estimation of High-Dimensional Low-Rank Matrices," *Annals of Statistics*, 39, 887–930. [621]

Salakhutdinov, R., and Srebro, N. (2010), "Collaborative Filtering in a Non-Uniform World: Learning With the Weighted Trace Norm," *NIPS*, 2056–2064. [621]

Scheet, P., and Stephens, M. (2006), "A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase," *The American Journal of Human Genetics*, 78, 629–644. [622]

Siegel, R., Naishadham, D., and Jemal, A. (2013), "Cancer Statistics, 2013," *CA: A Cancer Journal for Clinicians*, 63, 11–30. [630]

Singer, A., and Cucuringu, M. (2010), "Uniqueness of Low-Rank Matrix Completion by Rigidity Theory," *SIAM Journal on Matrix Analysis and Applications*, 31, 1621–1641. [621]

Toh, K.-C., and Yun, S. (2010), "An Accelerated Proximal Gradient Algorithm for Nuclear Norm Regularized Least Squares Problems," *Pacific Journal of Optimization*, 6, 615–640. [623,629]

Tomasi, C., and Kanade, T. (1992), "Shape and Motion from Image Streams: A Factorization Method," *Proceedings of the National Academy of Sciences*, 90, 9795–9802. [621]

Tothill, R. W., Tinker, A. V., George, J., Brown, R., Fox, S. B., Lade, S., Johnson, D. S., Trivett, M. K., Etemadmoghadam, D., Locandro, B., Traficante, N., Fereday, S., Hung, J. A., Chiew, Y.-E., Haviv, I., Australian Ovarian Cancer Study Group, Gertig, D., DeFazio, A., and Bowtell, D. D. L. (2008), "Novel Molecular Subtypes of Serous and Endometrioid Ovarian Cancer Linked to Clinical Outcome," *Clinical Cancer Research*, 14, 5198–5208. [630]

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001), "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics*, 17, 520–525. [622]

Vershynin, R. (2010), *Introduction to the Non-Asymptotic Analysis of Random Matrices*, Cambridge, UK: Cambridge University Press. [632]

——— (2013), "Spectral Norm of Products of Random and Deterministic Matrices," *Probability Theory and Related Fields*, 150, 471–509. [632]

Wang, X., Li, A., Jiang, Z., and Feng, H. (2006), "Missing Value Estimation for DNA Microarray Gene Expression Data by Support Vector Regression Imputation and Orthogonal Coding Scheme," *BMC Bioinformatics*, 7, 32. [622]

Yu, Z., and Schaid, D. J. (2007), "Methods to Impute Missing Genotypes for Population Data," *Human Genetics*, 122, 495–504. [622]