ASSVD: Adaptive Sparse Singular Value Decomposition for High Dimensional Matrices

Xiucai Ding[1], Xianyi Chen[2*], Mengling Zou[3] and Guangxing Zhang[4]
[1] Department of Mathematics, Duke University, 120 Science Drive, Physics Building Room 095, Durham, NC, 27710, USA
[e-mail: xiucai.ding@duke.edu]
[2] Mathematics and Computer Science, University of North Carolina at Pembroke, NC, 28372, USA
[e-mail: 0204622@163.com]
[3] Department of Computer Science, University of Drbrecen, Debrecen, H-4208, Hungary
[e-mail: 2404141560@qq.com]
[4]Nanjing Qisheng Cloud Information Technology Co., Ltd., Nanjing, Jiangsu, 211809, China
[e-mail: zgx@qscloudata.com]
*Corresponding author: Xianyi Chen

Received February 17, 2020; revised April 17, 2020; accepted April 21, 2020; published June 30, 2020

Abstract

In this paper, an adaptive sparse singular value decomposition (ASSVD) algorithm is proposed to estimate the signal matrix when only one data matrix is observed and there is high dimensional white noise, in which we assume that the signal matrix is low-rank and has sparse singular vectors, i.e. it is a simultaneously low-rank and sparse matrix. It is a structured matrix since the non-zero entries are confined on some small blocks. The proposed algorithm estimates the singular values and vectors separable by exploring the structure of singular vectors, in which the recent developments in Random Matrix Theory known as anisotropic Marchenko-Pastur law are used. And then we prove that when the signal is strong in the sense that the signal to noise ratio is above some threshold, our estimator is consistent and outperforms over many state-of-the-art algorithms. Moreover, our estimator is adaptive to the data set and does not require the variance of the noise to be known or estimated. Numerical simulations indicate that ASSVD still works well when the signal matrix is not very sparse.

Keywords: Matrix denoising, random matrix theory, adaptive sparse singular value decomposition (ASSVD), anisotropic Marchenko-Pastur law.

http://doi.org/10.3837/tiis.2020.06.017 **ISSN : 1976-7277**

1. Introduction

Matrix denoising is important in many scientific endeavors, such as the data cleansing of big data, image processing, information security and so on [1-3]. Consider that we can observe a $p \times n$ signal-plus-noise matrix

$$\tilde{S} = S + Z, \tag{1}$$

where S is the true signal matrix and Z is the noise matrix. In the classic setting when the dimension of data is much smaller than the sample size, the truncated singular value decomposition (TSVD) [4] is the default technique for estimating S from $\tilde{S}$. Two popular methods, for choosing the truncation level, are soft-thresholding [6] and hard-thresholding [7].

The advance of technology has led to high dimensional data set whose dimensionality diverges with sample size n. In this regime, the classic multivariate analysis [8] lose its validity and Random Matrix Theory (RMT) [9] serves as a powerful technical tool. In this paper, we study the high dimensional data set when p is comparable to n, i.e. there exists some small constant $\tau > 0$ such that

$$\tau \leq c_n \leq \tau^{-1},\ c_n := \frac{p}{n}. \tag{2}$$

It is remarkable that, unlike the standard results in RMT [10], we only require the boundedness of C_n instead of the convergence. This makes our algorithm more adaptive to test data set.

In this paper, we consider the estimation of S from its noisy estimation $\tilde{S}$ in the high dimensional setting when (eq.2) holds. A popular and practical assumption on S is simultaneously low rank and sparse in the sense that S has a finite number of nonzero singular values and sparse singular vectors. This type of data set is commonly encountered in many scientific disciplines [11-13]. A typical example is from the study of gene expression data. An microarray experiment typically assesses a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags) under multiple conditions. The gene expression data in an microarray experiment can be represented by a real-valued expression matrix S, where the rows of S correspond to the expression pattern of genes (e.g. cancer patient) and column correspond to the gene levels. A subset of gene patterns can be clustered together as a subtype of the same pattern, which in turn is determined by a subset of genes. The original gene expression matrix obtained from a scanning process contains noise, missing values, and systematic variations arising from the experimental procedure. Therefore, our discussion here provides an ideal model for the gene expression data.

In the literature of low-rank matrix estimation, nuclear norm minimization (NNM) [14,15] are proved to be useful. However, since we have sparse structure here, we expect to obtain better estimate. To handle this issue, one research line is to add more regularization items for optimization other than the nuclear norm, for instance l_1 penalty, to capture the structure of sparsity [16,17]. The other research line is to apply the two-way iterative thresholding method [12] to iteratively explore the low-rank and sparse structure of S. However, in [12], we need to estimate the variance of noise based on prior information which is usually very difficult in practice.

In the present paper, we assume that S has sparse structure in the sense that its singular vectors are sparse. As a consequence, the nonzero entries of S are confined on some small

blocks and hence S is highly structured. In **Fig. 1**, we illustrate this property using simulated synthetic data set. It can be seen that the non-zero entries are confined on some blocks of small sizes.

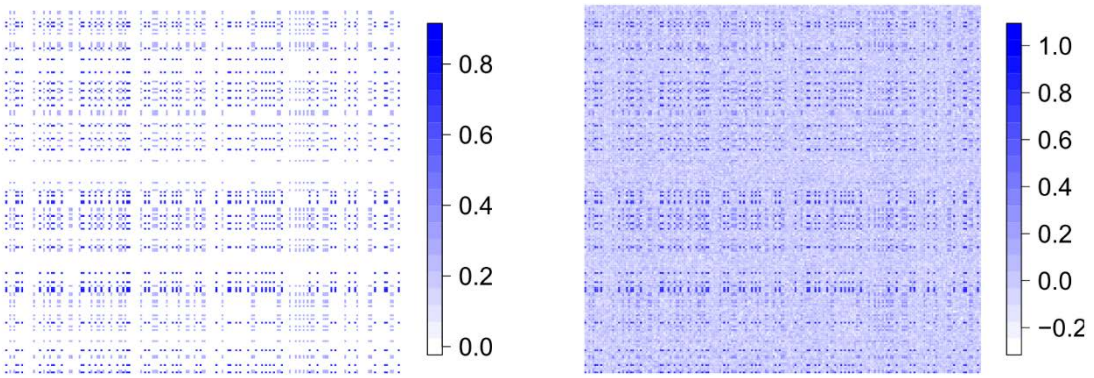

Fig. 1. Synthetic data set. Left panel is the image of S, whereas right panel is $\tilde{S}$. S is a rank-two matrix and its singular vectors have only 10% non-zero entries. Z is a random Gaussian matrix with unit variance.

For this type of data set, instead of investigating the sparse structure of S, the best solution is to explore the structure of the singular vectors, i.e. the positions of the non-zero entries. However, none of the existing methods explore the structure of singular vectors directly.

In this paper, we propose the ASSVD to estimate the singular values and vectors, separately. ASSVD will explore the positions of non-zero entries of the singular vectors directly. From the point of matrix decomposition, our method is rather straightforward and provide the estimates of singular values and vectors. Moreover, with the recent technical inputs from [5], we do not need to estimate the variance of the noise. We also prove that our ASSVD gives a consistent estimator when the S is strong (see Assumption 5 below). Numerical simulations show that ASSVD outperforms over many state-of-the-art algorithms even when the S is neither strong nor sparse.

We point out similar problems have been studied in [5] when the noise variance is one (known). Our ASSVD can deal with general noise situation without estimating the noise variance. Moreover, in this paper, we propose a novel adaptive estimator for the singular values. This estimator only uses the singular values of $\tilde{S}$.

The contributions of this paper can be summarized as follows:

- We propose ASSVD, an adaptive and simple algorithm that enables the estimation of a simultaneously low-rank and sparse matrix in presence of high dimensional noise. Our ASSVD does not need to estimate the variance of noise and is adaptive to the data matrix.
- We theoretically and numerically prove that ASSVD can well estimate the high dimensional data matrix and outperforms over many state-of-the-art algorithms.
- As a byproducts, ASSVD produces estimates for the singular values and vectors. Such results can be of independent interest.

The rest of this paper is organized as follows. In Section 2, we introduce the main assumptions and the proposed ASSVD. In Section 3, we design Monte-Carlo simulations to illustrate the use of ASSVD and compare with some state-of-the-art algorithms. In Section 4, we prove the theoretical properties of ASSVD. Finally, we summarize in Section 5.

2. Adaptive matrix denoising

In this section, we introduce the main assumptions will be used throughout the paper and then the algorithm: ASSVD.

2.1 Main assumptions

We assume that the entries of the white noise matrix $Z = (z_{ij})$ are i.i.d random variables such that

$$\mathbb{E}z_{ij} = 0,\ \mathbb{E}z_{ij}^2 = \frac{\sigma^2}{n}. \tag{3}$$

and the noise variance is bounded, i.e

$$\sigma < \infty. \tag{4}$$

Moreover, there exists a large constant $C > 0$ for $k \le C$, a constant $\mu_k > 0$,which makes

$$\mathbb{E}|\sqrt{n}z_{ij}|^k \le \mu_k,\ 3 \le k \le C. \tag{5}$$

Denote the singular value decomposition of S as

$$S = \sum_{i=1}^{r} d_i\, u_i v_i^{\mathrm{T}},\ d_1 > d_2 > \cdots > d_r, \tag{6}$$

where $r > 0$ is a fixed integer, d_i, u_i and v_i are the singular values, left and right singular vectors of S, respectively. We assume that

$$0 < d_i < \infty,\ 1 \le i \le r. \tag{7}$$

Moreover, we assume that u_i, v_i are sparse. Specifically, let m_u^i and m_v^i be the number of non-zero entries of u_i and v_i, respectively. Denote

$$w = \max_{1 \le i \le r}\{m_u^i, m_v^i\}.$$

then assume that there exists some constant $C_1 > 0$ such that

$$w \le C_1. \tag{8}$$

In light of (2), we define the sparsity level of S as

$$s = \frac{w}{n}.$$

We conclude from (8) that $s \to 0$ when $n \to \infty$.

For future reference, we summarize the assumptions of the present paper.

Assumption 1. For the model (1), we assume that (2), (3), (4), (5]), (6), (7) and (8) hold true.

2.2 Adaptive spare singular value decomposition (ASSVD)

We now introduce our algorithm, ASSVD. As mentioned in Section 1, our algorithm estimates the singular values and vectors, separately.

Algorithm 1 ASSVD

Require: Data matrix $\widetilde{S}$, ratio c_n, rank estimate q.
Ensure: Estimate of S
1: Do SVD for $\widetilde{S} = \sum_{i=1}^{p\wedge n} \lambda_i \tilde{u}_i \tilde{v}_i^{\mathrm{T}}$, and do the initialization $\widetilde{S}_1 = \widetilde{S} = \sum t_i^1 \tilde{u}_i^1 (\tilde{v}_i^1)^{\mathrm{T}}$.
2: **while** $1 \le j \le q$ **do**
3: Estimate the singular d_j using

$$\hat{d}_j = (t_1^j \widehat{m}_1(t_1^j)\widehat{m}_2(t_1^j))^{-1/2}. \tag{9}$$

4: Do K-means clustering to partition the entries of $\widetilde{u}_1^j$ and $\widetilde{v}_1^j$ into two classes, where

$$I_j := \{1 \le k \le p : |\tilde{u}_1^j(k)| \gg p^{-1/2}\},$$

and

$$J_j := \{1 \le k \le n : |\tilde{v}_1^j(k)| \gg n^{-1/2}\}.$$

5: Do SVD for the block matrix $\widetilde{S}_b = \widetilde{S}_j[I_j, J_j]$ and denote

$$\widetilde{S}_j[I_j, J_j] = \sum \rho_i u_i^j (v_i^j)^{\mathrm{T}}.$$

6: Assume $I_j = \{k_1, \cdots, k_l\}$, construct $\widehat{u}_j$ by letting

$$\widehat{u}_j(k_i) = \begin{cases} u_1^j(i), & k_i \in I_j, \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, we can construct $\widehat{v}_j$.
7: Let $\widetilde{S}_{j+1} = \widetilde{S}_j - \widehat{d}_j \widehat{u}_j \widehat{v}_j^{\mathrm{T}}$ and do SVD for $\widetilde{S}_{j+1} = \sum t_i^{j+1} \widetilde{u}_i^{j+1} (\widetilde{v}_i^{j+1})^{\mathrm{T}}$.
8: **end while**
9: Denote our estimator as

$$\widehat{S}_{assvd} = \sum_{i=1}^{q} \widehat{d}_i \widehat{u}_i \widehat{v}_i^{\mathrm{T}}. \tag{10}$$

We first introduce some notations. Denote the eigenvalues and eigenvectors of $\tilde{S}\tilde{S}^{\mathrm{T}}$ as $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$ and $\tilde{u}_1, \cdots, \tilde{u}_p$, respectively. Similarly, we define the eigenvalues and eigenvectors of $\tilde{S}^{\mathrm{T}}\tilde{S}$ as $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_n$ and $\tilde{v}_1, \cdots, \tilde{v}_n$, respectively. Since $\tilde{S}\tilde{S}^{T}$ and $\tilde{S}^{\mathrm{T}}\tilde{S}$ have the same non-zero eigenvalues, in order not to cause confusion, we define them as $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_{p\wedge n}$, $p \wedge n = \min\{p, n\}$. Note that $\{\tilde{u}_i\}$ and $\{\tilde{v}_i\}$ are the left and right singular vectors of $\tilde{S}$, respectively. We next make it precise of the strong signal.

Definition 2. For $1 \le i \le r$, we say d_i is a strong signal if

$$\frac{d_i}{\sigma} > c_n^{1/4} + \kappa,$$

where κ is a fixed small constant.
Next, we define some statistics. For a given parameter q and when $i \le q$, denote

$$\hat{m}_1(\lambda_i) = \frac{1}{p}\sum_{j=q+1}^{p}\frac{1}{\lambda_i - \lambda_j},\ \ \hat{m}_2(\lambda_i) = \frac{1}{n}\sum_{j=q+1}^{p}\frac{1}{\lambda_i - \mu_j}.$$

It will seen later that q is used to estimate the number of strong signals and we refer it as the rank estimate. Its definition and construction will be discussed in Section 2.3. Armed with the above preparation, we introduce ASSVD as Algorithm 1.

Remark 3. First of all, from the above procedure, when d_i is a strong signal satisfying Definition 2, we conclude that $\hat{d}_i, \hat{u}_i$ and $\hat{v}_i$ are the estimates of d_i, u_i and v_i, respectively. Secondly, our ASSVD is adaptive to our data matrix only and we do not need to estimate the variance of the noise.

2.3 Choice of parameter

As we have seen from the ASSVD algorithm, the number of strong signals of S needed to be estimated separately using the parameter q. In this paper, we employ the resampling procedure [21] to choose q. The main idea behind the construction is to use the information of magnitude of singular values of $\tilde{S}$. Heuristically, as we can see from Theorem 6 later, if d_i, d_{i+1} are both strong signals, then the ratio of their corresponding singular values λ_i/λ_{i+1} will be well-separate from one. On the other hand, if both of them are weak signals in the sense that Definition 2 fails, their ratios will be close to one. Hence, there exists a transition point for the ratio of consecutive singular values of $\tilde{S}$ and this happens between the q-th and $(q+1)$-th singular values, which information will be used to construct the statistic.

Remark 4. It can conclude from the above algorithm that with probability β (say $\beta = 0.98$), q will be a reasonably statistic for the number of strong signals. The ς is chosen to make precise of being far away from one.

Table 1. Estimate of singular values $d_1 = 10$ using (10). We report the averaged estimate over 10^4 simulations.

$\sigma/(p,n)$	(100, 200)	(200, 400)	(300, 400)	(400, 300)
1	9.8	9.83	9.9	10.12
2	9.6	10.13	9.96	9.82
3	9.45	10.05	9.86	9.87
4	9.83	9.65	9.88	10.51
5	9.35	10.76	9.93	10.56

3. Simulations

3.1 Performance of the estimates $\hat{d}_i, \hat{u}_i$ and $\hat{v}_i$

As mentioned before, ASSVD estimates the singular values and vectors separately. We first study the performance of such estimators using a rank-two example under various noise level when Assumption 5 holds. To generate sparse singular vectors, we use the R package

R1magic. The noise matrix is chosen to be a random Gaussian matrix generated from the R package mvtnorm. In the simulations below, we set

$$S = d_1 u_1 v_1^{\mathrm{T}} + d_2 u_2 v_2^{\mathrm{T}}.$$

Here $u_i \in \mathbb{R}^p$, $v_i \in \mathbb{R}^n, i = 1,2$ are generated using R1magic with $s = 0.1$ and $u_1 \perp u_2, v_1 \perp v_2$ with $d_1 = 10, d_2 = 7$. In **Table 1**, we report the estimation of d_1 using (9) for a variety choices of noise levels and combinations (p,n). It can be seen that we estimator is robust against all such combinations.

Next, we consider the accuracy of the estimation for singular vectors. We report the results of the left and right singular vectors for a fixed noise level in **Fig. 2** and **3** respectively. It can be seen that our estimate is quite accurate.

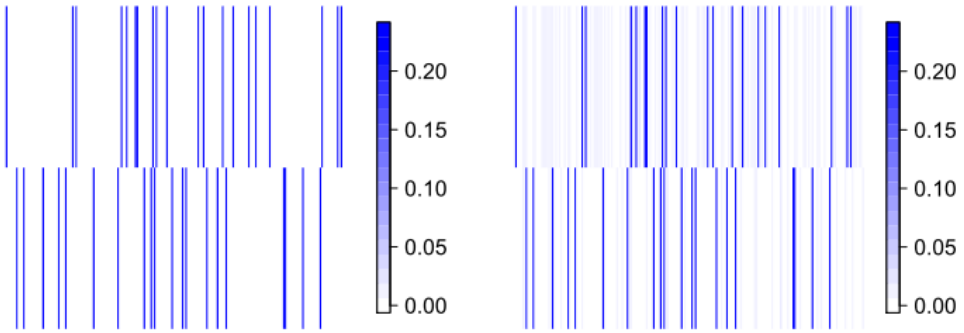

Fig. 2. Estimation of left singular subspace for σ = 2, p = 200, n = 400. Left panel is the true subspace whereras right panel is the estimated subspace.

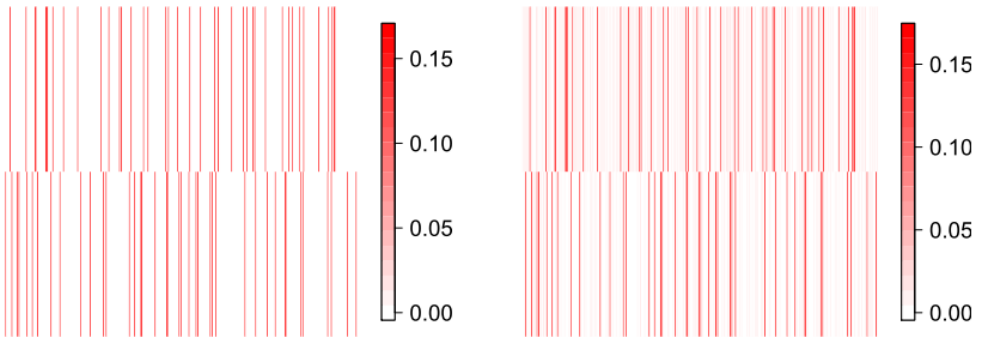

Fig. 3. Estimation of right singular subspace for σ = 2, p = 200, n = 400. Left panel is the true subspace whereras right panel is the estimated subspace.

3.2 Comparison with other algorithms

In this section, we compare ASSVD with some state-of-the-art algorithms. Specifically, we compare with the sparse singular value decomposition (SSVD) in [12], the nuclear norm minimization with l_1 penalty [16] (NSNM), optimal shrinkage of singular values (OSSVD) in [18], shrinkage estimates (OptShrink) in [19] and the truncated SVD (TSVD). For the implementation of SSVD, we use an R package **ssvd** contributed by the first author of [12].

For the shrinkage estimates in [18, 19], the Matlab codes can be found on the author's websites.

Algorithm 2 Resampling estimation

Require: Non-zero singular values of $\widetilde{S}$, covering probability β, simulation number m
Ensure: Rank estimate

1: Generate a sequence of m i.i.d $p \times n$ random Gaussian matrices satisfying Assumption 1 with $\sigma = 1$. We record them as $Z_k, k = 1, 2, \cdots, m$.
2: Calculate the ratios of the first and second eigenvalues of $Z_k Z_k^{\mathrm{T}}$ and record them as $\mathcal{R}_k, k = 1, 2, \cdots, m$.
3: For the given probability β, find the value ς such that

$$\frac{\#\{\mathcal{R}_k - 1 \leq \varsigma\}}{m} = \beta.$$

4: Estimate the rank r^+ using

$$q = \arg\max\{1 \leq i \leq p \wedge n : \mathcal{R}_i > 1 + \varsigma\}.$$

Table 2. Comparison of different algorithms using Frobenius norm. We record the estimation errors for different methods averaged over 10^4 simulations. We highlight the smallest error terms.

		$d_n = 0.5$		$d_n = 2$	
Method	s/p	300	500	300	500
ASSVD	0.05	**0.043**	**0.041**	**0.045**	**0.039**
	0.1	**0.614**	**0.1**	**0.6**	**0.16**
	0.2	**0.822**	**0.2**	**0.825**	**0.137**
	0.45	**1.1**	**0.45**	**1.09**	**0.09**
SSVD	0.05	4.01	4	4.01	4.01
	0.1	4.01	4.03	4.02	4.01
	0.2	4.04	4.04	4.03	4
	0.45	4.06	4.06	4.08	4.02
NSNM	0.05	4.67	4.65	4.51	4.48
	0.1	6.51	6.53	5.82	5.61
	0.2	8.98	8.95	7.93	7.99
	0.45	15.76	14.78	15.63	14.78
OSSVD	0.05	5.01	5.08	4.71	4.65
	0.1	5.41	5.38	6.02	5.98
	0.2	7.04	7.12	6.83	6.124
	0.45	11.06	11.76	10.08	10.95
OptShrink	0.05	4.1	3.98	3.96	3.5
	0.1	4.51	4.41	4.05	4
	0.2	5.04	4.98	4.63	4.41
	0.45	8	7.8	7.1	6.95
TSVD	0.05	53.9	65.24	53.75	65.93
	0.1	53.72	68.87	53.38	66.71
	0.2	52.33	63.16	52.2	64.65
	0.45	51.043	62.78	52.4	74.13

First of all, we study the performance of various methods for a fixed noise level $\sigma = 1$. We use the same setup as in Section 3 by varying the sparsity level s between 0.05 and 0.45. It can be concluded from **Table 2** that: (i). ASSVD outperforms the other algorithms at all levels of sparsity and combinations of p and n; (ii). Even though we assume (8) and subsequently $s \to 0$ asymptotically, numerical simulations indicate that our estimation is still reasonable accurate when S is not very sparse; (iii). TSVD has the worst performance and becomes worse with the increase of dimension; but it is stable under sparsity variation; (iv). SSVD has stable and smaller errors at all sparsity levels. However, we will show later that it will become worse (as indicated in when d_1, d_2 increases); (v). The penalty method becomes worse when the sparsity level increases.

We mention that, in this setting, both d_1 and d_2 are strong signals.

4 Theoretical properties

In this section, we state the main statistical properties of ASSVD. The key ingredients for our paper are the convergence limits and rates for the singular values and vectors.

4.1 Convergence of singular values and vectors of $\tilde{S}$

In [5], the author computed the convergence limits and rates for the singular values and vectors when $\sigma^2 = 1$. We extend the results for general noise level σ.

$$\theta(d) := \frac{(d^2 + \sigma^2)(d^2 + c_n\sigma^2)}{d^2},$$

and

$$a_1(d) := \frac{d^4 - c_n}{d^2(d^2 + c_n)},\ a_2(d) = \frac{d^4 - c_n}{d^2(d^2 + 1)}.$$

We next introduce the assumptions on the strength of the signals $d_i, 1 \le i \le r$.

Assumption 5. Suppose that for some $1 \le r^+ \le r$ and some small constant $\kappa > 0$, we have

$$d_i > \sigma d_n^{1/4} + \kappa,\ |d_i - d_j| \ge \kappa,\ 1 \le i \ne j \le r^+.$$

Moreover, when $r^+ + 1 \le k \le r$, we assume

$$d_k < \sigma c_n^{1/4}.$$

We next state the results for the singular values and vectors.

Theorem 6. We suppose that Assumptions 1 and 5 hold true. For any given small $\epsilon > 0$, there exists a large constant $D \equiv D(\epsilon) > 0$, such that for sufficiently large n, with probability at least $1 - n^{-D}$, we have

$$|\lambda_i - \theta(d_i)| \le n^{-1/2+\epsilon},\ 1 \le i \le r^+, \tag{11}$$

and

$$\left|\langle u_j, \tilde{u}_i\rangle^2 - \delta_{ij} a_1(d_i)\right| \le (\delta_{ij} n^{-1/2+\epsilon} + n^{-1+\epsilon}),$$

$$\left|\langle v_j, \tilde{v}_i\rangle^2 - \delta_{ij} a_2(d_i)\right| \le (\delta_{ij} n^{-1/2+\epsilon} + n^{-1+\epsilon}),$$

where $\delta_{ij} = 1$ when $i = j$ and $\delta_{ij} = 0$ otherwise. Furthermore, for $r^+ + 1 \le k \le r$, we have

$$|\lambda_k - \sigma^2(1 + d_n^{1/2})^2| \le n^{-2/3+\epsilon},$$

and

$$\langle u_l, \tilde{u}_k\rangle^2 \le n^{-1+\epsilon},\ \langle v_l, \tilde{v}_k\rangle^2 \le n^{-1+\epsilon},\ 1 \le l \le r.$$

Proof. Denote $\tilde{S}_1 = \tilde{S}/\sigma, S_1 = S/\sigma$ and $Z_1 = Z/\sigma$. The results for the model

$$\tilde{S}_1 = S_1 + Z_1,$$

have been established in [Section 2 Theorem 2.2 and 2.3]. Note that $\tilde{S}_1$ and $\tilde{S}$ have the same singular vectors and $\lambda(\tilde{S}) = \sigma\lambda(\tilde{S}_1)$. We can therefore conclude the proof using [Section 2 Theorem 2.2 and 2.3].

We remark that the almost surely convergence results have been established in [20] using Free Probability Theory. We provide convergent rates in the above theorems.

4.2 Convergence of the estimator $\hat{S}_{assvd}$

With the preparation of Theorem 6, we next establish the properties of our estimator $\hat{S}_{assvd}$ under Frobenius norm. Recall (10).

Theorem 7. Suppose that Assumptions 1 and 5 hold true. Then for some small constant $\epsilon > 0$, there exists a large constant $D \equiv D(\epsilon) > 0$, such that for a sufficiently large n, with probability at least $1 - n^{-D}$, we have

$$\left\|S - \hat{S}_{assvd}\right\|_F \le n^{-1/2+\epsilon} + \sqrt{\sum_{i=r^+ +1}^{r} d_i^2}.$$

Proof. We decompose $S = S_o + S_b$, where

$$S_o = \sum_{i=1}^{r^+} d_i\, u_i v_i^{\mathrm{T}},\ S_b = \sum_{i=r^+ +1}^{r} d_i\, u_i v_i^{\mathrm{T}}.$$

It is easy to see that

$$\left\|S - \hat{S}_{assvd}\right\|_F \le \left\|S_o - \hat{S}_{assvd}\right\|_F + \sqrt{\sum_{i=r^+ +1}^{r} d_i^2}.$$

From the proof of [2, Theorem 3.4] (see equation (5) there), we find that with probability at least $1 - n^{-D}$

$$\left\|S_o - \hat{S}_{assvd}\right\|_F^2 \le n^{-1+2\epsilon} + 2\sum_{i=1}^{r^+}(\hat{d}_i - d_i)^2.$$

Moreover, by [2, Proposition 3.3], we find that $q = r^+$ with probability at least $1 - n^{-D}$. Therefore, the proof follows from the following lemma and its proof can be found in the appendix.

Lemma 8. Recall the estimate $\hat{d}_i$ in (9). Assume the assumptions of Theorem 7 holds. Then with probability at least $1 - n^{-D}$, we have

$$|\hat{d}_i - d_i| \le n^{-1/2+\epsilon},\ i \le r^+.$$

We conclude from Theorem 7 that when $r^+ = r$, i.e. all signals are strong, ASSVD can provide us a consistent estimator. However, in this situation, the shrinkage algorithms (OSSVD [18] and OptShrink [19]) obtain bound

$$\sqrt{\sum_{i=1}^{r} d_i^2\,(1 - a_1(d_i)a_2(d_i))} > 0$$

since $0 < a_1(d_i), a_2(d_i) < 1$.

For the iterative thresholding method SSVD , even though we theoretically have the same rate with them, numerically simulations show better performance than them. Moreover, since our algorithm does not involve any iterations, ASSVD is more simple and fast in the implementation.

For the penalty method, there does not exist any literature on proving the optimal bounds. However, as we can see from the mini-max bound in [17] that it will be bounded by the nuclear norm, which is strictly positive.

5. Conclusions and discussions

In this paper, we study the problem of estimating a simultaneously low-rank and sparse matrix from a high dimensional noisy observation. We propose an efficient algorithm, adaptive sparse singular value decomposition (ASSVD), by exploring the structure of the singular values and vectors. The inputs of ASSVD are based on recent developments in Random Matrix Theory. An main advantage is that we do not need to estimate the variance of the noise.Theoretical analysis shows that ASSVD outperforms over many existing methods. Extensive experimental results demonstrate the efficiency and efficacy of our proposed method. Moreover, ASSVD still works very well even when the data matrix is not very sparse. One future direction is to generalize this idea to incorporate high dimensional heteroskedastic noise. It is also very interesting to explore the situation when the rank of S diverges with n.

References

[1] Z. H. Xia, L. H. Lu, T. Qiu, H. J. Shim, X. Y. Chen, and Byeungwoo Jeon, "A Privacy-Preserving Image Retrieval Based on AC-Coefficients and Color Histograms in Cloud Environment," *Computers, Materials & Continua*, Vol. 58, No. 1, pp. 27-43, 2019. Article(CrossRef Link)

[2] X. Y. Chen, H. D. Zhong, and Z. F. Bao, "A GLCM-Feature-Based Approach for Reversible Image Transformation," *Computers, Materials & Continua*, Vol. 59, No. 1, pp. 239-255, 2019. Article(CrossRef Link)

[3] L. Z. Xiong, and Y. Q. Shi, "On the Privacy-Preserving Outsourcing Scheme of Reversible Data Hiding Over Encrypted Image Data in Cloud Computing," *Computers, Materials & Continua*, Vol.55, No.3, pp.523-539, 2018. Article(CrossRef Link)

[4] Dennis I. Merino, "Topics in Matrix Analysis," *Cambridge University Press*, 2008. Article(CrossRef Link)

[5] X.C. Ding, "High dimensional deformed rectangular matrix with applications in matrix denoising," *Bernoulli*, Vol.26, pp.387-417, 2020. Article(CrossRef Link)

[6] D. Donoho, "De-noising by soft-thresdholding," *IEEE Trans. Inf. Theory*, vol. 41, pp.613–627,1995. Article(CrossRef Link)

[7] M. Gavish, and D. L. Donoho, "The Optimal Hard Threshold for Singular Values is p 4/3," *IEEE Trans. Inf. Theory*, vol. 60, pp.5040–5053, 2014. Article(CrossRef Link)

[8] N. H. Timm, *Applied Multivariate Analysis*, Springer Texts in Statistics, Springer Science & Business Media, 2007.

[9] L. Erd´os, and H.-T. Yau, "A Dynamical Approach to Random Matrix Theory," *Courant Lecture Notes, American Mathematical Soc*, Vol. 28, 2017. Article(CrossRef Link)

[10] Z. D. Bai, and J.W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, 2*nd Edition*, Springer Series in Statistics, Springer-Verlag New York, 2010. Article(CrossRef Link)

[11] B. Pontes, R. Giraldez, and J. Aguilar-Ruiz, "Biclustering on expression data: A review," *J. Biomed. Inform.*, vol. 57, pp. 163–180, 2015. Article(CrossRef Link)

[12] D. Yang, Z. Ma, and A. Buja, "Rate optimal denoising of simultaneously sparse and low rank matrices," *J. Mach. Learn. Res.*, vol. 17, pp. 1–27, 2016. Article(CrossRef Link)

[13] R. Otazo, E. Cand´es, and D. K. Sodickson, "Low-rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components," *Magn Reson Med.*, vol 73, pp. 1125-1136, 2014. Article(CrossRef Link)

[14] J. F. Cai, E. J. Cand´es, and Z. W. Shen, "A Singular Value Thresholding Algorithm for Matrix Completion," *SIAM J. Optim.*, vol. 20, pp.1956–1982, 2010. Article(CrossRef Link)

[15] M. Gavish, and D. L. Donoho, "Minimax risk of matrix denoising by singular value thresholding," *Ann. Statist.*, vol. 42, pp. 2413–2440, 2014. Article(CrossRef Link)

[16] P. V. Giampouras, K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "Simultaneously Sparse and Low-Rank Abundance Matrix Estimation for Hyperspectral Image Unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol 54, pp.4775–4789, 2016. Article(CrossRef Link)

[17] E. Richard, P-A. Savalle, and N. Vayatis, "Estimation of simultaneously sparse and low rank matrices," in *Proc. of ICML'12*, pp.51-58, 2012. Article(CrossRef Link)

[18] M. Gavish, and D. L. Donoho, "Optimal shrinkage of singular values," *IEEE Trans. Inf. Theory*, vol. 63, pp.2137–2152, 2017. Article(CrossRef Link)

[19] R. R. Nadakuditi, "OptShrink: An Algorithm for Improved Low-Rank Signal Matrix Denoising by Optimal, Data-Driven Singular Value Shrinkage," *IEEE Trans. Inf. Theory*, vol. 60, pp.3002–3018, 2014. Article(CrossRef Link)

[20] F. Benaych-Georges, and R. R. Nadakuditi, "The singular values and vectors of low rank perturbations of large rectangular random matrices," *J. Multivariate Anal.*, vol.111, pp. 120-135, 2012. Article(CrossRef Link)

[21] D. Passemiera, and J.F. Yao, "Estimation of the number of spikes, possibly equal, in the high-dimensional case," *J. Multivariate Anal.*, vol 127, pp.173-183, 2014. Article(CrossRef Link)
[22] A. Knowles, and J. Yin, "Anisotropic local laws for random matrices," *Probab. Theory Relat. Fields*, vol. 169, pp.257–352, 2017. Article(CrossRef Link)
[23] W. Li, B. Zhu, "A 2k-vertex kernel for Vertex Cover based on Crown Decomposition," *Theoretical computer science*, vol 739, pp.80-85, 2018. Article(CrossRef Link)

Appendix

In this appendix, we first summarize the recent results of Random Matrix Theory, the anisotropic local laws in [22]. Then we prove Lemma 8.

We investigate the spectrum properties of ZZ^{T} and $Z^{\mathrm{T}}Z$ provided Assumption 1 holds. For any symmetric $n \times n$ matrix H, the empirical spectral distribution (ESD) is defined as

$$\mu_H = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(H)},$$

where $\delta(\cdot)$ is the standard Dirac-Delta function. For any probability measure μ and complex value $z \in \mathbb{C}_+$, we denote the Stieltjes transform as

$$m_\mu(z) = \int \frac{1}{x - z}\mu(dx).$$

It is well-known that the limiting spectral distribution of $\frac{1}{\sigma^2}ZZ^{\mathrm{T}}$ satisfies the Marchenko-Pastur (MP) law denoted as

$$\mu_{mp}(I) = \max\{1 - c, 0\}\mathbf{1}_{0\in I} + \nu(I),$$

where $I \subset \mathbb{R}$ is a measurable set and $d\nu(x)$ satisfies

$$d\nu(x) = \frac{1}{2\pi}\frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{c_n x}dx,\ \lambda_\pm = (1 \pm c_n^{1/2})^2.$$

We denote the Stieltjes transform of the MP law as $m_1(z)$. Similar results hold for $\frac{1}{\sigma^2}Z^{\mathrm{T}}Z$ and we denote its Stieltjes transform as $m_2(z)$.

To study each individual eigenvalue, we need the local MP law. Denote the spectral parameter set as

$$\mathrm{S} := \{z = E + \mathrm{i}\eta \in \mathbb{C}:\ \lambda_+ < E \le \infty, \eta \ge 0\}.$$

The Stieltjes transforms of ESDs of $\sigma^{-2}ZZ^{\mathrm{T}}$ and $\sigma^{-2}Z^{\mathrm{T}}Z$ are defined as

$$m_{1n}(z) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i - z},$$

and

$$m_{2n}(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\mu_i - z}.$$

The local law states that $m_{1n}(z)$ and $m_{2n}(z)$ are close to those of $m_1(z)$ and $m_2(z)$, respectively.

Lemma 10. Suppose Assumption 1 holds true. Then for some small $\epsilon > 0$ and large $D \equiv D(\epsilon) > 0$, we have

$$\sup_{z\in \mathrm{S}}(|m_{1n}(z) - m_1(z)| + |m_{2n}(z) - m_2(z)|) \le n^{-1+\epsilon}.$$

Armed with the above lemma, we now head to prove Lemma 8.

From the proof of , we find that

$$f\left(\theta\left(\frac{d}{\sigma}\right)\right) = \left(\frac{d}{\sigma}\right)^{-2},\ d > \sigma c_n^{1/4},$$

where $f(x)$ is defined as

$$f(x) = xm_1(x)m_2(x).$$

Note that $f(x)$ is a continuously differentiable function. Therefore, we have with probability at least $1 - n^{-D}$

$$|f(\lambda_i/\sigma) - (d/\sigma)^{-2}| \le n^{-1/2+\epsilon}.$$

Xiucai Ding is currently a research associate at Duke University, and he will become a tenure-track Assistant Professor in the Department of Statistics, University of California Davis in the fall of 2020. He received his M.S. degree from the Courant Institute of Mathematical Sciences, New York University, New York, U.S., in 2014 and Ph.D. degree from the University of Toronto, Toronto, Canada in 2018. His main research interests include random matrix theory with applications in statistics, manifold learning, machine learning and deep learning, non-stationary time series analysis and statistical optimal transport theory.

Xianyi Chen received his PhD in Computer Science and Technology from Hunan University, China, in 2014. He is a visiting fellow of the Mathematics and Computer Science, The University of North Carolina at Pembroke, USA, in 2018-2019. He is currently a vice professor in the School of Computer and Software, Nanjing University of Information Science& Technology, China. His research interests include big data base information hiding, digital watermarking, cloud computing security and machine learning.

Mengling Zou is a software engineering student of Nanjing University of information engineering, Nanjing, China, and she is a visiting scholar of Department of Computer Science, University of Drbrecen, Debrecen, in 2019.9-2020.1. Her research interests focus on reversible data hiding and reversible mage transformation.

Guangxing Zhang received the M.S. degree from Chinese Academy of Sciences in 2011. Now, he is a software engineer at Nanjing Qisheng Cloud Information Technology Co., Ltd. His research interests mainly include remote sensing image processing and machine learning.