Bernoulli **26**(1), 2020, 387–417
https://doi.org/10.3150/19-BEJ1129

High dimensional deformed rectangular matrices with applications in matrix denoising

XIUCAI DING

Department of Statistics, University of Toronto, Toronto, ON, M5S 3G3, Canada.
E-mail: xiucai.ding@mail.utoronto.ca

We consider the recovery of a low rank $M \times N$ matrix S from its noisy observation $\tilde{S}$ in the high dimensional framework when M is comparable to N. We propose two efficient estimators for S under two different regimes. Our analysis relies on the local asymptotics of the eigenstructure of large dimensional rectangular matrices with finite rank perturbation. We derive the convergent limits and rates for the singular values and vectors for such matrices.

Keywords: matrix denoising; random matrices; rotation invariant estimation; singular value decomposition

1. Introduction

Matrix denoising is important in many scientific endeavors. They appear prominently in singal processing [36], image denoising [12], machine learning [37], statistics [13,14,16], empirical finance [6,20] and biology [31]. In these applications, researchers are interested in recovering the true deterministic matrix from a noisy observation. Consider that we can observe a noisy $M \times N$ data matrix $\tilde{S}_N$, where

$$\tilde{S}_N = S_N + X_N, \tag{1.1}$$

the deterministic matrix S_N is known as the signal matrix and X_N the noise matrix. In the classic framework where M is much smaller than N, the truncated singular value decomposition (TSVD) is the default technique, see for example [15]. This method recovers S_N with an estimator $\hat{S}_N = \sum_{i=1}^m \mu_i \tilde{u}_i \tilde{v}_i^*$ using the truncated singular value decomposition, where $m < \min\{M, N\}$ denotes the truncation level, $\mu_i, \tilde{u}_i, \tilde{v}_i, i = 1, 2, \ldots, m$ are the singular values and vectors of $\tilde{S}$. We usually need to provide a threshold γ to choose m and use the singular values only when $\mu_i \geq \gamma$. Two popular methods are the soft thresholding [11] and hard thresholding [13].

In recent years, the advance of technology has lead to the observation of massive scale data, where the dimension of the variable is comparable to the length of the observation. In this situation, the TSVD will lose its validity. To address this problem, in the present paper, we consider the matrix denoising problem (1.1) by assuming M is comparable to N and estimate S_N in the following two regimes:

Regime (1). S_N is of low rank and we have prior information that its singular vectors are sparse.

Regime (2). S_N is of low rank and we have no prior information on the singular vectors.

In Regime (1), S_N is called simultaneously low rank and sparse matrix. This type of matrix has been heavily used in biology. A typical example is from the study of gene expression data [23,31]. In [37], Yang, Ma and Buja also consider such problem but from a quite different perspective. They do not take the local behavior of singular values and vectors into consideration. Instead, they use an adaptive thresholding method to recover S_N in (1.1). In Regime (2), we are interested in looking at what is the best we can do in this case. A natural (and probably necessary) assumption is rotation invariance [5], as the only information we know about the singular vectors is orthonormality. It is notable that, in this case, our result coincides with the results proposed by Gavish and Donoho [14], where they consider the estimator from another perspective and restrict the estimator to be conservative (see Definition 3 in [14]).

In this paper, we will study the convergent limits and rates of the singular values and vectors for the sequence of matrices $\tilde{S}_N$ defined in (1.1). For the rest of the paper, we will omit the subscript N for convenience and write

$$\tilde{S} = S + X. \tag{1.2}$$

To avoid repetition, we summarize the technical assumptions of the noise matrix X.

Assumption 1.1. We assume X is a white noise matrix, where the entries x_{ij} of X are i.i.d. random variables such that

$$\mathbb{E}x_{ij} = 0, \qquad \mathbb{E}|x_{ij}|^2 = \frac{1}{N}.$$

Furthermore, we assume that for $l \in \mathbb{N}$, there exists some constant $C_l > 0$, such that

$$\mathbb{E}|\sqrt{N}x_{ij}|^l \leq C_l. \tag{1.3}$$

Denote the SVD of S as

$$S = UDV^* = \sum_{k=1}^{r} d_i u_i v_i^*, \tag{1.4}$$

where $D = \operatorname{diag}\{d_1, \ldots, d_r\}$, $U = (u_1, \ldots, u_r)$, $V = (v_1, \ldots, v_r)$, and where $u_i \in \mathbb{R}^M$, $v_i \in \mathbb{R}^N$ are orthonormal vectors and r is a fixed constant. We also assume $d_1 > d_2 > \cdots > d_r > 0$. Then (1.2) can be written as

$$\tilde{S} = X + UDV^*. \tag{1.5}$$

Throughout the paper, we are interested in the following setup

$$c_N := \frac{N}{M}, \qquad \lim_{N\to\infty} c_N = c \in (0, \infty). \tag{1.6}$$

It is well known that for the noise matrix X, the spectrum of XX^* satisfies the celebrated Marchenco–Pastur (MP) law [24] and the largest eigenvalue satisfies the Tracy–Widom (TW) distribution [35]. Specifically, denote $\lambda_i := \lambda_i(XX^*)$, $i = 1, 2, \dots, K$, where $K = \min\{M, N\}$, as the eigenvalues of XX^* in a decreasing fashion, we have that

$$\lambda_1 = \lambda_+ + O\big(N^{-2/3}\big), \qquad \lambda_+ = \big(1 + c^{-1/2}\big)^2, \tag{1.7}$$

holds with high probability. Furthermore, denote ξ_i, ζ_i as the singular vectors of X, for some large constant $C > 0$, with high probability, we have [9]

$$\max_k \big\{\big|\xi_i(k)\big|^2 + \big|\zeta_i(k)\big|^2\big\} = O\big(N^{-1}\big), \quad i \le C.$$

To sketch the behavior of $\tilde{S}$, we consider the case when $r = 1$ in (1.5). Assuming that the distribution of the entries of X is bi-unitarily invariant, Benaych-Georges and Nadakuditi established the convergent limits in [2] using free probability theory. Denote $\mu_i := \mu_i(\tilde{S}\tilde{S}^*)$, $i = 1, 2, \dots, K$, they proved that when $d > c^{-1/4}$, μ_1 would detach from the spectrum of the MP law and become an outlier. And when $d < c^{-1/4}$, μ_1 converges to λ_+ and sticks to the spectrum of the MP law. For the singular vectors, denote $\tilde{u}_i$, $\tilde{v}_i$ as the left and right singular vectors of $\tilde{S}$, $i = 1, 2, \dots, K$. They proved that when $d > c^{-1/4}$, $\tilde{u}_1$, $\tilde{v}_1$ would be concentrated on cones with axis parallel to u_1, v_1 respectively, and the apertures of the cones converged to some deterministic limits. And when $d < c^{-1/4}$, $\tilde{u}_1$, $\tilde{v}_1$ will be asymptotically perpendicular to u_1, v_1, respectively. We point out that similar results have been proved for the Wigner matrices with additive deformation and covaraince matrices with multiplication perturbation. For such results, we refer the readers to [1,4,7,17,18,28,30,32].

Our computation and proof rely on the isotropic local MP law [3,19,29]. These results say that the eigenvalue distribution of the sample covariance matrix XX^* is close to the MP law, down to the spectral scale containing slightly more than one eigenvalue. These local laws are formulated using the Green functions,

$$\mathcal{G}_1(z) := \big(XX^* - z\big)^{-1}, \qquad \mathcal{G}_2(z) := \big(X^*X - z\big)^{-1}, \qquad z = E + i\eta \in \mathbb{C}^+. \tag{1.8}$$

To illustrate our results and ideas, we give an overview of the present paper. As we have seen from [9,10], the self-adjoint linearization technique is quite useful in dealing with rectangular matrices. Hence, in a first step, we denote by

$$\begin{aligned} \tilde{H} &= \begin{bmatrix} 0 & z^{1/2}\tilde{S} \\ z^{1/2}\tilde{S}^* & 0 \end{bmatrix} = \begin{bmatrix} 0 & z^{1/2}X \\ z^{1/2}X^* & 0 \end{bmatrix} + \begin{bmatrix} 0 & z^{1/2}UDV^* \\ z^{1/2}VDU^* & 0 \end{bmatrix} \\ &= H + \mathbf{UDU}^*, \end{aligned} \tag{1.9}$$

where $\mathbf{D}$, $\mathbf{U}$ are defined as

$$\mathbf{D} := \begin{bmatrix} 0 & z^{1/2}D \\ z^{1/2}D & 0 \end{bmatrix}, \qquad \mathbf{U} := \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}. \tag{1.10}$$

Next we will give a heuristic description of our results. We will always denote $\mu_1 \ge \dots \ge \mu_K$, $K = \min\{M, N\}$ as the eigenvalues of $\tilde{S}\tilde{S}^*$ and $\tilde{u}_i$, $\tilde{v}_i$ as the singular vectors of $\tilde{S}$. And we denote $G(z)$ as the Green function of H. Consider $r=1$ in (1.5) and by a standard perturbation discussion (see Lemma 4.7), we find that μ_1 satisfies the equation $\det(\mathbf{U}^* G(\mu_1)\mathbf{U}+\mathbf{D}^{-1})=0$. Using the isotropic local law in [19], we find that (see Lemma 4.9) G has a deterministic limit Π when N is large enough. Heuristically, the convergent limit of μ_1 is determined by the equation $\det(\mathbf{U}^*\Pi(z)\mathbf{U}+\mathbf{D}^{-1})=0$. An elementary calculation shows that, when $d>c^{-1/4}$, $\mu_1 \to p(d)$, where $p(d)$ is defined in (2.6).

When $d>c^{-1/4}$, the largest eigenvalue μ_1 will detach from the bulk and become an outlier around its *classical location* $p(d)$. We would expect this happens under a scale of $N^{-1/3}$. This can be understood in the following ways: increasing d beyond the critical value $c^{-1/4}$, we expect μ_1 to become an outlier, where its location $p(d)$ is located at a distance greater than $O(N^{-2/3})$ from λ_+. By using mean value theorem, the phase transition will take place on the scale when

$$\left|d-c^{-1/4}\right| \geq O\big(N^{-1/3}\big). \tag{1.11}$$

When (1.11) happens, we also prove that

$$\mu_1 = p(d) + O\big(N^{-1/2}\big(d-c^{-1/4}\big)^{1/2}\big). \tag{1.12}$$

Below this scale, we would expect the spectrum of $\tilde{S}\tilde{S}^*$ to stick to that of XX^*. Especially, the largest eigenvalue μ_1 still has the Tracy–Widom distribution with the scale $N^{-2/3}$, which reads as

$$\mu_1 = \lambda_+ + O\big(N^{-2/3}\big). \tag{1.13}$$

For the singular vectors, when $d>c^{-1/4}$, we have $\langle u_1, \tilde{u}_1\rangle^2 \to a_1(d)$, $\langle v_1, \tilde{v}_1\rangle^2 \to a_2(d)$, where $a_1(d)$, $a_2(d)$ are deterministic functions of d and defined in (2.9). For the local behavior, we will use an integral representation of Greens functions (see (5.17)). Under the assumption that d_i's are well-separated and satisfy (1.11), we prove that

$$\langle u_1, \tilde{u}_1\rangle^2 = a_1(d) + O\big(N^{-1/2}\big), \qquad \langle v_1, \tilde{v}_1\rangle^2 = a_2(d) + O\big(N^{-1/2}\big). \tag{1.14}$$

Below the scale of (1.11), we prove that

$$\langle u_1, \tilde{u}_1\rangle^2 = O\big(N^{-1}\big), \qquad \langle v_1, \tilde{v}_1\rangle^2 = O\big(N^{-1}\big). \tag{1.15}$$

Armed with (1.12), (1.13), (1.14) and (1.15), we can go to the matrix denoising problem (1.5) under the two different regimes. In the first regime, we assume there exists sparse structure of the singular vectors, in the case when $d>c^{-1/4}$, we would expect $\tilde{u}_1$, $\tilde{v}_1$ to be sparse as well. Hence, $\tilde{S}$ will be of sparse structure. Therefore, by suitably choosing a submatrix of $\tilde{S}$ and doing SVD for the submatrix, we can get an estimator for the singular vectors. Our novelty is to truncate singular values and vectors simultaneously. For the estimation of singular values, we can reverse (1.12) to get the estimator for d. For the singular vectors, based on (1.15), the truncation level should be much larger than $N^{-1/2}$ and we will use K-means clustering algorithm to choose such level. However, when $d<c^{-1/4}$, we can estimate nothing according to (1.13) and (1.15).

In the second regime, as we have no prior information whatsoever on the true eigenbasis of S, the only possibility is to use the eigenbasis of $\tilde{S}$. This is equivalent to the assumption of rotation invariance. We will propose a consistent rotation invariant estimator (RIE) $\Xi(\tilde{S})$, which satisfies the following condition,

$$\Omega_1 \Xi(\tilde{S})\Omega_2 = \Xi(\Omega_1 \tilde{S} \Omega_2), \tag{1.16}$$

where Ω_1, Ω_2 are orthogonal (rotation) matrix in $\mathbb{R}^M$, $\mathbb{R}^N$, respectively. Before concluding this section, we list our main contributions of this paper:

(i). We systematically study the local behavior of the singular values and vectors for finite rank perturbation of large dimensional rectangular matrices of model (1.5). We compute the convergent limits and rates for them.

(ii). We provide two efficient estimators for the matrix denoising model (1.5) under two different regimes. We provide practical algorithms to compute the estimators. For the sparse estimation, as far as we know, our paper is the first one to truncate the singular values and vectors simultaneously.

This paper is organized as follows. In Section 2, we give the main results of this paper. In Section 3, we propose the estimators for (1.5) under two regimes. In Section 4, we record the basic tools for the proof of the main theorems. In Section 5, we prove the main theorems listed in Section 2.

Conventions. For two quantities a_N and b_N depending on N, the notation $a_N = O(b_N)$ means that $|a_N| \leq C|b_N|$ for some positive constant $C > 0$, and $a_N = o(b_N)$ means that $|a_N| \leq c_N|b_N|$ for some positive constants $c_N \to 0$ as $N \to \infty$. We also use the notation $a_N \sim b_N$ if $a_N = O(b_N)$ and $b_N = O(a_N)$. We define the minimum of any two reals a, b by $a \wedge b$. For any matrix A, we denote by A^* as the transpose of A and $\|A\|_F$ the Frobenius norm of A. We will also use $\boldsymbol{\sigma}(H)$ to denote the spectrum for any square matrix H. And for any rectangular matrix S we use $\sigma_i(S)$ to denote its i-th largest singular value.

2. Main results

Throughout the paper, we always use ϵ_1 for a small constant and D_1 for a large constant. Denote $\mathcal{R} := \{1, 2, \ldots, r\}$ and $\mathcal{O}$ as a subset of of $\mathcal{R}$ by

$$\mathcal{O} := \left\{i : d_i \geq c^{-1/4} + N^{-1/3+\epsilon_0}\right\}, \quad \epsilon_0 > \epsilon_1 \text{ is a small constant,} \tag{2.1}$$

and the number of outlier singular values as

$$k^+ = |\mathcal{O}|. \tag{2.2}$$

Our results can be extended to a more general domain by denoting $\mathcal{O}' := \{i : d_i \geq c^{-1/4} + N^{-1/3}\}$. We will not pursue this generalization. For more details, we refer to [4]. For any subset $A \subset \mathcal{O}$, we define the projections on the left and right singular subspace of $\tilde{S}$ by

$$\mathbf{P}_l := \sum_{i \in A} \tilde{u}_i \tilde{u}_i^*, \qquad \mathbf{P}_r := \sum_{j \in A} \tilde{v}_j \tilde{v}_j^*. \tag{2.3}$$

We also need the non-overlapping condition, which was firstly introduced in [4].

Definition 2.1. For $i = 1, 2, \ldots, M$, the non-overlapping condition is written as

$$\nu_i(A) \geq \left(d_i - c^{-1/4}\right)^{-1/2} N^{-1/2+\epsilon_0}, \tag{2.4}$$

where ϵ_0 is defined in (2.1) and $\nu_i(A)$ is defined by

$$\nu_i(A) := \begin{cases} \min_{j \notin A} |d_i - d_j|, & \text{if } i \in A, \\ \min_{j \in A} |d_i - d_j|, & \text{if } i \notin A. \end{cases} \tag{2.5}$$

With the above preparation, we state our main results of the singular values of $\tilde{S}$. Denote

$$p(d) = \frac{(d^2+1)(d^2+c^{-1})}{d^2}. \tag{2.6}$$

Recall $\tilde{S}$ defined in (1.5) and μ_i are the eigenvalues of $\tilde{S}\tilde{S}^*$.

Theorem 2.2. *Under Assumption 1.1 and the assumption of (1.6), for $i = 1, 2, \ldots, k^+$, where k^+ is defined in (2.2), there exists some large constant $C > 1$ such that $C\epsilon_1 < \epsilon_0$, when N is large enough, with $1 - N^{-D_1}$ probability, we have*

$$\left|\mu_i - p(d_i)\right| \leq N^{-1/2+C\epsilon_0}\left(d_i - c^{-1/4}\right)^{1/2}, \tag{2.7}$$

where $p(d_i)$ is defined in (2.6). Moreover, for $j = k^+ + 1, \ldots, r$, we have

$$|\mu_j - \lambda_+| \leq N^{-2/3+C\epsilon_0}, \tag{2.8}$$

where λ_+ is defined in (1.7).

The above theorem gives precise location of the outlier singular values and the extremal non-outlier singular values. For the outliers, they locate around their classical locations $p(d_i)$ and for the non-outliers, they locate around λ_+. The results of the singular vectors are given by the following theorem. Denote

$$a_1(d) = \frac{d^4 - c^{-1}}{d^2(d^2 + c^{-1})}, \qquad a_2(d) = \frac{d^4 - c^{-1}}{d^2(d^2+1)}. \tag{2.9}$$

Theorem 2.3. *Under Assumption 1.1 and the assumptions of (1.6) and (2.4), for all $i, j = 1, 2, \ldots, r$, there exists some constant $C > 0$, with $1 - N^{-D_1}$ probability, when N is large enough, we have*

$$\left|\langle u_i, \mathbf{P}_l u_j\rangle - \delta_{ij}\mathbf{1}(i \in A)a_1(d_i)\right| \leq N^{\epsilon_1} R(i, j, A, N), \tag{2.10}$$

$$\left|\langle v_i, \mathbf{P}_r v_j\rangle - \delta_{ij}\mathbf{1}(i \in A)a_2(d_i)\right| \leq N^{\epsilon_1} R(i, j, A, N), \tag{2.11}$$

where $a_1(d), a_2(d)$ *are defined in* (2.9) *and* $R(i, j, A, N)$ *is defined as*

$$
\begin{aligned}
R(i, j, A, N) := N^{-1/2}\Bigg[&\frac{\mathbf{1}(i \in A, j \in A)}{(d_i - c^{-1/4})^{1/2} + (d_j - c^{-1/4})^{1/2}} \\
&+ \mathbf{1}(i \in A, j \notin A)\frac{(d_i - c^{-1/4})^{1/2}}{|d_i - d_j|} \\
&+ \mathbf{1}(i \notin A, j \in A)\frac{(d_j - c^{-1/4})^{1/2}}{|d_i - d_j|}\Bigg] \\
&+ N^{-1}\Bigg[\bigg(\frac{1}{\nu_i} + \frac{\mathbf{1}(i \in A)}{|d_i - c^{-1/4}|}\bigg)\bigg(\frac{1}{\nu_j} + \frac{\mathbf{1}(j \in A)}{|d_j - c^{-1/4}|}\bigg)\Bigg].
\end{aligned}
$$

Moreover, fix a small constant $\tau > 0$, *for* $k^+ + 1 \leq j \leq (1 - \tau)K$, *denote* $\kappa_j^d := N^{-2/3}(j \wedge (K + 1 - j))^{2/3}$, *we have*

$$
\left|\langle u_i, \tilde{u}_j\rangle^2\right| \leq \frac{N^{C\epsilon_0}}{N((d_i - c^{-1/4})^2 + \kappa_j^d)}, \quad i = 1, 2, \ldots, r, \tag{2.12}
$$

and

$$
\left|\langle v_i, \tilde{v}_j\rangle\right|^2 \leq \frac{N^{C\epsilon_0}}{N((d_i - c^{-1/4})^2 + \kappa_j^d)}, \quad i = 1, 2, \ldots, r. \tag{2.13}
$$

Furthermore, if $c \neq 1$, (2.12) *and* (2.13) *hold for all* $j = k^+ + 1, \ldots, M$.

Remark 2.4. The assumption $j \leq (1 - \tau)K$ ensures that $\mu_j \geq \delta$, for some constant $\delta > 0$. When $c \neq 1$, it is guaranteed as we will see from Lemma 4.12 that $\mu_j \geq (1 - c^{-1/2})^2/2$. We need $\mu_j \geq \delta$ for the technical purpose of the application of the local laws.

Next, we will give some examples to illustrate our results. We assume that $c \neq 1$.

Example 2.5. (1) Consider the right singular vectors and let $A = \{i\}$, we have

$$
\left|\langle v_i, \tilde{v}_i\rangle^2 - a_2(d_i)\right| \leq N^{\epsilon_1}\left[\frac{1}{N^{1/2}(d_i - c^{-1/4})^{1/2}} + \frac{1}{N\nu_i^2(d_i - c^{-1/4})^2}\right].
$$

This implies that, the cone concentration of the singular vector holds if $i \in \mathcal{O}$ and the non-overlapping condition (2.4) holds. Furthermore, if d_i is well-separated from both the critical point $c^{-1/4}$ and the other outliers, the error bound is of order $\frac{1}{\sqrt{N}}$.

(2) Let $A = \{i\}$ and for $1 \leq j \neq i \leq r$, we have

$$
\left|\langle v_j, \tilde{v}_i\rangle^2\right| \leq \frac{N^{\epsilon_1}}{N(d_i - d_j)^2}.
$$

Hence, if $|d_i - d_j| = O(1)$, then $\tilde{v}_i$ will be completely delocalized in any direction orthogonal to v_i.

(3) If $i \in \mathcal{O}$, $j \notin \mathcal{O}$, then we have

$$\left|\langle v_i, \tilde{u}_j\rangle^2\right| \le \frac{N^{C\epsilon_0}}{N((d_i - c^{-1/4})^2 + \kappa_j^d)}.$$

Hence, when $|d_i - c^{-1/4}| = O(1)$ or $\kappa_j^d = O(1)$, $\tilde{u}_j$ will be completely delocalized in the direction of v_i. The first case reads as μ_i is an outlier and the second case as that μ_j is in the bulk of the spectrum of $\tilde{S}\tilde{S}^*$.

Before concluding this section, we use the following figure to illustrate the accuracy of the proposed bounds in (2.7), (2.10) and (2.11). We consider the rank one perturbation $\tilde{S} = duv^* + X$, where X is a Gaussian random matrix with mean zero and variance $1/N$ and u, v are sparse vectors generated from the R package *R1magic*.

To avoid the influence of the constant, we consider the ratio between the empirical bound and dominated part, that is, for $d > c^{-1/4}$, we will consider

$$R_1 = \Phi_1\big|\mu_1 - p(d)\big|, \qquad R_2 = \Phi_2\big|\langle u, \tilde{u}_1\rangle^2 - a_1(d)\big|, \qquad R_3 = \Phi_2\big|\langle v, \tilde{v}_1\rangle^2 - a_2(d)\big|,$$

where $\Phi_1 := \sqrt{N}(d - c^{-1/4})^{-1/2}$ and $\Phi_2 := \sqrt{N(d - c^{-1/4})}$. We consider the cases $c = 0.5$ and $c = 2$, and choose $d = 2$. For each N, we record the averaged ratios for $R_i, i = 1, 2, 3$, using 1000 repetitions and plot these ratios for a variety of choices (in total 181) of N between 200 and 2000. We can conclude from Figure 1 that these ratios are around some fixed constants independent of N.

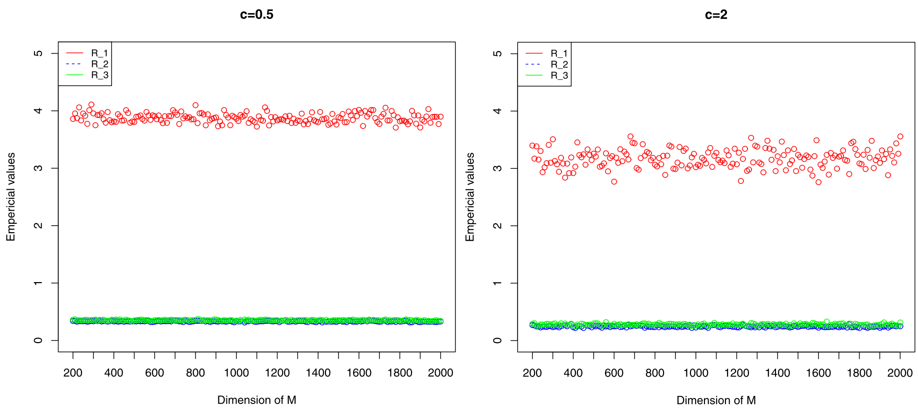

Figure 1. We can see from the above figure that R_1, R_2, R_3 are independent of N. Further, the left and right singular vectors have the same bounds.

3. Statistical applications

3.1. Sparse estimation

In the present application, we study the denoising problem (1.2), where S is sparse in the sense that the nonzero entries are assumed to be confined on a block. We assume that u_i, v_i are sparse and introduce the following definition to precisely describe the sparsity.

Definition 3.1. For any vector $v \in \mathbb{R}^N$, v is a sparse vector if there exists a subset $\mathbb{N}^* \subset \{1, 2, \ldots, N\}$ with $|N^*| = O(1)$, such that

$$\left|v(i)\right| = \begin{cases} O(1), & i \in \mathbb{N}^*; \\ O\left(N^{-1/2}\right), & \text{otherwise.} \end{cases}$$

Next we will propose an estimator for S by estimating the singular values and vectors separately. As can been see from Theorem 2.2, we can estimate the true outlier singular values from their corresponding sample values. To ease our discussion, we impose the following stronger assumptions on the outlier singular values of S.

Assumption 3.2. For $i, j = 1, 2, \ldots, k^+$, we assume that there exists some constant $\delta > 0$, such that

$$d_i > c^{-1/4} + \delta, \qquad |d_i - d_j| \geq \delta, \quad i \neq j.$$

Note that the above assumption is a stronger version of (2.1) and widely used in the practical applications [14,25–27]. We first estimate the number k^+ of outlier singular values. In [26], k^+ is referred as the *effective number of identifiable signals* and the author provided an information theoretic estimator by minimizing the Akaike Information Criterion (AIC). Furthermore, some other useful statistics have been proposed to effectively estimate the number of spikes in the spiked covariance matrix model, for instance the differences between consecutive eigenvalues in [27]. By Theorem 2.2, when $i \leq k^+$, we expect μ_i/μ_{i+1} will be away from one and when $i > k^+$, it will be close to one. In the present paper, we will employ the ratios of consecutive sample singular values [21] as our statistic. For $\tau := O(N^{-\alpha})$ satisfying

$$0 < \alpha < \frac{2}{3}, \tag{3.1}$$

we denote (Recall $K = \min\{M, N\}$.)

$$q = \arg\max_i \{1 \leq i \leq K : \mathcal{R}_i > 1 + \tau\}, \qquad \tau > 0, \mathcal{R}_i = \frac{\mu_i}{\mu_{i+1}}. \tag{3.2}$$

We summarize the property of q as the following proposition and its proof can be found in the supplementary material [8].

Proposition 3.3. *Under the assumptions of Theorem 2.2 and Assumption 3.2, for some* $\tau = O(N^{-\alpha})$ *satisfying* (3.1), *we have that*

$$\mathbb{P}\big(q = k^+\big) = 1 - o(1).$$

In practice, for the choice of τ, we employ the automatic calibration procedure of [27], Section 4. The idea is to use the ratio of the first two largest eigenvalues of a Wishart matrix, that is, an $M \times N$ random Gaussian matrix satisfying Assumption 1.1. Indeed, we need to search the eigenvalue index such that the ratio of two consecutive eigenvalues of $\tilde{S}\tilde{S}^*$ is much larger than $1+\tau$ corresponding to that of XX^*. In detail, we will use the following procedure to calibrate τ.

(1) Generate a sequence (say 1000) of $M \times N$ random Gaussian matrices Z_k, $k = 1, 2, \ldots, 1000$ satisfying Assumption 1.1. Calculate the ratios of the first and second eigenvalue of $Z_k Z_k^*$ and write them as $\mathcal{R}_{1,k}$, $k = 1, 2, \ldots, 1000$.
(2) For a given large probability β, (say $\beta = 0.98$ as suggested by [27]), find the value τ such that

$$\frac{\#\{k : \mathcal{R}_{1,k} - 1 \leq \tau\}}{1000} = \beta.$$

For $c = 2$, we find that $\tau = 0.0577$ for $M = 300$ and $\tau = 0.0372$ for $M = 500$. These will be used later for our simulation studies.

With the above notations, we provide the *stepwise SVD* Algorithm 1 to recover S in (1.2). As u_i, v_i are sparse, we need to find a submatrix of $\tilde{S}$ by a suitable truncation. Instead of simply truncating the singular values [14,37], we truncate the singular values and vectors simultaneously.

Algorithm 1 Stepwise SVD

Do SVD for $\tilde{S} = \sum_{i=1}^K \mu_i \tilde{u}_i \tilde{v}_i^*$, and do the initialization $\tilde{S}_1 = \tilde{S} = \sum t_i^1 \tilde{u}_i^1 (\tilde{v}_i^1)^*$.
while $1 \leq j \leq q$ **do**
  $\hat{d}_j = p^{-1}((t_1^j)^2)$, where $p^{-1}(x)$ is the inverse of the function defined in (2.6).
  Use two thresholds $\alpha_{u_j} \gg \frac{1}{\sqrt{M}}$, $\alpha_{v_j} \gg \frac{1}{\sqrt{N}}$, and denote

$$I_j := \big\{1 \leq k \leq M : \big|\tilde{u}_1^j(k)\big| \geq \alpha_{u_j}\big\}, \qquad J_j := \big\{1 \leq k \leq N : \big|\tilde{v}_1^j(k)\big| \geq \alpha_{v_j}\big\}. \tag{3.3}$$

  Do SVD for the block matrix $\tilde{S}_b = \tilde{S}_j[I_j, J_j] = \sum \rho_i u_i^j (v_i^j)^*$.
  Assume $I_j = \{k_1, \ldots, k_j\}$, construct $\hat{u}_j$ by letting

$$\hat{\mu}_j(k_j) = \begin{cases} \mu_1^j(j), & k_j \in I_j, \\ 0, & \text{otherwise}. \end{cases}$$

  Similarly, we can construct $\hat{v}_j$.
  Let $\tilde{S}_{j+1} = \tilde{S}_j - \hat{d}_j \hat{u}_j \hat{v}_j^*$ and do SVD for $\tilde{S}_{j+1} = \sum t_i^{j+1} \tilde{u}_i^{j+1} (\tilde{v}_i^{j+1})^*$.
end while
Denote $\hat{S} = \sum_{k=1}^q \hat{d}_k \hat{u}_k \hat{v}_k^*$ as our estimator.

Algorithm 1 provides us a way to recover S stepwise. We first estimate d_1, u_1, v_1 using the estimation $\hat{d}_1$, $\hat{u}_1$, $\hat{v}_1$, then d_2, u_2, v_2 by analyzing $\tilde{S} - \hat{d}_1 \hat{u}_1 \hat{v}_1^*$. In each step, we only need to look at the largest singular value and its associated singular vectors. It is notable that, we drop all the eigenvalues μ_i of $\tilde{S}\tilde{S}^*$ when $i < q$ and

$$\hat{d}_i = \mathbf{1}(i \geq q) p^{-1}(\mu_i). \tag{3.4}$$

Our methodology relies on truncating singular values and vectors simultaneously. As illustrated in (3.3), the thresholds α_u and α_v play the key roles in recovering the sparse structure of the singular vectors. It will be proved in Section 2 that any threshold satisfying (3.3) should work when N is sufficiently large. In the finite sample framework (when N is not quite large), we employ the K-means algorithm [16], Section 10.3.1, to stabilize the recovery of the sparse structure of S. The reason behind is, the entries in the singular vectors $\tilde{u}_i$, $\tilde{v}_i$ can be well classified into two categories. Denote the index sets C_u^j, C_v^j getting from the K-means algorithm, where they satisfy

$$\min_{k \in C_u^j} \left| \tilde{u}_1^j(k) \right| \gg \frac{1}{\sqrt{M}}, \qquad \min_{k \in C_v^j} \left| \tilde{v}_1^j(k) \right| \gg \frac{1}{\sqrt{N}}. \tag{3.5}$$

We now replace (3.3) with the following step:

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

$$\begin{aligned} I_j &:= \left\{ 1 \leq k \leq M : k \in C_u^j \right\}, \\ J_j &:= \left\{ 1 \leq k \leq N : k \in C_v^j \right\}, \end{aligned} \tag{3.6}$$

  where C_u^j, C_v^j satisfy (3.5).

Next, we summarize the theoretical properties of Algorithm 1 as the following theorem and leave its proof into the supplementary material [8].

Theorem 3.4. *With prior information that u_i, v_i are sparse in the sense of Definition* 3.1, *under the assumptions of Theorems* 2.2 *and* 2.3, *and Assumption* 3.2, *there exists some $C > 0$, with $1 - o(1)$ probability, for the estimator $\hat{S}$ getting from Algorithm* 1, *we have*

$$\|\hat{S} - S\|_F \leq N^{-1/2 + C\epsilon_0} + \sqrt{\sum_{i=k^+ + 1}^{r} d_i^2}.$$

Before concluding this subsection, we compare our method with other different algorithms. In [37], the authors proposed another algorithm from a quite different perspective. They did not take the properties of the singular values and vectors of $\tilde{S}$ into consideration. Instead, they used iterative thresholding on the rows of $\tilde{S}$ to get an estimator. The algorithm is called *sparse SVD*. Their algorithm can be regarded as the extension of TSVD on the submatrix of $\tilde{S}$.

Table 1. Comparison of the algorithms. We choose $r = 2$, $c = 2$, $d_1 = 7$, $d_2 = 4$ in (1.5). The noise matrix X is Gaussian. In the table, sparsity is defined as the ratio of non-zero entries and length of the vector and we assume that u_i, v_i, $i = 1, 2$ have the same sparsity. We highlight the smallest error norm

	$M = 300$			$M = 500$		
	Sparsity	L^2 error norm	Std	Sparsity	L^2 error norm	Std
SWSVD	0.05	**0.043**	0.175	0.05	**0.045**	0.189
	0.1	**0.614**	0.178	0.1	**0.6**	0.16
	0.2	**0.822**	0.126	0.2	**0.825**	0.137
	0.45	**1.1**	0.114	0.45	**1.09**	0.09
SSVD	0.05	4.01	0.002	0.05	4.01	0.002
	0.1	4.01	0.004	0.1	4.02	0.002
	0.2	4.04	0.004	0.2	4.03	0.004
	0.45	4.06	0.005	0.45	4.08	0.004
TSVD	0.05	53.9	6.872	0.05	53.75	6.63
	0.1	53.72	6.63	0.1	53.38	6.71
	0.2	52.33	7.01	0.2	52.2	6.65
	0.45	51.043	2.49	0.45	52.4	4.3

We use Table 1 to compare the results of three algorithms, our stepwise SVD (SWSVD), the sparse SVD (SSVD) proposed by [37] and the truncated SVD (TSVD). For the implementation of SSVD, we use the *ssvd* package in R which is contributed by the first author of [37]. From Table 1, we find that our method outperforms both the SSVD and TSVD in all the cases. Furthermore, the standard deviation is small, which implies that our estimation is quite stable.

3.2. Rotation invariant estimation

This subsection is devoted to recovering S in (1.2) assuming that no prior information about S is available. In this regime, we will consider the rotation invariant estimator (RIE) satisfying (1.16). We conclude from [5] that any RIE shares the same singular vectors as $\tilde{S}$. To construct the optimal estimator, we use the Frobenius norm as our loss function. Denote $\hat{S} = \Xi(\tilde{S})$, we have

$$\|S - \hat{S}\|_F^2 = \operatorname{Tr}(S - \hat{S})(S - \hat{S})^*. \tag{3.7}$$

Therefore, the form of the RIE can be written in the following way

$$\hat{S} = \underset{H \in \mathcal{M}(\tilde{U}, \tilde{V})}{\arg\min} \|H - S\|_F, \tag{3.8}$$

where $\mathcal{M}(\tilde{U}, \tilde{V})$ is the class of $M \times N$ matrices whose left singular vectors are $\tilde{U}$ and right singular vectors are $\tilde{V}$. Suppose $\hat{S} = \sum_{i=1}^K \eta_k \tilde{u}_k \tilde{v}_k^*$, denote $\mu_{k_1 k} = \langle u_{k_1}, \tilde{u}_k \rangle$, $\nu_{k_1 k} = \langle v_{k_1}, \tilde{v}_k \rangle$, then by an elementary computation, we find

$$\|S-\hat{S}\|_F^2=\sum_{k=1}^{r}(d_k^2+\eta_k^2)-2\sum_{k=1}^{r}d_k\eta_k\mu_{kk}\nu_{kk}+\sum_{k=r+1}^{K}\eta_k^2$$
$$-2\sum_{k_1\neq k_2}^{r}d_{k_1}\eta_{k_2}\mu_{k_1k_2}\nu_{k_1k_2}-2\sum_{k_1=r+1}^{K}\sum_{k_2=1}^{r}\eta_{k_1}d_{k_2}\mu_{k_2k_1}\nu_{k_2k_1}. \tag{3.9}$$

Therefore, $\hat{S}$ is optimal if

$$\eta_k=\langle\tilde{u}_k,S\tilde{v}_k\rangle=\sum_{k_1=1}^{r}d_{k_1}\mu_{k_1k}\nu_{k_1k},\quad k=1,\ldots,K. \tag{3.10}$$

In the present paper, we use the following estimator for η_k and will prove its consistency in Section 2. Recall (3.2), the estimator is denoted as

$$\hat{\eta}_k=\begin{cases}\hat{d}_k a_1(\hat{d}_k)a_2(\hat{d}_k), & k\le q;\\ 0, & k>q,\end{cases} \tag{3.11}$$

where $\hat{d}_k=p^{-1}(\mu_k)$ and $a_1(x),a_2(x)$ are defined in (2.9). Denote

$$\hat{\mathcal{S}}=\sum_{k=1}^{q}\hat{\eta}_k\tilde{u}_k\tilde{v}_k^*. \tag{3.12}$$

It is notable that the convergent limits for the shrinkage $\hat{\eta}_k$ and MSE for $\hat{\mathcal{S}}$ have already been computed in [25]. We next summarize the theoretical properties of our estimators as the following theorem. Its proof can be found in the supplementary material [8].

Theorem 3.5. (1). *Under the assumptions of Theorem 2.2 and 2.3, there exists some large constant $C>0$ and small constant $\tau>0$, with $1-o(1)$ probability, we have $\hat{\eta}_k\to\eta_k$, $k=1,2,\ldots,K$. Furthermore, for $1\le k\le(1-\tau)K$, we have*

$$|\hat{\eta}_k-\eta_k|\le\mathbf{1}(k\le k^+)N^{-1/2+C\epsilon_0}+\mathbf{1}(k>k^+)N^{-1+C\epsilon_0}. \tag{3.13}$$

Moreover, when $c\neq 1$, (3.13) holds for all $k=1,\ldots,K$. (2). *When $c\neq 1$, there exists some constant $C>0$, with $1-o(1)$ probability, for $\hat{\mathcal{S}}$ defined in (3.12), we have*

$$\|\hat{\mathcal{S}}-S\|_F^2\le\sum_{i=1}^{r}d_i^2-\sum_{i=1}^{k+}\big(d_ia_1(d_i)a_2(d_i)\big)^2+N^{-1/2+C\epsilon_0}.$$

Figure 2 are two examples of the estimations of η_k. From the graph, we find that our estimator $\hat{\eta}_k$ is quite accurate. Figure 3 records the relative improvement in average loss (RIAL) compared

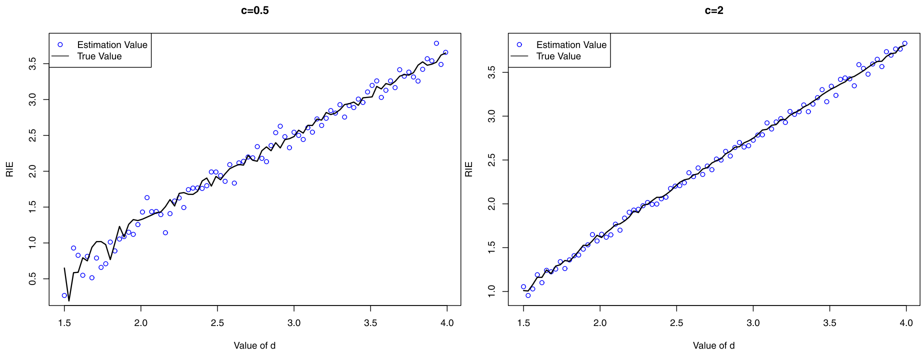

Figure 2. RIE. We choose $r = 1$ and $M = 300$ for (1.5). We estimate η_1 using the estimator (3.11) for $c = 0.5, 2$ with different values of d. The entries of X are Gaussian random variables and the singular vectors satisfy the exponential distribution with rate 1.

to TSVD, where the RIAL is defined as

$$\mathrm{RIAL}(N) = 1 - \frac{\mathbb{E}\|\hat{\mathcal{S}} - S\|_F}{\mathbb{E}\|S_T - S\|_F}, \tag{3.14}$$

and where S_T is the TSVD estimation and $\hat{\mathcal{S}}$ the RIE. We conclude from the figure that our method provides better estimation compared to the TSVD. Similar results have been shown for the estimation of covariance matrices by Ledoit and Péché in [22].

Remark 3.6. In [14], Donoho and Gavish get similar results from the perspective of optimal shrinkage. However, they need two more assumptions: (1). they drop the last two error terms in

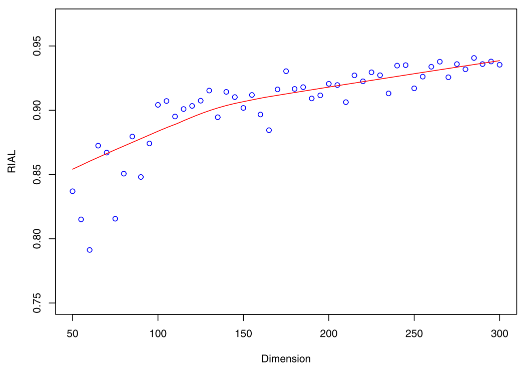

Figure 3. RIE compared to TSVD. We choose $r = 1$, $d = 4$, $c = 2$ in (1.2). X is a random Gaussian matrix and the entries of the singular vectors satisfy the exponential distribution with rate 1. We perform 1000 Monte-Carlo simulations for each M to simulate the RIAL defined in (3.14). The red line indicates the increasing trend as M increases.

(3.9) by assuming they are small enough (see Lemma 4 in their paper); (2). their estimators are assumed to be conservative, where they assume the shrinker vanishes when the sample singular values are below λ_+ defined in (1.7), that is, for some constant $\gamma > 0$,

$$\eta_k = 0, \quad \text{when } \mu_k \le \lambda_+ + \gamma.$$

However, we find that the estimator defined in (3.11) can still be consistent even without these assumptions.

4. Basic tools

In this section, we introduce some notations and tools which will be used in this paper. Recall that the empirical spectral distribution (ESD) of an $n \times n$ symmetric matrix H is defined as

$$F_H^{(n)}(\lambda) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\lambda_i(H) \le \lambda\}}.$$

We define the typical domain for $z = E + i\eta$ by

$$\mathbf{D}(\tau) \equiv \mathbf{D}(\tau, N) := \left\{ z \in \mathbb{C}^+ : \tau \le E \le \tau^{-1}, N^{-1+\tau} \le \eta \le \tau^{-1} \right\}, \tag{4.1}$$

where $\tau > 0$ is a small constant. Recall (1.6), we assume that $\tau < c_N < \tau^{-1}$.

Definition 4.1. The Stieltjes transform of the ESD of $X^* X$ is given by

$$m_2(z) \equiv m_2^{(N)}(z) := \int \frac{1}{x - z} dF_{X^*X}^{(N)}(x) = \frac{1}{N} \sum_{i=1}^{N} (\mathcal{G}_2)_{ii}(z) = \frac{1}{N} \operatorname{Tr} \mathcal{G}_2(z),$$

where $\mathcal{G}_2(z)$ is defined in (1.8). Similarly, we can also define $m_1(z) := M^{-1} \operatorname{Tr} \mathcal{G}_1(z)$.

Denote $m_{1c}(z) := \lim_{N \to \infty} m_1(z)$, $m_{2c}(z) := \lim_{N \to \infty} m_2(z)$ be the Stieltjes transforms of limiting spectral distributions of $m_1(z)$, $m_2(z)$. Using the identity $m_1(z) = -\frac{1-c_N}{z} + c_N m_2(z)$, we have

$$m_{1c}(z) = \frac{c-1}{z} + c m_{2c}(z). \tag{4.2}$$

Definition 4.2. For X satisfying (1.3), under the assumption (1.6), the ESD of XX^* converges weakly to the Marchenko–Pastur (MP) law as $N \to \infty$ [24]:

$$\rho_{1c}(x)\, dx = \frac{c}{2\pi} \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{x} dx, \qquad \lambda_\pm = \left(1 \pm c^{-\frac{1}{2}}\right)^2. \tag{4.3}$$

The Stieltjes transform of the MP law $m_{1c}(z)$ has the closed form expression (see (1.2) of [33])

$$m_{1c}(z) = \frac{1 - c^{-1} - z + i\sqrt{(\lambda_+ - z)(z - \lambda_-)}}{2zc^{-1}}. \tag{4.4}$$

Remark 4.3. From (4.2), we have that $m_2(z)$ converges to $m_{2c}(z)$ as $N \to \infty$, where

$$m_{2c}(z) = \frac{c^{-1}-1}{z} + c^{-1}m_{1c}(z) = \frac{c^{-1}-1-z+i\sqrt{(\lambda_+-z)(z-\lambda_-)}}{2z}. \tag{4.5}$$

It is notable that

$$-z^{-1}\big(1+m_{2c}(z)\big)^{-1} = m_{1c}(z). \tag{4.6}$$

Recall (1.9) and $G(z) = (H-z)^{-1}$, by Schur's complement [19], it is easy to check that

$$G(z) = \begin{pmatrix} \mathcal{G}_1(z) & z^{-1/2}\mathcal{G}_1(z)X \\ z^{-1/2}X^*\mathcal{G}_1(z) & \mathcal{G}_2(z) \end{pmatrix}, \tag{4.7}$$

for $\mathcal{G}_{1,2}$ defined in (1.8). Denote the index sets $\mathcal{I}_1 := \{1,\dots,M\}$, $\mathcal{I}_2 := \{M+1,\dots,M+N\}$, $\mathcal{I} := \mathcal{I}_1 \cup \mathcal{I}_2$. Then we have

$$m_1(z) = \frac{1}{M}\sum_{i\in\mathcal{I}_1} G_{ii}, \qquad m_2(z) = \frac{1}{N}\sum_{\mu\in\mathcal{I}_2} G_{\mu\mu}.$$

Similarly, we denote $\tilde{G}(z) = (\tilde{H}-z)^{-1}$, where $\tilde{H}$ is defined in (1.9). Next we introduce the spectral decomposition of $\tilde{G}(z)$. By (4.7), we have

$$\tilde{G}(z) = \sum_{k=1}^{K} \frac{1}{\mu_k - z}\begin{pmatrix} \tilde{u}\tilde{u}_k^* & z^{-1/2}\sqrt{\mu_k}\tilde{u}_k\tilde{v}_k^* \\ z^{-1/2}\sqrt{\mu_k}\tilde{v}_k\tilde{u}_k^* & \tilde{v}_k\tilde{v}_k^* \end{pmatrix}. \tag{4.8}$$

As we have seen in (2.6), the function $p(d)$ plays a key role in describing the convergent limits of the outlier singular values of $\tilde{S}$. An elementary computation yields that $p(d)$ attains its global minimum when $d = c^{-1/4}$ and $p(c^{-1/4}) = \lambda_+$, and

$$p'(x) \sim \big(x - c^{-1/4}\big). \tag{4.9}$$

To precisely locate the outlier singular values of $\tilde{S}$, we need to analyze

$$T^s(x) := \prod_{i=1}^{s}\big(xm_{1c}(x)m_{2c}(x) - d_i^{-2}\big). \tag{4.10}$$

By (4.4) and (4.5), when $x \geq \lambda_+$, we have

$$xm_{1c}(x)m_{2c}(x) = \frac{x-(1+c^{-1})-\sqrt{(x+c^{-1}-1)^2-4c^{-1}x}}{2c^{-1}}. \tag{4.11}$$

Next we collect the preliminary results of the properties of $T^s(x)$, whose proof will be provided in the supplementary material [8].

Lemma 4.4. *Suppose $d_1 > d_2 > \cdots > d_s > c^{-1/4}$, then we have that there exist s solutions of $T^s(x) = 0$ and they are $p_i := p(d_i)$, $i = 1, 2, \dots, s$, write*

$$T^s(p_i) = 0. \tag{4.12}$$

Furthermore, denote

$$\mathcal{T}(x) := x m_{1c}(x) m_{2c}(x), \tag{4.13}$$

$\mathcal{T}(x)$ is a strictly monotone decreasing function when $x > \lambda_+$.

For $z \in \mathbf{D}(\tau)$ defined in (4.1), denote

$$\kappa := |E - \lambda_+|. \tag{4.14}$$

By (4.11), it is easy to check that

$$\mathcal{T}(z) - c^{1/2} = \frac{z - \lambda_+ - i\sqrt{(\lambda_+ - z)(z - \lambda_-)}}{2c^{-1}}. \tag{4.15}$$

The following lemma summarizes the basic properties of $m_{2c}(z)$ and $\mathcal{T}(z)$, the estimates are based on the elementary calculations of (4.11) and (4.15). Their proofs can be found in [3], Lemma 3.3, and [4], Lemma 3.6.

Lemma 4.5. *For any $z \in \mathbf{D}(\tau)$ defined in (4.1), we have*

$$|\mathcal{T}(z)| \sim |m_{2c}(z)| \sim 1, \qquad |c^{1/2} - \mathcal{T}(z)| \sim |1 - m_{2c}^2(z)| \sim \sqrt{\kappa + \eta},$$

and

$$\operatorname{Im} \mathcal{T}(z) \sim \operatorname{Im} m_{2c}(z) \sim \begin{cases} \sqrt{\kappa + \eta}, & \text{if } E \in [\lambda_-, \lambda_+], \\ \dfrac{\eta}{\sqrt{\kappa + \eta}}, & \text{if } E \notin [\lambda_-, \lambda_+], \end{cases}$$

as well as

$$|\operatorname{Re} \mathcal{T}(z) - c^{1/2}| \sim \begin{cases} \dfrac{\eta}{\sqrt{\kappa + \eta}} + \kappa, & E \in [\lambda_-, \lambda_+], \\ \sqrt{\kappa + \eta}, & E \notin [\lambda_-, \lambda_+]. \end{cases} \tag{4.16}$$

The next lemma provides the local estimate on the derivative of $\mathcal{T}(x)$ on the real axis. We put its proof in the supplementary material [8].

Lemma 4.6. *For $d > c^{-1/4}$, denote $I_d := [x_-(d), x_+(d)]$, $x_\pm(d) := p(d) \pm N^{-1/2+\epsilon_0}(d - c^{-1/4})^{1/2}$, where ϵ_0 is defined in (2.1). Then $\forall x \in I_d$, we have that*

$$\mathcal{T}'(x) \sim (d - c^{-1/4})^{-1}. \tag{4.17}$$

The following perturbation identity plays the key role in our proof, as it naturally provides us a way to incorporate the Green functions using a deterministic equation. Its proof can be found in [17], Lemma 6.1.

Lemma 4.7. *Recall* (1.9), *assume* $\mu \in \mathbb{R}/\boldsymbol{\sigma}(H)$ *and* $\det \mathbf{D} \neq 0$, *then* $\mu \in \boldsymbol{\sigma}(\tilde{H})$ *if and only if*

$$\det\big(\mathbf{U}^* G(\mu)\mathbf{U} + \mathbf{D}^{-1}\big) = 0. \tag{4.18}$$

The following lemma establishes the connection between the Green functions of H and $\tilde{H}$ defined in (1.9), which is proved in the supplementary material [8].

Lemma 4.8. *For* $z \in \mathbb{C}^+$, *we have*

$$\tilde{G}(z) = G(z) - G(z)\mathbf{U}\big(\mathbf{D}^{-1} + \mathbf{U}^* G(z)\mathbf{U}\big)^{-1}\mathbf{U}^* G(z), \tag{4.19}$$

and

$$\mathbf{U}^* \tilde{G}(z)\mathbf{U} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\big(\mathbf{D}^{-1} + \mathbf{U}^* G(z)\mathbf{U}\big)^{-1}\mathbf{D}^{-1}. \tag{4.20}$$

One of the key ingredients of our computation are the local laws. We firstly introduce the anisotropic local law, which can be found in [19], Theorem 3.6. Denote

$$\Psi(z) := \sqrt{\frac{\operatorname{Im} m_{2c}(z)}{N\eta}} + \frac{1}{N\eta}, \qquad \underline{\Sigma} := \begin{pmatrix} z^{-1/2} & 0 \\ 0 & I \end{pmatrix}, \tag{4.21}$$

and $m(z) \equiv m_N(z)$ as the unique solution of the equation

$$f\big(m(z)\big) = z, \quad \operatorname{Im} m(z) \geq 0, \qquad f(x) = -\frac{1}{x} + \frac{1}{c_N}\frac{1}{x+1}.$$

Recall (4.7), the following lemma shows that $G(z)$ converges to a deterministic matrix $\Pi(z)$ with high probability.

Lemma 4.9. *Fix* $\tau > \epsilon_1$, *then for all* $z \in \mathbf{D}(\tau)$, *with* $1 - N^{-D_1}$ *probability, for any unit deterministic vectors* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{M+N}$, *we have*

$$\big|\big\langle \mathbf{u}, \underline{\Sigma}^{-1}\big(G(z) - \Pi(z)\big)\underline{\Sigma}^{-1}\mathbf{v}\big\rangle\big| \leq N^{\epsilon_1}\Psi(z), \qquad \big|m_2(z) - m(z)\big| \leq \frac{N^{\epsilon_1}}{N\eta}, \tag{4.22}$$

where $\Pi(z)$ *is defined as*

$$\Pi(z) := \begin{pmatrix} -z^{-1}\big(1 + m(z)\big)^{-1} & 0 \\ 0 & m(z) \end{pmatrix}. \tag{4.23}$$

It is notable that in general, $m(z)$ depends on N and Lemma 4.5 also holds for $m(z)$. However, in our computation, we can replace $m(z)$ with $m_{2c}(z)$ due to the following local MP law, which is proved in [29], Theorem 3.1.

Lemma 4.10. *Fix $\tau > \epsilon_1$, then for all $z \in \mathbf{D}(\tau)$, with $1 - N^{-D_1}$ probability, we have*

$$\left|m_2(z) - m_{2c}(z)\right| \le N^{\epsilon_1}\Psi(z).$$

Beyond the support of the limiting spectrum of the MP law, we have stronger results all the way down to the real axis. More precisely, define the region

$$\tilde{\mathbf{D}}(\tau, \epsilon_1) := \left\{z \in \mathbb{C}^+ : \lambda_+ + N^{-2/3+\epsilon_1} \le E \le \tau^{-1}, 0 < \eta \le \tau^{-1}\right\}, \tag{4.24}$$

then we have the following stronger control on $\tilde{\mathbf{D}}(\tau, \epsilon_1)$. The proof can be found in [3], Theorem 3.12, and [19], Theorem 3.7.

Lemma 4.11. *For $z \in \tilde{\mathbf{D}}(\tau, \epsilon_1)$, with $1 - N^{-D_1}$ probability, we have*

$$\left|\left\langle u, \mathcal{G}_2(z)v\right\rangle - m_{2c}(z)\langle u, v\rangle\right| \le N^{-1/2+\epsilon_1}(\kappa + \eta)^{-1/4},$$

for all unit vectors $u, v \in \mathbb{R}^N$. Similar result holds for $\mathcal{G}_1(z)$, $m_{1c}(z)$. Furthermore, for any deterministic vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{M+N}$, we have

$$\left|\left\langle \mathbf{u}, \underline{\Sigma}^{-1}\big(G(z) - \Pi(z)\big)\underline{\Sigma}^{-1}\mathbf{v}\right\rangle\right| \le N^{-1/2+\epsilon_1}(\kappa + \eta)^{-1/4}. \tag{4.25}$$

Denote the non-trivial classical eigenvalue locations $\gamma_1 \ge \gamma_2 \ge \cdots \ge \gamma_K$ of XX^* as $\int_{\gamma_i}^{\infty} d\rho_{1c} = \frac{i}{N}$, where ρ_{1c} is defined in (4.3). The consequent result of Lemma 4.9 is the rigidity of eigenvalues, which can be found in [4], Theorem 3.5.

Lemma 4.12. *Fix any small $\tau \in (0, 1)$, for $1 \le i \le (1 - \tau)K$, with $1 - N^{-D_1}$ probability, we have*

$$|\lambda_i - \gamma_i| \le N^{-2/3+\epsilon_1}\big(i \wedge (K + 1 - i)\big)^{-1/3}.$$

Furthermore, if $c \ne 1$, the above estimate holds for all $i = 1, 2, \ldots, K$.

Using Lemma 4.12, we find that κ_j^d defined in (2.12) is a deterministic version of $\kappa_{\mu_j} = |\mu_j - \lambda_+|$.

5. Proofs of Theorem 2.2 and 2.3

5.1. Singular values

In this subsection, we focus on the singular values of $\tilde{S}$ and prove Theorem 2.2. We will follow the basic idea of [17] and slightly modify the proof. A key deviation from their proof is that our matrix $\mathbf{D}$ defined in (1.10) is not diagonal, it appears that in order to analyze (4.18), they only need to deal with the diagonal elements but we need to control the whole matrix. We will make use of the following interlacing theorem for rectangular matrices, the proof can be found in [34], Exercise 1.3.22.

Lemma 5.1. *For any $M \times N$ matrices A, B, denote $\sigma_i(A)$ as the i-th largest singular value of A, then we have*

$$\sigma_{i+j-1}(A+B) \le \sigma_i(A) + \sigma_j(B), \quad 1 \le i, j, i+j-1 \le K.$$

The proof relies on two main steps: (i) fix a configuration independent of N, establish two permissible regions, $\Gamma(\mathbf{d})$ of k^+ components and I_0, where the outliers of $\tilde{S}\tilde{S}^*$ are allowed to lie in $\Gamma(\mathbf{d})$ and each component contains precisely one eigenvalue and the $r-k^+$ non-outliers lie in I_0; (ii) a continuity argument where the result of (i) can be extended to arbitrary N-dependent $\mathbf{D}$.

The following $2r \times 2r$ matrix plays the key role in our analysis

$$M^r(x) := \mathbf{U}^* G(x) \mathbf{U} + \mathbf{D}^{-1}. \tag{5.1}$$

By Lemma 4.7, $x \in \boldsymbol{\sigma}(\tilde{S}\tilde{S}^*)$ if and only if $\det M^r(z) = 0$. Using Lemma 4.10 and 4.11, we find that $x^{-r} T^r(x) \approx \det M^r(x)$, where $T^r(x)$ is defined in (4.10). As $T^r(x)$ behaves differently in $\Gamma(\mathbf{d})$ and I_0, we will use different strategies to prove (2.7) and (2.8).

We remark that, our discussion is slightly easier than [17], Section 6, in particular the counting argument of the non-outliers. The reason is, for the application purpose, we only need the result of (2.8) to locate the eigenvalues around λ_+. However, in [17], they have stronger results to stick the eigenvalues of $\tilde{S}\tilde{S}^*$ around those of XX^*. We will not pursue this generalization in this paper.

Proof of Theorem 2.2. Denote $k^0 := r - k^+$ and write

$$\mathbf{d} = (d_1, \dots, d_r) = \big(\mathbf{d}^0, \mathbf{d}^+\big), \qquad \mathbf{d}^\sigma = \big(d_1^\sigma, \dots, d_{k^\sigma}^\sigma\big), \quad \sigma = 0, +,$$

where we adapt the convention

$$d_{k^0}^0 \le \dots \le d_1^0 \le c^{1/4} < d_{k^+}^+ \le \dots \le d_1^+, \quad k^0 + k^+ = r.$$

Next, we define the sets

$$\mathcal{D}^+(\epsilon_0) := \big\{\mathbf{d}^+ : c^{-1/4} + N^{-1/3+\epsilon_0} \le d_i^+ \le \tau^{-1}, i = 1, \dots, k^+\big\}, \tag{5.2}$$

$$\mathcal{D}^0(\epsilon_0) := \big\{\mathbf{d}^0 : 0 < d_i^0 < c^{-1/4} + N^{-1/3+\epsilon_0}, i = 1, \dots, k_0\big\}, \tag{5.3}$$

and the sets of allowed $\mathbf{d}'s$, which is $\mathcal{D}(\epsilon_0) := \{(\mathbf{d}^0, \mathbf{d}^+) : \mathbf{d}^\sigma \in \mathcal{D}^\sigma(\epsilon_0), \sigma = +, 0\}$. Denote the following sequence of intervals

$$I_i^+(\mathbf{d}) := \big[p\big(d_i^+\big) - N^{-1/2+\epsilon_3}\big(d_i^+ - c^{-1/4}\big)^{1/2}, p\big(d_i^+\big) + N^{-1/2+\epsilon_3}\big(d_i^+ - c^{-1/4}\big)^{1/2}\big], \tag{5.4}$$

where ϵ_3 satisfies the following condition

$$C\epsilon_1 < \epsilon_3 < \frac{1}{4}\epsilon_0, \quad C > 2 \text{ is some large constant.} \tag{5.5}$$

For $\mathbf{d} \in \mathcal{D}(\epsilon_0)$, we denote $\Gamma(\mathbf{d}) := \bigcup_{i=1}^{k^+} I_i^+(\mathbf{d})$ and $I^0 := [\lambda_+ - N^{-2/3+C'\epsilon_0}, \lambda_+ + N^{-2/3+C'\epsilon_0}]$, where C' satisfies $2 < C' < 4$.

For a first step, we show that $\Gamma(\mathbf{d})$ is our permissible region which keeps track of the outlier eigenvalues of $\tilde{S}\tilde{S}^*$. And the rest of the eigenvalues corresponding to $\mathcal{D}^0(\epsilon_0)$ will lie in I^0. We fix a configuration $\mathbf{d}(0) \equiv \mathbf{d}$ that is independent of N in this step.

Lemma 5.2. *For any $\mathbf{d} \in \mathcal{D}(\epsilon_0)$, with $1 - N^{-D_1}$ probability, we have*

$$\boldsymbol{\sigma}^+\big(\tilde{S}\tilde{S}^*\big) \subset \Gamma(\mathbf{d}), \tag{5.6}$$

where $\boldsymbol{\sigma}^+(\tilde{S}\tilde{S}^*)$ is the set of the outlier eigenvalues of $\tilde{S}\tilde{S}^*$ associated with $\mathcal{D}^+(\epsilon_0)$. Moreover, each interval $I_i^+(\mathbf{d})$ contains precisely one eigenvalue of $\tilde{S}\tilde{S}^*$, $i = 1, 2, \ldots, k^+$. Furthermore, we have

$$\boldsymbol{\sigma}^{\boldsymbol{o}}\big(\tilde{S}\tilde{S}^*\big) \subset I^0, \tag{5.7}$$

where $\boldsymbol{\sigma}^{\boldsymbol{o}}(\tilde{S}\tilde{S}^*)$ is the set of the non-outlier eigenvalues corresponding to $\mathcal{D}^0(\epsilon_0)$.

Proof. First of all, it is easy to check that $\Gamma(\mathbf{d}) \cap I^0 = \varnothing$ using (4.9) and the fact $C' > 2$. Denote $S_b := p(d_{k^+}^+) - N^{-1/2+\epsilon_3}(d_{k^+}^+ - c^{-1/4})^{1/2}$. In order to prove (5.6), we first consider the case when $x > S_b$. It is notable that $x \notin \boldsymbol{\sigma}(XX^*)$ by Lemma 4.12, (4.9) and (5.5). Recall (4.23) and (5.1), using the fact r is bounded and Lemma 4.11, with $1 - N^{-D_1}$ probability, we have

$$M^r(x) = \mathbf{U}^*\Pi(x)\mathbf{U} + \mathbf{D}^{-1} + O\big(N^{-1/2+\epsilon_1}\kappa^{-1/4}\big). \tag{5.8}$$

It is well known that if $\lambda \in \boldsymbol{\sigma}(A+B)$ then $\operatorname{dist}(\lambda, \boldsymbol{\sigma}(A)) \leq \|B\|$; therefore, we have that $\mu_i(\tilde{S}\tilde{S}^*) \leq \tau^{-1}$, $i = 1, \ldots, K$ for $\tau > 0$ defined in (4.1). Recall (4.10), by (4.9), (4.17) and (5.5), with $1 - N^{-D_1}$ probability, we have

$$\big|T^r(x)\big| \geq N^{-1/2+(C-1)\epsilon_1}\kappa^{-1/4}, \quad \text{if } x \in \big[S_b, \tau^{-1}\big]/\Gamma(\mathbf{d}). \tag{5.9}$$

Using the formula

$$\det\begin{bmatrix} xI_r & \operatorname{diag}(\alpha_1, \ldots, \alpha_r) \\ \operatorname{diag}(\alpha_1, \ldots, \alpha_r) & yI_r \end{bmatrix} = \prod_{i=1}^r \big(xy - \alpha_i^2\big),$$

Lemma 4.10, (4.6) and (5.8), we conclude that

$$\det\big(\mathbf{D}^{-1} + \mathbf{U}^*\Pi(x)\mathbf{U}\big) = x^{-r}T^r(x) + O\big(N^{-1/2+\epsilon_1}\kappa^{-1/4}\big). \tag{5.10}$$

By (5.9) and (5.10), we conclude that $M^r(x)$ is non-singular when $x \in [S_b, \tau^{-1}]/\Gamma(\mathbf{d})$.

Next, we will use Roché's theorem to show that inside the permissible region, each interval $I_i^+(\mathbf{d})$ contains precisely one eigenvalue of $\tilde{S}\tilde{S}^*$. Let $i \in \{1, \ldots, k^+\}$ and pick a small N-independent counterclockwise (positive-oriented) contour $\mathcal{C} \subset \mathbb{C}/[(1 - c^{-1/2})^2, (1 + c^{-1/2})^2]$ that encloses $p(d_i^+)$ but no other $p(d_j^+)$, $j \neq i$. For large enough N, define $f(z) := \det(M^r(z))$, $g(z) := \det(T^r(z))$. By the definition of determinant, the functions g, f are holomorphic on and inside $\mathcal{C}$. And $g(z)$ has precisely one zero $z = p(d_i^+)$ inside $\mathcal{C}$. On $\mathcal{C}$, it is easy to check that

$$\min_{z \in \mathcal{C}}\big|g(z)\big| \geq c > 0, \quad \big|g(z) - f(z)\big| \leq N^{-1/2+\epsilon_1}\kappa^{-1/4},$$

where we use (5.8) and Lemma 4.10. Hence, $f(z)$ has only one zero in $I_i^+(\mathbf{d})$ according to Rouché's theorem. This concludes the proof of (5.6) using Lemma 4.7. In order to prove (5.7), using the following fact: for any two $M \times N$ rectangular matrices A, B, we have $\sigma_i(A+B) \geq \sigma_i(A) + \sigma_K(B)$, $i = 1, \dots, K$, and Lemma 4.12, we find that

$$\mu_i \geq \lambda_+ - N^{-2/3+C'\epsilon_0}, \quad i = k^+ + 1, \dots, r. \tag{5.11}$$

For the non-outliers, we assume that $S_b > \lambda_+ + N^{-2/3+C'\epsilon_0}$, otherwise the proof is already done. Now we assume $x \notin I_0$, by (5.6) and (5.11), we only need to discuss the case when $x \in (\lambda_+ + N^{-2/3+C'\epsilon_0}, S_b)$. In this case, we will prove that $M^r(x)$ is non-singular by comparing with $M^r(z)$, where $z = x + iN^{-2/3-\epsilon_4}$ and $\epsilon_4 < \epsilon_1$ is some small positive constant. Denote the spectral decomposition of $G(z)$ as

$$G(z) = \sum_k \frac{1}{\lambda_k - z} \mathbf{g}_\alpha \mathbf{g}_\alpha^*, \quad \mathbf{g}_\alpha \in \mathbb{R}^{M+N}.$$

Denote $\mathbf{u}_i, i = 1, \dots, 2r$ as the i-th column in $\mathbf{U}$ defined in (1.10) and abbreviate $\mathbf{u}_i^* G(z) \mathbf{u}_j$ as $G_{\mathbf{u}_i\mathbf{u}_j}(z)$, and $\eta := N^{-2/3-\epsilon_4}$, using spectral decomposition and the fact $x > \lambda_+ + N^{-2/3+C'\epsilon_0}$, we have

$$\left| G_{\mathbf{u}_i\mathbf{u}_j}(x) - G_{\mathbf{u}_i\mathbf{u}_j}(x+i\eta) \right| \leq \operatorname{Im} G_{\mathbf{u}_i\mathbf{u}_i}(x+i\eta) + \operatorname{Im} G_{\mathbf{u}_j\mathbf{u}_j}(x+i\eta).$$

Therefore, by Lemma 4.10 and 4.11, with $1 - N^{-D_1}$ probability, we have

$$M^r(x) = M^r(z) + O\left(N^{\epsilon_1} \left(\operatorname{Im} m_{2c}(z) + \sqrt{\frac{\operatorname{Im} m_{2c}(z)}{N\eta}} \right) \right).$$

Using Lemma 4.5 and a similar discussion to (5.9), we have

$$M^r(x) = T^r(z) + O\big(N^{-1/3} \big(N^{-C'\epsilon_0/4} + N^{\epsilon_1 - C'\epsilon_0/4} \big) \big).$$

By Lemmas 4.5 and 4.10, we find that $|T^r(z)| \geq N^{-1/3+\frac{C'\epsilon_0}{2}}$, where we use the assumption that $x > \lambda_+ + N^{-2/3+C'\epsilon_0}$. Therefore, $M^r(x)$ is non-singular as we have assumed $2 < C' < 4$. This concludes the proof of (5.7). □

In the second step, we will extend the proof to any configuration $\mathbf{d}(1)$ depending on N using the continuity argument. This is done by a bootstrap argument by choosing a continuous path connecting $\mathbf{d}(0)$ and $\mathbf{d}(1)$. It is recorded as the following lemma and its proof will be provided in the supplementary material [8].

Lemma 5.3. *For any N-dependent configuration $\mathbf{d}(1) \in \mathcal{D}(\epsilon_0)$, (2.7) and (2.8) hold true.*

□

5.2. Singular vectors

In this section, we focus on the local behavior of singular vectors. We will follow the discussion of [4], Section 5 and 6. We first deal with the outlier singular vectors and then the non-outlier ones. Due to similarity, we only prove (2.11) and (2.13), (2.10) and (2.12) can be handled similarly.

Proof of (2.11). It is notable that, by Lemma 4.11 and Theorem 2.2, for $i \in \mathcal{O}$, there exists a constant $C > 0$, for N large enough, with $1 - N^{-D_1}$ probability, we can choose an event Ξ such that for all $z \in \tilde{\mathbf{D}}(\tau, \epsilon_1)$ defined in (4.24)

$$\mathbf{1}(\Xi)\big|\big(V^* \mathcal{G}_2(z) V\big)_{ij} - m_{2c}(z)\delta_{ij}\big| \leq (\kappa + \eta)^{-1/4} N^{-1/2 + C\epsilon_1}. \tag{5.12}$$

Next, we will restrict our discussion on the event Ξ. Recall (2.5) and for $A \subset \mathcal{O}$, we define for each $i \in A$ the radius

$$\rho_i := \frac{\nu_i \wedge (d_i - c^{-1/4})}{2}. \tag{5.13}$$

Under the assumption of (2.4), we have (see the equation (5.10) of [4])

$$\rho_i \geq \frac{1}{2}\big(d_i - c^{-1/4}\big)^{-1/2} N^{-1/2+\epsilon_0}. \tag{5.14}$$

We define the contour $\Gamma := \partial \Upsilon$ as the boundary of the union of discs $\Upsilon := \bigcup_{i \in A} B_{\rho_i}(d_i)$, where $B_\rho(d)$ is the open disc of radius ρ around d. We summarize the basic properties of Υ as the following lemma, its proof can be found in [4], Lemmas 5.4 and 5.5.

Lemma 5.4. *Recall* (2.6) *and* (4.24), *we have* $\overline{p(\Upsilon)} \subset \tilde{\mathbf{D}}(\tau, \epsilon_1)$. *Moreover, each outlier* $\{\mu_i\}_{i \in A}$ *lies in* $p(\Upsilon)$, *and all the other eigenvalues of* $\tilde{S}\tilde{S}^*$ *lie in the complement of* $\overline{p(\Upsilon)}$.

Armed with the above results, we now start the proof of the outlier singular vectors. Our starting point is an integral representation of the singular vectors. By (4.7), we have

$$v_i^* \tilde{\mathcal{G}}_2 v_j = \mathbf{v}_i^* \tilde{G} \mathbf{v}_j, \tag{5.15}$$

where $\mathbf{v}_i \in \mathbb{R}^{M+N}$ is the natural embedding of v_i with $\mathbf{v}_i = (0, v_i)^*$. Recall (2.3), using the spectral decomposition of $\tilde{\mathcal{G}}_2(z)$, Lemma 5.4 and Cauchy's integral formula, we have

$$\mathbf{P}_r = -\frac{1}{2\pi i}\int_{p(\Gamma)} \tilde{\mathcal{G}}_2(z)\, dz = -\frac{1}{2\pi i}\int_{\Gamma} \tilde{\mathcal{G}}_2\big(p(\zeta)\big) p'(\zeta)\, d\zeta. \tag{5.16}$$

By Lemma 4.8, Cauchy's integral formula, (5.15) and (5.16), we have

$$\langle v_i, \mathbf{P}_r v_j \rangle = \frac{1}{2 d_i d_j \pi i}\int_{p(\Gamma)} \big(\mathbf{D}^{-1} + \mathbf{U}^* G(z) \mathbf{U}\big)^{-1}_{ij} \frac{dz}{z}, \tag{5.17}$$

where $\bar{i}$, $\bar{j}$ are defined as $\bar{i} := r + i$, $\bar{j} := r + j$. Recall (4.23), as $\mathbf{D}^{-1} + \mathbf{U}^* \Pi(z) \mathbf{U}$ is of finite dimension, by Lemma 4.10, 4.11, (4.6) and (5.12), we can now use $\Pi(z)$ as

$$\Pi(z) := \begin{pmatrix} m_{1c}(z) & 0 \\ 0 & m_{2c}(z) \end{pmatrix}.$$

Next, we decompose $\mathbf{D}^{-1} + \mathbf{U}^* G(z) \mathbf{U}$ by

$$\mathbf{D}^{-1} + \mathbf{U}^* G(z) \mathbf{U} = \mathbf{D}^{-1} + \mathbf{U}^* \Pi(z) \mathbf{U} - \Delta(z), \quad \Delta(z) = \mathbf{U}^* \Pi(z) \mathbf{U} - \mathbf{U}^* G(z) \mathbf{U}. \tag{5.18}$$

It is notable that $\Delta(z)$ can be controlled by Lemmas 4.10 and 4.11. Using the resolvent expansion to the order of one on (5.18), we have

$$\langle v_i, \mathbf{P}_r v_j \rangle = \frac{1}{d_i d_j} \big(S^{(0)} + S^{(1)} + S^{(2)} \big), \tag{5.19}$$

where

$$\begin{aligned} S^{(0)} &:= \frac{1}{2\pi i} \int_{p(\Gamma)} \left(\frac{1}{\mathbf{D}^{-1} + \mathbf{U}^* \Pi(z) \mathbf{U}} \right)_{ij} \frac{dz}{z}, \\ S^{(1)} &= \frac{1}{2\pi i} \int_{p(\Gamma)} \left[\frac{1}{\mathbf{D}^{-1} + \mathbf{U}^* \Pi(z) \mathbf{U}} \Delta(z) \frac{1}{\mathbf{D}^{-1} + \mathbf{U}^* \Pi(z) \mathbf{U}} \right]_{ij} \frac{dz}{z}, \\ S^{(2)} &= \frac{1}{2\pi i} \int_{p(\Gamma)} \left[\frac{1}{\mathbf{D}^{-1} + \mathbf{U}^* \Pi(z) \mathbf{U}} \Delta(z) \frac{1}{\mathbf{D}^{-1} + \mathbf{U}^* \Pi(z) \mathbf{U}} \Delta(z) \frac{1}{\mathbf{D}^{-1} + \mathbf{U}^* G(z) \mathbf{U}} \right]_{ij} \frac{dz}{z}. \end{aligned}$$

By an elementary computation, we have

$$\big(\mathbf{D}^{-1} + \mathbf{U}^* \Pi(z) \mathbf{U} \big)^{-1}_{ij} = \begin{cases} \delta_{ij} \dfrac{z m_{2c}(z)}{z m_{1c}(z) m_{2c}(z) - d_i^{-2}}, & 1 \le i, j \le r; \\ \delta_{ij} \dfrac{z m_{1c}(z)}{z m_{1c}(z) m_{2c}(z) - d_i^{-2}}, & r \le i, j \le 2r; \\ \delta_{\bar{i} j} (-1)^{i+j} \dfrac{z^{1/2} d_i^{-1}}{z m_{1c}(z) m_{2c}(z) - d_i^{-2}}, & 1 \le i \le r, r \le j \le 2r; \\ \delta_{i \bar{j}} (-1)^{i+j} \dfrac{z^{1/2} d_j^{-1}}{z m_{1c}(z) m_{2c}(z) - d_j^{-2}}, & r \le i \le 2r, 1 \le j \le r. \end{cases} \tag{5.20}$$

Using the fact $p_i m_{1c}(p_i) m_{2c}(p_i) = \frac{1}{d_i^2}$ and the residual theorem, we have

$$S^{(0)} = \delta_{ij} \frac{m_{2c}(p_i)}{\mathcal{T}'(p_i)} = \delta_{ij} \frac{d_i^4 - c^{-1}}{d_i^2 + 1}. \tag{5.21}$$

Next, we control the term $S^{(1)}$. Applying (5.20) on $S^{(1)}$, we have

$$S^{(1)} = \frac{1}{2\pi i} \int_{p(\Gamma)} \frac{f(z)}{(zm_{1c}(z)m_{2c}(z) - d_i^{-2})(zm_{1c}(z)m_{2c}(z) - d_j^{-2})}\, dz, \tag{5.22}$$

where $f(z) = f_1(z) + f_2(z)$ and $f_{1,2}(z)$ are defined as

$$f_1(z) := m_{2c}(z)\big[zm_{2c}(z)\Delta(z)_{ij} + (-1)^{i+\bar{i}} z^{1/2} d_i^{-1} \Delta(z)_{\bar{i}j}\big],$$
$$f_2(z) := d_j^{-1}\big[(-1)^{j+\bar{j}} z^{1/2} m_{2c}(z)\Delta(z)_{i\bar{j}} + (-1)^{i+j+\bar{i}+\bar{j}} d_i^{-1} \Delta(z)_{\bar{i}\bar{j}}\big].$$

We now use the change of variable as in (5.16) and rewrite $S^{(1)}$ as

$$\begin{aligned} S^{(1)} &= \frac{1}{2\pi i} \int_\Gamma \frac{f(p(\zeta))}{(\zeta^{-2} - d_i^{-2})(\zeta^{-2} - d_j^{-2})} p'(\zeta)\, d\zeta \\ &= d_i^2 d_j^2 \frac{1}{2\pi i} \int_\Gamma \frac{f(p(\zeta))\zeta^4}{(d_i^2 - \zeta^2)(d_j^2 - \zeta^2)} p'(\zeta)\, d\zeta, \end{aligned}$$

where we use the fact $p(\zeta)m_{1c}(p(\zeta))m_{2c}(p(\zeta)) = \zeta^{-2}$. By (4.9), Lemma 4.5 and 4.11, we conclude that

$$\big|f\big(p(\zeta)\big)p'(\zeta)\zeta^4\big| \le \big(\zeta - c^{-1/4}\big)^{1/2} N^{-1/2+\epsilon_1}. \tag{5.23}$$

Denote

$$f_{ij}(\zeta) = \frac{f(p(\zeta))p'(\zeta)\zeta^4}{(d_i + \zeta)(d_j + \zeta)}.$$

As f_{ij} is holomorphic inside the contour Γ, by Cauchy's differentiation formula, we have

$$f'_{ij}(\zeta) = \frac{1}{2\pi i} \int_{\mathcal{C}} \frac{f_{ij}(\xi)}{(\xi - \zeta)^2}\, d\xi, \tag{5.24}$$

where the contour $\mathcal{C}$ is the circle of radius $\frac{|\zeta - c^{-1/4}|}{2}$ centered at ζ. Hence, by (4.9), (5.23), (5.24) and the residual theorem, we have

$$\big|f'_{ij}(\zeta)\big| \le \big(\zeta - c^{-1/4}\big)^{-1/2} N^{-1/2+\epsilon_1}. \tag{5.25}$$

In order to estimate $S^{(1)}$, we consider the following three cases (i) $i, j \in A$, (ii) $i \in A, j \notin A$ (or $i \notin A, j \in A$), (iii) $i, j \notin A$. By the residual theorem, $S^{(1)} = 0$ when case (iii) happens. Hence, we only need to consider the cases (i) and (ii). For the case (i), when $i \neq j$, by the residual theorem and (5.25), we have

$$\begin{aligned}\left|S^{(1)}\right| &= d_i^2 d_j^2 \left|\frac{f_{ij}(d_i) - f_{ij}(d_j)}{d_i - d_j}\right| \leq \frac{d_i^2 d_j^2}{|d_i - d_j|}\left|\int_{d_i}^{d_j} \left|f'_{ij}(t)\right| dt\right| \\ &\leq \frac{d_i^2 d_j^2 N^{-1/2+\epsilon_1}}{(d_i - c^{-1/4})^{1/2} + (d_j - c^{-1/4})^{1/2}}.\end{aligned}$$

When $i = j$, by the residual theorem, we have $|S^{(1)}| \leq d_i^4 (d_i - c^{-1/4})^{-1/2} N^{-1/2+\epsilon_1}$. For the case (ii), when $i \in A$, $j \notin A$, by the residual theorem and (5.12), we have

$$\left|S^{(1)}\right| = \left|\frac{d_i^2 d_j^2 f_{ij}(d_i)}{d_i - d_j}\right| \leq \frac{d_i^2 d_j^2 (d_i - c^{-1/4})^{1/2}}{|d_i - d_j|} N^{-1/2+\epsilon_1}.$$

We can get similar results when $i \notin A$, $j \in A$. Putting all the cases together, we find that

$$\begin{aligned}\left|S^{(1)}\right| \leq N^{-1/2+\epsilon_1}\Bigg[&\frac{\mathbf{1}(i \in A, j \in A) d_i^2 d_j^2}{(d_i - c^{-1/4})^{1/2} + (d_j - c^{-1/4})^{1/2}} \\ &+ \mathbf{1}(i \in A, j \notin A) \frac{d_i^2 d_j^2 (d_i - c^{-1/4})^{1/2}}{|d_i - d_j|} \\ &+ \mathbf{1}(i \notin A, j \in A) \frac{d_i^2 d_j^2 (d_j - c^{-1/4})^{1/2}}{|d_i - d_j|}\Bigg].\end{aligned} \tag{5.26}$$

Finally, we need to estimate $S^{(2)}$. Here the residual calculations can not be applied directly as $\mathbf{U}^* G(z)\mathbf{U}$ is not necessary to be diagonal and a relation comparable to $p(\zeta) m_{1c}(p(\zeta)) \times m_{2c}(p(\zeta)) = \zeta^{-2}$ does not exist. Instead, we need to precisely choose the contour Γ. We record the result as the following lemma, whose proofs will be given in the supplementary material [8].

Lemma 5.5. *When N is large enough, with $1 - N^{-D_1}$ probability, for some constant $C > 0$, we have*

$$\left|S^{(2)}\right| \leq C N^{-1+2\epsilon_1} \left(\frac{1}{\nu_i} + \frac{\mathbf{1}(i \in A)}{|d_i - c^{-1/4}|}\right)\left(\frac{1}{\nu_j} + \frac{\mathbf{1}(j \in A)}{|d_j - c^{-1/4}|}\right). \tag{5.27}$$

Therefore, plugging (5.21), (5.26) and (5.27) into (5.19), we conclude the proof of (2.11). Before concluding this subsection, we briefly discuss the proof of (2.10). By Lemma 4.8 and Cauchy's integral formula, we have

$$\langle u_i, \mathbf{P}_l u_j \rangle = \frac{1}{2 d_i d_j \pi i} \int_{p(\Gamma)} \left(D^{-1} + \mathbf{U}^* G(z)\mathbf{U}\right)^{-1}_{\bar{i}\bar{j}} \frac{dz}{z}.$$

Then we can use a similar discussion as (5.19), computing the convergent limit from $S^{(0)}$ and controlling the bounds for $S^{(1)}$ and $S^{(2)}$. We remark that the convergent limit is different because we use $(\mathbf{D}^{-1}+\mathbf{U}^*\Pi(z)\mathbf{U})_{ij}, r \leq i,\ j \leq 2r$ in (5.20), which results in

$$S^{(0)} = \delta_{ij}\frac{m_{1c}(p_i)}{\mathcal{T}'(p_i)} = \delta_{ij}\frac{d_i^4 - c^{-1}}{d_i^2 + c^{-1}}.$$

This concludes the proof of (2.10). □

For the non-outliers, the proof strategy for the outlier singular vectors will not work as we cannot use the residual theorem. We will use a spectral decomposition for our proof.

Proof of (2.13). Denote

$$z = \mu_j + i\eta, \tag{5.28}$$

where η is defined as the smallest solution of

$$\operatorname{Im} m_{2c}(z) = N^{-1+6\epsilon_1}\eta^{-1}. \tag{5.29}$$

As we assume $j \leq (1-\tau)K$ or $c \neq 1$, we conclude that $|z|$ has a constant lower bound. Therefore, by Lemma 4.9, 4.10 and 4.11, with $1 - N^{-D_1}$ probability, we have

$$\left|\left\langle \mathbf{u}, \underline{\Sigma}^{-1}\big(G(z) - \Pi(z)\big)\underline{\Sigma}^{-1}\mathbf{v}\right\rangle\right| \leq \frac{N^{4\epsilon_1}}{N\eta}. \tag{5.30}$$

Recall (4.14), abbreviating $\kappa = |\mu_j - \lambda_+|$, by Lemma 4.5 and (2.8), we find that (see [4], (6.5) and (6.6))

$$\eta \sim \begin{cases} \dfrac{N^{6\epsilon_1}}{N\sqrt{\kappa} + N^{2/3+2\epsilon_1}}, & \text{if } \mu_j \leq \lambda_+ + N^{-2/3+4\epsilon_1}, \\ N^{-1/2+3\epsilon_1}\kappa^{1/4}, & \text{if } \mu_j \geq \lambda_+ + N^{-2/3+4\epsilon_1}. \end{cases} \tag{5.31}$$

For z defined in (5.28), by the spectral decomposition, we have

$$\langle v_i, \tilde{v}_j\rangle^2 \leq \eta\left\langle v_i, \operatorname{Im}\tilde{\mathcal{G}}_2(z)v_i\right\rangle = \eta\left\langle \mathbf{v}_i, \operatorname{Im}\tilde{G}(z)\mathbf{v}_i\right\rangle, \tag{5.32}$$

where $\mathbf{v}_i \in \mathbb{R}^{M+N}$ is the natural embedding of v_i. By Lemma 4.8, we have

$$\left\langle \mathbf{v}_i, \tilde{G}(z)\mathbf{v}_i\right\rangle = -\frac{1}{zd_i^2}\left(\mathbf{D}^{-1} + \mathbf{U}^*G(z)\mathbf{U}\right)^{-1}_{ii}.$$

Similar to (5.19), using a simple resolvent expansion and (5.20), we have

$$\begin{aligned}
&\left\langle \mathbf{v}_i, \tilde{G}(z)\mathbf{v}_i\right\rangle \\
&\quad = -\frac{1}{zd_i^2}\Bigg[\frac{zm_{2c}(z)}{zm_{1c}(z)m_{2c}(z) - d_i^{-2}} + \frac{zf(z)}{(zm_{1c}(z)m_{2c}(z) - d_i^{-2})^2} \\
&\qquad + \left(\left[\left(\mathbf{D}^{-1} + \mathbf{U}^*\Pi(z)\mathbf{U}\right)^{-1}\Delta(z)\right]^2\left(\mathbf{D}^{-1} + \mathbf{U}^*G(z)\mathbf{U}\right)^{-1}\right)_{ii}\Bigg],
\end{aligned} \tag{5.33}$$

where $f(z)$ is defined in (5.22). To estimate the right-hand side of (5.33), we use the following error estimate

$$\min_j \left| d_j^{-2} - \mathcal{T}(z) \right| \geq \operatorname{Im} \mathcal{T}(z) \sim \operatorname{Im} m_{2c}(z) = \frac{N^{6\epsilon_1}}{N\eta} \gg \frac{N^{4\epsilon_1}}{N\eta} \geq \left| \Delta(z) \right|,$$

where we use (5.30) and Lemma 4.10. By a similar resolvent expansion, there exists some constant $C > 0$, such that

$$\left\| \frac{1}{\mathbf{D}^{-1} + \mathbf{U}^* G(z) \mathbf{U}} \right\| \leq \frac{C}{\operatorname{Im} m_{2c}(z)} = C N^{1-6\epsilon_1} \eta.$$

We therefore get from (5.33), the definition of f and (5.30) that

$$\langle \mathbf{v}_i, \tilde{G}(z) \mathbf{v}_i \rangle = \frac{m_{2c}(z)}{1 - d_i^2 \mathcal{T}(z)} + O\left(\frac{d_i^2}{|1 - d_i^2 \mathcal{T}(z)|^2} \frac{N^{4\epsilon_1}}{N\eta} \right). \tag{5.34}$$

By (5.32), we have

$$\begin{aligned} \langle v_i, \tilde{v}_j \rangle^2 \leq \frac{\eta}{|1 - d_i^2 \mathcal{T}(z)|^2} \Big[& \operatorname{Im} m_{2c}(z) \big(1 - d_i^2 c^{1/2} + \operatorname{Re}\big(d_i^2 c^{1/2} - d_i^2 \mathcal{T}(z) \big) \big) \\ & + d_i^2 \operatorname{Re} m_{2c}(z) \operatorname{Im} \mathcal{T}(z) + \frac{C d_i^2 N^{4\epsilon_1}}{N\eta} \Big]. \end{aligned} \tag{5.35}$$

By (4.16), (5.29) and (5.31), we have

$$\begin{aligned} & \operatorname{Im} m_{2c}(z) \big[\big(1 - d_i^2 c^{1/2} \big) + \operatorname{Re}\big(d_i^2 c^{1/2} - d_i^2 \mathcal{T}(z) \big) \big] \\ & \quad \leq \frac{C N^{6\epsilon_1}}{N\eta} \left(\left| d_i - c^{-1/4} \right| + \max\left\{ \sqrt{\kappa + \eta}, \frac{\eta}{\sqrt{\kappa + \eta}} + \kappa \right\} \right). \end{aligned}$$

For the other item, by Lemma 4.5, we have $|\operatorname{Re} m_{2c}(z) \operatorname{Im} \mathcal{T}(z)| \sim \operatorname{Im} m_{2c}(z)$. Putting all these estimates together, we have

$$\langle v_i, \tilde{v}_j \rangle^2 \leq \frac{C N^{6\epsilon_1}}{N |1 - d_i^2 \mathcal{T}(z)|^2}.$$

The rest of the proof leaves to give an estimate of $1 - d_i^2 \mathcal{T}(z)$. We summarize it as the following lemma and put its proof in the supplementary material [8].

Lemma 5.6. *Recall* (4.3), *for all* $\mu_j \in [\lambda_-, \lambda_+ + N^{-2/3 + C\epsilon_0}]$, *there exists a constant* $\delta > 0$, *such that*

$$\left| 1 - d_i^2 \mathcal{T}(z) \right| \geq \delta d_i^2 \left(\left| d_i^{-2} - c^{1/2} \right| + \operatorname{Im} \mathcal{T}(z) \right).$$

Therefore, we have

$$\langle v_i, \tilde{v}_j\rangle^2 \leq \frac{N^{C\epsilon_0}}{N((d_i - c^{-1/4})^2 + \kappa_j^d)}, \quad \kappa_j^d := N^{-2/3}\big(j \wedge (K+1-j)\big)^{2/3},$$

where we use the fact that $\operatorname{Im} \mathcal{T}(z) \geq c\sqrt{\kappa_j^d}$ (see the equation (6.14) of [4]). This concludes the proof of (2.13). For the proof of (2.12), we will use the spectral decomposition

$$\langle u_i, \tilde{u}_j\rangle^2 \leq \eta\langle u_i, \operatorname{Im} \tilde{\mathcal{G}}_1(z) u_i\rangle = \eta\langle \mathbf{u}_i, \operatorname{Im} \tilde{G}(z)\mathbf{u}_i\rangle,$$

and

$$\langle \mathbf{u}_i, \tilde{G}(z)\mathbf{u}_i\rangle = -\frac{1}{zd_i^2}\big(\mathbf{D}^{-1} + \mathbf{U}^* G(z)\mathbf{U}\big)^{-1}_{\bar{i}\bar{i}}.$$

Then by the resolvent expansion similar to (5.33) and control the items using Lemmas 4.5, 4.9, 4.10 and 4.11, we can conclude the proof. □

Acknowledgements

The author would like to thank Zhigang Bao, Jeremy Quastel, Bálint Virág, Ke Wang and Zhou Zhou for fruitful discussions and valuable suggestions, which have greatly improved the paper. The author is also grateful to an anonymous referee, the associated editor and editor for providing detailed and constructive suggestions and comments, which have improved the paper significantly.

Supplementary Material

Supplement to "High dimensional deformed rectangular matrices with applications in matrix denoising" (DOI: 10.3150/19-BEJ1129SUPP; .pdf). This supplementary material contains auxiliary lemmas and proofs of Proposition 3.3, Theorems 3.4 and 3.5, Lemmas 4.4, 4.6, 4.8, 5.3, 5.5 and 5.6.

References

[1] Baik, J., Ben Arous, G. and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575 https://doi.org/10.1214/009117905000000233

[2] Benaych-Georges, F. and Nadakuditi, R.R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. MR2782201 https://doi.org/10.1016/j.aim.2011.02.007

[3] Bloemendal, A., Erdős, L., Knowles, A., Yau, H.-T. and Yin, J. (2014). Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.* **19** no. 33, 53. MR3183577 https://doi.org/10.1214/ejp.v19-3054

[4] Bloemendal, A., Knowles, A., Yau, H.-T. and Yin, J. (2016). On the principal components of sample covariance matrices. *Probab. Theory Related Fields* **164** 459–552. MR3449395 https://doi.org/10.1007/s00440-015-0616-x
[5] Bun, J., Allez, R., Bouchaud, J.-P. and Potters, M. (2016). Rotational invariant estimator for general noisy matrices. *IEEE Trans. Inform. Theory* **62** 7475–7490. MR3599095 https://doi.org/10.1109/TIT.2016.2616132
[6] Bun, J., Bouchaud, J.-P. and Potters, M. (2017). Cleaning large correlation matrices: Tools from random matrix theory. *Phys. Rep.* **666** 1–109. MR3590056 https://doi.org/10.1016/j.physrep.2016.10.005
[7] Ding, X. (2017). Asymptotics of empirical eigen-structure of general covariance matrices. Available at arXiv:1708.06296.
[8] Ding, X. (2020). Supplement to "High dimensional deformed rectangular matrices with applications in matrix denoising." https://doi.org/10.3150/19-BEJ1129SUPP.
[9] Ding, X. (2019). Singular vector distribution of sample covariance matrices. *Adv. in Appl. Probab.* **51** 236–267. MR3984017 https://doi.org/10.1017/apr.2019.10
[10] Ding, X. and Yang, F. (2018). A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *Ann. Appl. Probab.* **28** 1679–1738. MR3809475 https://doi.org/10.1214/17-AAP1341
[11] Donoho, D.L. (1995). De-noising by soft-thresholding. *IEEE Trans. Inform. Theory* **41** 613–627. MR1331258 https://doi.org/10.1109/18.382009
[12] Elad, M. (2010). *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. New York: Springer. MR2677506 https://doi.org/10.1007/978-1-4419-7011-4
[13] Gavish, M. and Donoho, D.L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inform. Theory* **60** 5040–5053. MR3245370 https://doi.org/10.1109/TIT.2014.2323359
[14] Gavish, M. and Donoho, D.L. (2017). Optimal shrinkage of singular values. *IEEE Trans. Inform. Theory* **63** 2137–2152. MR3626861 https://doi.org/10.1109/TIT.2017.2653801
[15] Golub, G.H. and Van Loan, C.F. (1996). *Matrix Computations*, 3rd ed. *Johns Hopkins Studies in the Mathematical Sciences*. Baltimore, MD: Johns Hopkins Univ. Press. MR1417720
[16] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. *Springer Texts in Statistics* **103**. New York: Springer. MR3100153 https://doi.org/10.1007/978-1-4614-7138-7
[17] Knowles, A. and Yin, J. (2013). The isotropic semicircle law and deformation of Wigner matrices. *Comm. Pure Appl. Math.* **66** 1663–1750. MR3103909 https://doi.org/10.1002/cpa.21450
[18] Knowles, A. and Yin, J. (2014). The outliers of a deformed Wigner matrix. *Ann. Probab.* **42** 1980–2031. MR3262497 https://doi.org/10.1214/13-AOP855
[19] Knowles, A. and Yin, J. (2017). Anisotropic local laws for random matrices. *Probab. Theory Related Fields* **169** 257–352. MR3704770 https://doi.org/10.1007/s00440-016-0730-4
[20] Laloux, L., Cizeau, P., Potters, M. and Bouchaud, J. (2000). Random matrix theory and financial correlations. *Int. J. Theor. Appl. Finance* **3** 391–397.
[21] Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40** 694–726. MR2933663 https://doi.org/10.1214/12-AOS970
[22] Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** 233–264. MR2834718 https://doi.org/10.1007/s00440-010-0298-3
[23] Lee, M., Shen, H., Huang, J.Z. and Marron, J.S. (2010). Biclustering via sparse singular value decomposition. *Biometrics* **66** 1087–1095. MR2758496 https://doi.org/10.1111/j.1541-0420.2010.01392.x
[24] Marčenko, V.A. and Pastur, L.A. (1967). Distribution of eigenvalues for some sets of random matrices. *Math. USSR, Sb.* **1** 457.

[25] Nadakuditi, R.R. (2014). OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Trans. Inform. Theory* **60** 3002–3018. MR3200641 https://doi.org/10.1109/TIT.2014.2311661

[26] Nadakuditi, R.R. and Edelman, A. (2008). Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *IEEE Trans. Signal Process.* **56** 2625–2638. MR1500236 https://doi.org/10.1109/TSP.2008.917356

[27] Passemier, D. and Yao, J. (2014). Estimation of the number of spikes, possibly equal, in the high-dimensional case. *J. Multivariate Anal.* **127** 173–183. MR3188885 https://doi.org/10.1016/j.jmva.2014.02.017

[28] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

[29] Pillai, N.S. and Yin, J. (2014). Universality of covariance matrices. *Ann. Appl. Probab.* **24** 935–1001. MR3199978 https://doi.org/10.1214/13-AAP939

[30] Pizzo, A., Renfrew, D. and Soshnikov, A. (2013). On finite rank deformations of Wigner matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 64–94. MR3060148 https://doi.org/10.1214/11-AIHP459

[31] Pontes, B., Giráldez, R. and Aguilar-Ruiz, J. (2015). Biclustering on expression data: A review. *BMC Bioinform.* **57** 163–180.

[32] Renfrew, D. and Soshnikov, A. (2013). On finite rank deformations of Wigner matrices II: Delocalized perturbations. *Random Matrices Theory Appl.* **2** 1250015, 36. MR3039820 https://doi.org/10.1142/S2010326312500153

[33] Silverstein, J.W. (2009). The Stieltjes transform and its role in eigenvalue behavior of large dimensional random matrices. In *Random Matrix Theory and Its Applications. Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.* **18** 1–25. Hackensack, NJ: World Sci. Publ. MR2603192 https://doi.org/10.1142/9789814273121_0001

[34] Tao, T. (2012). *Topics in Random Matrix Theory. Graduate Studies in Mathematics* **132**. Providence, RI: Amer. Math. Soc. MR2906465 https://doi.org/10.1090/gsm/132

[35] Tracy, C.A. and Widom, H. (1996). On orthogonal and symplectic matrix ensembles. *Comm. Math. Phys.* **177** 727–754. MR1385083

[36] Tufts, D. and Shah, A. (1993). Estimation of a signal waveform from noisy data using low-rank approximation to a data matrix. *IEEE Trans. Signal Process.* **41** 7475–7490.

[37] Yang, D., Ma, Z. and Buja, A. (2016). Rate optimal denoising of simultaneously sparse and low rank matrices. *J. Mach. Learn. Res.* **17** Paper No. 92, 27. MR3543498

Received August 2017 and revised November 2018