OPTSHRINK: AN ALGORITHM FOR IMPROVED LOW-RANK SIGNAL MATRIX DENOISING BY OPTIMAL, DATA-DRIVEN SINGULAR VALUE SHRINKAGE

RAJ RAO NADAKUDITI

Abstract. The truncated singular value decomposition (SVD) of the measurement matrix is the optimal solution to the *representation* problem of how to best approximate a noisy measurement matrix using a low-rank matrix. Here, we consider the (unobservable) *denoising* problem of how to best approximate a low-rank signal matrix buried in noise by optimal (re)weighting of the singular vectors of the measurement matrix. We exploit recent results from random matrix theory to exactly characterize the large matrix limit of the optimal weighting coefficients and show that they can be computed directly from data for a large class of noise models that includes the i.i.d. Gaussian noise case.

Our analysis brings into sharp focus the shrinkage-and-thresholding form of the optimal weights, the non-convex nature of the associated shrinkage function (on the singular values) and explains why matrix regularization via singular value thresholding with convex penalty functions (such as the nuclear norm) will always be suboptimal. We validate our theoretical predictions with numerical simulations, develop an implementable algorithm (OptShrink) that realizes the predicted performance gains and show how our methods can be used to improve estimation in the setting where the measured matrix has missing entries.

1. Introduction

Techniques for low-rank signal matrix extraction from a signal-plus-noise matrix appear prominently in many statistical signal processing [84, 77, 43], machine learning [29, 46], estimation and classification applications [49]. In many applications, the low-rank approximation is the first step in an inferential process (see, for e.g. [83, 26, 75, 86, 45, 38]). These techniques are necessary whenever the $n \times m$ signal-plus-noise data or measurement matrix formed by, for example lining up the m samples or measurements of $n \times 1$

2000 *Mathematics Subject Classification.* 15A52, 46L54, 60F99.

Key words and phrases. Random matrices, Haar measure, free probability, phase transition, random eigenvalues, random eigenvectors, random perturbation, sample covariance matrices.

This work was supported by an ONR Young Investigator Award N000141110660, an AFRL subcontract from Solid State Scientific (PM Mike Noyola), AFOSR Young Investigator Award FA9550-12-1-0266 and a ARO MURI grant W911NF-11-1-0391. The author thanks Jeff Fessler for teaching him about the role of non-convex penalty functions and Iain Johnstone for suggesting the phrase 'observable solution to an unobservable problem' to emphasize the surprising nature of the optimum. We thank Florent Benaych-Georges for many stimulating conversations and Jack Silverstein for discussions on the importance of the low-coherence condition in the missing entries setting. W e thank Brendan Farrell for his suggestion to analyze the missing data problem using the signal-plus-noise-plus-small-perturbation framework and the reviewers for their suggestions. A short (conference) version containing some of the ideas in this paper first appeared in [60].

observation vectors alongside each other, can be modeled as

$$\widetilde{X} = \sum_{i=1}^{r} \theta_i u_i v_i^H + X, \tag{1}$$

where ^{H} denotes the conjugate transpose and u_i and v_i are left and right "signal" singular vectors associated with singular values θ_i of the signal matrix

$$S = \sum_{i=1}^{r} \theta_i u_i v_i^H \tag{2}$$

and X is the noise-only matrix of random (not necessarily i.i.d.) noises. These models also arise in other graph signal processing type settings; see for example [62, Text before (9)], [63, Section V], [47, Section III.A] or the various models described in [21].

Relative to this model the objective is to form an estimate of the low-rank signal matrix assuming, for now, that its rank r is known. The truncated singular value decomposition (SVD) plays a prominent role in a widely-used 'optimal' solution to a problem that is addressed by the famous Eckart-Young-Mirsky (henceforth, EYM) theorems [30, 59, 35]. Specifically, if $||\cdot||_F$ denotes the matrix Frobenius norm then the solution to the constrained optimization problem

$$\widehat{S}_{\text{eym}} = \underset{\text{rank}(S)=r}{\arg\min} ||\widetilde{X} - S||_F, \tag{3}$$

is given by

$$\widehat{S}_{\text{eym}} = \sum_{i=1}^{r} \widehat{\sigma}_i \widehat{u}_i \widehat{v}_i^H,$$

where $\widetilde{X} = \sum_i \widehat{\sigma}_i \widehat{u}_i \widehat{v}_i^H$ is the SVD of $\widetilde{X}$. This is also the maximum likelihood (ML), rank r estimate when X is assumed to be a matrix with i.i.d. Gaussian entries since the negative log-likelihood function is precisely the right hand side of (3). Its use is also justified in the small n, large m (or vice versa) regime, whenever local asymptotic normality [55] has 'kicked in'.

A natural extension is to consider settings where the signal matrix is low rank and has some additional exploitable structure. Examples include low-rank and sparse (see the body of work on sparse principal component analysis. e.g. [89, 90, 40, 27, 39, 74, 88, 7]), low rank and Toeplitz structured (see e.g. [84, 13, 87]), low rank and Hankel structured [56] and low rank and nonnegative [10, 53, 23]; see [24, 58] for an excellent overview of these methods and additional references. As expected, by exploiting structure in the signal matrix we can improve estimation performance relative to the EYM estimator which assumes no structure besides the low-rank condition.

1.1. **Denoising by optimally weighted approximation.** Here we place ourselves in the setting where *no structure is assumed in the low-rank signal matrix* and ask how the EYM estimator can be improved. The starting point for our investigation is the observation that as formulated in (3), the EYM estimator solves the *representation problem* of finding the best rank r approximation of the *signal-plus-noise* measurement matrix. It says nothing about the *denoising* problem of how to best estimate the low-rank *signal matrix*, even though practitioners sometimes invoke it as though it does. Thus we should not expect the EYM estimator to be the optimal solution to the denoising problem.

Let $||w||_{\ell_0} = |\{\#i : w_i \neq 0\}|$ so that $||w||_{\ell_0} = r$ denotes a vector w with r non-zero entries. In this paper, we consider variations of the denoising problem formulated as a weighted approximation problem of the form

$$w^{\text{opt}} := \underset{||w||_{\ell_0}=r}{\arg\min} || \sum_{i=1}^{r} \theta_i u_i v_i^H - \sum_i w_i \widehat{u}_i \widehat{v}_i^H ||_F. \tag{4}$$

Note that in (4), we are trying to approximate the unknown signal matrix using the singular vectors estimated from the noisy measurement matrix. Our setup is different from other weighted low-rank approximation problems considered in the literature as in [66], which involve weighted modifications of the problem in (3). In our formulation, setting $w_i = \widehat{\sigma}_i$ recovers the EYM estimator so that by inspecting the solution we can directly assess when and the extent to which the EYM estimator will be suboptimal.

We prove, using recent results from random matrix theory [5], that for a large class of noise models, which includes but goes well beyond the i.i.d. Gaussian model, we can compute w^{opt} in closed-form in the large matrix limit. The computation shows that w^{opt} depends only on (an integral transform of) the limiting singular value distribution of the noise-only matrix X. We then exploit this fact to develop a concrete algorithm for computing a consistent (in a sense we make precise) estimate of the limiting oracle solution directly from measurement matrix.

1.2. **Form of the optimal shrinkage-and-thresholding operator.** The analysis shows that w_i^{opt} takes the form of a shrinkage-and-thresholding operator (on the singular values of $\widetilde{X}$) that is completely characterized by the limiting singular value distribution of the noise-only matrix. The resulting shrinkage function is non-convex with $w_i^{\text{opt}} \approx \widehat{\sigma}_i(1 - O(1/\widehat{\sigma}_i^2))$ for large $\widehat{\sigma}_i$ and $w_i^{\text{opt}} \to 0$ for $\widehat{\sigma}_i \leq b + o(1)$ where b is a critical threshold that depends on the limiting noise-only singular value distribution.

The shrinkage portion of the solution arises because $\widehat{\sigma}_i$ is positively biased relative to θ_i and because the corresponding singular vectors of $\widetilde{X}$ are biased, noisy estimates of the (true) singular vectors of the latent signal matrix [5]. The thresholding portion of the solution arises because of a phase transition in the 'informativeness' of the estimated singular vectors, relative to the latent singular vectors whereby for $\theta_i > \theta_c$ inner-products of the form $(\widehat{u}_i^H u_i)$ and $(v_i^H \widehat{v}_i)$ are $O(1)$ and tend to a constant, while for $\theta_i < \theta_c$, inner-products of the form $(\widehat{u}_i^H u_i)$ and $(v_i^H \widehat{v}_i)$ are $o(1)$ and tend to zero.

Our analysis of the structure of the optimal solution 1) brings into sharp focus the form of the optimal shrinkage-and-thresholding operator, 2) provides insight on why the EYM estimator is near optimal in the low noise regime but sub-optimal in the moderate to high noise regime and 3) explains why we can expect that soft thresholding (of singular value) operators with convex penalty functions (such as the nuclear norm [14]) that are tuned to be near-optimal in the small θ_i regime will be suboptimal in the large θ_i regime (and vice versa).

1.3. **Mitigating the effect of rank over-estimation.** It is a delightful fact that even though the optimization problem in (4) is unobservable, because it depends on the unknown matrix we are trying to estimate, the optimal solution itself is computable. We assume no structure, other than low rank, on the signal matrix; the exploitable structure

is present in the 'noise portion' of the eigen-spectrum, *i.e.*, the $\min(m,n)-r$ singular values of $\widetilde{X}$.

This makes contact with the important question of how to estimate r in (1) so that one may distinguish the 'signal portion' of the eigen-spectrum from the 'noise portion'. The problem has been completely solved for the setting where X has i.i.d. Gaussian entries. In this setting, the recentering and rescaling constants that must be applied to the largest eigenvalue of XX^H to produce the Tracy-Widom distribution can be precisely characterized and used to set the appropriate threshold; see [25, 41, 2, 3, 42, 31, 71, 85, 64, 67, 68, 51, 52, 65, 69]. Recent work on the universality of this limiting distribution [80, 34, 32, 73, 12, 72] provides a rigorous justification for using essentially the same method in the non-Gaussian setting.

Similarly, when the columns of X are i.i.d. and each column has a (non-identity) population covariance matrix with a known (limiting) eigen-distribution, then the results in [31] facilitate computation of the appropriate threshold for distinguishing the 'noise portion' of the eigen-spectrum from the 'signal portion'.

If the form of population covariance matrix is misspecified then applying the tests based on this theory will lead to an overestimation of the rank of the signal matrix. Developing robust estimators of the signal rank that "work" without having to specify the symmetry structure (e.g. i.i.d. elements, i.i.d. columns, variance profile, etc.) of the noise random matrix remains an important open problem. Such estimators will have to exploit (symmetry-independent) 'universal' features of the spectrum in a way that present estimators do not.

This is where the algorithm we have developed really shines. Our algorithm takes as its input an estimate of the rank of the signal matrix and returns a (re)weighted approximation that largely mitigates the effect of rank overestimation in a manner that the EYM estimate cannot. Thus, advances in robust rank estimation when used with our algorithm will lead to improved signal matrix approximation. If the rank is correctly estimated, then the algorithm will better estimate weak subspace components of the signal matrix than the EYM algorithm.

1.4. **Contributions.** Characterizing the limiting solution of (4), computing the resulting limiting squared error, quantifying the improvement relative to the EYM estimator and developing an implementable algorithm that realizes these performance gains are the main contributions of this paper. Some of the ideas in this paper were initially presented in a conference paper by the author [60], in the context of the i.i.d Gaussian noise setting. This version goes beyond the Gaussian setting considered there. We also treat the setting where measurement matrix has missing entries, as considered in [22, 33, 16, 18, 17, 48]. In addition to rigorous results, we formulate some (empirically validated and theoretically justified) conjectures for the structure of the solution for various 'rank-regularized' variations of (4).

In related work, Hachem et al [36] looked at the problem of structured subspace estimation arising in the context of parameter estimation in large arrays. They propose an oracle solution [36, Equation (13), pp. 435] and analyze its first and second order performance in the context of the MUSIC direction-of-arrival estimator.

If we were to apply the ideas and techniques developed in this paper to the problem

$$w^{\text{opt}} := \underset{||w||_{\ell_0}=r}{\arg\min} || \sum_{i=1}^{r} u_i u_i^H - w_i \widehat{u}_i \widehat{u}_i^H ||_F^2,$$

then, we would recover a solution that corresponds to their oracle solution. Here, we consider the problem of estimating the low-rank matrix; our results and our new algorithm can be analyzed using the techniques in [36] to provide insights on the first and second order convergence properties. We leave the extension of our techniques to the estimation of projection matrices is relatively straightforward as an exercise to the reader.

The paper is organized as follows. The setup, the main theoretical results and a new algorithm based on the theoretical analysis are presented in Section 2. Simulation results to validate the theoretical predictions and a comparison of our method to other matrix regularization methods are contained in Section 3.

2. Main results and a new algorithm

2.1. **Setup and Notation.** Let X_n be an $n \times m$ ($n \leq m$, without loss of generality[1]) random matrix whose ordered singular values we denote by $\sigma_1(X_n) \geq \cdots \geq \sigma_n(X_n)$. Let μ_{X_n} be the empirical singular value distribution, *i.e.*, the probability measure defined as

$$\mu_{X_n} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\sigma_i(X_n)}.$$

Assume that the probability measure μ_{X_n} converges almost surely weakly, as $n, m \longrightarrow \infty$, to a non-random compactly supported probability measure μ_X that is supported on $[a, b]$. We assume that $\sigma_1 \xrightarrow{\text{a.s.}} b$, where $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence. These conditions are satisfied by the model where X_n has i.i.d. entries mean zero entries with variance $1/m$ and bounded higher order moments.

For a given $r \geq 1$, let $\theta_1 > \cdots > \theta_r > 0$ be deterministic non-zero real numbers, chosen independently of n. For every n, let S_n be an $n \times m$ signal matrix having rank r with its r non-zero distinct singular values equal to $\theta_1, \ldots, \theta_r$.

We suppose that X_n and S_n are independent and that X_n, the noise-only matrix is bi-unitarily invariant while the low-rank signal matrix S_n is deterministic. Recall that a random matrix is said to be *bi-orthogonally invariant* (or *bi-unitarily invariant*) if its distribution is invariant under multiplication on the left and right by orthogonal (or unitary) matrices. Alternately, if S_n has isotropically random right (or left) singular vectors, then X_n need not be unitarily invariant under multiplication on the right (or left, resp.) by orthogonal or unitary matrices. Equivalently, X_n can have deterministic right and left singular vectors while S_n can have isotropically random left and right singular vectors and we would get the same result stated shortly.

A matrix X_n with i.i.d. Gaussian entries satisfies these assumption; our results extend well beyond the Gaussian setting. The main advantage of modeling the noise matrices as having isotropically random singular vectors is that it allows us to characterize the solution

[1]We choose this convention to simplify the definition of the empirical singular value distribution.

in terms of just the (marginal) singular value distribution of the noise-only matrix instead of having to model the full joint distribution of the elements of the noise-only matrix.

Since the singular value distribution of the noise-only part can be estimated from the singular value distribution of the signal-plus-noise matrix, we can develop a concrete, data-driven algorithm, presented in Section 2.5, that can applied to real-world datasets to improve low-rank signal matrix recovery.

We observe a signal-plus-noise matrix $\widetilde{X}_n$ modeled as,

$$\widetilde{X}_n = S_n + X_n,$$

where the signal matrix S is modeled as in (2). For $i = 1, \ldots, q = \min(n, m)$, let $\widehat{u}_i$ and $\widehat{v}_i$ denote the left and right singular vectors of $\widetilde{X}$ (we suppress the subscript n) associated with the singular value $\widehat{\sigma}_i$. The solution to the optimization problem

$$w^{\text{eym}} = \underset{||w||_{\ell_0}=r}{\arg\min} ||\widetilde{X} - \sum_i w_i \widehat{u}_i \widehat{v}_i^H||_F, \tag{5}$$

is given by $w_i^{\text{eym}} = \widehat{\sigma}_i$ for $i = 1, \ldots, r$. This yields the rank r signal matrix estimate $\sum_{i=1}^r w_i^{\text{eym}} \widehat{u}_i \widehat{v}_i^H$ which, by the EYM theorem, is also the solution to the representation problem in (3).

For $w \in \mathbb{R}^l$, define the squared error as

$$\text{SE}(w) = ||S - \sum_{i=1}^{l} w_i \, \widehat{u}_i \widehat{v}_i^H||_F^2. \tag{6}$$

Consider the denoising optimization problem

$$w^{\text{opt}} := \underset{w=[w_1 \ \cdots w_r]^T \in \mathbb{R}_+^r}{\arg\min} \text{SE}(w). \tag{7}$$

We now characterize w^{opt} exactly (for every n) and provide an expression for its limiting value. In what follows, for a function f and $c \in \mathbb{R}$, we set

$$f(c^+) := \lim_{z \downarrow c} f(z).$$

2.2. **Theoretical results.**

Theorem 2.1 (Weighting coefficients)**.** *The solution to (7) exhibits the following behavior in the asymptotic regime where $n, m \to \infty$ and $n/m \to c \in [0, \infty)$. We have that for every $1 \le i \le r$,*

a)

$$w_i^{\text{opt}} = \left(\Re\{ \sum_{j=1}^{r} \theta_j (\widehat{u}_i^H u_j) \, (v_j^H \widehat{v}_i) \} \right)_+ \xrightarrow{a.s.} -2 \frac{D_{\mu_X}(\rho_i)}{D'_{\mu_X}(\rho_i)} \qquad \text{if } \theta_i^2 > 1/D_{\mu_X}(b^+), \tag{8}$$

where $x_+ = \max(0, x)$ and $\rho_i = D_{\mu_X}^{-1}(1/\theta_i^2)$.

b)

$$w_i^{\text{eym}} = \widehat{\sigma}_i \xrightarrow{a.s.} \begin{cases} D_{\mu_X}^{-1}(1/\theta_i^2) = \rho_i & \text{if } \theta_i^2 > 1/D_{\mu_X}(b^+), \\ b & \text{otherwise.} \end{cases} \tag{9}$$

In a) and b), $D_{\mu_X}(\cdot)$ is the D-transform of μ_X defined as

$$D_{\mu_X}(z) := \left[\int \frac{z}{z^2 - t^2} \mathrm{d}\mu_X(t)\right] \times \left[c \int \frac{z}{z^2 - t^2} \mathrm{d}\mu_X(t) + \frac{1-c}{z}\right] \qquad \text{for } z \notin \operatorname{supp} \mu_X,$$

and $D^{-1}_{\mu_X}(\cdot)$ denotes its functional inverse.

c) A straightforward consequence of a) and b) is that

$$w_i^{\mathrm{eym}} > w_i^{\mathrm{opt}},$$

almost surely. Also, $w_i^{\mathrm{eym}} \xrightarrow{a.s.} w_i^{\mathrm{opt}}$ as $\theta_i \to \infty$.

The emergence of the D transform in the limit characterization of the EYM and optimal coefficients follows from the results in [5]. There it was shown that, in the large matrix limit, the principal singular values and singular vectors of $\widetilde{X}$ can be completely characterized in terms of the singular values of the signal matrix and the D-transform of the limiting noise-only singular value distribution. This is why, in Theorem 2.1, the limiting values of w_i^{eym} and w_i^{opt} only depend on the singular values θ_i (or ρ_i) of the signal matrix and the limiting noise-only singular value distribution μ_X.

The D-transform is the analog of the log-Fourier transform in the sense that it describes how the distribution of the singular values of the sums of 'freely' independent matrices are related to the distribution of the singular values of the individual matrices [6]. In that sense it is an *asymptotically sufficient statistic* and hence its appearance in Theorem 2.1 is rather natural. See Section 2.5 of [5] for additional remarks.

We now characterize the limiting squared error for the optimal, EYM and other estimators with arbitrary weights.

Theorem 2.2 (Limiting squared error)**.** *Assuming that for $i = 1, \ldots, r$, $\theta_i^2 > 1/D_{\mu_X}(b^+)$. Then in the asymptotic regime considered in Theorem 2.1, the squared error, defined as in (6), exhibits the following limiting behavior:*

$$\mathrm{SE}(w) \xrightarrow{a.s.} \sum_{i=1}^{r} \left(\theta_i^2 + w_i^2 + \frac{4w_i}{\theta_i^2 D'_{\mu_X}(\rho_i)}\right)$$

Consequently,

a)

$$\mathrm{SE}(w^{\mathrm{opt}}) \xrightarrow{a.s.} \sum_{i=1}^{r} \left(\theta_i^2 - \frac{4}{(\theta_i^2 D'_{\mu_X}(\rho_i))^2}\right),$$

b)

$$\mathrm{SE}(w^{\mathrm{eym}}) \xrightarrow{a.s.} \sum_{i=1}^{r} \left(\theta_i^2 + \rho_i^2 + \frac{4\rho_i}{\theta_i^2 D'_{\mu_X}(\rho_i)}\right).$$

More generally

c)

$$\mathrm{SE}(w) - \mathrm{SE}(w^{\mathrm{opt}}) \xrightarrow{a.s.} \sum_{i=1}^{r} \left(w_i + \frac{2}{\theta_i^2 D'_{\mu_X}(\rho_i)}\right)^2$$

so that by construction

$$\mathrm{SE}(w^{\mathrm{opt}}) < \mathrm{SE}(w^{\mathrm{eym}}),$$

almost surely.

Theorem 2.2 reveals that whenever $w_i - w_i^{\mathrm{opt}}$ is large, we can expect a significant increase in SE relative to the optimal estimator. The next result reveals the shrinkage-and-thresholding form of the optimal estimator.

Theorem 2.3 (Shrinkage-and-thresholding form and resulting SE)**.** *When $r = 1$, let the sole non-zero singular value of S be denoted by θ and assume that $D'_{\mu_X}(b^+) = -\infty$. Then in the asymptotic regime considered, we have that*

$$w_1^{\mathrm{opt}} \xrightarrow{a.s.} \begin{cases} \dfrac{-2}{\theta^2 D'_{\mu_X}(\rho)} & \text{if } \theta^2 > 1/D_{\mu_X}(b^+) \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\rho = D^{-1}_{\mu_X}(1/\theta^2)$.

Consequently,

$$\mathrm{SE}(w^{\mathrm{opt}}) \xrightarrow{a.s.} \begin{cases} \theta^2 - \dfrac{4}{(\theta^2 D'_{\mu_X}(\rho))^2} & \text{if } \theta^2 > 1/D_{\mu_X}(b^+) \\ \\ \theta^2 & \text{otherwise.} \end{cases}$$

whereas

$$\mathrm{SE}(w^{\mathrm{eym}}) \xrightarrow{a.s.} \begin{cases} \theta^2 + \rho^2 + \dfrac{4\rho}{\theta^2 D'_{\mu_X}(\rho)} & \text{if } \theta^2 > 1/D_{\mu_X}(b^+) \\ \\ \theta^2 + b^2 & \text{otherwise.} \end{cases}$$

Theorem 2.3 shows that when b (the a.s. limit of the largest noise-only singular value) is $O(1)$, we can expect an $O(1)$ decrease in SE, relative to the EYM estimator, by thresholding whenever $\theta^2 < 1/D_{\mu_X}(b^+)$. Note that when X is i.i.d. Gaussian with mean zero, variance $1/m$ entries, then $b = (1 + \sqrt{c})$ and $D'_{\mu_X}(b^+) = -\infty$ so that these results apply. More generally, whenever μ_X exhibits a square-root decay at b then $D'_{\mu_X}(b^+) = -\infty$ will be satisfied. Silverstein and Choi [79] show that a large class of (non i.i.d.) Gaussian noise models will satisfy this condition.

2.3. **The missing data with i.i.d. noise setting.** We now consider the setting where $\widetilde{X}$ has missing entries so that the signal-plus-noise matrix is modeled as

$$\widetilde{X} = \left(\sum_{i=1}^{r} \theta_i u_i v_i^H + X \right) \odot M \tag{10}$$

where

$$M_{ij} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

and $\odot$ denotes the Hadamard or element-wise product. Consider the optimization problem

$$w^{\text{opt}} := \underset{w=[w_1 \ \cdots w_r]^T \in \mathbb{R}^r_+}{\arg\min} \, || \sum_{i=1}^{r} p\,\theta_i u_i v_i^H - \sum_{i=1}^{r} w_i \widehat{u}_i \widehat{v}_i^H ||_F^2. \tag{11}$$

Note that here we are approximating pS instead of S as in (6) (so that we can use the data-driven algorithm as-is). Setting $w_i^{\text{opt}} \mapsto w_i^{\text{opt}}/p$ will yield a solution to the denoising problem in (6). Let $||w||_\infty = \max_i |w_i|$ denote the element of the vector w with the maximum absolute value.

Theorem 2.4. *Assume that the singular vectors u_i and v_i in (10) satisfy a 'low-coherence' condition in the following sense: we suppose that there exist non-negative constants η_u, C_u, η_v and C_v, independent of n, such that for $i = 1, \ldots, r$*

$$\max_i ||u_i||_\infty \leq \eta_u \frac{\log^{C_u} n}{\sqrt{n}} \qquad \text{and} \qquad \max_i ||v_i||_\infty \leq \eta_v \frac{\log^{C_v} m}{\sqrt{m}}. \tag{12}$$

Let the elements of X_{ij} be i.i.d with mean zero, variance $1/m$ and bounded higher order moments. Then the solution to (11) exhibits the following limiting behavior. We have that for $p \in (0, 1]$ and $i = 1, \ldots, r$

a)

$$w_i^{\text{eym}} = \sigma_i(\widetilde{X}) \xrightarrow{a.s.} \begin{cases} \sqrt{p} \cdot \sqrt{\dfrac{(1 + p\,\theta_i^2)(c + p\,\theta_i^2)}{p\,\theta_i^2}} & \text{if } \theta_i > \dfrac{c^{1/4}}{\sqrt{p}}, \\ \sqrt{p}\,(1 + \sqrt{c}) & \text{otherwise.} \end{cases}$$

b)

$$w_i^{\text{opt}} \xrightarrow{a.s.} p\,\theta_i \cdot \sqrt{1 - \frac{c(1 + p\,\theta_i^2)}{p\,\theta_i^2(p\,\theta_i^2 + c)}} \sqrt{1 - \frac{c + p\,\theta_i^2}{p\,\theta_i^2(p\,\theta_i^2 + 1)}} \qquad \text{if } \theta_i > \frac{c^{1/4}}{\sqrt{p}}.$$

c) When $r = 1$

$$w_i^{\text{opt}} \xrightarrow{a.s.} 0 \qquad \text{if } \theta_i \leq \frac{c^{1/4}}{\sqrt{p}},$$

Theorem 2.4 is a statement about the optimality of the shrinkage-and-thresholding form when there are missing entries in the signal-plus-noise matrix. Note that in this case, the equivalent noise-only matrix will not bi-unitarily invariant when X is non-Gaussian. The proof (see Section 6), however, reveals that it asymptotically behaves as though it does so that the results of Theorems 2.1 and 2.3 still apply. Note that as a consequence, Theorem 2.2 can applied to compute the result asymptotic squared error. After the submission of this paper, we learned of recent work by Shabalin and Nobel for the $p = 1$ setting of Theorem 2.4 with i.i.d. Gaussian noise; see [78].

2.4. **The asymptotic equivalence of various rank-regularized estimators.** Let us define the *effective rank*, r_{eff}, of the signal matrix as

$$r_{\text{eff}} = \text{the number of } i \in \{1, \dots r\} \text{ such that } \theta_i^2 > 1/D_{\mu_X}(b^+). \tag{13}$$

Thus, the effective rank quantifies the number of singular values in the signal-plus-noise matrix $\widetilde{X}$ that are 'informative', *i.e.*, reveal the existence of a low-rank signal matrix. Clearly, $r_{\text{eff}} \leq r$ but $r_{\text{eff}} < r$ whenever the number of singular values that separate from the right edge b of the spectrum is less than the latent signal matrix rank r. The following conjecture formalizes their relation to the number of 'informative' singular vectors in the signal-plus-noise matrix.

Conjecture 2.5 (Uninformativeness below phase transition)**.** *Assume that* $D'_{\mu_X}(b^+) = -\infty$ *and* [2] *that for fixed* $\widehat{r}$*,*

$$\max_i \left(\sigma_i(X) - \sigma_{i+1}(X)\right) \leq O\left(\frac{\log n \text{ factors}}{n^{2/3}}\right),$$

with very high probability. Then we have that for $r_{\text{eff}} < i \leq \widehat{r}$ *and* $j = 1, \dots r$*,*

$$\max_{i,j} |(\widehat{u}_i^H u_j)| \leq O\left(\frac{\log n \text{ factors}}{n^{1/6}}\right) \text{ and } \qquad \max_{i,j} |(v_j^H \widehat{v}_i)| \leq O\left(\frac{\log m \text{ factors}}{m^{1/6}}\right),$$

with high enough probability that we can establish their almost sure convergence to zero.

We now consider the principal rank-regularized optimization problem

$$w^{\text{opt}}(\widehat{r}) := \underset{w=[w_1 \ \cdots w_{\widehat{r}}]^T \in \mathbb{R}_+^{\widehat{r}}}{\arg\min} \ || \sum_{i=1}^{r} \theta_i u_i v_i^H - \sum_{i=1}^{\widehat{r}} w_i \widehat{u}_i \widehat{v}_i^H ||_F^2. \tag{14}$$

We characterize the structure of the optimal estimator and the resulting MSE next.

Corollary 2.6 (Performance with estimated principal component rank)**.** *Let* $\widehat{r}$ *be a fixed (with* n*) estimate of* r_{eff} *and* r_{eff} *be defined as in (13). Then, in the asymptotic regime considered, assuming Conjecture 2.5 holds, we have that*

$$w_i^{\text{opt}}(\widehat{r}) \xrightarrow{a.s.} \begin{cases} -2\dfrac{D_{\mu_X}(\rho_i)}{D'_{\mu_X}(\rho_i)} & \text{for } i \leq r_{\text{eff}} \\ \\ 0 & \text{otherwise.} \end{cases}$$

and hence

$$\text{SE}(w^{\text{opt}}) \xrightarrow{a.s.} \sum_{i=1}^{\min(r_{\text{eff}}, \widehat{r})} \left(\theta_i^2 - \frac{4}{(\theta_i^2 D'_{\mu_X}(\rho_i))^2}\right) + \sum_{i=\min(r_{\text{eff}}, \widehat{r})+1}^{\max(r, \widehat{r})} \theta_i^2,$$

whereas

$$\text{SE}(w^{\text{eym}}) \xrightarrow{a.s.} \sum_{i=1}^{\min(r_{\text{eff}}, \widehat{r})} \left(\theta_i^2 + \rho_i^2 + \frac{4\rho_i}{\theta_i^2 D'_{\mu_X}(\rho_i)}\right) + \sum_{i=\min(r_{\text{eff}}, \widehat{r})+1}^{\max(r, \widehat{r})} (\theta_i^2 + b^2),$$

[2] Note that these conditions are met when X has i.i.d. entries of variance $1/m$. See Theorem 2.10 of [8].

where $\rho_i = D_{\mu_X}^{-1}(1/\theta_i)$ *and we set* $\theta_i = 0$ *for* $i > r$. *Consequently,*

$$\mathrm{SE}(w^{\mathrm{eym}}) - \mathrm{SE}(w^{\mathrm{opt}}) > \sum_{i=1}^{\min(r_{\mathrm{eff}},\widehat{r})} \left(\rho_i + \frac{2}{\theta_i^2 D'_{\mu_X}(\rho_i)}\right)^2 + \sum_{i=\min(r_{\mathrm{eff}},\widehat{r})+1}^{\max(r,\widehat{r})} b^2 > 0,$$

almost surely.

Corollary 2.6 reveals that the optimal estimator can realize a significant improvement in performance relative to the EYM estimator whenever $b = O(1)$ and $r_{\mathrm{eff}} < r$. The corollary highlights the importance of reliably estimating r_{eff} instead of r. Now, consider the rank regularized optimization problem

$$\overline{w}^{\mathrm{opt}}(\widehat{r}) := \underset{w\in\mathbb{R}^q, ||w||_{\ell_0}=\widehat{r}}{\arg\min} || \sum_{i=1}^{r} \theta_i u_i v_i^H - \sum_{i=1}^{q} w_i \widehat{u}_i \widehat{v}_i^H ||_F \tag{15}$$

We characterize the exact solution next.

Theorem 2.7 (Optimal rank regularized solution)**.** *For arbitrary integer* $1 \le \widehat{r} \le q$, *the solution to (15) is given by*

$$\overline{w}^{\mathrm{opt}}(\widehat{r}) = \ell_{\widehat{r}} \left[\{ \Re(\sum_{j=1}^{r} \theta_j (\widehat{u}_i^H u_j) \, (v_j^H \widehat{v}_i))_+ \}_{i=1}^{q} \right],$$

where for $x \in \mathbb{R}_+^q$, $\ell_{\widehat{r}}(x)$ *returns a* $q \times 1$ *vector whose* $\widehat{r}$ *non-zero elements equal the* $\widehat{r}$ *largest entries of* x *while the remaining entries are identically zero.*

We state a conjecture on the delocalization of the bulk singular vectors and characterize the asymptotic limit of (15) next.

Conjecture 2.8 (Complete delocalization of bulk singular vectors)**.** *Define* $q = \min(m, n)$. *Assume that* $D'_{\mu_X}(b^+) = -\infty$ *and that for all* $1 \le i \le q$, *where* i *depends on* n

$$\max_i \left(\sigma_i(X) - \sigma_{i+1}(X)\right) \le O\left(\frac{\log n \text{ factors}}{n}\right),$$

with very high probability. Then, we have that for large enough n *and every* $i_n > r_{\mathrm{eff}}$ *and* $j = 1, \dots r$

$$\max_{i,j} |(\widehat{u}_i^H u_j)| \le O\left(\frac{\log n \text{ factors}}{n^{1/2}}\right) \text{ and } \qquad \max_{i,j} |(v_j^H \widehat{v}_i)| \le O\left(\frac{\log m \text{ factors}}{m^{1/2}}\right),$$

with high enough probability that we can establish their almost sure convergence to zero.

Corollary 2.9 (Limiting rank regularized weights)**.** *Assuming Conjectures 2.5 and 2.8 hold, we have that*

$$\overline{w}_i^{\mathrm{opt}}(\widehat{r}) \xrightarrow{a.s.} \begin{cases} -2\dfrac{D_{\mu_X}(\rho_i)}{D'_{\mu_X}(\rho_i)} & \text{for } i = 1, \dots, \min(r_{\mathrm{eff}}, \widehat{r}) \\ \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, even though, for finite n

$$\mathrm{SE}(\overline{w}^{\mathrm{opt}}(q)) \le \mathrm{SE}(\overline{w}^{\mathrm{opt}}(r_{\mathrm{eff}})) \le \mathrm{SE}(w^{\mathrm{opt}}(r_{\mathrm{eff}})),$$

Algorithm 1 OptShrink: A new algorithm for low-rank matrix denoising by optimal, data-driven singular value shrinkage.

Input: $\widetilde{X} = n \times m$ signal-plus-noise matrix
Input: $\widehat{r}$ = Estimate of the effective rank of the latent low-rank signal matrix
Compute $\widetilde{X} = \sum_{i=1}^{q} \widehat{\sigma}_i \widehat{u}_i \widehat{v}_i^H$
Compute $\widehat{\Sigma}_{\widehat{r}} = \mathrm{diag}(\widehat{\sigma}_{\widehat{r}+1}, \dots \widehat{\sigma}_q) \in \mathbb{R}^{(n-\widehat{r})\times(m-\widehat{r})}$
for $i = 1, \dots \widehat{r}$ **do**
    Compute $\widehat{D}(\widehat{\sigma}_i; \widehat{\Sigma}_{\widehat{r}})$ using (16a) and $\widehat{D}'(\widehat{\sigma}_i; \widehat{\Sigma}_{\widehat{r}})$ using (16b)
    Compute $\widehat{w}_{i,\widehat{r}}^{\mathrm{opt}} = -2 \, \dfrac{\widehat{D}(\widehat{\sigma}_i; \widehat{\Sigma}_{\widehat{r}})}{\widehat{D}'(\widehat{\sigma}_i; \widehat{\Sigma}_{\widehat{r}})}$
end for
return $\widehat{S}_{\mathrm{opt}} = \sum_{i=1}^{\widehat{r}} \widehat{w}_{i,\widehat{r}}^{\mathrm{opt}} \, \widehat{u}_i \, \widehat{v}_i^H$ = denoised estimate of the rank $\widehat{r}$ signal matrix
return (optional) Compute estimate of MSE using (17a)
return (optional) Compute estimate of relative MSE using (17b)

as $n \to \infty$ we have that

$$\mathrm{SE}(\overline{w}^{\mathrm{opt}}(q)) - \mathrm{SE}(\overline{w}^{\mathrm{opt}}(r_{\mathrm{eff}})) \xrightarrow{a.s.} 0 \qquad \text{and} \qquad \mathrm{SE}(\overline{w}^{\mathrm{opt}}(q)) - \mathrm{SE}(w^{\mathrm{opt}}(r_{\mathrm{eff}})) \xrightarrow{a.s.} 0.$$

Corollary 2.9 shows that when there is delocalization in the singular vectors then, in the large matrix limit, optimal performance is attained by estimating the effective rank r_{eff}, applying shrinkage to the informative r_{eff} components and thresholding (to zero) the remaining components. In other words, there are vanishing (with n) performance losses when the coefficients given by $w^{\mathrm{opt}}(r_{\mathrm{eff}})$ are used in place of $\overline{w}^{\mathrm{opt}}(q)$. We believe that Conjectures 2.5 and 2.8 hold in the signal-plus-noise matrix with missing entries setting considered in Section 2.3 so that Corollaries 2.6 and 2.9 will apply there as well. This is pertinent because we now describe an algorithm for consistently estimating w^{opt} *directly from data* by exploiting the information in the singular value spectrum of the signal-plus-noise matrix.

2.5. **A new algorithm for improved denoising.** Equation (8) shows that the optimal estimator in the large matrix limit is given by

$$w_i^{\mathrm{opt}} = -2 \frac{D_{\mu_X}(\rho_i)}{D'_{\mu_X}(\rho_i)} + o(1),$$

where ρ_i is the large matrix limit of the i-th largest singular value. In the finite n, m setting, for $i = 1, \dots, r_{\mathrm{eff}}$, $\widehat{\rho}_i = \widehat{\sigma}_i$ is a biased, but asymptotically consistent estimator of ρ_i. We now describe an algorithm for estimating w_i^{opt} using a single signal-plus-noise matrix.

For a matrix $X \in \mathbb{K}^{n\times m}$. Define

$$\widehat{D}(z; X) := \frac{1}{n} \mathrm{Tr}\left(z \, (z^2 \, I - X X^H)^{-1}\right) \cdot \frac{1}{m} \mathrm{Tr}\left(z \, (z^2 \, I - X^H X)^{-1}\right) \tag{16a}$$

and

$$\widehat{D}'(z;X) := \frac{1}{n}\operatorname{Tr}\left[z\left(z^2\,I - XX^H\right)^{-1}\right]\cdot\frac{1}{m}\operatorname{Tr}\left[-2z^2\left(z^2\,I - X^HX\right)^{-2} + \left(z^2\,I - X^HX\right)^{-1}\right]$$
$$+\frac{1}{m}\operatorname{Tr}\left[z\left(z^2\,I - X^HX\right)^{-1}\right]\cdot\frac{1}{n}\operatorname{Tr}\left[-2z^2\left(z^2\,I - XX^H\right)^{-2} + \left(z^2\,I - XX^H\right)^{-1}\right]. \quad (16b)$$

By construction (and the definition of the D-transform), $\widehat{D}(z;X) \xrightarrow{\text{a.s.}} D_{\mu_X}(z)$ and $\widehat{D}'(z;X) \xrightarrow{\text{a.s.}} D'_{\mu_X}(z)$ for z outside the support of μ_X. We now show how the spectrum of $\widetilde{X}$ can be used to estimate μ_X. To that end, we establish a useful identify by first defining

$$\mu_{X,\widehat{r}} = \frac{1}{n-\widehat{r}}\sum_{i=\widehat{r}+1}^{n}\delta_{\sigma_i(X_n)}.$$

Then, it is easy to see that for fixed (with n) $\widehat{r}$, $\mu_{X_{\widehat{r}}} \xrightarrow{\text{a.s.}} \mu_{X,0}$. Thus, if

$$\widehat{\Sigma}_{\widehat{r}} = \operatorname{diag}(\widehat{\sigma}_{\widehat{r}+1},\ldots\widehat{\sigma}_q) \in \mathbb{R}^{(n-\widehat{r})\times(m-\widehat{r})}$$

is a diagonal matrix containing the $q-\widehat{r}$ "noise" singular values of $\widetilde{X}$, then, by construction, and whenever $\widehat{\sigma}_i \xrightarrow{\text{a.s.}} \rho_i > b$, then $\widehat{D}(\widehat{\sigma}_i;\widehat{\Sigma}_{\widehat{r}}) \xrightarrow{\text{a.s.}} D_{\mu_X}(\rho_i)$ and $\widehat{D}'(\widehat{\sigma}_i;\widehat{\Sigma}_{\widehat{r}}) \xrightarrow{\text{a.s.}} D'_{\mu_X}(\rho_i)$. Hence, we form a consistent estimate of w_i^{opt} as described in Algorithm 1. The methods described in Section 1.3 can be used to form an estimate of $\widehat{r}$.

By Theorem 2.2, we can compute an estimate of the absolute and relative mean squared error (defined as $\text{MSE}/||S||_F^2$) as

$$\widehat{\text{MSE}}_{\widehat{r}} = \sum_{i=1}^{\widehat{r}}\frac{1}{\widehat{D}(\widehat{\sigma}_i;\widehat{\Sigma}_{\widehat{r}})} - \sum_{i=1}^{\widehat{r}}(\widehat{w}_{i,\widehat{r}}^{\text{opt}})^2 \quad (17a)$$

$$\widehat{\text{relMSE}}_{\widehat{r}} = 1 - \frac{\sum_{i=1}^{\widehat{r}}(\widehat{w}_{i,\widehat{r}}^{\text{opt}})^2}{\sum_{i=1}^{\widehat{r}}\frac{1}{\widehat{D}(\widehat{\sigma}_i;\widehat{\Sigma}_{\widehat{r}})}}, \quad (17b)$$

respectively. A value for $\widehat{\text{relMSE}}_{\widehat{r}}$ near 0 indicates very good low-rank signal matrix approximation while a value near 1 indicates a poor approximation. These metrics might be better proxies for the noisiness of a signal-plus-noise matrix than the condition number or the spectral gap. We conclude with a statement of the theoretical consistency of the w_i^{opt} produced by Algorithm 1.

Theorem 2.10. *Assume that $\widehat{r} = r_{\text{eff}}$. Then for $1 \le i \le \widehat{r}$, we have that*

$$\widehat{w}_{i,\widehat{r}}^{\text{opt}} \xrightarrow{a.s.} -2\frac{D_{\mu_X}(\rho_i)}{D'_{\mu_X}(\rho_i)}$$

Proof. This is a straightforward consequence of Theorem 2.1-a) and the fact that the almost sure limit of (16a) leads (as described in the introduction of [5]) directly to the D-transform. □

3. Numerical Validation, Discussion and Extensions

We now numerically validate our predictions. In the experiments that follow, we consider the model in (1) with $r = 1$, $n = m = 400$ and select X to be an $n \times m$ matrix with i.i.d. $\mathcal{N}(0, 1/m)$ entries. For various values of θ, Figure 1-a) compares empirically computed w_1^{opt} averaged over 100 trials with the (limiting) theoretical prediction given by the $p = 1$ result in Theorem 2.4. Figure 1-b) compares the realized normalized MSE and shows that the EYM solution is near-optimal for large values of θ but far from sub-optimal for small values of θ. The simulations validate the shrinkage-and-thresholding form of the solution for w_1^{opt} given by Theorem 2.3 and show that Algorithm 1 realizes the predicted performance gains.

We now consider the optimization problem in (14) and evaluate the performance of the various algorithms for various values of $\widehat{r}$ for $\theta = 10$ and $\theta = 2$. Here, $r_{\text{eff}} = 1$ and Corollary 2.6 predicts that the optimal (oracle) algorithm should significantly outperform the EYM algorithm whenever $\widehat{r} > r_{\text{eff}}$. Figure 2 shows the validity of this prediction and also shows that even though Algorithm 1 is suboptimal, relative to the oracle estimator, it is able to largely mitigate the effect of r_{eff} overestimation due to the shrinkage effect.

Figure 3 compares the normalized MSE estimates as a function of θ, produced by Algorithm 1 to the empirical values for the setting where $\widehat{r} = r = 1$ and $\theta_1 = \theta$ and where $\widehat{r} = r = 2$, $\theta_1 = 20$ and $\theta_2 = \theta$. As expected the estimates, produced are accurate whenever $r_{\text{eff}} = \widehat{r}$.

We now validate Theorem 2.4. We fix $r = 1$ and $\theta_1 = \theta = 2$ in (10) and vary p, the proportion of entries with missing data. We sample u_1 and v_1 uniformly at random from the unit hypersphere so that the low-coherence conditions in Theorem 2.4 are met. Theorem 2.4 predicts that $w_1^{\text{opt}} \to 0$ (asymptotically) when $p < \sqrt{n/m}/\theta^2 = 0.25$. Figure 4 shows the accuracy of the prediction and the significant improvement in performance of the oracle estimator and Algorithm 1 relative to the EYM estimator.

3.1. **Suboptimality of singular value thresholding.**

We now compare our algorithm to regularized matrix estimates obtained as the solution to the optimization problem

$$\widehat{S}_{\text{svt},\lambda} = \arg\min_S ||\widetilde{X} - S||_F^2 + 2\lambda||S||_*, \tag{18}$$

where $||\cdot||_*$ is the nuclear norm (or the sum of the singular values of the argument matrix). The optimization problem in (18) yields the closed-form solution [14]

$$\widehat{S}_{\text{svt},\lambda} = \widehat{U}\,\text{diag}((\widehat{\sigma} - \lambda)_+)\widehat{V}^H. \tag{19}$$

The resulting singular value thresholded (SVT) matrix corresponds to the weighting

$$w_{\text{svt},i}(\lambda) = \begin{cases} \widehat{\sigma}_i - \lambda & \text{if } \widehat{\sigma}_i > \lambda \\ 0 & \text{otherwise.} \end{cases}$$

Figure 5-a) and b) compare the resulting soft-thresholding operator associated with the SVT approximation with the optimal and the EYM solutions for $\lambda = 1, 2$ as a function of θ and w^{eym}, respectively for the same $r = 1$, $n = m$ setting in (1) with X_{ij} i.i.d. $\mathcal{N}(0, 1/m)$. Here $b = (1 + \sqrt{c}) = 2$.

While SVT with $\lambda = 2$ can yield comparable shrinkage (in the small θ regime) and thresholding (below $\theta = 1$) as the optimal estimator, $w_{\text{svt}}(2) - w^{\text{opt}}$ will be large for moderate θ so that by Theorem 2.2-c) we expect SVT to be suboptimal for larger values of θ. Figure 6 compares the performance of Algorithm 1 and the optimal estimator to the SVT algorithm with $\lambda = 1$ and $\lambda = 2$. SVT is significantly suboptimal as expected. Our results show that our algorithm would outperform SVT with convex shrinkage functions for any of the general family of noise models considered here.

3.2. **Better singular value shrinkage with non-convex potential functions?** A closer examination of Figure 5-a) and b) reveals that the optimal estimator shrinks less for larger values of θ than the SVT possibly can. In fact, the optimal estimator will generically yield a non-convex shrinkage function which scales as

$$w_i^{\text{opt}} \approx \widehat{\sigma}_i \left(1 - O\left(\frac{1}{\widehat{\sigma}_i^2}\right)\right),$$

for large $\widehat{\sigma}_i$. Might singular value shrinkage with other non-convex potential functions generically outperform convex potential functions as well? These would be the non-convex analogs in the matrix setting of the non-negative Garrotte estimator [11] in the vector setting. Fully understanding their benefits and shortfalls, relative to Algorithm 1, remains an open line of inquiry.

3.3. **Role of informative components.** We conclude by reexamining the role of the principal (or leading) r_{eff} singular vectors of $\widetilde{X}$ in the solution of the optimization problem (15). Theorem 2.7 shows that we should take the components $\widehat{u}_i$ and $\widehat{v}_i$ for which the inner product $(\widehat{u}_i^H u_i)$ and $(v_i^H \widehat{v}_i)$ is $O(1)$. The supposition in (7) is that the principal components are these components.

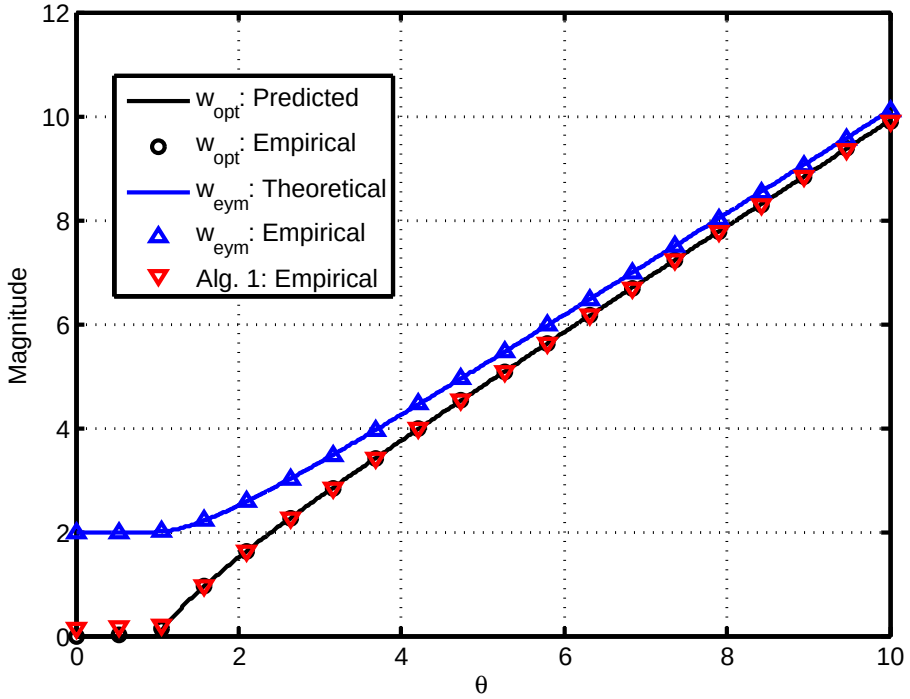
However, in an expository paper by the author [61], it is shown that if the (limiting) spectrum of the noise-only matrix is supported on two disconnected intervals, then the middle components can be more informative than the principal components. Thus, while this work (via Theorem 2.2) brings into focus the importance of accurately estimating r_{eff}, it is equally important to be able to identify the most informative components. The development of fast, accurate algorithms for the same for large matrix-valued datasets remains an important open problem.

3.4. **Extensions.** We have initial numerical evidence that the algorithm presented here outperforms the EYM estimator for the variety of applications described in [21], even though they do not exactly fit the noise matrix models analyzed here. Extending the analysis of our algorithm to these models would shed further insight on the limits of low-rank signal matrix approximation.

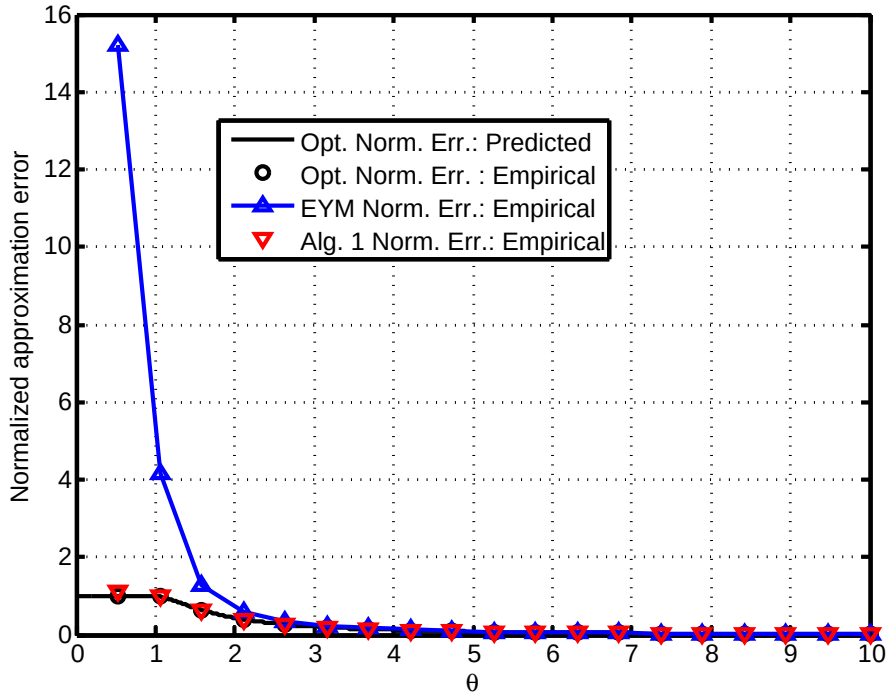
We conclude by listing some directions of future research. These include 1) rigorously establishing the delocalization conjectures, 2) designing penalty functions that are robust to noise model mismatch, 3) clarifying the benefits, if any, of matrix regularization [44, 28, 50] with convex or non-convex penalty functions relative to rank regularized solutions for the unstructured low-rank signal matrix setting, 4) extending the methods developed to problems involving estimation of signal matrices with an unstructured low-rank component and a sparse [15, 19, 20, 81, 70] or diagonal [76] component or low-rank

structured component [24] and 5) developing minimax estimators, along the lines of the work in [21], except for the more general class of noise models considered here.

Lastly, consider Theorem 2.3, where it is shown that for $\theta < 1/D_{\mu_X}(b^+)$, $\mathrm{SE}(w_1^{\mathrm{opt}}) \xrightarrow{\text{a.s.}} \theta^2$. In this regime, is there another (non-SVD based) algorithm that can estimate the signal matrix with mean-squared-error $\theta^2 - O(1)$? More generally, is there a non-SVD based algorithm that can (reliably) recover the (unstructured) low-rank signal matrix in the regime where the SVD based methods break down? This is a largely open question whose answer would better clarify the interplay between the limits of SVD-based estimation of the signal matrix singular vectors and the fundamental limits of estimation of the signal matrix itself. We leave these questions for future work.

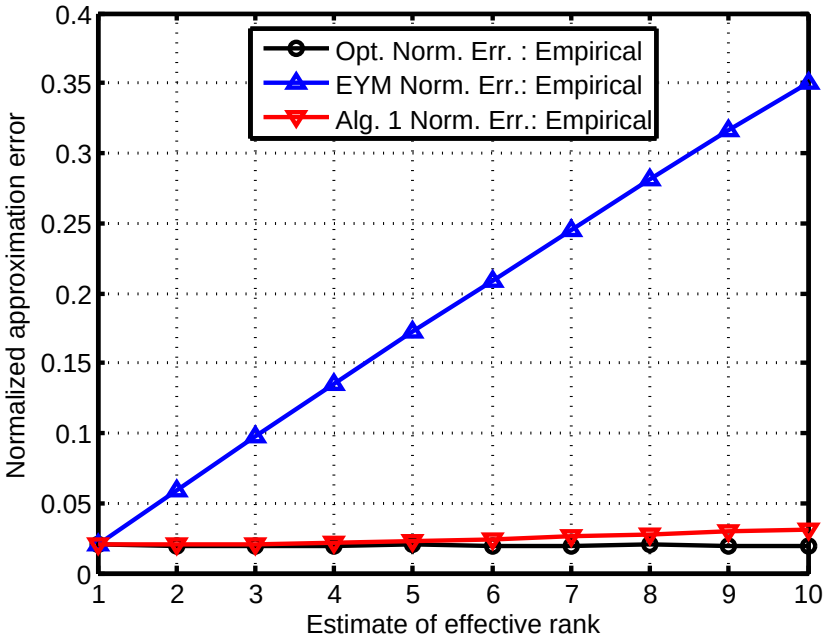


(A) w_1^{opt} and w_1^{eym} versus θ.

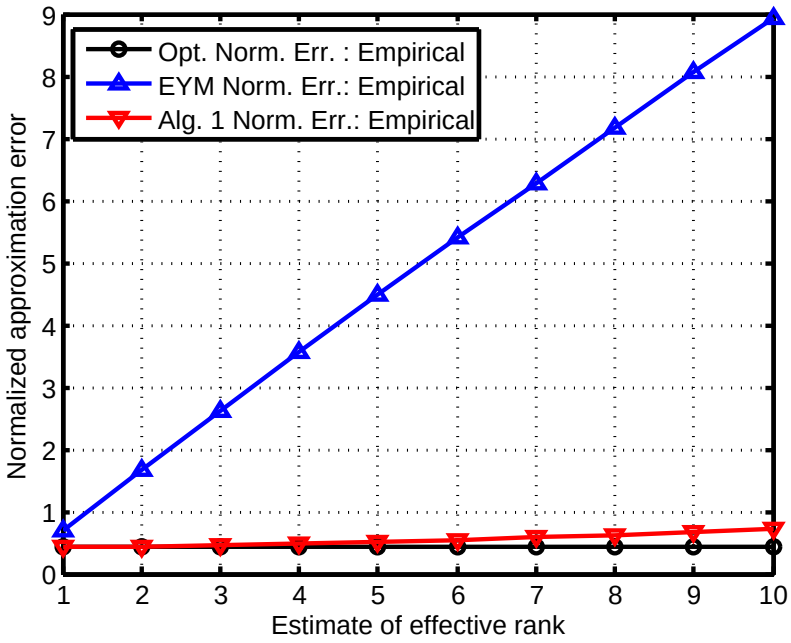


(B) $||\theta uv^H - w\widehat{u}_1\widehat{v}_1^H||_F^2/\theta^2$ versus θ.

FIGURE 1. For the model in (1) with $r = 1$, $n = m = 400$ and X an $n \times m$ matrix with i.i.d. $\mathcal{N}(0, 1/m)$ entries, for various values of $\theta := \theta_1$, (a) we compare the theoretically predicted w_1^{opt} using (8) with the w_1^{eym} computed using (9) (so that they precisely correspond to the $p = 1$ prediction in Theorem 2.4) with empirically computed values of the same (averaged over 100 trials). Here we set $\widehat{r} = 1$ in Algorithm 1. (b) plots the realized normalized approximation errors and compares them to the (oracle) performance of the optimal detector predicted in Theorem 2.2.

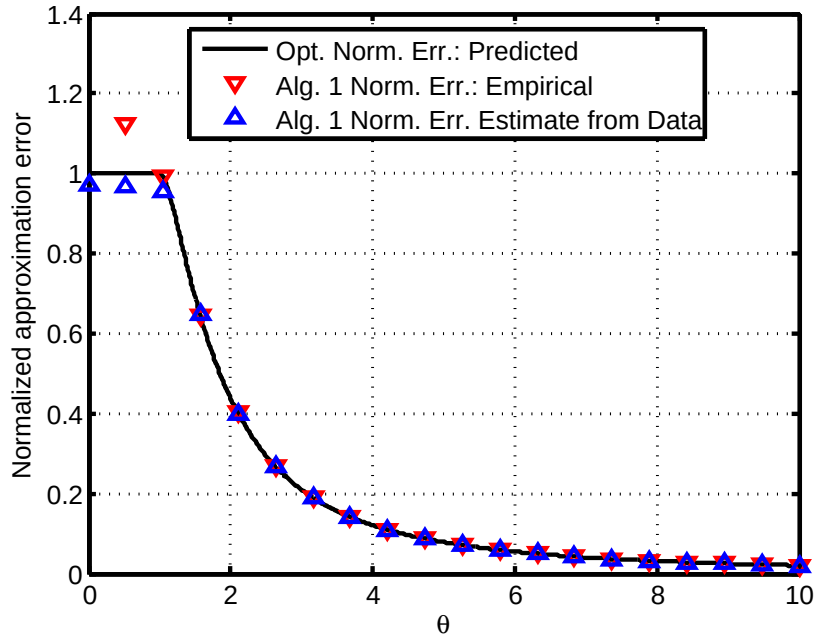


(A) Normalized MSE versus $\widehat{r}$: $\theta = 10$.

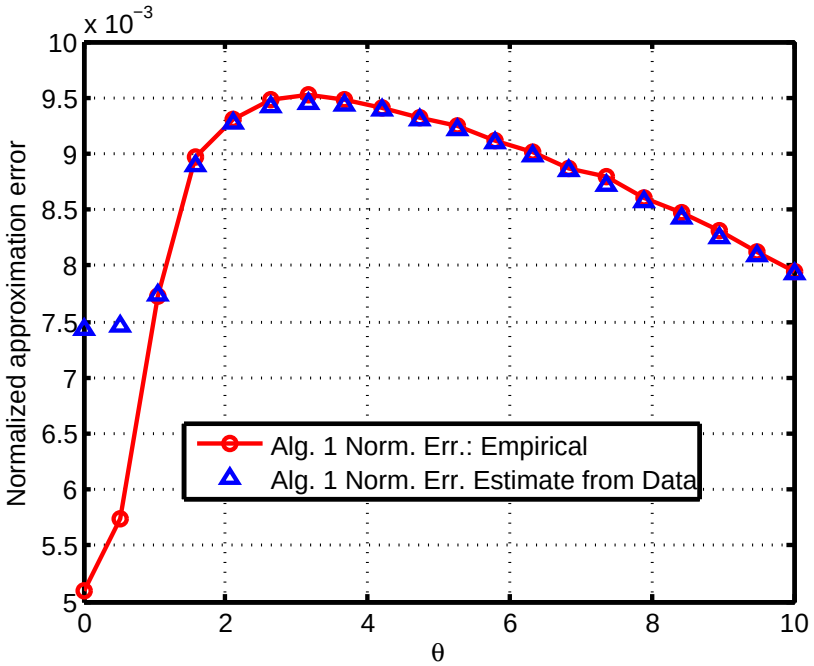


(B) Normalized MSE versus $\widehat{r}$: $\theta = 2$.

FIGURE 2. Here, we are in the same setting as in Figure 1, except, we evaluate the performance of Algorithm 1 and the EYM estimator of rank $\widehat{r}$ to that of the rank $\widehat{r}$ oracle optimal estimator computed using the left hand side of (8) for various values of $\widehat{r}$. In a), $\theta = 10$ while in b) $\theta = 2$.

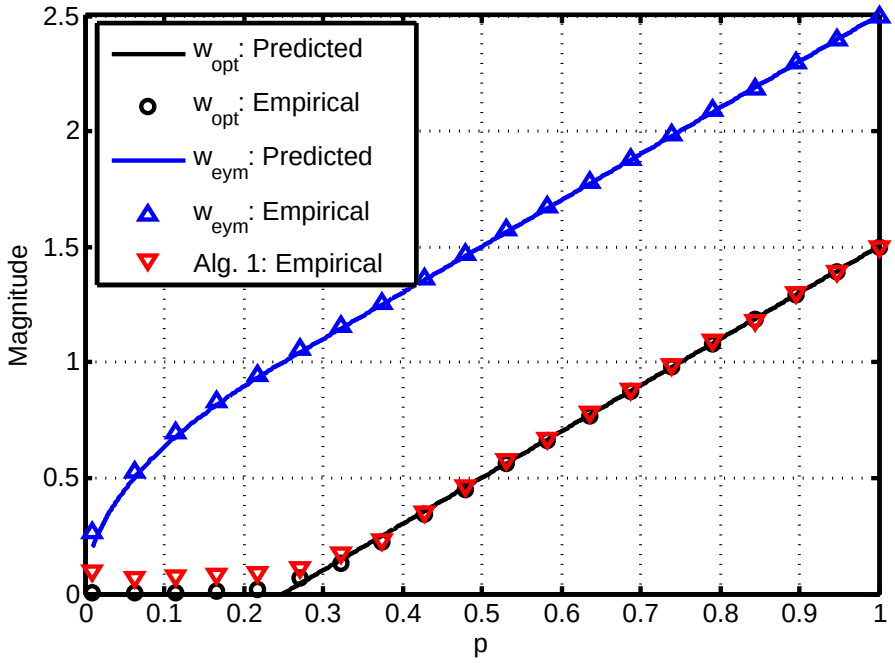


(A) $\widehat{r} = 1$, $S = \theta u v^H$

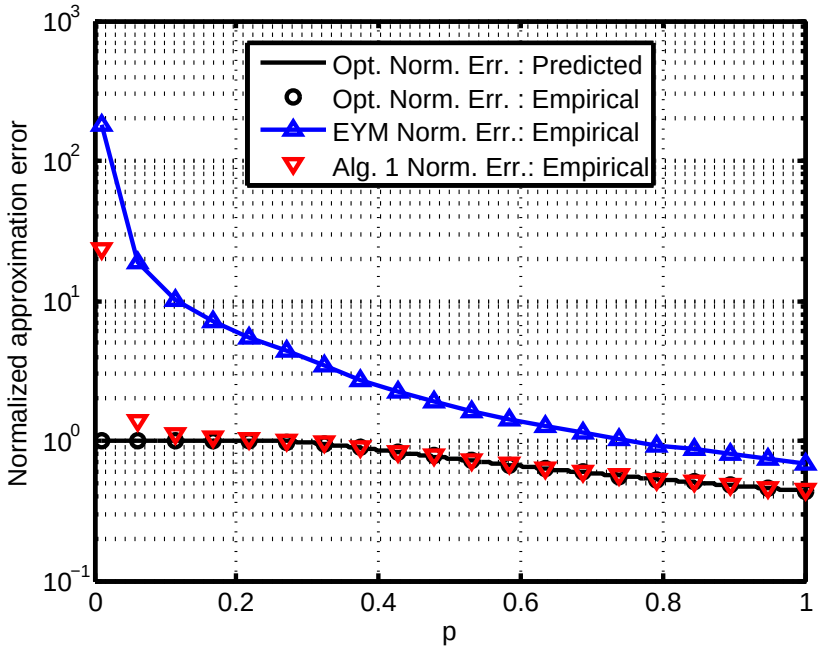


(B) $\widehat{r} = 2$, $S = 20\, u_1 v_1^H + \theta u_2 v_2^H$

FIGURE 3. Here, we compare the normalized approximation error computed empirically with the estimate $\widehat{\text{relMSE}}_{\widehat{r}}$ computed using (17b) as returned by Algorithm 1. When $\theta \leq 1$, $r_{\text{eff}} = r - 1$ so that one of the components becomes uninformative so that including it in the estimate will increase the realized error.

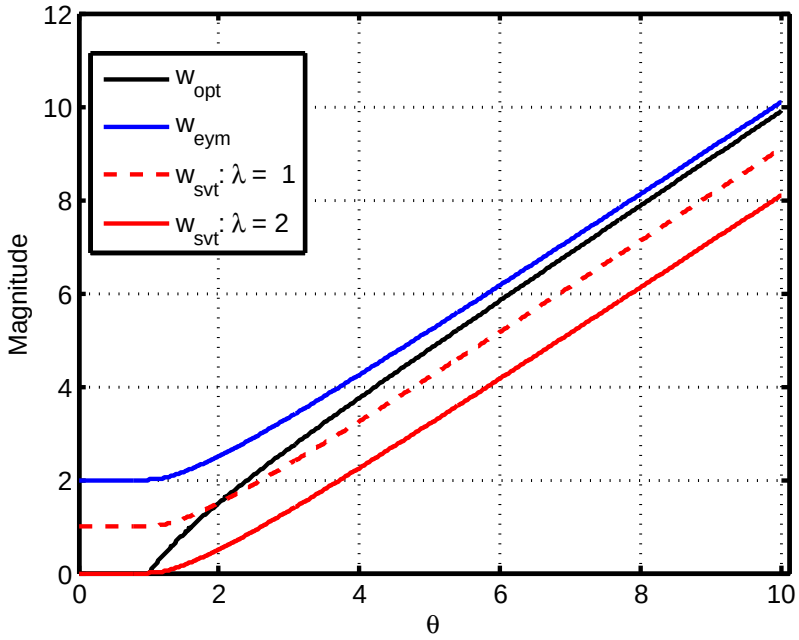


(A) w_1^{opt} and w_1^{eym} versus p.

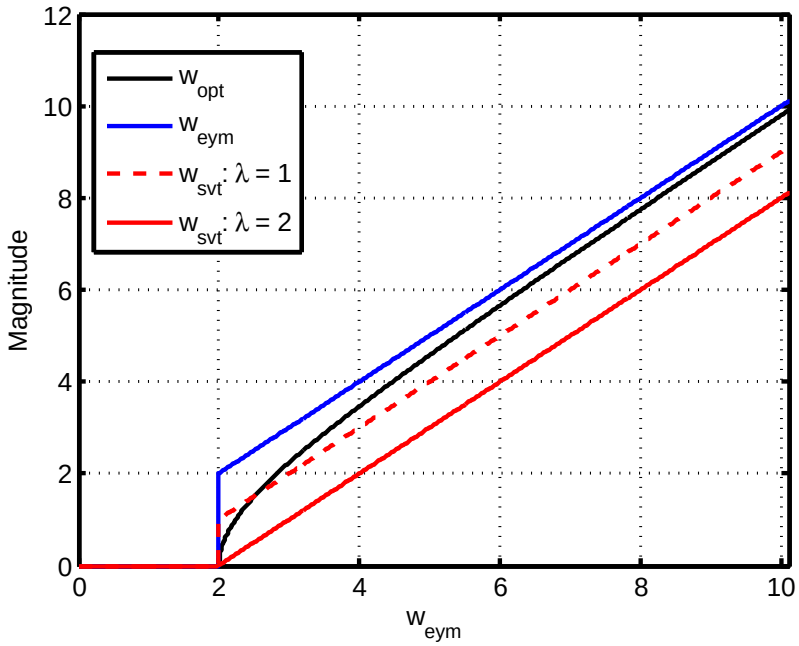


(B) $||p\theta uv^H - w\widehat{u}_1\widehat{v}_1^H||_F^2/p^2/\theta^2$ versus p; here $\theta = 2$.

FIGURE 4. For the model in (10), with $r = 1$, $n = m = 400$ and X an $n \times m$ matrix with i.i.d. $\mathcal{N}(0, 1/m)$ entries, we perform the same comparisons as in Figure 1 (averaged over 100 trials), except we fix $\theta = 2$ and instead vary p, the proportion of entries with missing data. We sample u_1 and v_1 uniformly at random from the unit hypersphere so that the low-coherence conditions in Theorem 2.4 are met. Theorem 2.4 predicts that $w^{\text{opt}} \to 0$ (asymptotically) when $p < \sqrt{n/m}/\theta^2 = 0.25$.



(A) Shrinkage and thresholding operators as a function of θ.



(B) Shrinkage and thresholding operators as a function of w^{eym}.

FIGURE 5. Here we are in the same setting as Figure 1. We plot w^{opt}, w^{eym} and $w_{\text{svt},\lambda}$ for $\lambda = 1, 2$ as a function of θ and w^{eym}. Note the non-convex nature of the shrinkage portion of the optimal shrinkage-and-thresholding operator, the optimality of the EYM solution for large values of θ (high SNR regime) and the sub-optimality of the SVT solution.

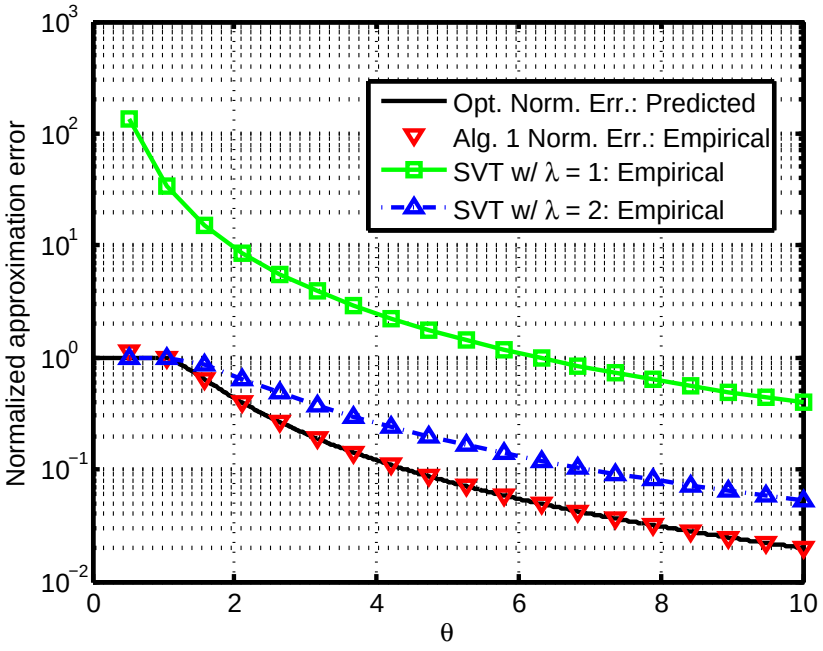


FIGURE 6. For the same setting as in Figure 1, we compare the performance of Algorithm 1 (with $\widehat{r} = 1$) with that of the SVT estimator in (19) for $\lambda = 1, 2$.

4. Proof of Theorems 2.1, 2.3 and 2.7

We first prove Theorem 2.1 -b). Since $w_i^{\text{eym}} = \widehat{\sigma}_i$, Theorem 2.1-b) follows immediately from Theorem 2.9 in [5]. Next, we prove the first part of Theorem 2.1-a) by showing that

$$w_i^{\text{opt}} = \left(\Re\{ \sum_{j=1}^{r} \theta_j (\widehat{u}_i^H u_j)\, (v_j^H \widehat{v}_i) \} \right)_+ .$$

Theorem 2.7 follows by adopting the exact same approach, with some minor modifications so we shall omit its proof. We first establish some intermediate results.

Lemma 4.1. *Let $A \in \mathbb{K}^{n\times m}$ and $q = \min(n, m)$. Consider the optimization problem*

$$D_{opt} := \underset{D=\text{diag}(\{d_1,\ldots,d_q\}),d_i\in\mathbb{R}_+}{\arg\min} ||A - D||_F,$$

where $\text{diag}(\cdot)$ *denotes a matrix with the arguments on the diagonal and zeros elsewhere (even for a rectangular matrix). Then*

$$(D_{opt})_{ii} = \max(0, \Re(A_{ii})).$$

Proof. We first solve the unconstrained problem

$$D_{opt} := \underset{D=\text{diag}(\{d_1,\ldots,d_q\})}{\arg\min} ||A - D||_F,$$

Note that

$$||A - D||_F^2 = \sum_{i=1}^{q} (A_{ii} - d_i)^2 + \underbrace{\sum_{i\neq j} A_{ij}^2}_{\text{constant}} \geq \sum_{i\neq j} A_{ij}^2,$$

so that setting $d_i = A_{ii}$ attains the lower bound. The additional constraint that $d_i \in \mathbb{R}_+$ yields the stated result which is simply a projection onto $\mathbb{R}_+$. □

Corollary 4.1. *For fixed r, the solution to the optimization problem*

$$D_{opt} := \underset{D=\text{diag}(\{d_1,\ldots d_r,0,\ldots,0\}),d_i\in\mathbb{R}_+}{\arg\min} ||A - D||_F,$$

is given by

$$(D_{\text{opt}})_{ii} = \max(0, A_{ii}) \qquad \text{for} i = 1, \ldots r.$$

Now consider the optimization problem

$$w^{\text{opt}} = \underset{w\in\mathbb{R}_+^r}{\arg\min} || \sum_{i=1}^{r} \theta_i u_i v_i^H - \sum_{i=1}^{r} w_i \widehat{u}_i \widehat{v}_i^H ||_F.$$

Let $U_r = \begin{bmatrix} u_1 & \ldots & u_r \end{bmatrix}$, $V_r = \begin{bmatrix} v_1 & \ldots & v_r \end{bmatrix}$, $\Theta_r = \text{diag}(\theta_1, \ldots, \theta_r)$, $\widehat{U} = \begin{bmatrix} \widehat{u}_i & \ldots \widehat{u}_n \end{bmatrix}$ and $\widehat{V} = \begin{bmatrix} \widehat{v}_i & \ldots \widehat{v}_m \end{bmatrix}$. Then for $W = \text{diag}(w_1, \ldots, w_r, 0, \ldots 0)$, the optimization problem can be rewritten as

$$w^{\text{opt}} = \underset{w\in\mathbb{R}_+^r, W=\text{diag}(w)}{\arg\min} ||U_r \Theta_r V_r^H - \widehat{U} W \widehat{V}^H||_F.$$

By the unitary invariance of the Frobenius norm we have that

$$||U_r \Theta_r V_r^H - \widehat{U} W \widehat{V}^H||_F = ||\widehat{U}^H U_r \Theta_r V_r^H \widehat{V} - W||_F.$$

Let $K = \widehat{U}^H U_r \Theta_r V_r^H \widehat{V}$. Then,

$$K = \begin{bmatrix} \widehat{u}_1^H u_1 & \dots & \widehat{u}_1^H u_r \\ \vdots & \vdots & \vdots \\ \widehat{u}_r^H u_1 & \dots & \widehat{u}_r^H u_r \\ \\ \widehat{u}_{r+1}^H u_1 & \dots & \widehat{u}_{r+1}^H u_r \\ \vdots & \vdots & \vdots \\ \widehat{u}_n^H u_1 & \dots & \widehat{u}_n^H u_r \end{bmatrix} \begin{bmatrix} \theta_1 & & \\ & \ddots & \\ & & \theta_r \end{bmatrix} \begin{bmatrix} v_1^H \widehat{v}_1 & \dots & v_1^H \widehat{v}_r & v_1^H \widehat{v}_{r+1} & \dots & v_1^H \widehat{v}_m \\ \dots & \dots & \dots & \dots & & \\ v_r^H \widehat{v}_1 & \dots & v_r^H \widehat{v}_r & v_r^H \widehat{v}_{r+1} & \dots & v_r^H \widehat{v}_m \end{bmatrix}$$

$$= \sum_{j=1}^{r} \theta_j \begin{bmatrix} \widehat{u}_1^H u_j \\ \vdots \\ \widehat{u}_r^H u_j \\ \\ \widehat{u}_{r+1}^H u_j \\ \vdots \\ \widehat{u}_n^H u_j \end{bmatrix} \begin{bmatrix} v_j^H \widehat{v}_1 & \dots & v_j^H \widehat{v}_r & v_j^H \widehat{v}_{r+1} & \dots & v_j^H \widehat{v}_m \end{bmatrix}$$

Expanding out the diagonal entries of K we get

$$K = \sum_{j=1}^{r} \begin{bmatrix} \theta_j \left(\widehat{u}_1^H u_j\right) \cdot \left(v_j^H \widehat{v}_1\right) & * & * \\ * & \ddots & * \\ * & * & \theta_j \left(\widehat{u}_m^H u_j\right) \cdot \left(v_j^H \widehat{v}_n\right) \end{bmatrix}$$

so that (deterministically),

$$K_{ii} = \sum_{j=1}^{r} \theta_j \widehat{u}_i^H u_j v_j^H \widehat{v}_i \tag{20}$$

and the solution

$$w_i^{\text{opt}} = \max(0, \Re \sum_{j=1}^{r} \theta_j \widehat{u}_i^H u_j v_j^H \widehat{v}_i) = (\Re \sum_{j=1}^{r} \theta_j \widehat{u}_i^H u_j v_j^H \widehat{v}_i)_+,$$

follows immediately from (20) by the application of Corollary 4.1. We have thus proved the equality on the left-hand side of Theorem 2.1-a). It is easy to see how this approach yields Theorem 2.7.

We now prove the limit characterization portion of Theorem 2.1-a). In [5, Theorem 2.10 c)], it was proved that for $j = 1, \dots, r$, and $i \neq j$ such that $\theta_i^2 > 1/D_{\mu_X}(b^+)$, $\widehat{u}_i^H u_j \xrightarrow{\text{a.s.}} 0$ and $v_j^H \widehat{v}_i \xrightarrow{\text{a.s.}} 0$. Consequently,

$$K_{ii} = \sum_{j=1}^{r} \theta_j \, \widehat{u}_i^H u_j \cdot v_j^H \widehat{v}_i \xrightarrow{\text{a.s.}} \theta_i \, \widehat{u}_i^H u_i v_i^H \widehat{v}_i.$$

Let $\rho_i = D_{\mu_X}^{-1}(1/\theta_i)$. In [5, Theorem 2.10 c)] it was shown that

$$|\widehat{u}_i^H u_i|^2 \xrightarrow{\text{a.s.}} \frac{-2\phi_{\mu_X}(\rho_i)}{\theta_i^2 D'_{\mu_X}(\rho_i)} \qquad \text{and} \qquad |\widehat{v}_i^H v_i|^2 \xrightarrow{\text{a.s.}} \frac{-2\phi_{\widetilde{\mu}_X}(\rho_i)}{\theta_i^2 D'_{\mu_X}(\rho_i)},$$

where $\widetilde{\mu}_X = c\mu_X + (1-c)\delta_0$ and for any probability measure μ,

$$\phi_\mu(z) := \int \frac{z}{z^2 - t^2} \mathrm{d}\mu(t). \tag{21}$$

While there is ambiguity in the sign (or phase, when complex valued) of the individual singular vectors, the proof in [5] shows that

$$\widehat{u}_i^H u_i \, v_i^H \widehat{v}_i \xrightarrow{\text{a.s.}} \sqrt{\frac{-2\phi_{\mu_X}(\rho_i)}{\theta_i^2 D'_{\mu_X}(\rho_i)} \cdot \frac{-2\phi_{\widetilde{\mu}_X}(\rho_i)}{\theta_i^2 D'_{\mu_X}(\rho_i)}} = \frac{2\sqrt{\phi_{\mu_X}(\rho_i) \cdot \phi_{\widetilde{\mu}_X}(\rho_i)}}{\theta_i^2 D'_{\mu_X}(\rho_i)}. \tag{22}$$

However, $D_{\mu_X}(z) = \phi_\mu(z) \cdot \phi_{\widetilde{\mu}}(z)$ so that $\phi_\mu(\rho_i) \cdot \phi_{\widetilde{\mu}}(\rho_i) = D_{\mu_X}(\rho_i) = D_{\mu_X}(D_{\mu_X}^{-1}(1/\theta_i^2)) = 1/\theta_i^2$, so that

$$\theta_i(\widehat{u}_i^H u_i)(v_i^H \widehat{v}_i) \xrightarrow{\text{a.s.}} \frac{-2}{\theta_i^2 D'_{\mu_X}(\rho_i)} = -2\frac{D_{\mu_X}(\rho_i)}{D'_{\mu_X}(\rho_i)}. \tag{23}$$

This gives the limit on the right hand side of part a).

To prove part c), we note that $w_i^{\text{eym}} > \theta_i$ (as a consequence of Horn's interlacing inequalities [37]) while, for large enough n, $w_i^{\text{opt}} = \theta_i \widehat{u}_i^H u_i v_i^H \widehat{v}_i + o(1) < \theta_i$. Thus $w_i^{\text{eym}} > w_i^{\text{opt}}$ for large enough n. Since $\widehat{u}_i^H u_i v_i^H \widehat{v}_i \to 1$ for $\theta_i \to \infty$, $w_i^{\text{eym}} \xrightarrow{\text{a.s.}} w_i^{\text{opt}}$ as $\theta_i \to \infty$.

We now prove Theorem 2.3. Note that when $r = 1$,

$$w_1^{\text{opt}} = \left(\Re \theta_1 \widehat{u}_1^H u_1 v_1^H \widehat{v}_1\right)_+.$$

When $r = 1$ and $\theta_1^2 \le 1/D_{\mu_X}(b^+)$ and $D'_{\mu_X}(b^+) = -\infty$, then by Theorem 2.11 of [5], $\widehat{u}_1^H u_1 \xrightarrow{\text{a.s.}} 0$ and $v_1^H \widehat{v}_1 \xrightarrow{\text{a.s.}} 0$. Consequently, $w_1^{\text{opt}} \xrightarrow{\text{a.s.}} 0$ and we have established the phase transition (or shrinkage-and-thresholding form) of w_1^{opt} in Theorem 2.3. The expressions for $\mathrm{SE}(w^{\text{opt}})$ and $\mathrm{SE}(w^{\text{eym}})$ are a straightforward consequence of Theorem 2.2.

5. Proof of Theorems 2.2 and Corollaries 2.6 and 2.9

Here, we have that

$$\begin{aligned}
\mathrm{SE}(w) &= || \sum_{i=1}^r \theta_i u_i v_i^H - \sum_{i=1}^r w_i \widehat{u}_i \widehat{v}_i^H ||_F^2 \\
&= \sum_i \theta_i^2 + \sum_j w_j^2 - 2\Re \operatorname{Tr} \sum_{i,j} \theta_i w_j u_i v_i^H \widehat{v}_j \widehat{u}_j^H \\
&= \sum_i \theta_i^2 + \sum_i w_i^2 - 2 \operatorname{Tr} \sum_i \theta_i w_i u_i v_i^H \widehat{v}_i \widehat{u}_i^H - 2\,\Re \operatorname{Tr} \sum_{i \neq j} \theta_i w_j u_i v_i^H \widehat{v}_j \widehat{u}_j^H
\end{aligned}$$

In [5, Theorem 2.10 c)], it was proved that for $j = 1, \ldots, r$, and $i \neq j$ such that $\theta_i^2 > 1/D_{\mu_X}(b^+)$, $\widehat{u}_i^H u_j \xrightarrow{\text{a.s.}} 0$ and $v_j^H \widehat{v}_i \xrightarrow{\text{a.s.}} 0$. Hence,

$$\begin{aligned}\text{SE}(w) &= \sum_i (\theta_i^2 + w_i^2 - 2\theta_i w_i \widehat{u}_i^H u_i \cdot \widehat{v}_i^H v_i) - 2\,\Re \sum_{i \neq j} \theta_i w_j \underbrace{\widehat{u}_j^H u_i}_{\xrightarrow{\text{a.s.}} 0} \underbrace{v_i^H \widehat{v}_j}_{\xrightarrow{\text{a.s.}} 0} \\ &\xrightarrow{\text{a.s.}} \sum_{i=1}^r \left(\theta_i^2 + \frac{4w_i}{\theta_i^2 D'_{\mu_X}(\rho_i)} + w_i^2 \right),\end{aligned}$$

where we have substituted (23) to give us the final expression in the stated result.

Theorem 2.2-a) and b) follow from substituting the limiting values of w_i^{opt} and w_i^{eym} given by Theorem 2.1 in the derived expression. Theorem 2.2-c) follows easily by simple algebraic manipulation of the limiting expressions for $\text{SE}(w)$ and $\text{SE}(w^{\text{opt}})$. The portions of Corollaries 2.6 and 2.9 that characterize the structure of the limiting weights follows immediately from Conjecture 2.5 and Conjecture 2.8 via an application of Theorem 2.7.

We now consider the asymptotic squared error. Note that

$$\sqrt{\text{SE}(w^{\text{opt}}(r_{\text{eff}}))} - \sqrt{\text{SE}(\overline{w}^{\text{opt}})} = ||\sum_{i=1}^r \theta_i u_i v_i^H - \sum_{i=1}^q \overline{w}_i^{\text{opt}} \widehat{u}_i \widehat{v}_i^H||_F - ||\sum_{i=1}^r \theta_i u_i v_i^H - \sum_{i=1}^{r_{\text{eff}}} w_i^{\text{opt}} \widehat{u}_i \widehat{v}_i^H||_F.$$

By the triangle inequality we have

$$\sqrt{\text{SE}(w^{\text{opt}}(r_{\text{eff}}))} - \sqrt{\text{SE}(\overline{w}^{\text{opt}})} \leq ||\sum_{i=1}^{r_{\text{eff}}} (\overline{w}_i^{\text{opt}} - w_i^{\text{opt}}) u_i v_i^H||_F + ||\sum_{i=r_{\text{eff}}+1}^{q} \overline{w}_i^{\text{opt}} \widehat{u}_i \widehat{v}_i^H||_F^2.$$

Since we have just shown that $\overline{w}_i^{\text{opt}} \xrightarrow{\text{a.s.}} w_i^{\text{opt}}$ for $i = 1, \ldots, r_{\text{eff}}$, we have

$$||\sum_{i=1}^{r_{\text{eff}}} (\overline{w}_i^{\text{opt}} - w_i^{\text{opt}}) u_i v_i^H||_F^2 \xrightarrow{\text{a.s.}} 0.$$

If we can show that

$$||\sum_{i=r_{\text{eff}}+1}^{q} \overline{w}_i^{\text{opt}} \widehat{u}_i \widehat{v}_i^H||_F^2 \xrightarrow{\text{a.s.}} 0,$$

then we can conclude that $\text{SE}(\overline{w}^{\text{opt}}) - \text{SE}(w^{\text{opt}}(r_{\text{eff}})) \xrightarrow{\text{a.s.}} 0$ and we are done. To that end, we shall utilize the claim from Conjecture 2.5 that the $o(n)$ leading coefficients of $\overline{w}_i^{\text{opt}}$ corresponding to the edge (or principal) singular vectors will be bounded by $O(\log n \text{ factors}/n^{1/3})$ and the claim from Conjecture 2.8 that $O(n)$ of $\overline{w}_i^{\text{opt}}$ coefficients corresponding to the bulk singular vectors will be bounded by $O(\log n \text{ factors}/n)$ with

very high probability. This gives us

$$\begin{aligned}
||\sum_{i=r_{\text{eff}}+1}^{q} \overline{w}_i^{\text{opt}} \widehat{u}_i \widehat{v}_i^H||_F^2 &= \sum_{i>r_{\text{eff}}, i\in\text{bulk}} (\overline{w}_i^{\text{opt}})^2 + \sum_{i>r_{\text{eff}}, i\in\text{edge}} (\overline{w}_i^{\text{opt}})^2 \\
&\leq O(n) O\left(\frac{\log n \text{ factors}}{n^2}\right) + o(n) O\left(\frac{\log n \text{ factors}}{n^{2/3}}\right) \\
&\leq \underbrace{O\left(\frac{\log n \text{ factors}}{n}\right) + O\left(\frac{\log n \text{ factors}}{n^{2/3}}\right)}_{\xrightarrow{\text{a.s.}} 0}.
\end{aligned}$$

If the probability is high enough we will be able to conclude that $\text{SE}(\overline{w}^{\text{opt}}(r_{\text{eff}})) \xrightarrow{\text{a.s.}} \text{SE}(w^{\text{opt}}(r_{\text{eff}}))$. Repeating this calculation with $w^{\text{opt}}(\widehat{r})$ and utilizing Conjecture 2.5 gives us the expression for the asymptotic squared error in Corollary 2.6.

6. Proof of Theorem 2.4

We begin by recalling that

$$\widetilde{X} = (\underbrace{U\Theta V^H}_{=:S} + X) \odot M = S \odot M + X \odot M, \tag{24}$$

where X is the noise-only matrix with $\mathbb{E}[X_{ij}] = 0$ and $\text{Var}[X_{ij}] = 1/m$ and

$$M_{ij} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases}$$

Note that $\mathbb{E}_M[S \odot M] = p\,S$, so that (24) can be rewritten in a signal-plus-noise-plus-small-perturbation form[3] given by

$$\widetilde{X} = \underbrace{\mathbb{E}[S \odot M]}_{=p\,S} + Z + \underbrace{(S \odot M - \mathbb{E}[S \odot M])}_{=:\Delta_S}, \tag{25}$$

where Z is the noise-only random matrix with missing entries given by

$$Z_{ij} = \begin{cases} X_{ij} & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases} \tag{26}$$

Let

$$\overline{X} = p\,S + Z, \tag{27}$$

so that, from (25), $\widetilde{X} = \overline{X} + \Delta_S$. Let $\overline{X} = \sum_i \overline{\sigma}_i \overline{u}_i \overline{v}_i^H$ be the SVD of $\overline{X}$. In lieu of (11), consider the slightly modified optimization problem

$$\overline{w}^{\text{opt}} := \underset{w=[w_1 \ \cdots w_r]^T \in \mathbb{R}_+^r}{\arg\min} ||\sum_{i=1}^{r} p\,\theta_i u_i v_i^H - \sum_{i=1}^{r} w_i \overline{u}_i \overline{v}_i^H||_F^2. \tag{28}$$

We will first show that $\overline{w}_i^{\text{opt}}$ is characterized by the stated expression in Theorem 2.4. Then we will show that $\sigma_1(\Delta_S) \xrightarrow{\text{a.s.}} 0$, which we will utilize to prove that $w_i^{\text{opt}} \xrightarrow{\text{a.s.}} \overline{w}_i^{\text{opt}}$.

[3]Thanks to Brendan Farrell for suggesting this approach to analyzing the problem.

Comparing (7) to (28) reveals that the left hand side of Theorem 2.1-a) still holds except with $\theta_i \longmapsto p\,\theta_i$. Consequently,

$$\overline{w}_i^{\text{opt}} = \left(\Re\{ \sum_{j=1}^{r} p\,\theta_j (\overline{u}_i^H u_j)\,(v_j^H \overline{v}_i) \} \right)_+ . \tag{29}$$

We now establish the almost sure limit of the right hand side of (29).

To that end, we first note that since $\mathbb{E}[X_{ij}] = 0$ and $\mathbb{E}[X_{ij}^2] = 1/m$, from (26), we have that $\mathbb{E}[Z_{ij}] = 0$ and $\mathbb{E}[Z_{ij}^2] = p/m$. Moreover, since the higher order moments of the entries of X were assumed to be bounded, the higher order moments of the entries of Z will be bounded as well. Consequently, it can be shown [1] that

$$\mathrm{d}\mu_{Z_n}(x) \xrightarrow{\text{a.s.}} \mathrm{d}\mu_Z(x) = \frac{\sqrt{4\,p^2\,c - (x^2 - p - p\,c)^2}}{\pi\,p\,c\,x} \mathbb{1}_{(a,b)}(x) dx + \max\left(0, 1 - \frac{1}{c}\right) \delta_0, \tag{30}$$

where $a = \sqrt{p}(1-\sqrt{c})$ and $b = \sqrt{p}(1+\sqrt{c})$ are the end points of the support of μ_Z. Here, μ_Z is the famous Marčenko-Pastur distribution [57]. It is known [1], that $\sigma_1(Z) \xrightarrow{\text{a.s.}} b = \sqrt{p}(1+\sqrt{c})$. Moreover, from the results of Bloemendal et al [8, Theorems 2.4 and 2.5], we have that for any $\{u_i\}_{i=1}^r$ and $\{v_i\}_{i=1}^r$, independent of Z,

$$u_i^H (w^2 I_n - Z Z^H)^{-1} u_j \xrightarrow{\text{a.s.}} \int \frac{\mathrm{d}\mu_Z(t)}{w^2 - t^2}\,\delta_{ij} \tag{31a}$$

and

$$v_i^H (w^2 I_m - Z^H Z)^{-1} v_j \xrightarrow{\text{a.s.}} \int \frac{\mathrm{d}\mu_{\widetilde{Z}}(t)}{w^2 - t^2}\,\delta_{ij}, \tag{31b}$$

where $\mu_{\widetilde{Z}} = c\mu_Z + (1-c)\delta_0$ (when $c < 1$). An inspection of the proofs in [5] reveals that the almost sure limits of these bilinear forms determine the almost sure limits of $\sigma_i(\overline{X})$ and $(\overline{u}_i^H u_j)$ and $(v_j^H \overline{v}_i)$ for $i = 1, \ldots, r$. Equation (31) asserts that these limits are the same as the limits that we would have obtained if Z were i.i.d. Gaussian (and hence bi-unitarily invariant) with matching mean and variance as the Z in (26). Consequently, the almost sure limit of $\overline{w}_i^{\text{opt}}$ in (29) will be the same as though Z were i.i.d. Gaussian with mean zero and variance p/m entries. Hence, by Theorem 2.1-b)

$$\overline{w}_i^{\text{eym}} := \sigma_i(\overline{X}) \xrightarrow{\text{a.s.}} \begin{cases} \rho_i = D_{\mu_Z}^{-1}(1/p^2\theta_i^2) & \text{if } p^2\,\theta_i^2 > \dfrac{1}{D_{\mu_Z}(b^+)} = p\,\sqrt{c} \\ \\ \sqrt{p}\,(1+\sqrt{c}) & \text{otherwise,} \end{cases}$$

while by Theorem 2.1-a),

$$\overline{w}_i^{\text{opt}} \xrightarrow{\text{a.s.}} -2\frac{D_{\mu_Z}(\rho_i)}{D'_{\mu_Z}(\rho_i)} \qquad \text{if } \theta_i^2 > \frac{\sqrt{c}}{p}.$$

Computing the D-transform of μ_Z in (30) (see Example 3.1 in [5] for the computation when $p = 1$ from which the general p answer can be easily deduced) gives us the pertinent expression for $\overline{w}_i^{\text{opt}}$ and $\overline{w}_i^{\text{eym}}$ which match the expressions in Theorem 2.4. The $r = 1$ phase transition behavior for $\overline{w}_1^{\text{opt}}$ follows from Theorem 2.3.

From the perturbation theory of singular values [37, Theorem 3.3.16-(c), pp. 178], we have that

$$|\sigma_i(pS + Z + \Delta_S) - \sigma_i(pS + Z)| \leq \sigma_1(\Delta_S), \tag{32}$$

for $i = 1, \ldots, \min(m, n)$. Consequently

$$|w_i^{\text{eym}} - \overline{w}_i^{\text{eym}}| \leq \sigma_1(\Delta_S),$$

so if we can show that $\sigma_1(\Delta_S) \xrightarrow{\text{a.s.}} 0$ then we will have shown that $w_i^{\text{eym}} \xrightarrow{\text{a.s.}} \overline{w}_i^{\text{eym}}$ and we have proved Theorem 2.4-a).

To prove that $w_i^{\text{opt}} \xrightarrow{\text{a.s.}} \overline{w}_i^{\text{opt}}$ we need a more involved argument that requires showing that we get the same limiting behavior when $Z + \Delta_S$ is substituted for Z in the bilinear forms on the left hand side of (31). We begin by noting that

$$\begin{aligned}|u_i^H(wI_n - (Z+\Delta_S)(Z+\Delta_S)^H)^{-1}u_j - u_i^H(wI_n - ZZ^H)^{-1}u_j| \\ \leq \sigma_1((wI_n - ZZ^H)^{-1} - (wI_n - (Z+\Delta_S)(Z+\Delta_S)^H)^{-1})\end{aligned}$$

as a consequence of the variational characterization of the largest singular value. To make further progress, we shall utilize the resolvent identity[4] which states that

$$(wI - B)^{-1} - (wI - A)^{-1} = (wI - B)^{-1}(B - A)(wI - A)^{-1},$$

where $\Im w > 0$ and A and B are Hermitian matrices. Applying this identity with $A = ZZ^H$ and $B = (Z + \Delta_S)(Z + \Delta_S)^H$ yields

$$\begin{aligned}|u_i^H(wI_n - (Z+\Delta_S)(Z+\Delta_S)^H)^{-1}u_j - u_i^H(wI_n - ZZ^H)^{-1}u_j| \\ \leq \sigma_1((wI_n - (Z+\Delta_S)(Z+\Delta_S)^H)^{-1}(\Delta_S\Delta_S^H + \Delta_S Z^H + Z\Delta_S^H)(wI_n - ZZ^H)^{-1}) \\ \leq \frac{1}{|\Im w|^2} \cdot \sigma_1(\Delta_S\Delta_S^H + \Delta_S Z^H + Z\Delta_S^H).\end{aligned}$$

Since $\sigma_1(AB) \leq \sigma_1(A) \cdot \sigma_1(B)$ [37, Theorem 3.3.16-(d), pp. 178] and $\sigma_1(A + B) \leq \sigma_1(A) + \sigma_1(B)$ [37, Theorem 3.3.16-(a), pp. 178], we have that

$$\sigma_1(\Delta_S\Delta_S^H + \Delta_S Z^H + Z\Delta_S^H) \leq \sigma_1^2(\Delta_S) + 2\sigma_1(Z)\sigma_1(\Delta_S) \leq 3\sigma_1(Z)\sigma_1(\Delta_S), \tag{33}$$

if $\sigma_1(\Delta_S) \leq \sigma_1(Z)$ thus leading to the inequality

$$|u_i^H(wI_n - ZZ^H)^{-1}u_j - u_i^H(wI_n - (Z+\Delta_S)(Z+\Delta_S)^H)^{-1}u_j| \leq \frac{3 \cdot \sigma_1(Z)}{|\Im w|^2}\sigma_1(\Delta_S). \tag{34}$$

Since $\sigma_1(Z) \xrightarrow{\text{a.s.}} b = \sqrt{p}\,(1 + \sqrt{c}) < \infty$, if we can show that $\sigma_1(\Delta_S) \xrightarrow{\text{a.s.}} 0$ we will have shown that the bilinear forms involving u_i and u_j exhibits the same limiting behavior as though Z had i.i.d. Gaussian entries with zero mean and variance p/m. Repeating the argument would give us the analogous statement for the bilinear forms involving v_i and v_j. To prove that $\sigma_1(\Delta_S) \xrightarrow{\text{a.s.}} 0$, we first characterize $\mathbb{E}[\sigma_1(\Delta_S)]$. From a theorem by Latała [54], we have that

$$\mathbb{E}[\sigma_1(\Delta_S)] \leq C\left(\max_i \sqrt{\sum_j \mathbb{E}[\Delta S_{ij}^2]} + \max_j \sqrt{\sum_i \mathbb{E}[\Delta S_{ij}^2]} + \sqrt[4]{\sum_{ij} \mathbb{E}[\Delta S_{ij}^4]}\right),$$

where C is a universal constant (that does not depend on n or m). This gives us

$$\mathbb{E}[\sigma_1(\Delta_S)] \leq O\left(\frac{\log n \text{ factors}}{\sqrt{n}}\right). \tag{35}$$

[4]This identity can be verified by multiplying by $(wI - B)$ on the left and $(wI - A)$ on the right of the expressions on either side of the equality.

We note that

$$|S_{ij}| \leq O\left(\frac{\log n \text{ factors}}{n}\right),$$

while

$$\max_{i,j} |\Delta S_{ij}| \leq O\left(\frac{\log n \text{ factors}}{n}\right) =: K. \tag{36}$$

Plugging in $i = 1$ in (32) we have

$$|\sigma_1(pS + Z + \Delta_S) - \sigma_1(pS + Z)| \leq 1 \cdot \sigma_1(\Delta_S),$$

which implies that the largest singular value of a matrix is a 1-Lipschitz function of the nm entries of the matrix. Moreover, $\sigma_1(t\,A + (1-t)\,B) \leq t\sigma_1(A) + (1-t)\sigma_1(B)$, implying that the largest singular value is a convex, 1-Lipschitz function. Since, by (36), the entries of the Δ_S are bounded, independent random variables, we can apply Talagrand's concentration inequality (see [82, Theorem 2.1.13, pp. 73]) to obtain the tail bound

$$\text{Prob}\left(|\sigma_1(\Delta_S) - \mathbb{E}[\sigma_1(\Delta_S)]| > \epsilon\right) \leq 2\exp\left(-c\,\frac{\epsilon^2}{K^2}\right) = 2\exp\left(-c\,\frac{\epsilon^2 n^2}{\log n \text{ factors}}\right). \tag{37}$$

From (35), we have that $\mathbb{E}[\sigma_1(\Delta_S)] \to 0$ as $n \to \infty$. Moreover, the right-hand side of (37) is absolutely summable, *i.e.*,

$$\sum_n 2\exp\left(-c\,\frac{\epsilon^2 n^2}{\log n \text{ factors}}\right) < \infty,$$

which implies, via the Borel-Cantelli lemma, that

$$\sigma_1(\Delta_S) \xrightarrow{\text{a.s.}} 0 \tag{38}$$

Applying (38) to (32) yields the result that

$$w_i^{\text{eym}} = \sigma_i(pS + Z + \Delta_S) \xrightarrow{\text{a.s.}} \sigma_i(pS + Z) = \overline{w}_i^{\text{eym}}.$$

This proves Theorem 2.4-a). Moreover, from (34), we have that

$$u_i^H(wI_n - (Z + \Delta_S)(Z + \Delta_S)^H)^{-1}u_j \xrightarrow{\text{a.s.}} u_i^H(wI_n - ZZ^H)^{-1}u_j,$$

and by repeating the same argument we can show that

$$v_i^H(wI_m - (Z + \Delta_S)^H(Z + \Delta_S))^{-1}v_j \xrightarrow{\text{a.s.}} v_i^H(wI_m - Z^HZ)^{-1}v_j.$$

Using the same argument it can be shown that

$$u_i^H(wI_n - (Z + \Delta_S)(Z + \Delta_S)^H)^{-1}(Z + \Delta_S)v_j \xrightarrow{\text{a.s.}} 0,$$

and

$$v_i^H(wI_m - (Z + \Delta_S)^H(Z + \Delta_S))^{-1}(Z + \Delta_S)^H u_j \xrightarrow{\text{a.s.}} 0.$$

Following the proofs in [5], the convergence of these bilinear forms implies that the almost sure limits of $\sigma_i(\widetilde{X})$ and $(\widehat{u}_i^H u_j)$ and $(v_j^H \widehat{v}_i)$ for $i, j = 1, \ldots, r$ are identical to the almost sure limits of $\sigma_i(\overline{X})$ and $(\overline{u}_i^H u_j)$ and $(v_j^H \overline{v}_i)$ for $i, j = 1, \ldots, r$. Consequently, $w_i^{\text{opt}} \xrightarrow{\text{a.s.}} \overline{w}_i^{\text{opt}}$ and we have proved Theorem 2.4-b) and c).

7. Justification for assumptions in Conjectures 2.5 and 2.8

A key aspect (see [5, Lemma 4.1]) in rigorously proving Conjectures 2.5 and 2.8 is understanding the behavior of expressions of the form

$$u_i^H (z_j^2 I_n - XX^H)^{-2} u_i,$$

where z_j is a singular value of $\widetilde{X}$ but not of X. Let $X = U\Sigma V^H$ and $w = U^H u_i$. Then

$$u_i^H (z_j^2 I_n - XX^H)^{-2} u_i = \sum_i \frac{|w_i|^2}{(z_j^2 - \sigma_i^2(XX^H))^2} \geq \frac{|w_j|^2}{(\sigma_{i+r}^2(XX^H) - \sigma_i^2(XX^H))^2}.$$

When X has isotropically random singular vectors, $w_j = O(1/n)$ with high probability so if $z_j \in [a, b]$ and $\max_i \sigma_i(XX^H) - \sigma_{i+1}(XX^H)$ is bounded with probability by $O(\log n/n)$ in the bulk and the right hand side of the above expression will get unbounded (with n) resulting delocalization of the associated singular vectors. When μ_X exhibits a square root decay at the edge, then we expect the singular values at the edge to be spaced $O(n^{-2/3})$ apart with high probability so we might delocalization via the same argument. See [61] for an exposition of some of these issues and [4, 8] for recent results on the fine details of the spacing distribution of Wigner and Wishart random matrices.

References

[1] Z Bai and J. W. Silverstein. Spectral analysis of large dimensional random matrices, 2010.

[2] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[3] J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.

[4] G. Ben Arous and P. Bourgade. Extreme gaps between eigenvalues of random matrices. *The Annals of Probability*, vol. 41, no. 4, pp. 2648–2681, 2013.

[5] F. Benaych-Georges and R. R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.

[6] F. Benaych-Georges. Rectangular random matrices, related convolution. *Probab. Theory Related Fields*, 144(3-4):471–515, 2009.

[7] A. Birnbaum, I. M. Johnstone, B. Nadler, and D. Paul. Minimax bounds for sparse pca with noisy high-dimensional data. *arXiv:1203.0967*, 2012.

[8] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Isotropic local laws for sample covariance and generalized wigner matrices. arXiv: 1308.5729.

[9] S. Boucheron, G. Lugosi and P. Massart. Concentration Inequalities: A Nonasymptotic Theory of Independence, Oxford University Press, 2013.

[10] C. Boutsidis and E. Gallopoulos. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.

[11] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[12] C. Cacciapuoti, A. Maltsev, and B. Schlein. Local Marchenko-Pastur law at the hard edge of sample covariance matrices. *arXiv:1206.1730*, 2012.

[13] J. A. Cadzow and D. M. Wilkes. Enhanced rational signal modeling. *Signal processing*, 25(2):171–188, 1991.

[14] J. F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal of Optimization*, 20(4):1956–1982, 2010.

[15] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011.

[16] E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[17] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational ¡athematics*, 9(6):717–772, 2009.
[18] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. on Information Theory*, 56(5):2053–2080, 2010.
[19] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Sparse and low-rank matrix decompositions. In *Proc. 47th Annual Allerton Conference on Communication, Control, and Computing*, pages 962–967. IEEE, 2009.
[20] V. Chandrasekaran, S. Sanghavi, P. A.Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.
[21] S. Chatterjee. Matrix estimation by universal singular value thresholding. *arXiv:1212.1247*, 2012.
[22] P. Chen and D. Suter. Recovering the missing components in a large noisy low-rank matrix: Application to sfm. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(8):1051–1063, 2004.
[23] M. T. Chu, F. Diele, R. Plemmons, and S. Ragni. Optimality, computation, and interpretation of nonnegative matrix factorizations. In *SIAM Journal on Matrix Analysis*. Citeseer, 2004.
[24] M. T. Chu, R. E. Funderlic, and R. J. Plemmons. Structured low rank approximation. *Linear algebra and its applications*, 366:157–172, 2003.
[25] P. L. Combettes and J. W. Silverstein. Signal detection via spectral theory of large dimensional random matrices. *IEEE Trans. on Sig. Proc.*, vol. 8(40), pp. 2100–2105, 1992.
[26] S. Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.
[27] A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294, 2008.
[28] C.-A. Deledalle, S. Vaiter, G. Peyré, J. Fadili, and C. Dossal. Risk estimation for matrix recovery with spectral regularization. *arXiv:1205.1482*, 2012.
[29] P. Drineas, R. Kannan, and M. W. Mahoney. Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36(1):158–183, 2006.
[30] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.
[31] N. El Karoui. Tracy–widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability*, 35(2):663–714, 2007.
[32] L. Erdős and H.-T. Yau. Universality of local spectral statistics of random matrices. *Bull. Amer. Math. Soc*, 49:377–414, 2012.
[33] M. Fazel, E. j. Candes, B. Recht, and P. Parrilo. Compressed sensing and robust recovery of low rank matrices. In *42nd Asilomar Conference on Signals, Systems and Computers, 2008* , pp. 1043–1047. IEEE, 2008.
[34] D. Féral and S. Péché. The largest eigenvalues of sample covariance matrices for a spiked population: diagonal case. *Journal of Mathematical Physics*, 50:073302, 2009.
[35] G. H. Golub, A. Hoffman, and G. W. Stewart. A generalization of the Eckart-Young-Mirsky matrix approximation theorem. *Linear Algebra and Its Applications*, 88:317–327, 1987.
[36] W. Hachem, P. Loubaton, X. Mestre, J. Najim, and P. Vallet. A subspace estimator for fixed rank perturbations of large random matrices. *Journal of Multivariate Analysis*, 2012.
[37] R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1991.
[38] D. Hsu and S. M. Kakade. Learning gaussian mixture models: Moment methods and spectral decompositions. *arXiv:1206.5766*, 2012.
[39] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.
[40] I. M. Johnstone and A. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), 2009.
[41] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Ann. of Statistics*, 29(2):295–327, 2001.
[42] I. M. Johnstone. High dimensional statistical inference and random matrices. In *Proceedings of the International Congress of Mathematicians: Madrid*, pages 307–333, 2006.
[43] I. T. Jolliffe. *Principal component analysis*, volume 2. Wiley Online Library, 2002.

[44] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 98888:1865–1890, 2012.
[45] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. *Learning Theory*, pages 155–199, 2005.
[46] R. Kannan and S. Vempala. *Spectral algorithms*. Now Publishers Inc, 2009.
[47] D. R. Karger, S. Oh, and D. Shah. Budget-optimal crowdsourcing using low-rank matrix approximations. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 284–291. IEEE, 2011.
[48] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. , *IEEE Trans. on Information Theory*, 56(6):2980–2998, 2010.
[49] V. C. Klema and A. Laub. The singular value decomposition: Its computation and some applications. *IEEE Trans. on Automatic Control*, 25(2):164–176, 1980.
[50] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.
[51] S. Kritchman and B. Nadler. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):19–32, 2008.
[52] S. Kritchman and B. Nadler. Non-parametric detection of the number of signals: hypothesis testing and random matrix theory. *IEEE Trans. on Signal Processing*, 57(10):3930–3941, 2009.
[53] A. N. Langville, C. D. Meyer, R. Albright, J. Cox, and D. Duling. Initializations for the nonnegative matrix factorization. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 23–26. Citeseer, 2006.
[54] R. Latała. Some estimates of norms of random matrices. *Proc. of the American Math. Soc.*, vol. 133, no. 5, pp. 1273–1282, 2005.
[55] L. M. Le Cam Locally asymptotically normal families of distributions: certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses. In *University of California Press*, vol. 3, no. 2, 1960.
[56] Y. Li, K. J. R. Liu, and J. Razavilar. A parameter estimation scheme for damped sinusoidal signals based on low-rank Hankel approximation. *IEEE Trans. on Signal Processing*, 45(2):481–486, 1997.
[57] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.
[58] I. Markovsky. Structured low-rank approximation and its applications. *Automatica*, 44(4):891–909, 2008.
[59] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50, 1960.
[60] R. R. Nadakuditi. Exploiting random matrix theory to improve noisy low-rank matrix approximation. In *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, pages 769–773. IEEE, 2011.
[61] R. R. Nadakuditi. When are the most informative components for inference also the principal components? *arXiv:1302.1232*, 2013.
[62] R. R. Nadakuditi and M. E. J. Newman. Graph spectra and the detectability of community structure in networks. *Physical Review Letters*, 108(18):188701, 2012.
[63] R. R. Nadakuditi and M. E. J. Newman. Spectra of random graphs with arbitrary expected degrees. *Physical Review E*, 87(1):012803, 2013.
[64] R. R. Nadakuditi and A. Edelman. Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *IEEE Trans. on Signal Processing*, 56(7):2625–2638, 2008.
[65] B. Nadler. Nonparametric detection of signals by information theoretic criteria: performance analysis and an improved estimator. *IEEE Trans. on Signal Processing*, 58(5):2746–2756, 2010.
[66] N. Srebro and T. Jaakkola. Weighted low-rank approximations. In *In 20th International Conference on Machine Learning*, volume 20, page 720, 2003.
[67] A. Onatski. Testing hypotheses about the numbers of factors in large factor models. *Econometrica*, 77(5):1447–1479, 2009.
[68] A. Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016, 2010.
[69] D. Passemier and J.-F. Yao. On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices: Theory and Applications*, 1(1):1150002, 2012.

[70] S. Oymak and B. Hassibi. Finding dense clusters via "low rank+ sparse" decomposition. *arXiv:1104.5186*, 2011.
[71] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617, 2007.
[72] S. Péché. Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probability Theory and Related Fields*, 143(3-4):481–516, 2009.
[73] N. S. Pillai and J. Yin. Universality of covariance matrices. *arXiv preprint arXiv:1110.2501*, 2011.
[74] A. Rohde and A. B. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.
[75] A. Sanjeev and R. Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.
[76] J. Saunderson, V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1395–1416, 2012.
[77] L. L. Scharf. The svd and reduced rank signal processing. *Signal Processing*, 25(2):113–133, 1991.
[78] A. .A. Shabalin and A. B. Nobel. Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis*, 2013.
[79] J. W. Silverstein and S.-I. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. Journal of Multivariate Analysis, vol. 54, no. 2, pp. 295–309, 1995.
[80] A. Soshnikov. A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *Journal of Statistical Physics*, 108(5-6):1033–1056, 2002.
[81] M. Tao and X. Yuan. Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM Journal on Optimization*, 21(1):57–81, 2011.
[82] T. Tao. Topics in random matrix theory. vol. 132, AMS, 2012.
[83] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.
[84] D. W. Tufts and A. A. Shah. Estimation of a signal waveform from noisy data using low-rank approximation to a data matrix. *Signal Processing, IEEE Transactions on*, 41(4):1716–1721, 1993.
[85] M. O. Ulfarsson and V. Solo. Dimension estimation in noisy PCA with SURE and random matrix theory. *IEEE Trans. on Sig. Proc.*, vol. 56(12), pp. 5804–5816, 2008.
[86] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, pages 113–122. IEEE, 2002.
[87] D. M. Wikes and M. H. Hayes. Iterated toeplitz approximation of covariance matrices. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 1663–1666. IEEE, 1988.
[88] Y. Zhang, A. dAspremont, and L. El Ghaoui. Sparse PCA: Convex relaxations, algorithms and applications. In *Handbook on Semidefinite, Conic and Polynomial Optimization*, pages 915–940. Springer, 2012.
[89] Z. Zhang, H. Zha, and H. Simon. Low-rank approximations with sparse factors I: Basic algorithms and error analysis. *SIAM Journal on Matrix Analysis and Applications*, 23(3):706–727, 2002.
[90] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Raj Rao Nadakuditi, Department of Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109. USA.

E-mail address: rajnrao@eecs.umich.edu

URL: http://www.eecs.umich.edu/~rajnrao/