Optimal Shrinkage of Singular Values Under Random Data Contamination

Danny Barash
School of Computer Science and Engineering
Hebrew University
Jerusalem, Israel
danny.barash@mail.huji.ac.il

Matan Gavish
School of Computer Science and Engineering
Hebrew University
Jerusalem, Israel
gavish@cs.huji.ac.il

Abstract

A low rank matrix X has been contaminated by uniformly distributed noise, missing values, outliers and corrupt entries. Reconstruction of X from the singular values and singular vectors of the contaminated matrix Y is a key problem in machine learning, computer vision and data science. In this paper, we show that common contamination models (including arbitrary combinations of uniform noise, missing values, outliers and corrupt entries) can be described efficiently using a single framework. We develop an asymptotically optimal algorithm that estimates X by manipulation of the singular values of Y, which applies to any of the contamination models considered. Finally, we find an explicit signal-to-noise cutoff, below which estimation of X from the singular value decomposition of Y must fail, in a well-defined sense.

1 Introduction

Reconstruction of low-rank matrices from noisy and otherwise contaminated data is a key problem in machine learning, computer vision and data science. Well-studied problems such as dimension reduction [3], collaborative filtering [24, 28], topic models [13], video processing [21], face recognition [35], predicting preferences [26], analytical chemistry [29] and background-foreground separation [4] all reduce, under popular approaches, to low-rank matrix reconstruction. A significant part of the literature on these problems is based on the singular value decomposition (SVD) as the underlying algorithmic component, see e.g. [7, 19, 23].

Understanding and improving the behavior of SVD in the presence of random data contamination therefore arises as a crucially important problem in machine learning. While this is certainly a classical problem [14, 17, 20], it remains of significant interest, owing in part to the emergence of low-rank matrix models for matrix completion and collaborative filtering [9, 34].

Let X be an m-by-n unknown low-rank matrix of interest ($m \leq n$), and assume that we only observe the data matrix Y, which is a contaminated or noisy version of X. Let

$$Y = \sum_{i=1}^{m} y_i \mathbf{u}_i \mathbf{v}_i' \tag{1}$$

be the SVD of the data matrix Y. Any algorithm based on the SVD essentially aims to obtain an estimate for the target matrix X from (1). Most practitioners simply form the Truncated SVD (TSVD) estimate [18]

$$\hat{X}_r = \sum_{i=1}^{r} y_i \mathbf{u}_i \mathbf{v}_i' \tag{2}$$

where r is an estimate of $rank(X)$, whose choice in practice tends to be ad hoc [15].

Recently, [10, 16, 32] have shown that under white additive noise, it is useful to apply a carefully designed *shrinkage* function $\eta : \mathbb{R} \to \mathbb{R}$ to the data singular values, and proposed estimators of the form

$$\hat{X}_\eta = \sum_{i=1}^{n} \eta(y_i)\mathbf{u}_i\mathbf{v}_i' \,. \tag{3}$$

Such estimators are extremely simple to use, as they involve only simple manipulation of the data singular values. Interestingly, in the additive white noise case, it was shown that a unique optimal shrinkage function $\eta(y)$ exists, which asymptotically delivers the same performance as the best possible rotation-invariant estimator based on the data Y [16]. Singular value shrinkage thus emerged as a simple yet highly effective method for improving the SVD in the presence of white additive noise, with the unique optimal shrinker as a natural choice for the shrinkage function. A typical form of optimal singular value shrinker is shown in Figure 1 below, left panel.

Shrinkage of singular values, an idea that can be traced back to Stein's groundbreaking work on covariance estimation from the 1970's [33], is a natural generalization of the classical TSVD. Indeed, $\hat{X}_r$ is equivalent to shrinkage with the *hard thresholding* shrinker $\eta(y) = \mathbf{1}_{y \geq \lambda}$, as (2) is equivalent to

$$\hat{X}_\lambda = \sum_{i=1}^{n} \mathbf{1}_{y_i \geq \lambda}\mathbf{u}_i\mathbf{v}_i' \tag{4}$$

with a specific choice of the so-called *hard threshold* λ. While the choice of the rank r for truncation point TSVD is often ad hoc and based on gut feeling methods such as the Scree Plot method [11], its equivalent formulation, namely hard thresholding of singular values, allows formal and systematic analysis. In fact, restricting attention to hard thresholds alone [15] has shown that under white additive noise there exists a unique asymptotically optimal choice of hard threshold for singular values. The optimal hard threshold is a systematic, rational choice for the number of singular values that should be included in a truncated SVD of noisy data. [27] has proposed an algorithm that finds η^* in presence of additive noise and missing values, but has not derived an explicit shrinker.

1.1 Overview of main results

In this paper, we extend this analysis to common data contaminations that go well beyond additive white noise, including an arbitrary combination of additive noise, multiplicative noise, missing-at-random entries, uniformly distributed outliers and uniformly distributed corrupt entries.

The primary contribution of this paper is formal proof that there exists a unique asymptotically optimal shrinker for singular values under uniformly random data contaminations, as well a unique asymptotically optimal hard threshold. Our results are based on a novel, asymptotically precise description of the effect of these data contaminations on the singular values and the singular vectors of the data matrix, extending the technical contribution of [16, 27, 32] to the setting of general uniform data contamination.

General contamination model. We introduce the model

$$Y = A \odot X + B \tag{5}$$

where X is the target matrix to be recovered, and A, B are random matrices with i.i.d entries. Here, $(A \odot B)_{i,j} = A_{i,j}B_{i,j}$ is the Hadamard (entrywise) product of A and B.

Assume that $A_{i,j} \stackrel{\text{iid}}{\sim} (\mu_A, \sigma_A^2)$, meaning that the entries of A are i.i.d drawn from a distribution with mean μ_A and variance σ_A^2, and that $B_{i,j} \stackrel{\text{iid}}{\sim} (0, \sigma_B^2)$. In Section 2 we show that for various choices of the matrix A and B, this model represents a broad range of uniformly distributed random contaminations, including an arbitrary combination of additive noise, multiplicative noise, missing-at-random entries, uniformly distributed outliers and uniformly distributed corrupt entries. As a simple example, if $B \equiv 0$ and $P(A_{i,j} = 1) = \kappa$, then the Y simply has missing-at-random entries.

To quantify what makes a "good" singular value shrinker η for use in (3), we use the standard Mean Square Error (MSE) metric and

$$L(\eta|X) = \left|\left|\hat{X}_\eta(Y) - X\right|\right|_F^2 .$$

Using the methods of [16], our results can easily be extended to other error metrics, such as the nuclear norm or operator norm losses. Roughly speaking, an optimal shrinker η^* has the property that, asymptotically as the matrix size grows,

$$L(\eta^*|X) \leq L(\eta|X)$$

for any other shrinker η and any low-rank target matrix X.

The design of optimal shrinkers requires a subtle understanding of the random fluctuations of the data singular values $y_1, \ldots, y_n$, which are caused by the random contamination. Such results in random matrix theory are generally hard to prove, as there are nontrivial correlations between y_i and y_j, $i \neq j$. Fortunately, in most applications it is very reasonable to assume that the target matrix X is low rank. This allows us to overcome this difficulty by following [15, 27, 32] and considering an asymptotic model for low-rank X, inspired by Johnstone's Spiked Covariance Model [22], in which the correlation between y_i and y_j, for $i \neq j$ vanish asymptotically.

We state our main results informally at first. The first main result of this paper is the existence of a unique asymptotically optimal hard threshold λ^* in (4).

Importantly, as $\mathbb{E}(Y) = \mu_A X$, to apply hard thresholding to $Y = A \odot X + B$ we must from now on define

$$\hat{X}_\lambda = \frac{1}{\mu_A} \sum_{i=1}^n \mathbf{1}_{y_i > \lambda} \mathbf{u}_i \mathbf{v}_i' .$$

Theorem 1. *(Informal.) Let X be an m-by-n low-rank matrix and assume that we observe the contaminated data matrix Y given by the general contamination model* (5). *Then there exists a unique optimal (def. 3) hard threshold λ^* for the singular values of Y, given by*

$$\lambda^* = \sigma_B \sqrt{\left(c + \frac{1}{c}\right)\left(c + \frac{\beta}{c}\right)}$$

where $\beta = m/n$ and $c = \sqrt{1 + \beta + \sqrt{1 + 14\beta + \beta^2}}/\sqrt{2}$.

Our second main result is the existence of a unique asymptotically optimal shrinkage function η^* in (equation (3)). We calculate this shrinker explicitly:

Theorem 2. *(Informal.) Assume everything as in Theorem 1. Then there exists a unique optimal (def. 3) shrinker η^* for the singular values of Y given by*

$$\eta^*(y) = \begin{cases} \dfrac{\sigma_B^2}{y\mu_A} \sqrt{\left(\left(\dfrac{y}{\sigma_B}\right)^2 - \beta - 1\right)^2 - 4\beta} & y \geq \sigma_B(1 + \sqrt{\beta}) \\ 0 & y < \sigma_B(1 + \sqrt{\beta}) \end{cases}$$

We also discover that for each contamination model, there is a critical signal-to-noise cutoff, below which X cannot be reconstructed from the singular values and vectors of Y. Specifically, let η_0 be the zero singular value shrinker, $\eta_0(y) \equiv 0$, so that $\hat{X}_{\eta_0}(Y) \equiv 0$. Define the *critical signal level* for a shrinker η by

$$x^{critical}(\eta) = \inf_x \left\{x : L(\eta|X) < L(\eta_0|X)\right\}$$

where $X = x\tilde{\mathbf{u}}\tilde{\mathbf{v}}'$ is an arbitrary rank-1 matrix with singular value x. In other words, $x^{critical}(\eta)$ is the smallest singular value of the target matrix, for which η still outperforms the trivial zero shrinker η_0. As we show in Section 4, a target matrix X with a singular value below $x^{critical}(\eta)$ cannot be reliably reconstructed using η. The critical signal level for the optimal shrinker η^* is of special importance, since a target matrix X with a singular value below $x^{critical}(\eta^*)$ cannot be reliably reconstructed using *any* shrinker η. Restricting attention to hard thresholds only, we define $x^{critical}(\lambda)$, the critical level for a hard threshold, similarly. Again, singular values of X that fall below $x^{critical}(\lambda^*)$ cannot be reliably reconstructed using *any* hard threshold.

Our third main result is the explicit calculation of these critical signal levels:

Theorem 3. *(Informal.) Assume everything as in Theorem 1 and let c be as in Theorem 1. Let η^* be the optimal shrinker from Theorem 2 and let λ^* be the optimal hard threshold from Theorem 1. The critical signal levels for η^* and λ^* are given by:*

$$
\begin{aligned}
x^{critical}(\eta^*) &= (\sigma_B/\mu_A) \cdot \beta^{\frac{1}{4}} \\
x^{critical}(\lambda^*) &= (\sigma_B/\mu_A) \cdot c.
\end{aligned}
$$

Finally, one might ask what the improvement is in terms of the mean square error that is guaranteed by using the optimal shrinker and optimal threshold. As discussed below, existing methods are either infeasible in terms of running time on medium and large matrices, or lack a theory that can predict the reconstruction mean square error. For lack of a better candidate, we compare the optimal shrinker and optimal threshold to the default method, namely, TSVD.

Theorem 4. *(Informal.) Consider $\beta = 1$, and denote the worst-case mean square error of TSVD, η^* and λ^* by M_{TSVD}, M_{η^*} and M_{λ^*}, respectively, over a target matrix of low rank r. Then*

$$
\begin{aligned}
M_{TSVD} &= \left(\frac{\sigma_B}{\mu_A}\right)^2 5r \\
M_{\eta^*} &= \left(\frac{\sigma_B}{\mu_A}\right)^2 2r \\
M_{\lambda^*} &= \left(\frac{\sigma_B}{\mu_A}\right)^2 3r.
\end{aligned}
$$

Indeed, the optimal shrinker offers a significant performance improvement (specifically, an improvement of $3r(\sigma_B/\mu_A)^2$, over the TSVD baseline.
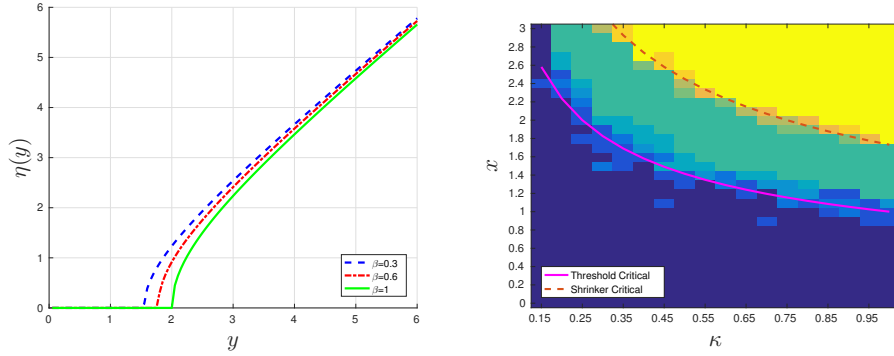

Figure 1: Left: Optimal shrinker for additive noise and missing-at-random contamination. Right: Phase plane for critical signal levels, see Section 6, Simulation 2.

Our main results allow easy calculation of the optimal threshold, optimal shrinkage and signal-to-noise cutoffs for various specific contamination models. For example:

1. **Additive noise and missing-at-random.** Let X be an m-by-n low-rank matrix. Assume that some entries are completely missing and the rest suffer white additive noise. Formally, we observe the contaminated matrix

$$
Y_{i,j} = \begin{cases} X_{i,j} + Z_{i,j} & \text{w.p. } \kappa \\ 0 & \text{w.p. } 1 - \kappa \end{cases},
$$

where $Z_{i,j} \overset{\text{iid}}{\sim} (0, \sigma^2)$, namely, follows an unknown distribution with mean 0 and variance σ^2. Let $\beta = m/n$. Theorem 1 implies that in this case, the optimal hard threshold for the singular values of Y is

$$
\lambda^* = \sqrt{\sigma^2 \kappa \left(c + 1/c\right) \left(c + \beta/c\right)}
$$

where $c = \sqrt{1 + \beta + \sqrt{1 + 14\beta + \beta^2}}/\sqrt{2}$. In other words, the optimal location (w.r.t mean square error) to truncate the singular values of Y, in order to recover X, is given by λ^*. The

4

optimal shrinker from Theorem 2 for this contamination mode may be calculated similarly, and is shown in Figure 1, left panel. By Theorem 4, the improvement in mean square error obtained by using the optimal shrinker, over the TSVD baseline, is $3r\sigma^2/\kappa$, quite a significant improvement.

2. **Additive noise and corrupt-at-random.** Let X be an m-by-n low-rank matrix. Assume that some entries are irrecoverably corrupt (replaced by random entries), and the rest suffer white additive noise. Formally,

$$Y_{i,j} = \begin{cases} X_{i,j} + Z_{i,j} & \text{w.p. } \kappa \\ W_{i,j} & \text{w.p. } 1 - \kappa \end{cases} .$$

Where $Z_{i,j} \stackrel{\text{iid}}{\sim} (0, \sigma^2)$, $W_{i,j} \stackrel{\text{iid}}{\sim} (0, \tau^2)$, and τ is typically large. Let $\tilde{\sigma} = \sqrt{\kappa\sigma^2 + (1 - \kappa)\tau^2}$. The optimal shrinker, which should be applied to the singular values of Y, is given by:

$$\eta^*(y) = \begin{cases} \tilde{\sigma}^2/(y\kappa)\sqrt{\left((y/\tilde{\sigma})^2 - \beta - 1\right)^2 - 4} & y \geq \tilde{\sigma}(1 + \sqrt{\beta}) \\ 0 & y < \tilde{\sigma}(1 + \sqrt{\beta}) \end{cases} .$$

By Theorem 4, the improvement in mean square error, obtained by using the optimal shrinker, over the TSVD baseline, is $3r(\kappa\sigma^2 + (1 - \kappa)\tau^2)/\kappa^2$.

1.2   Related Work

The general data contamination model we propose includes as special cases several modes extensively studied in the literature, including missing-at-random and outliers. While it is impossible to propose a complete list of algorithms to handle such data, we offer a few pointers, organized around the notions of robust principal component analysis (PCA) and matrix completion. To the best of our knowledge, the precise effect of general data contamination on the SVD (or the closely related PCA) has not been documented thus far. The approach we propose, based on careful manipulation of the data singular values, enjoys three distinct advantages. One, its running time is not prohibitive; indeed, it involves a small yet important modification on top of the SVD or TSVD, so that it is available whenever the SVD is available. Two, it is well understood and its performance (say, in mean square error) can be reliably predicted by the available theory. Three, to the best of our knowledge, none of the approaches below have become mainstream, and most practitioners still turn to the SVD, even in the presence of data contamination. Our approach can easily be used in practice, as it relies on the well-known and very widely used SVD, and can be implemented as a simple modification on top of the existing SVD implementations.

Robust Principle Component Analysis (RPCA). In RPCA, one assumes $Y = X + W$ where X is the low rank target matrix and W is a sparse outliers matrix. Classical approaches such as influence functions [20], multivariate trimming [17] and random sampling techniques [14] lack a formal theoretical framework and are not well understood. More modern approaches based on convex optimization [9, 34] proposed reconstructing X from Y via the nuclear norm minimization

$$\min_X \|X\|_* + \lambda \|Y - X\|_1 \, ,$$

whose runtime and memory requirements are both prohibitively large in medium and large matrices.

Matrix Completion. There are numerous heuristic approaches for data analysis in the presence of missing values [5, 30, 31]. To the best of our knowledge, there are no formal guarantees of their performance. When the target matrix is known to be low rank, the reconstruction problem is known as matrix completion. [7–9] and numerous other authors have shown that a semi-definite program may be used to stably recover the target matrix, even in the presence of additive noise. Here too, the runtime and memory requirements are both prohibitively large in medium and large matrices, making these algorithms infeasible in practice.

2   A Unified Model for Uniformly Distributed Contamination

Contamination modes encountered in practice are best described by a combination of primitive modes, shown in Table 1 below. These primitive contamination modes fit into a single template:

Definition 1. *Let A and B be two random variables, and assume that all moments of A and B are bounded. Define the* contamination link function

$$f_{A,B}(x) = Ax + B.$$

Given a matrix X, define the corresponding contaminated matrix Y with entries

$$Y_{i,j} \overset{indep.}{\sim} f_{A,B}(X_{i,j}).\tag{6}$$

Now observe that each of the primitive modes above corresponds to a different choice of random variables A and B, as shown in Table 1. Specifically, each of the primitive modes is described by a different assignment to A and B. We employ three different random variables in these assignments: $Z \overset{iid}{\sim} (0, \sigma^2/n)$, a random variable describing multiplicative or additive noise; $W \overset{iid}{\sim} (0, \tau^2/n)$, a random variable describing a large "outlier" measurement; and $M \overset{iid}{\sim} Bernoulli(\kappa)$ describing a random choice of "defective" entries, such as a missing value, an outlier and so on.

Table 1: *Primitive modes fit into the model* (6). *By convention, Y is m-by-n, $Z \overset{iid}{\sim} (0, \sigma^2/n)$ denotes a noise random variable, $W \overset{iid}{\sim} (0, \tau^2/n)$ denotes an outlier random variable and $M \overset{iid}{\sim} Bernoulli(\kappa)$ is a contaminated entry random variable.*

mode	model	A	B	levels
i.i.d additive noise	$Y_{i,j} = X_{i,j} + Z_{i,j}$	1	Z	σ
i.i.d multiplicative noise	$Y_{i,j} = X_{i,j} Z_{i,j}$	Z	0	σ
missing-at-random	$Y_{i,j} = M_{i,j} X_{i,j}$	M	0	κ
outliers-at-random	$Y_{i,j} = X_{i,j} + M_{i,j}W_{i,j}$	1	MW	κ,τ
corruption-at-random	$Y_{i,j} = M_{i,j}X_{i,j} + (1 - M_{i,j})W_{i,j}$	M	$(1 - M)W$	κ,τ

Actual datasets rarely demonstrate a single primitive contamination mode. To adequately describe contamination observed in practice, one usually needs to combine two or more of the primitive contamination modes into a composite mode. While there is no point in enumerating all possible combinations, Table 2 offers a few notable composite examples, using the framework (6). Many other examples are possible of course.

3 Signal Model

Following [32] and [15], as we move toward our formal results we are considering an asymptotic model inspired by Johnstone's Spiked Model [22]. Specifically, we are considering a sequence of increasingly larger data target matrices X_n, and corresponding data matrices $Y_n \overset{iid}{\sim} f_{A_n,B_n}(X_n)$. We make the following assumptions regarding the matrix sequence $\{X_n\}$:

A1 *Limiting aspect ratio:* The matrix dimension $m_n \times n$ sequence converges: $m_n/n \to \beta$ as $n \to \infty$. To simplify the results, we assume $0 < \beta \le 1$.

A2 *Fixed signal column span:* Let the rank $r > 0$ be fixed and choose a vector $\mathbf{x} \in \mathbb{R}^r$ with coordinates $\mathbf{x} = (x_1, \dots x_r)$ such that $x_1 > \dots > x_r > 0$. Assume that for all n

$$X_n = \tilde{U}_n \, diag(x_1, \dots, x_r)\tilde{V}_n$$

is an arbitrary singular value decomposition of X_n,

Table 2: *Some examples of composite contamination modes and how they fit into the model* (6). *Z,W,M are the same as in Table 1.*

mode	A	B	levels
Additive noise and missing-at-random	M	ZM	σ,κ
Additive noise and corrupt-at-random	M	$ZM + W(1 - M)$	σ,κ,τ
multiplicative noise and corrupt-at-random	ZM	$W(1 - M)$	σ,κ,τ
Additive noise and outliers	1	$Z + W(1 - M)$	σ,κ,τ

A3 *Incoherence of the singular vectors of X_n:* We make one of the following two assumptions regarding the singular vectors of X_n:

A3.1 X_n is random with an orthogonally invariant distribution. Specifically, $\tilde{U}_n$ and $\tilde{V}_n$, which follow the Haar distribution on orthogonal matrices of size m_n and n, respectively.

A3.2 The singular vectors of X_n are non-concentrated. Specifically, each left singular vector $\tilde{\mathbf{u}}_{n,i}$ of X_n (the i-th column of $\tilde{U}_n$) and each right singular vector $\tilde{\mathbf{v}}_{n,j}$ of X_n (the j-th column of $\tilde{V}_n$) satisfy[1]

$$||\tilde{\mathbf{u}}_{n,i}||_\infty \leq C\frac{\log^D(m_n)}{\sqrt{m_n}} \qquad \text{and} \qquad ||\tilde{\mathbf{v}}_{n,j}||_\infty \leq C\frac{\log^D(n)}{\sqrt{n}}$$

for any i, j and fixed constants C, D.

Definition 2. (Signal model.) Let $A_n \overset{\text{iid}}{\sim} (\mu_A, \sigma_A^2/n)$ and $B_n \overset{\text{iid}}{\sim} (0, \sigma_B^2/n)$ have bounded moments. Let X_n follow assumptions **[A1]–[A3]** above. We say that the matrix sequence $Y_n = f_{A_n,B_n}(X_n)$ follows our signal model, where $f_{A,B}(X)$ is as in Definition 1. We further denote $X_n = \sum_{i=1}^r x_i\tilde{\mathbf{u}}_{n,i}\tilde{\mathbf{v}}_{n,i}$ for the singular value decomposition of X_n and $Y_n = \sum_{i=1}^m y_{n,i}\mathbf{u}_{n,i}\mathbf{v}_{n,i}$ for the singular value decomposition of Y_n.

4 Main Results

Having described the contamination and the signal model, we can now formulate our main results. All proofs are deferred to the Supporting Information. Let X_n and Y_n follow our signal model, Definition 2, and write $\mathbf{x} = (x_1, \ldots, x_r)$ for the non-zero singular values of X_n. For a shrinker η, we write

$$L_\infty(\eta|\mathbf{x}) \overset{a.s.}{=} \lim_{n\to\infty} \left\|\hat{X}_n(Y_n) - X_n\right\|_F^2.$$

assuming the limit exists almost surely. The special case of hard thresholding at λ is denoted as $L_\infty(\eta|\mathbf{x})$.

Definition 3. Optimal shrinker and optimal threshold. A shrinker η^* is called *optimal* if

$$L_\infty(\eta|\mathbf{x}) \leq L_\infty(\eta|\mathbf{x})$$

for any shrinker η, any $r \geq 1$ and any $\mathbf{x} = (x_1, \ldots, x_r)$. Similarly, a threshold λ is called optimal if $L_\infty(\lambda^*|\mathbf{x}) \leq L_\infty(\lambda|\mathbf{x})$ for any threshold λ, any $r \geq 1$ and any $\mathbf{x} = (x_1, \ldots, x_r)$.

With these definitions, our main results Theorem 2 and Theorem 1 become formal. To make Theorem 3 formal, we need the following lemma and definition.

Lemma 1. Decomposition of the asymptotic mean square error. Let X_n and Y_n follow our signal model (Definition 2) and write $\mathbf{x} = (x_1, \ldots, x_r)$ for the non-zero singular values of X_n, and let η be the optimal shrinker. Then the limit $L_\infty(\eta|\mathbf{x})$ a.s. exists, and $L_\infty(\eta|\mathbf{x}) \overset{a.s.}{=} \sum_{i=1}^r L_1(\eta|x)$, where

$$L_1(\eta|x) = \begin{cases} x^2\left(1 - \dfrac{(t^4-\beta)^2}{(t^4+\beta t^2)(t^4+t^2)}\right) & t \geq \beta^{\frac{1}{4}} \\ x^2 & t < \beta^{\frac{1}{4}} \end{cases}$$

where $t = (\mu_A \cdot x)/\sigma_B$. Similarly, for a threshold λ we have $L_\infty(\lambda|\mathbf{x}) = \sum_{i=1}^r L_1(\lambda|x)$ with

$$L_1(\lambda|x) = \begin{cases} \left(\dfrac{\sigma_B}{\mu_A}\right)^2\left(\left(t+\dfrac{1}{t}\right)\left(t+\dfrac{\beta}{t}\right) - \left(t^2 - \dfrac{2\beta}{t^2}\right)\right) & \mu_A x \geq x(\lambda) \\ x^2 & \mu_A x < x(\lambda) \end{cases}$$

Where

$$x(y) = \begin{cases} (\sigma_B/\sqrt{2}\mu_A)\sqrt{(y/\sigma_B)^2 - \beta - 1 + \sqrt{\left(1+\beta-(y/\sigma_B)^2\right)^2 - 4\beta}} & t \geq \beta^{\frac{1}{4}} \\ 0 & t < \beta^{\frac{1}{4}} \end{cases} \tag{7}$$

---

[1]The incoherence assumption is widely used in related literature [6, 12, 27], and asserts that the singular vectors are spread out so X is not sparse and does not share singular subspaces with the noise.

Definition 4. Let η_0 be the zero singular value shrinker, $\eta_0(y) \equiv 0$, so that $\hat{X}_{\eta_0}(Y) \equiv 0$. Let η be a singular value shrinker. The critical signal level for η is

$$x^{critical}(\eta) = \inf_x \left\{ L_1(\eta|X) < L_1(\eta_0|X) \right\}$$

As we can see, the asymptotic mean square error decomposes over the singular values of the target matrix, $x_1, \ldots, x_r$. Each value x_i that falls below $x^{critical}(\eta)$ is better estimated with the zero shrinker η_0 than with η. It follows that any x_i that falls below $x^{critical}(\eta^*)$, where η^* is the optimal shrinker, cannot be reliably estimated by any shrinker η, and its corresponding data singular value y_i should simply be set to zero. This makes Theorem 2 formal.

5 Estimating the model parameters

In practice, using the optimal shrinker we propose requires an estimate of the model parameters. In general, σ_B is easy to estimate from the data via a median-matching method [15], namely

$$\hat{\sigma}_B = \frac{y_{med}}{\sqrt{n\mu_\beta}},$$

where y_{med} is the median singular value of Y, and μ_β is the median of the Marčenko-Pastur distribution. However, estimation of μ_A and σ_A must be considered on a case-by-case basis. For example, in the "Additive noise and missing at random" mode (table 2), $\sigma_A \equiv 1$ is known, and μ_A is estimated by dividing the amount of missing values by the matrix size.

6 Simulation

Simulations were performed to verify the correctness of our main results[2]. For more details, see Supporting Information.

1. **Critical signal level $x^{critical}(\lambda^*)$ under increasing noise.** Figure 2, left panel, shows the amount of data singular values y_i above $x^{critical}(\lambda^*)$, as a function of the fraction of missing values κ. Theorem 3 correctly predicts the exact values of κ at which the "next" data singular value falls below $x^{critical}(\lambda^*)$.

2. **Phase plane for critical signal levels $x^{critical}(\eta^*)$ and $x^{critical}(\lambda^*)$.** Figure 1, right panel, shows the x, κ plane, where x is the signal level and κ is the fraction of missing values. At each point in the plane, several independent data matrices were generated. Heatmap shows the fraction of the experiments at which the data singular value y_1 was above $x^{critical}(\eta^*)$ and $x^{critical}(\lambda^*)$. The overlaid graphs are theoretical predictions of the critical points.

3. **Brute-force verification of the optimal shrinker shape.** Figure 2, right panel, shows the shape of the optimal shrinker (Theorem 1). We performed a brute-force search for the value of $\eta(y)$ that produces the minimal mean square error. A brute force search, performed with a relatively small matrix size, matches the asymptotic shape of the optimal shrinker.

7 Simulation details

1. **Critical signal level $x^{critical}(\lambda^*)$ under increasing noise.** Consider the composite mode "i.i.d additive noise and missing-at-random"(sec. 2, mode 2) with noise level $\sigma = 1$, singular values $\mathbf{x} = (2, 3, 4, 5, 6)$ and matrix size $m = n = 1000$. To test whether our main results correctly predict the values of κ at which the signal singular values stop being estimable, we scanned the κ axis. At each value of κ we counted the number of estimable singular values and compared with the cutoff values of κ, at which this number should change. Figure 2 shows the number of measured estimable singular values against contamination level κ, with overlaid cut levels predicted by our main results (vertical lines).

---

[2]The full Matlab code that generated the figures in this paper and in the Supporting Information is permanently available at `https://purl.stanford.edu/kp113fq0838`.
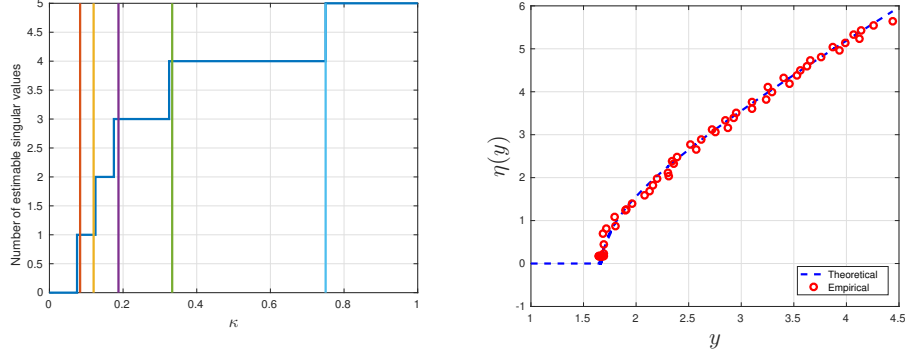
Figure 2: Left: empirical validation of the predicted critical signal level (Simulation 1). Right: Empirical validation of the optimal shrinker shape (Simulation 3).

2. **Phase plane for critical signal levels $x^{critical}(\eta^*)$ and $x^{critical}(\lambda^*)$.** Figure 1, right panel, shows the x, κ plane, where x is the signal level and κ is the fraction of missing values. At each point in the plane, several independent experiments (data matrices) were generated. The heatmap shows the the fraction of the experiments at which the data singular value y_1 was above $x^{critical}(\eta^*)$ and $x^{critical}(\lambda^*)$. Overlaid are the theoretical prediction of the critical points. Matrix size is 600×600, Monte carlo is 5 and $\beta = 1$.

3. **Brute-force verification of the optimal shrinker shape.** Figure (2,right) shows the shape of the optimal shrinker (Theorem 1). We performed a brute-force calculation scanning for the value of $\eta(y)$ that produces the lower mean square error. Brute force search, peformed in a 250×250 matrix, matches the asymptotic shape of the optimal shrinker. The scan set the noise level σ to 1, missing data level to $1 - \mu_A = 0.3$, $\beta = m/n = 1$, and scanned signal values in range [0,6].

8 Proofs

In this section we prove our main results: Theorem 1, Theorem 2, Theorem 3 and Theorem 4, as well as Lemma 1. The proofs rely on the following fundamental lemmas.

Lemma 2. Let $Y_n = f_{A_n, B_n}(X_n)$ be a matrix sequence following our signal model (Definition 2) with $A_n \overset{iid}{\sim} (\mu_A, \sigma_A^2/n)$ and $B_n \overset{iid}{\sim} (0, \sigma_B^2/n)$. Write $\bar{x}_i = \mu_A \cdot x_i$. Then the following holds:

1. For each $1 \leq i \leq r$ we have

$$\lim_{n \to \infty} y_{n,i} \overset{a.s.}{=} \begin{cases} \sigma_B \sqrt{\left(\dfrac{\bar{x}_i}{\sigma_B} + \dfrac{\sigma_B}{\bar{x}_i}\right)\left(\dfrac{\bar{x}_i}{\sigma_B} + \dfrac{\beta \sigma_B}{\bar{x}_i}\right)} & \bar{x}_i > \sigma_B \beta^{\frac{1}{4}} \\ \sigma_B(1 + \sqrt{\beta}) & \bar{x}_i \leq \sigma_B \beta^{\frac{1}{4}} \end{cases}$$

2. Let $1 \leq i \leq r$ and $1 \leq j \leq m_n$. If $\bar{x}_i > \sigma_B \beta^{\frac{1}{4}}$ and $j \leq r$, we have

$$d \cdot \lim_{n \to \infty} |\langle \tilde{\mathbf{u}}_i, \mathbf{u}_{n,j} \rangle|^2 \overset{a.s.}{=} \begin{cases} \dfrac{(\frac{\bar{x}_i}{\sigma})^4 - \beta}{(\frac{\bar{x}_i}{\sigma})^4 + \beta(\frac{\bar{x}_i}{\sigma})^2} & \bar{x}_i = \bar{x}_j \\ 0 & \bar{x}_i \neq \bar{x}_j \end{cases} \qquad (8)$$

and

$$d \cdot \lim_{n \to \infty} |\langle \tilde{\mathbf{v}}_i, \mathbf{v}_{n,j} \rangle|^2 \overset{a.s.}{=} \begin{cases} \dfrac{(\frac{\bar{x}_i}{\sigma})^4 - \beta}{(\frac{\bar{x}_i}{\sigma})^4 + (\frac{\bar{x}_i}{\sigma})^2} & \bar{x}_i = \bar{x}_j \\ 0 & \bar{x}_i \neq \bar{x}_j \end{cases}$$

Where $d_i = |\{x_j | x_j = x_i\}|$. Otherwise, both quantities converge a.s to 0 as $n \to \infty$.

9

Proof of Theorem 1. By lemma 3 we have that for any singular value hard threshold estimator λ, AMSE is given by (11). As for the bottom case in (11), clearly $\overline{x}^2$ is strictly increasing in $\overline{x}$, and therefor in x as well. As for the top case in (11), direct differentiation shows that it is strictly decreasing. It follows that the two functions of $\overline{x}$ intersect at a unique point. Denote their intersection point by $\overline{x^{est}} = \mu_A x^{est}$:

$$\overline{x^{est}} = \sigma_B \sqrt{\frac{1 + \beta + \sqrt{1 + 14\beta + \beta^2}}{2}}.$$

From the displacement formula 18, it follows that the following hard threshold for Y's singular values achieves the minimum AMSE between the above two expressions:

$$\lambda^* = \sigma_B \left(\sqrt{\left(\frac{\overline{x^{est}}}{\sigma_B} + \frac{\sigma_B}{\overline{x^{est}}} \right) \left(\frac{\overline{x^{est}}}{\sigma_B} + \frac{\beta \sigma_B}{\overline{x^{est}}} \right)} \right)$$

. $\hfill\square$

Proof of Theorem 2. Recall that for any matrix X, $\|X\|_F^2 = \langle X, X \rangle$ where $\langle \cdot, \cdot \rangle$ is the Hilbert–Schmidt matrix inner product. The notation $\langle X, X \rangle$ refers to Hilbert–Schmidt matrix inner product or to the standard Euclidean inner product. Denoting our shrinker by $\eta = \eta(y)$, and let $\overline{\eta} = \frac{\eta(y_i)}{\mu_A}$, we evaluate the mean square error:

$$\left\| \frac{1}{\mu_A} \hat{X}_\eta(Y) - X \right\|_F^2 = \langle \frac{1}{\mu_A} \hat{X}_\eta(Y) - X, \frac{1}{\mu_A} \hat{X}_\eta(Y) - X \rangle_{HS} =$$

$$\langle \frac{1}{\mu_A} \hat{X}_\eta(Y), \frac{1}{\mu_A} \hat{X}_\eta(Y) \rangle + \langle X_n, X_n \rangle - 2 \langle \frac{1}{\mu_A} \hat{X}_\eta(Y_n), X_n \rangle =$$

$$\sum_{i=1}^{m_n} \overline{\eta}(y_{n,i})^2 + \sum_{i=1}^{r} x_i^2 - 2 \sum_{i,j=1}^{r} x_i \overline{\eta}(y_{n,i}) \langle \mathbf{a_i b_i'}, \mathbf{u_{n,j} v_{n,j}'} \rangle =$$

$$\sum_{i=r+1}^{m_n} \overline{\eta}(y_{n,i})^2 + \sum_{i=1}^{r} \left[\overline{\eta}(y_{n,i})^2 + x_i^2 - 2x_i \sum_{j=1}^{r} \overline{\eta}(y_{n,j}) \langle \mathbf{a_i b_i'}, \mathbf{u_{n,j} v_{n,j}'} \rangle \right].$$

We now assume that $\forall i, y_i < \sigma(1 + \sqrt{\beta}), \overline{\eta}(y_i) = 0$; this is justified inside the proof of 3. As the first step of our analysis, we also assume that $rank(X) = 1$ to get when $n \to \infty$:

$$\left\| \frac{1}{\mu_A} \hat{X}_\eta(Y) - X \right\|_F^2 \overset{a.s.}{=} \overline{\eta}(y_{n,1})^2 + x_1^2 - 2x_1 \overline{\eta}(y_{n,1}) \langle \mathbf{a_1 b_1'}, \mathbf{u_{n,1} v_{n,1}'} \rangle =$$

$$\overline{\eta}(y_{n,1})^2 + x_1^2 - 2x_1 \overline{\eta}(y_{n,1}) \langle \mathbf{a_1}, \mathbf{u_{n,1}} \rangle \langle \mathbf{b_1'}, \mathbf{v_{n,1}'} \rangle.$$

Differentiating w.r.t η and comparing to zero we get: $\overline{\eta} = x_1 \langle \mathbf{a_1}, \mathbf{u_{n,1}} \rangle \langle \mathbf{b_1'}, \mathbf{v_{n,1}'} \rangle$. We would like to express $\overline{\eta}$ as a function of $y = y_1$ instead of x_1. When $n \to \infty$, the expression $\langle \mathbf{a_1}, \mathbf{u_{n,1}} \rangle \langle \mathbf{b_1'}, \mathbf{v_{n,1}'} \rangle$ is known. x_1 can be expressed as a function of y as in 7. Denote $t = \frac{\mu_A \cdot x}{\sigma_B}$, then we have:

$$\overline{\eta} = x_1 \langle \mathbf{a_1}, \mathbf{u_{n,1}} \rangle \langle \mathbf{b_1'}, \mathbf{v_{n,1}'} \rangle = \sigma_B \cdot \frac{t^4 - \beta}{\sqrt{(t^4 + \beta t^2)(t^4 + t^2)}} = \frac{\sigma_B^2}{y} \frac{t^4 - \beta}{t^2} = \frac{\sigma_B^2}{y} \left(\sqrt{t^2 - \frac{\beta}{t^2}} \right)^2$$

$$= \frac{\sigma_B^2}{y} \sqrt{(t^2 + \frac{\beta}{t^2})^2 - 4\beta} = \frac{\sigma_B^2}{y} \sqrt{((t + \frac{1}{t})(t + \frac{\beta}{t}) - \beta - 1))^2 - 4\beta} = \frac{\sigma_B^2}{y} \sqrt{(\frac{y}{\sigma_B}^2 - \beta - 1)^2 - 4\beta}.$$

Finally:

$$\eta^* = \begin{cases} \dfrac{\sigma_B^2}{\mu_A y} \sqrt{((\dfrac{y}{\sigma_B})^2 - \beta - 1)^2 - 4\beta} & y \geq \sigma_B(1 + \sqrt{\beta}) \\ 0 & y < \sigma_B(1 + \sqrt{\beta}) \end{cases}$$

Now without the assumption that $rank(X) = 1$, we have:

$$\left\|\hat{X}_\eta(Y) - X\right\|_F \overset{a.s.}{=} \sum_{i=i}^{r} \eta(y_{n,i})^2 + x_i^2 - 2x_i\eta(y_{n,i})\langle\mathbf{a_i}, \mathbf{u_{n,i}}\rangle\langle\mathbf{b_i'}, \mathbf{v_{n,i}'}\rangle.$$

Here there are r independent positive summands; the minimum of each is achieved by η^* above 8. $\qquad\square$

Proof of Theorem 3. We first calculate $x^{critical}(\lambda^*)$.

Let $\overline{x} = \mu_A x$, and λ a threshold operator s.t. $\lambda \geq \sigma_B(1 + \sqrt{\beta})$. Denote by $\overline{x^{est}} = \mu_A x^{est}$ be the unique solution to the intersection equality of (3):

$$\overline{x}^2 = \sigma_B^2((\frac{\overline{x}}{\sigma_B} + \frac{\sigma_B}{\overline{x}})(\frac{\overline{x}}{\sigma_B} + \frac{\beta\sigma_B}{\overline{x}}) - ((\frac{\overline{x}}{\sigma_B})^2 - \frac{2\beta\sigma_B^2}{\overline{x}^2}))$$

The left and right hand expressions denote the zero estimator loss $L_1(\lambda_0|X)$ and $L(\lambda|X)$ for a given λ accordingly. For $x < x^{est} = \frac{\sigma_B}{\mu_A}c$, $L_1(\lambda_0|X) < L(\lambda|X)$ for any λ, and the proof for $x^{critical}(\lambda^*)$ follows.

We now turn to $x^{critical}(\eta^*)$. According to the fundamental displacement lemma (lemma 2), x_i is asymptotically undetectable– displaced to a value independent of x – if its matching limit data singular value $y_i \leq \sigma_B(1 + \sqrt{\beta})$. The transition between detectable and undetectable is at:

$$\overline{x_i} = \sigma_B\beta^{\frac{1}{4}}$$

Assume a shrinker η different from the zero shrinker, s.t. for $y_i \leq \sigma_B(1 + \sqrt{\beta})$, $\eta(y_i) \neq 0$. The AMSE expression for any shrinker is:

$$\left\|\frac{1}{\mu_A}\hat{X}_\eta(Y) - X\right\|_F^2 = \langle\frac{1}{\mu_A}\hat{X}_\eta(Y) - X, \frac{1}{\mu_A}\hat{X}_\eta(Y) - X\rangle_{HS} =$$

$$\langle\frac{1}{\mu_A}\hat{X}_\eta(Y), \frac{1}{\mu_A}\hat{X}_\eta(Y)\rangle + \langle X_n, X_n\rangle - 2\langle\frac{1}{\mu_A}\hat{X}_\eta(Y_n), X_n\rangle =$$

$$\sum_{i=1}^{m_n} \overline{\eta}(y_{n,i})^2 + \sum_{i=1}^{r} x_i^2 - 2\sum_{i,j=1}^{r} x_i\overline{\eta}(y_{n,i})\langle\mathbf{a}_i\mathbf{b}_i', \mathbf{u}_{n,j}\mathbf{v}_{n,j}'\rangle$$

For η as assumed above, the expression $\sum_{i=1}^{m_n} \overline{\eta}(y_{n,i})^2$ tends to ∞ with n: there are infinitely many noise singular values that are shrunk to a positive value. The other summands are bounded, so $L(\eta|X) < L(\eta_0|X)$ and therefore,

$$x^{critical}(\eta^*) = \frac{\sigma_B\beta^{\frac{1}{4}}}{\mu_A} .$$

$\qquad\square$

Proof of Theorem 4. We start with M_{λ^*}. From lemma 3 the maximal AMSE is at the intersection point of the AMSE expressions. This point is given by:

$$\overline{x^{est}} = \mu_A x^{est} = \sigma_B\sqrt{\frac{1 + \beta + \sqrt{1 + 14\beta + \beta^2}}{2}}$$

Given $\beta = 1$, plug in x^{est} and the calculation of M_{λ^*} follows.

Turning to M_{TSVD}, TSVD is actually hard thresholding at the bulk edge. According to the displacement formula 1, $\lambda_{TSVD} = \sigma_B(1 + \sqrt{(\beta)})$, thus signal slightly larger than $x^{worst} = \sigma_B\beta^{\frac{1}{4}}/\mu_A$, will displace into a matching y value at slightly larger size than the bulk edge, and so the top loss term in

(11) applies. Direct differentiation shows that it is strictly decreasing. Noting that the first term is larger than x^2,the proof follows.

Turning finally to M_{η^*}, similarly to when we derived the optimal shrinker, we begin by focusing on a single singular value. Considering the optimal shrinker for a rank one matrix:

$$\eta = x_1 \langle \mathbf{a_1}, \mathbf{u_{n,1}} \rangle \langle \mathbf{b'_1}, \mathbf{v'_{n,1}} \rangle,$$

and let $t = (\mu_A \cdot x)/\sigma_B$, the AMSE is:

$$\left\| \hat{X}_\eta(Y) - \mu_A X \right\|_F^2 = \eta(y_{n,1})^2 + \overline{x_1}^2 - 2\overline{x_1}\eta(y_{n,1})\langle \mathbf{a_1}, \mathbf{u_{n,1}} \rangle \langle \mathbf{b'_1}, \mathbf{v'_{n,1}} \rangle \overset{a.s.}{=}$$
$$\overline{x}^2 - \eta^2 =$$
$$\mu^2 x^2 \left(1 - \frac{(t^4 - \beta)^2}{(t^4 + \beta t^2)(t^4 + t^2)} \right)$$

Choosing $t^2 = z$, we get

$$\sigma^2 z \left(1 - \frac{(z^2 - \beta)^2}{(z^2 + \beta z)(z^2 + z)} \right)$$

According to the assumption on x, $t > \beta^{\frac{1}{4}}$ and therefore $z > \sqrt{\beta}$ and the expression is monotonically increasing with β, in $\beta \in (0, 1]$. Setting $\beta = 1$, the simplified expression is

$$\sigma^2 \left(2 - \frac{1}{z} \right) \underset{z \to \infty}{\to} 2\sigma^2$$

This is the expression for the squared asymptotic loss for every singular value, yielding worst case AMSE of $r2\sigma^2$ in total. Since the loss is measured against X and not $\mu_A X$, we multiply the expressions by a factor of $\frac{1}{\mu_A^2}$ and the proof follows.

$\square$

Proof of Lemma 1. We begin with the shrinker loss.

$$L_\infty(\eta|\mathbf{x}) = \left\| \frac{1}{\mu_A}\hat{X}_\eta(Y) - X \right\|_F^2 = \langle \frac{1}{\mu_A}\hat{X}_\eta(Y) - X, \frac{1}{\mu_A}\hat{X}_\eta(Y) - X \rangle_{HS}$$
$$= \langle \frac{1}{\mu_A}\hat{X}_\eta(Y), \frac{1}{\mu_A}\hat{X}_\eta(Y) \rangle + \langle X_n, X_n \rangle - 2\langle \frac{1}{\mu_A}\hat{X}_\eta(Y_n), X_n \rangle$$
$$= \sum_{i=1}^{m_n} \overline{\eta}(y_{n,i})^2 + \sum_{i=1}^{r} x_i^2 - 2\sum_{i,j=1}^{r} x_i\overline{\eta}(y_{n,i})\langle \mathbf{a_i b'_i}, \mathbf{u_{n,j} v'_{n,j}} \rangle$$
$$= \sum_{i=r+1}^{m_n} \overline{\eta}(y_{n,i})^2 + \sum_{i=1}^{r} \left[\overline{\eta}(y_{n,i})^2 + x_i^2 - 2x_i \sum_{j=1}^{r} \overline{\eta}(y_{n,j})\langle \mathbf{a_i b'_i}, \mathbf{u_{n,j} v'_{n,j}} \rangle \right].$$

Assuming that $\forall y_i < \sigma_B(1 + \sqrt{B}), \overline{\eta}(y_i) = 0$,when $n \to \infty$:

$$\overset{a.s.}{=} \sum_{i=i}^{r} \eta(y_{n,i})^2 + x_i^2 - 2x_i\eta(y_{n,i})\langle \mathbf{a_i}, \mathbf{u_{n,i}} \rangle \langle \mathbf{b'_i}, \mathbf{v'_{n,i}} \rangle$$

The AMSE decomposition for a hard threshold λ is a result of lemma 3.

$\square$

Lemma 3. AMSE of hard threshold λ. Fix the signal matrix's rank r and let $\underline{x} \in \mathbb{R}^r$ the fixed rank singular values vector of the signal X. Let $\{X_n(\underline{x})\}_{n=1}^{\infty}$, $\{A_n\}_{n=1}^{\infty}$ and $\{B_n\}_{n=1}^{\infty}$, Let β be the sequence asymptotic ratio $X_n \in \mathbb{R}^{m_n \times n}(\mathbb{R})$ where $\lim_{n \to \infty} m_n/n = \beta$. Let Y_n matrix

sequence s.t. $Y = A \odot X + B$ as in the basic framework (2) with matching β, σ_B, μ_A. Let Y's SVD decomposition be:

$$Y = \sum_{i=1}^{m} y_i \mathbf{u}_i \mathbf{v}_i' \tag{9}$$

Assume we estimate X by the hard threshold estimator with parameter k, i.e. by setting to zero Y's singular values that are smaller than k:

$$\hat{X}_\lambda(Y) = \sum_{i=1}^{m} \eta_H(y_i; \lambda) \mathbf{u}_i \mathbf{v}_i', \text{ , where } \eta_H(y; \lambda) = \begin{cases} 0 & y < \lambda, \\ y & \text{otherwise} \end{cases}$$

and let λ be a selected hard threshold s.t. $\lambda \geq \sigma_B(1 + \sqrt{\beta})$.

Define the Asymptotic MSE:

$$AMSE(\hat{X}_\lambda(Y)) = \lim_{n \to \infty} \left\| \frac{1}{\mu_A} \hat{X}_\lambda(Y_n) - X_n \right\|_F^2$$

and . Then

$$AMSE(\hat{X}_\lambda, \mathbf{x}) = \sum_{i=1}^{r} M(\hat{X}_\lambda, x_i) \tag{10}$$

and

$$M(\hat{X}_\lambda, x) = \begin{cases} \left(\frac{\sigma_B}{\mu_A} \right)^2 \left(\left(t + \frac{1}{t} \right) \left(t + \frac{\beta}{t} \right) - \left(t^2 - \frac{2\beta}{t^2} \right) \right) & \overline{x} \geq x(\lambda) \\ x^2 & \overline{x} < x(\lambda) \end{cases} \tag{11}$$

where $t = \mu_A x / \sigma_B$.

$$x(y) = \begin{cases} (\sigma_B/\sqrt{2}\mu_A)\sqrt{(y/\sigma_B)^2 - \beta - 1 + \sqrt{(1 + \beta - (y/\sigma_B)^2)^2 - 4\beta}} & y \geq \sigma_B\sqrt{1 + \beta + \sqrt{\beta}(1 + \mu_A^2)} \\ 0 & else \end{cases} \tag{12}$$

Proof of lemma 3   We'll denote the SVD of a signal matrix X, which is an element of $\{X_n\}$, by:

$$X = \sum_{i=1}^{r} x_i \mathbf{a}_i \mathbf{b}_i'$$

$$\left\| \hat{X}_\lambda(Y) - \mu_A X \right\|_F^2 = \langle \hat{X}_\lambda(Y) - \mu_A X, \hat{X}_\lambda(Y) - \mu_A X \rangle_{HS} =$$

$$\langle \hat{X}_\lambda(Y), \hat{X}_\lambda(Y) \rangle + \langle \mu_A X_n, \mu_A X_n \rangle - 2 \langle \hat{X}_\lambda(Y_n), \mu_A X_n \rangle =$$

$$\sum_{i=1}^{m_n} \eta_H(y_{n,i}; \lambda)^2 + \sum_{i=1}^{r} \overline{x_i}^2 - 2 \sum_{i,j=1}^{r} \overline{x_i} \eta_H(y_{n,j}; \lambda) \langle \mathbf{a}_i \mathbf{b}_i', \mathbf{u}_{n,j} \mathbf{v}_{n,j}' \rangle =$$

$$\sum_{i=r+1}^{m_n} \mu_H(y_{n,i}; \lambda)^2 + \sum_{i=1}^{r} \left[(\mu_H(y_{n,i}; \lambda)^2 + \overline{x_i}^2) - 2\overline{x_i} \sum_{j=1}^{r} \mu_H(y_{n,j}; \lambda) \langle \mathbf{a_i}\mathbf{b_i'}, \mathbf{u_{n,j}}\mathbf{v_{n,j}'} \rangle \right].$$

According to the displacement formula 1

1. $y_{n,r+1} \overset{a.s.}{=} \sigma_B(1 + \sqrt{\beta}) < \lambda$, the leftmost term above converges almost surely to zero.

13

2. When $0 \leq \overline{x_i} \leq \sigma_B \beta^{\frac{1}{4}}$, all that remains is $\sum_{i=1}^{r} \overline{x_i}^2$ which proves the second case of lemma 3.

Assuming now $\overline{x_i} \geq \sigma_B \beta^{\frac{1}{4}}$, we calculate the limit of every expression in:

$$\sum_{i=1}^{r} \left[(\mu_H(y_{n,i}; \lambda)^2 + \overline{x_i}^2) - 2\overline{x_i} \sum_{j=1}^{r} \mu_H(y_{n,j}; \lambda) \langle \mathbf{a_i b_i'}, \mathbf{u_{n,j} v_{n,j}'} \rangle \right].$$

1. According to the displacement formula:

$$\lim_{n \to \infty} \eta_H(y_{n,i}; \lambda)^2 \stackrel{a.s.}{=} \begin{cases} \sigma_B^2 (\frac{\overline{x_i}}{\sigma_B} + \frac{\sigma_B}{\overline{x_i}})(\frac{\overline{x_i}}{\sigma_B} + \frac{\beta \sigma_B}{\overline{x_i}}) & \text{for } y_{n_i} > \lambda ; \\ 0 & \text{otherwise} \end{cases}$$

2. According to the rotation formula:

$$\lim_{n \to \infty} \langle \mathbf{a_i b_i'}, \mathbf{u_{n,j} v_{n,j}'} \rangle = \lim_{n \to \infty} \langle \mathbf{a_j}, \mathbf{u_{n,j}} \rangle \langle \mathbf{b_i}, \mathbf{v_{n,j}} \rangle \stackrel{a.s.}{=} \begin{cases} \dfrac{(\frac{\overline{x_i}}{\sigma_B})^4 - \beta}{d_i \sqrt{((\frac{\overline{x_i}}{\sigma_B})^4 + \beta (\frac{\overline{x_i}}{\sigma_B})^2)((\frac{\overline{x_i}}{\sigma_B})^4 + (\frac{\overline{x_i}}{\sigma_B})^2)}} & \text{for } \overline{x_i} = \overline{x_j}; \\ 0 & \text{otherwise} \end{cases}$$

   where $d_i = |\{x_j | x_j = \overline{x_i}\}|$.

3.

$$\lim_{n \to \infty} \mu_H(y_{n,j}; \lambda) \langle \mathbf{a_i b_i'}, \mathbf{u_{n,j} v_{n,j}'} \rangle \stackrel{a.s.}{=} \begin{cases} \dfrac{\sigma_B((\frac{\overline{x_i}}{\sigma_B})^4 - \beta)}{(\frac{\overline{x_i}}{\sigma_B})^3} & \text{for } y_{n_i} > \lambda; \\ 0 & \text{otherwise} \end{cases}$$

4. For the last and rightmost element of the AMSE, assuming $\forall i, d_i = 1$:

$$-2\overline{x_i} \mu_H(y_{n,j}; \lambda) \langle \mathbf{a_i}, \mathbf{v_{n,j}} \rangle \langle \mathbf{b_i}, \mathbf{u_{n,j}} \rangle \stackrel{a.s.}{=} -\sigma_B^2 \left(\frac{2x_i^2}{\sigma_B^2} - 2\beta \frac{\sigma^2}{x_i^2} \right)$$

Adding all expressions yields for $\overline{x_i} \geq \sigma_B \beta^{\frac{1}{4}}$ and $\overline{x_i} > x(\lambda)$:

$$M(\hat{X}_\lambda, x) = \sigma_B^2 ((\frac{\overline{x_i}}{\sigma_B} + \frac{\sigma_B}{\overline{x_i}})(\frac{\overline{x_i}}{\sigma_B} + \frac{\beta \sigma_B}{\overline{x_i}}) - ((\frac{\overline{x_i}}{\sigma_B})^2 - \frac{2\beta \sigma_B^2}{\overline{x_i}^2}))$$

. Which concludes the calculation by showing the first expression in the lemma. Since the loss is measured against X and not $\mu_A X$, we multiply the expressions by a factor of $\frac{1}{\mu_A^2}$ and the proof follows.

$\square$

It remains to prove Lemma 2. The proof follows a strategy proposed by [27]. It is convenient to break the proof into a sequence of lemmas, which are of independent interest.

Definition 5. Let Z be an m-by-n matrix. For $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^m$ and $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n$, define

$$H(w|\mathbf{u}_1, \mathbf{u}_2, Z) \quad := \quad \mathbf{u}_1'(w^2 I_n - ZZ')^{-1} \mathbf{u}_2 \tag{13}$$

$$Q(w|\mathbf{v}_1, \mathbf{v}_2, Z) \quad := \quad \mathbf{v}_1'(w^2 I_n - Z'Z)^{-1} \mathbf{v}_2 . \tag{14}$$

Definition 6. Let $\mathcal{Z} = \{Z_n\}_{n=0}^{\infty}$ be a sequence of matrices s.t. Z_n is m_n-by-n and $\lim_{n \to \infty} m_n/n = \beta$. Let $\mathcal{U} = \{U_n\}_{n=0}^{\infty}$ be any sequence of m_n-by-m_n orthonormal matrices with columns $U_n = (\mathbf{u}_{n,1}, \dots, \mathbf{u}_{n,m_n})$. Similarly let $\mathcal{V} = \{V_n\}_{n=0}^{\infty}$ be any sequence of n-by-n orthonormal matrices with columns $V_n = (\mathbf{v}_{n,1}, \dots, \mathbf{v}_{n,n})$. Define

$$H_{i,j}(w|\mathcal{U}, \mathcal{Z}) \quad := \quad \lim_{n \to \infty} H(w|\mathbf{u}_{n,i}, \mathbf{u}_{n,j}, Z_n) \tag{15}$$

$$Q_{i,j}(w|\mathcal{V}, \mathcal{Z}) \quad := \quad \lim_{n \to \infty} H(w|\mathbf{v}_{n,i}, \mathbf{v}_{n,j}, Z_n) \tag{16}$$

assuming these limits exist almost surely.

The following result is due to Bloemendal et al [2]:

Lemma 4. Let $\mathcal{Z} = \{Z_n\}_{n=0}^{\infty}$ a sequence of matrices s.t. Z_n is m_n-by-n and $\lim_{n \to \infty} m_n/n = \beta \leq 1$. Further assume that $(Z_n)_{i,j} \overset{iid}{\sim} \mathcal{F}$, where $\mathcal{F}$ is some distribution with bounded moments, mean 0 and variance σ^2/n. Let $\mathcal{U}$ and $\mathcal{V}$ be arbitrary sequences of orthonormal matrices as in Definition 6, which are either nonrandom or independent of $\{Z_n\}$. Then for all $1 \leq i \leq m_n$ and $1 \leq j \leq n$,

$$H_{i,j}(w \,|\, \mathcal{U}, \mathcal{Z}) \overset{a.s.}{=} \int \frac{d\mu_{\mathcal{Z}}(t)}{w^2 - t^2} \delta_{i,j}$$

$$Q_{i,j}(w \,|\, \mathcal{V}, \mathcal{Z}) \overset{a.s.}{=} \int \frac{d\mu_{\mathcal{Z}}(t)}{w^2 - t^2} \delta_{i,j}$$

where $\mu_{\mathcal{Z}}(t)$ is the density of the Marčenko-Pastur distribution [25] given by

$$\mu_{\mathcal{Z}}(t) = \frac{4\sigma^4 \beta - (t^2 - \sigma^2 - \sigma^2 \beta)^2}{2\pi\sigma^2 \beta t} \mathbf{1}_{\beta_-, \beta_+}(t), \tag{17}$$

where $\beta_{\pm} = \sigma^2(1 \pm \sqrt{\beta})^2$.

The next lemma shows that the matrix with entries $A_{i,j} X_{i,j}$, with $A_{i,j} \overset{iid}{\sim} (\mu_A, \sigma_A^2)$, and X_n from our signal model, is well approximated by the matrix with entries $\mathbb{E}[A] X_{i,j}$.

Lemma 5. Let A be a random variable with mean μ_A and variance σ_A^2. Let $\{X_n\}$ be a matrix sequence satisfying assumptions **A1**–**A3** (Section 2 in the main text). Let $\delta_{n,1}$ be the largest singular value of the matrix Δ_n with entries

$$(\Delta_n)_{i,j} = A_{i,j} X_{i,j} - \mu_A X_{i,j} .$$

Then $\delta_{n,1} \overset{a.s.}{\to} 0$ as $n \to \infty$.

Proof. See [27] equations no. 35–38. □

The next lemma shows that adding a "small perturbation" to Z_n does not change the value of H and Q from Definition 6.

Lemma 6. Let $\mathcal{Z} = \{Z_n\}_{n=0}^{\infty}$ a sequence of matrices s.t. Z_n is m_n-by-n and $\lim_{n \to \infty} m_n/n = \beta \leq 1$, and let $\mathcal{U}$ and $\mathcal{V}$ be arbitrary sequences of orthonormal matrices as in Definition 6, which are either nonrandom or independent of $\{Z_n\}$. Assume that for some $1 \leq i \leq m_n$ and $1 \leq j \leq n$,

$$H_{i,j}(w \,|\, \mathcal{U}, \mathcal{Z}) \overset{a.s.}{=} f(w)\delta_{i,j} \qquad \text{and}$$

$$Q_{i,j}(w \,|\, \mathcal{V}, \mathcal{Z}) \overset{a.s.}{=} f(w)\delta_{i,j} .$$

Let $\{\Delta_n\}_{n=0}^{\infty}$ be a sequece of matrices of the same sizes and assume that $\delta_{n,1} \overset{a.s.}{\to} 0$ as $n \to \infty$, where $\delta_{n,1}$ is the largest singular value of Δ_n ($n = 1, 2, \ldots$). Denote by $\bar{\mathcal{Z}}$ the sequence of matrices $\{Z_n + \Delta_n\}$. Then also

$$H_{i,j}(w \,|\, \mathcal{U}, \bar{\mathcal{Z}}) \overset{a.s.}{=} f(w)\delta_{i,j} \qquad \text{and}$$

$$Q_{i,j}(w \,|\, \mathcal{V}, \bar{\mathcal{Z}}) \overset{a.s.}{=} f(w)\delta_{i,j} .$$

Proof. See [27] equations no. 33 and 34. □

The next lemma in this chain of arguments is due to [1]:

Lemma 7. Let $\mathcal{Z} = \{Z_n\}_{n=0}^{\infty}$ a sequence of matrices s.t. Z_n is m_n-by-n and $\lim_{n \to \infty} m_n/n = \beta \leq 1$. Assume that X_n is a sequence of matrices satisfying assumptions **[A1]**–**[A3]** from Section 2. Define $Y_n = X_n + Z_n$ and let $X_n = \sum_{i=1}^{r} x_i \tilde{\mathbf{u}}_i \tilde{\mathbf{v}}_i'$ and $Y_n = \sum_{i=1}^{m_n} y_{n,i} \mathbf{u}_{n,i} \mathbf{v}_{n,i}'$ denote their singular value decompositions, respectively. Assume that for any $\mathcal{U}$ and $\mathcal{V}$, which are arbitrary sequences of orthonormal matrices as in Definition 6, either nonrandom or independent of $\{Z_n\}$, we have for all $1 \leq i \leq m_n$ and $1 \leq j \leq n$,

$$H_{i,j}(w \,|\, \mathcal{U}, \mathcal{Z}) \overset{a.s.}{=} \int \frac{d\mu_{\mathcal{Z}}(t)}{w^2 - t^2} \delta_{i,j}$$

$$Q_{i,j}(w \,|\, \mathcal{V}, \mathcal{Z}) \overset{a.s.}{=} \int \frac{d\mu_{\mathcal{Z}}(t)}{w^2 - t^2} \delta_{i,j} ,$$

where $\mu_{\mathcal{Z}}(t)$ is given by (17). Then

15

1. For each $1 \leq i \leq r$ we have

$$\lim_{n \to \infty} y_{n_i} \overset{a.s.}{=} \begin{cases} \sigma(\sqrt{(\frac{x_i}{\sigma} + \frac{\sigma}{x_i})(\frac{x_i}{\sigma} + \frac{\beta\sigma}{x_i})}) & \text{for } x_i > \sigma\beta^{\frac{1}{4}} \\ \sigma(1 + \sqrt{\beta}) & \text{for } x_i \leq \sigma\beta^{\frac{1}{4}}. \end{cases} \tag{18}$$

2. Let $1 \leq i \leq r$ and $1 \leq j \leq m_n$. If $x_i > \sigma_B\beta^{\frac{1}{4}}$ and $j \leq r$, we have

$$d \cdot \lim_{n \to \infty} |\langle u_{n_i}^0, u_{n_j} \rangle|^2 \overset{a.s.}{=} \begin{cases} \dfrac{(\frac{x_i}{\sigma})^4 - \beta}{(\frac{x_i}{\sigma})^4 + \beta(\frac{x_i}{\sigma})^2} & x_i = x_j \\ 0 & x_i \neq x_j \end{cases} \tag{19}$$

and

$$d \cdot \lim_{n \to \infty} |\langle v_{n_i}^0, v_{n_j} \rangle|^2 \overset{a.s.}{=} \begin{cases} \dfrac{(\frac{x_i}{\sigma})^4 - \beta}{(\frac{x_i}{\sigma})^4 + (\frac{x_i}{\sigma})^2} & x_i = x_j \\ 0 & x_i \neq x_j \end{cases} \tag{20}$$

Otherwise, both quantities converge a.s to 0 as $n \to \infty$.

We can finally connect the dots and prove Lemma 2.

Proof of Lemma 2. . Per the lemma statement, let $Y_n = f_{A_n,B_n}(X_n)$ be a matrix sequence following our signal model (Definition 2) with $A_n \overset{iid}{\sim} (\mu_A, \sigma_A^2/n)$ and $B_n \overset{iid}{\sim} (0, \sigma_B^2/n)$. Let $\mathbf{A}_n$ be an m_n-by-n matrix with entries $(\mathbf{A}_n)_{i,j} \overset{iid}{\sim} A_n$ and let $\mathbf{B}n$ be an m_n-by-n matrix with entries $(\mathbf{B}_n)_{i,j} \overset{iid}{\sim} B_n$. We can write $(Y_n)_{i,j} = (\mathbf{A}_n)_{i,j}(X_n)_{i,j} + (\mathbf{B}_n)_{i,j}$. Letting Δ_n be the m_n-by-n matrix with entries $(\Delta_n)_{i,j} = (\mathbf{A}_n)_{i,j}(X_n)_{i,j} - \mu_A(X_n)_{i,j}$ we have

$$Y_n = \mu_A X_n + \mathbf{B}_n + \Delta_n.$$

By Lemma 5, the top singular value $\delta_{n,1}$ of Δ_n satisfies $\delta_{n,1} \overset{a.s.}{\to} 0$ as $n \to \infty$.

Now, choose arbitrary sequences of orthonormal matrices $\mathcal{U}$ and $\mathcal{V}$ as in Definition 6, either nonrandom or independent of $\{\mathbf{B}_n\}$. Let $\mathcal{B}$ denote the matrix sequence $\{\mathbf{B}_n\}$ and let $\bar{\mathcal{B}}$ denote the matrix sequence $\{\mathbf{B}_n + \Delta_n\}$. Invoking Lemma 6 we obtain

$$H_{i,j}(w \,|\, \mathcal{U}, \mathcal{B}) \overset{a.s.}{=} H_{i,j}(w \,|\, \mathcal{U}, \bar{\mathcal{B}})$$
$$Q_{i,j}(w \,|\, \mathcal{V}, \mathcal{B}) \overset{a.s.}{=} Q_{i,j}(w \,|\, \mathcal{V}, \bar{\mathcal{B}}).$$

for all $1 \leq i \leq m_n$ and $1 \leq j \leq n$. However, by Lemma 4, for all $1 \leq i \leq m_n$ and $1 \leq j \leq n$ we have

$$H_{i,j}(w \,|\, \mathcal{U}, \mathcal{B}) \overset{a.s.}{=} \int \frac{d\mu_{\mathcal{Z}}(t)}{w^2 - t^2} \delta_{i,j}$$
$$Q_{i,j}(w \,|\, \mathcal{V}, \mathcal{B}) \overset{a.s.}{=} \int \frac{d\mu_{\mathcal{Z}}(t)}{w^2 - t^2} \delta_{i,j},$$

where $\mu_{\mathcal{Z}}(t)$ is given by (17) with $\sigma \equiv \sigma_B$. It follows that the sequence $Y_n = \mu_A X_n + (\mathbf{B}_n + \Delta_n)$ satisfies all the assumptions of Lemma 7, for arbitrary $\mathcal{U}$ and $\mathcal{V}$. We therefore conlude that, for Y_n, equations (18), (19) and (20) hold, replacing σ with σ_B and x_i with $\bar{x}_i$ ($1 \leq i \leq r$). The lemma follows. $\square$

9 Contamination modes

We describe here contamination modes that did not enter the main text.

Basic contamination modes

1. **Additive noise.** The simplest form of contamination is noise added to each entry. Assume $Y_{i,j} = X_{i,j} + Z_{i,j}$. The matching parameters for the model $A \odot X + B$ are $\mu_A = 1, \sigma_A = 0, \sigma_B = \sigma_Z$.

2. **Missing-at-random.** Assume data with entries that are missing-at-random, where missing entries are replaced by zeros with probability (w.p.) $1 - \kappa$. Assume

$$Y_{i,j} = \begin{cases} X_{i,j} & \text{w.p. } \kappa \\ 0 & \text{w.p. } 1 - \kappa \end{cases},$$

.

   The matching parameters for the model $A \odot X + B$ are $\mu_A = \kappa, \sigma_A = 0, \sigma_B = 0$.

3. **Outliers-at-random.** When some entries of Y contain inordinate level of noise, the corresponding entries are said to be outliers:
   Assuming that this noise is additive, we can write

$$Y_{i,j} = \begin{cases} X_{i,j} & \text{w.p. } \kappa \\ X_{i,j} + W_{i,j} & \text{w.p. } 1 - \kappa \end{cases}.$$

   With $W_{i,j} \overset{\text{iid}}{\sim} (0, \tau^2)$. The matching parameters for the model $A \odot X + B$ are $\mu_A = 1, \sigma_A = 0, \sigma_B = \sqrt{(1 - \kappa)\tau^2}$.

4. **Multiplicative Noise.** Each signal entry is multiplied by a random noise distribution sample. $Y_{i,j} = Z_{i,j} \cdot X_{i,j}$. The matching parameters for the model $A \odot X + B$ are $\mu_A = \mu_Z, \sigma_A = \sigma_Z, \sigma_B = 0$.

5. **Corrupt-at-random.** In some measurement processes some entries are completely destroyed and *replaced* by randomly generated noise. To distinguish this form of contamination from simple outliers, where the original entry is not replaced, we refer to this as corruption. Assume

$$Y_{i,j} = \begin{cases} X_{i,j} & \text{w.p. } \kappa \\ W_{i,j} & \text{w.p. } 1 - \kappa \end{cases}.$$

   ,with $W_{i,j} \overset{\text{iid}}{\sim} (0, \tau^2)$. The matching parameters for the model $A \odot X + B$ are $\mu_A = \kappa, \sigma_A = 0, \sigma_B = \sqrt{(1 - \kappa)\tau^2}$.

Composite contamination modes

1. **Additive noise and missing-at-random.** Assume that additive noise has been added and then some entries were deleted and replaced with zeros. Then

$$Y_{i,j} = \begin{cases} X_{i,j} + Z_{i,j} & \text{w.p. } \kappa \\ 0 & \text{w.p. } 1 - \kappa \end{cases}.$$

   The equivalent A,B selection is: $\mu_A = \kappa, \sigma_A = 0, \sigma_B = \sqrt{\kappa \sigma_Z^2}$.

2. **Additive noise and outliers-at-random.**

$$Y_{i,j} = \begin{cases} X_{i,j} + Z_{i,j} & \text{w.p. } \kappa \\ X_{i,j} + W_{i,j} & \text{w.p. } 1 - \kappa \end{cases}.$$

   Where $W_{i,j} \overset{\text{iid}}{\sim} (0, \tau^2)$, and τ is large compared to σ_Z. The equivalent A,B selection is: $\mu_A = 1, \sigma_A = 0, \sigma_B = \sqrt{\kappa \sigma^2 + (1 - \kappa)\tau^2}$.

3. **Multiplicative noise and corrupt-at-random.**

$$Y_{i,j} = \begin{cases} Z_{i,j} X_{i,j} & \text{w.p. } \kappa \\ W_{i,j} & \text{w.p. } 1 - \kappa \end{cases}.$$

   The equivalent A,B selection is: $\mu_A = \kappa \cdot \mu_A, \sigma_A = 0, \sigma_B = \sqrt{(1 - \kappa)\tau^2}$.

10   Conclusions

Singular value shrinkage emerges as an effective method to reconstruct low-rank matrices from contaminated data that is both practical and well understood. Through simple, carefully designed manipulation of the data singular values, we obtain an appealing improvement in the reconstruction mean square error. While beyond our present scope, following [16], it is highly likely that the optimal shrinker we have developed offers the same mean square error, asymptotically, as the best rotation-invariant estimator based on the data, making it asymptotically the best SVD-based estimator for the target matrix.

Acknowledgements

References

[1] Benaych-Georges, Florent and Nadakuditi, Raj Rao. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111: 120–135, 2012. ISSN 0047259X.

[2] Bloemendal, Alex, Erdos, Laszlo, Knowles, Antti, Yau, Horng Tzer, and Yin, Jun. Isotropic local laws for sample covariance and generalized Wigner matrices. *Electronic Journal of Probability*, 19(33):1–53, 2014. ISSN 10836489.

[3] Boutsidis, Christos, Zouzias, Anastasios, Mahoney, Michael W, and Drineas, Petros. Randomized dimensionality reduction for k-means clustering. *IEEE Transactions on Information Theory*, 61(2):1045–1062, 2015.

[4] Bouwmans, Thierry, Sobral, Andrews, Javed, Sajid, Ki, Soon, and Zahzah, El-hadi. Decomposition into low-rank plus additive matrices for background / foreground separation : A review for a comparative evaluation with a large-scale dataset. *Computer Science Review*, 2016. ISSN 1574-0137.

[5] Buuren, Stef and Groothuis-Oudshoorn, Karin. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3), 2011.

[6] Cai, Jian-Feng, Candes, Emmanuel J., and Zuowei, Shen. A singular value thresholding algorithm for matrix completion. *2010 Society for Industrial and Applied Mathematics*, 20(4): 1956–1982, 2010.

[7] Candes, Emmanuel J. and Plan, Yaniv. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010. ISSN 00189219.

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

[9] Candès, Emmanuel J., Li, Xiaodong, Ma, Yi, and Wright, John. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, may 2011. ISSN 00045411.

[10] Candes, Emmanuel J, Sing-Long, Carlos A, and Trzasko, Joshua D. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE transactions on signal processing*, 61 (19):4643–4657, 2013.

[11] Cattell, Raymond B. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

[12] Chandrasekaran, Venkat, Sanghavi, Sujay, Parrilo, Pablo a., and Willsky, Alan S. Rank-Sparsity Incoherence for Matrix Decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011. ISSN 1052-6234.

[13] Das, Rajarshi, Zaheer, Manzil, and Dyer, Chris. Gaussian lda for topic models with word embeddings. In *ACL (1)*, pp. 795–804, 2015.

[14] Fischler, Martin A and Bolles, Robert C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[15] Gavish, Matan and Donoho, David L. The optimal hard threshold for singular values is 4/sqrt(3). *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014. ISSN 00189448.

[16] Gavish, Matan and Donoho, David L. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152, 2017.

[17] Gnanadesikan, Ramanathan and Kettenring, John R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, pp. 81–124, 1972.

[18] Golub, Gene and Kahan, William. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, 2(2):205–224, 1965.

[19] Hastie, Trevor, Tibshirani, Robert, Sherlock, Gavin, Brown, Patrick, Botstein, David, and Eisen, Michael. Imputing Missing Data for Gene Expression Arrays Imputation using the SVD. *Technical Report*, pp. 1–9, 1999.

[20] Huber, Peter J. *Robust statistics*. Springer, 2011.

[21] Ji, Hui, Liu, Chaoqiang, Shen, Zuowei, and Xu, Yuhong. Robust video denoising using low rank matrix completion. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1791–1798, 2010. ISSN 1063-6919.

[22] Johnstone, Iain M. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[23] Lin, Zhouchen, Chen, Minming, and Ma, Yi. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. 2013.

[24] Luo, Xin, Zhou, Mengchu, Xia, Yunni, and Zhu, Qingsheng. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Transactions on Industrial Informatics*, 10(2):1273–1284, 2014.

[25] Marcenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sbornik*, 1(4):457–483, 1967.

[26] Meloun, Milan, Capek, Jindrich, Miksk, Petr, and Brereton, Richard G. Critical comparison of methods predicting the number of components in spectroscopic data. *Analytica Chimica Acta*, 423(1):51–68, 2000.

[27] Nadakuditi, Raj Rao. OptShrink: An algorithm for improved low-rank signal matrix Denoising by optimal, data-driven singular value shrinkage. *IEEE Transactions on Information Theory*, 60 (5):3002–3018, 2014. ISSN 00189448.

[28] Rao, Nikhil, Yu, Hsiang-Fu, Ravikumar, Pradeep K, and Dhillon, Inderjit S. Collaborative filtering with graph information: Consistency and scalable methods. In *Advances in neural information processing systems*, pp. 2107–2115, 2015.

[29] Rennie, Jasson Dm M and Srebro, Nathan. Fast Maximum Margin Matrix Factorization for Collaborative Prediction. *Proceedings of the 22Nd International Conference on Machine Learning*, pp. 713–719, 2005. ISSN 1595931805. doi: 10.1145/1102351.1102441. URL `http://doi.acm.org/10.1145/1102351.1102441`.

[30] Rubin, Donald B. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489, 1996.

[31] Schafer, Joseph L. *Analysis of incomplete multivariate data*. CRC press, 1997.

[32] Shabalin, Andrey A and Nobel, Andrew B. Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis*, 118:67–76, 2013. ISSN 0047-259X.

[33] Stein, Charles M. Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics*, 74(5), 1986. URL `http://link.springer.com/article/10.1007/BF01085007`.

[34] Wright, John, Peng, Yigang, Ma, Yi, Ganesh, Arvind, and Rao, Shankar. Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices. *Advances in Neural Information Processing Systems (NIPS)*, pp. 2080—-2088, 2009. ISSN 0010-3640.

[35] Yang, Jian, Qian, Jianjun, Luo, Lei, Zhang, Fanlong, and Gao, Yicheng. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE Transactions on Pattern Analysis and Machine Intelligence Machine Intelligence*, pp(99): 1–1, 2016. ISSN 0162-8828.