Matrix denoising for weighted loss functions and heterogeneous signals

William Leeb[*]

Abstract

We consider the problem of estimating a low-rank matrix from a noisy observed matrix. Previous work has shown that the optimal method depends crucially on the choice of loss function. In this paper, we use a family of weighted loss functions, which arise naturally for problems such as submatrix denoising, denoising with heteroscedastic noise, and denoising with missing data. However, weighted loss functions are challenging to analyze because they are not orthogonally-invariant. We derive optimal spectral denoisers for these weighted loss functions. By combining different weights, we then use these optimal denoisers to construct a new denoiser that exploits heterogeneity in the signal matrix to boost estimation with unweighted loss.

1 Introduction

This paper is concerned with estimating a low-rank signal matrix $\mathbf{X}$ from an observed matrix $\mathbf{Y} = \mathbf{X} + \mathbf{G}$, where $\mathbf{G}$ is a full-rank matrix of noise. We consider two distinct aspects of the matrix denoising problem. First, we study methods designed for a broader family of loss functions, known as *weighted* loss functions, than considered in earlier works. Second, we design a new denoiser for *unweighted* loss that improves upon previous work by exploiting *heterogeneity* in the target matrix's singular vectors. Like many works on matrix denoising, our methods are designed for an asymptotic regime where the number of rows and columns of $\mathbf{X}$ grow infinitely large, and where the energy in the noise swamps the energy in the signal. This setting is often referred to as the *spiked model* [4, 3, 5, 47, 29, 8].

The methods introduced in this paper extend singular value shrinkage [51, 20, 19, 43, 18, 38], which modifies $\mathbf{Y}$'s singular values to mitigate the effects of noise. Our method of *spectral denoising* agrees with singular value shrinkage with unweighted loss, but performs better with weighted loss. While weighted loss functions arise in a number of applications which we describe, they are challenging as they are not orthogonally-invariant. To derive optimal spectral denoisers for weighted loss, we extend the asymptotic theory of the spiked model, building on work from [40].

Our new method of *localized denoising* is designed for unweighted loss. Unlike singular value shrinkage, however, localized denoising exploits heterogeneity in $\mathbf{X}$'s singular vectors; when certain blocks of coordinates of $\mathbf{X}$ are known to contain more of the signal's energy than others, localized denoising outperforms shrinkage. At the same time, localized denoising's asymptotic performance is never worse than shrinkage's, and so localized denoising inherits shrinkage's well-known optimality properties.

1.1 Main ideas

In the high-noise, high-dimensional spiked model, the energy of the noise $\mathbf{G}$ is unbounded as $p, n \to \infty$, while the energy of $\mathbf{X}$ is fixed. Consistent estimation of $\mathbf{X}$ from $\mathbf{Y}$ is therefore not possible, so the "best" denoiser depends on the choice of loss function. The weighted loss functions we use arise in a variety of applications, described in Section 6; the new method of spectral denoising is adapted to each of these. Table 1 lists these algorithms and their locations in the paper.

The optimal spectral denoiser for weighted loss solves a least-squares problem parameterized by weighted inner products between the singular vectors of $\mathbf{X}$ and $\mathbf{Y}$. Though formulas for unweighted inner products

---

[*]School of Mathematics, University of Minnesota, Twin Cities. Minneapolis, MN.

#	Description	Reference
1	Optimal spectral denoising for weighted loss	Section 4
2	Localized denoising for unweighted loss	Section 5
3	Submatrix denoising	Section 6.1
4	Matrix denoising with doubly-heteroscedastic noise	Section 6.2
5	Matrix denoising with missing data	Section 6.3

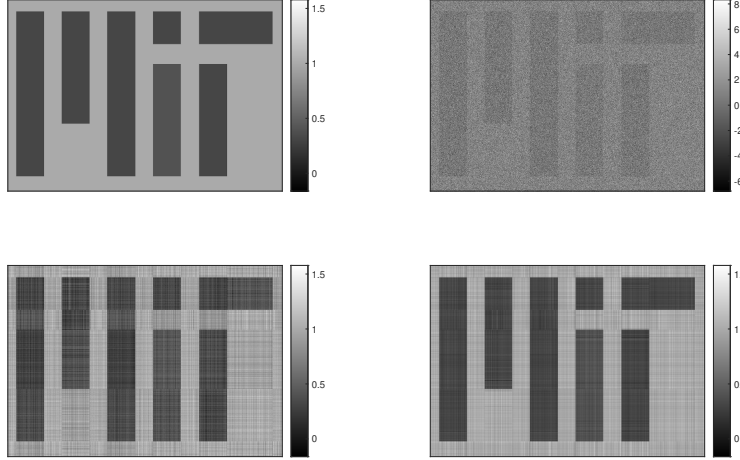Table 1: Algorithms introduced in this paper.

Figure 1: Denoising the MIT logo; see Section 1.2 for details. Upper left: the rank 5 signal matrix. Upper right: the observed noisy matrix. Lower left: the matrix denoised by optimal singular value shrinkage [20, 51]. Lower right: the matrix denoised by localized denoising (Algorithm 2). The relative error of singular value shrinkage is approximately 1.25×10^{-1}, whereas the relative error of localized denoising is approximately 7.41×10^{-2}.

are well-known [47, 8], the results we need require a new analysis extending our earlier work in [40]. While we leave the details to Theorem 3.2, the key idea is that a singular vector $\hat{\mathbf{u}}_j$ of $\mathbf{Y}$ may be written as a combination of its projection onto the corresponding singular vector $\mathbf{u}_j$ of $\mathbf{X}$ and a residual unit vector $\tilde{\mathbf{u}}_j$, $\hat{\mathbf{u}}_j = c_j \mathbf{u}_j + s_j \tilde{\mathbf{u}}_j$. Here, c_j and s_j are known from the classical theory of the spiked model [47]. Because the noise $\mathbf{G}$ is orthogonally invariant, the $\tilde{\mathbf{u}}_j$ are uniformly random in the subspace orthogonal to $\mathbf{X}$'s singular vectors. Consequently, inner products of the form $\tilde{\mathbf{u}}_j^T \mathbf{A} \tilde{\mathbf{u}}_k$ have predictable behavior when the dimension is large [21, 57, 49].

1.2 Illustrative example

The method of *localized denoising*, introduced in Section 5, uses the optimal spectral denoiser for *weighted* loss to construct a matrix denoiser for *unweighted* loss. The matrix is broken into submatrices, each of which is denoised by applying the optimal spectral denoiser with weights projecting onto that submatrix's coordinates. In Figure 1, we illustrate the performance of localized denoising on the MIT logo, which is a 1574-by-2800 matrix with rank 5. The logo is corrupted by iid Gaussian noise with standard deviation $\sigma = t_5/(1.5\gamma^{1/4})$, t_5 being the smallest singular value of the clean image. We apply optimal singular value shrinkage and localized denoising, the latter by breaking the rows into 15 equispaced segments and the columns into 30 equispaced segments. The relative error $\|\widehat{\mathbf{X}}^{\mathrm{loc}} - \mathbf{X}\|_{\mathrm{F}}/\|\mathbf{X}\|_{\mathrm{F}}$ of localized denoising is approximately 7.41×10^{-2}; the relative error of singular value shrinkage is approximately 1.25×10^{-1}, which is significantly larger. The improvement from localized denoising is due to the signal matrix's heterogeneity along the rows and columns. However, the row and column subdivisions are not chosen to extract any specific structure in the image, and localized denoising does not appear to be very sensitive to the choice of subdivisions; similar results may be obtained with other subdivisions as well.

1.3 Outline of the paper

Section 2 contains the problem statement and key definitions. Section 3 presents the new asymptotic results. Section 4 derives the optimal spectral denoiser for weighted loss. Section 5 introduces localized denoising. Section 6 describes three applications of weighted loss functions. Section 7 reports on numerical experiments. Section 8 concludes by discussing potential applications.

2 Preliminaries

2.1 The observation model

We observe a p-by-n data matrix $\mathbf{Y} = \mathbf{X} + \mathbf{G}$, consisting of a low-rank signal matrix $\mathbf{X}$ and a full-rank isotropic Gaussian noise matrix $\mathbf{G}$. We write $\mathbf{X}$ as $\mathbf{X} = \sum_{k=1}^{r} t_k \mathbf{u}_k \mathbf{v}_k^T$, where the $\mathbf{u}_k$ and $\mathbf{v}_k$ are orthonormal vectors in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively, and $t_1 > \cdots > t_r > 0$. The entries of the noise matrix $\mathbf{G}$ are iid $N(0, 1/n)$. We write $\mathbf{Y}$ as $\mathbf{Y} = \sum_{k=1}^{\min(n,p)} \lambda_k \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T$, where the $\hat{\mathbf{u}}_k$ and $\hat{\mathbf{v}}_k$ are orthonormal vectors in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively, and $\lambda_1 \geq \cdots \geq \lambda_{\min(n,p)} \geq 0$.

We let $\Omega = \Omega_p$ be one of a sequence of matrices with p columns, and $\Pi = \Pi_n$ be one of a sequence of matrices with n columns. In Section 2.3, these matrices will be used to define the loss function for estimating $\mathbf{X}$. In order to have a well-defined asymptotic theory when $p, n \to \infty$, we assume that certain quantities defined in terms of Ω, Π, and the singular vectors of $\mathbf{X}$ have definite, finite limits. We define:

$$\mu = \lim_{p \to \infty} \frac{1}{p} \mathrm{tr}(\Omega^T \Omega), \quad \nu = \lim_{n \to \infty} \frac{1}{n} \mathrm{tr}(\Pi^T \Pi). \tag{1}$$

For $1 \leq j, k \leq r$, we assume that the weighted inner products between the population singular vectors converge almost surely in the large p, large n limits:

$$e_{jk} = \lim_{p \to \infty} \langle \Omega \mathbf{u}_j, \Omega \mathbf{u}_k \rangle, \quad \tilde{e}_{jk} = \lim_{n \to \infty} \langle \Pi \mathbf{v}_j, \Pi \mathbf{v}_k \rangle. \tag{2}$$

For $1 \leq k \leq r$, we will let $\alpha_k = e_{kk}$ and $\beta_k = \tilde{e}_{kk}$.

We assume that these limits exist and are finite and positive. We also assume that the operator norms of the matrices $\Omega = \Omega_p$ and $\Pi = \Pi_n$ remain bounded as $p, n \to \infty$; that is, $\|\Omega_p\|_{\mathrm{op}}, \|\Pi_n\|_{\mathrm{op}} \leq C < \infty$ for all p and n, where C does not depend on p or n. These conditions are the only assumptions we make on the matrices Ω and Π.

We will parametrize the problem size by the number of columns n, and let the number of rows $p = p_n$ grow with n. Specifically, we will assume that the limit

$$\gamma = \lim_{n \to \infty} \frac{p_n}{n} \tag{3}$$

is well-defined and finite. In all statements where $n \to \infty$, it will be implicitly assumed as well that $p \to \infty$ and $p/n \to \gamma$. We assume that the number of population components r and the singular values $t_1, \ldots, t_r$ stay fixed with p and n.

Remark 1. All quantities that depend on p and n, such as $\mathbf{u}_k$ and $\mathbf{v}_k$, are actually elements of a *sequence* indexed by p and/or n. However, for notational simplicity, we drop the explicit dependence on p and n unless it is needed for clarity.

Remark 2. There are counterexamples to the existence of the limits in (1) and (2). For example, one may take $\mathbf{u}_1$ to be the first standard unit vector $(1, 0, \ldots, 0)^T$ when p is even, and the constant vector $(1, \ldots, 1)^T / \sqrt{p}$ when p is odd; and take $\Omega_p = \mathrm{diag}(0, 1, \ldots, 1)$. The limit defining α_1 will not exist in this case, as odd terms in the sequence $\|\Omega_p \mathbf{u}_1\|^2$ converge to 1, and even terms converge to 0. By contrast, the examples in Section 7 satisfy the asymptotic conditions.

Remark 3. The values μ, ν, e_{jk}, and $\tilde{e}_{jk}$ from equations (1) and (2) are used to characterize the weighted inner products between singular vectors of $\mathbf{X}$ and $\mathbf{Y}$; see Theorem 3.2. These weighted inner products are needed to evaluate optimal spectral denoisers for weighted loss, as described in Section 4.

3

2.2 Heterogeneity, genericity, and weighted orthogonality

One of the aspects of the theory of matrix denoising we will explore is the role of the signal matrix $\mathbf{X}$'s singular vectors, $\mathbf{u}_1, \ldots, \mathbf{u}_r$ and $\mathbf{v}_1, \ldots, \mathbf{v}_r$. To that end, we introduce two definitions we will be using throughout the paper. We say that a unit vector $\mathbf{x} \in \mathbb{R}^m$ is *generic* with respect to an m-by-m positive-semidefinite matrix $\mathbf{A}_m \in \mathbb{R}^{m \times m}$ if $\mathbf{x}^T \mathbf{A} \mathbf{x} \sim \frac{1}{m} \text{tr}(\mathbf{A})$, where "$\sim$" indicates that the difference between the two sides vanishes almost surely as $m \to \infty$ (to be precise, $\mathbf{x}$ and $\mathbf{A}$ are elements of a sequence of vectors and matrices, respectively, indexed by m; but following the convention described in Remark 1 we will drop the explicit dependence on m).

By contrast, we say that $\mathbf{x}$ is *heterogeneous* if it is not generic. This means that the energy of the vector $\mathbf{x}$ is not uniformly distributed across its coordinates in the eigenbasis of $\mathbf{A}$. Indeed, if $\mathbf{A} = \sum_{k=1}^m h_k \mathbf{w}_k \mathbf{w}_k^T$ is the eigendecomposition of $\mathbf{A}$, then

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{k=1}^m h_k \langle \mathbf{x}, \mathbf{w}_k \rangle^2. \tag{4}$$

If the energy of $\mathbf{x}$ were equally spread out across the $\mathbf{w}_k$, then $\langle \mathbf{x}, \mathbf{w}_k \rangle \sim 1/\sqrt{m}$, and so $\mathbf{x}^T \mathbf{A} \mathbf{x} \sim \text{tr}(\mathbf{A})/m$.

Given a collection of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^m$, we will say that they satisfy the *weighted orthogonality condition* (or are *weighted orthogonal*) with respect to a positive-semidefinite matrix $\mathbf{A}$ if

$$\mathbf{x}_i^T \mathbf{A} \mathbf{x}_j \sim 0 \tag{5}$$

whenever $i \neq j$. In other words, the $\mathbf{x}_j$ are asymptotically orthogonal with respect to the weighted inner product defined by $\mathbf{A}$.

Remark 4. From the Hanson-Wright inequality [21, 57, 49], random unit vectors $\mathbf{x}$ from suitably regular distributions are generic, with respect to any $\mathbf{A}$ with bounded operator norm. Furthermore, the weighted orthogonality condition will also hold for independent random unit vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ from a suitable distribution (see [7]).

2.3 Spectral denoisers and weighted loss functions

For the top r empirical singular vectors $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_r$ and $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_r$ of $\mathbf{Y}$, define the matrices $\widehat{\mathbf{U}} = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_r]$ and $\widehat{\mathbf{V}} = [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_r]$. Consider the class of estimators defined by

$$\mathcal{S} = \left\{ \widehat{\mathbf{U}} \widehat{\mathbf{B}} \widehat{\mathbf{V}}^T : \widehat{\mathbf{B}} \in \mathbb{R}^{r \times r} \right\}. \tag{6}$$

Each matrix in $\mathcal{S}$ has the same singular subspaces as the observed matrix $\mathbf{Y}$, though not necessarily the same singular vectors. We call $\mathcal{S}$ the family of *spectral denoisers*.

We consider estimating the low-rank signal matrix $\mathbf{X}$ with respect to the *weighted Frobenius loss* defined by

$$\mathcal{L}_n(\widehat{\mathbf{X}}, \mathbf{X}) = \| \Omega (\widehat{\mathbf{X}} - \mathbf{X}) \Pi^T \|_{\text{F}}^2, \tag{7}$$

where $\| \cdot \|_{\text{F}}$ denotes the matrix Frobenius norm, and Ω and Π are matrices satisfying the conditions in Section 2.1. This type of loss function is used when the user pays different prices for errors in different rows and columns.

We now define the precise estimation problem we will consider. For any deterministic r-by-r matrix $\widehat{\mathbf{B}}$, we define the asymptotic error

$$\mathcal{L}(\widehat{\mathbf{U}} \widehat{\mathbf{B}} \widehat{\mathbf{V}}^T, \mathbf{X}) = \lim_{n \to \infty} \mathcal{L}_n(\widehat{\mathbf{U}} \widehat{\mathbf{B}} \widehat{\mathbf{V}}^T, \mathbf{X}). \tag{8}$$

Our goal is then to find the matrix $\widehat{\mathbf{B}}$ to minimize this loss, and show how $\widehat{\mathbf{B}}$ may be consistently estimated from the observed matrix $\mathbf{Y}$. That is, we define

$$\widehat{\mathbf{B}} = \underset{\widehat{\mathbf{B}}' \in \mathbb{R}^{r \times r}}{\text{argmin}} \, \mathcal{L}(\widehat{\mathbf{U}} \widehat{\mathbf{B}}' \widehat{\mathbf{V}}^T, \mathbf{X}) \tag{9}$$

and define $\widehat{\mathbf{X}} = \widehat{\mathbf{U}} \widehat{\mathbf{B}} \widehat{\mathbf{V}}^T$.

Remark 5. For any deterministic $\widehat{\mathbf{B}}$, the asymptotic loss (8) exists and is finite almost surely, even though the matrices $\widehat{\mathbf{U}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}^T$ and $\mathbf{X}$ are growing in size. It will be shown in Section 4 that since $\widehat{\mathbf{U}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}^T$ and $\mathbf{X}$ each have rank at most r, $\|\Omega(\widehat{\mathbf{X}} - \mathbf{X})\Pi^T\|_{\mathrm{F}}^2$ depends only on $t_1, \ldots, t_r$; the r^2 entries of $\widehat{\mathbf{B}}$; and the weighted inner products between the top r singular vectors of $\mathbf{Y}$ and $\mathbf{X}$. It will follow from Theorem 3.2 that these inner products converge almost surely to finite limits, and consequently that the asymptotic loss (8) is well-defined almost surely.

3 Asymptotic theory for the spiked model

In this section, we derive the limits of inner products between the weighted population and empirical vectors. We define the cosines between the unweighted empirical and population vectors:

$$c_{jk} = \lim_{p \to \infty} \langle \hat{\mathbf{u}}_j, \mathbf{u}_k \rangle, \quad \tilde{c}_{jk} = \lim_{n \to \infty} \langle \hat{\mathbf{v}}_j, \mathbf{v}_k \rangle. \tag{10}$$

Next we define *weighted* inner products between the population and empirical vectors:

$$c_{jk}^{\omega} = \lim_{p \to \infty} \langle \Omega \hat{\mathbf{u}}_j, \Omega \mathbf{u}_k \rangle, \quad \tilde{c}_{jk}^{\omega} = \lim_{n \to \infty} \langle \Pi \hat{\mathbf{v}}_j, \Pi \mathbf{v}_k \rangle. \tag{11}$$

These are inner products with weight matrices $\Omega^T\Omega$ and $\Pi^T\Pi$, respectively. We also define the weighted inner products between the empirical singular vectors:

$$d_{jk} = \lim_{p \to \infty} \langle \Omega \hat{\mathbf{u}}_j, \Omega \hat{\mathbf{u}}_k \rangle, \quad \tilde{d}_{jk} = \lim_{n \to \infty} \langle \Pi \hat{\mathbf{v}}_j, \Pi \hat{\mathbf{v}}_k \rangle. \tag{12}$$

We let $c_k^{\omega} = c_{kk}^{\omega}$ and $\tilde{c}_k^{\omega} = \tilde{c}_{kk}^{\omega}$, and similarly for the other terms.

Remark 6. From Theorem 3.2 below, the limits (11) and (12) exist almost surely and are finite so long as the assumptions on Ω and Π from Section 2.1 hold.

The first result provides formulas for c_{jk} and $\tilde{c}_{jk}$, and relates the singular values of $\mathbf{X}$ to those of $\mathbf{Y}$. It is well-known in the literature (see e.g. [47, 8]).

Proposition 3.1. *For $1 \le k \le r$, the kth singular value of $\mathbf{Y}$ converges almost surely as $p, n \to \infty$ to λ_k, defined by:*

$$\lambda_k^2 = \begin{cases} (t_k^2 + 1)\left(1 + \frac{\gamma}{t_k^2}\right), & \text{if } t_k > \gamma^{1/4}, \\ (1 + \sqrt{\gamma})^2, & \text{if } t_k \le \gamma^{1/4}, \end{cases} \tag{13}$$

For $1 \le j, k \le r$, the limits (10) defining c_{jk} and $\tilde{c}_{jk}$ almost surely exist and are given by the following expressions:

$$c_{jk}^2 = \begin{cases} \frac{1 - \gamma/t_k^4}{1 + \gamma/t_k^2}, & \text{if } j = k \text{ and } t_k > \gamma^{1/4}, \\ 0, & \text{if } j \ne k \text{ or } t_k \le \gamma^{1/4}, \end{cases} \tag{14}$$

and

$$\tilde{c}_{jk}^2 = \begin{cases} \frac{1 - \gamma/t_k^4}{1 + 1/t_k^2}, & \text{if } j = k \text{ and } t_k > \gamma^{1/4}, \\ 0, & \text{if } j \ne k \text{ or } t_k \le \gamma^{1/4}. \end{cases} \tag{15}$$

Remark 7. While the signs of c_k and $\tilde{c}_k$ are arbitrary, their product satisfies $c_k \tilde{c}_k \ge 0$. We may therefore assume that $c_k \ge 0$ and $\tilde{c}_k \ge 0$ (see, e.g., [43]).

Theorem 3.2. *Suppose $1 \le j, k \le r$. Then the limits (11) and (12) almost surely exist and are equal to the following expressions:*

$$c_{jk}^{\omega} = \begin{cases} e_{jk}c_j, & \text{if } t_j > \gamma^{1/4}, \\ 0, & \text{if } t_j \le \gamma^{1/4}, \end{cases} \tag{16}$$

$$\tilde{c}_{jk}^{\omega} = \begin{cases} \tilde{e}_{jk}\tilde{c}_j, & \text{if } t_j > \gamma^{1/4}, \\ 0, & \text{if } t_j \le \gamma^{1/4}, \end{cases} \tag{17}$$

$$d_{jk} = \begin{cases} c_k^2 \alpha_k + s_k^2 \mu, & \text{if } j = k \text{ and } t_k > \gamma^{1/4}, \\ e_{jk} c_j c_k, & \text{if } j \ne k \text{ and } \min\{t_j, t_k\} > \gamma^{1/4}, \\ 0, & \text{if } j \ne k \text{ and } \min\{t_j, t_k\} \le \gamma^{1/4}, \end{cases} \tag{18}$$

$$\tilde{d}_{jk} = \begin{cases} \tilde{c}_k^2 \beta_k + \tilde{s}_k^2 \nu, & \text{if } j = k \text{ and } t_k > \gamma^{1/4}, \\ \tilde{e}_{jk} \tilde{c}_j \tilde{c}_k, & \text{if } j \ne k \text{ and } \min\{t_j, t_k\} > \gamma^{1/4}, \\ 0, & \text{if } j \ne k \text{ and } \min\{t_j, t_k\} \le \gamma^{1/4}. \end{cases} \tag{19}$$

The proof of Theorem 3.2 may be found in Section A.

Remark 8. While the signs of inner products between singular vectors are arbitrary, Theorem 3.2 states that once the signs of e_{jk} and $\tilde{e}_{jk}$ are fixed, the signs of c_{jk}^{ω}, $\tilde{c}_{jk}^{\omega}$, d_{jk} and $\tilde{d}_{jk}$ are determined.

4 Optimal spectral denoising

In this section, we derive the asymptotically optimal spectral denoiser with respect to the weighted loss (8), and show how to consistently estimate it from $\mathbf{Y}$. We define the r-by-r weighted inner product matrices $\mathbf{D} = (d_{kl})$, $\widetilde{\mathbf{D}} = (\tilde{d}_{jk})$, $\mathbf{E} = (e_{jk})$, $\widetilde{\mathbf{E}} = (\tilde{e}_{jk})$, $\mathbf{C} = (c_{jk}^{\omega})$, and $\widetilde{\mathbf{C}} = (\tilde{c}_{jk}^{\omega})$, and the vector $\mathbf{t} = (t_1, \ldots, t_r)^T$ of population singular values.

Theorem 4.1. *The optimal choice of $\widehat{\mathbf{B}}$ is given by $\widehat{\mathbf{B}} = \mathbf{D}^+ \mathbf{C} \mathrm{diag}(\mathbf{t}) \widetilde{\mathbf{C}}^T \widetilde{\mathbf{D}}^+$, with weighted AMSE almost surely equal to*

$$\lim_{n \to \infty} \|\Omega(\widehat{\mathbf{X}} - \mathbf{X})\Pi^T\|_F^2 = \langle \mathbf{E} \mathrm{diag}(\mathbf{t}) \widetilde{\mathbf{E}} - \mathbf{C}^T \mathbf{D}^+ \mathbf{C} \mathrm{diag}(\mathbf{t}) \widetilde{\mathbf{C}}^T \widetilde{\mathbf{D}}^+ \widetilde{\mathbf{C}}, \mathrm{diag}(\mathbf{t}) \rangle_F. \tag{20}$$

The proof of Theorem 4.1 may be found in Section B.

The matrices $\mathbf{D}$, $\widetilde{\mathbf{D}}$, $\mathbf{E}$, $\widetilde{\mathbf{E}}$, $\mathbf{C}$, and $\widetilde{\mathbf{C}}$ and the singular values $t_1, \ldots, t_r$ may be estimated using Proposition 3.1 and Theorem 3.2. First, from Proposition 3.1, we can estimate t_k, c_k and $\tilde{c}_k$, so long as $t_k > \gamma^{1/4}$, i.e. if $\lambda_k > 1 + \sqrt{\gamma}$:

$$t_k = \sqrt{\frac{\lambda_k^2 - 1 - \gamma + \sqrt{(\lambda_k^2 - 1 - \gamma)^2 - 4\gamma}}{2}}, \quad c_k = \sqrt{\frac{1 - \gamma/t_k^4}{1 + \gamma/t_k^2}}, \quad \tilde{c}_k = \sqrt{\frac{1 - \gamma/t_k^4}{1 + 1/t_k^2}}. \tag{21}$$

Remark 9. From Remark 7, we can take both c_k and $\tilde{c}_k$ to be positive.

The values $d_{jk} = \hat{\mathbf{u}}_j^T \Omega^T \Omega \hat{\mathbf{u}}_j$ and $\tilde{d}_{jk} = \hat{\mathbf{v}}_j^T \Pi^T \Pi \hat{\mathbf{v}}_k$ are directly estimable, as they are the weighted inner products between the empirical singular vectors. We then estimate α_k and β_k, assuming $t_k > \gamma^{1/4}$:

$$\alpha_k = \frac{d_k - s_k^2 \mu}{c_k^2}, \quad \beta_k = \frac{\tilde{d}_k - \tilde{s}_k^2 \nu}{\tilde{c}_k^2}. \tag{22}$$

When $j \ne k$, we take $e_{jk} = d_{jk}/(c_j c_k)$ and $\tilde{e}_{jk} = \tilde{d}_{jk}/(\tilde{c}_j \tilde{c}_k)$ (so long as t_j and t_k both exceed $\gamma^{1/4}$, i.e. λ_j and λ_k both exceed $1 + \sqrt{\gamma}$). Finally, for all j, k, we take $c_{jk}^{\omega} = e_{jk} c_j$ and $\tilde{c}_{jk} = \tilde{e}_{jk} \tilde{c}_j$. The method is summarized in Algorithm 1.

Algorithm 1 Optimal spectral denoising for weighted loss

1: Input: $\mathbf{Y}$; weights Ω and Π
2: rank r SVD of $\mathbf{Y}$:

$\lambda_1 \geq \cdots \geq \lambda_r > 1 + \sqrt{\gamma}$
$\widehat{\mathbf{U}} = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_r], \ \widehat{\mathbf{V}} = [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_r]$

3: $\mu = \mathrm{tr}(\Omega^T \Omega)/p, \ \nu = \mathrm{tr}(\Pi^T \Pi)/n$
4: for $1 \leq k \leq r$:

$t_k = \sqrt{\dfrac{\lambda_k^2 - 1 - \gamma + \sqrt{(\lambda_k^2 - 1 - \gamma)^2 - 4\gamma}}{2}}$

$c_k = \sqrt{\dfrac{1 - \gamma/t_k^4}{1 + \gamma/t_k^2}}, \ \tilde{c}_k = \sqrt{\dfrac{1 - \gamma/t_k^4}{1 + 1/t_k^2}}$

$s_k = \sqrt{1 - c_k^2}, \tilde{s}_k = \sqrt{1 - \tilde{c}_k^2}$
$d_k = \|\Omega \hat{\mathbf{u}}_k\|^2, \ \tilde{d}_k = \|\Pi \hat{\mathbf{v}}_k\|^2$
$\alpha_k = (d_k - s_k^2 \mu)/c_k^2, \ \beta_k = (\tilde{d}_k - \tilde{s}_k^2 \nu)/\tilde{c}_k^2$
$c_k^\omega = \alpha_k c_k, \ \tilde{c}_k^\omega = \beta_k \tilde{c}_k$

5: for $1 \leq j \neq k \leq r$:

$d_{jk} = \hat{\mathbf{u}}_j^T \Omega^T \Omega \hat{\mathbf{u}}_k, \ \tilde{d}_{jk} = \hat{\mathbf{v}}_j^T \Pi^T \Pi \hat{\mathbf{v}}_k$
$e_{jk} = d_{jk}/(c_j c_k), \ \tilde{e}_{jk} = \tilde{d}_{jk}/(\tilde{c}_j \tilde{c}_k), \ j \neq k$
$c_{jk}^\omega = e_{jk} c_j, \ \tilde{c}_{jk}^\omega = \tilde{e}_{jk} \tilde{c}_j$

6: $\mathbf{D} = (d_{jk}), \ \widetilde{\mathbf{D}} = (\tilde{d}_{jk})$
$\mathbf{E} = (e_{jk}), \ \widetilde{\mathbf{E}} = (\tilde{e}_{jk})$
$\mathbf{C} = (c_{jk}^\omega), \ \widetilde{\mathbf{C}} = (\tilde{c}_{jk}^\omega)$
7: $\mathbf{t} = (t_1, \ldots, t_r)^T$
8: $\widehat{\mathbf{B}} = \mathbf{D}^+ \mathbf{C} \mathrm{diag}(\mathbf{t}) \widetilde{\mathbf{C}}^T \widetilde{\mathbf{D}}^+$
9: $\widehat{\mathbf{X}} = \widehat{\mathbf{U}} \widehat{\mathbf{B}} \widehat{\mathbf{V}}^T$
10: $\mathrm{AMSE} = \langle \mathbf{E} \mathrm{diag}(\mathbf{t}) \widetilde{\mathbf{E}} - \mathbf{C}^T \mathbf{D}^+ \mathbf{C} \mathrm{diag}(\mathbf{t}) \widetilde{\mathbf{C}}^T \widetilde{\mathbf{D}}^+ \widetilde{\mathbf{C}}, \mathrm{diag}(\mathbf{t}) \rangle_{\mathrm{F}}$

4.1 Diagonal denoisers

In this section, we consider a subset of spectral denoisers, in which the matrix $\widehat{\mathbf{B}}$ is required to be diagonal. More precisely, we search for a vector $\hat{\mathbf{t}} = (\hat{t}_1, \ldots, \hat{t}_r)^T$ of real numbers, so that the estimator

$$\widehat{\mathbf{X}}^{\mathrm{dd}} = \widehat{\mathbf{U}} \mathrm{diag}(\hat{\mathbf{t}}) \widehat{\mathbf{V}}^T = \sum_{k=1}^r \hat{t}_k \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T \tag{23}$$

minimizes the AMSE $\mathcal{L}(\widehat{\mathbf{X}}^{\mathrm{dd}}, \mathbf{X}) = \lim_{n \to \infty} \|\Omega(\widehat{\mathbf{X}}^{\mathrm{dd}} - \mathbf{X})\Pi^T\|_{\mathrm{F}}^2$.

Remark 10. Optimal diagonal denoising cannot have better asymptotic performance than optimal spectral denoising, as the diagonal denoiser is a spectral denoiser. However, Theorem 4.2 below shows that under weighted orthogonality, the methods coincide; and the simplicity of $\widehat{\mathbf{X}}^{\mathrm{dd}}$ makes it easier to analyze, which will be exploited in the proofs of Theorem 5.2 and Proposition 6.1 and the analysis of Section 4.2.

Theorem 4.2. *Suppose that either* $\mathbf{u}_1, \ldots, \mathbf{u}_r$ *are weighted orthogonal with respect to* $\Omega^T \Omega$, *or* $\mathbf{v}_1, \ldots, \mathbf{v}_r$ *are weighted orthogonal with respect to* $\Pi^T \Pi$. *Suppose too that* $t_k > \gamma^{1/4}$, $1 \leq k \leq r$. *Then the singular values* $\hat{t}_k$, $1 \leq k \leq r$, *of* $\widehat{\mathbf{X}}^{\mathrm{dd}}$ *are:*

$$\hat{t}_k = t_k c_k \tilde{c}_k \cdot \frac{\alpha_k}{c_k^2 \alpha_k + s_k^2 \mu} \cdot \frac{\beta_k}{\tilde{c}_k^2 \beta_k + \tilde{s}_k^2 \nu}, \tag{24}$$

where t_k, c_k *and* $\tilde{c}_k$ *are given by* (21), *and* α_k *and* β_k *are given by* (22). *The weighted AMSE is almost surely equal to*

$$\lim_{n \to \infty} \|\Omega(\widehat{\mathbf{X}} - \mathbf{X})\Pi^T\|_{\mathrm{F}}^2 = \sum_{k=1}^r t_k^2 \alpha_k \beta_k \left(1 - c_k^2 \tilde{c}_k^2 \cdot \frac{\alpha_k}{c_k^2 \alpha_k + s_k^2 \mu} \cdot \frac{\beta_k}{\tilde{c}_k^2 \beta_k + \tilde{s}_k^2 \nu}\right). \tag{25}$$

7

If $\mathbf{u}_1, \ldots, \mathbf{u}_r$ and $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are both weighted orthogonal with respect to $\Omega^T \Omega$ and $\Pi^T \Pi$, respectively, then $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}^{\mathrm{dd}}$.

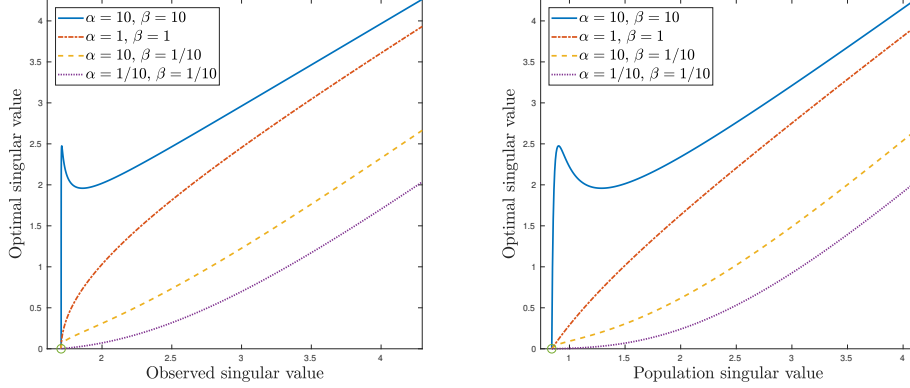The proof of Theorem 4.2 is found in Section C.

Figure 2: The optimal singular value $\hat{t}$, plotted as a function of the observed singular value λ (left) and the population singular value t (right), for varying values of α and β and $\mu = \nu = 1$.

4.2 Behavior of the optimal singular values

In this section, we assume either that $r = 1$; or that $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are weighted orthogonal with respect to $\Omega^T \Omega$ and $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are weighted orthogonal with respect to $\Pi^T \Pi$. In either case, the optimal spectral denoiser coincides with the optimal diagonal denoiser, and both are given by Theorem 4.2. Though this setting is quite restrictive, it permits us to exploit formula (24) for the optimal singular values to gain insight into the behavior of the optimal spectral denoiser. Propositions 4.3 and 4.4 describe the behavior of the optimal singular value $\hat{t}_k$ in this setting. Because each $\hat{t}_k$ depends only on the information specific to component k, we will drop the subscript k from the notation.

Proposition 4.3. *If either $\alpha \leq \mu$ or $\beta \leq \nu$, then $\hat{t} \leq \lambda$. Conversely, for any fixed value of t, there are sufficiently large values of α and β for which $\hat{t} > \lambda$.*

Proposition 4.4. *If $\alpha \leq \mu$ or $\beta \leq \nu$, then $\hat{t}$ is an increasing function of λ.*

The proofs of Propositions 4.3 and 4.4 may be found in Section D and Section E, respectively.

Remark 11. From [51, 20, 43], the optimal singular value for unweighted Frobenius loss is $\hat{t}^{\mathrm{shr}} = tc\tilde{c}$, which is smaller than the observed singular value λ. Proposition 4.3 shows that with *weighted* loss, such shrinkage only occurs for small α or β.

The conclusion of Proposition 4.4 need not hold if $\alpha > \mu$ and $\beta > \mu$. In Figure 2 we plot the optimal $\hat{t}$, both as a function of the observed singular value λ and the population singular value t, for various values of α and β (and $\mu = \nu = 1$). The non-monotonicity is apparent when $\alpha = \beta = 10$.

5 Localized denoising

In this section, we introduce a new procedure called *localized denoising* for estimating $\mathbf{X}$ with *unweighted* Frobenius loss. As we will show, localized denoising is asymptotically never worse than optimal singular value shrinkage [20, 51], defined by $\widehat{\mathbf{X}}^{\mathrm{shr}} = \sum_{k=1}^{r} \hat{t}_k^{\mathrm{shr}} \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T$, where $\hat{t}_k^{\mathrm{shr}} = t_k c_k \tilde{c}_k$. Since singular value shrinkage is optimal for unweighted loss both in the minimax sense and when averaging over a uniform prior on $\mathbf{u}_k$ and $\mathbf{v}_k$ [18, 51], localized denoising inherits these same optimality properties. Furthermore, localized denoising can outperform singular value shrinkage when the singular vectors of $\mathbf{X}$ are heterogeneous.

8

5.1 Definition of localized denoising

To define localized denoising, we expand the identity matrices $\mathbf{I}_p = \sum_{i=1}^{I} \Omega_i$ and $\mathbf{I}_n = \sum_{j=1}^{J} \Pi_j$ into sums of pairwise orthogonal projections $\Omega_i \in \mathbb{R}^{p \times p}$ and $\Pi_j \in \mathbb{R}^{n \times n}$, where I and J are fixed. We require that $\Omega_i = \Omega_i^T = \Omega_i^2$ and $\Omega_{i'}\Omega_i = \mathbf{O}_{p \times p}$ for $i \neq i'$; and similarly for the Π_j.

We let $\widehat{\mathbf{X}}_{(i,j)}^{\mathrm{loc}}$ denote the optimal spectral denoiser with respect to the weight matrices Ω_i and Π_j. We then define the *locally-denoised matrix*:

$$\widehat{\mathbf{X}}^{\mathrm{loc}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \Omega_i \widehat{\mathbf{X}}_{(i,j)}^{\mathrm{loc}} \Pi_j. \tag{26}$$

We summarize the localized denoising procedure in Algorithm 2.

Algorithm 2 Localized denoising for unweighted loss

1: Input: $\mathbf{Y}$; pairwise orthogonal projections $\Omega_1, \ldots, \Omega_I, \Pi_1, \ldots, \Pi_J$
   $\sum_{i=1}^{I} \Omega_i = \mathbf{I}_p$; $\sum_{j=1}^{J} \Pi_j = \mathbf{I}_n$
2: for $1 \leq i \leq I$, $1 \leq j \leq J$:
   $\widehat{\mathbf{X}}_{(i,j)}^{\mathrm{loc}}$ is output of Algorithm 1 with weights Ω_i and Π_j
   $\mathrm{AMSE}_{(i,j)}^{\mathrm{loc}}$ is estimated mean squared error
3: $\widehat{\mathbf{X}}^{\mathrm{loc}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \Omega_i \widehat{\mathbf{X}}_{(i,j)}^{\mathrm{loc}} \Pi_j$
4: $\mathrm{AMSE}^{\mathrm{loc}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \mathrm{AMSE}_{(i,j)}^{\mathrm{loc}}$

5.2 Performance of localized denoising

The following results compare the behavior of the localized denoiser $\widehat{\mathbf{X}}^{\mathrm{loc}}$ to the optimal singular value shrinker $\widehat{\mathbf{X}}^{\mathrm{shr}}$.

Theorem 5.1. $\|\widehat{\mathbf{X}}^{\mathrm{loc}} - \mathbf{X}\|_{\mathrm{F}}^2 \leq \|\widehat{\mathbf{X}}^{\mathrm{shr}} - \mathbf{X}\|_{\mathrm{F}}^2$ *almost surely as* $p, n \to \infty$.

Theorem 5.2. *Suppose that either* $\mathbf{u}_1, \ldots, \mathbf{u}_r$ *are weighted orthogonal with respect to all* Ω_i, *or* $\mathbf{v}_1, \ldots, \mathbf{v}_r$ *are weighted orthogonal with respect to all* Π_j. *Then almost surely as* $p, n \to \infty$, $\|\widehat{\mathbf{X}}^{\mathrm{loc}} - \mathbf{X}\|_{\mathrm{F}}^2 \leq \|\widehat{\mathbf{X}}^{\mathrm{shr}} - \mathbf{X}\|_{\mathrm{F}}^2 - \xi$, *where* $\xi \geq 0$, *and* $\xi > 0$ *if some* $\mathbf{u}_k$ *is heterogeneous with respect to some* Ω_i *or some* $\mathbf{v}_k$ *is heterogeneous with respect to some* Π_j.

In other words, unless all the $\mathbf{u}_k$ are generic with respect to all of the Ω_i *and* all the $\mathbf{v}_k$ are generic with respect to all of the Π_i, localized denoising will outperform singular value shrinkage asymptotically. The proofs of Theorems 5.1 and 5.2 are found in Section F and Section G, respectively.

Remark 12. The weighted orthogonality condition of Theorem 5.2 will hold if the columns of $\mathbf{X}$ are drawn iid from a sufficiently well-behaved distribution in $\mathbb{R}^p$.

Remark 13. To apply Theorem 5.2, the user must select projection matrices Ω_i and Π_j with respect to which the singular vectors of $\mathbf{X}$ are heterogeneous. Datasets are often drawn from different experimental regimes in genetic microarray experiments [27, 41, 50], single-cell RNA processing [52, 55], and medical imaging [35]. In these settings, it is known a priori that signal components will likely be heterogeneous across the different subpopulations, and localized shrinkage is a natural tool.

Remark 14. Theorem 5.1 guarantees that even if the projection matrices Ω_i and Π_j are not chosen judiciously (see Remark 13), the asymptotic performance of localized denoising is never worse than singular value shrinkage. In practice, localized denoising requires estimating more parameters than does shrinkage, and if I and J are sizeable relative to p and n its performance might be worse due to finite sample fluctuations in these parameter estimates, especially when the singular vectors of $\mathbf{X}$ do not exhibit strong heterogeneity with respect to the projections. For such an example, see Section 7.1, and specifically Remark 21. Though a detailed analysis of this topic is beyond the scope of the present work, in practice the user can compare these trade-offs via simulation to determine if localized denoising is appropriate for their problem size and the expected level of heterogeneity with respect to the projections.

6 Applications of weighted denoising

In this section, we describe three applications of weighted loss functions: submatrix denoising, denoising with heteroscedastic noise, and denoising with missing data. In these problems, we estimate a low-rank matrix with respect to *unweighted* Frobenius loss. However, an intermediate step of the estimation procedure requires the use of a *weighted* loss function.

6.1 Submatrix denoising

We suppose we observe a data matrix $\mathbf{Y} = \mathbf{X} + \mathbf{G}$, but our goal is to estimate only a p_0-by-n_0 submatrix of $\mathbf{X}$, where $p_0/p \sim \mu$ and $n_0/n \sim \nu$. Denoting by $\Omega \in \mathbb{R}^{p_0 \times p}$ the coordinate selection operator for the p_0 rows of the submatrix, and $\Pi \in \mathbb{R}^{n_0 \times n}$ the coordinate selection operator for the n_0 columns of the submatrix, we may write the target submatrix as $\mathbf{X}_0 = \Omega \mathbf{X} \Pi^T$.

One approach is to estimate the entire matrix $\mathbf{X}$ with respect to the weighted loss $\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X}) = \|\Omega(\widehat{\mathbf{X}} - \mathbf{X})\Pi^T\|_{\mathrm{F}}^2$. This loss only penalizes errors in the p_0 rows and n_0 columns of $\mathbf{X}_0$. If $\widehat{\mathbf{X}}$ denotes the optimal spectral denoiser minimizing $\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X})$, we define our estimator $\widehat{\mathbf{X}}_0 = \Omega \widehat{\mathbf{X}} \Pi^T$. The method is summarized in Algorithm 3.

Algorithm 3 Submatrix denoising

1: Input: $\mathbf{Y}$; projections Ω, Π
2: $\widehat{\mathbf{X}}$ is output of Algorithm 1 with weights Ω and Π
   AMSE is estimated mean squared error
3: $\widehat{\mathbf{X}}_0 = \Omega \widehat{\mathbf{X}} \Pi^T$

Another natural approach is to simply ignore the $p - p_0$ rows and $n - n_0$ columns outside of $\mathbf{X}_0$, and denoise $\mathbf{X}_0$ directly by optimal singular value shrinkage to the matrix $\mathbf{Y}_0 = \mathbf{X}_0 + \mathbf{G}_0$ (defining $\mathbf{G}_0 = \Omega \mathbf{G} \Pi^T$). We let $\widehat{\mathbf{X}}_0^{\mathrm{shr}}$ denote this estimator.

In the following result, we make the same assumptions on Ω and Π from Section 2.1. Note that $p_0 = \mathrm{tr}(\Omega^T \Omega)$, and $n_0 = \mathrm{tr}(\Pi^T \Pi)$.

Proposition 6.1. *Suppose $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are weighted orthogonal with respect to $\Omega^T \Omega$, and $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are weighted orthogonal with respect to $\Pi^T \Pi$. Suppose $\alpha_k < \sqrt{\mu}$ and $\beta_k < \sqrt{\nu}$, for $1 \leq k \leq r$. Then $\|\widehat{\mathbf{X}}_0 - \mathbf{X}_0\|_{\mathrm{F}}^2 < \|\widehat{\mathbf{X}}_0^{\mathrm{shr}} - \mathbf{X}_0\|_{\mathrm{F}}^2$, where the strict inequality holds almost surely in the limit $p, n \to \infty$.*

The proof of Proposition 6.1 is found in Section H.

Remark 15. If $\mathbf{u}_k$ and $\mathbf{v}_k$ are generic with respect to $\Omega^T \Omega$ and $\Pi^T \Pi$, respectively, then $\alpha_k = \mu$ and $\beta_k = \nu$. Proposition 6.1 requires the much weaker condition that $\alpha_k \leq \sqrt{\mu}$ and $\beta_k \leq \sqrt{\nu}$ (note that $\mu \leq \sqrt{\mu}$ and $\nu \leq \sqrt{\nu}$). Informally, even if the fraction of the signal's energy contained in $\mathbf{X}_0$ is disproportionately large, it still pays to denoise $\mathbf{X}_0$ using the entire observed matrix $\mathbf{Y}$, rather than the submatrix $\mathbf{Y}_0$ alone.

It will follow from the proof of Proposition 6.1 that if $\alpha_k < \sqrt{\mu}$ and $\beta_k < \sqrt{\nu}$, then the singular vectors of $\mathbf{X}_0$ are better approximated by computing the singular vectors of $\mathbf{Y}$ and projecting onto the images of Ω and Π, respectively, rather than computing the singular vectors of the submatrix $\mathbf{Y}_0$ itself. More precisely, we will show that $\mathbf{u}_k^0 = \frac{\Omega \mathbf{u}_k}{\|\Omega \mathbf{u}_k\|}$ and $\mathbf{v}_k^0 = \frac{\Pi \mathbf{v}_k}{\|\Pi \mathbf{v}_k\|}$ are the singular vectors of $\mathbf{X}_0$, and that the vectors $\hat{\mathbf{u}}_k^\omega = \frac{\Omega \hat{\mathbf{u}}_k}{\|\Omega \hat{\mathbf{u}}_k\|}$ and $\hat{\mathbf{v}}_k^\omega = \frac{\Pi \hat{\mathbf{v}}_k}{\|\Pi \hat{\mathbf{v}}_k\|}$, are better correlated with $\mathbf{u}_k^0$ and $\mathbf{v}_k^0$, respectively, then are the singular vectors of $\mathbf{Y}_0$.

6.2 Doubly-heteroscedastic noise

We consider estimating a low-rank matrix $\mathbf{X}$ from an observed matrix $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, where $\mathbf{N}$ is a noise matrix of the form $\mathbf{N} = \mathbf{S}^{1/2} \mathbf{G} \mathbf{T}^{1/2}$, $\mathbf{G}$ has iid entries with distribution $N(0, 1/n)$, and $\mathbf{S} \in \mathbb{R}^{p \times p}$ and $\mathbf{T} \in \mathbb{R}^{n \times n}$ are positive-definite matrices. We assume the eigenvalues of $\mathbf{S} = \mathbf{S}_p$ and $\mathbf{T} = \mathbf{T}_n$ remain in an interval $[a, b]$ for all p and n, where $a > 0$ and $b < \infty$ are fixed independently of p and n. We refer to the matrix $\mathbf{N}$ as *doubly-heteroscedastic* noise.

Remark 16. This noise model generalizes two previous models of heteroscedastic noise in the context of principal component analysis [22, 23, 24, 58, 40]. In both, the matrix $\mathbf{X}$ consists of random, iid signal vectors $X_1, \ldots, X_n$ of the form $X_j = \sum_{k=1}^r \ell_k^{1/2} z_{jk} \mathbf{u}_k$, where $\ell_1 > \cdots > \ell_r > 0$, $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are orthonormal vectors, and the z_{jk} are iid random variables with variance 1 and mean 0. The model from [58] and [40] takes $\mathbf{T} = \mathbf{I}_n$, in which case the observations are of the form $Y_j = X_j + \mathbf{S}^{1/2} G_j$, where $G_j \sim N(\mathbf{0}, \mathbf{I}_p)$. By contrast, the papers [22, 23, 24] take $\mathbf{S} = \mathbf{I}_p$, in which case the observations are of the form $Y_j = X_j + b_j^{1/2} G_j$. The doubly-heteroscedastic noise model takes $Y_j = X_j + b_j^{1/2} \mathbf{S}^{1/2} G_j$, which generalizes both these models.

We consider the following three-step procedure. First, we *whiten* the noise, replacing $\mathbf{Y}$ by $\widetilde{\mathbf{Y}}$ defined by $\widetilde{\mathbf{Y}} = \mathbf{S}^{-1/2} \mathbf{Y} \mathbf{T}^{-1/2}$. We may write $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{X}} + \mathbf{G}$, where $\widetilde{\mathbf{X}} = \mathbf{S}^{-1/2} \mathbf{X} \mathbf{T}^{-1/2}$ and $\mathbf{G}$ has iid $N(0, 1/n)$ entries. Next, we apply a denoiser to $\widetilde{\mathbf{Y}}$ to estimate $\widetilde{\mathbf{X}}$; we denote this by $\psi(\widetilde{\mathbf{Y}})$, for ψ tailored to removing *white* noise. Finally, we *unwhiten* $\psi(\widetilde{\mathbf{Y}})$ to obtain our final estimate $\widehat{\mathbf{X}} = \mathbf{S}^{1/2} \psi(\widetilde{\mathbf{Y}}) \mathbf{T}^{1/2}$ of $\mathbf{X}$.

The Frobenius loss between $\widehat{\mathbf{X}}$ and $\mathbf{X}$ may be written as follows:

$$\|\widehat{\mathbf{X}} - \mathbf{X}\|_{\mathrm{F}}^2 = \|\mathbf{S}^{1/2} \psi(\widetilde{\mathbf{Y}}) \mathbf{T}^{1/2} - \mathbf{S}^{1/2} \widetilde{\mathbf{X}} \mathbf{T}^{1/2}\|_{\mathrm{F}}^2 = \|\mathbf{S}^{1/2} (\psi(\widetilde{\mathbf{Y}}) - \widetilde{\mathbf{X}}) \mathbf{T}^{1/2}\|_{\mathrm{F}}^2, \tag{27}$$

which is a *weighted* loss between $\psi(\widetilde{\mathbf{Y}})$ and $\widetilde{\mathbf{X}}$, with weights $\mathbf{S}^{1/2}$ and $\mathbf{T}^{1/2}$. The denoiser $\psi(\widetilde{\mathbf{Y}})$ should be chosen to minimize this weighted Frobenius loss. The procedure is summarized in Algorithm 4, where ψ is taken to be optimal spectral denoising,

---

Algorithm 4 Matrix denoising with doubly-heteroscedastic noise

---

1: Input: $\mathbf{Y}$; positive-definite $\mathbf{S}$, $\mathbf{T}$
2: $\widetilde{\mathbf{Y}} = \mathbf{S}^{-1/2} \mathbf{Y} \mathbf{T}^{-1/2}$
3: $\psi(\widetilde{\mathbf{Y}})$ is output of Algorithm 1 with weights $\mathbf{S}^{1/2}$ and $\mathbf{T}^{1/2}$
   AMSE is estimated mean squared error
4: $\widehat{\mathbf{X}} = \mathbf{S}^{1/2} \psi(\widetilde{\mathbf{Y}}) \mathbf{T}^{1/2}$

Remark 17. The procedure of whitening, denoising, and unwhitening has been employed in recent papers on the spiked model; see, for instance, [42, 40, 16]. In particular, [40] shows several advantages of working with the whitened matrix when the noise is one-sided, such as improved estimation of the singular vectors of $\mathbf{X}$. By contrast, the paper [24] shows that whitening is suboptimal in certain settings.

6.2.1 Estimating S and T

The signal/noise decomposition $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ is obviously not well-defined unless the user possesses some additional knowledge about the noise matrix $\mathbf{N} = \mathbf{S}^{1/2} \mathbf{G} \mathbf{T}^{1/2}$. While a detailed treatment of this problem is outside the scope of this paper, we observe that in the large p, large n asymptotic limit, the matrices $\mathbf{S} = \mathbf{S}_p$ and $\mathbf{T} = \mathbf{T}_n$ may be replaced by estimators $\widehat{\mathbf{S}} = \widehat{\mathbf{S}}_p$ and $\widehat{\mathbf{T}} = \widehat{\mathbf{T}}_n$ consistent in operator norm; that is, almost surely

$$\lim_{p \to \infty} \|\mathbf{S}_p - \widehat{\mathbf{S}}_p\|_{\mathrm{op}} = \lim_{n \to \infty} \|\mathbf{T}_n - \widehat{\mathbf{T}}_n\|_{\mathrm{op}} = 0. \tag{28}$$

Remark 18. The matrices $\mathbf{S}$ and $\mathbf{T}$ may be replaced by, respectively, $\theta \mathbf{S}$ and $\mathbf{T}/\theta$ for any $\theta > 0$. Without loss of generality, we may therefore assume that $\mathrm{tr}(\mathbf{T})/n = 1$.

The next result describes a simple method for estimating $\mathbf{S}$ and $\mathbf{T}$ consistently in operator norm when both are diagonal and the singular vectors of $\mathbf{X}$ are delocalized.

Proposition 6.2. *Suppose* $\max_{1 \le k \le r} \|\mathbf{u}_k\|_\infty \|\mathbf{v}_k\|_\infty = o(n^{-1/2})$, $\mathbf{S} = \mathrm{diag}(a_1, \ldots, a_p)$ *and* $\mathbf{T} = \mathrm{diag}(b_1, \ldots, b_n)$. *For* $1 \le i \le p$ *and* $1 \le j \le n$, *define the estimators*

$$\hat{a}_i = \sum_{j=1}^n Y_{ij}^2, \quad \hat{b}_j = \frac{\sum_{i=1}^p Y_{ij}^2}{\frac{1}{n} \sum_{i=1}^p \hat{a}_i}, \tag{29}$$

and and let $\widehat{\mathbf{S}} = \widehat{\mathbf{S}}_p = \mathrm{diag}(\hat{a}_1, \ldots, \hat{a}_p)$ *and* $\widehat{\mathbf{T}} = \widehat{\mathbf{T}}_n = \mathrm{diag}(\hat{b}_1, \ldots, \hat{b}_p)$. *Then* $\widehat{\mathbf{S}}$ *and* $\widehat{\mathbf{T}}$ *are consistent estimators of* $\mathbf{S}$ *and* $\mathbf{T}$, *respectively; that is,* (28) *holds almost surely.*

The proof of Proposition 6.2 may be found in Section I.

Remark 19. The values $\hat{a}_i$ in (29) are the sample standard deviations of the rows of $\sqrt{n}\mathbf{Y}$. Normalizing $\mathbf{Y}$ by $\widehat{\mathbf{S}}^{1/2}$ is then an instance of *standardization* of the rows, a commonly used method in principal component analysis [30].

Remark 20. The estimates $\hat{a}_i$ and $\hat{b}_j$ capture the variation of both the noise and the signal. Proposition 6.2 states that if the signal is delocalized, then in the large p, large n limit its contribution becomes negligible. However, for finite p and n, $\widehat{\mathbf{S}}$ and $\widehat{\mathbf{T}}$ will still see the effects of the signal, and may not be good estimates of $\mathbf{S}$ and $\mathbf{T}$. The experiment in Section 7.3 compares the use of the true $\mathbf{S}$ and $\mathbf{T}$ to their estimates.

6.2.2 Whitening increases the SNR for generic signal matrices

We show that the whitening transformation increases a natural signal-to-noise ratio of the observed matrix. We will assume throughout this section that the $\mathbf{u}_k$ (respectively, $\mathbf{v}_k$) are generic with respect to $\mathbf{S}$ (respectively, $\mathbf{T}$), and that they satisfy the pairwise orthogonality condition with respect to $\mathbf{S}$ (respectively $\mathbf{T}$). Writing the SVD of $\mathbf{X}$ as $\mathbf{X} = \sum_{k=1}^{r} t_k \mathbf{u}_k \mathbf{v}_k^T$, we define the signal-to-noise ratio (SNR) for component k of $\mathbf{X}$:

$$\mathrm{SNR}_k = \frac{t_k^2}{\|\mathbf{N}\|_{\mathrm{op}}^2}, \tag{30}$$

which is the ratio of the squared operator norm of the component $t_k \mathbf{u}_k \mathbf{v}_k^T$ of $\mathbf{X}$ and the squared operator norm of the noise.

Whitening turns $\mathbf{Y}$ into $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{X}} + \mathbf{G}$, with $\widetilde{\mathbf{X}} = \sum_{k=1}^{r} t_k (\mathbf{S}^{-1/2}\mathbf{u}_k)(\mathbf{T}^{-1/2}\mathbf{v}_k)^T = \sum_{k=1}^{r} \tilde{t}_k \tilde{\mathbf{u}}_k \tilde{\mathbf{v}}_k^T$, where $\tilde{t}_k = t_k\|\mathbf{S}^{-1/2}\mathbf{u}_k\|\|\mathbf{T}^{-1/2}\mathbf{v}_k\|$, $\tilde{\mathbf{u}}_k = \mathbf{S}^{-1/2}\mathbf{u}_k/\|\mathbf{S}^{-1/2}\mathbf{u}_k\|$, and $\tilde{\mathbf{v}}_k = \mathbf{T}^{-1/2}\mathbf{v}_k/\|\mathbf{T}^{-1/2}\mathbf{v}_k\|$. The SNR after whitening is then:

$$\widetilde{\mathrm{SNR}}_k = \frac{\tilde{t}_k^2}{\|\mathbf{G}\|_{\mathrm{op}}^2}, \tag{31}$$

Define

$$\tau = \left(\frac{1}{p}\mathrm{tr}(\mathbf{S})\right)\left(\frac{1}{p}\mathrm{tr}(\mathbf{S}^{-1})\right)\left(\frac{1}{n}\mathrm{tr}(\mathbf{T})\right)\left(\frac{1}{n}\mathrm{tr}(\mathbf{T}^{-1})\right). \tag{32}$$

Note that from Jensen's inequality, $\tau \geq 1$, and $\tau > 1$ if either $\mathbf{S}$ or $\mathbf{T}$ is not a multiple of the identity. The following result extends an analogous finding from [40]:

Proposition 6.3. *Suppose $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are generic and weighted orthogonal with respect to $\mathbf{S}$, and $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are generic and weighted orthogonal with respect to $\mathbf{T}$. Then $\widetilde{\mathrm{SNR}}_k \geq \tau\mathrm{SNR}_k$, $1 \leq k \leq r$, almost surely as $p, n \to \infty$. In particular, $\widetilde{\mathrm{SNR}}_k$ is larger than SNR_k if either $\mathbf{S}$ or $\mathbf{T}$ is not a multiple of the identity.*

In other words, the SNR increases after whitening the noise by at least a factor of τ; energy is transferred from the noise component to the signal component. The proof of Proposition 6.3, which extends an analogous result in [40], is in Section J.

6.3 Matrices with missing/unobserved values

We consider the setting where $\mathbf{X}$ is a low-rank target matrix we wish to recover and $\mathbf{G}$ is a matrix of iid Gaussian $N(0,1)$ entries, but rather than observe $\mathbf{X} + \mathbf{G}$, we observe only some subset of the entries. The problem of estimating a matrix from a subset of its entries is known as *matrix completion* [48, 32, 25, 13, 14, 12, 13, 33, 31, 11, 16, 34, 44, 53].

In this section, we will adopt a heterogeneous, rank 1 sampling model, as in [14]. We suppose that the rows and columns are sampled independently, with row i sampled with probability q_i^r, and column j sampled with probability q_j^c; that is, entry (i,j) of $\mathbf{X} + \mathbf{G}$ is sampled with probability $q_i^r q_j^c$. We observe the vector $\mathbf{y} = \mathcal{F}(\mathbf{X} + \mathbf{G})$ of M sampled entries, where $\mathcal{F} : \mathbb{R}^{p \times n} \to \mathbb{R}^M$ is the subsampling operator.

Following the approach from [16], we consider the backprojected matrix $\mathbf{Y} = \mathcal{F}^*(\mathbf{y}) \in \mathbb{R}^{p \times n}$, in which the unobserved entries are replaced by 0's. We write $\mathbf{Y} = \mathcal{F}^*(\mathcal{F}(\mathbf{X})) + \mathcal{F}^*(\mathcal{F}(\mathbf{G}))$. We show that asymptotically, $\mathcal{F}^*(\mathcal{F}(\mathbf{X}))$ behaves like the matrix $\mathbf{PXQ}$. More precisely, we have the following result:

Proposition 6.4. *Suppose* $\max_{1 \le k \le r} \|\mathbf{u}_k\|_\infty \|\mathbf{v}_k\|_\infty = o(n^{-1/2})$. *Then in the limit* $p/n \to \gamma$, $\|\mathcal{F}^*(\mathcal{F}(\mathbf{X})) - \mathbf{PXQ}\|_{\mathrm{op}} \to 0$ *almost surely.*

The proof of Proposition 6.4 may be found in Section K. It is a straightforward generalization of the analogous one-sided result in [16].

Let $\mathbf{N} = \mathcal{F}^*(\mathcal{F}(\mathbf{G}))$. Writing $\mathbf{N} = (N_{ij})$, we have $N_{ij} = \delta_{ij}G_{ij}$, where δ_{ij} is 1 if entry (i,j) is sampled, and 0 otherwise. Then N_{ij} has variance $q_i^r q_j^c$. Consequently, we can whiten the noise by applying $\mathbf{P}^{-1/2}$ and $\mathbf{Q}^{-1/2}$; Proposition 6.3 suggests this will improve estimation of the matrix. To that end, we define $\widetilde{\mathbf{Y}} = \mathbf{P}^{-1/2}\mathbf{Y}\mathbf{Q}^{-1/2} = \widetilde{\mathbf{X}} + \widetilde{\mathbf{G}}$, where $\widetilde{\mathbf{X}} = \mathbf{P}^{-1/2}\mathcal{F}^*(\mathcal{F}(\mathbf{X}))\mathbf{Q}^{-1/2}$, and $\widetilde{\mathbf{G}} = \mathbf{P}^{-1/2}\mathbf{N}\mathbf{Q}^{-1/2}$. Then $\widetilde{\mathbf{G}}$ is a random matrix where each entry has mean zero and variance 1.

From Proposition 6.4, asymptotically the matrix $\widetilde{\mathbf{X}}$ behaves like $\mathbf{P}^{1/2}\mathbf{X}\mathbf{Q}^{1/2}$, and so denoising $\widetilde{\mathbf{Y}}$ estimates $\psi(\widetilde{\mathbf{Y}})$ of $\mathbf{P}^{1/2}\mathbf{X}\mathbf{Q}^{1/2}$. To estimate $\mathbf{X}$ we should perform denoising to $\widetilde{\mathbf{Y}}$ with respect to the weighted loss function $\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X}) = \|\mathbf{P}^{-1/2}(\widehat{\mathbf{X}} - \mathbf{X})\mathbf{Q}^{-1/2}\|_{\mathrm{F}}^2$, with weight matrices $\mathbf{P}^{-1/2}$ and $\mathbf{Q}^{-1/2}$. We then apply $\mathbf{P}^{-1/2}$ and $\mathbf{Q}^{-1/2}$ to the resulting matrix, to obtain an estimator of $\mathbf{X}$ itself. The method is summarized in Algorithm 5, where ψ is the optimal spectral denoiser.

Algorithm 5 Matrix denoising with missing data

1: Input: Samples $\mathbf{y}$; sampling operator $\mathcal{F}$; sampling matrices $\mathbf{P}$, $\mathbf{Q}$
2: Normalize, backproject observations $\widetilde{\mathbf{Y}} = \mathbf{P}^{-1/2}\mathcal{F}^*(\mathbf{y})\mathbf{Q}^{-1/2}$
3: $\psi(\widetilde{\mathbf{Y}})$ is output of Algorithm 1 with matrix $\widetilde{\mathbf{Y}}$, weights $\mathbf{P}^{-1/2}$ and $\mathbf{Q}^{-1/2}$
   AMSE is estimated mean squared error
4: $\widehat{\mathbf{X}} = \mathbf{P}^{-1/2}\psi(\widetilde{\mathbf{Y}})\mathbf{Q}^{-1/2}$
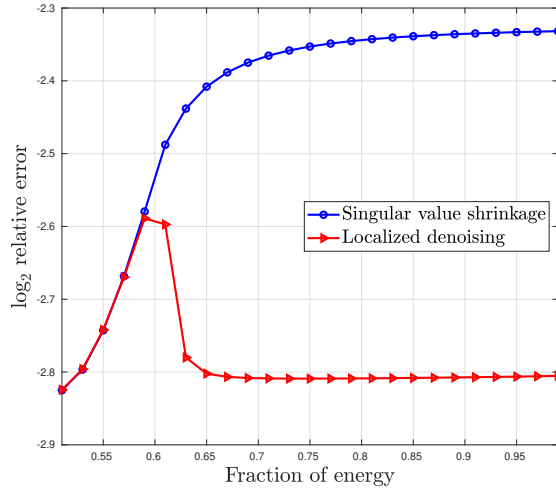

Figure 3: Localized denoising versus singular value shrinkage for the checkerboard matrix, shown in Figure 4; see Section 7.1 for simulation details. The x-axis is parametrized by the fraction of signal energy contained in the light squares. Localized denoising outperforms shrinkage when the fraction is large enough that the rank 2 block structure is detectable.

7 Numerical results

In this section, we report on numerical simulations demonstrating the performance of the algorithms from this paper.
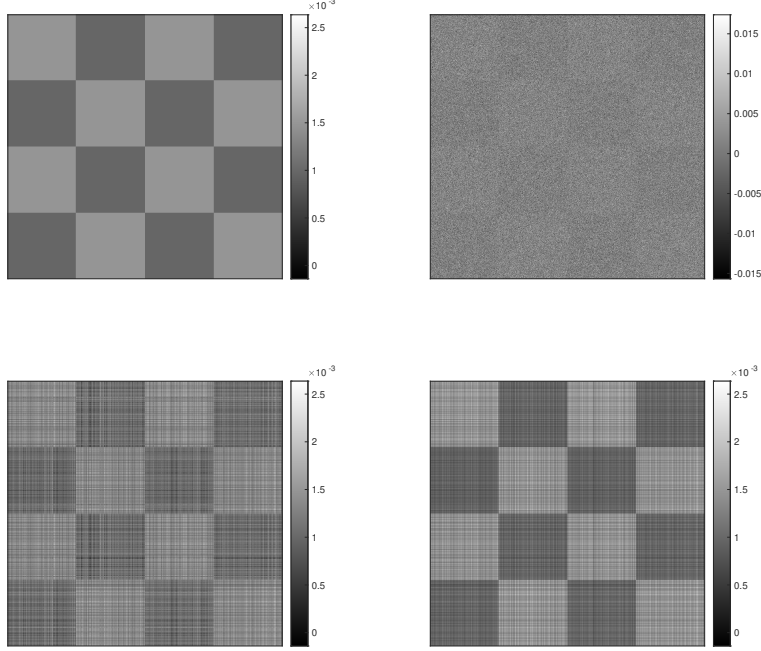
Figure 4: Localized denoising versus singular value shrinkage; see Section 7.1 for simulation details. Upper left: the rank 2 signal matrix; the fraction of signal energy in the light squares is $f = 0.7$. Upper right: the observed noisy matrix. Lower left: the matrix denoised by optimal singular value shrinkage [20, 51]. Lower right: the matrix denoised by localized denoising (Algorithm 2). The relative error of singular value shrinkage is approximately 1.92×10^{-1}, whereas the relative error of localized denoising is approximately 1.40×10^{-1}.

7.1 Localized denoising

We evaluate the performance of localized denoising (Algorithm 2). We generate a "checkerboard" signal matrix $\mathbf{X}$ of size p-by-n, $p = n = 800$, shown in the top left panel of Figure 4. Each light square has the same value, as does each dark square. For a specified number $f \in [1/2, 1]$, the total energy of the light squares is $f \times 100\%$ of the total energy of $\mathbf{X}$. The Frobenius norm of $\mathbf{X}$ is normalized to be 1. Whenever $f > 1/2$, $\mathbf{X}$ is rank 2; when $f = 1/2$, $\mathbf{X}$ has constant value and is rank 1. We add a matrix $\mathbf{G}$ of Gaussian noise, whose entries have standard deviation $1/(10\sqrt{n})$.

We estimate $\mathbf{X}$ from $\mathbf{Y}$ using two methods: singular value shrinkage [20, 51] and localized denoising. Localized denoising is applied with row projection matrices Ω_i, $i = 1, 2, 3, 4$, that project onto equispaced blocks of rows, and column projection matrices Π_i, $i = 1, 2, 3, 4$, that project onto equispaced blocks of columns. For each f, the experiment is repeated 50 times; the $\log_2$ mean errors are plotted in Figure 3.

As f increases, localized denoising outperforms singular value shrinkage more dramatically. This is because localized denoising uses a priori knowledge of $\mathbf{X}$'s block structure, which becomes more pronounced as f grows. Figure 4 shows an example of images of the true matrix $\mathbf{X}$, the noisy matrix $\mathbf{Y}$, and the two denoised matrices, when $f = 0.7$. In this example, the relative error $\|\widehat{\mathbf{X}}^{\mathrm{loc}} - \mathbf{X}\|_{\mathrm{F}}/\|\mathbf{X}\|_{\mathrm{F}}$ of localized denoising is approximately 1.40×10^{-1}, whereas the shrinkage error is 1.92×10^{-1}.

Remark 21. The error curves in Figure 3 both appear nearly identical when $f \lesssim 0.6$, after which localized denoising begins to outperform singular value shrinkage. This is because for small values of f the smallest singular value of $\mathbf{X}$ is not detectable, and so both methods treat the matrix as a constant, rank 1 matrix. Though not apparent from the plot, when $f \leq 0.55$ the performance of singular value shrinkage is slightly better than localized denoising, due to finite sample fluctuations (see Remark 14). For example, when $f = 0.51$, the mean relative error of localized denoising is approximately 1.4118×10^{-1}, while that of shrinkage is approximately 1.4112×10^{-1}.
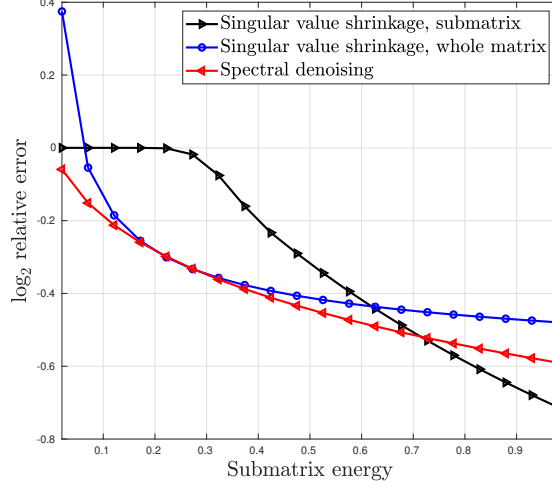
14

Figure 5: Submatrix denoising; see Section 7.2 for simulation details. The left and right singular vectors of $\mathbf{X}$ each contain $\sqrt{f} \times 100\%$ of their energy in the submatrix coordinates, and the x-axis is parametrized by the fraction of energy $f = \|\mathbf{X}_0\|_{\mathrm{F}}^2 / \|\mathbf{X}\|_{\mathrm{F}}^2$ in the submatrix $\mathbf{X}_0$. The $\log_2$ relative error of singular value shrinkage on the submatrix plateaus to 0 when f is small, since the signal in the submatrix alone is too weak to be detected. Singular value shrinkage on the submatrix outperforms spectral denoising (Algorithm 3) when f is very large, but is otherwise inferior; this is the behavior expected from Proposition 6.1.

7.2 Submatrix denoising

We evaluate the performance of spectral denoising for estimating a submatrix $\mathbf{X}_0$ contained within a larger matrix $\mathbf{X}$ (Algorithm 3). We generate a rank 1 signal matrix $\mathbf{X}$ of size p-by-n, $p = 500$, $n = 1000$, with singular values $\gamma^{1/4} + 1/2$, where $\gamma = 1/4$. For a specified $f \in (0, 1)$, the left singular vector $\mathbf{u} = \mathbf{u}_1$ of $\mathbf{X}$ is piecewise constant on the two sets of coordinates $\{1, \ldots, p/2\}$ and $\{p/2 + 1, \ldots, p\}$; the values are such that the energy of $\mathbf{u}$ on coordinates $\{1, \ldots, p/2\}$ is equal to $\sqrt{f}$. Similarly, the right singular vector $\mathbf{v} = \mathbf{v}_1$ of $\mathbf{X}$ is piecewise constant on the two sets of coordinates $\{1, \ldots, n/2\}$ and $\{n/2 + 1, \ldots, n\}$, with values such that the energy of $\mathbf{v}$ on $\{1, \ldots, n/2\}$ is also equal to $\sqrt{f}$. Denoting by $\mathbf{X}_0$ the $p/2$-by-$n/2$ upper-left submatrix of $\mathbf{X}$, $f = \|\mathbf{X}_0\|_{\mathrm{F}}^2 / \|\mathbf{X}\|_{\mathrm{F}}^2$.

The noise matrix has Gaussian entries with variance $1/n$. We denoise the submatrix $\mathbf{X}_0$ using Algorithm 3; optimal singular value shrinkage [20, 51] on $\mathbf{X}_0$ alone ("submatrix shrinkage"); and optimal singular value shrinkage on $\mathbf{X}$ followed by projection onto the rows and columns of $\mathbf{X}_0$ ("global shrinkage"). For each f, the experiment is repeated for 50 draws. Figure 5 plots the $\log_2$ mean relative errors.

Optimal spectral denoising outperforms global shrinkage for all f, since singular value shrinkage is an instance of spectral denoising and hence will not do better than the optimal spectral denoiser. Optimal spectral denoising and global shrinkage perform nearly identically when $f \approx 1/4$, since in this regime the singular vectors of $\mathbf{X}$ are constant, and hence generic with respect to the weight matrices.

For small f, the relative error of global shrinkage exceeds 1, since the submatrix's energy is very small compared to the rest of the matrix. By contrast, optimal spectral denoising with weights Ω and Π highlights the rows and columns in $\mathbf{X}_0$.

Optimal spectral denoising outperforms submatrix shrinkage except when f is close to 1. This is consistent with Proposition 6.1, which states that unless the energy of $\mathbf{X}$'s singular vectors are highly concentrated in the rows and columns of $\mathbf{X}_0$, optimal spectral denoising will outperform singular value shrinkage on the submatrix.

Finally, optimal singular value shrinkage on the submatrix has relative error 1 when f is small. This is because singular value shrinkage on the submatrix only computes the SVD of $\mathbf{Y}_0$, not $\mathbf{Y}$; when the energy in the submatrix $\mathbf{X}_0$ is too weak (i.e. f is too small), no signal will be detected in the submatrix $\mathbf{Y}_0$ alone. By contrast, the singular values of the full matrix $\mathbf{Y}$ always reveal the presence of signal.
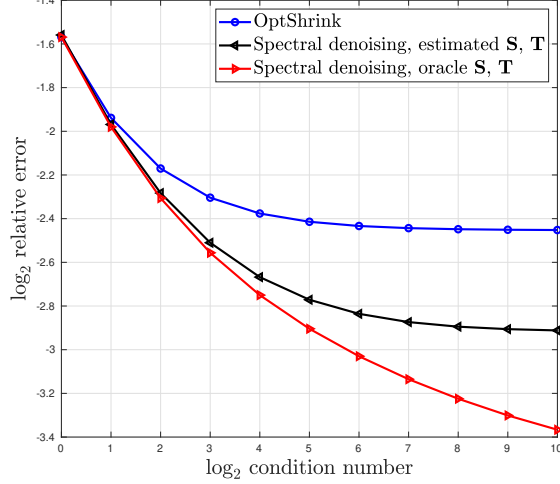
15

Figure 6: Denoising a matrix with doubly-heteroscedastic noise; see Section 7.3 for simulation details. The x-axis is parameterized by $\log_2$ of the condition number κ of $\mathbf{S}$ and $\mathbf{T}$. Proposition 6.3 suggests that noise whitening will enhance performance, and increasingly so as the condition number κ grows, as is the case comparing spectral denoising with oracle whitening (Algorithm 4) and OptShrink [43]. Interestingly, this appears to still hold even when $\mathbf{S}$ and $\mathbf{T}$ are estimated using the procedure from Proposition 6.2.

7.3 Doubly-heteroscedastic noise

We examine the performance of Algorithm 4. We generate a p-by-n signal matrix $\mathbf{X}$, $p = 1000$, $n = 2000$, of rank $r = 5$, with singular values $t^* + 1/2 + k$, $k = 0, 1, 2, 3, 4$, where t^* is the smallest singular detectable value of $\mathbf{X}$, evaluated using the method in [39]. Both the left and right singular vectors of $\mathbf{X}$ are random orthonormal vectors in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively.

For specified $\kappa \geq 1$, we generate row and column diagonal covariance matrices $\mathbf{S}$ and $\mathbf{T}$, each with eigenvalues linearly spaced between $1/\kappa$ and 1. The noise matrix is $\mathbf{S}^{1/2}\mathbf{G}\mathbf{T}^{1/2}$, where $\mathbf{G}$ has iid Gaussian entries with variance $1/n$. We apply three denoising schemes: Algorithm 4 with the true $\mathbf{S}$ and $\mathbf{T}$; Algorithm 4 with $\mathbf{S}$ and $\mathbf{T}$ estimated using the method described in Proposition 6.2; and OptShrink [43]. The experiment is repeated 50 times for each value of κ.

Figure 6 shows the $\log_2$ mean relative errors of each method as a function of $\log_2(\kappa)$. For this model of $\mathbf{S}$ and $\mathbf{T}$, the condition number κ is an increasing function of the parameter τ from Section 6.2.2. Consequently, Proposition 6.3 suggests that whitening will improve the matrix SNR, and that the improvement should increase as κ grows. This is precisely what Figure 6 demonstrates; optimal spectral denoising with whitening by $\mathbf{S}$ and $\mathbf{T}$ does indeed outperform OptShrink, and the performance gap grows with κ. Using the estimated covariances, the performance is degraded but still outperforms OptShrink when κ is large.

7.4 Missing data

We test spectral denoising for missing data (Algorithm 5) by comparing it to nuclear-norm regularized least-squares [11], which estimates $\mathbf{X}$ by:

$$\widehat{\mathbf{X}}^{\mathrm{nuc}} = \underset{\widehat{\mathbf{X}} \in \mathbb{R}^{p \times n}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathcal{F}(\widehat{\mathbf{X}}) - \mathbf{y}\|^2 + \theta \|\mathbf{P}^{1/2}\widehat{\mathbf{X}}\mathbf{Q}^{1/2}\|_* \right\}. \tag{33}$$

Here, $\|\cdot\|_*$ denotes the nuclear norm; $\mathcal{F} : \mathbb{R}^{p \times n} \to \mathbb{R}^M$ is the projection operator onto the M observed samples; and $\mathbf{P}$ and $\mathbf{Q}$ are the diagonal matrices of sampling probabilities for rows and columns, respectively. We weight the nuclear norm by the square root of the sampling probabilities, as suggested in [14]. Following [11], we choose the parameter θ so that when $\mathbf{y}$ is pure noise, $\widehat{\mathbf{X}}^{\mathrm{nuc}}$ is set to zero. It follows from the KKT conditions [10] that this is equivalent to $\theta = \|\mathbf{P}^{-1/2}\mathcal{F}^*(\mathbf{y})\mathbf{Q}^{-1/2}\|_*$, which is approximated by $1 + \sqrt{\gamma}$. We solve (33) using the algorithm in [26].
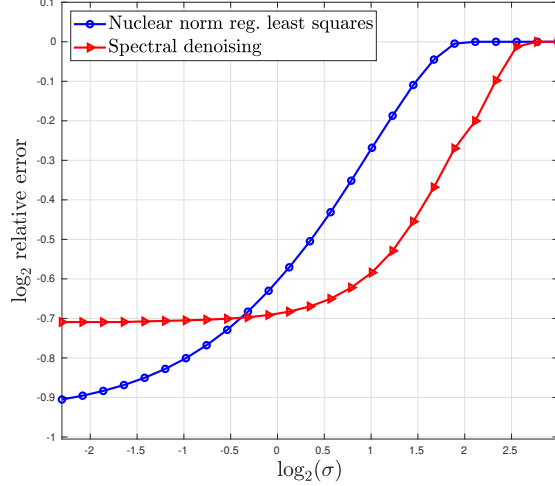
16

Figure 7: Denoising with missing data; see Section 7.4 for simulation details. The x-axis is parameterized by the $\log_2$ noise level. The $\log_2$ relative errors of both methods plateau to 0 at large σ, because the signal is undetectable in this regime. Spectral denoising for missing data (Algorithm 5) does better when σ is large, but underperforms when σ is small.

We generate a rank $r = 5$ signal matrix $\mathbf{X}$ of size p-by-n, $p = 200$, $n = 400$, with singular values $\sqrt{\sqrt{\gamma} + 200k}$, $k = 1, \ldots, 5$, $\gamma = 1/2$. Both the left and right singular vectors of $\mathbf{X}$ are uniformly random. We add to $\mathbf{X}$ a Gaussian noise matrix $\mathbf{G}$, where each entry has variance σ^2/n for a specified value of σ. $\mathbf{X} + \mathbf{G}$ is then subsampled with row and column sampling probabilities each equispaced between 0.3 and 0.7. For each value of σ, the experiment is repeated 50 times. Figure 7 displays the $\log_2$ mean relative errors. When σ is large, spectral denoising is superior, whereas in the small σ regime nuclear-norm regularized least-squares is better.

7.5   Non-Gaussian noise

The optimal spectral denoiser requires estimation of the weighted inner product matrices $\mathbf{D}$, $\widetilde{\mathbf{D}}$, $\mathbf{C}^\omega$ and $\widetilde{\mathbf{C}}^\omega$. The formulas for the entries of these matrices provided by Theorem 3.2 assumes that the noise matrix $\mathbf{G}$ is Gaussian. However, it is of interest whether the same formulas may be applied to non-Gaussian noise. To partially address this question, we compare the finite sample accuracy of the formulas in Theorem 3.2 for different noise distributions.

For each n, we generate $\mathbf{Y} = \mathbf{X} + \mathbf{G}$ of size p-by-n, where $p = 2n$. The signal has rank $r = 2$, with singular values $\gamma^{1/4} + 2$ and $\gamma^{1/4} + 3$; $\mathbf{u}_1$ is uniformly equal to $1/\sqrt{p}$, and $\mathbf{u}_2$ is $1/\sqrt{p}$ on entries $1, \ldots, p/2$, and $-1/\sqrt{p}$ on entries $p/2 + 1, \ldots, p$. $\mathbf{v}_1$ and $\mathbf{v}_2$ are generated similarly, with n in place of p. The noise matrix has iid entries of variance $1/n$, drawn from a specified distribution: Gaussian, Rademacher, t_{10} or t_3, where the t distributions are normalized to have variance $1/n$.

The p-by-p weight matrix Ω is diagonal with diagonal entries $1, \ldots, 3p/4$ equal to 1, and the remaining entries 0. We evaluate the true matrix $\mathbf{E}$ and use formulas (16) and (18) to predict $\mathbf{D}$ and $\mathbf{C}^\omega$. For each draw, we compute the actual inner product matrices $\widehat{\mathbf{D}}$ and $\widehat{\mathbf{C}}^\omega$ using the left singular vectors $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$ of $\mathbf{Y}$. Due to the ambiguity in signs, we make all entries of the matrices positive. We then compute the relative errors $\|\widehat{\mathbf{D}} - \mathbf{D}\|_{\mathrm{F}}/\|\mathbf{D}\|_{\mathrm{F}}$ and $\|\widehat{\mathbf{C}}^\omega - \mathbf{C}^\omega\|_{\mathrm{F}}/\|\mathbf{C}^\omega\|_{\mathrm{F}}$.

For each noise type and each value of $n = 500k$, $k = 1, 2, 4, 8, 16$, the experiment is repeated 1000 times. The average and maximum relative errors are recorded in Table 2 for $\mathbf{C}^\omega$, and in Table 3 for $\mathbf{D}$. Both the average and maximum errors for Gaussian noise very nearly match those for the Rademacher and t_{10} distributions. However, the errors for the heavier tailed t_3 distribution are much larger, indicating that the theory breaks down for this noise model. The errors for the Gaussian, Rademacher, and t_{10} distributions appear to decay approximately like $O(n^{-1/2})$; this is the rate we expect from [6] and Theorem 2.19 in [8]. The errors for the t_3 distribution do not exhibit such decay with increasing n, indicating that the model

17

	Mean relative error, $\mathbf{C}^\omega$			
n	Gaussian	Rademacher	t, df=10	t, df=3
500	2.956e-02	3.057e-02	2.928e-02	3.284e-01
1000	2.050e-02	2.127e-02	2.042e-02	4.208e-01
2000	1.459e-02	1.514e-02	1.466e-02	5.280e-01
4000	1.033e-02	1.019e-02	1.030e-02	6.535e-01
8000	7.459e-03	7.693e-03	7.495e-03	7.640e-01

	Max relative error, $\mathbf{C}^\omega$			
n	Gaussian	Rademacher	t, df=10	t, df=3
500	1.117e-01	1.126e-01	1.241e-01	9.702e-01
1000	7.895e-02	8.411e-02	7.878e-02	9.837e-01
2000	6.520e-02	5.106e-02	6.907e-02	9.953e-01
4000	4.490e-02	4.056e-02	4.638e-02	9.950e-01
8000	3.035e-02	3.172e-02	3.174e-02	9.951e-01

Table 2: Average and maximum relative errors $\|\widehat{\mathbf{C}}^\omega - \mathbf{C}^\omega\|_{\mathrm{F}}/\|\mathbf{C}^\omega\|_{\mathrm{F}}$; see Section 7.5 for simulation details. For Gaussian, Rademacher, and t_{10} distributions, the average errors decay approximately like $O(n^{-1/2})$. The errors for the t_3 distribution do not decay, indicating poor model fit.

	Mean relative error, $\mathbf{D}$			
n	Gaussian	Rademacher	t, df=10	t, df=3
500	1.991e-02	2.039e-02	1.975e-02	1.757e-01
1000	1.414e-02	1.437e-02	1.412e-02	2.219e-01
2000	9.917e-03	1.015e-02	1.001e-02	2.788e-01
4000	7.027e-03	6.875e-03	7.005e-03	3.455e-01
8000	5.040e-03	5.224e-03	5.087e-03	4.141e-01

	Max relative error, $\mathbf{D}$			
n	Gaussian	Rademacher	t, df=10	t, df=3
500	6.954e-02	7.491e-02	7.354e-02	9.663e-01
1000	5.044e-02	4.533e-02	5.505e-02	9.238e-01
2000	3.772e-02	3.578e-02	4.101e-02	9.529e-01
4000	2.468e-02	2.564e-02	2.712e-02	9.741e-01
8000	1.584e-02	1.836e-02	1.668e-02	9.779e-01

Table 3: Average and maximum relative errors $\|\widehat{\mathbf{D}} - \mathbf{D}\|_{\mathrm{F}}/\|\mathbf{D}\|_{\mathrm{F}}$; see Section 7.5 for simulation details. For Gaussian, Rademacher, and t_{10} distributions, the average errors decay approximately like $O(n^{-1/2})$. The errors for the t_3 distribution do not decay, indicating poor model fit.

| | Mean relative error | | |
Noise type	Oracle	K-N	Naive
Gaussian	4.010e-01	4.010e-01	4.012e-01
Rademacher	4.012e-01	4.012e-01	4.013e-01
t, df=10	4.022e-01	4.022e-01	4.024e-01
t, df=5	4.047e-01	4.086e-01	4.099e-01
t, df=4	4.337e-01	4.484e-01	4.539e-01
t, df=3	6.606e-01	7.459e-01	7.611e-01
t, df=2.5	1.026e+00	1.287e+00	1.300e+00

| | Max relative error | | |
Noise type	Oracle	K-N	Naive
Gaussian	4.420e-01	4.420e-01	4.421e-01
Rademacher	4.459e-01	4.459e-01	4.459e-01
t, df=10	4.349e-01	4.349e-01	4.349e-01
t, df=5	1.280e+00	1.262e+00	1.262e+00
t, df=4	2.149e+00	2.113e+00	2.113e+00
t, df=3	4.922e+00	4.940e+00	4.943e+00
t, df=2.5	6.693e+00	6.703e+00	6.706e+00

Table 4: Average and maximum relative errors of estimation; see Section 7.6 for simulation details. The naive rank estimate $\hat{r}^{\mathrm{naive}}$ from (34) tends to overestimate the true rank $r = 2$, whereas the estimate $\hat{r}^{\mathrm{KN}}$ of Kritchman and Nadler [36] is more accurate. However, the difference between the errors in the resulting estimates of $\mathbf{X}$ is not large. Both methods perform poorly for heavy tailed distributions.

| | Mean rank | | |
Noise type	Oracle	K-N	Naive
Gaussian	2.000e+00	2.000e+00	2.084e+00
Rademacher	2.000e+00	2.000e+00	2.037e+00
t, df=10	2.000e+00	2.000e+00	2.094e+00
t, df=5	2.000e+00	2.128e+00	2.443e+00
t, df=4	2.000e+00	2.858e+00	3.577e+00
t, df=3	2.000e+00	7.875e+00	8.967e+00
t, df=2.5	2.000e+00	1.643e+01	1.716e+01

| | Max rank | | |
Noise type	Oracle	K-N	Naive
Gaussian	2	2	3
Rademacher	2	2	3
t, df=10	2	2	3
t, df=5	2	4	5
t, df=4	2	7	8
t, df=3	2	16	17
t, df=2.5	2	26	26

Table 5: Average and maximum rank estimates; see Section 7.6 for simulation details. The naive rank estimate $\hat{r}^{\mathrm{naive}}$ from (34) tends to overestimate the true rank $r = 2$, whereas the estimate $\hat{r}^{\mathrm{KN}}$ of Kritchman and Nadler [36] is more accurate. However, the difference between the errors in the resulting estimates of $\mathbf{X}$ is not large. Both methods perform poorly for heavy tailed distributions.

does not match.

7.6 Rank estimation

In this section, we explore estimation of the rank r of $\mathbf{X}$ from the observed matrix $\mathbf{Y}$, a topic that has received considerable attention [36, 37, 17, 15, 46, 45]. The naive estimator $\hat{r}^{\text{naive}}$ is defined by

$$\hat{r}^{\text{naive}} = \#\{k : \lambda_k > 1 + \sqrt{\gamma}\}; \tag{34}$$

this simply counts the number of $\mathbf{Y}$'s singular values exceeding $1 + \sqrt{\gamma}$, the asymptotically largest singular value of the noise matrix $\mathbf{G}$. It is known that $\hat{r}^{\text{naive}}$ may overestimate the rank; see, e.g., [28]. The rank estimator of Kritchman and Nadler from [36], denoted by $\hat{r}^{\text{KN}}$, is designed to prevent attributing noisy singular values to signal. We compare the performance of $\hat{r}^{\text{naive}}$ and $\hat{r}^{\text{KN}}$ for different noise distributions in terms of the accuracy of estimating r and the effect on the denoising error.

For $p = 300$ and $n = 600$, we generate a p-by-n signal matrix $\mathbf{X}$ with rank $r = 2$ and singular values $\gamma^{1/4} + 1$ and $\gamma^{1/4} + 2$. The left singular vector $\mathbf{u}_1$ is uniformly equal to $1/\sqrt{p}$, and $\mathbf{u}_2$ is $1/\sqrt{p}$ on entries $1, \ldots, p/2$, and $-1/\sqrt{p}$ on entries $p/2 + 1, \ldots, p$. The right singular vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are generated the same way, with n in place of p. The noise matrix $\mathbf{G}$ had iid entries of variance $1/n$, drawn from one a specified distribution, namely: Gaussian, Rademacher, or the t distributions with 10, 5, 4, 3 and 2.5 degrees of freedom, where the t distributions are normalized to have variance $1/n$. The p-by-p weight matrix Ω is diagonal with diagonal entries linearly spaced between $1/p$ and 1. The n-by-n weight matrix Π is also diagonal, with diagonal entries linearly spaced between $1/p$ and $1/\gamma$. In each run, we apply Algorithm 1 with the oracle $r = 2$, the naive $\hat{r}^{\text{naive}}$ from (34), and $\hat{r}^{\text{KN}}$ from [36], with 0.1 confidence level[1]. For each noise distribution, the experiment is repeated 1000 times. In Tables 4 and 5 we record the relative errors $\|\Omega(\widehat{\mathbf{X}} - \mathbf{X})\Pi^T\|_{\text{F}}/\|\Omega\mathbf{X}\Pi^T\|_{\text{F}}$ and the estimated ranks.

For the Gaussian, Rademacher, and t_{10} distributions, the Kritchman-Nadler estimate $\hat{r}^{\text{KN}}$ is typically closer to the true rank, $r = 2$, than is the naive estimate $\hat{r}^{\text{naive}}$. However, the average and maximum errors are close regardless of the rank estimator used, since even when $\hat{r}^{\text{naive}} = 3$, the third singular value of $\mathbf{Y}$ is so close to the detection edge $1 + \sqrt{\gamma}$ that the estimates of the cosines c_3 and $\tilde{c}_3$ are nearly 0. With the t_5 distribution, both $\hat{r}^{\text{naive}}$ and $\hat{r}^{\text{KN}}$ are more likely to overestimate the true rank. While the average errors are close to those for the Gaussian, Rademacher, and t_{10} distributions, the maximum errors are much larger, indicating that while this noise distribution's "typical" behavior may be close to the thinner tailed ones, a small number of extreme draws of $\mathbf{G}$ can result in very poor performance. For the t distributions with 4, 3, and 2.5 degrees of freedom, both $\hat{r}^{\text{naive}}$ and $\hat{r}^{\text{KN}}$ drastically overestimate the rank, and the resulting relative errors are enormous.

8   Conclusion

This paper has introduced a family of spectral denoisers for low-rank matrix estimation, which generalizes singular value shrinkage. We have derived optimal spectral denoisers for weighted loss functions, and discussed applications. By judiciously combining these denoisers for different weights we contructed the method of localized denoising, which outperforms singular value shrinkage under heterogeneity. While this paper has focused on theoretical and algorithmic development, in future work we plan to apply the methods to problems where related but suboptimal methods have previously been employed. This includes the problems of denoising and deconvolution of images from cryoelectron microscopy [9]; 3-D reconstruction of heterogeneous molecules from noisy images [1]; and denoising XFEL images [42, 59].

Acknowledgements

---

[1] The code for computing $\hat{r}^{\text{KN}}$ was taken from Boaz Nadler's website: `www.wisdom.weizmann.ac.il/~nadler/Rank_Estimation/rank_estimation.html`

References

[1] Joakim Andén and Amit Singer. Structural variability from noisy tomographic projections. *SIAM Journal on Imaging Sciences*, 11(2):1441–1492, 2018.

[2] Zhidong Bai and Jack W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, 2009.

[3] Zhidong Bai and Jian-feng Yao. Central limit theorems for eigenvalues in a spiked population model. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 44(3):447–474, 2008.

[4] Zhidong Bai and Jian-feng Yao. On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167–177, 2012.

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

[6] Zhigang Bao, Xiucai Ding, and Ke Wang. Singular vector and singular subspace distribution for the matrix denoising model. *arXiv preprint arXiv:1809.10476*, 2018.

[7] Florent Benaych-Georges, Alice Guionnet, and Myléne Maida. Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electronic Journal of Probability*, 16:1621–1662, 2011.

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

[9] Tejal Bhamre, Teng Zhang, and Amit Singer. Denoising and covariance estimation of single particle cryo-EM images. *Journal of Structural Biology*, 195(1):72–81, 2016.

[10] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

[12] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.

[13] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[14] Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Completing any low-rank matrix, provably. *The Journal of Machine Learning Research*, 16(1):2999–2034, 2015.

[15] Edgar Dobriban. Permutation methods for factor analysis and PCA. *Annals of Statistics*, to appear.

[16] Edgar Dobriban, William Leeb, and Amit Singer. Optimal prediction in the linearly transformed spiked model. *Annals of Statistics*, 48(1):491–513, 2020.

[17] Edgar Dobriban and Art B. Owen. Deterministic parallel analysis: an improved method for selecting factors and principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.

[18] Matan Gavish and David L. Donoho. Minimax risk of matrix denoising by singular value thresholding. *The Annals of Statistics*, 42(6):2413–2440, 2014.

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

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

[21] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

[22] David Hong, Laura Balzano, and Jeffrey A. Fessler. Towards a theoretical analysis of PCA for heteroscedastic data. In *54th Annual Allerton Conference on Communication, Control, and Computing*, pages 496–503. IEEE, 2016.

[23] David Hong, Laura Balzano, and Jeffrey A. Fessler. Asymptotic performance of PCA for high-dimensional heteroscedastic data. *Journal of Multivariate Analysis*, 2018.

[24] David Hong, Laura Balzano, and Jeffrey A. Fessler. Optimally Weighted PCA for High-Dimensional Heteroscedastic Data. *arXiv preprint arXiv:1810.12862*, 2018.

[25] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, pages 665–674. ACM, 2013.

[26] Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 457–464. ACM, June 2009.

[27] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1), 2007.

[28] Iain Johnstone. Tail sums of Wishart and Gaussian eigenvalues beyond the bulk edge. *Australian and New Zealand Journal of Statistics*, 60(1):65–74, 2018.

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

[30] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[31] Raghunandan H. Keshavan and Andrea Montanari. Regularization for matrix completion. In *Proceedings of International Symposium on Information Theory*, pages 1503–1507. IEEE, 2010.

[32] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[33] Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[34] Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, pages 2302–2329, 2011.

[35] Sonal Kothari, John H. Phan, Richard A. Moffitt, Todd H. Stokes, Shelby E. Hassberger, Qaiser Chaudry, Andrew N. Young, and May D. Wang. Automatic batch-invariant color segmentation of histological cancer images. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 657–660. IEEE, 2011.

[36] Shira Kritchman and Boaz Nadler. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):19–32, 2008.

[37] Shira Kritchman and Boaz Nadler. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Transactions on Signal Processing*, 57(10):3930–3941, 2009.

[38] William Leeb. Optimal singular value shrinkage for operator norm loss. *arXiv preprint arXiv:2005.11807*, 2020.

[39] William Leeb. Rapid evaluation of the spectral signal detection threshold and Stieltjes transform. *arXiv preprint arXiv:1904.11665*, 2020.

[40] William Leeb and Elad Romanov. Optimal spectral denoising and PCA with heteroscedastic noise. *arXiv preprint arXiv:1811.02201v2*, 2019.

[41] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, Benjamin Langmead David Simcha, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10), 2010.

[42] Lydia T. Liu, Edgar Dobriban, and Amit Singer. *e*PCA: High dimensional exponential family PCA. *The Annals of Applied Statistics*, 12(4):2121–2150, 2018.

[43] Raj Rao Nadakuditi. 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.

[44] Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

[45] Damien Passemier and Jianfeng Yao. On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices: Theory and Applications*, 1(1):1150002, 2012.

[46] Damien Passemier and Jianfeng Yao. Estimation of the number of spikes, possibly equal, in the high-dimensional case. *Journal of Multivariate Analysis*, 127:173–183, 2014.

[47] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.

[48] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, December 2011.

[49] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[50] Andreas Scherer. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*, volume 868. John Wiley & Sons, 2009.

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

[52] Uri Shaham, Kelly P. Stanton, Jun Zhao, Huamin Li, Khadir Raddassi, Ruth Montgomery, and Yuval Kluger. Removal of batch effects using distribution-matching residual networks. *Bioinformatics*, 33(16):2539–2546, 2017.

[53] N. Srebro and R. R. Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems*, pages 2056–2064, 2010.

[54] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Society, 2012.

[55] Po-Yuan Tung, John D. Blischak, Chiaowen Joyce Hsiao, David A. Knowles, Jonathan E. Burnett, Jonathan K. Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports*, 7(39921), 2017.

[56] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[57] Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *The Annals of Probability*, 1(6):1068–1070, 1973.

[58] Anru Zhang, T. Tony Cai, and Yihong Wu. Heteroskedastic PCA: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*, 2018.

[59] Zhizhen Zhao, Lydia T. Liu, and Amit Singer. Steerable *e*PCA: Rotationally invariant exponential family PCA. *IEEE Transactions on Image Processing*, 29(1):6069–6081, 2020.

A Proof of Theorem 3.2

The proof of Theorem 3.2 is similar to the analysis found in [40], in that it rests on the same decomposition of the empirical singular vectors $\hat{\mathbf{u}}_j$ and $\hat{\mathbf{v}}_j$ into the signal and residual components. If $\mathbf{a}$ and $\mathbf{b}$ are vectors of the same dimension, we will write $\mathbf{a} \sim \mathbf{b}$ as a short-hand for $\|\mathbf{a} - \mathbf{b}\| \to 0$ almost surely as $p, n \to \infty$. The statements are symmetric in the left and right singular vectors, so for compactness we will only prove them for the left ones. The proofs for the other side are identical.

Because the noise matrix $\mathbf{G}$ has an isotropic distribution, we can write:

$$\hat{\mathbf{u}}_j \sim c_j \mathbf{u}_j + s_j \tilde{\mathbf{u}}_j, \tag{35}$$

where $\tilde{\mathbf{u}}_j$ is a unit vector that is uniformly random over the sphere in the subspace orthogonal to $\mathbf{u}_1, \ldots, \mathbf{u}_r$ (see [47]). Because $\tilde{\mathbf{u}}_j$ is uniformly random, it is asymptotically orthogonal to any independent unit vector $\mathbf{w}$; that is,

$$\tilde{\mathbf{u}}_j^T \mathbf{w} \sim 0. \tag{36}$$

Furthermore, $\tilde{\mathbf{u}}_j$ satisfies the normalized trace formula, namely if $\mathbf{A}$ is any matrix with bounded operator norm, then

$$\tilde{\mathbf{u}}_j^T \mathbf{A} \tilde{\mathbf{u}}_j \sim \frac{1}{p} \mathrm{tr}(\mathbf{A}). \tag{37}$$

We refer the reader to [7, 21, 57, 49] for details. We will use (36) and (37) repeatedly. Furthermore, when $j \neq k$ it follows from Lemma A.2 in [40] that

$$\tilde{\mathbf{u}}_j^T \mathbf{A} \tilde{\mathbf{u}}_k \sim 0. \tag{38}$$

Applying Ω to each side of (35), we have:

$$\Omega \hat{\mathbf{u}}_j \sim c_j \Omega \mathbf{u}_j + s_j \Omega \tilde{\mathbf{u}}_j. \tag{39}$$

The proofs of the identities in Theorem 3.2 now follow by manipulating the asymptotic equation (39) appropriately, in conjunction with (36), (37) and (38).

We first show the formulas for c_{jk}^ω. We take inner products of each side of (39) with $\Omega \mathbf{u}_k$:

$$c_{jk}^\omega \sim \langle \Omega \hat{\mathbf{u}}_j, \Omega \mathbf{u}_k \rangle \sim c_j \langle \Omega \mathbf{u}_j, \Omega \mathbf{u}_k \rangle + s_j \langle \Omega \tilde{\mathbf{u}}_j, \Omega \mathbf{u}_k \rangle \sim c_j \langle \Omega \mathbf{u}_j, \Omega \mathbf{u}_k \rangle \sim c_j e_{jk}, \tag{40}$$

where we have used (36).

To derive the formula for d_j, we take the squared norm of each side of (39):

$$d_j \sim \|\Omega \hat{\mathbf{u}}_j\|^2 \sim c_j^2 \|\Omega \mathbf{u}_j\|^2 + s_j^2 \|\Omega \tilde{\mathbf{u}}_j\|^2 \sim c_j^2 \alpha_j + s_j^2 \mu. \tag{41}$$

The first asymptotic equivalence follows from (36), and the second from (37).

Finally, we derive the formula for d_{jk}, $j \neq k$. From (39), we have

$$\langle \Omega \hat{\mathbf{u}}_j, \Omega \hat{\mathbf{u}}_k \rangle \sim c_j c_k \langle \Omega \mathbf{u}_j, \Omega \mathbf{u}_k \rangle + s_j s_k \langle \Omega \tilde{\mathbf{u}}_j, \Omega \tilde{\mathbf{u}}_k \rangle + s_j c_k \langle \Omega \tilde{\mathbf{u}}_j, \Omega \mathbf{u}_k \rangle + c_j s_k \langle \Omega \mathbf{u}_j, \Omega \tilde{\mathbf{u}}_k \rangle. \tag{42}$$

From (36) and (38), the terms involving $\tilde{\mathbf{u}}_j$ and $\tilde{\mathbf{u}}_k$ vanish, and we are left with

$$d_{jk} \sim \langle \Omega \hat{\mathbf{u}}_j, \Omega \hat{\mathbf{u}}_k \rangle \sim c_j c_k \langle \Omega \mathbf{u}_j, \Omega \mathbf{u}_k \rangle \sim c_j c_k e_{jk}. \tag{43}$$

This completes the proof of Theorem 3.2.

B   Proof of Theorem 4.1

The target matrix $\mathbf{X}$ may be written

$$\mathbf{X} = \sum_{k=1}^{r} t_k \mathbf{u}_k \mathbf{v}_k^T = \mathbf{U} \text{diag}(\mathbf{t}) \mathbf{V}^T, \tag{44}$$

and our estimate $\widehat{\mathbf{X}}$ is of the form

$$\widehat{\mathbf{X}} = \widehat{\mathbf{U}} \widehat{\mathbf{B}} \widehat{\mathbf{V}}^T, \tag{45}$$

where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$, $\widehat{\mathbf{U}} = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_r]$, $\widehat{\mathbf{V}} = [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_r]$, and $\mathbf{t} = (t_1, \ldots, t_r)^T$.

Define $\mathbf{W} = \Omega\mathbf{U}$, $\mathbf{Z} = \Pi\mathbf{V}$, $\widehat{\mathbf{W}} = \Omega\widehat{\mathbf{U}}$, and $\widehat{\mathbf{Z}} = \Pi\widehat{\mathbf{V}}$. We may then write the weighted loss as follows:

$$\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X}) = \|\Omega(\widehat{\mathbf{X}} - \mathbf{X})\Pi^T\|_{\mathrm{F}}^2 = \|\Omega\widehat{\mathbf{X}}\Pi^T - \Omega\mathbf{X}\Pi^T\|_{\mathrm{F}}^2 = \|\widehat{\mathbf{W}}\widehat{\mathbf{B}}\widehat{\mathbf{Z}}^T - \mathbf{W}\text{diag}(\mathbf{t})\mathbf{Z}^T\|_{\mathrm{F}}^2, \tag{46}$$

which is the *unweighted* Frobenius loss between $\widehat{\mathbf{W}}\widehat{\mathbf{B}}\widehat{\mathbf{Z}}^T$ and $\mathbf{W}\text{diag}(\mathbf{t})\mathbf{Z}^T$. Continuing, we have:

$$\begin{aligned}
\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X}) &= \|\mathbf{W}\text{diag}(\mathbf{t})\mathbf{Z}^T - \widehat{\mathbf{W}}\widehat{\mathbf{B}}\widehat{\mathbf{Z}}^T\|_{\mathrm{F}}^2 \\
&= \|\mathbf{W}\text{diag}(\mathbf{t})\mathbf{Z}^T\|_{\mathrm{F}}^2 + \|\widehat{\mathbf{W}}\widehat{\mathbf{B}}\widehat{\mathbf{Z}}^T\|_{\mathrm{F}}^2 - 2\langle\mathbf{W}\text{diag}(\mathbf{t})\mathbf{Z}^T, \widehat{\mathbf{W}}\widehat{\mathbf{B}}\widehat{\mathbf{Z}}^T\rangle_{\mathrm{F}} \\
&= \langle\mathbf{W}^T\mathbf{W}\text{diag}(\mathbf{t})\mathbf{Z}^T\mathbf{Z}, \text{diag}(\mathbf{t})\rangle_{\mathrm{F}} + \langle\widehat{\mathbf{W}}^T\widehat{\mathbf{W}}\widehat{\mathbf{B}}\widehat{\mathbf{Z}}^T\widehat{\mathbf{Z}}, \widehat{\mathbf{B}}\rangle_{\mathrm{F}} - 2\langle\widehat{\mathbf{W}}^T\mathbf{W}\text{diag}(\mathbf{t})\mathbf{Z}^T\widehat{\mathbf{Z}}, \widehat{\mathbf{B}}\rangle_{\mathrm{F}} \\
&\sim \langle\mathbf{E}\text{diag}(\mathbf{t})\widetilde{\mathbf{E}}, \text{diag}(\mathbf{t})\rangle_{\mathrm{F}} + \langle\mathbf{D}\widehat{\mathbf{B}}\widetilde{\mathbf{D}}, \widehat{\mathbf{B}}\rangle_{\mathrm{F}} - 2\langle\mathbf{C}\text{diag}(\mathbf{t})\widetilde{\mathbf{C}}^T, \widehat{\mathbf{B}}\rangle_{\mathrm{F}}.
\end{aligned} \tag{47}$$

Defining the operator $\mathcal{T}$ by $\mathcal{T}(\widehat{\mathbf{B}}) = \mathbf{D}\widehat{\mathbf{B}}\widetilde{\mathbf{D}}$, the pseudoinverse of $\mathcal{T}$ is given by $\mathcal{T}^+(\mathbf{B}) = \mathbf{D}^+\mathbf{B}\widetilde{\mathbf{D}}^+$. Consequently, the choice of $\widehat{\mathbf{B}}$ that minimizes $\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X})$ is given by:

$$\widehat{\mathbf{B}} = \mathbf{D}^+\mathbf{C}\text{diag}(\mathbf{t})\widetilde{\mathbf{C}}^T\widetilde{\mathbf{D}}^+. \tag{48}$$

The error may then be evaluated by substituting this expression for $\widehat{\mathbf{B}}$ into (47), completing the proof.

C   Proof of Theorem 4.2

Under weighted orthogonality, $e_{jk} = \tilde{e}_{jk} = 0$ whenever $j \neq k$, and so $d_{jk} = \tilde{d}_{jk} = c_{jk}^\omega = \tilde{c}_{jk}^\omega = 0$ when $j \neq k$ as well. Consequently, the matrices $\mathbf{E}$, $\widetilde{\mathbf{E}}$, $\mathbf{D}$, $\widetilde{\mathbf{D}}$, $\mathbf{C}$, and $\widetilde{\mathbf{C}}$ are diagonal. The optimal $\widehat{\mathbf{B}}$ is given by:

$$\widehat{\mathbf{B}} = \mathbf{D}^+\mathbf{C}\text{diag}(\mathbf{t})\widetilde{\mathbf{C}}\widetilde{\mathbf{D}}^+, \tag{49}$$

which is also diagonal, with diagonal entries

$$\hat{t}_k = \frac{t_k c_k^\omega \tilde{c}_k^\omega}{d_k \tilde{d}_k} = \frac{t_k c_k \alpha_k \tilde{c}_k \beta_k}{(c_k^2 \alpha_k + s_k^2 \mu)(\tilde{c}_k^2 \beta_k + \tilde{s}_k^2 \nu)} = t_k c_k \tilde{c}_k \frac{\alpha_k \beta_k}{(c_k^2 \alpha_k + s_k^2 \mu)(\tilde{c}_k^2 \beta_k + \tilde{s}_k^2 \nu)}, \tag{50}$$

which is the desired expression.

D   Proof of Proposition 4.3

Suppose a coordinate has signal strength $t = t_k$ (we drop the subscript as we are only considering one component). We may assume without loss of generality (and by rescaling α and β) that $\mu = \nu = 1$. Consequently, the optimal singular value is equal to:

$$\hat{t} = tc\tilde{c} \cdot \frac{\alpha}{c^2\alpha + s^2} \cdot \frac{\beta}{\tilde{c}^2\beta + \tilde{s}^2}. \tag{51}$$

By taking α and β sufficiently large, this value can be made arbitrarily close to

$$\frac{t}{c\tilde{c}} = \frac{t\sqrt{(1+\gamma/t^2)(1+1/t^2)}}{1-\gamma/t^4} = \frac{\lambda}{1-\gamma/t^4} > \lambda. \tag{52}$$

That is, the optimal singular value $\hat{t}$ will be greater than the observed singular value λ in this parameter regime.

On the other hand, if $\beta \leq 1 = \nu$, we have:

$$\frac{\hat{t}}{\lambda} = \frac{1}{\lambda}tc\tilde{c} \cdot \frac{\alpha}{\alpha c^2 + s^2} \cdot \frac{\beta}{\beta \tilde{c}^2 + \tilde{s}^2} \leq \frac{1}{\lambda}t\frac{\tilde{c}}{c} = \frac{t}{\sqrt{(t^2+1)(t^2+\gamma)}}t\sqrt{\frac{t^2+\gamma}{t^2+1}} = \frac{t^2}{t^2+1} \leq 1, \tag{53}$$

which shows that $\hat{t} \leq \lambda$. A nearly identical proof works if $\alpha \leq \mu$. This completes the proof.

E   Proof of Proposition 4.4

Without loss of generality, we will assume $\mu = \nu = 1$. We consider the functions $c(t) = \sqrt{(1-\gamma/t^4)/(1+\gamma/t^2)}$ and $\tilde{c}(t) = \sqrt{(1-\gamma/t^4)/(1+1/t^2)}$. Define the functions $\varphi(t)$ and $\psi(t)$ by

$$\varphi(t) = \frac{\alpha c(t)}{\alpha c(t)^2 + 1 - c(t)^2} \tag{54}$$

and

$$\psi(t) = \frac{\beta \tilde{c}(t)}{\beta \tilde{c}(t)^2 + 1 - \tilde{c}(t)^2}. \tag{55}$$

Then we may write the optimal singular value $\hat{t}$ as a function $f(t)$ as follows:

$$f(t) = t\varphi(t)\psi(t). \tag{56}$$

Let us assume that $\alpha \leq 1$; the proof for $\beta \leq 1$ will be nearly identical. We wish to show that $f'(t) \geq 0$, for $t > \gamma^{1/4}$. We have

$$\frac{f'(t)}{f(t)} = \frac{\varphi'(t)}{\varphi(t)} + \frac{\psi'(t)}{\psi(t)} + \frac{1}{t}, \tag{57}$$

and since $f(t) > 0$, we must show that the right side is positive. It is straightforward to verify that

$$\varphi'(t) = \frac{\alpha c'(t)[1 - (\alpha-1)c(t)^2]}{[1 + (\alpha-1)c(t)^2]^2} \tag{58}$$

from which it follows that

$$\frac{\varphi'(t)}{\varphi(t)} = \frac{c'(t)}{c(t)}\frac{1 - (\alpha-1)c(t)^2}{1 + (\alpha-1)c(t)^2} \geq \frac{c'(t)}{c(t)}. \tag{59}$$

Similarly, we can show

$$\frac{\psi'(t)}{\psi(t)} = \frac{\tilde{c}'(t)}{\tilde{c}(t)}\frac{1 - (\beta-1)\tilde{c}(t)^2}{1 + (\beta-1)\tilde{c}(t)^2} \geq -\frac{\tilde{c}'(t)}{\tilde{c}(t)}. \tag{60}$$

Consequently, it is enough to show

$$\frac{c'(t)}{c(t)} - \frac{\tilde{c}'(t)}{\tilde{c}(t)} + \frac{1}{t} \geq 0. \tag{61}$$

26

Direction computation shows

$$\frac{c'(t)}{c(t)} = \gamma \frac{t^4 + 2t^2 + \gamma}{t(t^2 + \gamma)(t^4 - \gamma)} \tag{62}$$

and

$$\frac{\tilde{c}'(t)}{\tilde{c}(t)} = \frac{t^4 + 2\gamma t^2 + \gamma}{t(t^2 + 1)(t^4 - \gamma)}. \tag{63}$$

Substituting (62) and (63) into the left side of (61) and multiplying by $t(t^2 + \gamma)(t^2 + 1)$, we get:

$$\begin{aligned}
t(t^2 + \gamma)(t^2 + 1) &\left(\frac{c'(t)}{c(t)} - \frac{\tilde{c}'(t)}{\tilde{c}(t)} + \frac{1}{t} \right) \\
&= t(t^2 + \gamma)(t^2 + 1) \left(\gamma \frac{t^4 + 2t^2 + \gamma}{t(t^2 + \gamma)(t^4 - \gamma)} - \frac{t^4 + 2\gamma t^2 + \gamma}{t(t^2 + 1)(t^4 - \gamma)} + \frac{1}{t} \right) \\
&= t^4 + 2\gamma t^2 + \gamma > 0,
\end{aligned} \tag{64}$$

which is the desired result.

F   Proof of Theorem 5.1

We denote by $\hat{t}_1^{\mathrm{shr}}, \ldots, \hat{t}_r^{\mathrm{shr}}$ the singular values of $\widehat{\mathbf{X}}^{\mathrm{shr}}$, and $\hat{\mathbf{t}}^{\mathrm{shr}} = (\hat{t}_1^{\mathrm{shr}}, \ldots, \hat{t}_r^{\mathrm{shr}})^T$. We may then write

$$\widehat{\mathbf{X}}^{\mathrm{shr}} = \widehat{\mathbf{U}} \mathrm{diag}(\hat{\mathbf{t}}^{\mathrm{shr}}) \widehat{\mathbf{V}}^T. \tag{65}$$

This is a spectral denoiser (in the set $\mathcal{S}$), and hence its weighted loss with weights Ω_i and Π_j cannot be less than that of the optimal spectral denoiser $\widehat{\mathbf{X}}_{(i,j)}^{\mathrm{loc}}$. That is,

$$\|\Omega_i(\widehat{\mathbf{X}}_{(i,j)}^{\mathrm{loc}} - \mathbf{X})\Pi_j^T\|_{\mathrm{F}}^2 \le \|\Omega_i(\widehat{\mathbf{X}}^{\mathrm{shr}} - \mathbf{X})\Pi_j^T\|_{\mathrm{F}}^2 \tag{66}$$

Because the Ω_i and Π_j are pairwise orthogonal projections which sum to the identity, the total Frobenius loss can be decomposed:

$$\begin{aligned}
\|\widehat{\mathbf{X}}^{\mathrm{loc}} - \mathbf{X}\|_{\mathrm{F}}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \|\Omega_i(\widehat{\mathbf{X}}^{\mathrm{loc}} - \mathbf{X})\Pi_j^T\|_{\mathrm{F}}^2 &= \sum_{i=1}^{I} \sum_{j=1}^{J} \|\Omega_i(\widehat{\mathbf{X}}_{(i,j)}^{\mathrm{loc}} - \mathbf{X})\Pi_j^T\|_{\mathrm{F}}^2 \\
&\le \sum_{i=1}^{I} \sum_{j=1}^{J} \|\Omega_i(\widehat{\mathbf{X}}^{\mathrm{shr}} - \mathbf{X})\Pi_j^T\|_{\mathrm{F}}^2 = \|\widehat{\mathbf{X}}^{\mathrm{shr}} - \mathbf{X}\|_{\mathrm{F}}^2,
\end{aligned} \tag{67}$$

which is the desired inequality.

G   Proof of Theorem 5.2

For $1 \le k \le r$, $1 \le i \le I$, and $1 \le j \le J$, let $\alpha_k^{(i)} = \|\Omega_i \mathbf{u}_k\|^2$, $\mu^{(i)} = \mathrm{tr}(\Omega_i)/p$, $\beta_k^{(j)} = \|\Pi_j \mathbf{v}_k\|^2$, and $\nu^{(j)} = \mathrm{tr}(\Pi_j)/n$. Then

$$\sum_{i=1}^{I} \alpha_k^{(i)} = \sum_{i=1}^{I} \mu^{(i)} = \sum_{j=1}^{J} \beta_k^{(j)} = \sum_{j=1}^{J} \nu^{(j)} = 1. \tag{68}$$

Let $\widehat{\mathbf{X}}_{(i,j)}^{\mathrm{dd}}$ be the optimal diagonal denoiser with weights Ω_i and Π_j. Because of the weighted orthogonality condition, Theorem 4.2 states that the AMSE for $\widehat{\mathbf{X}}_{(i,j)}^{\mathrm{dd}}$ is

$$\|\Omega_i(\widehat{\mathbf{X}}_{(i,j)}^{\mathrm{dd}} - \mathbf{X})\Pi_j^T\|_{\mathrm{F}}^2 = \sum_{k=1}^{r} t_k^2 \alpha_k^{(i)} \beta_k^{(j)} \left(1 - c_k^2 \tilde{c}_k^2 \cdot \frac{\alpha_k^{(i)}}{c_k^2 \alpha_k^{(i)} + s_k^2 \mu^{(i)}} \cdot \frac{\beta_k^{(j)}}{\tilde{c}_k^2 \beta_k^{(j)} + \tilde{s}_k^2 \nu^{(j)}} \right). \tag{69}$$

Since $\widehat{\mathbf{X}}^{\mathrm{loc}}_{(i,j)}$ minimizes the weighted error with weights Ω_i and Π_j, we have:

$$\|\widehat{\mathbf{X}}^{\mathrm{loc}} - \mathbf{X}\|_{\mathrm{F}}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \|\Omega_i(\widehat{\mathbf{X}}^{\mathrm{loc}} - \mathbf{X})\Pi_j^T\|_{\mathrm{F}}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \|\Omega_i(\widehat{\mathbf{X}}^{\mathrm{loc}}_{(i,j)} - \mathbf{X})\Pi_j^T\|_{\mathrm{F}}^2$$

$$\leq \sum_{i=1}^{I}\sum_{j=1}^{J} \|\Omega_i(\widehat{\mathbf{X}}^{\mathrm{dd}}_{(i,j)} - \mathbf{X})\Pi_j^T\|_{\mathrm{F}}^2$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{r} t_k^2 \alpha_k^{(i)}\beta_k^{(j)}\left(1 - c_k^2\tilde{c}_k^2 \cdot \frac{\alpha_k^{(i)}}{c_k^2\alpha_k^{(i)} + s_k^2\mu^{(i)}} \cdot \frac{\beta_k^{(j)}}{\tilde{c}_k^2\beta_k^{(j)} + \tilde{s}_k^2\nu^{(j)}}\right) \cdot$$

$$= \sum_{k=1}^{r} t_k^2\left(1 - c_k^2\tilde{c}_k^2 \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(\alpha_k^{(i)})^2}{c_k^2\alpha_k^{(i)} + s_k^2\mu^{(i)}} \cdot \frac{(\beta_k^{(j)})^2}{\tilde{c}_k^2\beta_k^{(j)} + \tilde{s}_k^2\nu^{(j)}}\right) \cdot \tag{70}$$

On the other hand, the error obtained by $\widehat{\mathbf{X}}^{\mathrm{shr}}$ is equal to

$$\|\widehat{\mathbf{X}}^{\mathrm{shr}} - \mathbf{X}\|_{\mathrm{F}}^2 = \sum_{k=1}^{r} t_k^2(1 - c_k^2\tilde{c}_k^2). \tag{71}$$

Comparing (70) and (71), the result will follow if we can show that for each $1 \leq k \leq r$,

$$\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(\alpha_k^{(i)})^2}{c_k^2\alpha_k^{(i)} + s_k^2\mu^{(i)}} \cdot \frac{(\beta_k^{(j)})^2}{\tilde{c}_k^2\beta_k^{(j)} + \tilde{s}_k^2\nu^{(j)}} \geq 1, \tag{72}$$

where the inequality is strict so long as one of $\mathbf{u}_k$ or $\mathbf{v}_k$ is not generic with respect to some Ω_i or Π_j; or equivalently, either $\alpha_k^{(i)} \neq \mu^{(i)}$ for some i, or $\beta_k^{(j)} \neq \nu^{(j)}$ for some j. Because

$$\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(\alpha_k^{(i)})^2}{c_k^2\alpha_k^{(i)} + s_k^2\mu^{(i)}} \cdot \frac{(\beta_k^{(j)})^2}{\tilde{c}_k^2\beta_k^{(j)} + \tilde{s}_k^2\nu^{(j)}} = \left(\sum_{i=1}^{I} \frac{(\alpha_k^{(i)})^2}{c_k^2\alpha_k^{(i)} + s_k^2\mu^{(i)}}\right) \cdot \left(\sum_{j=1}^{J} \cdot \frac{(\beta_k^{(j)})^2}{\tilde{c}_k^2\beta_k^{(j)} + \tilde{s}_k^2\nu^{(j)}}\right), \tag{73}$$

it is enough to show that

$$\sum_{i=1}^{I} \frac{(\alpha_k^{(i)})^2}{c_k^2\alpha_k^{(i)} + s_k^2\mu^{(i)}} \geq 1, \tag{74}$$

with the inequality being strict so long as $\alpha_k^{(i)} \neq \mu^{(i)}$ for some i.

For each $1 \leq i \leq I$, let $r_i = \alpha_k^{(i)}/\mu^{(i)}$. Then

$$\sum_{i=1}^{I} \frac{(\alpha_k^{(i)})^2}{c_k^2\alpha_k^{(i)} + s_k^2\mu^{(i)}} = \sum_{i=1}^{I} \mu^{(i)} \frac{r_i^2}{c_k^2 r_i + s_k^2}. \tag{75}$$

The function $F(r) = r^2/(c_k^2 r + s_k^2)$ is convex. Since $\sum_{i=1}^{I}\mu^{(i)} = 1$, Jensen's inequality implies

$$\sum_{i=1}^{I} \mu^{(i)} \frac{r_i^2}{c_k^2 r_i + s_k^2} = \sum_{i=1}^{I} \mu^{(i)} F(r_i) \geq F\left(\sum_{i=1}^{I}\mu^{(i)} r_i\right) = F\left(\sum_{i=1}^{I}\alpha_k^{(i)}\right) = F(1) = 1, \tag{76}$$

which is the desired inequality. The inequality will be strict so long as $r_i = \alpha_k^{(i)}/\mu^{(i)}$ is not constantly equal to 1 over i, or equivalently if $\alpha_k^{(i)} \neq \mu^{(i)}$ for some i. This is the desired result.

H Proof of Proposition 6.1

Since $\mathbf{Y}_0 = \Omega\mathbf{Y}\Pi^T$ has only n_0 columns, to ensure that the scaling of the noise matches that of the standard spiked model, we must multiply it by $\sqrt{n/n_0} = 1/\sqrt{\nu}$. We define $\widetilde{\mathbf{Y}}_0 = \mathbf{Y}_0/\sqrt{\nu}$ and $\widetilde{\mathbf{X}}_0 = \mathbf{X}_0/\sqrt{\nu}$. Then $\widetilde{\mathbf{Y}}_0$ follows a standard spiked model with signal matrix $\widetilde{\mathbf{X}}_0$.

For $1 \le k \le r$, we let $\mathbf{u}_k$ and $\mathbf{v}_k$ denote the kth singular vectors of $\mathbf{X}$; $\hat{\mathbf{u}}_k$ and $\hat{\mathbf{v}}_k$ denote the kth singular vectors of $\mathbf{Y}$; $\mathbf{u}_k^0$ and $\mathbf{v}_k^0$ denote the kth singular vectors of $\mathbf{X}_0$ (and $\widetilde{\mathbf{X}}_0$); and $\hat{\mathbf{u}}_k^0$ and $\hat{\mathbf{v}}_k^0$ denote the kth singular vectors of $\mathbf{Y}_0$ (and $\widetilde{\mathbf{Y}}_0$). We let $t_1^0, \ldots, t_r^0$ denote the singular values of $\widetilde{\mathbf{X}}_0$. We also let $\gamma_0 = p_0/n_0 = (\mu/\nu)\gamma$ be the aspect ratio of the submatrix.

If $t_1, \ldots, t_r$ are the singular values of the full p-by-n signal matrix $\mathbf{X}$, then we may write the rescaled submatrix $\widetilde{\mathbf{X}}_0$ as

$$\widetilde{\mathbf{X}}_0 = \Omega\mathbf{X}\Pi^T/\sqrt{\nu} = \frac{1}{\sqrt{\nu}}\sum_{k=1}^{r} t_k\Omega\mathbf{u}_k\mathbf{v}_k^T\Pi^T = \sum_{k=1}^{r} t_k\sqrt{\frac{\alpha_k\beta_k}{\nu}} \frac{\Omega\mathbf{u}_k}{\|\Omega\mathbf{u}_k\|}\left(\frac{\Pi\mathbf{v}_k}{\|\Pi\mathbf{v}_k\|}\right)^T. \tag{77}$$

Because the $\Omega\mathbf{u}_k$ and $\Pi\mathbf{v}_k$ are assumed to by pairwise orthogonal, (77) is the SVD of $\widetilde{\mathbf{X}}_0$. Consequently:

$$\mathbf{u}_k^0 = \frac{\Omega\mathbf{u}_k}{\|\Omega\mathbf{u}_k\|}, \quad \mathbf{v}_k^0 = \frac{\Pi\mathbf{v}_k}{\|\Pi\mathbf{v}_k\|}, \quad t_k^0 = t_k\sqrt{\frac{\alpha_k\beta_k}{\nu}}. \tag{78}$$

We define the cosines

$$c_k^0 = \langle\hat{\mathbf{u}}_k^0, \mathbf{u}_k^0\rangle, \quad \tilde{c}_k^0 = \langle\hat{\mathbf{v}}_k^0, \mathbf{v}_k^0\rangle. \tag{79}$$

Following Remark 7, we may assume that the singular vectors have been chosen so that both c_k^0 and $\tilde{c}_k^0$ are non-negative. Then the AMSE obtained by first applying optimal singular value shrinkage to $\widetilde{\mathbf{Y}}_0$, and then rescaling by ν, is

$$\|\widehat{\mathbf{X}}_0^{\text{shr}} - \mathbf{X}_0\|_F^2 \sim \nu\sum_{k=1}^{r}(t_k^0)^2(1 - (c_k^0\tilde{c}_k^0)^2) \sim \sum_{k=1}^{r}t_k^2\alpha_k\beta_k(1 - (c_k^0\tilde{c}_k^0)^2) \tag{80}$$

We now turn to the weighted estimator $\widehat{\mathbf{X}}_0 = \Omega\widehat{\mathbf{X}}\Pi^T$. From the weighted orthogonality condition, $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}^{\text{dd}}$, the optimal diagonal denoiser. From Theorem 4.2, the AMSE of $\widehat{\mathbf{X}}_0$ may be written

$$\|\widehat{\mathbf{X}}_0 - \mathbf{X}_0\|_F^2 = \|\Omega(\widehat{\mathbf{X}}^{\text{shr}} - \mathbf{X})\|_F^2 \sim \sum_{k=1}^{r}t_k^2\alpha_k\beta_k\left(1 - c_k^2\tilde{c}_k^2 \cdot \frac{\alpha_k}{c_k^2\alpha_k + s_k^2\mu} \cdot \frac{\beta_k}{\tilde{c}_k^2\beta_k + \tilde{s}_k^2\nu}\right). \tag{81}$$

Comparing (80) and (81), the result will be proven if we can show

$$(c_k^0)^2 < c_k^2 \cdot \frac{\alpha_k}{c_k^2\alpha_k + s_k^2\mu}, \quad 1 \le k \le r, \tag{82}$$

and

$$(\tilde{c}_k^0)^2 < \tilde{c}_k^2 \cdot \frac{\beta_k}{\tilde{c}_k^2\beta_k + \tilde{s}_k^2\nu}, \quad 1 \le k \le r. \tag{83}$$

By the symmetry in the problem, it is enough to prove (82). Because we are working with each singular component separately, we will drop the subscript k. From Proposition 3.1, the formula for $(c^0)^2$ is given by

$$(c^0)^2 = \begin{cases} \frac{1 - \gamma_0/(t^0)^4}{1 + \gamma_0/(t^0)^2}, & \text{if } t^0 > \gamma_0^{1/4}, \\ 0, & \text{if } t^0 \le \gamma_0^{1/4}. \end{cases} \tag{84}$$

If $t^0 \le \gamma_0^{1/4}$, then (82) is trivial. Consequently, we assume $t^0 > \gamma_0^{1/4}$. Because $t^0 = t\sqrt{\alpha\beta/\nu}$ and $\gamma_0 = \gamma\mu/\nu$, this is equivalent to the condition

$$t^4 > \gamma\frac{\mu\nu}{\alpha^2\beta^2}. \tag{85}$$

29

Defining $R = \gamma/t^4$, we may consequently assume that

$$R < \frac{\alpha^2\beta^2}{\mu\nu} \leq 1. \tag{86}$$

We may rewrite $(c^0)^2$ in terms of R as follows:

$$(c^0)^2 = \frac{\alpha^2\beta^2 - R\mu\nu}{\alpha^2\beta^2 + t^2\alpha\beta\mu R}. \tag{87}$$

From the formula

$$c^2 = \frac{1 - \gamma/t^4}{1 + \gamma/t^2}, \tag{88}$$

we may rewrite the right side of (82) as

$$\frac{c^2\alpha}{c^2\alpha + s^2\mu} = \frac{\alpha^2\beta^2 - R\mu\nu + R(\mu\nu - \alpha^2\beta^2)}{\alpha^2\beta^2 + t^2\alpha\beta\mu R + \alpha\beta R[t^2\mu\beta + \mu\beta - \alpha\beta - t^2\mu]}. \tag{89}$$

Comparing (87) to (89), the inequality (82) is equivalent to showing

$$(\alpha^2\beta^2 - R\mu\nu)[t^2\mu\beta + \mu\beta - \alpha\beta - t^2\mu] < (\mu\nu - \alpha^2\beta^2)(\alpha\beta + t^2\mu R). \tag{90}$$

Because each side is affine linear in R, and $0 \leq R \leq \alpha^2\beta^2/(\mu\nu)$, it is enough to verify (90) at $R = 0$ and $R = \alpha^2\beta^2/(\mu\nu)$. When $R = \alpha^2\beta^2/(\mu\nu)$, the left side of (90) is 0, whereas the right side is non-negative because $\alpha < \sqrt{\mu}$ and $\beta < \sqrt{\nu}$, verifying the inequality in this case. When $R = 0$, the difference between the right side and left side of (90), divided by $\alpha\beta\mu$, is equal to

$$\begin{aligned}
\nu - \alpha^2\beta^2/\mu - \alpha\beta[t^2\beta + \beta - \alpha\beta/\mu - t^2] &= \nu - \alpha\beta[t^2\beta + \beta - t^2] \\
&= \nu - \alpha\beta^2 t^2 - \alpha\beta^2 + \alpha\beta t^2 \\
&= t^2\alpha\beta(1 - \beta) + \nu - \alpha\beta^2.
\end{aligned} \tag{91}$$

Since $\beta^2 \leq \nu \leq 1$ and $\alpha \leq 1$, this expression is positive, verifying (90) and completing the proof.

I   Proof of Proposition 6.2

From a standard Bernstein-type inequality for subexponential random variables (e.g. Proposition 5.16 from [56]), for every $\theta > 0$ we have:

$$\mathbb{P}(\max_i |\hat{a}_i - \mathbb{E}[\hat{a}_i]| > \theta) \leq Cp\exp\{-C'n\min(\theta^2, \theta)\}, \tag{92}$$

for constants $C, C' > 0$. Since $p \sim \gamma n$, the right hand side is summable in n; it follows from the Borel-Cantelli Lemma that $\max_i |\hat{a}_i - \mathbb{E}[\hat{a}_i]| \to 0$ almost surely as $n \to \infty$. From the delocalization of $\mathbf{X}$'s singular vectors, $\sup_{i,j} |X_{ij}|^2 = o(1/n)$. Since $\mathrm{tr}(\mathbf{T})/n = 1$, we then have

$$\mathbb{E}[\hat{a}_i] = \mathbb{E}\left[\sum_{j=1}^n X_{ij}^2 + 2\sum_{j=1}^n X_{ij}\sqrt{a_i b_j} G_{ij} + \sum_{j=1}^n a_i b_j G_{ij}^2\right] = a_i + o(1). \tag{93}$$

Consequently, $\|\widehat{\mathbf{S}}_p - \mathbf{S}_p\|_{\mathrm{op}} = \max_{1 \leq i \leq p} |\hat{a}_i - a_i| \to 0$ almost surely, as desired. Similar reasoning also shows that $\sum_{i=1}^p (\hat{a}_i - a_i)/n \to 0$ almost surely as $n \to \infty$. A nearly identical argument applied to the numerator of $\hat{b}_j$ then shows $\|\widehat{\mathbf{T}}_n - \mathbf{T}_n\|_{\mathrm{op}} \to 0$ almost surely, completing the proof.

J Proof of Proposition 6.3

To prove Proposition 6.3, we begin by deriving a lower bound on the operator norm of the noise matrix $\mathbf{N} = \mathbf{S}^{1/2}\mathbf{G}\mathbf{T}^{1/2}$. We let $\mathbf{a}$ and $\mathbf{b}$ be unit vectors so that $\mathbf{G}^T\mathbf{b} = \|\mathbf{G}\|_{\mathrm{op}}\mathbf{a}$. Then

$$\|\mathbf{G}\mathbf{T}^{1/2}\|_{\mathrm{op}} \geq \|\mathbf{T}^{1/2}\mathbf{G}^T\mathbf{b}\| = \|\mathbf{G}\|_{\mathrm{op}}\|\mathbf{T}^{1/2}\mathbf{a}\|. \tag{94}$$

Next, we take unit vectors $\mathbf{c}$ and $\mathbf{d}$ so that $\mathbf{G}\mathbf{T}^{1/2}\mathbf{d} = \|\mathbf{G}\mathbf{T}^{1/2}\|_{\mathrm{op}}\mathbf{c}$. Then we have

$$\|\mathbf{N}\|_{\mathrm{op}}^2 \geq \|\mathbf{S}^{1/2}\mathbf{G}\mathbf{T}^{1/2}\mathbf{d}\|^2 = \|\mathbf{G}\mathbf{T}^{1/2}\|_{\mathrm{op}}^2\|\mathbf{S}^{1/2}\mathbf{c}\|^2 \geq \|\mathbf{G}\|_{\mathrm{op}}^2 \cdot \|\mathbf{T}^{1/2}\mathbf{a}\|^2 \cdot \|\mathbf{S}^{1/2}\mathbf{c}\|^2. \tag{95}$$

Since the distribution of $\mathbf{G}$ is orthogonally-invariant, the distributions of $\mathbf{a}$ and $\mathbf{c}$ are uniform over the unit spheres in $\mathbb{R}^n$ and $\mathbb{R}^p$, respectively. Consequently, $\|\mathbf{T}^{1/2}\mathbf{a}\|^2 \sim \mathrm{tr}(\mathbf{T})/n$ and $\|\mathbf{S}^{1/2}\mathbf{c}\|^2 \sim \mathrm{tr}(\mathbf{S})/p$. Therefore,

$$\|\mathbf{N}\|_{\mathrm{op}}^2 \gtrsim (\mathrm{tr}(\mathbf{S})/p) \cdot (\mathrm{tr}(\mathbf{T})/n) \cdot \|\mathbf{G}\|_{\mathrm{op}}^2 \tag{96}$$

where the inequality holds almost surely in the large p, large n limit. Note that $\|\mathbf{G}\|_{\mathrm{op}} \sim 1 + \sqrt{\gamma}$ (see, e.g., [2]), though we do not need to use this fact.

Furthermore, we also have

$$\tilde{t}_k^2 = t_k^2\|\mathbf{S}^{-1/2}\mathbf{u}_k\|^2\|\mathbf{T}^{-1/2}\mathbf{v}_k\|^2 \sim t_k^2 \cdot \frac{1}{p}\mathrm{tr}(\mathbf{S}^{-1}) \cdot \frac{1}{n}\mathrm{tr}(\mathbf{T}^{-1}). \tag{97}$$

Consequently,

$$\begin{aligned}
\mathrm{SNR}_k = \frac{t_k^2}{\|\mathbf{N}\|_{\mathrm{op}}^2} &\lesssim \frac{t_k^2}{(\mathrm{tr}(\mathbf{S})/p) \cdot (\mathrm{tr}(\mathbf{T})/n) \cdot \|\mathbf{G}\|_{\mathrm{op}}^2} \\
&= \frac{t_k^2 \cdot (\mathrm{tr}(\mathbf{S}^{-1})/p) \cdot (\mathrm{tr}(\mathbf{T}^{-1})/n)}{\tau\|\mathbf{G}\|_{\mathrm{op}}^2} = \frac{\tilde{t}_k^2}{\tau\|\mathbf{G}\|_{\mathrm{op}}^2} = \frac{1}{\tau}\widetilde{\mathrm{SNR}}_k,
\end{aligned} \tag{98}$$

completing the proof.

K Proof of Proposition 6.4

Let δ_{ij} be 1 if entry (i,j) is sampled, and 0 otherwise. Then $\delta_{ij} \sim \mathrm{Bernoulli}(p_iq_j)$. Let $\Delta = (\delta_{ij})$; then $\mathcal{F}^*(\mathcal{F}(\mathbf{X})) = \Delta \odot \mathbf{X}$, where $\odot$ denotes the Hadamard product. Let $\mathbf{q}_r = (q_1^r, \ldots, q_p^r)^T$ and $\mathbf{q}_c = (q_1^c, \ldots, q_n^c)^T$. The matrix $\Delta - \mathbf{q}_r\mathbf{q}_c^T$ is a random matrix with mean zero, whose entries are uniformly bounded. It follows from Corollary 2.3.5 of [54] that $\|\Delta - \mathbf{q}_r\mathbf{q}_c^T\|_{\mathrm{op}}/\sqrt{n} \leq A$ a.s. as $n \to \infty$, for some constant $A > 0$.

We may write $\mathbf{P}\mathbf{X}\mathbf{Q} = \mathbf{X} \odot (\mathbf{q}_r\mathbf{q}_c^T)$, and consequently $\Delta \odot \mathbf{X} - \mathbf{P}\mathbf{X}\mathbf{Q} = (\Delta - \mathbf{q}_r\mathbf{q}_c^T) \odot \mathbf{X}$. Since $\mathbf{X} = \sum_{k=1}^r t_k\mathbf{u}_k\mathbf{v}_k^T$, it is enough to show that

$$\|(\Delta - \mathbf{q}_r\mathbf{q}_c^T) \odot \mathbf{u}_k\mathbf{v}_k^T\|_{\mathrm{op}} \to 0 \tag{99}$$

almost surely, for each k.

Suppose $\mathbf{a}$ and $\mathbf{b}$ are two unit vectors. Then

$$\left|\mathbf{a}^T[(\Delta - \mathbf{q}_r\mathbf{q}_c^T) \odot \mathbf{u}_k\mathbf{v}_k^T]\mathbf{b}\right| = \left|(\mathbf{a} \odot \mathbf{u}_k)^T(\Delta - \mathbf{q}_r\mathbf{q}_c^T)(\mathbf{b} \odot \mathbf{v}_k)\right| \leq A\sqrt{n}\|\mathbf{a} \odot \mathbf{u}_k\|\|\mathbf{b} \odot \mathbf{v}_k\| \tag{100}$$

almost surely as $n \to \infty$.

Now, since $\mathbf{a}$ is a unit vector,

$$\|\mathbf{a} \odot \mathbf{u}_k\| = \sqrt{\sum_{j=1}^p a_j^2 u_{jk}^2} \leq \|\mathbf{u}_k\|_\infty \tag{101}$$

and similarly,

$$\|\mathbf{b} \odot \mathbf{v}_k\| \leq \|\mathbf{v}_k\|_\infty. \tag{102}$$

Since $\max_{1 \leq k \leq r}\|\mathbf{u}_k\|_\infty\|\mathbf{v}_k\|_\infty = o(n^{-1/2})$, the result follows.