Matrix Denoising with Partial Noise Statistics: Optimal Singular Value Shrinkage of Spiked F-Matrices

Matan Gavish*[1], William Leeb†[2], and Elad Romanov‡[3]

[1]School of Computer Science and Engineering, Hebrew University of Jerusalem
[2]School of Mathematics, University of Minnesota
[3]Department of Statistics, Stanford University

Abstract

We study the problem of estimating a large, low-rank matrix corrupted by additive noise of unknown covariance, assuming one has access to additional side information in the form of noise-only measurements. We study the Whiten-Shrink-reColor (WSC) workflow, where a "noise covariance whitening" transformation is applied to the observations, followed by appropriate singular value shrinkage and a "noise covariance re-coloring" transformation. We show that under the mean square error loss, a unique, asymptotically optimal shrinkage nonlinearity exists for the WSC denoising workflow, and calculate it in closed form. To this end, we calculate the asymptotic eigenvector rotation of the random spiked F-matrix ensemble, a result which may be of independent interest. With sufficiently many pure-noise measurements, our optimally-tuned WSC denoising workflow outperforms, in mean square error, matrix denoising algorithms based on optimal singular value shrinkage which do not make similar use of noise-only side information; numerical experiments show that our procedure's relative performance is particularly strong in challenging statistical settings with high dimensionality and large degree of heteroscedasticity.

1 Introduction

Low-rank matrix reconstruction from partial or corrupted measurements is a well-studied problem in machine learning and statistics, with applications ranging from computer vision [45] and structural biology [2, 12] to medical imaging [16] and medical signal processing [42]. This paper considers recovery of a p-by-n matrix $\mathbf{X}$ of rank $r \ll n, p$ from an additively corrupted measured matrix $\mathbf{Y} = \mathbf{X} + \mathbf{R}$. Here, $\mathbf{R}$ is a noise matrix (independent of $\mathbf{X}$) whose columns are assumed independent and identically-distributed (i.i.d.), with an arbitrary between-row correlation structure $\mathbf{\Sigma}$.

Such matrix denoising problems occur, for example, in principal component analysis (PCA) under a low-rank factor model [57]. Consider n data points in p-dimensional Euclidean space, denoted $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^p$ of the form $\mathbf{y}_i = \mathbf{x}_i + \boldsymbol{\varepsilon}_i$. The i.i.d. "signal" vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are assumed to lie on some unknown r-dimensional subspace, and the i.i.d. "noise" vectors $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ have a full rank covariance matrix $\mathbf{\Sigma}$, and so spread over the entire ambient space. One would like to reconstruct the low dimensional signal vectors from the noisy observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$. Let $\mathbf{Y} \in \mathbb{R}^{p \times n}$ be the data matrix, formed by stacking the observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ as its columns, so that $\mathbf{Y} = \mathbf{X} + \mathbf{R}$, where $\mathbf{X} \in \mathbb{R}^{p \times n}$ is a matrix whose columns are $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $\mathbf{R}$ is a noise

*gavish@cs.huji.ac.il
†wleeb@umn.edu
‡elad.romanov@gmail.com

matrix. This signal estimation problem becomes a matrix denoising problem: estimate $\mathbf{X}$ from $\mathbf{Y}$.

This paper approaches the matrix denoising problem in the well-known *spiked model* [31], wherein the matrix dimensions $p, n \to \infty$ with a limiting aspect ratio $\gamma := \lim_{p,n\to\infty} p/n > 0$ while the signal rank $r = \text{rank}(\mathbf{X})$ is fixed. The spiked model captures the key features of the matrix denoising problem in a regime where both underlying signal rank and signal-to-noise ratio (SNR) are small; in particular, the ground truth $\mathbf{X}$ is not consistently estimable from $\mathbf{Y}$ (see Section 2 below).

From the perspective of random matrix theory, the spiked model has relatively simple and well-understood asymptotic behavior [7, 8, 11, 49]. Specifically, the behavior of $\mathbf{Y}$ is described by the following phenomena:

1. *Singular value displacement:* The singular values of $\mathbf{Y}$ are divided into a "bulk", which has a deterministic limiting shape and corresponds to pure noise, and at most r "outliers" which exceed the bulk and correspond to "signal". The locations of the outliers are asymptotically deterministic, and the i-th largest singular value of $\mathbf{Y}$ depends only on the i-th largest singular value of $\mathbf{X}$. The presence (or lack thereof) of an outlier is a threshold phenomenon: there is some SNR level σ^*, a detection threshold, such that the i-th singular value σ_i creates an observable outlier if and only if $\sigma_i > \sigma^*$.

2. *Principal component angles:* The angles between the leading observed PCs and the signal PCs are essentially deterministic. The i-th signal PC is essentially orthogonal to the the j-th ($j \neq i$) observed PC. The angle between the i-th signal and i-th observed PCs concentrates around a deterministic number $\in [0, 1)$, which may be consistently estimated from the observed i-th singular value of $\mathbf{Y}$.

A popular and practical approach to the matrix denoising problem is *singular value shrinkage*, where $\mathbf{X}$ is estimated by taking the singular value decomposition (SVD) of $\mathbf{Y}$, retaining its singular vectors while systematically deflating the singular values to correct for the noise [22, 24, 26, 47, 50, 52]. Leveraging the above spiked model asymptotic phenomena, which are entirely quantifiable, several authors have derived optimal singular value shrinkers under various settings, cf. [22, 23, 25, 26, 28, 30, 38–40, 47, 52, 56].

For isotropic noise ($\mathbf{\Sigma} = \mathbf{I}$), matrix denoising in the spiked model is well-understood. However, the general case of heteroscedastic noise offers interesting problems of practical interest. Several recent works have studied PCA, singular value shrinkage and related spectral methods in the presence of heteroscedastic noise, either across rows, columns or both; see for example [1, 9, 20, 23, 28–30, 35, 39–41, 47, 56, 65]. Under our present setting of independent columns and inter-row covariance matrix $\mathbf{\Sigma}$, formulas for the optimal singular value shrinker are available: when the noise covariance $\mathbf{\Sigma}$ is known, the optimal shrinker can be computed exactly, and when it is not, one can consistently estimate the optimal shrinkage rule from the observed spectrum of $\mathbf{Y}$; the resulting procedure is known as *OptShrink* [47] (see also [23]).

When denoising a signal sampled with additive heteroscedastic noise, one sometimes has the opportunity to sample the noisy channel in the absence of any signal. A natural approach – indeed a classical idea in signal estimation – is to use noise-only measurements to "whiten" the measurements. According to this approach, one should (i) "whiten" the data, that is multiply by $\mathbf{\Sigma}^{-1/2}$ (assuming $\mathbf{\Sigma}$ is somehow available); (ii) apply an estimation procedure calibrated for uncorrelated noise, yielding an estimator $\widehat{\mathbf{X}}^w$; and (iii) apply a "recoloring" transformation $\widehat{\mathbf{X}} = \mathbf{\Sigma}^{1/2}\widehat{\mathbf{X}}^w$. The popularity of this approach in signal processing is due both to its conceptual simplicity, and to the ubiquity of linear filtering, under which it is often optimal. The influential textbook of van Trees [58] advocates, for example: *"Many of our models assumed the received signal, either a waveform or a vector, was observed [in] the presence of "white noise" [...] We demonstrated, first with vectors and then with waveforms, that we could always find a "whitening transformation" that mapped the original process into a signal plus white noise problem. The*

reader should remember to consider this approach when dealing with more general problems." ([58, Epilogue]).

In our present denoising problem, under the spiked model, it is natural to consider a ***W**hiten-**S**hrink-re**C**olor* (WSC) workflow, where a "signal covariance whitening" transformation is applied to the observations, followed by appropriate singular value shrinkage and a "signal covariance recoloring" transformation. The simplest scenario where WSC may be considered is when the noise (population) covariance $\mathbf{\Sigma}$ is available as side information. This case was studied recently by two of the authors, who derived the optimal shrinkage rule of the WSC procedure [40]. As one can expect, denoising performance were shown to improve by incorporating the side information $\mathbf{\Sigma}$ into the WSC workflow, compared with optimal singular value shrinkage [47]; interestingly, the optimal shrinker for WSC was found to be different from the singular value shrinker optimally tuned for white noise.

Clearly, the assumption that the noise covariance $\mathbf{\Sigma}$ is known or consistently estimable is typically unrealistic in high dimensions [13]. As [40] demonstrated, complete knowledge of the noise covariance offers significant improvement for matrix denoising under heteroscedastic noise. One then naturally wonders whether *partial* information of the noise $\mathbf{R}$ could be similarly leveraged. Specifically, assume access to side information in the form of $m > p$ *pure-noise* samples. Following [39], the present paper considers matrix denoising in the spiked model under a WSC workflow, where instead of the noise population covariance, "whitening" and "recoloring" are done using a sample covariance matrix of pure-noise samples $\widehat{\mathbf{\Sigma}}$. This problem is fascinating in part owing to the fact that, as $\widehat{\mathbf{\Sigma}}$ is an inconsistent estimate of $\mathbf{\Sigma}$, whitening by $\widehat{\mathbf{\Sigma}}$ injects additional "noise" into the estimation procedure, which should be systematically corrected for.

Contributions.

1. **Optimal WSC denoiser.** Our main contribution is the derivation of the optimal WSC denoiser in mean square error. Specifically, we show that an asymptotically optimal shrinker exists for this WSC workflow, and derive it in closed form.

2. **Asymptotic singular vector rotations for the spiked F-matrix.** Curiously, the whitened data matrix

$$\mathbf{Y}^w = \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{Y} = \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{X} + \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{R}\,.$$

   is an object of independent interest known in the random matrix theory literature as a *spiked F-matrix* [6, 48]. Prior works in signal processing and statistics have studied the problem of signal detection under spiked F-matrix ensemble, focusing on the behavior of the largest eigenvalues in the presence or absence of a signal [18, 33, 34, 48, 55, 66, 67]. In particular, it was shown [48] that the singular values of $\mathbf{Y}^w$ exhibit the same basic phenomenology similar to the "classical" spiked model:[1] its singular values are arranged in the form of a "bulk" plus at most r outliers exceeding the bulk. (The limiting distribution of the bulk was calculated already in the classical work of Wachter [61].) Formulas for the spike detection threshold, as well as the spike-forward map (singular value displacements) have also been computed. As discussed above, derivation of optimal shrinkers for WSC worflow in our scenario requires calculation of the limiting angles between the population (signal) principal components and their empirical counterparts. A secondary contribution of the present paper is closed-form formulas for the limiting angles of the spiked F-matrix ensemble.

[1] Note that $\mathbf{Y}^w$ does *not* follow a generalized spiked model in the sense of e.g. [11], since the low-rank and noise parts are statistically dependent on one another through their mutual dependence on $\widehat{\mathbf{\Sigma}}$.

Paper outline. This paper is organized as follows. In Section 2, we state the precise observation model and estimation problem; provide key definitions of functions; and describe the proposed denoising algorithm in detail. In Section 3, we state the theoretical results on spiked F-matrices, and show how these may be used to derive the optimal denoisers. In Section 4 we briefly survey several known results on the properties of the F-matrix ensemble, which shall be crucial in the derivation to follow. Section 5 is devoted to the proofs of our technical results, with some details deferred to the Appendix. Lastly, in Section 6 we report on numerical experiments illustrating the behavior of the proposed denoising method. In particular, we numerically compare the performance of our method to that of optimal singular value shrinkage (OptShrink [47]) under different model configurations. Section 7 is devoted to conclusion and some additional discussion.

2 Notation and problem setup

2.1 Observation model

Let $\sigma_1, \ldots, \sigma_r > 0$, be positive numbers, and $\mathbf{u}_1, \ldots, \mathbf{u}_r \in \mathbb{R}^p$ and $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{R}^n$ be vectors. Denote by $\mathbf{U} \in \mathbb{R}^{p\times r}$ the matrix whose columns are $\mathbf{u}_1, \ldots, \mathbf{u}_r$, and similarly for $\mathbf{V} \in \mathbb{R}^{n\times r}$. Also, let $\mathbf{\Lambda} \in \mathbb{R}^{r\times r}$ be a diagonal matrix with diagonal given by $(\sigma_1, \ldots, \sigma_r)$. The signal matrix, the object to be estimated, is

$$\mathbf{X} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top . \tag{1}$$

Clearly, $\mathrm{rank}(\mathbf{X}) \le r$. Let $\mathbf{Z} \in \mathbb{R}^{p\times n}$ be a random matrix, independent of $\mathbf{X}$, with i.i.d. Gaussian entries $Z_{i,j} \sim \mathcal{N}(0,1)$, and let $\mathbf{R} = \mathbf{\Sigma}^{1/2}\mathbf{Z}$, where $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ is positive definite. One observes $\mathbf{Y} \in \mathbb{R}^{p\times n}$,

$$\mathbf{Y} = \mathbf{X} + \frac{1}{\sqrt{n}}\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top + \frac{1}{\sqrt{n}}\mathbf{\Sigma}^{1/2}\mathbf{Z} . \tag{2}$$

That is, each column of $\mathbf{X}$ is corrupted by additive noise of mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}/n$. We remark that this normalization (dividing the noise by $\sqrt{n}$) is such that the singular values of the signal $\mathbf{X}$ and the noise $n^{-1/2}\mathbf{R}$ are of the same order, cf. [22, 23, 26, 39, 47, 52].

The noise covariance $\mathbf{\Sigma}$ is assumed to be unknown. Instead, one is given side information in the form of m pure-noise samples, which are independent of the measurement matrix $\mathbf{Y}$. That is, one observes a noise-only matrix

$$\mathbf{R}' = \mathbf{\Sigma}^{1/2}\mathbf{Z}' \in \mathbb{R}^{p\times m}, \qquad \mathbf{Z}'_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1) . \tag{3}$$

We will assume that $m \ge p$ so that $\mathrm{rank}(\mathbf{R}') = \mathrm{rank}(\mathbf{\Sigma}^{1/2})$ with probability (w.p.) 1. Having observed the signal-plus-noise measurement matrix $\mathbf{Y}$ and the side information $\mathbf{R}'$, our goal is to estimate the signal matrix $\mathbf{X}$. For an estimator $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}(\mathbf{Y}, \mathbf{R}')$, we measure its error using the Frobenius loss (MSE): $\mathbb{E}\|\mathbf{X} - \widehat{\mathbf{X}}\|_F^2$. The matrices in the observation model are summarized in Table 1.

We study this denoising problem under the spiked model [11, 31]. Formally, we consider a sequence of denoising problems, $n, m, p \to \infty$, with the following specifications:

1. **"High-dimensional" asymptotics:** For constants $\gamma \in (0, \infty)$ and $\beta \in (0, 1)$,

$$\frac{p}{n} \to \gamma, \quad \frac{p}{m} \to \beta, \quad \text{as} \quad n, m, p \to \infty . \tag{4}$$

2. **Regularity of noise covariance sequence:**

Symbol(s)	Description	Size(s)	Observed?
$\mathbf{U}$, $\mathbf{V}$, $\mathbf{\Lambda}$	Signal SVD	$p \times r$, $p \times n$, $r \times r$	Not observed
$\mathbf{X}$	Signal $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$	$p \times n$	Not observed
$\mathbf{Z}$, $\mathbf{Z}'$	White noise	$p \times n$, $p \times m$	Not observed
$\mathbf{\Sigma}$	Noise covariance	$p \times p$	Not observed
$\mathbf{R}$	Noise $\mathbf{\Sigma}^{1/2}\mathbf{Z}$	$p \times n$	Not observed
$\mathbf{R}'$	Out-of-sample noise $\mathbf{\Sigma}^{1/2}\mathbf{Z}'$	$p \times m$	Observed
$\mathbf{Y}$	Signal-plus-noise $\mathbf{X} + \mathbf{R}$	$p \times n$	Observed
$\widehat{\mathbf{U}}$, $\widehat{\mathbf{V}}$, $\widehat{\mathbf{\Lambda}}$	r-SVD of $\mathbf{Y}$	$p \times r$, $p \times n$, $r \times r$	Observed
$\widehat{\mathbf{\Sigma}}$	Sample covariance $\mathbf{R}'\mathbf{R}'^\top/m$	$p \times p$	Observed
$\mathbf{N}$	Pseudo-whitened noise $\widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{R}/\sqrt{n}$	$p \times m$	Not observed
$\mathbf{E}$	Wishart matrix $\mathbf{Z}\mathbf{Z}^\top/n$	$p \times p$	Not observed
$\mathbf{S}$	Wishart matrix $\mathbf{Z}'\mathbf{Z}'^\top/m$	$p \times p$	Not observed
$\mathbf{D}_1, \dots, \mathbf{D}_r$	Signal PC weights	$p \times p$	Not observed

Table 1: Matrices used in this paper.

(a) The empirical spectral distribution (ESD) of $\mathbf{\Sigma}$ converges weakly almost surely to some deterministic, *compactly supported* law dH. To wit, for every bounded continuous $f(\cdot)$, $p^{-1}\mathrm{tr}(f(\mathbf{\Sigma})) := p^{-1}\sum_{i=1}^{p} f(\lambda_i(\mathbf{\Sigma})) \longrightarrow \int f(\lambda) dH(\lambda)$ a.s. as $p \to \infty$.

(b) The extremal eigenvalues of $\mathbf{\Sigma}$ converge to the edges of the support of dH. To wit, if $\mathrm{supp}(dH) = [a, b]$ then $\lambda_{\min}(\mathbf{\Sigma}) \to a$, $\lambda_{\max}(\mathbf{\Sigma}) \to b$. We further require that $a > 0$.

We denote the first moment of the limiting empirical spectral distribution (LESD) by

$$\mu := \lim_{p\to\infty} p^{-1}\,\mathrm{tr}(\mathbf{\Sigma}) = \int_a^b \lambda dH(\lambda)\,. \tag{5}$$

3. **Generative assumptions on signal:** The number of spikes r and the intensities $\sigma_1, \dots, \sigma_r > 0$ are fixed as $n, m, p \to \infty$. The signal principal directions satisfy the following:

(a) **Right principal directions:** The matrix $\mathbf{V} \in \mathbb{R}^{n\times r}$ is such that $\mathbf{V}^\top\mathbf{V} \to \mathbf{I}_{r\times r}$ almost surely. In other words, $\{\mathbf{v}_1, \dots, \mathbf{v}_r\}$ are (asymptotically) orthonormal vectors.

(b) **Left principal directions:** The vectors $\mathbf{u}_1, \dots, \mathbf{u}_r$ have the form[2]

$$\mathbf{u}_i = p^{-1/2}\mathbf{D}_i\mathbf{w}_i\,, \quad \text{for} \quad \mathbf{w}_1, \dots, \mathbf{w}_r \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p\times p})\,. \tag{6}$$

Furthermore, the (sequences of) matrices $\mathbf{D}_i$ satisfy:

i. *Boundedness:* For some $C > 0$, $\max_{1\le i\le r} \|\mathbf{D}_i\| < C$ almost surely.

ii. *Unit energy:* $\lim_{p\to\infty} p^{-1}\|\mathbf{D}_i\|_F^2 = 1$.

iii. *Limiting joint law:* The algebra generated by $\{\mathbf{D}_i\mathbf{D}_i^\top, \mathbf{\Sigma}, \mathbf{\Sigma}^{-1}\}$ has a limiting joint law in the sense of free probability theory (see Section A.2).

The following quantities will play an important role in our formulas:

$$\tau_i := \lim_{p\to\infty} p^{-1}\mathrm{tr}\left(\mathbf{D}_i^\top\mathbf{\Sigma}^{-1}\mathbf{D}_i\right)\,, \quad 1 \le i \le r. \tag{7}$$

We assume that each τ_i is finite and strictly positive.

[2]The Gaussianity of the vectors $\mathbf{w}_i$ is not strictly necessary; one could replace it by any other isotropic distribution of sufficiently light tail.

(c) **Simple signal spectrum:** The following "effective" spike intensities contain no multiplicities. By way of notation, they are ordered as $\sqrt{\tau_1}\sigma_1 > \ldots > \sqrt{\tau_r}\sigma_r$. We remark that this assumption is standard throughout the literature on singular value shrinkage, see for example [23, 26, 39, 47, 52].

Remark 1. *Much of the existing literature on singular value shrinkage either assumes isotropic noise ($\mathbf{\Sigma} = \mathbf{I}_{p\times p}$) or an isotropic prior ($\mathbf{D}_i = \mathbf{I}_{p\times p}$) on the left signal directions $\mathbf{u}_1, \ldots, \mathbf{u}_r$ (cf. [23, 26, 47, 52]). Similar to [39], we relax this assumption by allowing in (6) for an alignment between the signal direction and the noise, whose "strength" is captured by the parameter τ_i from (7). Note that (6) implies that $\mathbf{u}_i^\top \mathbf{u}_j \to \mathbb{1}\{i = j\}$ a.s., thus the $\mathbf{u}_i$-s may be interpreted as the principal components of $\mathbf{X}$ (though this statement is only precise asymptotically). In the language of factor analysis, the $\mathbf{v}_i$-s may be interpreted as the factor values [3, 4, 21, 51].*

Remark 2. *Throughout the paper we assume that $\mathbf{Z}, \mathbf{Z}'$ have Gaussian entries; this assumption is used explicitly in the proofs (specifically, the bi-orthogonal invariance of these matrices). In Section 6 we give numerical evidence indicating that our results should be universal with respect to the noise distribution: they continue to hold when $\mathbf{Z}, \mathbf{Z}'$ are i.i.d. with sufficiently light-tailed isotropic entries.*

2.2 Whiten-Shrink-reColor (WSC) denoisers

Our task is to estimate the signal matrix $\mathbf{X}$ from the observed signal-plus-noise matrix $\mathbf{Y} = \mathbf{X}+\mathbf{R}$ and the noise-only samples $\mathbf{R}'$. We next describe the class of procedures we consider for this problem.

Let $\widehat{\mathbf{\Sigma}}$ be an estimate of the covariance matrix $\mathbf{\Sigma}$. Later, we will take $\widehat{\mathbf{\Sigma}} = \mathbf{R}'\mathbf{R}'^\top/m$, but for now any estimate would suffice. We use $\widehat{\mathbf{\Sigma}}$ to *pseudo-whiten* the noise on the observation matrix, constructing a new matrix $\mathbf{Y}^w$:

$$\mathbf{Y}^w = \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{Y} = \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top + \frac{1}{\sqrt{n}}(\widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{\Sigma}^{1/2})\mathbf{Z}\,. \tag{8}$$

We consider the following family of estimators $\mathcal{F}$, computed from the SVD of $\mathbf{Y}^w$:

$$\mathbf{Y}^w \overset{\text{SVD}}{=} \sum_{k=1}^{\min\{p,n\}} \widehat{\theta}_k \widehat{\mathbf{u}}_k^w \widehat{\mathbf{v}}_k^w, \quad \mathcal{F} = \left\{ \sum_{k=1}^{\widehat{r}} \eta_k \widehat{\mathbf{\Sigma}}^{1/2} \widehat{\mathbf{u}}_k^w \widehat{\mathbf{v}}_k^w : \eta_1, \ldots, \eta_{\widehat{r}} \in \mathbb{R} \right\}, \tag{9}$$

where $\widehat{r}$ denotes a data-driven estimator of $\text{rank}(\mathbf{X})$, to be described in Section 2.4. Let $\widehat{\mathbf{X}}^{\boldsymbol{\eta}} = \sum_{k=1}^{\widehat{r}} \eta_k \widehat{\mathbf{\Sigma}}^{1/2} \widehat{\mathbf{u}}_k^w \widehat{\mathbf{v}}_k^w$, where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{\widehat{r}})$. It will be shown later (in Section 3.2) that for any deterministic $\boldsymbol{\eta}$, the asymptotic loss

$$\text{AMSE}(\boldsymbol{\eta}) = \lim_{p\to\infty} \|\mathbf{X} - \widehat{\mathbf{X}}^{\boldsymbol{\eta}}\|_{\text{F}}^2 \tag{10}$$

almost surely exists, and is finite. The goal, then, is to find $\eta_1, \ldots, \eta_r$ so to minimize the asymptotic loss:

$$\boldsymbol{\eta} = \operatorname*{argmin}_{\widetilde{\boldsymbol{\eta}}} \text{AMSE}(\widetilde{\boldsymbol{\eta}}). \tag{11}$$

We will derive the optimal choice of $\eta_1, \ldots, \eta_{\widehat{r}}$ in Section 3.2.

Remark 3. *The values $\eta_1, \ldots, \eta_{\widehat{r}}$ are known as the generalized singular values of $\widehat{\mathbf{X}}$ with respect to the matrix $\widehat{\mathbf{\Sigma}}^{-1}$; correspondingly, the vectors $\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_1^w, \ldots, \widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_r^w$, $\widehat{\mathbf{v}}_1^w, \ldots, \widehat{\mathbf{v}}_r^w$ are the generalized singular vectors [43]. That is, $\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_1^w, \ldots, \widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_r^w$ are orthonormal with respect to the weighted inner product $\langle \mathbf{x}, \tilde{\mathbf{x}} \rangle_{\widehat{\mathbf{\Sigma}}^{-1}} = \mathbf{x}^\top \widehat{\mathbf{\Sigma}}^{-1} \tilde{\mathbf{x}}$.*

An equivalent formulation of the estimator, which may be more natural, is given as follows. Define

$$\widehat{\mathbf{u}}_k = \frac{\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w}{\|\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w\|}, \quad \widehat{\mathbf{v}}_k = \widehat{\mathbf{v}}_k^w \quad 1 \le k \le r, \tag{12}$$

so that we may write the estimator in the form

$$\widehat{\mathbf{X}} = \sum_{k=1}^{r} t_k \widehat{\mathbf{u}}_k \widehat{\mathbf{v}}_k, \quad \text{where} \quad t_k = \|\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w\| \cdot \eta_k. \tag{13}$$

As we will show in Theorem 2, the vectors $\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_r$ are asymptotically orthonormal as $p, n, m \to \infty$; consequently, equation (13) is an approximate SVD of $\widehat{\mathbf{X}}$, with approximate singular values $t_1, \ldots, t_r$. The vectors $\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_r$ may be interpreted as estimates of the population principal components, $\mathbf{u}_1, \ldots, \mathbf{u}_r$. Accordingly, we will refer to $\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_r$ as the *empirical principal components*.

Rationale for the estimation procedure. The estimator family $\mathcal{F}$ in (9) was studied in the authors' previous work [39] when the estimator $\widehat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ is asymptotically consistent in operator norm as $p, n, m \to \infty$; in the setting of the present paper, this occurs when, for example, $\widehat{\mathbf{\Sigma}}$ is a sample covariance and $p/m \to 0$. In this setting, it is shown that under a uniform prior on the singular vectors of $\mathbf{X}$, the estimator $\widehat{\mathbf{X}}$ outperforms the optimal singular value shrinkage estimator (OptShrink) described in [47]. It follows that so long as $\widehat{\mathbf{\Sigma}}$ is sufficiently close to $\mathbf{\Sigma}$, the optimal $\widehat{\mathbf{X}}$ will outperform OptShrink as well.

2.3 Key definitions

We introduce several parameters and functions that will be used to evaluate the optimal $\eta_1, \ldots, \eta_r$. Their significance will be explained in Section 4; for now, we simply present their definitions. The key parameters and functions, along with others that are introduced later in the paper, are summarized in Tables 2 and 3, respectively.

Define the following values:

$$\sigma_{\text{thresh}} = \sqrt{\frac{\beta + \sqrt{\beta + \gamma - \beta\gamma}}{1 - \beta}}, \tag{14}$$

and

$$\theta_{\text{max}} = \frac{1 + \sqrt{\beta + \gamma - \beta\gamma}}{1 - \beta}, \quad \theta_{\text{min}} = \frac{1 - \sqrt{\beta + \gamma - \beta\gamma}}{1 - \beta}. \tag{15}$$

The quantities $\theta_{\text{max}}, \theta_{\text{min}}$ are the almost-sure limits of, respectively the largest and smallest nonzero singular values of $\widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{Y}$ in the pure-noise case, that is, when there are no spikes. The value σ_{thresh} is the smallest singular value of $\mathbf{X}$ that can be reliably detected as $p \to \infty$ in the case where $\mathbf{\Sigma} = \mathbf{I}_p$. We discuss this in more detail in Section 3.1 below.

On the ray $(\theta_{\text{max}}^2, \infty) \subset \mathbb{R}$, we define the *Stieltjes transform of Wachter's distribution* as follows:

$$\overline{s}(z) = \frac{1}{\gamma z} - \frac{1}{z} - \frac{\gamma\left(z(1-\beta) + (1-\gamma)\right) + 2\beta z - \gamma\sqrt{\left(z(1-\beta) + (1-\gamma)\right)^2 - 4z}}{2\gamma z(\gamma + \beta z)}. \tag{16}$$

We also define the *associated Stieltjes transform of Wachter's distribution* as follows:

$$\underline{s}(z) = -\frac{\gamma\left(z(1-\beta) + (1-\gamma)\right) + 2\beta z - \gamma\sqrt{\left(z(1-\beta) + (1-\gamma)\right)^2 - 4z}}{2z(\gamma + \beta z)}. \tag{17}$$

We also define a related function on the same domain $(\theta_{\max}^2, \infty)$:

$$\zeta(z) = -\frac{z(1-\beta) - (1-\gamma) - \sqrt{((1-\beta)z + (1-\gamma))^2 - 4z}}{2(\gamma + \beta - \gamma\beta)z}\,. \tag{18}$$

It is also convenient to define

$$\psi(z) = z \cdot \underline{s}(z) \cdot \zeta(z) = \frac{(1-\beta)z - 1 - \gamma - \sqrt{(z(1-\beta) + (1-\gamma))^2 - 4z}}{2(\beta z + \gamma)}\,, \tag{19}$$

and the function

$$\varphi(z) = z \cdot |\underline{s}(z)| \cdot (\overline{s}(z))^2\,. \tag{20}$$

We also define the *spike-forward* map $\Xi : [\sigma_{\text{thresh}}, \infty) \to [\theta_{\max}, \infty)$ by

$$\Xi(\sigma) = \sqrt{\frac{(1+\sigma^2)(\gamma+\sigma^2)}{(1-\beta)\sigma^2 - \beta}}\,. \tag{21}$$

Note that $\Xi(\cdot)$ is strictly increasing and smooth. Its inverse $\Xi^{-1} : [\theta_{\max}, \infty) \to [\sigma_{\text{thresh}}, \infty)$ is

$$\Xi^{-1}(\theta) = \frac{1}{\sqrt{\psi(\theta^2)}}\,, \tag{22}$$

where $\psi(z)$ is defined in (19).

Let $\mathsf{m}_\beta(\cdot)$ be the Stieltjes transform of the Marchenko-Pastur law with shape parameter β. When $0 < z < (1-\sqrt{\beta})^2$, namely it is smaller than the left edge edge of the Marchenko-Pastur bulk, it is given by the following formula, cf. [6, Lemma 3.11]:

$$\mathsf{m}_\beta(z) = \frac{1-\beta-z-\sqrt{(z-1-\beta)^2 - 4\beta}}{2\beta z}\,. \tag{23}$$

Denote the functions $\Upsilon_1(\cdot), \Upsilon_2(\cdot)$, $z \in (\theta_{\max}^2, \infty)$,

$$\Upsilon_1(z) = \frac{1}{z}\left[1 - \underline{s}(z) + (\underline{s}(z))^2 \mathsf{m}_\beta(-\underline{s}(z))\right] \tag{24}$$

$$\Upsilon_2(z) = \frac{1 + \gamma(\zeta(z) + z\zeta'(z))}{z^2}\left[1 - 2\underline{s}(z)\mathsf{m}_\beta(-\underline{s}(z)) + (\underline{s}(z))^2 \mathsf{m}'_\beta(-\underline{s}(z))\right]\,, \tag{25}$$

where $\mathsf{m}'_\beta(\cdot)$ is the derivative of $\mathsf{m}_\beta(\cdot)$. One can verify that whenever $z \in (\theta_{\max}^2, \infty)$, one has $0 < -\underline{s}(z) < (1-\sqrt{\beta})^2$ and so (23) is indeed applicable. We show this explicitly in Lemma 10, Appendix C.

Finally, define the function

$$E(z) = -\gamma|\zeta(z)| \cdot \left(\Upsilon_1(z) + z\Upsilon_1'(z)\right) + z \cdot |\underline{s}(z)| \cdot \left(\Upsilon_2(z) - \overline{s}(z)^2\right)\,. \tag{26}$$

2.4 Detailed description of the estimator

We now give the exact details of our proposed estimator, including estimates of the formulas for the optimal choice of $\eta_1, \ldots, \eta_r$. This process will be derived formally in Section 3.2; we present it now for the reader's convenience.

Inputs:

- The data matrix $\mathbf{Y} \in \mathbb{R}^{p \times n}$ to be denoised.
- $m > p$ samples of pure noise $\mathbf{R}' \in \mathbb{R}^{p \times m}$.
- A small parameter $\varepsilon > 0$.

Symbol	Description
$\overline{s}$	Stieltjes transform
$\underline{s}$	Associated Stieltjes transform
ζ	Resolvent-like function
ψ	$\psi(z) = z \cdot \underline{s}(z) \cdot \zeta(z)$
φ	$\varphi(z) = z \cdot \lvert\underline{s}(z)\rvert \cdot (\overline{s}(z))^2$
Ξ	Spike-forward map
Ξ^{-1}	Spike-backward map
Υ_1, Υ_2, E	Auxiliary functions

Table 2: Functions used in this paper.

Symbol(s)	Description
σ_{thresh}	Detection threshold
$\theta_{\min}$	Left bulk edge
$\theta_{\max}$	Right bulk edge
$\widehat{\theta}_1, \ldots, \widehat{\theta}_r$	Singular values of $\mathbf{Y}^w$
$\sigma_1, \ldots, \sigma_r$	Singular values of $\mathbf{X}$
$c_1, \ldots, c_r$	Left weighted inner products
$\underline{c}_1, \ldots, \underline{c}_r$	Right inner products
$\tau_1, \ldots, \tau_r$	Signal/noise alignments
μ	Noise covariance trace
p	Dimensionality
n	Number of signal-plus-noise samples
m	Number of noise-only samples
γ	Signal-plus-noise aspect ratio p/n
β	Noise-only aspect ratio p/m
$\eta_1, \ldots, \eta_r$	Optimal generalized singular values
$t_1, \ldots, t_r$	Optimal approximate singular values

Table 3: Parameters used in this paper.

Algorithm:

1. Form the sample covariance of pure noise samples $\widehat{\mathbf{\Sigma}} = \mathbf{R}'\mathbf{R}'^{\top}/m$.

2. Estimate the normalized trace of the noise covariance:

$$\widehat{\mu} = \frac{1}{p}\operatorname{tr}\widehat{\mathbf{\Sigma}}. \tag{27}$$

3. Form the matrix $\mathbf{Y}^w = \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{Y}$ and take its SVD,

$$\mathbf{Y}^w = \sum_{i=1}^{\min(n,p)} \widehat{\theta}_i \widehat{\mathbf{u}}_i^w (\widehat{\mathbf{v}}_i^w)^{\top}. \tag{28}$$

4. Estimate the "effective" signal rank[3]

$$\widehat{r} = \max\left\{1 \le i \le n \,:\, \widehat{\theta}_i > \theta_{\max} + \varepsilon\right\}, \tag{29}$$

   (if the set is empty, set $\widehat{r} = 0$), where $\theta_{\max}$ appears in (15).

5. Estimate the parameters $\tau_1, \ldots, \tau_{\widehat{r}}$ via

$$\widehat{\tau}_j = \frac{\varphi(\widehat{\theta}_j^2)}{|\psi'(\widehat{\theta}_j^2)|\|\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_j\|^2 - \widehat{\mu}E(\widehat{\theta}_j^2)}, \tag{30}$$

6. For $1 \le k \le \widehat{r}$, estimate the following parameters:

$$\widehat{\sigma}_k = \frac{1}{\sqrt{\widehat{\tau}_k}}\Xi^{-1}(\widehat{\theta}_k), \quad \widehat{c}_k = \sqrt{\frac{1}{\widehat{\tau}_k}\cdot\frac{\varphi(\widehat{\theta}_k^2)}{\psi'(\widehat{\theta}_k^2)}}, \quad \widehat{\underline{c}}_k = \sqrt{\underline{s}(\widehat{\theta}_k^2)\cdot\frac{\psi(\widehat{\theta}_k^2)}{\psi'(\widehat{\theta}_k^2)}}. \tag{31}$$

7. Define

$$\widehat{\eta}_k = \frac{\widehat{\sigma}_k\widehat{c}_k\widehat{\underline{c}}_k}{\|\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w\|^2}. \tag{32}$$

   Return the estimator

$$\widehat{\mathbf{X}} = \sum_{k=1}^{\widehat{r}} \widehat{\eta}_k \widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w(\widehat{\mathbf{v}}_k^w)^{\top}, \tag{33}$$

   and the estimated AMSE:

$$\widehat{\mathrm{AMSE}} = \sum_{k=1}^{r}\widehat{\sigma}_k^2\left(1 - \widehat{c}_k^2\widehat{\underline{c}}_k^2\frac{1}{\|\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w\|^2}\right). \tag{34}$$

3 Main results

In this section, we provide the mathematical results that justify the algorithm in Section 2.4. Specifically, we provide formulas for the quantities necessary to estimate the asymptotically optimal coefficients $\eta_1, \ldots, \eta_{\widehat{r}}$.

[3]It will shown later that for any fixed $\varepsilon > 0$, we have $\widehat{r}(\varepsilon) \le r$ asymptotically almost surely; furthermore, for ε sufficiently small, $\widehat{r}(\varepsilon)$ will be exactly the number of spikes r^* whose intensity is strong enough so as to not be "swallowed" completely by the noise.

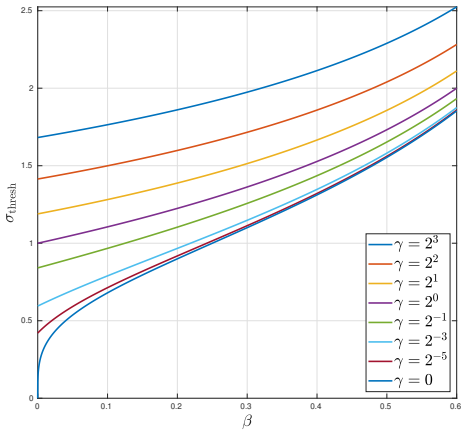

Figure 1: The value of σ_{thresh} as a function of β, for different values of γ.

3.1 The limiting spectrum of a spiked F-matrix

As mentioned before, the random matrix $\mathbf{Y}^w$ is an instance of a spiked F-matrix. The results of this section quantify the relevant phenomenology surrounding these matrices, namely we: 1) compute the spike detection threshold, and a formula for the singular value displacement (spike-forward) map; 2) compute the limiting cosines between the population and empirical PCs. These are the two components necessary to derive the optimal shrinkage rule, as presented above; we do this in Section 3.2.

We start with a limiting formula for the singular values of $\mathbf{Y}^w$:

Theorem 1. *Set* $r^* = \max\left\{1 \le k \le r \,:\, \sqrt{\tau_k}\sigma_k > \sigma_{\text{thresh}}\right\}$. *Then, for any fixed* k, *almost surely,*

$$\lim_{p\to\infty} \widehat{\theta}_k = y_k := \begin{cases} \Xi(\sqrt{\tau_k}\sigma_k) & \text{if } 1 \le k \le r^* \\ \theta_{\max} & \text{if } k > r^* \end{cases} . \tag{35}$$

Theorem 1 identifies a detectability phase-transition: a spike can be identified consistently by looking at the leading empirical singular value (as an outlier, separated from the "main bulk" of other singular values) whenever its "effective intensity" $\sqrt{\tau_i}\sigma_i$ exceeds the threshold σ_{thresh}.[4] When this is the case, this effective intensity can in fact be estimated consistently from the data as $\Xi^{-1}(\widehat{\theta}_i)$. Figure 1 plots the detection threshold σ_{thresh} as a function of β, for different values of γ.

The result in Theorem 1 is not new. To our knowledge, it first appeared in [48], where it is stated for $\mathbf{\Sigma} = \mathbf{I}$; an extension for the non-white case is straightforward. More sophisticated results have since appeared in the literature. For a spike above the detectability threshold, assuming $\mathbf{\Sigma} = \mathbf{I}$ and Gaussian $\mathbf{Z}, \mathbf{Z}'$, [18] proved a CLT for the empirical singular values: $\{\sqrt{p}(\widehat{\theta}_k - y_k)\}_{i\le r^*}$ converge jointly to a centered multivariate Gaussian (with an explicitly given covariance matrix). This result was extended by [62], which only assumed $\mathbf{Z}, \mathbf{Z}'$ with independent entries and finite 4-th moment. Another related work is [34] which studied, assuming Gaussian noise, the distribution of the leading empirical singular values in the non-asymptotic (finite n), high SNR regime ($\sigma \to \infty$) by perturbation theory methods. A follow-up work [19] extended the analysis for complex Gaussian matrices. Lastly, [63] studied the leading eigenvalues in a model where the numbers of spikes r is divergent (but not too large compared to n, p, m). Although Theorem 1 not new, in Section 5.1 we will nonetheless provide a self-contained proof, which shall also function as an important preparatory step towards proving Theorem 2 below.

[4] Note that Theorem 1 does not imply that detection is impossible below the threshold. Detection in this so-called "sub-critical regime" was studied in [33], and their results imply that consistent detection is indeed impossible (assuming $\mathbf{\Sigma} = \mathbf{I}$ and uniformly random signal spikes).

Lastly, Theorem 1 readily justifies the rank estimation procedure described in (29). It implies that for any fixed small enough ε, specifically, $0 < \varepsilon < \Xi(\sqrt{\tau_{r^*}}\sigma_{r^*}) - \theta_{\max}$, one has $\lim_{p\to\infty} \widehat{r}(\varepsilon) = r^*$ almost surely. We remark in passing that one may in fact choose $\varepsilon = o(1)$, so to always ensure (asymptotically) consistent rank estimation. Indeed, when there is no signal ($\mathbf{X} = \mathbf{0}$) the stochastic fluctuations $n^{2/3}(\lambda_1(n^{-1}\mathbf{Y}\mathbf{Y}^\top) - \theta_{\max})$ converge in distribution to a Tracy-Widom law [27,32]. In particular, in the presence of spikes, one may verify (e.g. by singular value interlacing) that $\lambda_{r+1}(n^{-1}\mathbf{Y}^\top\mathbf{Y}) = \theta_{\max} + O_{\mathbb{P}}(n^{-2/3})$. Consequently, for any vanishing $\varepsilon = \omega(n^{-2/3})$, a.s. $r^* \le \liminf_{p\to\infty} r^*(\varepsilon) \le \limsup_{p\to\infty} r^*(\varepsilon) \le r$.

The next theorem calculates the limiting cosines between the population signal spikes and their empirical counterparts:

Theorem 2. *Let $1 \le k \le r$, and set $y_k = \Xi(\sqrt{\tau_k}\sigma_k)$. Suppose $\sqrt{\tau_k}\sigma_k > \sigma_{\text{thresh}}$. Then*

$$c_k^2 \equiv \lim_{p\to\infty} \left(\mathbf{u}_k^\top \widehat{\boldsymbol{\Sigma}}^{1/2} \widehat{\mathbf{u}}_k^w\right)^2 = \frac{1}{\tau_k} \cdot \frac{y_k^2 \left(\overline{s}(y_k^2)\right)^2 \underline{s}(y_k^2)}{\psi'(y_k^2)}, \tag{36}$$

$$\underline{c}_k^2 \equiv \lim_{p\to\infty} \left(\mathbf{v}_k^\top \widehat{\mathbf{v}}_k\right)^2 = \underline{s}(y_k^2) \cdot \frac{\psi(y_k^2)}{\psi'(y_k^2)}, \tag{37}$$

$$\lim_{p\to\infty} \left(\mathbf{u}_k^\top \widehat{\boldsymbol{\Sigma}}^{1/2} \widehat{\mathbf{u}}_k^w\right) \cdot \left(\mathbf{v}_k^\top \widehat{\mathbf{v}}_k\right) = -\frac{1}{\sqrt{\tau_k}} \cdot \frac{y_k \overline{s}(y_k^2) \underline{s}(y_k^2) \cdot \sqrt{\psi(y_k^2)}}{\psi'(y_k^2)} \ge 0, \tag{38}$$

$$\lim_{p\to\infty} \|\widehat{\boldsymbol{\Sigma}}^{1/2} \widehat{\mathbf{u}}_k^w\|^2 = \frac{1}{|\psi'(y_k^2)|} \left[\mu \cdot E(y_k^2) + \frac{1}{\tau_k}\varphi(y_k^2)\right], \tag{39}$$

and

$$\lim_{p\to\infty} \left(\mathbf{u}_k^\top \widehat{\mathbf{u}}_k\right)^2 = \frac{\varphi(y_k^2)}{\tau_k \mu \cdot E(y_k^2) + \varphi(y_k^2)}, \tag{40}$$

where the limits hold almost surely. Moreover, if $1 \le k, l \le r$ and $k \ne l$, then

$$\lim_{p\to\infty} \left(\mathbf{u}_k^\top \widehat{\boldsymbol{\Sigma}}^{1/2} \widehat{\mathbf{u}}_l^w\right)^2 = \lim_{p\to\infty} \left(\mathbf{v}_k^\top \widehat{\mathbf{v}}_l\right)^2 = 0. \tag{41}$$

The proof of Theorem 2 is found in Section 5.2.

3.2 Derivation of the optimal shrinker

Recall the form of the estimators (9) and (13). Equipped with Theorems 1 and 2, it is straightforward to evaluate the optimal weights $\eta_1, \dots, \eta_r$ in terms of the population parameters.

Theorem 3. *Let $y_k = \Xi(\sqrt{\tau_k}\sigma_k)$, and suppose $\sqrt{\tau_k}\sigma_k > \sigma_{\text{thresh}}$, $1 \le k \le r$. Then the optimal $\boldsymbol{\eta} = (\eta_1, \dots, \eta_r)$ that minimizes* $\text{AMSE}(\boldsymbol{\eta})$ *is given by the following:*

$$\eta_k = \sigma_k \cdot c_k \cdot \underline{c}_k \cdot |\psi'(y_k^2)| \left[\mu \cdot E(y_k) + \frac{1}{\tau_k}\varphi(y_k)\right]^{-1}, \tag{42}$$

where c_k and $\underline{c}_k$ are defined in (36) *and* (37)*, respectively. The optimal $t_1, \dots, t_k$ are given by*

$$t_k = \sigma_k \cdot c_k \cdot \underline{c}_k \cdot |\psi'(y_k^2)| \left[\mu \cdot E(y_k) + \frac{1}{\tau_k}\varphi(y_k)\right]^{-1/2}. \tag{43}$$

The AMSE at these optimal values is equal to

$$\text{AMSE}(\boldsymbol{\eta}) = \sum_{k=1}^r \sigma_k^2 \left(1 - c_k^2 \underline{c}_k^2 |\psi'(y_k^2)| \left[\mu \cdot E(y_k) + \frac{1}{\tau_k}\varphi(y_k)\right]^{-1}\right). \tag{44}$$

Proof. Using the asymptotics in Theorem 2, the mean squared error may be written as:

$$\|\widehat{\mathbf{X}} - \mathbf{X}\|_F^2 = \sum_{k=1}^{r^*} \left\{ \eta_k^2 \|\widehat{\mathbf{\Sigma}}^{1/2} \widehat{\mathbf{u}}_k^w\|^2 - 2\eta_k \sigma_k (\mathbf{u}_k^\top \widehat{\mathbf{\Sigma}}^{1/2} \widehat{\mathbf{u}}_k^w)(\mathbf{v}_k^\top \widehat{\mathbf{v}}_k) \right\} + \sum_{k=1}^{r} \sigma_k^2 - \sum_{j \neq k} 2\eta_j \sigma_k (\mathbf{u}_k^\top \widehat{\mathbf{\Sigma}}^{1/2} \widehat{\mathbf{u}}_j^w)(\mathbf{v}_k^\top \widehat{\mathbf{v}}_j)$$

$$\simeq \sum_{k=1}^{r^*} \left\{ \eta_k^2 \frac{1}{|\psi'(y_k^2)|} \left[\mu \cdot E(y_k^2) + \frac{1}{\tau_k} \varphi(y_k^2) \right] - 2\eta_k \sigma_k c_k \underline{c}_k \right\} + \sum_{k=1}^{r} \sigma_k^2, \tag{45}$$

and minimizing this gives the desired expression for η_k. The optimal t_k follow from $t_k = \eta_k \cdot \|\widehat{\mathbf{\Sigma}}^{1/2} \widehat{\mathbf{u}}_k^w\|$, and (39). The formula for the AMSE is obtained substituting η_k into the expression for the AMSE in (45). □

The optimal η_k may be consistently estimated using only the observed matrices $\mathbf{Y}$ and $\mathbf{R}'$. Indeed, suppose that $\widehat{\theta} > \theta_{\max}$. Then from Theorems 1 and 2, the values τ_k may be estimated via

$$\widehat{\tau}_k = \frac{\varphi(\widehat{\theta}_k^2)}{|\psi'(\widehat{\theta}_k^2)| \|\widehat{\mathbf{\Sigma}}^{1/2} \widehat{\mathbf{u}}_k\|^2 - \widehat{\mu} E(\widehat{\theta}_k^2)}, \quad 1 \leq k \leq r. \tag{46}$$

while the population intensities σ_k may be consistently estimated as

$$\widehat{\sigma}_k = \frac{1}{\sqrt{\widehat{\tau}_k}} \Xi^{-1}(\widehat{\theta}_k), \quad 1 \leq k \leq r. \tag{47}$$

With these estimates, using Theorem 2 c_k and $\underline{c}_k$ may be estimated as follows:

$$\widehat{c}_k = \sqrt{\frac{1}{\widehat{\tau}_k} \cdot \frac{\widehat{\theta}_k^2 \left(\overline{s}(\widehat{\theta}_k^2) \right)^2 \underline{s}(\widehat{\theta}_k^2)}{\psi'(\widehat{\theta}_k^2)}}, \qquad \widehat{\underline{c}}_k = \sqrt{\underline{s}(\widehat{\theta}_k^2) \cdot \frac{\psi(\widehat{\theta}_k^2)}{\psi'(\widehat{\theta}_k^2)}}. \tag{48}$$

Finally, the factor

$$|\psi'(y_k^2)| \left[\mu \cdot E(y_k) + \frac{1}{\tau_k} \varphi(y_k) \right]^{-1} \tag{49}$$

may be estimated as $\|\widehat{\mathbf{\Sigma}}^{1/2} \widehat{\mathbf{u}}_k^w\|^{-2}$. Putting these together, we estimate the optimal η_k as

$$\widehat{\eta}_k = \widehat{\sigma}_k \cdot \widehat{c}_k \cdot \widehat{\underline{c}}_k \,. \tag{50}$$

This completes the derivation of the optimal $\widehat{\mathbf{X}}$.

4 Background on F-Matrices

Before moving on to the proofs of our main results, we briefly review several known result about the spectrum of the matrix $\mathbf{Y}^w$ in the absence of a signal. In other words, we consider the pseudo-whitened noise matrix $\mathbf{N} \in \mathbb{R}^{p \times n}$

$$\mathbf{N} = \frac{1}{\sqrt{n}} \widehat{\mathbf{\Sigma}}^{-1/2} \mathbf{R} = \frac{1}{\sqrt{n}} \widehat{\mathbf{\Sigma}}^{-1/2} \mathbf{\Sigma}^{1/2} \mathbf{Z} \,. \tag{51}$$

Define the following Wishart matrices:

$$\mathbf{E} = \mathbf{Z}\mathbf{Z}^\top / n, \quad \mathbf{S} = \mathbf{Z}'\mathbf{Z}'^\top / m, \tag{52}$$

Importantly, observe that the singular values of $\mathbf{N}$ do not depend on the population covariance $\mathbf{\Sigma}$. To see this, write $\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma}^{1/2} \mathbf{S} \mathbf{\Sigma}^{1/2}$, and recall that the squared singular values of $\mathbf{N}$ are

the (non-zero) eigenvalues of $\mathbf{N}^\top\mathbf{N} = \mathbf{Z}^\top\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^{1/2}\mathbf{Z} = \mathbf{Z}^\top\mathbf{S}^{-1}\mathbf{Z}$, which does not depend on $\boldsymbol{\Sigma}$. Thus, when studying the singular values of $\mathbf{N}$ we may assume w.l.o.g. that $\boldsymbol{\Sigma} = \mathbf{I}$, as we shall do throughout this section. The non-zero eigenvalues of $\mathbf{N}^\top\mathbf{N} \in \mathbb{R}^{n\times n}$ and $\mathbf{N}\mathbf{N}^\top \in \mathbb{R}^{p\times p}$ are the same. With $\mathbf{E}$ defined as in (52), we write $\mathbf{N}\mathbf{N}^\top = \mathbf{S}^{-1/2}\mathbf{E}\mathbf{S}^{-1/2}$. Applying a similarity transformation, its eigenvalues are identical to those of the matrix

$$\mathbf{F} = \mathbf{S}^{-1/2}\mathbf{E}\mathbf{S}^{1/2} = \mathbf{S}^{-1}\mathbf{E}\,. \tag{53}$$

The random matrix ensemble (53) is known as an F-matrix/Fisher matrix; see, for example, [6]. It is closely related to the multivariate Beta ensemble, which features in classical multivariate statistics, in particular Multivariate Analysis of Variance (MANOVA) [46].[5]

We briefly recall some definitions. For a diagonalizable matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$, its empirical spectral distribution (ESD) is the counting measure of the eigenvalues: $\mu_{\mathbf{A}} := n^{-1}\sum_{i=1}^{p}\delta_{\lambda_i(\mathbf{A})}$. For a probability measure μ on $\mathbb{R}$, its Stieltjes transform $s_\mu(\cdot)$ is the function

$$s_\mu(z) = \int_{-\infty}^{\infty}\frac{1}{\lambda - z}d\mu(z)\,, \quad \text{where} \quad z \in \mathbb{C}\setminus\mathbb{R}\,. \tag{54}$$

The limiting empirical spectral distribution (LESD) for F-matrices was first derived by Wachter [61] with subsequent generalizations in [53, 54, 64]; for a textbook reference, see [6, Theorem 4.10].

Theorem 4 (Wachter's distribution)**.** *Let $\theta_{\min}, \theta_{\max}$ be as in (15). The ESD of the F-matrix (53) converges weakly almost surely to a deterministic distribution. When $\gamma \le 1$, the LESD is continuous, supported on $[\theta_{\min}, \theta_{\max}]$ and has the density,*

$$F_{\gamma,\beta}(\lambda) = \frac{(1-\beta)\sqrt{(\theta_{\max}^2 - \lambda)(\lambda - \theta_{\min}^2)}}{2\pi\lambda(\gamma + \beta\lambda)}\mathbb{1}\{\theta_{\min}^2 \le \lambda \le \theta_{\max}^2\}\,. \tag{55}$$

When $\gamma > 1$ the LESD has a continuous density (55), but also an atom at $\lambda = 0$ with weight $1 - \gamma^{-1}$.

Note that when $\beta = 0$, (55) collapses to the well-known Marchenko-Pastur law with shape γ.

Theorem 55 implies, in particular, that all but a vanishing fraction of the non-zero eigenvalues of $\mathbf{F}$ are contained in $[\theta_{\min}^2, \theta_{\max}^2]$. The next result, which follows from [5, Theorem 1.1], implies that in fact, asymptotically almost surely there are no eigenvalues outside the support:

Theorem 5 (Extreme eigenvalues of F-matrix)**.** *Almost surely,*

$$\lim_{p\to\infty}\lambda_{\min}(\mathbf{F}) = \theta_{\min}^2, \quad \lim_{p\to\infty}\lambda_{\max}(\mathbf{F}) = \theta_{\max}^2\,.$$

In our proofs, we shall use a well-known formula for the Stieltjes transform of Wachter's law. The following appears, for example, in [6, Theorem 4.10].

Proposition 1 (Stieltjes transform of Wachter's law)**.** *The Stieltjes transform of Wachter's law is given by (16). Furthermore, we have the following convergence for the empirical Stieltjes transform:*

$$\begin{aligned}
&p^{-1}\mathrm{tr}(\mathbf{N}\mathbf{N}^\top - z\mathbf{I})^{-1} \longrightarrow \overline{s}(z),\\
&\frac{d}{dz}p^{-1}\mathrm{tr}(\mathbf{N}\mathbf{N}^\top - z\mathbf{I})^{-1} = p^{-1}\mathrm{tr}(\mathbf{N}\mathbf{N}^\top - z\mathbf{I})^{-2} \longrightarrow \overline{s}'(z)\,,
\end{aligned}$$

[5] One can show that the eigenvalues of $\mathbf{F}$ and the MANOVA matrix $(\mathbf{S}+\mathbf{E})^{-1/2}\mathbf{E}(\mathbf{S}+\mathbf{E})^{-1/2}$ are bijectively mapped to one another via the mapping $z \mapsto z/(1+z)$.

for all $z \in (\theta_{\max}^2, \infty)$. Above, convergence is a.s. and uniform on compact subintervals.

Furthermore,

$$n^{-1}\mathrm{tr}(\mathbf{N}^\top\mathbf{N} - z\mathbf{I})^{-1} = n^{-1}\left[\mathrm{tr}(\mathbf{N}\mathbf{N}^\top - z\mathbf{I})^{-1} + (n-p)\frac{1}{z}\right] \longrightarrow \underline{s}(z),$$
$$\frac{d}{dz}n^{-1}\mathrm{tr}(\mathbf{N}^\top\mathbf{N} - z\mathbf{I})^{-1} = n^{-1}\mathrm{tr}(\mathbf{N}^\top\mathbf{N} - z\mathbf{I})^{-2} \longrightarrow \underline{s}'(z)\,.$$

We also need a limiting expression for the following resolvent-like expression from [18, Lemma 1]:

Proposition 2. *Let $\zeta(z)$ be as in (18). Then*

$$p^{-1}\,\mathrm{tr}\,(\mathbf{E} - z\mathbf{S})^{-1} \longrightarrow \zeta(z),$$
$$\frac{d}{dz}p^{-1}\,\mathrm{tr}\,(\mathbf{E} - z\mathbf{S})^{-1} = p^{-1}\,\mathrm{tr}\,\mathbf{S}\,(\mathbf{E} - z\mathbf{S})^{-2} \longrightarrow \zeta'(z)\,,$$

for all $z \in (\theta_{\max}^2, \infty)$. Above, convergence is a.s. and uniform on compact subintervals.

Notation. For (possibly random) sequences $a_p, b_p \in \mathbb{R}$, we write $a_p \simeq b_p$ to indicate $a_p - b_p \to 0$ a.s. For a sequence of matrices $\mathbf{M}_p \in \mathbb{R}^{p\times p}$, we denote $\overline{\mathrm{tr}}(\mathbf{M}_p) := \lim_{p\to\infty} p^{-1}\mathrm{tr}(\mathbf{M}_p)$, where it is understood that the limit exists a.s.

5 Proofs of main results

5.1 Proof of Theorem 1

We study the leading eigenvalues of the pseudo-whitened data matrix (8),

$$\mathbf{Y}^w = \mathbf{P} + \mathbf{N}\,, \tag{56}$$

where $\mathbf{P} = \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{U}\mathbf{\Lambda}\mathbf{V}$ is the pseudo-whitened signal matrix, and $\mathbf{N}$ is the pseudo-whitened noise matrix (51). Since $\mathrm{rank}(\mathbf{P}) = r$, by Weyl's interlacing inequalities, for every i,

$$\sigma_{i+r+1}(\mathbf{Y}^w) \le \sigma_{i+1}(\mathbf{N}) + \sigma_{r+1}(\mathbf{P}) = \sigma_{i+1}(\mathbf{N})\,, \quad \sigma_{(r+i)+r+1}(\mathbf{N}) \le \sigma_{r+i+1}(\mathbf{Y}^w) + \sigma_{r+1}(-\mathbf{P}) = \sigma_{r+i+1}(\mathbf{Y}^w)\,.$$

Thus, for any i, $\widehat{\theta}_{i+r+1} := \sigma_{i+r+1}(\mathbf{Y})$ satisfies $\sigma_{2r+i+1}(\mathbf{N}) \le \sigma_{i+r+1}(\mathbf{Y}) \le \sigma_{i+1}(\mathbf{N})$. By Theorems 4 and 5, $\sigma_{2r+i+1}(\mathbf{N}), \sigma_{i+1}(\mathbf{N}) \to \theta_{\max}$ a.s. Consequently, for fixed $k \ge r+1$, $\widehat{\theta}_k \to \theta_{\max}$.

It remains to check whether $\mathbf{Y}^w$ has singular values which are asymptotically larger than $\theta_{\max}$. Note that, appealing to Theorem 5, such singular values necessarily cannot be singular values of the noise matrix $\mathbf{N}$. By [11, Lemma 4.1], the singular values of $\mathbf{Y}^w$ which are not singular values of $\mathbf{N}$ are exactly (with multiplicities) the solutions to $\det(\widehat{\mathbf{M}}(y)) = 0$, where $\widehat{\mathbf{M}}(y)$ is the $2r$-by-$2r$ symmetric matrix

$$\widehat{\mathbf{M}}(y) = \begin{bmatrix} y\cdot\mathbf{U}^\top\widehat{\mathbf{\Sigma}}^{-1/2}\left(y^2\mathbf{I}_p - \mathbf{N}\mathbf{N}^\top\right)^{-1}\widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{U} & \mathbf{U}^\top\widehat{\mathbf{\Sigma}}^{-1/2}\left(y^2\mathbf{I}_p - \mathbf{N}\mathbf{N}^\top\right)^{-1}\mathbf{N}\mathbf{V} \\ \mathbf{V}^\top\mathbf{N}^\top\left(y^2\mathbf{I}_p - \mathbf{N}\mathbf{N}^\top\right)^{-1}\widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{U} & y\cdot\mathbf{V}^\top\left(y^2\mathbf{I}_n - \mathbf{N}^\top\mathbf{N}\right)^{-1}\mathbf{V} \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{\Lambda}^{-1} \\ \mathbf{\Lambda}^{-1} & 0 \end{bmatrix}. \tag{57}$$

Define the matrix $\boldsymbol{\mathcal{T}} = \mathrm{diag}(\tau_1, \ldots, \tau_r) \in \mathbb{R}^{r\times r}$. The next lemma asserts that the random matrix $\widehat{\mathbf{M}}(y)$ converges a.s. to a deterministic limit as $p \to \infty$:

Lemma 1. *Suppose that $y \notin \mathbb{C} \setminus [\theta_{\min}, \theta_{\max}]$. Then a.s., $\widehat{\mathbf{M}}(y) \longrightarrow \mathbf{M}(y)$, where*

$$\mathbf{M}(y) = \begin{bmatrix} -y\zeta(y^2)\cdot\boldsymbol{\mathcal{T}} & \mathbf{0} \\ \mathbf{0} & -y\underline{s}(y^2)\cdot\mathbf{I} \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{\Lambda}^{-1} \\ \mathbf{\Lambda}^{-1} & 0 \end{bmatrix}. \tag{58}$$

Moreover, convergence (of each entry) is uniform on compact subsets in $y \notin \mathbb{C} \setminus [\theta_{\min}, \theta_{\max}]$.

Proof. We start with the off-diagonal elements of (57). First, note that the matrix $\mathbf{N}$ is invariant to multiplication by $O(n)$ from the right. Consequently, if $\mathcal{O} \sim \mathrm{Haar}(O(n))$, then $\mathbf{U}^\top \widehat{\mathbf{\Sigma}}^{-1/2} \left(y^2 \mathbf{I}_p - \mathbf{N}\mathbf{N}^\top\right)^{-1} \mathbf{N}\mathbf{V} \stackrel{d}{=} \mathbf{U}^\top \widehat{\mathbf{\Sigma}}^{-1/2} \left(y^2 \mathbf{I}_p - \mathbf{N}\mathbf{N}^\top\right)^{-1} \mathbf{N}\mathcal{O}\mathbf{V} \simeq \mathbf{0} \in \mathbb{R}^{r\times r}$. Now, considering the top left block and using (6),

$$\begin{aligned}\left(\mathbf{U}^\top \widehat{\mathbf{\Sigma}}^{-1/2} \left(y^2 \mathbf{I}_p - \mathbf{N}\mathbf{N}^\top\right)^{-1} \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{U}\right)_{j,k} &= p^{-1}\mathbf{w}_j^\top \mathbf{D}_j^\top \widehat{\mathbf{\Sigma}}^{-1/2} \left(y^2 \mathbf{I}_p - \mathbf{N}\mathbf{N}^\top\right)^{-1} \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{D}_k \mathbf{w}_k \\ &\stackrel{(\star)}{\simeq} p^{-1}\mathrm{tr}\left(\mathbf{D}_k^\top \widehat{\mathbf{\Sigma}}^{-1/2} \left(y^2 \mathbf{I}_p - \mathbf{N}\mathbf{N}^\top\right)^{-1} \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{D}_k\right) \mathbb{1}\{j=k\}\,,\end{aligned}$$

where $(\star)$ follows by the independence of $\mathbf{w}_j, \mathbf{w}_k$ and, e.g., the Hanson-Wright inequality. The term under the trace is $\widehat{\mathbf{\Sigma}}^{-1/2} \left(y^2 \mathbf{I}_p - \mathbf{N}\mathbf{N}^\top\right)^{-1} \widehat{\mathbf{\Sigma}}^{-1/2} = \left(y^2 \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^{1/2}\mathbf{N}\mathbf{N}^\top \widehat{\mathbf{\Sigma}}^{1/2}\right)^{-1} = \mathbf{\Sigma}^{-1/2} \left(y^2\mathbf{S} - \mathbf{E}\right)^{-1} \mathbf{\Sigma}^{-1/2}$. Crucially, the matrix $(y^2\mathbf{S} - \mathbf{E})^{-1}$ is orthogonally invariant, and a.s. has bounded operator norm, since $y > \sigma_{\mathrm{thresh}}$. Consequently, $(y^2\mathbf{S} - \mathbf{E})^{-1}$ and $\mathbf{\Sigma}^{-1/2}\mathbf{D}_k\mathbf{D}_k^\top\mathbf{\Sigma}^{-1/2}$ are asymptotically free (see Section A.2), and so

$$\begin{aligned}\overline{\mathrm{tr}}\left(\mathbf{D}_k^\top \widehat{\mathbf{\Sigma}}^{-1/2} \left(y^2 \mathbf{I}_p - \mathbf{N}\mathbf{N}^\top\right)^{-1} \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{D}_k\right) &= \overline{\mathrm{tr}}\left((y^2\mathbf{S} - \mathbf{E})^{-1}\mathbf{\Sigma}^{-1/2}\mathbf{D}_k\mathbf{D}_k^\top\mathbf{\Sigma}^{-1/2}\right) \\ &= \overline{\mathrm{tr}}\left((y^2\mathbf{S} - \mathbf{E})^{-1}\right)\overline{\mathrm{tr}}\left(\mathbf{\Sigma}^{-1/2}\mathbf{D}_k\mathbf{D}_k^\top\mathbf{\Sigma}^{-1/2}\right) \\ &= -\zeta(y^2)\cdot\tau_k\,, \end{aligned} \tag{59}$$

where we used Proposition 2 and (7). This establishes pointwise convergence; uniform convergence follows from the Arzela-Ascoli theorem, where both equicontinuity and uniform boundedness of the entries clearly hold, since y is bounded away from the support of Wachter's distribution, $[\theta_{\min}, \theta_{\max}]$. Finally, the bottom right block of (57) may be analyzed similarly, using Proposition 1. □

Since the operator norm of $\mathbf{Y}^w$ is a.s. bounded by a constant, applying an argument analogous to that in Lemma 6.1 from [10], we deduce that for each $1 \le k \le r$, either: 1) $\sigma_k(\mathbf{Y}^w) \to \theta_{\max}$ a.s.; or 2) $\sigma_k(\mathbf{Y}^w)$ tends a.s. to a root of the deterministic polynomial equation $\det(\mathbf{M}(y)) = 0$. Moreover, the roots of this equation match exactly (including their multiplicities) the limiting locations and counts of the outlying singular values of $\mathbf{Y}^w$. Now, one may verify that $\det(\mathbf{M}(y)) = 0$ if and only if for some $1 \le k \le r$,

$$\psi(y^2) := y^2 \cdot \zeta(y^2) \cdot \underline{s}(y^2) = 1/(\tau_k\sigma_k^2)\,, \tag{60}$$

where an explicit formula for $\psi(z)$ is given in (19). One can show that $\psi(\cdot)$ is strictly decreasing, and maps $(\theta_{\max}^2, \infty)$ bijectively to $(1/\sigma_{\mathrm{thresh}}^2, 0)$. Consequently, a solution to (60) exists if and only if $\sqrt{\tau_k}\sigma_k > \sigma_{\mathrm{thresh}}$, in other words, $k \le r^*$. In this case, one may also compute explicitly the functional inverse, so that (60) is equivalent to $y = \Xi(\sqrt{\tau_k}\sigma_k)$, where the spike-forward map $\Xi(\cdot)$ is given in (21). This concludes the proof of Theorem 1. □

5.2 Proof of Theorem 2

The first step of the proof consists of computing the limiting inner products $\langle \mathbf{u}_i, \widehat{\mathbf{\Sigma}}^{-1/2}\widehat{\mathbf{u}}_k^w\rangle$ and $\langle \mathbf{v}_i, \widehat{\mathbf{v}}_k\rangle$. This is an important intermediate step towards calculating $\langle \mathbf{u}_i, \widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w\rangle$ and $\|\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w\|$. To this end, define the vectors $\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k \in \mathbb{R}^r$, $1 \le k \le r$,

$$\boldsymbol{\alpha}_k = \mathbf{U}^\top \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{u}_k^w\,, \quad \boldsymbol{\beta}_k = \mathbf{V}^\top \widehat{\mathbf{v}}_k\,.$$

Lemma 2. *Let $1 \le k \le r$ be such that $\sqrt{\tau_k}\sigma_k > \sigma_{\mathrm{thresh}}$. For every $i \in [r] \setminus \{k\}$,*

$$\lim_{p\to\infty} (\boldsymbol{\alpha}_k)_i = 0, \quad \lim_{p\to\infty} (\boldsymbol{\beta}_k)_i = 0\,. \tag{61}$$

Proof. By [11, Lemma 5.1], the vector $\boldsymbol{x}_k = (\boldsymbol{\Lambda}\boldsymbol{\beta}_k, \boldsymbol{\Lambda}\boldsymbol{\alpha}_k) \in \mathbb{R}^{2r}$ is in the kernel of $\widehat{\mathbf{M}}(\widehat{\theta}_k)$, where $\widehat{\mathbf{M}}(\cdot)$ is defined in (57). Since $\boldsymbol{x}_k$ has bounded norm, and $\lim_{p\to\infty} \|\widehat{\mathbf{M}}(\widehat{\theta}_k) - \mathbf{M}(\mathbf{y}_k)\| = 0$ (by Lemma 1), we have $\lim_{p\to\infty} \mathbf{M}(y_k)\boldsymbol{x}_k = \mathbf{0}$ a.s. Considering only coordinates i and $i+r$, where $i \neq k$, yields the equation

$$\mathbf{0} \simeq \begin{bmatrix} y_k\zeta(y_k^2)\tau_i & 1/\sigma_i \\ 1/\sigma_i & y_k\underline{s}(y_k^2) \end{bmatrix} \begin{bmatrix} (\boldsymbol{x}_k)_i \\ (\boldsymbol{x}_k)_{i+r} \end{bmatrix}. \tag{62}$$

Observe that the matrix in (62) is invertible: its determinant is $\tau_i y_k^2 \zeta(y_k^2)\underline{s}(y_k^2) - 1/\sigma_i^2 = \tau_i\psi(y_k^2) - 1/\sigma_i^2 = \tau_i/(\tau_k\sigma_k^2) - 1/\sigma_i^2$, where we used (60); now recall that by assumption, $\tau_k\sigma_k^2 \neq \tau_i\sigma_i^2$, so the determinant is not zero. Thus, $(\boldsymbol{x}_k)_i, (\boldsymbol{x}_k)_{i+r} \simeq 0$, and the claim follows. □

By the definition of $\widehat{\mathbf{u}}_k^w, \widehat{\mathbf{v}}_k$ as singular vectors of $\mathbf{Y}^w$, we have $\mathbf{Y}^w\widehat{\mathbf{v}}_k = \widehat{\theta}_k\widehat{\mathbf{u}}_k^w$ and $\mathbf{Y}^{w\top}\widehat{\mathbf{u}}_k^w = \widehat{\theta}_k\widehat{\mathbf{v}}_k$. Decomposing $\mathbf{Y}^w = \widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top + \mathbf{N}$ yields $\widehat{\theta}_k\widehat{\mathbf{u}}_k = (\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top + \mathbf{N})\widehat{\mathbf{v}}_k = \widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\beta}_k + \mathbf{N}\widehat{\mathbf{v}}_k$. Rearranging, $\mathbf{N}\widehat{\mathbf{v}}_k = \widehat{\theta}_k\widehat{\mathbf{u}}_k^w - \widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\beta}_k$. Similarly, $\mathbf{N}^\top\widehat{\mathbf{u}}_k^w = \widehat{\theta}_k\widehat{\mathbf{v}}_k - \mathbf{V}\boldsymbol{\Lambda}\boldsymbol{\alpha}_k$. Now,

$$\begin{aligned}
\widehat{\theta}_k^2\widehat{\mathbf{v}}_k &= \mathbf{Y}^{w\top}\mathbf{Y}^w\widehat{\mathbf{v}}_k = (\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top + \mathbf{N})^\top(\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top + \mathbf{N})\widehat{\mathbf{v}}_k \\
&= (\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top + \mathbf{N})^\top(\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\beta}_k + \mathbf{N}\mathbf{v}_k) \\
&= \mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^\top\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\beta}_k + \mathbf{N}^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\beta}_k + \mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{N}\widehat{\mathbf{v}}_k) + \mathbf{N}^\top\mathbf{N}\widehat{\mathbf{v}}_k \\
&= \mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^\top\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\beta}_k + \mathbf{N}^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\beta}_k + \mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}(\widehat{\theta}_k\widehat{\mathbf{u}}_k^w - \widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\beta}_k) + \mathbf{N}^\top\mathbf{N}\widehat{\mathbf{v}}_k \\
&= \mathbf{N}^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\beta}_k + \widehat{\theta}_k\mathbf{V}\boldsymbol{\Lambda}\boldsymbol{\alpha}_k + \mathbf{N}^\top\mathbf{N}\widehat{\mathbf{v}}_k\,.
\end{aligned}$$

When $\widehat{\theta}_k$ is an outlier, namely when, $\sqrt{\tau_k}\sigma_k > \sigma_{\text{thresh}}$, we have $\widehat{\theta}_k \simeq y_k := \Xi(\sqrt{\tau_k}\sigma_k)$ and moreover that $\widehat{\theta}_k^2\mathbf{I} - \mathbf{N}^\top\mathbf{N}$ is invertible. Thus, $\widehat{\mathbf{v}}_k = (\widehat{\theta}_k^2\mathbf{I} - \mathbf{N}^\top\mathbf{N})^{-1}\left(\mathbf{N}^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\beta}_k + \widehat{\theta}_k\mathbf{V}\boldsymbol{\Lambda}\boldsymbol{\alpha}_k\right)$. Finally, recall that by Lemma 2, $\boldsymbol{\alpha}_k \simeq (\boldsymbol{\alpha}_k)_k\boldsymbol{e}_k$, $\boldsymbol{\beta}_k \simeq (\boldsymbol{\beta}_k)_k\boldsymbol{e}_k$, where $\boldsymbol{e}_k$ is the k-th standard basis element. Note that one could repeat the same calculation above, starting with $\widehat{\theta}_k^2\widehat{\mathbf{u}}_k^w = \mathbf{Y}^w\mathbf{Y}^{w\top}\widehat{\mathbf{u}}_k$. We deduce the following formulas for the outlying singular vectors:

$$\begin{aligned}
\widehat{\mathbf{v}}_k &\simeq \sigma_k \cdot (\boldsymbol{\beta}_k)_k \cdot (y_k^2\mathbf{I} - \mathbf{N}^\top\mathbf{N})^{-1}\mathbf{N}^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{u}_k + \sigma_k y_k \cdot (\boldsymbol{\alpha}_k)_k \cdot (y_k^2\mathbf{I} - \mathbf{N}^\top\mathbf{N})^{-1}\mathbf{v}_k\,, \\
\widehat{\mathbf{u}}_k^w &\simeq \sigma_k \cdot (\boldsymbol{\alpha}_k)_k \cdot \left(y_k^2\mathbf{I} - \mathbf{N}\mathbf{N}^\top\right)^{-1}\mathbf{N}\mathbf{v}_k + \sigma_k y_k \cdot (\boldsymbol{\beta}_k)_k \cdot \left(y_k^2\mathbf{I} - \mathbf{N}\mathbf{N}^\top\right)^{-1}\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{u}_k\,.
\end{aligned} \tag{63}$$

Lemma 3. *Suppose that $\sqrt{\tau_k}\sigma_k > \sigma_{\text{thresh}}$, and set $y_k = \Xi(\sqrt{\tau_k}\sigma_k)$. Let*

$$\underline{c}_k^2 = \underline{s}(y_k^2) \cdot \frac{\psi(y_k^2)}{\psi'(y_k^2)}, \quad \tilde{c}_k^2 = \tau_k \cdot \zeta(y_k^2) \cdot \frac{\psi(y_k^2)}{\psi'(y_k^2)}\,, \tag{64}$$

where $\underline{c}_k$ also appears in Theorem 2. Then a.s.,

$$\lim_{p\to\infty}(\boldsymbol{\alpha}_k)_k^2 = \tilde{c}_k^2, \quad \lim_{p\to\infty}(\boldsymbol{\beta}_k)_k^2 = \underline{c}_k^2, \quad \lim_{p\to\infty}(\boldsymbol{\alpha}_k)_k(\boldsymbol{\beta}_k)_k = \tilde{c}_k\underline{c}_k\,. \tag{65}$$

Proof. By definition, $\|\widehat{\mathbf{v}}_k\|^2 = 1$. Consequently, by (63),

$$\begin{aligned}
1 \simeq{}& \sigma_k^2(\boldsymbol{\beta}_k)_k^2 \cdot \mathbf{u}_k^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{N}(y_k^2\mathbf{I} - \mathbf{N}^\top\mathbf{N})^{-2}\mathbf{N}^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{u}_k + (\sigma_k y_k)^2(\boldsymbol{\alpha}_k)_k^2 \cdot \mathbf{v}_k^\top(y_k^2\mathbf{I} - \mathbf{N}^\top\mathbf{N})^{-2}\mathbf{v}_k \\
&+ \sigma_k^2 y_k(\boldsymbol{\beta}_k)_k(\boldsymbol{\alpha}_k)_k \cdot \mathbf{v}_k^\top(y_k^2\mathbf{I} - \mathbf{N}^\top\mathbf{N})^{-2}\mathbf{N}^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{u}_k\,.
\end{aligned}$$

Note the third term above vanishes asymptotically, by the independence of $\mathbf{u}_k = \mathbf{D}_k\mathbf{w}_k$. Similarly, using $\|\mathbf{u}_k^w\|^2 = 1$ in (63), and replacing the quadratic forms by the corresponding traces (as in the proof of Theorem 1 above) yields the system of equations,

$$\begin{bmatrix} 1/\sigma_k^2 \\ 1/\sigma_k^2 \end{bmatrix} \simeq \begin{bmatrix} p^{-1}\mathrm{tr}\left(\mathbf{D}_k^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{N}(y_k^2\mathbf{I} - \mathbf{N}^\top\mathbf{N})^{-2}\mathbf{N}^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{D}_k\right) & y_k^2 \cdot n^{-1}\mathrm{tr}(y_k^2\mathbf{I} - \mathbf{N}^\top\mathbf{N})^{-2} \\ y_k^2 \cdot p^{-1}\mathrm{tr}\left(\mathbf{D}_k^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}(y_k^2\mathbf{I} - \mathbf{N}\mathbf{N}^\top)^{-2}\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{D}_k\right) & n^{-1}\mathrm{tr}\left(\mathbf{N}^\top(y_k^2\mathbf{I} - \mathbf{N}\mathbf{N}^\top)^{-2}\mathbf{N}\right) \end{bmatrix} \begin{bmatrix} (\boldsymbol{\beta}_k)_k^2 \\ (\boldsymbol{\alpha}_k)_k^2 \end{bmatrix}. \tag{66}$$

Next, we calculate limiting expressions for the coefficients. Using Proposition 1, the right column of (66) is

$$y_k^2 \cdot n^{-1}\mathrm{tr}(y_k^2\mathbf{I} - \mathbf{N}^\top\mathbf{N})^{-2} \simeq y_k^2\underline{s}'(y_k^2), \tag{67}$$

$$n^{-1}\mathrm{tr}\left(\mathbf{N}^\top(y_k^2\mathbf{I} - \mathbf{N}\mathbf{N}^\top)^{-2}\mathbf{N}\right) = y_k^2 n^{-1}\mathrm{tr}(y_k^2\mathbf{I} - \mathbf{N}\mathbf{N}^\top)^{-2} - n^{-1}\mathrm{tr}(y_k^2\mathbf{I} - \mathbf{N}\mathbf{N}^\top)^{-1} \simeq y_k^2\underline{s}'(y_k^2) + \underline{s}(y_k^2)\,.$$

In (59) we have calculated:

$$H_k(z) = \lim_{p\to\infty} p^{-1}\mathrm{tr}\left(\mathbf{D}_k^\top\widehat{\mathbf{\Sigma}}^{-1/2}(z\mathbf{I} - \mathbf{N}\mathbf{N}^\top)^{-1}\widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{D}_k\right) = -\tau_k\zeta(z)$$
$$-H_k'(z) = \lim_{p\to\infty} p^{-1}\mathrm{tr}\left(\mathbf{D}_k^\top\widehat{\mathbf{\Sigma}}^{-1/2}(z\mathbf{I} - \mathbf{N}\mathbf{N}^\top)^{-2}\widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{D}_k\right) = \tau_k\zeta'(z)\,.$$

Thus, the bottom left entry of (66) is $\simeq \tau_k y_k^2\zeta'(y_k^2)$. As for the top left entry, observe that $\mathbf{N}(y_k^2\mathbf{I} - \mathbf{N}^\top\mathbf{N})^{-2}\mathbf{N}^\top = (y_k^2\mathbf{I} - \mathbf{N}\mathbf{N}^\top)^{-2}\mathbf{N}\mathbf{N}^\top$. Consequently, this entry tends to $\simeq -y_k^2 H_k'(y_k^2) - H_k(y_k^2) = \tau_k y_k^2\zeta'(y_k^2) + \tau_k\zeta(y_k^2)$. Recalling that $1/\sigma_k^2 = \tau_k\psi(y_k^2) = y_k^2\zeta(y_k^2)\underline{s}(y_k^2)$ and plugging the aforementioned limits in (66), we deduce that $(\boldsymbol{\alpha}_k)_k^2, (\boldsymbol{\beta}_k)_k^2$ asymptotically satisfy

$$\begin{bmatrix}\tau_k\psi(y_k^2)\\ \tau_k\psi(y_k^2)\end{bmatrix} \simeq \begin{bmatrix}\tau_k y_k^2\zeta'(y_k^2) + \tau_k\zeta(y_k^2) & y_k^2\underline{s}'(y_k^2)\\ \tau_k y_k^2\zeta'(y_k^2) & y_k^2\underline{s}'(y_k^2) + \underline{s}(y_k^2)\end{bmatrix}\begin{bmatrix}(\boldsymbol{\beta}_k)_k^2\\ (\boldsymbol{\alpha}_k)_k^2\end{bmatrix}\,. \tag{68}$$

Solving this system yields the claimed expressions. Lastly, to conclude the calculation of $(\boldsymbol{\alpha}_k)_k(\boldsymbol{\beta}_k)_k$, it suffices to show that it is non-negative (since we already computed its modulus). Indeed, by definition,

$$\widehat{\theta}_k = \mathbf{u}_k^{w\top}\mathbf{Y}^w\mathbf{v}_k = \mathbf{u}_k^{w\top}(\widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top + \mathbf{N})\mathbf{v}_k \le \boldsymbol{\alpha}_k^\top\mathbf{\Lambda}\boldsymbol{\beta}_k + \|\mathbf{N}\|\,.$$

Now, since $\widehat{\theta}_k$ is an outlier, $\|\mathbf{N}\| \simeq \theta_{\max} < y_k \simeq \widehat{\theta}_k$, hence $\boldsymbol{\alpha}_k^\top\mathbf{\Lambda}\boldsymbol{\beta}_k > 0$. By Lemma 2, $\boldsymbol{\alpha}_k^\top\mathbf{\Lambda}\boldsymbol{\beta}_k \simeq \sigma_k(\boldsymbol{\alpha}_k)_k(\boldsymbol{\beta}_k)_k$, and so we are done. □

Next, we calculate the correlations between the recolored empirical singular vectors $\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_i^w$ and their corresponding population spikes $\mathbf{u}_k$.

Lemma 4. *Suppose that $\sqrt{\tau_k}\sigma_k > \sigma_{\mathrm{thresh}}$. Then Eqs. (36), (38) and (41) hold.*

Proof. We start with the representation (63) and take an inner product with $\widehat{\mathbf{\Sigma}}^{1/2}\mathbf{u}_\ell$:

$$\mathbf{u}_\ell^\top\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w \simeq \sigma_k\cdot(\boldsymbol{\alpha}_k)_k\cdot\mathbf{u}_\ell^\top\widehat{\mathbf{\Sigma}}^{1/2}\left(y_k^2\mathbf{I} - \mathbf{N}\mathbf{N}^\top\right)^{-1}\mathbf{N}\mathbf{v}_k + \sigma_k y_k\cdot(\boldsymbol{\beta}_k)_k\cdot\mathbf{u}_\ell^\top\widehat{\mathbf{\Sigma}}^{1/2}\left(y_k^2\mathbf{I} - \mathbf{N}\mathbf{N}^\top\right)^{-1}\widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{u}_k\,.$$

The first term is asymptotically vanishing, and the second term is possibly non-vanishing only when $k = \ell$ (since the $\mathbf{u}_k$-s are independent), and so (41) is established. We have

$$\mathbf{u}_k^\top\widehat{\mathbf{\Sigma}}^{1/2}\left(y_k^2\mathbf{I} - \mathbf{N}\mathbf{N}^\top\right)^{-1}\widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{u}_k = \mathbf{u}_k^\top\widehat{\mathbf{\Sigma}}\left(y_k^2\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^{1/2}\mathbf{E}\mathbf{\Sigma}^{1/2}\right)^{-1}\mathbf{u}_k = \mathbf{u}_k^\top\mathbf{\Sigma}^{1/2}\mathbf{S}\left(y_k^2\mathbf{S} - \mathbf{E}\right)^{-1}\mathbf{\Sigma}^{-1/2}\mathbf{u}_k$$
$$\overset{(i)}{\simeq} p^{-1}\mathrm{tr}\left(\mathbf{S}\left(y_k^2\mathbf{S} - \mathbf{E}\right)^{-1}\right)\cdot\langle\mathbf{\Sigma}^{1/2}\mathbf{u}_k, \mathbf{\Sigma}^{-1/2}\mathbf{u}_k\rangle \overset{(ii)}{\simeq} -\overline{s}(y_k^2)\,,$$

where (i) follows from the orthogonal invariance of $\mathbf{S}(y_k^2\mathbf{S} - \mathbf{E})^{-1}$; and (ii) follows from Proposition 1, writing $\mathrm{tr}(\mathbf{S}(y_k^2\mathbf{S} - \mathbf{E})^{-1}) = \mathrm{tr}(y_k^2\mathbf{I} - \mathbf{S}^{-1/2}\mathbf{E}\mathbf{S}^{-1/2})^{-1} = \mathrm{tr}(y_k^2\mathbf{I} - \mathbf{N}\mathbf{N}^\top)^{-1}$ (recall that the Stieltjes transform does not depend on $\mathbf{\Sigma}$) and the normalization $\|\mathbf{u}_k\|^2 \simeq p^{-1}\|\mathbf{D}_k\|_F^2 \simeq 1$. Recall that $(\mathbf{v}_k^\top\widehat{\mathbf{v}}_k^w)^2 = (\boldsymbol{\beta}_k)_k^2 \simeq \underline{c}_k^2$ was computed in Lemma 3. Thus,

$$(\mathbf{u}_k^\top\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w)^2 \simeq \sigma_k^2 y_k^2\cdot\underline{c}_k^2\cdot\overline{s}(y_k)^2, \quad (\mathbf{u}_k^\top\widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w)(\mathbf{v}_k^\top\widehat{\mathbf{v}}_k^w) \simeq -\sigma_k y_k\cdot\underline{c}_k^2\cdot\overline{s}(y_k)\,.$$

Finally, use $\sigma_k^2 = 1/\left(\tau_k\psi(y_k^2)\right)$ to get the expressions in Eqs. (36) and (38). □

It remains to compute $\|\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\mathbf{u}}_i^w\|^2$. To this end, multiply (63) by $\widehat{\boldsymbol{\Sigma}}^{1/2}$ and take the norm,

$$\begin{aligned}
\|\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w\|^2 &\simeq \left\|\sigma_k(\boldsymbol{\alpha}_k)_k\widehat{\boldsymbol{\Sigma}}^{1/2}\left(y_k^2\mathbf{I}-\mathbf{N}\mathbf{N}^\top\right)^{-1}\mathbf{N}\mathbf{v}_k+\sigma_k y_k(\boldsymbol{\beta}_k)_k\widehat{\boldsymbol{\Sigma}}^{1/2}\left(y_k^2\mathbf{I}-\mathbf{N}\mathbf{N}^\top\right)^{-1}\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{u}_k\right\|^2 \\
&\overset{(i)}{\simeq} \sigma_k^2\tilde{c}_k^2\mathbf{v}_k^\top\mathbf{N}^\top\left(y_k^2\mathbf{I}-\mathbf{N}\mathbf{N}^\top\right)^{-1}\widehat{\boldsymbol{\Sigma}}\left(y_k^2\mathbf{I}-\mathbf{N}\mathbf{N}^\top\right)^{-1}\mathbf{N}\mathbf{v}_k \\
&\quad+\sigma_k^2y_k^2\underline{c}_k^2\mathbf{u}_k^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}(y_k^2-\mathbf{N}\mathbf{N}^\top)^{-1}\widehat{\boldsymbol{\Sigma}}(y_k^2-\mathbf{N}\mathbf{N}^\top)^{-1}\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{u}_k \\
&\overset{(ii)}{\simeq} \sigma_k^2\tilde{c}_k^2\cdot n^{-1}\mathrm{tr}\left(\mathbf{N}^\top\left(y_k^2\mathbf{I}-\mathbf{N}\mathbf{N}^\top\right)^{-1}\widehat{\boldsymbol{\Sigma}}\left(y_k^2\mathbf{I}-\mathbf{N}\mathbf{N}^\top\right)^{-1}\mathbf{N}\right) \\
&\quad+\sigma_k^2y_k^2\underline{c}_k^2\cdot p^{-1}\mathrm{tr}\left(\mathbf{D}_k^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}(y_k^2\mathbf{I}-\mathbf{N}\mathbf{N}^\top)^{-1}\widehat{\boldsymbol{\Sigma}}(y_k^2\mathbf{I}-\mathbf{N}\mathbf{N}^\top)^{-1}\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{D}_k\right),
\end{aligned}$$

where: (i) we discarded the cross term, which is asymptotically vanishing; and (ii) similarly to previous calculations, the quadratic forms concentrate around the traces. Define

$$\begin{aligned}
G_1(z) &= \lim_{p\to\infty} n^{-1}\mathrm{tr}\left(\mathbf{N}^\top\left(z\mathbf{I}-\mathbf{N}\mathbf{N}^\top\right)^{-1}\widehat{\boldsymbol{\Sigma}}\left(z\mathbf{I}-\mathbf{N}\mathbf{N}^\top\right)^{-1}\mathbf{N}\right), \\
G_{2,k}(z) &= \lim_{p\to\infty} p^{-1}\mathrm{tr}\left(\mathbf{D}_k^\top\widehat{\boldsymbol{\Sigma}}^{-1/2}(z\mathbf{I}-\mathbf{N}\mathbf{N}^\top)^{-1}\widehat{\boldsymbol{\Sigma}}(z\mathbf{I}-\mathbf{N}\mathbf{N}^\top)^{-1}\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{D}_k\right),
\end{aligned} \tag{69}$$

so that

$$\|\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w\|^2 \simeq \sigma_k^2\tilde{c}_k^2\cdot G_1(y_k^2)+\sigma_k^2y_k^2\underline{c}_k^2\cdot G_{2,k}(y_k^2). \tag{70}$$

Also define the following mixed traces

$$\Upsilon_1(z) = \lim_{p\to\infty} p^{-1}\mathrm{tr}(z\mathbf{S}-\mathbf{E})^{-1}\mathbf{S}^2, \tag{71}$$

$$\Upsilon_2(z) = \lim_{p\to\infty} p^{-1}\mathrm{tr}(z\mathbf{S}-\mathbf{E})^{-2}\mathbf{S}^2, \tag{72}$$

where $z\in\mathbb{C}\setminus[\theta_{\min}^2,\theta_{\max}^2]$ and the limit exists a.s. We show in Appendix, Section C that $\Upsilon_1(\cdot),\Upsilon_2(\cdot)$ have the closed-form formulas given in Eqs. (24) and (25).

Lemma 5. *Let $\Upsilon_1(\cdot)$ be defined in (71), and recall the notation $\mu=p^{-1}\mathrm{tr}(\boldsymbol{\Sigma})$. Then*

$$G_1(z) = -\gamma\mu\cdot\left[z\Upsilon_1(z)+\Upsilon_1'(z)\right], \quad z\in\mathbb{C}\setminus[\theta_{\min}^2,\theta_{\max}^2]. \tag{73}$$

Proof. Let

$$\tilde{G}_1(z) = \lim_{n\to\infty} n^{-1}\mathrm{tr}\left((z\mathbf{I}-\mathbf{N}\mathbf{N}^\top)^{-1}\widehat{\boldsymbol{\Sigma}}\right), \quad \text{so that} \quad G_1(z) = -z\tilde{G}_1'(z)-\tilde{G}_1(z).$$

Straightforward algebraic manipulation yields $\mathrm{tr}\left((z\mathbf{I}-\mathbf{N}\mathbf{N}^\top)^{-1}\widehat{\boldsymbol{\Sigma}}\right) = \mathrm{tr}\left(\mathbf{S}(z\mathbf{S}-\mathbf{E})^{-1}\mathbf{S}\boldsymbol{\Sigma}\right)$. Since $\mathbf{S}(z\mathbf{S}-\mathbf{E})^{-1}\mathbf{S}$ is orthogonally invariant, $\{\mathbf{S}(z\mathbf{S}-\mathbf{E})^{-1}\mathbf{S},\boldsymbol{\Sigma}\}$ are asymptotically free (see Section A.2). Thus,

$$\lim_{p\to\infty} p^{-1}\mathrm{tr}\left(\mathbf{S}(z\mathbf{S}-\mathbf{E})^{-1}\mathbf{S}\boldsymbol{\Sigma}\right) = \overline{\mathrm{tr}}(\mathbf{S}(z\mathbf{S}-\mathbf{E})^{-1}\mathbf{S})\cdot\overline{\mathrm{tr}}(\boldsymbol{\Sigma}) = \mu\cdot\Upsilon_1(z),$$

and the lemma follows. □

Lemma 6. *Let $\Upsilon_2(\cdot)$ be defined in (72). Then*

$$G_{2,k}(z) = \mu\tau_k\left[\Upsilon_2(z)-(\overline{s}(z))^2\right]+(\overline{s}(z))^2, \quad z\in\mathbb{C}\setminus[\theta_{\min}^2,\theta_{\max}^2]. \tag{74}$$

We provide a proof in Appendix, Section B. To conclude the computation of $\|\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\mathbf{u}}_k^w\|^2$, and thereby the proof of Theorem 2, combine (70) with Lemmas 5 and 6. □

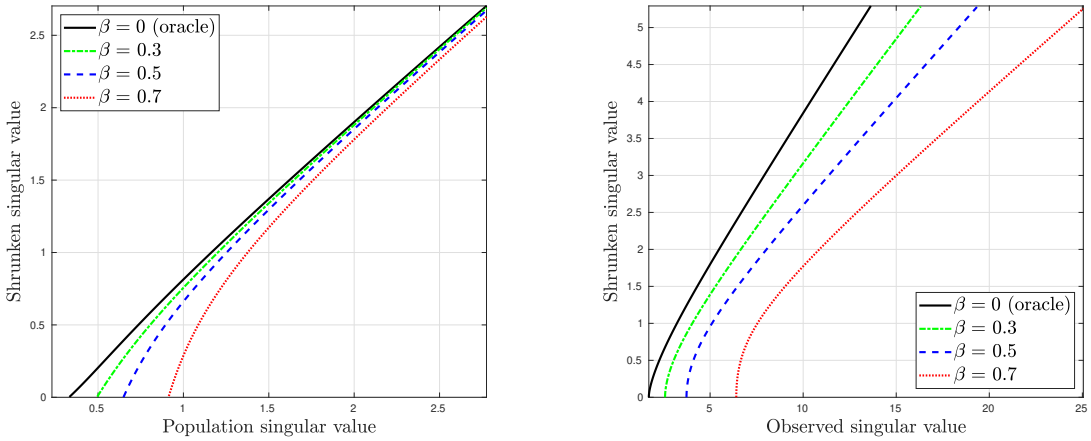

Figure 2: Plots of the asymptotically optimal singular values as functions of the population singular value (left) and of the observed singular value (right), for different values of β.

6 Numerical results

We report on several numerical experiments that illustrate both the performance of the denoising algorithm relative to other methods, and the agreement between the asymptotic results and finite sample estimates for both Gaussian and non-Gaussian noise. One of the questions addressed in this section is under what parameter settings the optimal Whiten-Shrink- re-Color (WSC) denoiser outperforms OptShrink, which is optimal shrinkage without any pre-transformation [47]. It was shown in [39] that when the signal principal components are uniformly random, optimal shrinkage with *oracle* whitening (corresponding to $\beta = 0$) outperforms OptShrink for heteroscedastic noise; consequently, optimal shrinkage with pseudo-whitening (when $\beta > 0$) will still outperform OptShrink when β is sufficiently small, though the precise value of β will depend on the other problem parameters, such as γ and the heteroscedasticity of the noise.

Sections 6.2 and 6.3 numerically illustrate this behavior, by comparing the errors of the two methods in different parameter regimes and evaluating the β at which the two methods have identical error. Section 6.4 explores a different but related phenomenon, comparing the minimum detectable signal singular value of the two methods. Section 6.5 examines the two methods' performances in estimating the signal principal components. Section 6.6 compares different shrinkage methods on finite-sample data, again showing that the relative performance of pseudo-whitening over OptShrink increases as the heteroscedasticity of the noise increases.

The experiments in Section 6.7 illustrate two phenomena: first, that the asymptotic results appear to hold for non-Gaussian noise with sufficiently many moments ("universality"), though they break down when the noise becomes too fat-tailed; and second, that the spiked model parameters appear to converge at the rate of approximately $O(p^{-1/2})$ for thin-tailed distributions.

All experiments reported in this section were performed in Matlab. Code may be found online at `https://github.com/wleeb/FShrink`.

6.1 Plots of optimal shrinkers

We plot the asymptotically optimal singular values t_1 as functions of the population singular values and of the observed singular values, for $\gamma = 1/2$, covariance $\mathbf{\Sigma} \in \mathbb{R}^{1000\times 1000}$ with eigenvalues linearly spaced between $1/500$ and 1, and $\mathbf{D}_1 = \mathbf{I}_{1000}$. The optimal singular values are computed for different values of β, and plotted in Figure 2. Note that $\beta = 0$ corresponds to the oracle singular values (where $\mathbf{\Sigma}$ is known exactly), used in [39]. For larger β, the optimal singular values decrease in value; that is, more shrinkage is needed to account for the increased uncertainty in the estimate of $\mathbf{\Sigma}$.

6.2 Asymptotic errors

The previous work [39] shows, assuming isotropic random spike directions, that an oracle WSC denoiser (which has access to $\mathbf{\Sigma}$; equivalently, $\beta = 0$ under our setup) outperforms optimal singular value shrinkage (OptShrink) [47]. Consequently, pseudo-whitening must also outperform OptShrink provided that β is sufficiently small, that is, one has access to sufficiently many pure-noise samples. Though the precise value of β at which optimal shrinkage with pseudo-whitening has superior performance is not immediately transparent from the error formulas, it can be evaluated numerically. In this set of experiments, we compare the asymptotic errors of optimal shrinkage with pseudo-whitening, comparing them to the asymptotic errors obtained by OptShrink; for context, we also show a comparison with the oracle WSC denoiser from [39] (which outperforms both methods, but is not a viable procedure in the setting of this paper).

6.2.1 Linearly-spaced eigenvalue spectrum

For a specified aspect ratio $\gamma > 0$ and condition number $\kappa \geq 1$, we consider a diagonal noise covariance $\mathbf{\Sigma}$ with $p = 2000$ linearly spaced eigenvalues whose maximum and minimum elements have ratio κ, and that are normalized so that $\tau = 1$; that is,

$$\frac{1}{p} \sum_{i=1}^{p} \frac{1}{\mathbf{\Sigma}_{ii}} = 1. \tag{75}$$

For $\kappa = 50,000$, we compute σ_{BBP}, the asymptotically largest singular value of the noise matrix $\mathbf{\Sigma}^{1/2}\mathbf{Z}/\sqrt{n}$, using the numerical method introduced in [37] (the term BBP is from the paper [7]); this method evaluates σ_{BBP} by numerically solving the equations that implicitly characterize σ_{BBP}. We consider a rank 1 p-by-n signal matrix $\mathbf{X}$ with i.i.d. singular vectors $\mathbf{u}$ and $\mathbf{v}$ and singular value $\sigma = \sigma_{\mathrm{BBP}} + 1$. For these parameters, the asymptotic errors for optimal shrinkage with exact whitening are then computed using the formula from [39], while the asymptotic error for OptShrink is evaluated numerically using the method from [37], which numerically solves the Stieltjes transform of the limiting spectral distribution of the sample covariance of the noise. For optimal shrinkage with pseudo-whitening, we consider a range of aspect ratios β between 0 and 0.95, and evaluate the AMSE for each value of β. These values are plotted as functions of β in Figure 3, alongside the asymptotic errors of the other two methods. As the plots demonstrate, for each value of γ and κ the error of shrinkage with pseudo-whitening approaches that of shrinkage with exact whitening as $\beta \to 0$. Furthermore, for each value of γ, as κ grows (that is, as the noise becomes more heteroscedastic) the performance of shrinkage with pseudo-whitening improves relative to OptShrink; when $\kappa = 50,000$, shrinkage with pseudo-whitening outperforms OptShrink for the entire considered range of β.

6.2.2 Polynomially-decaying eigenvalue spectrum

For a specified aspect ratio $\gamma > 0$ and parameter α, we consider a diagonal noise covariance $\mathbf{\Sigma}$ with $p = 2000$ eigenvalues of the form $C \cdot t^{\alpha}$, where t are equispaced between 1 and 3, and where C is chosen so that $\tau = 1$, i.e. (75) holds. For $\alpha = 5$, we compute σ_{BBP}, the asymptotically largest singular value of the noise matrix $\mathbf{\Sigma}^{1/2}\mathbf{Z}/\sqrt{n}$, using the numerical method introduced in [37]. We consider a rank 1 p-by-n signal matrix $\mathbf{X}$ with i.i.d. singular vectors $\mathbf{u}$ and $\mathbf{v}$ and singular value $\sigma = \sigma_{\mathrm{BBP}} + 1$. For these parameters, the asymptotic errors for optimal shrinkage with exact whitening are then computed using the formula from [39], while the asymptotic error for OptShrink is evaluated numerically using the method from [37]. For optimal shrinkage with pseudo-whitening, we consider a range of aspect ratios β between 0 and 0.95, and evaluate the AMSE for each value of β. These values are plotted as functions of β in Figure 4, alongside the asymptotic errors of the other two methods. As the plots demonstrate, for each value of γ and α the error of shrinkage with pseudo-whitening approaches that of shrinkage with exact

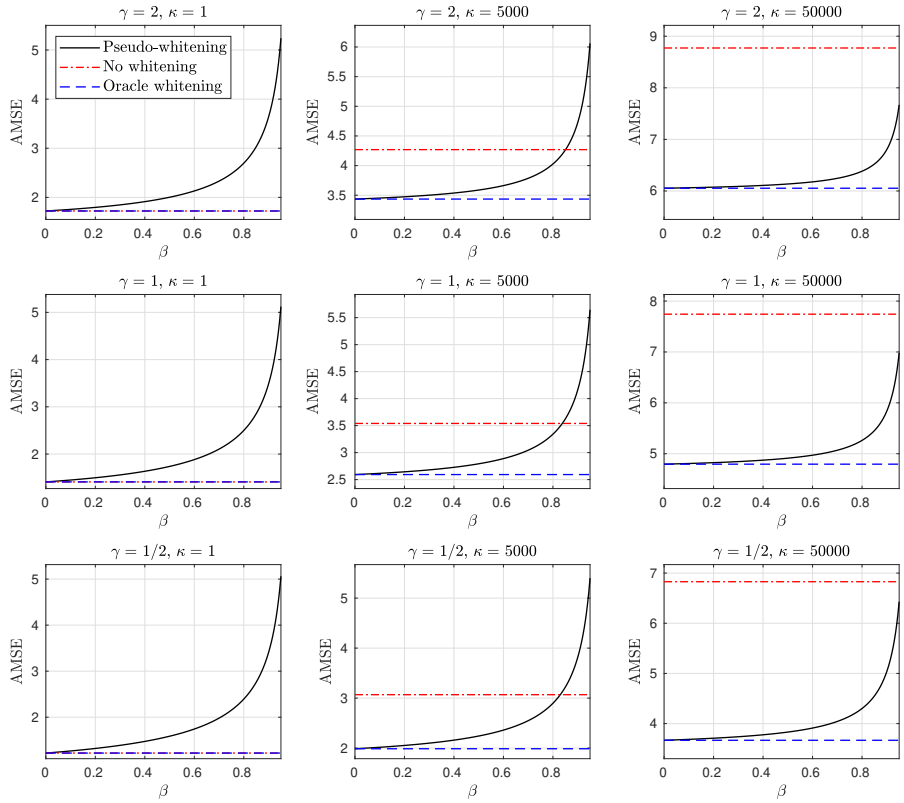

Figure 3: Asymptotic errors for the optimal WSC denoiser with pseudo-whitening as a function of $\beta \sim p/m$ for different values of $\gamma \sim p/n$ and parameter κ, where the noise covariance has linearly spaced eigenvalues and condition number κ. For reference, we also plot the asymptotic errors for the oracle WSC denoiser (from [39]) and optimal shrinkage without whitening (OptShrink, from [47]). The errors for OptShrink are numerically evaluated using the method from [37]. Our numerical results demonstrate that the performance gains offered by whitening become more pronounced as the condition number κ increases.

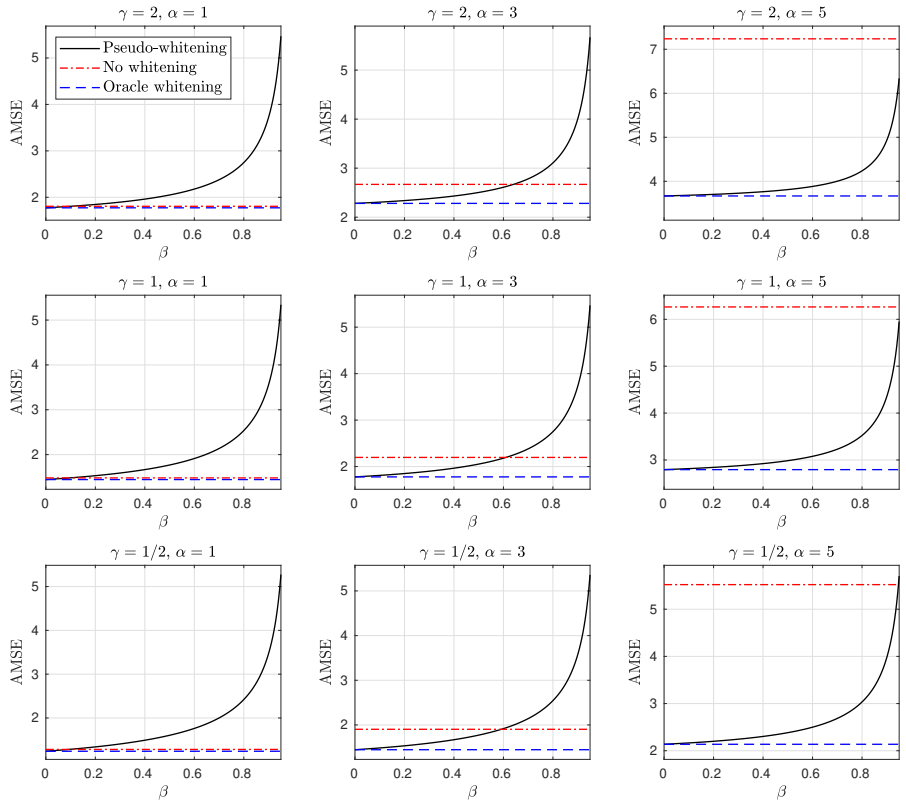

Figure 4: Asymptotic errors for optimal shrinkage with pseudo-whitening as a function of $\beta \sim p/m$ for different values of $\gamma \sim p/n$ and parameter α, with noise eigenvalues of the form $C \cdot t^{\alpha}$ with t equispaced between 1 and 3. For reference, we also plot the asymptotic errors for the oracle WSC denoiser (from [39]) and optimal shrinkage without whitening (OptShrink, from [47]). The errors for OptShrink are numerically evaluated using the method from [37]. Our numerical results demonstrate that the performance gains offered by whitening become more pronounced as α increases, that is the eigenvalues of the noise covariance deviate further from a constant profile.

whitening as $\beta \to 0$. Furthermore, for each value of γ, as α grows (that is, as the noise becomes more heteroscedastic) the performance of shrinkage with pseudo-whitening improves relative to OptShrink; for example, when $\gamma = 2$ and $\alpha = 5$, shrinkage with pseudo-whitening outperforms OptShrink for the entire considered range of β.

6.3 Critical value of β

For another view of the experiments in Section 6.2, we evaluate the value of β at which optimal shrinkage with pseudo-whitening outperforms OptShrink. When $\beta = 0$ (i.e. when oracle whitening is used), pseudo-whitening is guaranteed to outperform OptShrink [39]; the performance of pseudo-whitening degrades as β approaches 1, and for some parameter values pseudo-whitening will perform worse than OptShrink when β is sufficiently close to 1. To illustrate this phenomenon, we consider the same two families of noise covariances described in Sections 6.2.1 and 6.2.2. Holding all other parameter values fixed, we consider the error of optimal shrinkage with pseudo-whitening as a function of β, and solve for the value of β at which the pseudo-whitening error is equal to that of OptShrink when the signal strength is equal to $1.1\sigma_{\text{BBP}}$. The value of σ_{BBP} and the errors for OptShrink are evaluated using the method from [37], as in Section 6.2; and the critical value of β is numerically evaluated using the secant method.

The left panel of Figure 5 plots the critical value of β for the Section 6.2.1 model as a

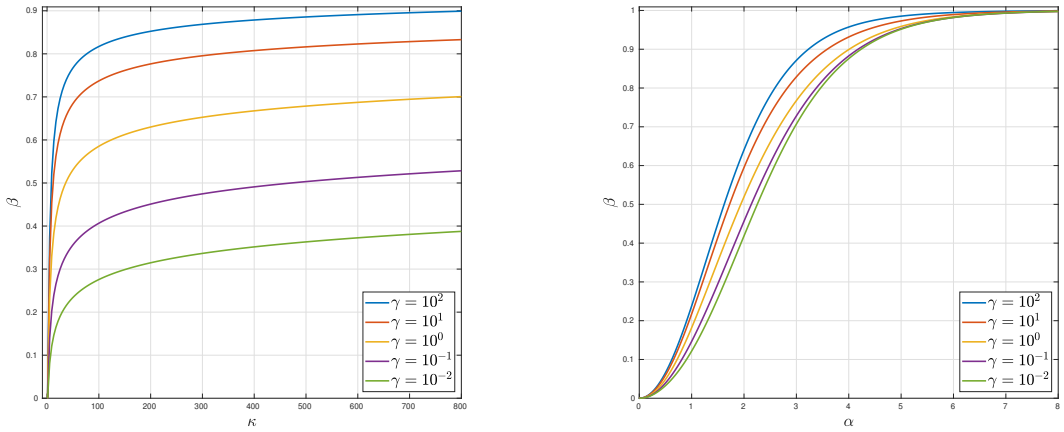

Figure 5: Value of β for which the AMSE of OptShrink is equal to the AMSE of optimal singular value shrinkage with pseudo-whitening. Left: critical value of β plotted as a function of κ, where the noise covariance has linearly spaced eigenvalues and condition number κ. Right: critical value of β plotted as a function of α, with noise eigenvalues of the form $C \cdot t^{\alpha}$ with t equispaced between 1 and 3. It is observed that the critical value of β grows as the heteroscedasticity parameters (κ or α) grow; that is, the more heteroscedastic the noise, the fewer noise-only samples are needed for optimal shrinkage with pseudo-whitening to outperform OptShrink.

function of the condition number κ; the right panel plots the critical value of β for the Section 6.2.2 model as a function of the decay parameter α. These curves are shown for a range of values of γ. Both plots reveal similar qualititative behavior: the critical value of β grows as the heteroscedasticity parameters (κ or α) grow; that is, the more heteroscedastic the noise, the fewer noise-only samples are needed for optimal shrinkage with pseudo-whitening to outperform OptShrink. Furthermore, the relative performance of optimal shrinkage with pseudo-whitening also increases with γ.

6.4 Minimum detectable signal

To further illustrate the differences in performance between optimal shrinkage with pseudo-whitening and OptShrink as a function of β, we consider the smallest singular value of $\mathbf{X}$ that is detectable under each noise model. We denote by θ_{BBP} the smallest detectable singular value of $\mathbf{X}$; for any specified variance profile, this may be numerically evaluated using the method from [37], which solves the equations that implicitly characterize the value. We also consider the minimum detectable signal singular value by pseudo-whitening, given by the formula (14); this value obviously grows with β. Figures 6 and 7 plot the minimum detectable values under OptShrink as functions of the heteroscedasticity of the noise, as measured by the parameter κ for the model in Section 6.2.1 and the parameter α for the model in Section 6.2.2, respectively. These figures also display the values of the largest β where the two methods have identical signal detection thresholds, again as functions of the heteroscedasticity of the noise covariances.

6.5 Principal component estimation

Next, we compare the methods of pseudo-whitening to OptShrink in estimating the principal components of the signal vector. We fix the parameter $\gamma = 1/2$ and consider the model from Section 6.2.1, where the noise covariance has $p = 2000$ linearly spaced spectrum and is normalized so that $\tau = 1$. Figure 8 plots the asymptotic absolute inner products between the estimated PCs under pseudo-whitening, for several values of β. Also shown are the asymptotic absolute inner products between the true PCs and the estimated PCs used by OptShrink (the top left singular vector of the unnormalized data matrix). The signal singular value is taken to be

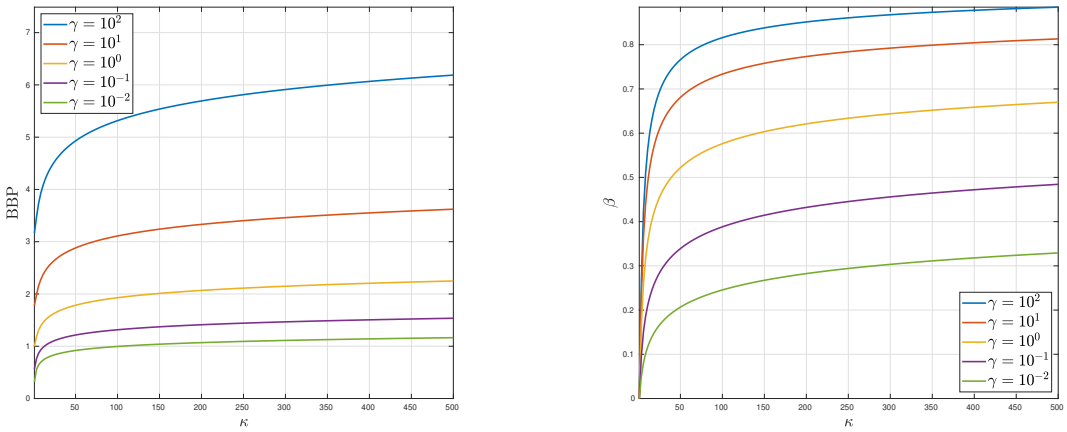

Figure 6: Left: the minimum detectable signal singular value by OptShrink. Right: value of β for which the minimum detectable signal singular value for the pseudo-whitened matrix equals the minimum detectable value for OptShrink. Values are plotted as functions of κ, where the noise covariance has linearly spaced eigenvalues and condition number κ.

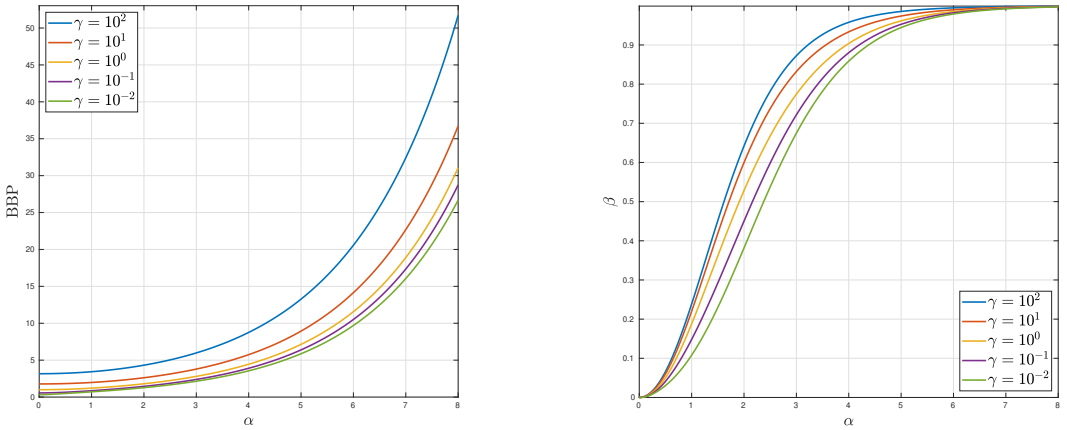

Figure 7: Left: the minimum detectable signal singular value by OptShrink. Right: value of β for which the minimum detectable signal singular value for the pseudo-whitened matrix equals the minimum detectable value for OptShrink. Values are plotted as functions of α, where the noise covariance eigenvalues are of the form $C \cdot t^{\alpha}$ with t equispaced between 1 and 3.

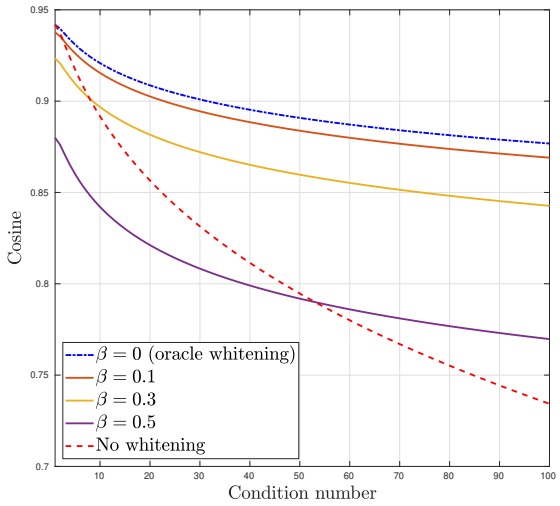

Figure 8: The cosines of the angles between the true principal component and the estimated principal component, plotted as functions the condition number κ of the noise covariance matrix $\mathbf{\Sigma}$ with linearly spaced eigenvalues.

$\max\{\sigma_{\text{thresh}}, \sigma_{\text{BBP}}\} + 1$, where σ_{thresh} is evaluated for the largest value of β considered ($\beta = .5$). Both σ_{BBP} and the cosines for the unwhitened PCs are evaluated using the method from [39].

As is apparent from the plot, larger values of β result in smaller cosines; that is, when pseudo-whitening is performed with fewer noise-only samples, the resulting estimates are worse. Second, as the condition number grows – that is, as the noise becomes more heteroscedastic – the relative performance of pseudo-whitening improves relative to estimation without whitening. At a certain condition number, each pseudo-whitening curve begins to outperform the estimates without any whitening.

6.6 Comparison with other singular value shrinkers

We compare the shrinker described in Section 2.4 to three other shrinkage methods. The first method is OptShrink [47], the optimal singular value shrinker of the data matrix $\mathbf{Y}$ itself without any transformation. OptShrink uses knowledge of the operator norm of the pure noise matrix $\mathbf{\Sigma}^{1/2}\mathbf{Z}$, which we evaluate numerically using the method from [37]. We also compare to optimal shrinkage with exact whitening from [39]. In the present context this is an oracle method, that assumes exact knowledge of the noise covariance $\mathbf{\Sigma}$ (equivalently, it may be viewed it as instance of the algorithm from Section 2.4 with $\beta = 0$). Finally, we compare with the shrinker from [39] with the sample covariance $\widehat{\mathbf{\Sigma}}$ used as a plug-in estimator; that is, this method uses pseudo-whitening, but treats it as if it were oracle whitening.

The details of the experiment are are follows. We set the problem size parameters $p = 600$; $n = 1200$; $m = 1800$; and $r = 3$. We generate the covariance $\mathbf{\Sigma}$ with diagonal entries equally spaced between 1 and a specified condition number $\kappa \geq 1$. The signal singular values σ_1, σ_2, and σ_3 are defined as follows:

$$\sigma_j = \sigma_{\text{thresh}} \cdot \frac{1 + (r - j + 1)/r}{\sqrt{\tau}}, \quad 1 \leq j \leq 3, \tag{76}$$

where τ is given by

$$\tau = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{\mathbf{\Sigma}_{ii}}. \tag{77}$$

These value of σ ensures that the signal components are asymptotically detectable, but very close to the edge of detectability. We generate the vectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ in $\mathbb{R}^n$ and $\mathbf{u}_1$, $\mathbf{u}_2$ and

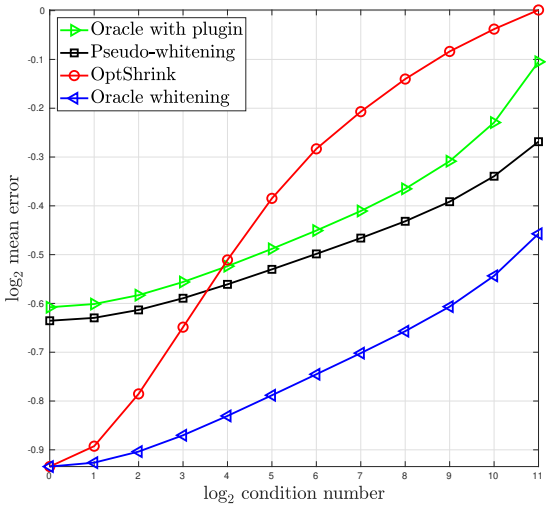

Figure 9: Plots of the mean relative error as a function of the condition number of the noise covariance matrix $\mathbf{\Sigma}$ (plotted in log scale). Four shrinkage methods are considered: (i) Oracle with plugin: the whiten-shrink-recolor procedure of [40], with a plugin estimate for the noise covariance matrix; (ii) Pseudo-whitening: the method proposed in this paper; (iii) OptShrink: the optimal singular value shrinker of [47]; (iv) Oracle whitening: the procedure if [40] using the true noise covariance matrix – note that this is an oracle method, and not practically implementable under the setting considered in this paper. The experiment shows that, unsurprisingly, oracle whitening attains the smallest error, and outperforms optimal shrinkage (OptShrink) by an increasing margin as the noise covariance becomes more and more ill-conditioned. Pseudo-whitening using the number of available noise-only measurements ($m = 1600$, the data dimensions being $n = 1200, p = 600$) is worse than OptShrink for small condition number, but attains markedly better performance when it is large. Furthermore pseudo-whitening, which is optimal over all Whiten-Shrink-reColor procedures, outperforms [40] with a plugin estimate of the noise covariance.

$\mathbf{u}_3$ in $\mathbb{R}^p$ to be random orthonormal vectors (obtained by applying Gram-Schmidt to random vectors). We then define the $p \times n$ signal matrix $\mathbf{X}$

$$\mathbf{X} = \sum_{k=1}^{3} \sigma_k \mathbf{u}_k \mathbf{v}_k^\top. \tag{78}$$

We generate the $p \times n$ in-sample noise matrix $\mathbf{\Sigma}^{1/2}\mathbf{Z}$, where $\mathbf{Z}$ has i.i.d. entries with variance 1 from a Gaussian distribution. We construct the $p \times n$ data matrix $\mathbf{Y} = \mathbf{X} + \mathbf{\Sigma}^{1/2}\mathbf{Z}$. Finally, for each value of κ, we generate the $p \times m$ out-of-sample noise matrix $\mathbf{\Sigma}^{1/2}\mathbf{Z}'$, where $\mathbf{Z}'$ has i.i.d. Gaussian entries of variance 1. Each of the four methods is then applied to this input data, with the oracle rank $r = 3$ supplied. For each value of κ, the data is generated $N = 200$ times, and the relative errors are averaged over all runs. The results of this experiment are shown Figure 9, which plots the $\log_2$ mean relative errors of each method against $\log_2(\kappa)$. Optimal shrinkage with oracle whitening outperforms each method, as is expected; the optimal shrinker with pseudo-whitening always outperforms the oracle shrinker with a plug-in covariance, and outperforms OptShrink at high condition numbers.

6.7 Convergence for Gaussian and non-gaussian noise

We check the convergence rates of the observed cosines and singular values to their limiting values. The simulation was run as follows. We set the parameters $\gamma = 2/3$ and $\beta = 1/4$. For each fixed value of p, we take $n = p/\gamma$ and $m = p/\beta$. For each p, we generate the noise covariance $\mathbf{\Sigma}$ with diagonal entries linearly spaced between 1 and 50. We generate the matrix $\mathbf{D} = \mathbf{D}_1$ as

$$\mathbf{D} = \sqrt{p} \cdot \mathrm{diag}(1/p^2, 4/p^2, \ldots, (p-1)^2/p^2, 1). \tag{79}$$

We then generate the vector $\mathbf{u} = \mathbf{u}_1$ as

$$\mathbf{u} = \frac{\mathbf{D}\mathbf{w}}{\|\mathbf{D}\mathbf{w}\|}, \tag{80}$$

where $\mathbf{w}$ has entries that are i.i.d. Gaussian. We also set the value $\sigma = \sigma_1$ to be:

$$\sigma = \sigma_{\text{thresh}} \cdot \frac{1.8}{\sqrt{\tau}}, \tag{81}$$

where

$$\tau = \frac{1}{p} \sum_{i=1}^{p} \frac{\mathbf{D}_{1i}^2}{\mathbf{\Sigma}_{ii}}. \tag{82}$$

This value of σ ensures that the signal is detectable. We also generate the vector $\mathbf{v} = \mathbf{v}_1$ with i.i.d. Gaussian entries. With these parameters set, we define the $p \times n$ signal matrix $\mathbf{X} = \sigma \mathbf{u}\mathbf{v}^\top$. We generate the $p \times n$ in-sample noise matrix $\mathbf{\Sigma}^{1/2}\mathbf{Z}$, where $\mathbf{Z}$ has i.i.d. entries with variance 1 from a specified distribution, either Gaussian, Rademacher, normalized t_{10}, normalized $t_{4.5}$, or normalized t_3 (where the t distributions are normalized to have unit variance). We construct the $p \times n$ data matrix $\mathbf{Y} = \mathbf{X} + \mathbf{\Sigma}^{1/2}\mathbf{Z}$. Finally, we generate the $p \times m$ out-of-sample noise matrix $\mathbf{\Sigma}^{1/2}\mathbf{Z}'$, where $\mathbf{Z}'$ has i.i.d. entries from the same distribution as the entries of $\mathbf{\Sigma}^{1/2}\mathbf{Z}$.

With the data generated in this way, we compute the values $\widehat{\mathbf{v}}^\top \mathbf{v}$, $(\widehat{\mathbf{u}}^w)^\top \widehat{\mathbf{\Sigma}}^{1/2} \mathbf{u}$, $(\widehat{\mathbf{u}}^w)^\top \widehat{\mathbf{\Sigma}}^{-1/2} \mathbf{u}$, and $\widehat{\theta}$. For each value of p, the experiment is run $N = 10,000$ times, and the errors are averaged as follows:

$$\mathrm{Error}_p(\xi) = \frac{1}{N} \sum_{i=1}^{N} \frac{|\widehat{\xi}_i - \xi|}{|\xi|}, \tag{83}$$

	Mean error, inner cosines				
p	Gaussian	Rademacher	t, df=10	t, df=4.5	t, df=3
550	9.857e-03	9.920e-03	9.979e-03	1.991e-02	4.841e-01
1100	6.998e-03	7.043e-03	7.027e-03	1.470e-02	6.068e-01
2200	4.926e-03	4.933e-03	4.946e-03	1.073e-02	7.302e-01
4400	3.508e-03	3.491e-03	3.440e-03	8.147e-03	8.476e-01

Table 4: Mean error of $\widehat{\mathbf{v}}^\top \mathbf{v}$.

	Mean error, outer cosines, whitened				
p	Gaussian	Rademacher	t, df=10	t, df=4.5	t, df=3
550	1.559e-02	1.600e-02	1.603e-02	2.604e-02	4.968e-01
1100	1.133e-02	1.131e-02	1.123e-02	1.907e-02	6.146e-01
2200	7.974e-03	8.020e-03	7.896e-03	1.366e-02	7.346e-01
4400	5.662e-03	5.534e-03	5.611e-03	1.026e-02	8.498e-01

Table 5: Mean error of $(\widehat{\mathbf{u}}^w)^\top \widehat{\mathbf{\Sigma}}^{1/2}\mathbf{u}$.

	Mean error, outer cosines, unwhitened				
p	Gaussian	Rademacher	t, df=10	t, df=4.5	t, df=3
550	2.121e-02	2.130e-02	2.123e-02	3.085e-02	4.794e-01
1100	1.498e-02	1.500e-02	1.496e-02	2.259e-02	6.022e-01
2200	1.058e-02	1.071e-02	1.060e-02	1.614e-02	7.266e-01
4400	7.529e-03	7.417e-03	7.477e-03	1.211e-02	8.452e-01

Table 6: Mean error of $(\widehat{\mathbf{u}}^w)^\top \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{u}$.

	Mean error, singular values				
p	Gaussian	Rademacher	t, df=10	t, df=4.5	t, df=3
550	1.653e-02	1.673e-02	1.677e-02	1.944e-02	2.857e-01
1100	1.169e-02	1.188e-02	1.164e-02	1.352e-02	3.609e-01
2200	8.267e-03	8.343e-03	8.350e-03	9.464e-03	4.464e-01
4400	5.918e-03	5.907e-03	5.928e-03	6.811e-03	6.014e-01

Table 7: Mean error of $\widehat{\theta}$.

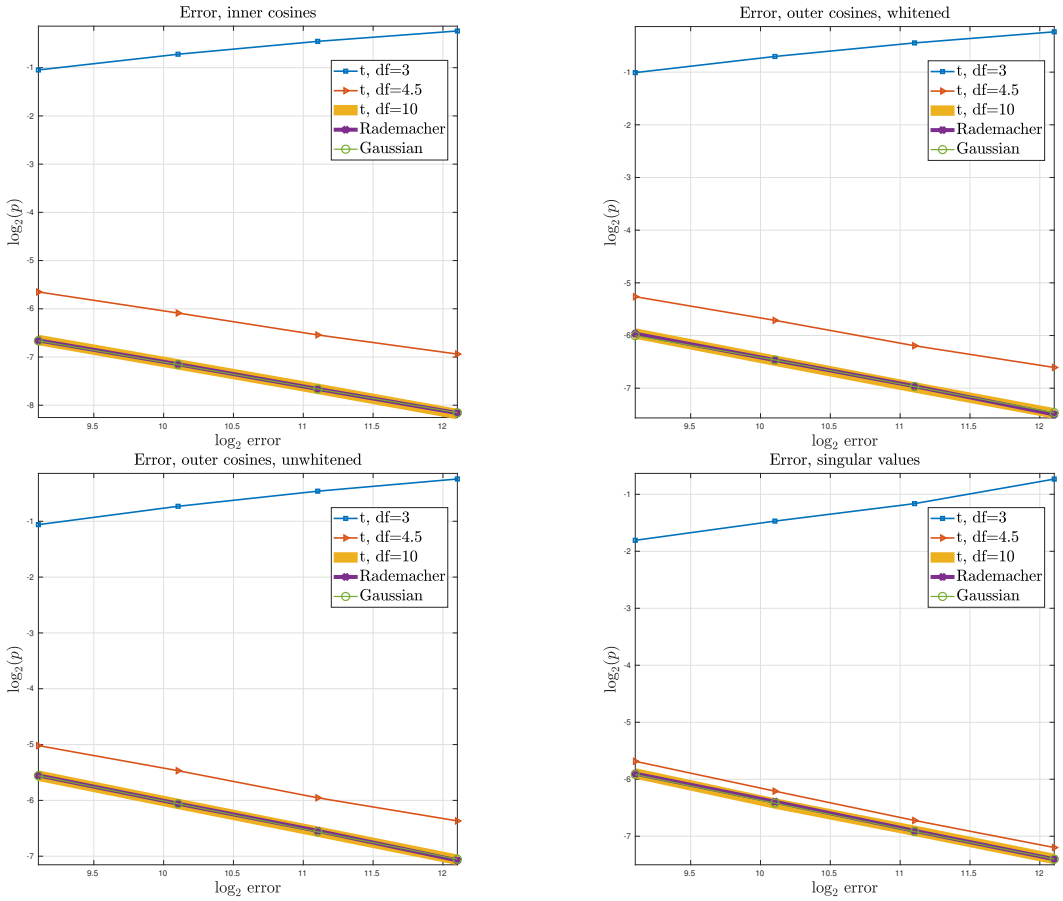

Figure 10: The $\log_2$ of the average relative errors between the spiked model parameters and their asymptotic values, plotted against $\log_2(p)$, for five different noise types. The errors for the Gaussian, Rademacher, and t_{10} noise are nearly indistinguishable, and appears to decay at a rate of approximately $O(p^{-1/2})$, since their curves have slopes close to $1/2$; the errors for $t_{4.5}$ noise are larger, but still appear to decay at approximately the same rate. By contrast, the errors for the heavy-tailed t_3 noise diverge.

where ξ denotes the asymptotic value of the parameter in question and $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ denote the N realizations of the parameter. These average errors are presented in Tables 4 through 7. Figure 10 shows the errors are plotted in log scale.

For noise with Gaussian, Rademacher, and t_{10} distributions the errors between the observed values and the true values decays at approximately the rate $O(p^{-1/2})$; furthermore, the errors themselves are nearly identical regardless of the noise distribution. By contrast, the errors for the $t_{4.5}$ distribution are larger, though they still shrink at a similar rate; whereas the errors for the fat-tailed t_3 distribution grow with p.

7 Discussion and conclusion

This paper considered the problem of low-rank matrix denoising under additive noise, assuming that the columns of the noise matrix are i.i.d. but have an otherwise arbitrary and unknown inter-row covariance structure. Under our setup, one has access to side information in the form of pure-noise samples, and wishes to make use of this additional information in denoising. Crucially, the number of available pure-noise samples is such that one cannot consistently estimate the noise covariance matrix $\mathbf{\Sigma}$; thus, one can think of $\mathbf{\Sigma}$ as being only partially known.

We proposed to tackle this problem by means of singular value shrinkage, under a Whiten-Shrink- re-Color framework. To wit, one forms the pure-noise sample covariance $\widehat{\mathbf{\Sigma}}$, and 1) uses $\widehat{\mathbf{\Sigma}}$ to pseudo-whiten the observed signal-plus-noise matrix; 2) applies optimal singular value shrinkage to the resulting pseudo-whitened matrix; and finally 3) performs a re-coloring step. Our main contribution is the derivation of the optimal singular shrinker to be used in this compound procedure. To this end, we proved new results on the spectrum of the spiked F-matrix ensemble, which may be of independent interest.

As one would expect, we demonstrated that our optimally-tuned WSC denoiser outperforms OptShrink, the optimally-tuned singular value shrinkage without noise whitening [47]), provided that the number of pure-noise samples is sufficiently large (see Section 6). A particularly appealing quality of the proposed estimator is its simplicity, both conceptually and implementation-wise. Indeed, the idea of noise whitening is classical within signal processing. Furthermore, the method comes with precise performance guarantees, including estimates of the AMSE and the inner products between the signal PCs and their estimates, which can be useful in practice. There is no reason to believe, however, that this approach is optimal among all denoising procedures. Devising new methods to make better use of the available side information (and/or deriving tight lower bounds on the attainable MSE) is an interesting direction for future research.

A natural extension of our method would be to replace the noise sample covariance $\widehat{\mathbf{\Sigma}}$ with a better covariance estimator: note that while the sample covariance has generically an optimal *estimation rate*, one can often construct estimators that attain a smaller estimation error (e.g., smaller constant prefactors); see for example the shrinkage estimator of [36]. To implement such modified WSC scheme, one needs to calculate precise analytic formulas for the singular values and singular vector angles of the whitened and recolored data matrix. For more sophisticated covariance estimators (such as [36]), this calculation appears to be challenging.

Acknowledgements

ER was partially supported by a Hebrew University of Jerusalem Einstein-Kaye scholarship. MG was partially supported by ISF grant no. 871/22. WL acknowledges support from NSF BIGDATA award IIS-1837992, BSF award 2018230, and NSF CAREER award DMS-2238821.

ER wishes to thank David Donoho, Apratim Dey and Michael Feldman for interesting discussions regarding this manuscript.

References

[1] Joshua Agterberg, Zachary Lubberts, and Carey Priebe. Entrywise estimation of singular vectors of low-rank matrices with heteroskedasticity and dependence. *IEEE Transactions on Information Theory*, 2022.

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

[3] Theodore Wilbur Anderson. Estimating linear statistical relationships. *Annals of Statistics*, 12(1):1–45, 03 1984.

[4] Theodore Wilbur Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003.

[5] Zhi-Dong Bai and Jack W Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345, 1998.

[6] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

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

[8] 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.

[9] Joshua K Behne and Galen Reeves. Fundamental limits for rank-one matrix estimation with groupwise heteroskedasticity. In *International Conference on Artificial Intelligence and Statistics*, pages 8650–8672. PMLR, 2022.

[10] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[11] 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.

[12] 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.

[13] T Tony Cai, Zhao Ren, and Harrison H Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016.

[14] Benoît Collins. Moments and cumulants of polynomial random variables on unitary groups, the Itzykson-Zuber integral, and free probability. *International Mathematics Research Notices*, 2003(17):953–982, 2003.

[15] Benoît Collins and Piotr Śniady. Integration with respect to the haar measure on unitary, orthogonal and symplectic group. *Communications in Mathematical Physics*, 264(3):773–795, 2006.

[16] Lucilio Cordero-Grande, Daan Christiaens, Jana Hutter, Anthony N. Price, and Jo V. Hajnal. Complex diffusion-weighted image estimation via matrix recovery under general noise models. *NeuroImage*, 200:391–404, 2019.

[17] Romain Couillet and Merouane Debbah. *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

[18] Prathapasinghe Dharmawansa, Iain M Johnstone, and Alexei Onatski. Local asymptotic normality of the spectrum of high-dimensional spiked F-ratios. *arXiv preprint arXiv:1411.3875*, 2014.

[19] Prathapasinghe Dharmawansa, Boaz Nadler, and Ofer Shwartz. Roy's largest root under rank-one perturbations: The complex valued case and applications. *Journal of multivariate analysis*, 174:104524, 2019.

[20] Xiucai Ding and Fan Yang. Spiked separable covariance matrices and principal components. *The Annals of Statistics*, 49(2):1113–1138, 2021.

[21] Edgar Dobriban. Permutation methods for factor analysis and PCA. *Annals of Statistics*, 48(5):2824–2847, 2020.

[22] David L Donoho, Matan Gavish, and Iain M Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of statistics*, 46(4):1742, 2018.

[23] David L Donoho, Matan Gavish, and Elad Romanov. ScreeNOT: Exact MSE-optimal singular value thresholding in correlated noise. *arXiv preprint arXiv:2009.12297*, 2020.

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

[25] 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.

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

[27] Xiao Han, Guangming Pan, Bo Zhang, et al. The Tracy-Widom law for the largest eigenvalue of F type matrices. *Annals of Statistics*, 44(4):1564–1592, 2016.

[28] David Hong, Laura Balzano, and Jeffrey A Fessler. Asymptotic performance of PCA for high-dimensional heteroscedastic data. *Journal of multivariate analysis*, 167:435–452, 2018.

[29] David Hong, Kyle Gilman, Laura Balzano, and Jeffrey A Fessler. HePPCAT: Probabilistic PCA for data with heteroscedastic noise. *IEEE Transactions on Signal Processing*, 69:4819–4834, 2021.

[30] David Hong, Fan Yang, Jeffrey A Fessler, and Laura Balzano. Optimally weighted PCA for high-dimensional heteroscedastic data. *arXiv preprint arXiv:1810.12862*, 2018.

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

[32] Iain M Johnstone. Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy-Widom limits and rates of convergence. *Annals of statistics*, 36(6):2638, 2008.

[33] Iain M Johnstone and Alexei Onatski. Testing in high-dimensional spiked models. *The Annals of Statistics*, 48(3):1231–1254, 2020.

[34] IM Johnstone and Boaz Nadler. Roy's largest root test under rank-one alternatives. *Biometrika*, 104(1):181–193, 2017.

[35] Boris Landa, Thomas TCK Zhang, and Yuval Kluger. Biwhitening reveals the rank of a count matrix. *arXiv preprint arXiv:2103.13840*, 2021.

[36] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012.

[37] William Leeb. Rapid evaluation of the spectral signal detection threshold and Stieltjes transform. *Advances in Computational Mathematics*, 47(4):1–29, 2021.

[38] William Leeb. Optimal singular value shrinkage for operator norm loss: Extending to non-square matrices. *Statistics & Probability Letters*, 186:109472, 2022.

[39] William Leeb and Elad Romanov. Optimal spectral shrinkage and PCA with heteroscedastic noise. *IEEE Transactions on Information Theory*, 67(5):3009–3037, 2021.

[40] William E Leeb. Matrix denoising for weighted loss functions and heterogeneous signals. *SIAM Journal on Mathematics of Data Science*, 3(3):987–1012, 2021.

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

[42] Tzu-Chi Liu, Yi-Wen Liu, and Hau-Tieng Wu. Denoising click-evoked otoacoustic emission signals by optimal shrinkage. *The Journal of the Acoustical Society of America*, 149(4):2659–2670, 2021.

[43] Charles F. Van Loan. Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis*, 13(1):76–83, 1976.

[44] James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

[45] Brian E. Moore, Raj Rao Nadakuditi, and Jeffrey A. Fessler. Improved robust PCA using low-rank denoising with optimal singular value shrinkage. In *Statistical Signal Processing (SSP), 2014 IEEE Workshop on*. IEEE, 2014.

[46] Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

[47] 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.

[48] Raj Rao Nadakuditi and Jack W Silverstein. Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples. *IEEE Journal of Selected Topics in Signal Processing*, 4(3):468–480, 2010.

[49] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

[50] Patrick O Perry. *Cross-validation for unsupervised learning*. Stanford University, 2009.

[51] Mark J. Schervish. A review of multivariate analysis. *Statistical Science*, 2(4):396–413, 1987.

[52] 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.

[53] Jack W Silverstein. The limiting eigenvalue distribution of a multivariate F matrix. *SIAM Journal on Mathematical Analysis*, 16(3):641–646, 1985.

[54] Jack W Silverstein and ZD Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192, 1995.

[55] Petre Stoica and Mats Cedervall. Detection tests for array processing in unknown correlated noise fields. *IEEE transactions on signal processing*, 45(9):2351–2362, 1997.

[56] Pei-Chun Su and Hau-Tieng Wu. Optimal shrinkage of singular values under high-dimensional noise with separable covariance structure. *arXiv preprint arXiv:2207.03466*, 2022.

[57] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[58] Harry L Van Trees. *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons, 2004.

[59] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[60] Dan V Voiculescu, Ken J Dykema, and Alexandru Nica. *Free random variables*. Number 1. American Mathematical Soc., 1992.

[61] Kenneth W Wachter et al. The limiting empirical measure of multiple discriminant ratios. *The Annals of Statistics*, 8(5):937–957, 1980.

[62] Qinwen Wang, Jianfeng Yao, et al. Extreme eigenvalues of large-dimensional spiked Fisher matrices with application. *The Annals of Statistics*, 45(1):415–460, 2017.

[63] Junshan Xie, Yicheng Zeng, and Lixing Zhu. Limiting laws for extreme eigenvalues of large-dimensional spiked Fisher matrices with a divergent number of spikes. *Journal of Multivariate Analysis*, page 104742, 2021.

[64] YQ Yin, ZD Bai, and PR Krishnaiah. Limiting behavior of the eigenvalues of a multivariate F matrix. *Journal of Multivariate Analysis*, 13(4):508–516, 1983.

[65] Anru R Zhang, T Tony Cai, and Yihong Wu. Heteroskedastic PCA: Algorithm, optimality, and applications. *The Annals of Statistics*, 50(1):53–80, 2022.

[66] LC Zhao, Paruchuri R Krishnaiah, and ZD Bai. On detection of the number of signals when the noise covariance matrix is arbitrary. *Journal of Multivariate Analysis*, 20(1):26–49, 1986.

[67] Zhaoda Zhu, Simon Haykin, and Xinping Huang. Estimating the number of signals using reference noise samples. *IEEE Transactions on Aerospace and Electronic Systems*, 27(3):575–579, 1991.

A Auxiliary Technical Results

A.1 Elementary concentration lemmas

The following elementary concentration results are used throughout the paper:

Lemma 7. *Let $\mathbf{A}_p \in \mathbb{R}^{p\times p}$ be a sequence of bounded matrices, $\mathbf{a} \in \mathbb{R}^p$ a bounded vector, and $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, p^{-1}\mathbf{I}_{p\times p})$. Then*

$$\|\mathbf{g}\|^2 \simeq 1, \quad \mathbf{g}^\top \mathbf{A}_p \mathbf{g} \simeq p^{-1}\mathrm{tr}(\mathbf{A}_p), \quad \mathbf{g}^\top \mathbf{a} \simeq 0\,.$$

Lemma 8. *Let $\mathbf{A}_p \in \mathbb{R}^{p\times p}$ be a sequence of bounded, orthogonally invariant, random matrices, namely, for any $\mathbf{O} \in O(p)$, $\mathbf{O}\mathbf{A}_p\mathbf{O}^\top \overset{d}{=} \mathbf{A}_p$. Let $\mathbf{a}_p, \mathbf{b}_p \in \mathbb{R}^n$ be two sequences of bounded vectors. Then*

$$\mathbf{a}_p^\top \mathbf{A}_p \mathbf{b}_p \simeq \left(\mathbf{a}_p^\top \mathbf{b}_p\right) \cdot p^{-1}\,\mathrm{tr}(\mathbf{A}_p)\,.$$

Lemma 9. *Let $A_p \in \mathbb{R}^{p\times n}$ be a sequence of left orthogonally invariant random matrices, namely, for any $\mathbf{O} \in O(p)$, $\mathbf{O}\mathbf{A}_p \overset{d}{=} \mathbf{A}_p$. Let $\mathbf{a}_p \in \mathbb{R}^p, \mathbf{b}_p \in \mathbb{R}^n$ be bounded. Then $\mathbf{a}_p^\top \mathbf{A}\mathbf{b}_p \simeq 0$.*

A.2 Free Probability and calculation of mixed moments

Free probability is a theory of noncommutative random variables, originally introduced by Voiculescu [60]. We briefly describe several elementary results that are used in our calculations. For more background, see e.g. [17, 44].

Definition 1 (Limiting joint law)**.** *Let $\mathbb{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_l\}$ be l (sequences of) of p-by-p matrices with bounded operator norm. We say $\mathbb{X}$ has an a.s. limiting joint law if for any non-commutative polynomial $\mathcal{P} \in \mathbb{C}\langle x_1, \ldots, x_l, y_1, \ldots, y_l\rangle$,[6] the a.s. limit*

$$\overline{\mathrm{tr}}\,\mathcal{P}\left(\mathbf{X}_1, \ldots, \mathbf{X}_l, \mathbf{X}_1^\top, \ldots, \mathbf{X}_l^\top\right) = \lim_{p\to\infty} \frac{1}{p}\,\mathrm{tr}\,\mathcal{P}\left(\mathbf{X}_1, \ldots, \mathbf{X}_l, \mathbf{X}_1^\top, \ldots, \mathbf{X}_l^\top\right) \tag{84}$$

exists. The mapping $\mathbb{C}\langle x_1, \ldots, x_l, y_1, \ldots, y_l\rangle \to \mathbb{R}$ taking a polynomial $\mathcal{P}$ to its corresponding limit is called the joint law of $\mathcal{S}$.

Remark 4. *Note that the joint law of $\mathcal{S}_p$ is completely determined by its (non-commutative) joint moments:*

$$z_{i_1} \cdots z_{i_t}\,, z \in \{x, y\}\,,\ i_j \in \{1, \ldots, l\}\ .$$

Moreover, if $\mathbb{X} = \{\mathbf{X}\}$ is a single symmetric matrix with an LESD dF, then it limiting law is

$$x^k \mapsto \int x^k dF(x)\,.$$

Definition 2 (Free independence)**.** *Suppose that $\mathbb{X}_1, \ldots, \mathbb{X}_k$ each have a joint limiting law. They are (asymptotically) freely independent if the following holds: let $\mathcal{P}_j(\mathbb{X}_{i_j})$ be any non-commutative polynomials in $\mathbb{X}_{i_j}$, then*

$$\overline{\mathrm{tr}}\,\left\{\left[\mathcal{P}_1(\mathbb{X}_{i_1}) - \overline{\mathrm{tr}}(\mathcal{P}_1(\mathbb{X}_{i_1}))\mathbf{I}\right] \cdots \left[\mathcal{P}_t(\mathbb{X}_{i_t}) - \overline{\mathrm{tr}}(\mathcal{P}_t(\mathbb{X}_{i_t}))\mathbf{I}\right]\right\} = 0\,, \tag{85}$$

whenever adjacent indices are always different: $i_1 \neq i_2$, $i_2 \neq i_3$ etc.

[6] By that, we mean a polynomial in $2l$ non-commutative variables. For example, $x_1x_2 \neq x_2x_1$.

Free independence is a powerful tool for computing traces involving multiple random matrices. A fundamental result connecting free probability with random matrices states that independent unitarily/orthogonally invariant random matrices are asymptotically freely independent. To our knowledge, the complex (unitary) first appeared in the work of Voiculescu [60]; for the real case, we cite a result of Collins and Śniady [15] (see also [14]). The following is essentially [15, Theorem 5.2]:

Theorem 6 (Asymptotic freedom for orthogonally invariant matrices)**.** *Let* $(\mathbf{X}_1), \ldots, (\mathbf{X}_k)$ *be* k *(sequences of)* p*-by-*p *matrices, so that each matrix has an a.s. limiting law. Let* $\mathbf{O}_2, \ldots, \mathbf{O}_k \sim \mathrm{Haar}(O(p))$*. Then*

$$(\mathbf{X}_1), (\mathbf{O}_2\mathbf{X}_2\mathbf{O}_2^\top), \ldots, (\mathbf{O}_k\mathbf{X}_k\mathbf{O}_k^\top)$$

are asymptotically freely independent.

B Proof of Lemma 6

The proof consists of repeated applications of (85). It is straightforward to verify that

$$G_{2,k}(z) = \overline{\mathrm{tr}}\left(\mathbf{A}_z\mathbf{\Sigma}\mathbf{A}_z^\top\mathbf{C}_k\right), \quad \text{where} \quad \mathbf{A}_z := (z\mathbf{S}-\mathbf{E})^{-1}\mathbf{S},\ \mathbf{C}_k = \mathbf{\Sigma}^{-1/2}\mathbf{D}_k\mathbf{D}_k^\top\mathbf{\Sigma}^{-1/2}. \quad (86)$$

Since $\mathbf{A}_z$ is orthogonal invariant and has bounded norm (since $z \notin [\theta_{\min}, \theta_{\max}]$), it is freely independent of all deterministic matrices. To carry out the computation, we apply (85) successively. For brevity, denote $\mathbf{\Delta}_z = \mathbf{A}_z - \overline{\mathrm{tr}}(\mathbf{A}_z)\mathbf{I}$. Then

$$\begin{aligned} G_{2,k}(z) &= \overline{\mathrm{tr}}\left(\mathbf{\Delta}_z\mathbf{\Sigma}\mathbf{A}_z^\top\mathbf{C}_k\right) + \overline{\mathrm{tr}}(\mathbf{A}_z)\overline{\mathrm{tr}}\left(\mathbf{\Sigma}\mathbf{A}_z^\top\mathbf{C}_k\right) \\ &= \overline{\mathrm{tr}}\left(\mathbf{\Delta}_z\mathbf{\Sigma}\mathbf{A}_z^\top\mathbf{C}_k\right) + \left(\overline{\mathrm{tr}}(\mathbf{A}_z)\right)^2\overline{\mathrm{tr}}(\mathbf{C}_k\mathbf{\Sigma}) = \overline{\mathrm{tr}}\left(\mathbf{\Delta}_z\mathbf{\Sigma}\mathbf{A}_z^\top\mathbf{C}_k\right) + (\overline{s}(z))^2, \end{aligned}$$

where the last equality uses the assumption $\overline{\mathrm{tr}}(\mathbf{D}_k\mathbf{D}_k^\top)$. We simplify further the first term on the r.h.s.,

$$\overline{\mathrm{tr}}\left(\mathbf{\Delta}_z\mathbf{\Sigma}\mathbf{A}_z^\top\mathbf{C}_k\right) = \overline{\mathrm{tr}}\left(\mathbf{\Delta}_z\mathbf{\Sigma}\mathbf{\Delta}_z^\top\mathbf{C}_k\right) + \overline{\mathrm{tr}}(\mathbf{A}_z^\top)\underbrace{\overline{\mathrm{tr}}\left(\mathbf{\Delta}_z\mathbf{\Sigma}\mathbf{C}_k\right)}_{=0} = \overline{\mathrm{tr}}\left(\mathbf{\Delta}_z\mathbf{\Sigma}\mathbf{\Delta}_z^\top\mathbf{C}_k\right).$$

Next,

$$\begin{aligned} \overline{\mathrm{tr}}\left(\mathbf{\Delta}_z\mathbf{\Sigma}\mathbf{\Delta}_z^\top\mathbf{C}_k\right) &= \overline{\mathrm{tr}}\left(\mathbf{\Delta}_z(\mathbf{\Sigma}-\mu\mathbf{I})\mathbf{\Delta}_z^\top\mathbf{C}_k\right) + \mu\cdot\overline{\mathrm{tr}}\left(\mathbf{\Delta}_z\mathbf{\Delta}_z^\top\mathbf{C}_k\right) \\ &= \underbrace{\overline{\mathrm{tr}}\left(\mathbf{\Delta}_z(\mathbf{\Sigma}-\mu\mathbf{I})\mathbf{\Delta}_z^\top(\mathbf{C}_k-\tau_k\mathbf{I})\right)}_{=0} + \tau_k\underbrace{\overline{\mathrm{tr}}\left(\mathbf{\Delta}_z(\mathbf{\Sigma}-\mu\mathbf{I})\mathbf{\Delta}_z^\top\right)}_{=0} + \mu\cdot\overline{\mathrm{tr}}\left(\mathbf{\Delta}_z\mathbf{\Delta}_z^\top\mathbf{C}_k\right) \\ &= \mu\tau_k\cdot\overline{\mathrm{tr}}(\mathbf{\Delta}_z\mathbf{\Delta}_z^\top). \end{aligned}$$

Lastly, $\overline{\mathrm{tr}}(\mathbf{\Delta}_z\mathbf{\Delta}_z^\top) = \overline{\mathrm{tr}}(\mathbf{A}_z\mathbf{A}_z^\top) - (\overline{\mathrm{tr}}(\mathbf{A}_z))^2 = \Upsilon_2(z) - (\overline{s}(z))^2$. Collecting all the terms above, we obtain the claimed formula for $G_{2,k}(z)$. ■

C Closed-form formulas for Υ_1, Υ_2

In this section we prove the closed-form formulas (24) and (25) for the mixed traces

$$\Upsilon_1(z) = \lim_{p\to\infty} p^{-1}\mathrm{tr}(z\mathbf{S}-\mathbf{E})^{-1}\mathbf{S}^2, \quad \Upsilon_2(z) = \lim_{p\to\infty} p^{-1}\mathrm{tr}(z\mathbf{S}-\mathbf{E})^{-2}\mathbf{S}^2,$$

where $z \in \mathbb{C} \setminus [\theta_{\min}^2, \theta_{\max}^2]$.

First, because the formula (23) for the Stieltjes transform is only applicable for arguments smaller than the left edge of the Marchenko-Pastur law, $0 < z < (1-\sqrt{\beta})^2$, we begin with:

Lemma 10. *For any $z \in (\theta_{\max}^2, \infty)$, one has $0 < -\underline{s}(z) < (1-\sqrt{\beta})^2$.*

Proof. Since $-\underline{s}(z)$ is positive and decreasing for $z > \theta_{\max}^2$, it suffices to show that $-\underline{s}(\theta_{\max}^2) \leq (1-\sqrt{\beta})^2$. A straightforward calculation gives $-\underline{s}(\theta_{\max}^2) = f(\gamma;\beta) = f_1(\gamma;\beta) + f_2(\gamma;\beta)$ where

$$f_1(\gamma;\beta) = \frac{\beta(\beta - \sqrt{\beta+\gamma-\beta\gamma})}{\beta+\gamma}, \tag{87}$$

and

$$f_2(\gamma;\beta) = \frac{1-\sqrt{\beta+\gamma-\beta\gamma}}{1-\gamma}\,; \tag{88}$$

note that $f_2(\gamma;\beta)$ is well-defined and differentiable at $\gamma = 1$, where its value is $f_2(1;\beta) = \frac{1}{2}(1-\beta)$.

We claim that the function $\gamma \mapsto f(\gamma;\beta)$, $\gamma \in [0,\infty)$, is maximized at $\gamma = 0$. This, in turn, implies $-\underline{s}(\theta_{\max}^2) \leq f(0;\beta) = \beta - \sqrt{\beta} + 1 - \sqrt{\beta} = (1-\sqrt{\beta})^2$ as required. To show this, let us compute the derivative of $f(\cdot;\beta)$. One has

$$\partial_\gamma f_1(\gamma;\beta) = \beta \frac{-\frac{1-\beta}{2\sqrt{\beta+\gamma-\beta\gamma}}(\beta+\gamma) - (\beta - \sqrt{\beta+\gamma-\beta\gamma})}{(\beta+\gamma)^2} = \frac{\beta\left(\beta - \sqrt{\beta+\gamma-\beta\gamma}\right)^2}{2(\beta+\gamma)^2\sqrt{\beta+\gamma-\beta\gamma}}\,, \tag{89}$$

and

$$\partial_\gamma f_2(\gamma;\beta) = \frac{-\frac{1-\beta}{2\sqrt{\beta+\gamma-\beta\gamma}}(1-\gamma) + (1 - \sqrt{\beta+\gamma-\beta\gamma})}{(1-\gamma)^2} = -\frac{\left(1 - \sqrt{\beta+\gamma-\beta\gamma}\right)^2}{2(1-\gamma)^2\sqrt{\beta+\gamma-\beta\gamma}}\,, \tag{90}$$

(which may also be continuously extended to $\gamma = 1$). Since $f(\gamma;\beta) \to 0$ as $\gamma \to \infty$, it is enough to show that the only solution to $\partial_\gamma f(\gamma;\beta) = 0$ for $\gamma \in [0,\infty]$ is $\gamma = 0$. First, let us consider $\gamma = 1$. One may readily calculate (e.g., using L'Hôpital's rule): $\partial_\gamma f_2(1;\beta) = -\frac{1}{8}(1-\beta)^2$ and $\partial_\gamma f_1(1;\beta) = \frac{\beta(1-\beta)^2}{2(1+\beta)^2}$, so $\partial_\gamma f(1;\beta) = -\frac{(1-\beta)^4}{8(1+\beta)^2} < 0$. We next consider values $\gamma \neq 1$. From (89) and (90), if $\partial_\gamma f(\gamma;\beta) = 0$ then either

$$(1-\gamma)\sqrt{\beta}\left(\beta - \sqrt{\beta+\gamma-\beta\gamma}\right) = (\beta+\gamma)\left(1 - \sqrt{\beta+\gamma-\beta\gamma}\right), \tag{91}$$

$$(1-\gamma)\sqrt{\beta}\left(\beta - \sqrt{\beta+\gamma-\beta\gamma}\right) = -(\beta+\gamma)\left(1 - \sqrt{\beta+\gamma-\beta\gamma}\right). \tag{92}$$

Making a change of variables $r = \beta+\gamma-\beta\gamma$, so that $\gamma = \frac{r-\beta}{1-\beta}$, equations (91) and (92) may be rewritten in terms of r as follows:

$$\sqrt{\beta}\left(\beta - \sqrt{r}\right)\left(\frac{1-r}{1-\beta}\right) \pm \left(1-\sqrt{r}\right)\left(\frac{r-\beta^2}{1-\beta}\right) = 0\,. \tag{93}$$

Using $r - \beta^2 = (\sqrt{r}-\beta)(\sqrt{r}+\beta)$ and $1 - r = (1-\sqrt{r})(1+\sqrt{r})$, (93) is equivalent to:

$$(\beta - \sqrt{r})(1-\sqrt{r})\left[\sqrt{\beta}(1+\sqrt{r}) \pm (\sqrt{r}+\beta)\right] = 0. \tag{94}$$

The roots are $r = \beta^2$, $r = 1$, and $r = \beta$. Since $\gamma = (r-\beta)/(1-\beta)$, and we only consider $\gamma \in [0,\infty) \setminus \{1\}$ (recall that $\gamma = 1$ was treated separately) it follows that the only solution to $\partial_\gamma f(\gamma;\beta) = 0$ with $\gamma \in [0,\infty)$ is $\gamma = 0$, concluding the proof. □

Let $\mathbf{S} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^\top$ be an eigen-decomposition, $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Since $\mathbf{E}$ is orthogonally invariant and independent of $\mathbf{S}$, $\mathbf{W}^\top\mathbf{E}\mathbf{W} \stackrel{d}{=} \mathbf{E}$ and is independent of $\mathbf{\Lambda}$. Consider the resolvent

$$\mathbf{R}_z(w) = (w\mathbf{I} + z\mathbf{\Lambda} - \mathbf{E})^{-1}\,. \tag{95}$$

Clearly,

$$\Upsilon_1(z) = \lim_{p\to\infty} p^{-1}\mathrm{tr}\left[\mathbf{R}_z(0)\mathbf{\Lambda}^2\right] = \lim_{p\to\infty} p^{-1}\sum_{i=1}^{p}[\mathbf{R}_z(0)]_{ii}\lambda_i^2,$$
$$\Upsilon_2(z) = -\lim_{p\to\infty} p^{-1}\mathrm{tr}\left[\frac{\partial}{\partial w}\mathbf{R}_z(0)\mathbf{\Lambda}^2\right] = -\lim_{p\to\infty} p^{-1}\sum_{i=1}^{p}\frac{\partial}{\partial w}[\mathbf{R}_z(0)]_{ii}\lambda_i^2\,,$$

Recall: since $z > \theta_{\max}^2$, assuming small enough w, $\mathbf{R}_z(w)$ is a.s. well-defined and has bounded operator norm. The main ingredient of the proof consists of deriving formulas for the *individual* diagonal entries of $\mathbf{R}_z(w)$. We remark that the calculation below uses rather standard ideas from random matrix theory, see for example the book [6].

For brevity, denote $\mathbf{A}_{z,w} = w\mathbf{I} + z\mathbf{\Lambda} - \mathbf{E}$ and so $\mathbf{R}_z(w) = \mathbf{A}_{z,w}^{-1}$. Let $i \in [n]$ be any coordinate. Denote by $\mathbf{P}_i \in \mathbb{R}^{(p-1)\times p}$ the projection onto the coordinate set $[p]\setminus\{i\}$. Let $\mathbf{e}_1, \ldots, \mathbf{e}_p \in \mathbb{R}^p$ be the standard basis vectors. Up to a permutation of the coordinates, $\mathbf{A}_{z,w}$ has the block form:

$$\mathbf{A}_{z,w} = \begin{bmatrix} [\mathbf{A}_{z,w}]_{i,i} & \mathbf{e}_i^\top\mathbf{A}_{z,w}\mathbf{P}_i^\top \\ \mathbf{P}_i\mathbf{A}_{z,w}\mathbf{e}_i & \mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top \end{bmatrix} = \begin{bmatrix} w + z\lambda_i - n^{-1}[\mathbf{Z}\mathbf{Z}^\top]_{ii} & -n^{-1}\mathbf{e}_i^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{P}_i^\top \\ -n^{-1}\mathbf{P}_i\mathbf{Z}\mathbf{Z}^\top\mathbf{e}_i & \mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top\,. \end{bmatrix}$$

Applying the block matrix inversion formula,

$$[\mathbf{A}_{z,w}^{-1}]_{ii} = \left[w + z\lambda_i - n^{-1}[\mathbf{Z}\mathbf{Z}^\top]_{ii} - (n^{-1}\mathbf{e}_i^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{P}_i^\top)(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}(n^{-1}\mathbf{P}_i\mathbf{Z}\mathbf{Z}^\top\mathbf{e}_i)\right]^{-1}\,,$$

equivalently,

$$\frac{1}{[\mathbf{R}_z(w)]_{ii}} = w + z\lambda_i - n^{-1}[\mathbf{Z}\mathbf{Z}^\top]_{ii} - (n^{-1}\mathbf{e}_i^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{P}_i^\top)(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}(n^{-1}\mathbf{P}_i\mathbf{Z}\mathbf{Z}^\top\mathbf{e}_i)\,. \tag{96}$$

We will show that the r.h.s. of (96) concentrates around a deterministic quantity.

We start with the following.

Lemma 11. *A.s.,*

$$\max_{1\le i\le n}\left|(n^{-1}\mathbf{e}_i^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{P}_i^\top)(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}(n^{-1}\mathbf{P}_i\mathbf{Z}\mathbf{Z}^\top\mathbf{e}_i) - n^{-1}\mathrm{tr}\left[\mathbf{P}_i^\top(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}\mathbf{P}_i\mathbf{E}\right]\right| \longrightarrow 0\,,$$
$$\max_{1\le i\le n}\left|(n^{-1}\mathbf{e}_i^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{P}_i^\top)\frac{\partial}{\partial w}(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}(n^{-1}\mathbf{P}_i\mathbf{Z}\mathbf{Z}^\top\mathbf{e}_i) - n^{-1}\mathrm{tr}\left[\mathbf{P}_i^\top\frac{\partial}{\partial w}(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}\mathbf{P}_i\mathbf{E}\right]\right| \longrightarrow 0\,.$$

Proof. Observe that $\mathbf{Z}^\top\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$. Moreover, this random vector is independent of $\mathbf{P}_i\mathbf{Z}$, hence also of $\mathbf{A}_{z,w}$. Furthermore, the random matrices $n^{-1}\mathbf{Z}^\top\mathbf{P}_i^\top(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}\mathbf{P}_i\mathbf{Z}$, as well as their w-derivatives, have operator norm bounded (a.s.) by a constant[7]. The result follows by the Hanson-Wright inequality (cf. [59, Theorem 6.2.1]) and a union bound over $1 \le i \le n$. □

We next consider the trace in Lemma 11.

Lemma 12. *A.s.,*

$$\max_{1\le i\le n}\left|n^{-1}\mathrm{tr}\left[\mathbf{P}_i^\top(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}\mathbf{P}_i\mathbf{E}\right] - n^{-1}\mathrm{tr}\left[\mathbf{A}_{z,w}^{-1}\mathbf{E}\right]\right| \longrightarrow 0\,,$$
$$\max_{1\le i\le n}\left|n^{-1}\mathrm{tr}\left[\mathbf{P}_i^\top\frac{\partial}{\partial w}(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}\mathbf{P}_i\mathbf{E}\right] - n^{-1}\mathrm{tr}\left[\frac{\partial}{\partial w}\mathbf{A}_{z,w}^{-1}\mathbf{E}\right]\right| \longrightarrow 0\,.$$

[7]Note that by eigenvalue interlacing, $\lambda_{\min}(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top) \ge \lambda_{\min}(\mathbf{A}_{z,w})$, hence $\|(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}\| \le \|\mathbf{A}_{z,w}^{-1}\|$.

Proof. By the block matrix inversion formula,

$$\mathbf{P}_i\mathbf{A}_{z,w}^{-1}\mathbf{P}_i^\top = \left(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top - [\mathbf{A}_{z,w}]_{ii}{}^{-1}\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{e}_i\mathbf{e}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top\right)^{-1}.$$

Thus,

$$\mathbf{P}_i\mathbf{A}_{z,w}^{-1}\mathbf{P}_i^\top - (\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1} = \mathbf{P}_i\mathbf{A}_{z,w}^{-1}\mathbf{P}_i^\top\left([\mathbf{A}_{z,w}]_{ii}{}^{-1}\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{e}_i\mathbf{e}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top\right)(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}$$

is rank 1, and clearly has bounded operator norm. Write $\mathbf{P}_i\mathbf{A}_{z,w}^{-1}\mathbf{P}_i^\top - (\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1} = \mathbf{q}\mathbf{q}^\top$, so

$$\begin{aligned}\max_{1\le i\le n}\left|n^{-1}\mathrm{tr}\left[\mathbf{P}_i^\top(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}\mathbf{P}_i\mathbf{E}\right] - n^{-1}\mathrm{tr}\left[\mathbf{P}_i^\top(\mathbf{P}_i(\mathbf{A}_{z,w})^{-1}\mathbf{P}_i^\top)\mathbf{P}_i\mathbf{E}\right]\right| &= \max_{1\le i\le n}\left|n^{-1}\mathbf{q}_i^\top(\mathbf{P}_i\mathbf{E}\mathbf{P}_i^\top)\mathbf{q}_i\right| \\ &\le \max_{1\le i\le n} n^{-1}\|\mathbf{q}_i\|^2\|\mathbf{P}_i\mathbf{E}\mathbf{P}_i^\top\| \longrightarrow 0\,.\end{aligned}$$

Moreover, $\mathbf{I} - \mathbf{P}_i^\top\mathbf{P}_i = \mathbf{e}_i\mathbf{e}_i^\top$ is rank 1, allowing us to deduce

$$\max_{1\le i\le n}\left|n^{-1}\mathrm{tr}\left[\mathbf{P}_i^\top\mathbf{P}_i(\mathbf{A}_{z,w})^{-1}\mathbf{P}_i^\top\mathbf{P}_i\mathbf{E}\right] - n^{-1}\mathrm{tr}\left[(\mathbf{A}_{z,w})^{-1}\mathbf{E}\right]\right| \longrightarrow 0$$

by a similar argument. This establishes the first claim of the Lemma. The second claim (pertaining to the w-derivatives) follows by a similar calculation. □

Now,

$$\frac{1}{n}\mathrm{tr}[\mathbf{A}_{z,w}^{-1}\mathbf{E}] = -\frac{p}{n} + \frac{1}{n}\mathrm{tr}\left[\mathbf{A}_{z,w}^{-1}(z\mathbf{\Lambda} + w\mathbf{I})\right] = \gamma\left(-1 + w\frac{1}{p}\mathrm{tr}(\mathbf{R}_z(w)) + z\frac{1}{p}\mathrm{tr}(\mathbf{R}_z(w)\mathbf{\Lambda})\right). \quad (97)$$

Setting $w = 0$, by Propositions 1 and 2,

$$\begin{aligned}&\frac{1}{p}\mathrm{tr}(\mathbf{R}_z(0)) \overset{d}{=} \frac{1}{p}\mathrm{tr}(z\mathbf{S} - \mathbf{E})^{-1} \longrightarrow -\zeta(z)\,, \\ &\frac{1}{p}\mathrm{tr}(\mathbf{R}_z(0)\mathbf{\Lambda}) \overset{d}{=} \frac{1}{p}\mathrm{tr}((z\mathbf{S} - \mathbf{E})^{-1}\mathbf{S}) \longrightarrow -\overline{s}(z) \\ &\frac{\partial}{\partial w}\frac{1}{p}\mathrm{tr}(\mathbf{R}_z(w)\mathbf{\Lambda})\Big|_{w=0} \overset{d}{=} -\frac{1}{p}\mathrm{tr}(z\mathbf{S} - \mathbf{E})^{-2}\mathbf{S}) \longrightarrow -\zeta'(z)\,.\end{aligned} \quad (98)$$

Combining (96) with Lemmas 11, 12 and Eqs. (97), (98), along with $\max_{1\le i\le n}|[n^{-1}\mathbf{Z}\mathbf{Z}^\top]_{ii} - 1| \to 0$, yields $\max_{1\le i\le n}|[\mathbf{R}_z(0)]_{ii} - \rho_i(z)| \to 0$, where

$$\frac{1}{\rho_i(z)} = z\lambda_i - 1 + \gamma(1 + z\overline{s}(z)) = z(\lambda_i + \underline{s}(z))\,. \quad (99)$$

The second equality in (99) uses the relation $\underline{s}(z) = \gamma\overline{s}(z) - (1-\gamma)\frac{1}{z}$ between the Stieltjes and the associated transforms. Furthermore, differentiating (96) with respect to w yields

$$-\frac{\frac{\partial}{\partial w}[\mathbf{R}_z(0)]_{ii}}{[\mathbf{R}_z(0)]_{ii}^2} = 1 - \frac{\partial}{\partial w}(n^{-1}\mathbf{e}_i^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{P}_i^\top)(\mathbf{P}_i\mathbf{A}_{z,w}\mathbf{P}_i^\top)^{-1}(n^{-1}\mathbf{P}_i\mathbf{Z}\mathbf{Z}^\top\mathbf{e}_i)\Big|_{w=0}.$$

Consequently, $\max_{1\le i\le n}|\frac{\partial}{\partial w}[\mathbf{R}_z(0)]_{ii} - \tilde{\rho}_i(z)| \to 0$ where

$$\tilde{\rho}_i(z) = -R_i(z)^2\left(1 + \gamma[\zeta(z) + z\zeta'(z)]\right) = -\frac{1 + \gamma(\zeta(z) + z\zeta'(z))}{z^2}\cdot\frac{1}{(\lambda_i + \underline{s}(z))^2}\,. \quad (100)$$

Equipped with Eqs. (99) and (100), we now conclude the calculation. Starting with Υ_1,

$$\Upsilon_1(z) \simeq \frac{1}{p}\sum_{i=1}^{p} \rho_i(z)\lambda_i^2 = \frac{1}{z}\cdot\frac{1}{p}\sum_{i=1}^{p}\frac{\lambda_i^2}{\lambda_i + \underline{s}(z)} = \frac{1}{z}\cdot\frac{1}{p}\sum_{i=1}^{p}\left(\lambda_i - \underline{s}(z) + \frac{(\underline{s}(z))^2}{\lambda_i + \underline{s}(z)}\right).$$

Note that $p^{-1}\sum_{i=1}^{p}\lambda_i = p^{-1}\mathrm{tr}(\mathbf{S}) \simeq 1$, and $p^{-1}\sum_{i=1}^{p}\frac{1}{\lambda_i+\underline{s}(z)} \simeq \mathsf{m}_\beta(-\underline{s}(z))$, where $\mathsf{m}_\beta(\cdot)$ is the Stieltjes transform of a Marchenko-Pastur law with shape β. Thus, formula (24) is obtained.

Next,

$$\begin{aligned}\Upsilon_2(z) \simeq -\frac{1}{p}\sum_{i=1}^{p}\tilde{\rho}_i(z)\lambda_i^2 &= \frac{1+\gamma(\zeta(z)+z\zeta'(z))}{z^2}\cdot\frac{1}{p}\sum_{i=1}^{p}\frac{\lambda_i^2}{(\lambda_i+\underline{s}(z))^2}\\ &= \frac{1+\gamma(\zeta(z)+z\zeta'(z))}{z^2}\cdot\frac{1}{p}\sum_{i=1}^{p}\left(1 - \frac{2\underline{s}(z)}{\lambda_i+\underline{s}(z)} + (\underline{s}(z))^2\frac{1}{(\lambda_i+\underline{s}(z))^2}\right).\end{aligned}$$

Observing that $p^{-1}\sum_{i=1}^{p}\frac{1}{(\lambda_i+\underline{s}(z))^2} \simeq \mathsf{m}'_\beta(-\underline{s}(z))$, we deduce (25). Thus the computation is concluded.