Optimal spectral shrinkage and PCA with heteroscedastic noise

William Leeb* and Elad Romanov†

Abstract

This paper studies the related problems of prediction, covariance estimation, and principal component analysis for the spiked covariance model with heteroscedastic noise. We consider an estimator of the principal components based on whitening the noise, and we derive optimal singular value and eigenvalue shrinkers for use with these estimated principal components. Underlying these methods are new asymptotic results for the high-dimensional spiked model with heteroscedastic noise, and consistent estimators for the relevant population parameters. We extend previous analysis on out-of-sample prediction to the setting of predictors with whitening. We demonstrate certain advantages of noise whitening. Specifically, we show that in a certain asymptotic regime, optimal singular value shrinkage with whitening converges to the best linear predictor, whereas without whitening it converges to a suboptimal linear predictor. We prove that for generic signals, whitening improves estimation of the principal components, and increases a natural signal-to-noise ratio of the observations. We also show that for rank one signals, our estimated principal components achieve the asymptotic minimax rate.

1 Introduction

Singular value shrinkage and eigenvalue shrinkage are popular methods for denoising data matrices and covariance matrices. Singular value shrinkage is performed by computing a singular value decomposition of the observed matrix Y, adjusting the singular values, and reconstructing. The idea is that when $Y = X + N$, where X is a low-rank signal matrix we wish to estimate, the additive noise term N inflates the singular values of X; by shrinking them we can move the estimated matrix closer to X, even if the singular vectors remain inaccurate. Similarly, eigenvalue shrinkage for covariance estimation starts with the sample covariance of the data, and shrinks its eigenvalues. There has been significant recent activity on deriving optimal shrinkage methods [48, 25, 44, 23, 24, 21, 22], and applying them to various scientific problems [12, 2, 43, 17].

A standard setting for analyzing the performance of these methods is the *spiked covariance model* [31, 7, 46, 6, 21]. Here, the observation matrix is composed of iid columns Y_j in $\mathbb{R}^p$, $j = 1, \dots, n$ from some distribution consisting of signal vectors X_j lying on a low-dimensional subspace, plus independent noise vectors ε_j with some covariance matrix Σ_ε. The theory for prediction of $X_1, \dots, X_n$ in the spiked model with orthogonally invariant noise, i.e., when $\Sigma_\varepsilon = \nu I_p$, is very well-developed [23, 48, 25, 36]. Singular value shrinkage is known to be minimax optimal, and asymptotically optimal shrinkers have been derived for a wide variety of loss functions.

Many applications in signal processing, imaging, and related fields involve noise that is *heteroscedastic* [45, 40, 11, 12, 34, 1, 2]. This paper studies the effect of *whitening* the noise; that is, working in rescaled coordinates, in which the noise is white. We first estimate the noise covariance matrix Σ_ε. We then normalize, or *whiten*, the observations Y_j by applying $\Sigma_\varepsilon^{-1/2}$; the resulting vectors Y_j^{w} consist of a transformed signal component $X_j^{\mathrm{w}} = \Sigma_\varepsilon^{-1/2} X_j$, plus *isotropic* noise $G_j = \Sigma_\varepsilon^{-1/2} \varepsilon_j$. Singular value shrinkage is then performed on this new, whitened observation matrix, after which the inverse transformation $\Sigma_\varepsilon^{1/2}$ is applied. Similarly, we perform eigenvalue shrinkage to the sample covariance of the whitened data, and then apply the inverse transformation.

While this approach is restricted to cases when Σ_ε can be consistently estimated, when it does apply it has a number of advantages over competing methods. First, in the classical "large n" asymptotic limit, our method of singular value prediction with whitening, while non-linear in the observed data, converges to the best linear predictor of the data, an oracle method that requires knowledge of the

*School of Mathematics, University of Minnesota, Twin Cities, Minneapolis, MN, USA.
†School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel.

population principal components. By contrast, singular value shrinkage without whitening (as in [44]) converges to a suboptimal linear filter. Further, we show that under certain modelling assumptions, whitening improves the estimation of the population singular vectors, and achieves the same rate of subspace estimation as the minimax optimal method derived in [58]. Next, because we compute the SVD of a matrix with isotropic noise, our method requires weaker assumptions on the principal components of the signal vectors than those in [44].

As the key step in our procedures is performing spectral shrinkage to the whitened data or covariance matrices, the question arises: what are the optimal singular values/eigenvalues? While whitening has been used with shrinkage in previous works (e.g. in [38, 19, 12]) it appears that the question of optimal shrinkage has not been fully addressed. This paper derives the precise choice of optimal singular values and eigenvalues, and shows, using new asymptotic results, how to consistently estimate them from the observed data.

1.1 Overview of results

1.1.1 Spectral shrinkage with noise whitening

We introduce a new method for predicting X from Y when the noise matrix N is heteroscedastic. We first perform a linear transformation to the observations to *whiten* the noise. The resulting vectors are still of the form "low rank plus noise", but the noise term has been transformed into an isotropic Gaussian, while the low-rank signal component has been rescaled along the principal components of the noise covariance.

Next, we shrink the singular values of the transformed matrix. Intuitively, this step removes the effect of the noise from the spectrum of the observed matrix. Finally, we arrive at a predictor of the signal matrix X by applying the inverse change of variables, i.e., we *unwhiten*.

This three-step procedure — whiten, shrink, unwhiten — depends on the choice of singular values used in the middle shrinkage step. As it turns out, there are precise, optimal, and consistently estimable formulas for the optimal singular values. These are derived in Section 4.1, and the resulting method summarized in Algorithm 1.

For covariance estimation, we introduce an analogous procedure in which eigenvalue shrinkage is applied to the sample covariance of the whitened observations. After shrinkage, we then apply the inverse whitening transformation. As with singular value shrinkage, this three-step procedure of whitening, shrinking the eigenvalues, and unwhitening depends crucially on the choice of eigenvalues for the middle step. In Section 4.2, we will explain the method in detail, including the derivation of consistent estimators for the optimal eigenvalues for a variety of loss functions. The method is summarized in Algorithm 2.

1.1.2 Singular value shrinkage and linear prediction

In Section 5, we show that in the classical regime (when $p \ll n$), singular value shrinkage with whitening converges to the optimal linear predictor of the data, while shrinkage without whitening will converge to a different, typically suboptimal, linear filter. In this sense, not only is shrinkage with whitening preferable to no whitening, but the whitening transform is an asymptotically optimal change of coordinates to apply to the data before shrinking in the classical setting.

In Section 6, we also derive the optimal coefficients for the out-of-sample prediction problem, described in [19]. In this problem, the PCs estimated from a set of in-sample data $Y_1, \ldots, Y_n$ are used to denoise an independently drawn out-of-sample observation. We show that the AMSE for singular value shrinkage with whitening is identical to the asymptotic expected loss achieve by out-of-sample denoising, which extends the analogous result from [19]. The out-of-sample predictor is summarized in Algorithm 3.

1.1.3 Subspace estimation and PCA

The eigenspace of the estimated covariance $\hat{\Sigma}_x$ (equivalently, the left singular subspace of $\hat{X}$) is not spanned by the singular vectors of the raw data matrix Y. Rather, they are spanned by the vectors $\hat{u}_k$ obtained by applying the inverse whitening transformation to the top r singular vectors of the whitened observation matrix.

In Section 7, we will show under a generic model for the signal PCs, the estimated PCs $\hat{u}_1, \ldots, \hat{u}_r$ improve upon estimation of the population PCs $u_1, \ldots, u_r$, as compared to the left singular vectors of Y. We will show too that when $r = 1$, $\hat{u}_1$ achieves the minimax rate of principal subspace estimation derived in [58]. That is, in a certain sense it is an optimal estimator of the signal principal subspace.

1.1.4 Spiked model asymptotics

The methods and analysis of this paper rely on precise descriptions of the asymptotic behavior of the singular values and singular vectors of the whitened matrix Y^{w}. While some of the necessary results are already found in the literature [46, 10], we have also needed to derive several new results as well, which may be found in Theorems 3.1 and 3.2 in Section 3. Whereas earlier work has characterized the angles between the singular vectors of X^{w} and Y^{w}, we have provided formulas for the cosines of the angles between the singular vectors after the inverse whitening transformation has been performed – that is, we characterize the change in angles resulting from unwhitening. These parameters are a key ingredient for deriving the optimal spectral shrinkers in Section 4.

1.2 Related work

1.2.1 Singular value shrinkage

The prediction method in this paper is a generalization of a standard method for predicting the matrix X from the observed matrix Y, known as singular value shrinkage. Briefly, it is performed by leaving fixed the singular vectors of Y, while adjusting its singular values, to mitigate the effects of noise on the spectrum. It is shown in [23] that when the noise matrix N is white Gaussian noise, or in other words $\Sigma_\varepsilon = I_p$, then singular value shrinkage is minimax optimal for predicting X from Y.

The paper [48] considers optimal singular value shrinkage for Frobenius loss and white noise. In [25], optimal singular value shrinkers are derived for isotropic noise, for a much broader family of loss functions; the special case of operator norm loss is considered in [36]. The effectiveness of these methods rests on the asymptotic spectral theory of the data matrix Y developed in [46, 10] among others.

In the paper [44], optimal singular value shrinkage (known as 'OptShrink') is derived under much more general conditions on the noise matrix N, by exploiting the general asymptotic spectral theory developed in [10] for non-isotropic noise. While OptShrink may be effectively applied when the noise is non-isotropic, it requires the signal principal components to be vectors with iid random entries (or orthonormalized versions thereof).

1.2.2 Eigenvalue shrinkage

Covariance estimation is a well-studied problem in statistics and its applications. A standard method for estimating the population covariance Σ_x is *eigenvalue shrinkage* [51, 52, 21, 22]. Analogously to singular value shrinkage for predicting X, eigenvalue shrinkage leaves fixed the eigenvectors of the sample covariance $\hat{\Sigma}_y = \sum_{j=1}^n Y_j Y_j^\top / n = YY^\top / n$, or equivalently the left singular vectors of Y, and replaces the eigenvalues by estimated values to reduce the effect of the noise.

As we will discuss in Section 2.2, it is often natural to consider different loss functions for measuring the error in covariance estimation [22]. The paper [21] derives optimal eigenvalue shrinkers for a very large collection of loss functions. Their method is restricted to white noise, i.e., where Σ_ε is a multiple of the identity matrix.

1.2.3 Heteroscedastic noise

There have been a number of recent papers on the spiked model with heteroscedastic noise. The paper [58] devises an iterative algorithm for estimating the principal subspace of X_j in this setting, and proves that their method achieves the optimal error rate. Our method uses a different estimator for the population PCs, which achieves an error that matches the optimal rate of [58] under an additional assumption (19) (which is vacuous when $r = 1$).

The papers [28, 26, 27] consider a different but related model, in which each observation Y_j has white noise but with noise strengths varying across the observations. In [27], they show that when the signal energy and noise energy are fixed, subspace estimation is optimal when the noise is white. The proof of our Theorem 7.2 builds on this result, by combining it with our analysis of the change in angles between the empirical and population PCs after whitening. The work [28] shows that an alternative choice of weighting is optimal for estimating the signal principal components. The aforementioned paper [44] designs optimal singular value shrinkers without whitening for a broad range of noise distributions, which include our noise model as a special case.

When working in the eigenbasis of the noise covariance, the whitening procedure we describe in this work is an example of what is called *weighted PCA*, in which weights are applied to individual variables before the principal components are computed [32, 30]. The inverse standard deviation of the noise is

a standard choice of weights [54, 57, 55]; in that sense, the present work can be seen as providing a theoretical analysis of this already widely-used choice.

1.2.4 Shrinkage with whitening

Previous works have proposed pairing the whitening transformation with spectral shrinkage, which we study in this work. The paper [38] proposes the use of whitening in conjunction with exponential family noise models for covariance estimation. The paper [19] proposes whitening in the context of transformed spiked models for data prediction. The papers [12, 2] use whitening and eigenvalue shrinkage for covariance estimation.

However, previous works on singular value shrinkage with whitening employed suboptimal shrinkers, developed from heuristic considerations. In this paper, we undertake a systematic study of this problem, and rigorously derive the optimal shrinkers, under Frobenius loss (in an asymptotic sense). For covariance estimation, [38] derives the optimal eigenvalue shrinker for the special case of operator norm loss, but their method does not apply to more general loss functions.

1.3 Outline of the paper

The rest of the paper is organized as follows. Section 2 contains a detailed description of the model and assumptions; statements of the prediction and estimation problems to be studied; and a review of known results on the spiked model and spectral shrinkage. Section 3 provides the asymptotic theory on the spiked model that will be used throughout the rest of the paper. Section 4 presents the optimal spectral shrinkers with whitening. Section 5 analyzes the behavior of weighted singular value shrinkage schemes in the classical ($p \ll n$) setting, and shows the optimality of whitening in this regime. Section 6 describes and solves the out-of-sample prediction problem. Section 7 derives several results on the theoretical benefits of whitening for principal component analysis. Section 8 presents the results of numerical experiments illuminating the theoretical analysis and demonstrating the performance of the proposed methods. Finally, Section 9 provides a conclusion and suggestions for future research.

2 Preliminaries

In this section, we will introduce the details of the spiked model with heteroscedastic noise, describe the problems we focus on in this paper, and review known results on the asymptotic spectral theory of the spiked model, singular value shrinkage, and eigenvalue shrinkage. This will also serve to introduce notation we will use throughout the text.

2.1 The observation model

We now specify the precise model we will be studying in this paper. We observe iid vectors $Y_1, \dots, Y_n$ in $\mathbb{R}^p$, of the form:

$$Y_j = X_j + \varepsilon_j. \tag{1}$$

The random *signal* vectors X_j are assumed to be mean zero and to have a rank r covariance matrix $\Sigma_x = \sum_{k=1}^{r} \ell_k u_k u_k^\top$, where the vectors u_k are taken to be orthonormal, and are called the *principal components (PCs)* of the random vectors X_j. More precisely, and to distinguish them from estimated vectors we will introduce later, we will call them the *population* PCs. The numbers ℓ_k, which are the variances of the X_j along u_k, are positive; we will specify their ordering later, in equation (16) below.

The random *noise* vectors ε_j are of the form

$$\varepsilon_j = \Sigma_\varepsilon^{1/2} G_j, \tag{2}$$

where $G_j \in \mathbb{R}^p$ is a mean-zero Gaussian noise vector with covariance I_p, and Σ_ε is a full-rank positive definite covariance matrix, assumed to be known (though see Remark 3). The noise vectors G_j are drawn independently from the X_j.

We can write

$$X_j = \sum_{k=1}^{r} \ell_k^{1/2} z_{jk} u_k \tag{3}$$

Symbol	Description	Reference
X_j	Signal	(3)
ε_j	Heteroscedastic noise	(2)
Y_j	Observed	(1)
X_j^{w}	Whitened signal	(5)
G_j	Whitened noise	(2)
Y_j^{w}	Whitened observation	(5)
z_k	Signal factor values	(3), (11)
z_k^{w}	Whitened signal factor values	(6), (11)
u_k	PC of X_j's	(3)
u_k^{w}	PC of X_j^{w}'s	(6)
$\overline{u}_k$	$W^{-1}u_k^{\mathrm{w}}/\|W^{-1}u_k^{\mathrm{w}}\|$	(9)
$\hat{u}_k^{\mathrm{w}}$	Left singular vector of Y^{w}	Preceding (8)
$\hat{u}_k$	$W^{-1}\hat{u}_k^{\mathrm{w}}/\|W^{-1}\hat{u}_k^{\mathrm{w}}\|$	(8)
$\overline{u}_k^{\mathrm{w}}$	$Wu_k/\|Wu_k\|$	(10)
v_k	Right singular vector of X	Preceding (8)
v_k^{w}	Right singular vector of X^{w}	Preceding (8)
$\hat{v}_k^{\mathrm{w}}$	Right singular vector of Y^{w}	Preceding (8)

Table 1: Vectors used in this paper.

where z_{jk} are uncorrelated (though not necessarily independent) random variables, with $\mathbb{E}z_{jk} = 0$ and $\mathrm{Var}(z_{jk}) = 1$. We remark that the assumption that X_j has mean zero is not essential; all the results of this paper will go through almost without modification if we first estimate the mean of X by the sample mean and subtract it from each observation Y_j. We also note that in the terminology of factor analysis, the z_{jk} may be called the factor values; for background on factor analysis, see, for instance, [3, 4, 47, 18].

In addition to the original observations Y_j, we will also be working with the *whitened* (or *homogenized* [38]) observations Y_j^{w}, defined by $Y_j^{\mathrm{w}} = WY_j$, where

$$W = \Sigma_\varepsilon^{-1/2} \tag{4}$$

is the *whitening matrix*. The vectors Y_j^{w} can be decomposed into a transformed signal $X_j^{\mathrm{w}} = WX_j$ plus white noise G_j. The whitened vectors X_j^{w} have rank r covariance

$$\Sigma_x^{\mathrm{w}} = W\Sigma_x W, \tag{5}$$

and lie in the r-dimensional subspace $\mathrm{span}\{Wu_1, \dots Wu_r\}$. We will let $u_1^{\mathrm{w}}, \dots, u_r^{\mathrm{w}}$ be the orthonormal PCs of X_j^{w} – that is, the leading r eigenvectors (up to sign) of Σ_x^{w} – and write

$$X_j^{\mathrm{w}} = \sum_{k=1}^{r} (\ell_k^{\mathrm{w}})^{1/2} z_{jk}^{\mathrm{w}} u_k^{\mathrm{w}}, \tag{6}$$

where again $\mathbb{E}z_{jk}^{\mathrm{w}} = 0$ and $\mathrm{Var}(z_{jk}^{\mathrm{w}}) = 1$, the ℓ_k^{w} are strictly positive, and

$$\ell_1^{\mathrm{w}} > \cdots > \ell_r^{\mathrm{w}} > 0. \tag{7}$$

In general, there is not a simple relationship between the PCs $u_1, \dots, u_r$ of X_j and the PCs $u_1^{\mathrm{w}}, \dots, u_r^{\mathrm{w}}$ of X_j^{w}, or between the eigenvalues $\ell_1, \dots, \ell_r$ and the eigenvalues $\ell_1^{\mathrm{w}}, \dots, \ell_r^{\mathrm{w}}$.

We introduce some additional notation. We will denote the normalized matrices by $Y = [Y_1, \dots Y_n]/\sqrt{n}$, $Y^{\mathrm{w}} = [Y_1^{\mathrm{w}}, \dots, Y_n^{\mathrm{w}}]/\sqrt{n}$, $X = [X_1, \dots, X_n]/\sqrt{n}$, $X^{\mathrm{w}} = [X_1^{\mathrm{w}}, \dots, X_n^{\mathrm{w}}]/\sqrt{n}$, $G = [G_1, \dots, G_n]/\sqrt{n}$ and $N = [\varepsilon_1, \dots, \varepsilon_n]/\sqrt{n}$. Note that $Y = X + N$ and $Y^{\mathrm{w}} = X^{\mathrm{w}} + G$.

We will denote by $v_1, \dots, v_r$ the right singular vectors of the matrix X, and denote by $v_1^{\mathrm{w}}, \dots, v_r^{\mathrm{w}}$ the right singular vectors of the matrix X^{w}. We denote by $\hat{u}_1^{\mathrm{w}}, \dots, \hat{u}_r^{\mathrm{w}}$ and $\hat{v}_1^{\mathrm{w}}, \dots \hat{v}_r^{\mathrm{w}}$ the top r left and right singular vectors of the matrix Y^{w}. We define, for $1 \le k \le r$, the empirical vectors:

$$\hat{u}_k = \frac{W^{-1}\hat{u}_k^{\mathrm{w}}}{\|W^{-1}\hat{u}_k^{\mathrm{w}}\|}. \tag{8}$$

We also define the population counterparts,

$$\overline{u}_k = \frac{W^{-1}u_k^{\mathrm{w}}}{\|W^{-1}u_k^{\mathrm{w}}\|}. \tag{9}$$

Similarly, for $1 \le k \le r$ we define

$$\overline{u}_k^{\mathrm{w}} = \frac{Wu_k}{\|Wu_k\|}. \tag{10}$$

Note that $\mathrm{span}\{\overline{u}_1, \ldots, \overline{u}_r\} = \mathrm{span}\{u_1, \ldots, u_r\}$, and $\mathrm{span}\{\overline{u}_1^{\mathrm{w}}, \ldots, \overline{u}_r^{\mathrm{w}}\} = \mathrm{span}\{u_1^{\mathrm{w}}, \ldots, u_r^{\mathrm{w}}\}$. However, the vectors $\overline{u}_1, \ldots, \overline{u}_r$ will not, in general, be pairwise orthogonal; and similarly for $\overline{u}_1^{\mathrm{w}}, \ldots, \overline{u}_r^{\mathrm{w}}$.

Finally, we define the factor vectors z_k and z_k^{w} by

$$z_k = (z_{1k}, \ldots, z_{nk})^\top, \quad z_k^{\mathrm{w}} = (z_{1k}^{\mathrm{w}}, \ldots, z_{nk}^{\mathrm{w}})^\top. \tag{11}$$

We formally consider a *sequence* of problems, where n and $p = p_n$ both tend to ∞ with a limiting aspect ratio, γ:

$$\gamma = \lim_{n\to\infty} \frac{p_n}{n}, \tag{12}$$

which is assumed to be finite and positive. The number of population components r and the variances $\ell_1, \ldots, \ell_r$ are assumed to be fixed with n. Because p and n are increasing, all quantities that depend on p and n are elements of a sequence, which will be assumed to follow some conditions which we will outline below and summarized in Section 2.1.1. Though we might denote, for instance, the PC u_k by $u_k^{(p)}$, X by $X^{(p,n)}$, and so forth, to keep the notation to a minimum – and in keeping with standard practice with the literature on the spiked model – we will typically drop the explicit dependence on p and n.

Remark 1. Because r is fixed as p and n grow, the left singular vectors of the p-by-n population matrix $X = [X_1, \ldots, X_n]/\sqrt{n}$ are asymptotically consistent estimators (up to sign) of the population PCs $u_1, \ldots, u_r$. More precisely, if $\tilde{u}_1, \ldots, \tilde{u}_r$ are the left singular vectors of X, then almost surely

$$\lim_{p\to\infty} |\langle u_k, \tilde{u}_k\rangle| = 1. \tag{13}$$

Similarly, if $\tilde{u}_1^{\mathrm{w}}, \ldots, \tilde{u}_r^{\mathrm{w}}$ are the left singular vectors of X^{w}, then almost surely

$$\lim_{p\to\infty} |\langle u_k^{\mathrm{w}}, \tilde{u}_k^{\mathrm{w}}\rangle| = 1. \tag{14}$$

The limits (13) and (14) may be easily derived from, for example, Corollary 5.50 in [53] (restated as Lemma B.2 in Appendix B), since the effective dimension of the X_j is r, not p. Because this paper is concerned only with first-order phenomena, we will not distinguish between u_k (respectively, u_k^{w}) and $\tilde{u}_k$ (respectively, $\tilde{u}_k^{\mathrm{w}}$).

Remark 2. The unnormalized vectors $W^{-1}u_k^{\mathrm{w}}$ are the *generalized singular vectors* of the matrix X, with respect to the weight matrix W^2 [39]. In particular, they are orthonormal with respect to the weighted inner product defined by W^2. Similarly, the vectors $W^{-1}\hat{u}_k^{\mathrm{w}}$ are generalized singular vectors of Y with respect to W^2.

We assume that the values $\|W^{-1}u_k^{\mathrm{w}}\|$, $1 \le k \le r$, have well-defined limits as $p \to \infty$, and we define the parameters τ_k, $1 \le k \le r$, by

$$\tau_k = \lim_{p\to\infty} \|W^{-1}u_k^{\mathrm{w}}\|^{-2}. \tag{15}$$

Note that the τ_k are *not* known a priori; we will show, however, how they may be consistently estimated from the observed data.

With the τ_k's defined, we now specify the ordering of the principal components of X_j that will be used throughout:

$$\ell_1\tau_1 > \cdots > \ell_r\tau_r > 0. \tag{16}$$

We will also assume that the spectrum of Σ_ε stays bounded between $a_{\min} > 0$ and $a_{\max} < \infty$. In order to have well-defined asymptotics in the large p, large n regime, we will assume that the normalized trace of Σ_ε has a well-defined limit, which we will denote by μ_ε:

$$\mu_\varepsilon = \lim_{p\to\infty} \frac{\mathrm{tr}(\Sigma_\varepsilon)}{p} \in (0, \infty). \tag{17}$$

For the convenience of the reader, Tables 1 and 2 summarize the notation for vectors and scalar parameters that will be used throughout this paper.

Symbol	Description	Reference
ℓ_k	Signal variances	(3), (16)
ℓ_k^{w}	Whitened signal variances	(6), (7)
γ	Aspect ratio	(12)
τ_k	$\lim_{p\to\infty} \|W^{-1}u_k^{\mathrm{w}}\|^{-2}$	(15)
$\bar{\ell}_k$	$\ell_k^{\mathrm{w}}/\tau_k$	(62)
μ_ε	Normalized trace of Σ_ε	(17)
σ_k^{w}	Singular value of Y^{w}	(41)
c_k^{w}	Cosine between u_k^{w} and $\hat{u}_k^{\mathrm{w}}$	(39)
$\tilde{c}_k^{\mathrm{w}}$	Cosine between v_k^{w} and $\hat{v}_k^{\mathrm{w}}$	(40)
c_k	Cosine between u_k and $\hat{u}_k$ under (19)	(47)

Table 2: Scalar parameters used in this paper.

Remark 3. We will assume for most of the paper that the noise covariance Σ_ε is known a priori (though see Section 4.3). However, all of the theoretical results, and resulting algorithms, go through unchanged if the true Σ_ε is replaced by any estimator $\hat{\Sigma}_\varepsilon$ that is consistent in operator norm, i.e.,

$$\lim_{p\to\infty} \|\Sigma_\varepsilon - \hat{\Sigma}_\varepsilon\|_{\mathrm{op}} = 0. \tag{18}$$

Examples of such estimators $\hat{\Sigma}_\varepsilon$ are discussed in Section 4.3.

2.1.1 The asymptotic assumptions

We enumerate the assumptions we have made on the asymptotic model:

1. $p, n \to \infty$ and the aspect ratio p/n converges to $\gamma > 0$.
2. The eigenvalues of Σ_ε lie between $a_{\min} > 0$ and $a_{\max} < \infty$.
3. The limit $\lim_{p\to\infty} \mathrm{tr}(\Sigma_\varepsilon)/p$ is well-defined, finite, and non-zero.
4. The limits $\lim_{p\to\infty} \|W^{-1}u_k^{\mathrm{w}}\|$ are well-defined, finite, and non-zero.

Assumptions 1–4 will be in effect throughout the entire paper. In addition, some of the results, namely Theorems 3.2 and 7.3, will require an additional assumption, which we refer to as *weighted orthogonality* of the PCs $u_1, \ldots, u_r$:

5. For $j \neq k$, the vectors u_j and u_k are asymptotically orthogonal with respect to the $W^2 = \Sigma_\varepsilon^{-1}$ inner product:

$$\lim_{p\to\infty} u_j^\top W^2 u_k = 0. \tag{19}$$

The assumptions 1–4 listed above are conceptually very benign. In applications, the practitioner will be faced with a finite p and n, for which all the listed quantities exist and are finite. The asymptotic assumptions 1–4 allow us to precisely quantify the behavior when p and n are large. By contrast, assumption 5 is stronger than assumptions 1–4, in that it posits not only that certain limits exist, but also their precise values (namely, 0). Note that assumption 5 is trivially satisfied when $r = 1$.

2.1.2 Weighted orthogonality and random PCs

At first glance, the weighted orthogonality condition (5), which will be used in Theorems 3.2 and 7.3, may seem quite strong. However, it is a considerably weaker assumption than what is often assumed by methods on the spiked model. For instance, the method of OptShrink in [44] assumes that the PCs $u_1, \ldots, u_r$ be themselves random vectors with iid entries (or orthonormalized versions thereof). Under this model, the inner products $u_j^\top W^2 u_k$ almost surely converge to 0; see Proposition 6.2 in [9].

In fact, we may introduce a more general random model for random PCs, under which assumption 5 will hold. For each $1 \leq k \leq r$, we assume there is a p-by-p symmetric matrix B_k with bounded operator norm ($\|B_k\|_{\mathrm{op}} \leq C < \infty$, where C does not depend on p), and $\mathrm{tr}(B_k)/p = 1$. We then take $u_1, \ldots, u_r$ to be the output of Gram-Schmidt performed on the vectors $B_k w_k$, where the w_k are vectors with iid subgaussian entries with variance $1/p$. Then $u_j^\top W^2 u_k = w_j^\top B_j^\top W^2 B_k w_k$, which converges to zero almost surely, again using [9] and the bounded operator norm of $B_j W^2 B_k$.

Remark 4. Under the random model just described the parameters τ_k are well-defined and equal to $\lim_{p\to\infty} \mathrm{tr}(B_k^\top W^2 B_k)/p$, so long as this limit exists. Indeed, it follows from (19) that u_k^{w} is asymptotically identical to $Wu_k/\|Wu_k\|$ (see Theorem 3.2), and so $\lim_{p\to\infty} \|W^{-1}u_k^{\mathrm{w}}\|^{-2} = \lim_{p\to\infty} \|Wu_k\|^2 = \lim_{p\to\infty} \mathrm{tr}(B_k^\top W^2 B_k)/p$, where we have once again invoked [9].

2.2 The prediction and estimation problems

This paper considers three central tasks: denoising the observations Y_j to recover X_j – what we refer to as *prediction*, since the X_j's are themselves random – estimating the population covariance Σ_x, and estimating the principal subspace $\mathrm{span}\{u_1, \dots, u_r\}$.

For predicting the signal vectors X_j, or equivalently the normalized signal matrix $X = [X_1, \dots, X_n]/\sqrt{n}$, we will use the asymptotic mean squared error to measure the accuracy of a predictor $\hat{X}$:

$$\mathrm{AMSE} = \lim_{n\to\infty} \mathbb{E}\|\hat{X} - X\|_{\mathrm{F}}^2 = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\|\hat{X}_j - X_j\|^2. \tag{20}$$

For covariance estimation, our goal is to estimate the covariance of the signal vectors, $\Sigma_x = \mathbb{E}[X_j X_j^\top]$ (under the convention that the X_j are mean zero; otherwise, we subtract off the mean). While the Frobenius loss, or MSE, is natural for signal estimation, for covariance estimation it is useful to consider a wider range of loss functions depending on the statistical problem at hand; see [22] and the references within for an elucidation of this point.

We will denote our covariance estimator as $\hat{\Sigma}_x$. Denote the loss function by $\mathcal{L}(\hat{\Sigma}_x, \Sigma_x)$; for instance, Frobenius loss $\mathcal{L}(\hat{\Sigma}_x, \Sigma_x) = \|\hat{\Sigma}_x - \Sigma_x\|_{\mathrm{F}}^2$, or operator norm loss $\mathcal{L}(\hat{\Sigma}_x, \Sigma_x) = \|\hat{\Sigma}_x - \Sigma_x\|_{\mathrm{op}}$. For a specified loss function $\mathcal{L}$, we seek to minimize the asymptotic values of these loss functions for our estimator,

$$\lim_{n\to\infty} \mathbb{E}\mathcal{L}(\hat{\Sigma}_x, \Sigma_x). \tag{21}$$

For both the data prediction and covariance estimation problems, it will be a consequence of our analysis that the limits of the errors are, in fact, well-defined quantities.

Finally, we are also concerned with principal component analysis (PCA), or estimating the principal subspace $\mathcal{U} = \mathrm{span}\{u_1, \dots, u_r\}$, in which the signal vectors X_j lie. We measure the discrepancy between the estimated subspace $\hat{\mathcal{U}}$ and the true subspace $\mathcal{U}$ by the angle $\Theta(\mathcal{U}, \hat{\mathcal{U}})$ between these subspaces, defined by

$$\sin\Theta(\mathcal{U}, \hat{\mathcal{U}}) = \|\hat{U}_\perp^\top U\|_{\mathrm{op}}, \tag{22}$$

where $\hat{U}_\perp$ and U are matrices whose columns are orthonormal bases of $\hat{\mathcal{U}}^\perp$ and $\mathcal{U}$, respectively.

2.3 Review of the spiked model

2.3.1 Asymptotic spectral theory of the spiked model

The spectral theory of the observed matrix Y has been thoroughly studied in the large p, large n regime, when $p = p_n$ grows with n. We will offer a brief survey of the relevant results from the literature [46, 10, 19].

In the case of isotropic Gaussian noise (that is, when $\Sigma_\varepsilon = I_p$), the r largest singular values of the matrix Y converge to σ_k, defined by:

$$\sigma_k^2 = \begin{cases} (\ell_k + 1)(1 + \gamma/\ell_k), & \text{if } \ell_k > \sqrt{\gamma}, \\ (1 + \sqrt{\gamma})^2, & \text{if } \ell_k \le \sqrt{\gamma} \end{cases}. \tag{23}$$

Furthermore, the top singular vectors $\hat{u}_k^y$ and $\hat{v}_k^y$ of Y make asymptotically deterministic angles with the singular vectors u_k and v_k of X. More precisely, the absolute cosines $|\langle \hat{u}_j^y, u_k \rangle|$ converge to $c_k = c_k(\gamma, \ell_k)$, defined by

$$c_k^2 = \begin{cases} \frac{1-\gamma/\ell^2}{1+\gamma/\ell} & \text{if } j = k \text{ and } \ell_k > \sqrt{\gamma} \\ 0 & \text{otherwise} \end{cases}, \tag{24}$$

and the absolute cosines $|\langle \hat{v}_j^y, v_k \rangle|$ converge to $\tilde{c}_k = \tilde{c}_k(\gamma, \ell_k)$, defined by

$$\tilde{c}_k^2 = \begin{cases} \frac{1-\gamma/\ell^2}{1+1/\ell} & \text{if } j = k \text{ and } \ell_k > \sqrt{\gamma} \\ 0 & \text{otherwise} \end{cases}. \tag{25}$$

When $\ell_k > \sqrt{\gamma}$, the population variance ℓ_k can be estimated consistently from the observed singular value σ_k. Since c_k and $\tilde{c}_k$ are functions of ℓ_k and the aspect ratio γ, these quantities can then also be consistently estimated.

Remark 5. Due to the orthogonal invariance of the noise matrix $N = G$ when $\Sigma_\varepsilon = I_p$, formulas (23), (24) and (25) are valid for any rank r matrix X, so long as X's singular values do not change with p and n. The paper [10] derive the asymptotics for more general noise matrices N, but with the additional assumption that the singular vectors of X are themselves random (see the discussion in Section 2.1.2). The formulas for the asymptotic singular values and cosines found in [10] are in terms of the Stieltjes transform [5] of the asymptotic distribution of singular values of Y, which can be estimated consistently using the observed singular values of Y.

2.3.2 Optimal shrinkage with Frobenius loss and white noise

We review the theory of shrinkage with respect to Frobenius loss; we briefly mention that the paper [25] extends these ideas to a much wider range of loss functions for the spiked model.

We suppose that our predictor of X is a rank r matrix of the form

$$\hat{X} = \sum_{k=1}^{r} t_k \hat{u}_k \hat{v}_k^\top, \tag{26}$$

where $\hat{u}_k$ and $\hat{v}_k$ are estimated vectors. We will assume that the vectors $\hat{v}_k$ are orthogonal, and that their cosines with the population vectors v_k of X are asymptotically deterministic. More precisely, we assume that $\langle v_j, \hat{v}_k \rangle^2 \to \tilde{c}_k^2$ when $j = k$, and converges to 0 when $j \neq k$. Similarly, we will assume that $\langle u_k, \hat{u}_k \rangle^2 \to c_k^2$; however, we do not need to assume any orthogonality condition on the u_j's and $\hat{u}_j$'s for the purposes of this derivation.

Expanding the squared Frobenius loss between $\hat{X}$ and X and using the orthogonality conditions on the v_j's and $\hat{v}_k$'s, we get:

$$\begin{aligned} \|\hat{X} - X\|_{\mathrm{F}}^2 &= \left\| \sum_{k=1}^{r} \left(t_k \hat{u}_k \hat{v}_k^\top - \ell_k^{1/2} u_k v_k^\top \right) \right\|_{\mathrm{F}}^2 \\ &= \sum_{k=1}^{r} \left\| t_k \hat{u}_k \hat{v}_k^\top - \ell_k^{1/2} u_k v_k^\top \right\|_{\mathrm{F}}^2 + \sum_{j \neq k} \left\langle t_j \hat{u}_j \hat{v}_j^\top - \ell_j^{1/2} u_j v_j^\top, t_k \hat{u}_k \hat{v}_k^\top - \ell_k^{1/2} u_k v_k^\top \right\rangle_{\mathrm{F}} \\ &\sim \sum_{k=1}^{r} \| t_k \hat{u}_k \hat{v}_k^\top - \ell_k^{1/2} u_k v_k^\top \|_{\mathrm{F}}^2, \end{aligned} \tag{27}$$

where $\sim$ denotes almost sure equality as $p, n \to \infty$.

Since the loss separates over the different components, we may consider each component separately. Using the asymptotic cosines, we have:

$$\| t_k \hat{u}_k \hat{v}_k^\top - \ell_k^{1/2} u_k v_k^\top \|_{\mathrm{F}}^2 \sim t_k^2 + \ell_k - 2\ell_k^{1/2} c_k \tilde{c}_k t_k, \tag{28}$$

which is minimized by taking

$$t_k = \ell_k^{1/2} c_k \tilde{c}_k. \tag{29}$$

These values of t_k, therefore, are the optimal ones for predicting X in Frobenius loss.

Furthermore, we can also derive an estimable formula for the AMSE. Indeed, plugging in $t_k = \ell_k^{1/2} c_k \tilde{c}_k$ to (28), we get:

$$\mathrm{AMSE} = \sum_{k=1}^{r} \ell_k^2 (1 - c_k^2 \tilde{c}_k^2). \tag{30}$$

Note that this derivation of the optimal t_k and the AMSE does not require the vectors $\hat{u}_k$ and $\hat{v}_k$ to be the singular vectors of Y. Rather, we just require the asymptotic cosines to be well-defined, and the v_j's and $\hat{v}_j$'s to be orthogonal across different components. Implementing this procedure, however, requires consistent estimates of ℓ_k, c_k and $\tilde{c}_k$.

2.3.3 Eigenvalue shrinkage for covariance estimation

Similar to the task of predicting the data matrix X is estimating the covariance matrix $\Sigma_x = \mathbb{E}[X_j X_j^\top] = \sum_{k=1}^r \ell_k u_k u_k^\top$. The procedure we consider in this setting is known as *eigenvalue shrinkage*. Given orthonormal vectors $\hat{u}_1, \ldots, \hat{u}_r$ estimating the PCs $u_1, \ldots, u_r$, we consider estimators of the form

$$\hat{\Sigma}_x = \sum_{k=1}^r t_k^2 \hat{u}_k \hat{u}_k^\top, \tag{31}$$

where t_k^2 are estimated population eigenvalues, which it is our goal to determine.

In [21], a large family of loss functions are considered for estimating Σ_x in white noise. All these loss functions satisfy two conditions. First, they are *orthogonally-invariant*, meaning that if both the estimated and population PCs are rotated, the loss does not change. Second, they are *block-decomposable*, meaning that if both the estimated and population covariance matrices are in block-diagonal form, the loss can be written as functions of the losses between the individual blocks.

The method of [21] rests on an observation from linear algebra. If (asymptotically) the $\langle \hat{u}_k, u_k \rangle = c_k$, and $\hat{u}_j \perp u_k$ for all $1 \le j \ne k \le r$, then there is an orthonormal basis of $\mathbb{R}^p$ with respect to which both Σ_x and any rank r covariance $\hat{\Sigma}_x$ are simultaneously block-diagonalizable, with r blocks of size 2-by-2. More precisely, there is a p-by-p orthogonal matrix O so that:

$$O \Sigma_x O^\top = \bigoplus_{k=1}^r A_k, \tag{32}$$

and

$$O \hat{\Sigma}_x O^\top = \bigoplus_{k=1}^r \hat{\ell}_k B_k, \tag{33}$$

where

$$A_k = \begin{pmatrix} \ell_k & 0 \\ 0 & 0 \end{pmatrix}, \tag{34}$$

and

$$B_k = \begin{pmatrix} c_k^2 & c_k \sqrt{1 - c_k^2} \\ c_k \sqrt{1 - c_k^2} & 1 - c_k^2 \end{pmatrix}. \tag{35}$$

If $\mathcal{L}(\hat{\Sigma}, \Sigma)$ is a loss function that is orthogonally-invariant and block-decomposable, then the loss between Σ_x and $\hat{\Sigma}_x$ decomposes into the losses between each A_k and B_k, which depend only on the one parameter $\hat{\ell}_k$. Consequently,

$$\hat{\ell}_k = \arg\min_{\ell} \mathcal{L}(A_k, \ell B_k). \tag{36}$$

The paper [21] contains solutions for $\hat{\ell}_k$ for a wide range of loss functions $\mathcal{L}$. For example, with Frobenius loss, the optimal value is $\hat{\ell}_k = \ell_k c_k^2$, whereas for operator norm loss the optimal value is $\hat{\ell}_k = \ell_k$. Even when closed form solutions are unavailable, one may perform the mimimization (36) numerically.

3 Asymptotic theory

A precise understanding of the asymptotic behavior of the spiked model is crucial for deriving optimal spectral shrinkers, as we have seen in Sections 2.3.2 and 2.3.3. In this section, we provide expressions for the asymptotic cosines between the empirical PCs and the population PCs, as well as limiting values for other parameters. The formulas from Theorem 3.1 below will be employed in Section 4.1 for optimal singular value shrinkage with whitening; and the formulas from Theorem 3.2 below will be employed in Section 4.2 for optimal eigenvalue shrinkage with whitening.

The first result, Theorem 3.1, applies to the standard spiked model with white noise. It gives a characterization of the asymptotic angles of the population PCs and empirical PCs with respect to an inner product $x^\top A y$ given by a symmetric positive-definite matrix A. Parts 1 and 4 are standard results on the spiked covariance model [46, 10]; we include them here for easy reference. A special case of part 2 appears in [38], in a somewhat different form; and part 3 appear to be new.

Theorem 3.1. *Suppose $Y_1^{\mathrm{w}}, \ldots, Y_n^{\mathrm{w}}$ are iid vectors in $\mathbb{R}^p$ from the spiked model with white noise, with $Y_j^{\mathrm{w}} = X_j^{\mathrm{w}} + G_j$ where X_j^{w} is of the form (6) and $G_j \sim N(0, I)$. Let $A = A_p$ be an element of a sequence of symmetric, positive-definite p-by-p matrices with bounded operator norm ($\|A_p\|_{\mathrm{op}} \leq C < \infty$ for all p), whose asymptotic normalized trace is well-defined and finite:*

$$\mu_a = \lim_{p\to\infty} \frac{1}{p}\mathrm{tr}(A_p) < \infty. \tag{37}$$

Suppose too that for $1 \leq k \leq r$, the following quantity τ_k^a is also well-defined and finite:

$$\tau_k^a = \lim_{p\to\infty} \|A_p^{1/2} u_k^{\mathrm{w}}\|^{-2} < \infty. \tag{38}$$

Define $c_k^{\mathrm{w}} > 0$ by:

$$(c_k^{\mathrm{w}})^2 = \begin{cases} \frac{1-\gamma/(\ell_k^{\mathrm{w}})^2}{1+\gamma/\ell_k^{\mathrm{w}}}, & \text{if } j = k \text{ and } \ell_k^{\mathrm{w}} > \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases}, \tag{39}$$

and let $s_k^{\mathrm{w}} = \sqrt{1-(c_k^{\mathrm{w}})^2}$. Also define $\tilde{c}_k^{\mathrm{w}} > 0$ by:

$$(\tilde{c}_k^{\mathrm{w}})^2 = \begin{cases} \frac{1-\gamma/(\ell_k^{\mathrm{w}})^2}{1+1/\ell_k^{\mathrm{w}}}, & \text{if } j = k \text{ and } \ell_k^{\mathrm{w}} > \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases}, \tag{40}$$

and $\tilde{s}_k^{\mathrm{w}} = \sqrt{1-(\tilde{c}_k^{\mathrm{w}})^2}$.

Then for any $1 \leq j, k \leq r$, we have, as $n \to \infty$ and $p/n \to \gamma$:

1. *The k^{th} largest singular value of Y^{w} converges almost surely to*

$$\sigma_k^{\mathrm{w}} = \begin{cases} \sqrt{(\ell_k^{\mathrm{w}}+1)\left(1+\frac{\gamma}{\ell_k^{\mathrm{w}}}\right)}, & \text{if } \ell_k^{\mathrm{w}} > \sqrt{\gamma} \\ 1+\sqrt{\gamma}, & \text{otherwise} \end{cases}. \tag{41}$$

2. *The A-norm of $\hat{u}_k^{\mathrm{w}}$ converges almost surely:*

$$\lim_{p\to\infty} \|A_p^{1/2}\hat{u}_k^{\mathrm{w}}\|^2 = \frac{(c_k^{\mathrm{w}})^2}{\tau_k^a} + (s_k^{\mathrm{w}})^2\mu_a. \tag{42}$$

3. *The A-inner product between u_k^{w} and $\hat{u}_k^{\mathrm{w}}$ converges almost surely:*

$$\lim_{p\to\infty} \langle A_p u_k^{\mathrm{w}}, \hat{u}_k^{\mathrm{w}}\rangle^2 = \begin{cases} (c_k^{\mathrm{w}}/\tau_k^a)^2, & \text{if } \ell_k^{\mathrm{w}} > \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases}. \tag{43}$$

4. *The inner product between v_j^{w} and $\hat{v}_k^{\mathrm{w}}$ converges almost surely:*

$$\lim_{n\to\infty} \langle v_j^{\mathrm{w}}, \hat{v}_k^{\mathrm{w}}\rangle^2 = \begin{cases} (\tilde{c}_k^{\mathrm{w}})^2, & \text{if } j = k \text{ and } \ell_k^{\mathrm{w}} > \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases}. \tag{44}$$

Remark 6. In fact, as will be evident from its proof Theorem 3.1 is applicable to any rank r matrix X^{w}, viewing u_k^{w} and v_k^{w} as the singular vectors of X^{w}. In particular, the columns of X^{w} need not be drawn iid from a mean zero distribution. All that is needed for Theorem 3.1 is that the singular values of X^{w} remain constant as p and n grow, and that the parameters τ_k are well-defined.

Theorem 3.1 is concerned only with the standard spiked model with white noise, $Y_j^{\mathrm{w}} = X_j^{\mathrm{w}} + G_j$. By contrast, the next result, Theorem 3.2, deals with the spiked model with colored noise, $Y_j = X_j + \varepsilon_j$, where $\varepsilon_j \sim N(0, \Sigma_\varepsilon)$. In Section 2.1, we defined the whitening matrix $W = \Sigma_\varepsilon^{-1/2}$ that transforms Y_j into the standard white-noise model Y_j^{w}; that is, $Y_j^{\mathrm{w}} = WY_j = WX_j + W\varepsilon_j = X_j^{\mathrm{w}} + G_j$. In stating and applying Theorem 3.2, we refer to the parameters for both models described in Section 2.1.

Theorem 3.2. *Assume that the PCs $u_1, \ldots, u_r$ satisfy the weighted orthogonality condition (19), i.e., for $1 \leq j \neq k \leq r$,*

$$\lim_{p\to\infty} u_j^\top W^2 u_k = 0. \tag{45}$$

Order the principal components of X_j by decreasing value of $\ell_k\tau_k$, as in (16); that is, we assume $\Sigma_x = \sum_{k=1}^r \ell_k u_k u_k^\top$, with

$$\ell_1\tau_1 > \cdots > \ell_r\tau_r > 0, \tag{46}$$

where $\tau_k = \lim_{p\to\infty} \|W^{-1}u_k^{\mathrm{w}}\|^{-2}$ as in (15).

Define $c_k > 0$, $1 \le k \le r$, by:

$$c_k^2 \equiv \begin{cases} \frac{(c_k^{\mathrm{w}})^2}{(c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \cdot \mu_\varepsilon \cdot \tau_k}, & \text{if } \ell_k^{\mathrm{w}} > \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases}, \tag{47}$$

where c_k^{w} is given by (39), ℓ_k^{w} is defined from (6) with $X_j^{\mathrm{w}} = WX_j$, and $\mu_\varepsilon = \lim_{p\to\infty} \frac{\mathrm{tr}(\Sigma_\varepsilon)}{p}$ as in (17).

Then for any $1 \le j, k \le r$,

1. *The vectors $\overline{u}_k$ and u_k are almost surely asymptotically identical:*

$$\lim_{p\to\infty} \langle u_k, \overline{u}_k \rangle^2 = 1. \tag{48}$$

2. *The vectors v_k^{w} and v_k are almost surely asymptotically identical:*

$$\lim_{n\to\infty} \langle v_k, v_k^{\mathrm{w}} \rangle^2 = 1. \tag{49}$$

3. *The inner product between u_j and $\hat{u}_k$ converges almost surely:*

$$\lim_{p\to\infty} \langle u_j, \hat{u}_k \rangle^2 = \begin{cases} c_k^2, & \text{if } j = k \text{ and } \ell_k^{\mathrm{w}} > \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases}, \tag{50}$$

   where c_k^2 is defined in (47).

4. *The vectors $\hat{u}_j$ and $\hat{u}_k$ are asymptotically orthogonal if $j \neq k$:*

$$\lim_{p\to\infty} \langle \hat{u}_j, \hat{u}_k \rangle^2 = \delta_{jk}. \tag{51}$$

5. *The parameter τ_k is almost surely asymptotically equal to $\|Wu_k\|^2$:*

$$\lim_{p\to\infty} (\tau_k - \|Wu_k\|^2) = 0. \tag{52}$$

6. *The variance ℓ_k^{w} of X_j^{w} along u_k^{w} is almost surely asymptotically equal to $\ell_k\tau_k$:*

$$\lim_{p\to\infty} (\ell_k^{\mathrm{w}} - \ell_k\tau_k) = 0. \tag{53}$$

The proofs for both Theorem 3.1 and Theorem 3.2 may be found in Appendix A.

4 Optimal spectral shrinkage with whitening

In this section, we will derive the optimal spectral shrinkers for signal prediction and covariance estimation to be used in conjunction with whitening.

4.1 Singular value shrinkage

Given the noisy matrix $Y = X + N$, we consider a class of predictors of X defined as follows. First, we whiten the noise, replacing Y with $Y^{\mathrm{w}} = WY$. We then apply singular value shrinkage to the transformed matrix Y^{w}. That is, if $\hat{u}_1^{\mathrm{w}}, \ldots, \hat{u}_r^{\mathrm{w}}$ and $\hat{v}_1^{\mathrm{w}}, \ldots, \hat{v}_r^{\mathrm{w}}$ are the top left and right singular vectors of Y^{w}, we define the new matrix

$$\hat{X}^{\mathrm{w}} = \sum_{k=1}^r t_k \hat{u}_k^{\mathrm{w}} (\hat{v}_k^{\mathrm{w}})^\top, \tag{54}$$

for some positive scalars t_k which we have yet to determine.

Finally, we recolor the noise, to bring the data back to its original scaling. That is, we define our final predictor $\hat{X}$ by

$$\hat{X} = W^{-1}\hat{X}^{\mathrm{w}}. \tag{55}$$

In this section, we will show how to optimally choose the singular values $t_1, \ldots, t_r$ in (54) to minimize the AMSE:

$$\mathrm{AMSE} = \lim_{n\to\infty} \mathbb{E}\|\hat{X} - X\|_{\mathrm{F}}^2. \tag{56}$$

Remark 7. Loss functions other than Frobenius loss (i.e., mean-squared error) may be considered as well. This will be done for the problem of covariance estimation in Section 4.2, where it is more natural [22]. For recovering the data matrix X itself, however, the MSE is the natural loss, and the optimal t_k can be derived for minimizing the AMSE without any additional assumptions on the model.

Once we have whitened the noise, our resulting matrix $Y^{\mathrm{w}} = X^{\mathrm{w}} + G$ is from the standard spiked model and consequently satisfies the conditions of Theorem 3.1, since G is a Gaussian matrix with iid $N(0,1)$ entries. We will apply the asymptotic results of Theorem 3.1, taking the matrix $A = W^{-1}$. Recalling the definitions of $\hat{u}_k$ and $\overline{u}_k$ from (8) and (9), respectively, we obtain an immediate corollary to Theorem 3.1:

Corollary 4.1. *For $1 \le k \le r$, the cosine between the vectors $\overline{u}_k$ and $\hat{u}_k$ converges almost surely:*

$$\lim_{p\to\infty} \langle \overline{u}_k, \hat{u}_k \rangle^2 = c_k^2 \equiv \begin{cases} \frac{(c_k^{\mathrm{w}})^2}{(c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \cdot \mu_\varepsilon \cdot \tau_k}, & \text{if } \ell_k^{\mathrm{w}} > \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases}. \tag{57}$$

We derive the optimal t_k. We write:

$$X^{\mathrm{w}} \sim \sum_{k=1}^{r} (\ell_k^{\mathrm{w}})^{1/2} u_k^{\mathrm{w}} (v_k^{\mathrm{w}})^\top, \tag{58}$$

and so

$$X = W^{-1}X^{\mathrm{w}} \sim \sum_{k=1}^{r} (\ell_k^{\mathrm{w}})^{1/2} W^{-1} u_k^{\mathrm{w}} (v_k^{\mathrm{w}})^\top = \sum_{k=1}^{r} (\ell_k^{\mathrm{w}}/\tau_k)^{1/2} \overline{u}_k (v_k^{\mathrm{w}})^\top. \tag{59}$$

Furthermore,

$$\hat{X}^{\mathrm{w}} = \sum_{k=1}^{r} t_k \hat{u}_k^{\mathrm{w}} (\hat{v}_k^{\mathrm{w}})^\top \tag{60}$$

and so

$$\hat{X} = W^{-1}\hat{X}^{\mathrm{w}} = \sum_{k=1}^{r} t_k W^{-1} \hat{u}_k^{\mathrm{w}} (\hat{v}_k^{\mathrm{w}})^\top = \sum_{k=1}^{r} t_k \|W^{-1}\hat{u}_k^{\mathrm{w}}\| \hat{u}_k (\hat{v}_k^{\mathrm{w}})^\top. \tag{61}$$

It is convenient to reparametrize the problem in terms of

$$\overline{\ell}_k \equiv \ell_k^{\mathrm{w}}/\tau_k, \tag{62}$$

and

$$\tilde{t}_k \equiv t_k \|W^{-1}\hat{u}_k^{\mathrm{w}}\| \sim t_k \left(\frac{(c_k^{\mathrm{w}})^2}{\tau_k} + (s_k^{\mathrm{w}})^2 \mu_\varepsilon \right)^{1/2}, \tag{63}$$

where we have used Theorem 3.1.

In this notation, we have $X = \sum_{k=1}^{r} \overline{\ell}_k^{1/2} \overline{u}_k (v_k^{\mathrm{w}})^\top$, and $\hat{X} = \sum_{k=1}^{r} \tilde{t}_k \hat{u}_k (\hat{v}_k^{\mathrm{w}})^\top$. From Theorem 3.1, the vectors v_j^{w} and $\hat{v}_k^{\mathrm{w}}$ are orthogonal if $j \neq k$, and the cosine between v_k^{w} and $\hat{v}_k^{\mathrm{w}}$ is $\tilde{c}_k \equiv \tilde{c}_k^{\mathrm{w}}$. The derivation from Section 2.3.2 shows that the optimal values $\tilde{t}_k$ are then given by

$$\tilde{t}_k = \overline{\ell}_k^{1/2} c_k \tilde{c}_k \tag{64}$$

For this to define a valid estimator, we must show how to estimate the values $\overline{\ell}_k$, c_k and $\tilde{c}_k$ from the observed data itself.

To that end, from Theorem 3.1 ℓ_k^{w} can be estimated by

$$\ell_k^{\mathrm{w}} = \frac{(\sigma_k^{\mathrm{w}})^2 - 1 - \gamma + \sqrt{((\sigma_k^{\mathrm{w}})^2 - 1 - \gamma)^2 - 4\gamma}}{2} \tag{65}$$

where σ_k^{w} is the k^{th} singular value of Y^{w}. The cosines c_k^{w} and $\tilde{c}_k^{\mathrm{w}}$ can then be estimated by formulas (39) and (40).

Now, rearranging part 2 from Theorem 3.1, we can solve for τ_k in terms of the estimable quantities c_k^{w}, s_k^{w}, μ_ε and $\|\Sigma_\varepsilon^{1/2}\hat{u}_k^{\mathrm{w}}\|^2$:

$$\tau_k \sim \frac{(c_k^{\mathrm{w}})^2}{\|\Sigma_\varepsilon^{1/2}\hat{u}_k^{\mathrm{w}}\|^2 - (s_k^{\mathrm{w}})^2\mu_\varepsilon}. \tag{66}$$

Indeed, this quantity can be estimated consistently: c_k^{w} and s_k^{w} are estimable from (39), $\|\Sigma_\varepsilon^{1/2}\hat{u}_k^{\mathrm{w}}\|^2$ is directly observed, and $\mu_\varepsilon \sim \mathrm{tr}(\Sigma_\varepsilon)/p$.

Having estimated τ_k, we apply formula $\bar{\ell}_k = \ell_k^{\mathrm{w}}/\tau_k$, and formula (50) for c_k. This completes the derivation of the optimal singular value shrinker. The entire procedure is described in Algorithm 1.

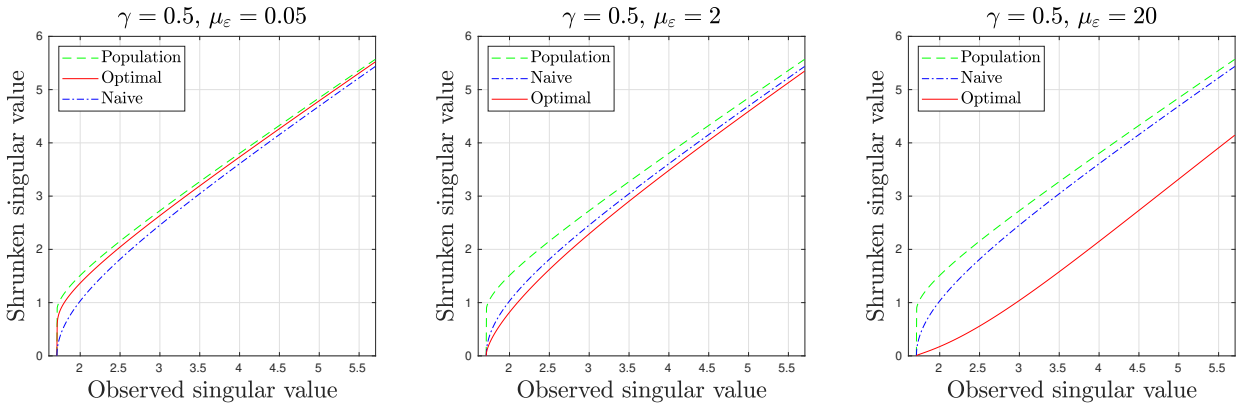

Figure 1: Optimal shrinker, naive shrinker, and population shrinker, for $\tau = 1$ and $\gamma = 0.5$.

Figures 1 and 2 plot the optimal shrinker, i.e., the function that sends each top observed singular value σ_k^{w} of Y^{w} to the optimal t_k. For contrast, we also plot the "population" shrinker, which maps σ_k^{w} to the corresponding $\sqrt{\ell_k^{\mathrm{w}}}$; and the "naive" shrinker, which maps σ_k^{w} to $\sqrt{\ell_k^{\mathrm{w}}}c_k^{\mathrm{w}}\tilde{c}_k^{\mathrm{w}}$. This latter shrinker is considered in the paper [19], and is naive in that it optimizes the Frobenius loss before the unwhitening step without accounting for the change in angles between singular vectors resulting from unwhitening. In Figure 1 we set $\gamma = 0.5$, while in Figure 2 we set $\gamma = 2$. We fix $\tau = 1$ but consider different values of μ_ε (the behavior depends only on the ratio of μ_ε and τ).

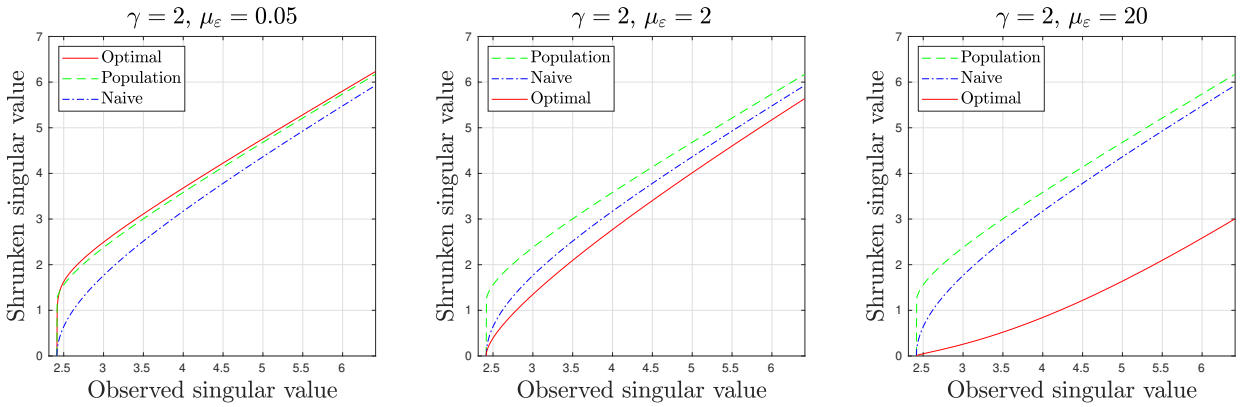

Figure 2: Optimal shrinker, naive shrinker, and population shrinker, for $\tau = 1$ and $\gamma = 2$.

Algorithm 1 Optimal singular value shrinkage with whitening

Input: observations $Y_1, \ldots, Y_n$; noise covariance Σ_ε; rank r
Define $Y = [Y_1, \ldots, Y_n]/\sqrt{n}$; $W = \Sigma_\varepsilon^{-1/2}$;$Y^{\mathrm{w}} = WY$
Compute rank r SVD of Y^{w}: $\hat{u}_1^{\mathrm{w}}, \ldots, \hat{u}_r^{\mathrm{w}}$; $\hat{v}_1^{\mathrm{w}}, \ldots, \hat{v}_r^{\mathrm{w}}$; $\sigma_1^{\mathrm{w}}, \ldots, \sigma_r^{\mathrm{w}}$
for all $k = 1, \ldots, r$ **do**
    if $\sigma_k^{\mathrm{w}} > 1 + \sqrt{\gamma}$ **then**
        $\ell_k^{\mathrm{w}} = \left[(\sigma_k^{\mathrm{w}})^2 - 1 - \gamma + \sqrt{((\sigma_k^{\mathrm{w}})^2 - 1 - \gamma)^2 - 4\gamma}\right]/2$
        $c_k^{\mathrm{w}} = \sqrt{(1 - \gamma/(\ell_k^{\mathrm{w}})^2) \,/\, (1 + \gamma/\ell_k^{\mathrm{w}})}$
        $s_k^{\mathrm{w}} = \sqrt{1 - (c_k^{\mathrm{w}})^2}$
        $\tilde{c}_k = \sqrt{(1 - \gamma/(\ell_k^{\mathrm{w}})^2) \,/\, (1 + 1/\ell_k^{\mathrm{w}})}$
        $\mu_\varepsilon = \mathrm{tr}(\Sigma_\varepsilon)/p$
        $\tau_k = (c_k^{\mathrm{w}})^2 \Big/ \left[\|\Sigma_\varepsilon^{1/2} \hat{u}_k^{\mathrm{w}}\|^2 - (s_k^{\mathrm{w}})^2 \mu_\varepsilon\right]$
        $t_k = (\ell_k^{\mathrm{w}})^{1/2} c_k^{\mathrm{w}} \tilde{c}_k \Big/ \left[(c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \mu_\varepsilon \tau_k\right]$
    else if $\sigma_k^{\mathrm{w}} \leq 1 + \sqrt{\gamma}$ **then**
        $t_k = 0$
Output: $\hat{X} = W^{-1} \sum_{k=1}^r t_k \hat{u}_k^{\mathrm{w}} (\hat{v}_k^{\mathrm{w}})^\top$

Remark 8. In practice, the rank r may not be known a priori. In Section 4.4, we describe several methods for estimating r from the data.

Remark 9. Algorithm 1 may be applied to denoising any rank r matrix X from the observed matrix $Y = X + N$. As pointed out in Remark 6, the assumption that the columns of X are drawn iid from a mean zero distribution with covariance Σ_x is not needed for the parameter estimates used by Algorithm 1 to be applicable, so long as the singular values of the whitened matrix X^{w} stay fixed (or converge almost surely) as p and n grow, and the parameters τ_k are well-defined.

4.2 Eigenvalue shrinkage

We turn now to the task of estimating the covariance Σ_x of X_j. Throughout this section, we will assume the conditions of Theorem 3.2, namely conditon (19).

Analogous to the procedure for singular value shrinkage with whitening, we consider the procedure of eigenvalue shrinkage with whitening. We first whiten the observations Y_j, producing new observations $Y_j^{\mathrm{w}} = WY_j$. We then form the sample covariance $\hat{\Sigma}_y^{\mathrm{w}}$ of the Y_j^{w}. We apply eigenvalue shrinkage to $\hat{\Sigma}_y^{\mathrm{w}}$, forming a matrix of the form

$$\hat{\Sigma}_x^{\mathrm{w}} = \sum_{k=1}^r t_k^2 \hat{u}_k^{\mathrm{w}} (\hat{u}_k^{\mathrm{w}})^\top, \tag{67}$$

where $\hat{u}_1, \ldots, \hat{u}_r^{\mathrm{w}}$ are the top r eigenvectors of $\hat{\Sigma}_y^{\mathrm{w}}$, or equivalently the top r left singular vectors of the whitened data matrix Y^{w}; and the t_k^2 are the parameters we will determine. Finally, we form our final estimator of Σ_x by unwhitening:

$$\hat{\Sigma}_x = W^{-1} \hat{\Sigma}_x^{\mathrm{w}} W^{-1}. \tag{68}$$

It remains to define the eigenvalues $t_1^2, \ldots, t_r^2$ of the matrix $\hat{\Sigma}_x^{\mathrm{w}}$. We let $\mathcal{L}$ denote any of the loss functions considered in [21]. As a reminder, all these loss functions satisfy two conditions. First, they are *orthogonally-invariant*, meaning that if both the estimated and population PCs are rotated, the loss does not change. Second, they are *block-decomposable*, meaning that if both the estimated and population covariance matrices are in block-diagonal form, the loss can be written as functions of the losses between the individual blocks.

The estimated covariance matrix $\hat{\Sigma}_x = W^{-1} \hat{\Sigma}_x^{\mathrm{w}} W^{-1}$ can be written as:

$$\hat{\Sigma}_x = W^{-1} \hat{\Sigma}_x^{\mathrm{w}} W^{-1} = \sum_{k=1}^r t_k^2 W^{-1} \hat{u}_k^{\mathrm{w}} (W^{-1} \hat{u}_k^{\mathrm{w}})^\top = \sum_{k=1}^r t_k^2 \|W^{-1} \hat{u}_k^{\mathrm{w}}\|^2 \hat{u}_k \hat{u}_k^\top = \sum_{k=1}^r \tilde{t}_k^2 \hat{u}_k \hat{u}_k^\top, \tag{69}$$

where we have defined $\tilde{t}_k^2$ by:

$$\tilde{t}_k^2 \equiv t_k^2 \|W^{-1}\hat{u}_k^{\mathrm{w}}\|^2 \sim t_k^2 \left(\frac{(c_k^{\mathrm{w}})^2}{\tau_k} + (s_k^{\mathrm{w}})^2 \mu_\varepsilon \right). \tag{70}$$

We also write out the eigendecomposition of Σ_x:

$$\Sigma_x = \sum_{k=1}^{r} \ell_k u_k u_k^\top. \tag{71}$$

From Theorem 3.2, the empirical PCs $\hat{u}_1, \ldots, \hat{u}_r$ are asymptotically pairwise orthonormal, and $\hat{u}_j$ and u_k are asymptotically orthogonal if $j \neq k$, and have absolute inner product c_k when $j = k$, given by (47).

Consequently, from Section 2.3.3 the optimal $\tilde{t}_k^2$ are defined by:

$$\tilde{t}_k^2 = \arg\min_{\ell} \mathcal{L}(A_k, \ell B_k), \tag{72}$$

where:

$$A_k = \begin{pmatrix} \ell_k & 0 \\ 0 & 0 \end{pmatrix}, \tag{73}$$

and

$$B_k = \begin{pmatrix} c_k^2 & c_k\sqrt{1-c_k^2} \\ c_k\sqrt{1-c_k^2} & 1-c_k^2 \end{pmatrix}. \tag{74}$$

As noted in Section 2.3.3, [21] provides closed form solutions to this minimization problem for many loss functions $\mathcal{L}$. For example, when operator norm loss is used the optimal $\tilde{t}_k^2$ is ℓ_k, and when Frobenius norm loss is used, the optimal $\tilde{t}_k^2$ is $\ell_k c_k^2$. When no such closed formula is known, the optimal values may be obtained by numerical minimization of (72).

Finally, the eigenvalues t_k^2 are obtained by inverting formula (70):

$$t_k^2 = \tilde{t}_k^2 \left(\frac{(c_k^{\mathrm{w}})^2}{\tau_k} + (s_k^{\mathrm{w}})^2 \mu_\varepsilon \right)^{-1}. \tag{75}$$

We summarize the covariance estimation procedure in Algorithm 2.

Remark 10. As stated in Remark 8, in practice the rank r will likely not be known a priori. We refer to Section 4.4 for a description of data-driven methods that may be used to estimate r.

4.3 Estimating the noise covariance Σ_ε

Algorithms 1 and 2 require access to the whitening transformation $W = \Sigma_\varepsilon^{-1/2}$, or equivalently the noise covariance matrix Σ_ε. However, the same method and analysis goes through unscathed if Σ_ε is replaced with an estimate $\hat{\Sigma}_\varepsilon$ that is consistent in operator norm, i.e., where

$$\lim_{p\to\infty} \|\Sigma_\varepsilon - \hat{\Sigma}_\varepsilon\|_{\mathrm{op}} = 0 \tag{76}$$

almost surely as $p/n \to \gamma$. Indeed, the distribution of the top r singular values and singular vectors of Y^{w} will be asymptotically identical whether the true $W = \Sigma_\varepsilon^{-1/2}$ is used to perform whitening or the estimated $\hat{W} = \hat{\Sigma}_\varepsilon^{-1/2}$ is used instead.

Remark 11. Because we assume that the maximum eigenvalue of Σ_ε is bounded and the minimum eigenvalue is bounded away from 0, (76) is equivalent to consistent estimation of the whitening matrix $W = \Sigma_\varepsilon^{-1/2}$ by $\hat{W} = \hat{\Sigma}_\varepsilon^{-1/2}$.

An estimator $\hat{\Sigma}_\varepsilon$ satisfying (76) may be obtained when we have access to an iid sequence of pure noise vectors $\varepsilon_1, \ldots, \varepsilon_{n'}$ in addition to the n signal-plus-noise vectors $Y_1, \ldots, Y_n$. This is the setting considered in [45], where a number of applications are also discussed. Here, we assume that $n' = n'(n)$ grows faster than $p = p(n)$, that is,

$$\lim_{n\to\infty} \frac{p(n)}{n'(n)} = 0. \tag{77}$$

Algorithm 2 Optimal eigenvalue shrinkage with whitening

Input: observations $Y_1, \dots, Y_n$; noise covariance Σ_ε; rank r
Define $Y = [Y_1, \dots, Y_n]/\sqrt{n}$; $W = \Sigma_\varepsilon^{-1/2}$; $Y^{\mathrm{w}} = WY$
Compute top r left singular vectors/values of Y^{w}: $\hat{u}_1^{\mathrm{w}}, \dots, \hat{u}_r^{\mathrm{w}}$; $\sigma_1^{\mathrm{w}}, \dots, \sigma_r^{\mathrm{w}}$
for all $k = 1, \dots, r$ **do**
    if $\sigma_k^{\mathrm{w}} > 1 + \sqrt{\gamma}$ **then**
        $\ell_k^{\mathrm{w}} = \left[(\sigma_k^{\mathrm{w}})^2 - 1 - \gamma + \sqrt{((\sigma_k^{\mathrm{w}})^2 - 1 - \gamma)^2 - 4\gamma}\right]/2$
        $c_k^{\mathrm{w}} = \sqrt{(1 - \gamma/(\ell_k^{\mathrm{w}})^2) \,/\, (1 + \gamma/\ell_k^{\mathrm{w}})}$
        $\mu_\varepsilon = \mathrm{tr}(\Sigma_\varepsilon)/p$
        $\tau_k = (c_k^{\mathrm{w}})^2 \Big/ \left[\|\Sigma_\varepsilon^{1/2} \hat{u}_k^{\mathrm{w}}\|^2 - (1 - (c_k^{\mathrm{w}})^2)\mu_\varepsilon\right]$
        $\ell_k = \ell_k^{\mathrm{w}}/\tau_k$
        $c_k = c_k^{\mathrm{w}} \Big/ \sqrt{(c_k^{\mathrm{w}})^2 + (1 - (c_k^{\mathrm{w}})^2)\mu_\varepsilon \tau_k}$
        $A_k = \begin{pmatrix} \ell_k & 0 \\ 0 & 0 \end{pmatrix}$
        $B_k = \begin{pmatrix} c_k^2 & c_k\sqrt{1-c_k^2} \\ c_k\sqrt{1-c_k^2} & 1 - c_k^2 \end{pmatrix}$
        $\tilde{t}_k^2 = \arg\min_\ell \mathcal{L}(A_k, \ell B_k)$
        $t_k^2 = \tilde{t}_k^2 \tau_k / [(c_k^{\mathrm{w}})^2 + (1 - (c_k^{\mathrm{w}})^2)\mu_\varepsilon \tau_k]$
    else if $\sigma_k^{\mathrm{w}} \le 1 + \sqrt{\gamma}$ **then**
        $t_k^2 = 0$
Output: $\hat{\Sigma}_x = \sum_{k=1}^r t_k^2 (W^{-1}\hat{u}_k^{\mathrm{w}})(W^{-1}\hat{u}_k^{\mathrm{w}})^\top$

In this case, we replace Σ_ε by the sample covariance:

$$\hat{\Sigma}_\varepsilon = \frac{1}{n'} \sum_{j=1}^{n'} \varepsilon_j \varepsilon_j^\top, \tag{78}$$

which converges to Σ_ε in operator norm; that is, (76) holds. In Section 8.5, we will illustrate the use of this method in simulations.

Remark 12. If p/n' does not converge to 0, then $\hat{\Sigma}_\varepsilon$ given by (78) is *not* a consistent estimator of Σ_ε in operator norm. Indeed, when $\Sigma_\varepsilon = I_p$ the distribution of $\hat{\Sigma}_\varepsilon$'s eigenvalues converges to the Marchenko-Pastur law [42], and more generally converges to a distribution whose Stieltjes transform is implicitly defined by a fixed point equation [5, 50, 49].

4.3.1 Alternative estimators of Σ_ε

Without access to an independent sequence of $n' \gg p$ pure noise samples, estimating the noise covariance Σ_ε consistently (with respect to operator norm) is usually hard as $p \to \infty$. However, it may still be practical when Σ_ε is structured. Examples include: when Σ_ε is sparse [13]; when Σ_ε^{-1} is sparse [56]; when Σ_ε is a circulant or Toeplitz matrix, corresponding to stationary noise [16]; and more generally, when the eigenbasis of Σ_ε is known a priori.

To elaborate on the last condition, let us suppose that the eigenbasis of Σ_ε is known, and without loss of generality that Σ_ε is diagonal; and suppose that and the u_k's are *delocalized* in that $\|u_k\|_\infty \to 0$ as $p \to \infty$. Write $\Sigma_\varepsilon = \mathrm{diag}(\nu_1, \dots, \nu_p)$, for unknown ν_i. In this setting, the sample variance of each coordinate will converge almost surely to the variance of the noise in that coordinate; that is, for $i = 1, \dots, p$, we have:

$$\hat{\nu}_i = \frac{1}{n}\sum_{j=1}^n Y_{ij}^2 = \frac{1}{n}\sum_{j=1}^n \left(\sum_{k=1}^r \ell_k u_{ki} z_{jk}\right)^2 + \frac{1}{n}\sum_{j=1}^n \varepsilon_{ij}^2 + 2\frac{1}{n}\sum_{j=1}^n \varepsilon_{ij} \sum_{k=1}^r \ell_k u_{ki} z_{jk} \to \nu_i, \tag{79}$$

where the limit is almost sure as $p, n \to \infty$. We have made use of the strong law of large numbers and the limit $\|u_k\|_\infty \to 0$.

Let $\hat{\Sigma}_\varepsilon$ have i^{th} diagonal entry $\hat{\nu}_i$. Then $\hat{\Sigma}_\varepsilon - \Sigma_\varepsilon$ is a mean-zero diagonal matrix, with diagonal entries $\hat{\nu}_i - \nu_i$; and the operator norm $\|\hat{\Sigma}_\varepsilon - \Sigma_\varepsilon\|_{\mathrm{op}} = \max_{1 \le i \le p} |\hat{\nu}_i - \nu_i|$, which is easily shown to go to 0 almost surely as $p \to \infty$ using the subgaussianity of the observations.

4.4 Estimating the rank r

A challenging question in principal component analysis is selecting the number of components corresponding to signal, and separating these from the noise. In our model, this corresponds to estimating the rank r of the matrix X, which is an input to Algorithms 1 and 2. A simple and natural estimate $\hat{r}$ of the rank is the following:

$$\hat{r} = \min\{k : \sigma_k^{\mathrm{w}} > 1 + \sqrt{\gamma} + \epsilon_n\}. \tag{80}$$

That is, we estimate the rank as the number of singular values of $Y^{\mathrm{w}} = X^{\mathrm{w}} + G$ exceeding the largest singular value of the noise matrix G, plus a small finite-sample correction factor $\epsilon_n > 0$. Any singular value exceeding $1 + \sqrt{\gamma} + \epsilon_n$ is attributable to signal, whereas any value below is consistent with pure noise.

When $\epsilon_n \equiv \epsilon$ for all n, it may be shown that in the large p, large n limit, $\hat{r}$ converges almost surely to the number of singular values of X^{w} exceeding $1 + \sqrt{\gamma} + \epsilon$. For small enough ϵ, this will recover all singular values of X^{w} exceeding $\sqrt{\gamma}$, and is likely sufficient for many applications. Furthermore, the correction ϵ_n may be calibrated using the Tracy-Widom distribution of the operator norm of $GG^\top$ by taking $\epsilon_n \sim n^{-2/3}$. Though a detailed discussion is beyond the scope of this paper, we refer to [35] for an approach along these lines.

An alternative procedure is similar to $\hat{r}$, but uses the original matrix Y rather than the whitened matrix Y^{w}:

$$\hat{r}' = \min\{k : \sigma_k > b_+ + \epsilon_n\}, \tag{81}$$

where b_+ is the asymptotic operator norm of the noise matrix N, and ϵ_n is a finite-sample correction factor. The value b_+ may be evaluated using, for example, the method from [37]. An estimator like this is proposed in [44]. In Section 8.8, we present numerical evidence that $\hat{r}$ may outperform $\hat{r}'$. More precisely, it appears that whitening can increase the gap between the smallest signal singular value and the bulk edge of the noise, making detection of the signal components more reliable.

Remark 13. We also remark that a widely-used method for rank estimation in non-isotropic noise is known as *parallel analysis* [29, 15, 14], which has been the subject of recent investigation [18, 20]. Other methods have also been explored [33].

5 Singular value shrinkage and linear prediction

In this section, we examine the relationship between singular value shrinkage and linear prediction. A linear predictor of X_j from Y_j is of the form AY_j, where A is a fixed matrix. It is known (see, e.g. [41]) that to minimize the expected mean-squared error, the best linear predictor, also called the *Wiener filter*, takes $A = \Sigma_x (\Sigma_x + \Sigma_\varepsilon)^{-1}$, and hence is of the form:

$$\hat{X}_j^{\mathrm{opt}} = \Sigma_x (\Sigma_x + \Sigma_\varepsilon)^{-1} Y_j. \tag{82}$$

We will prove the following result, which shows that in the classical regime $\gamma \to 0$, optimal shrinkage with whitening converges to the Wiener filter.

Theorem 5.1. *Suppose $Y_1, \dots, Y_n$ are drawn from the spiked model with heteroscedastic noise, $Y_j = X_j + \varepsilon_j$. Let $\hat{X}_1, \dots, \hat{X}_n$ be the predictors of $X_1, \dots, X_n$ obtained from singular value shrinkage with whitening, as described in Section 4.1 and Algorithm 1. Then almost surely in the limit $p/n \to 0$,*

$$\lim_{n \to \infty} \|\hat{X}^{\mathrm{opt}} - \hat{X}\|_{\mathrm{F}}^2 = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \|\hat{X}_j^{\mathrm{opt}} - \hat{X}_j\|^2 = 0. \tag{83}$$

In other words, the predictor $\hat{X}_j$ is asymptotically equivalent to the best linear predictor $\hat{X}_j^{\mathrm{opt}}$.

Theorem 5.1 is a consequence of the following result.

Theorem 5.2. *Suppose that the numbers s_k, $1 \le k \le r$ satisfy*

$$\lim_{\gamma \to 0} \frac{s_k}{\sigma_k^{\mathrm{w}}} = \frac{\ell_k^{\mathrm{w}}}{\ell_k^{\mathrm{w}} + 1}. \tag{84}$$

Then the predictor defined by

$$\hat{X}' = \sum_{k=1}^{r} s_k W^{-1} \hat{u}_k^{\mathrm{w}} (\hat{v}_k^{\mathrm{w}})^{\top} \tag{85}$$

satisfies

$$\lim_{n \to \infty} \|\hat{X}^{\mathrm{opt}} - \hat{X}'\|_{\mathrm{F}}^2 = 0, \tag{86}$$

where the limit holds almost surely as $p/n \to 0$.

We will also show that in the context of shrinkage methods, whitening is an *optimal* weighting of the data. To make this precise, we consider the following class of weighted shrinkage methods, which subsumes both ordinary singular value shrinkage and singular value shrinkage with noise whitening. For a fixed weight matrix Q, we multiply Y by Q, forming the matrix $Y^q = [QY_1, \ldots, QY_n]/\sqrt{n}$. We then apply singular value shrinkage to Y^q, with singular values $s_1^q, \ldots, s_r^q$, after which we apply the inverse weighting Q^{-1}. Clearly, ordinary shrinkage is the special case when $Q = I_p$, whereas singular value shrinkage with whitening is the case when $Q = W = \Sigma_\varepsilon^{-1/2}$.

When the singular values $s_1^q, \ldots, s_r^q$ are chosen optimally to minimize the AMSE, we will call the resulting predictor $\hat{X}_Q$, and denote by $\hat{X}_{Q,j}$ the denoised vectors so that $\hat{X}_Q = [\hat{X}_{Q,1}, \ldots, \hat{X}_{Q,n}]/\sqrt{n}$. In this notation, $\hat{X} = \hat{X}_W$ is optimal shrinkage with whitening, whereas $\hat{X}_I$ is ordinary shrinkage without whitening. The natural question is, what is the optimal matrix Q?

To answer this question, we introduce the linear predictors $\hat{X}_{Q,j}^{\mathrm{lin}}$, defined by

$$\hat{X}_{Q,j}^{\mathrm{lin}} = \sum_{k=1}^{r} \eta_k^q \langle QY_j, u_k^q \rangle Q^{-1} u_k^q, \tag{87}$$

where the $u_1^q, \ldots, u_r^q$ are the eigenvectors of $Q\Sigma_x Q$, and the η_k^q are chosen optimally to minimize the average AMSE across all n observations. We prove the following result, which is again concerned with the classical $\gamma \to 0$ regime.

Theorem 5.3. *Let $Q = Q_p$ be an element of a sequence of symmetric, positive-definite p-by-p matrices with bounded operator norm ($\|Q_p\|_{\mathrm{op}} \le C < \infty$ for all p). Then in the limit $p/n \to 0$, we have almost surely:*

$$\lim_{n \to \infty} \|\hat{X}_Q^{\mathrm{lin}} - \hat{X}_Q\|_{\mathrm{F}}^2 = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \|\hat{X}_{Q,j}^{\mathrm{lin}} - \hat{X}_{Q,j}\|^2 = 0. \tag{88}$$

In other words, the weighted shrinkage predictor $\hat{X}_{Q,j}$ is asymptotically equal to the linear predictor $\hat{X}_{Q,j}^{\mathrm{lin}}$.

Furthermore, $Q = W$ minimizes the AMSE:

$$W = \arg\min_{Q} \lim_{n \to \infty} \mathbb{E}\|\hat{X}_Q - X\|_{\mathrm{F}}^2. \tag{89}$$

The first part of Theorem 5.3, namely (88), states that any weighted shrinkage method converges to a linear predictor when $\gamma \to 0$. The second part of Theorem 5.3, specifically (89), states that of all weighted shrinkage schemes, whitening is optimal in the $\gamma \to 0$ regime.

Remark 14. A special case of Theorem 5.2 is the suboptimal "naive" shrinker with whitening, which uses singular values $\sqrt{\ell_k^{\mathrm{w}}}\, c_k^{\mathrm{w}} \bar{c}_k^{\mathrm{w}}$; see Figures 1 and 2 and the accompanying text. It is easily shown that Theorem 5.2 applies to this shrinker, and consequently that in the $\gamma \to 0$ limit this shrinker converges to the BLP. This fact will be illustrated numerically in Section 8.2.

We give detailed proofs of Theorems 5.1, 5.2 and 5.3 in Appendix B. In Section 5.1, we make a simple observation which underlies the proofs, which is of independent interest.

5.1 Columns of weighted singular value shrinkage

In this section, we show how to write the predictor $\hat{X}_Q$ in terms of the individual columns of $Y^q = [QY_1, \ldots, QY_n]/\sqrt{n}$. This observation will be used in the proofs of Theorems 5.1, 5.2 and 5.3, and also motivates the form of the out-of-sample predictor we will study in Section 6.

Let $m = \min(p, n)$. Consistent with our previous notation (when $Q = W$), we will denote by $\hat{u}_1^q, \ldots, \hat{u}_m^q$ the left singular vectors of the matrix Y^q, and we will denote by $\hat{v}_1^q, \ldots, \hat{v}_m^q$ the right singular vectors and $\sigma_1^q, \ldots, \sigma_m^q$ the corresponding singular values.

Lemma 5.4. *Each column $\hat{X}_{Q,j}$ of $\sqrt{n} \cdot \hat{X}_Q$ is given by the formula*

$$\hat{X}_{Q,j} = Q^{-1} \sum_{k=1}^{r} \eta_k^q \langle QY_j, \hat{u}_k^q \rangle \hat{u}_k^q, \tag{90}$$

where $\eta_k^q = s_k^q / \sigma_k^q$ is the ratio of the new and old singular values.

To see this, observe that we can write the j^{th} column of the matrix $\sqrt{n} \cdot Y^q$ as:

$$QY_j = \sum_{k=1}^{m} \sigma_k^q \hat{u}_k^q \hat{v}_{jk}^q, \tag{91}$$

and so by the orthogonality of $\hat{u}_k^q$, $\hat{v}_{jk}^q = \langle QY_j, \hat{u}_k^q \rangle / \sigma_k^q$. Consequently, when $\hat{X}_Q$ is obtained from Y^q by singular value shrinkage with singular values $s_1^q, \ldots, s_r^q$, followed by multiplication with Q^{-1}, we obtain formula (90).

6 Out-of-sample prediction

We now consider the problem of *out-of-sample* prediction. In Section 5.1, specifically Lemma 5.4, we saw that when applying the method of shrinkage with whitening, as described in Algorithm 1, each denoised vector $\hat{X}_j$ can be written in the form:

$$\hat{X}_j = \sum_{k=1}^{r} \eta_k \langle WY_j, \hat{u}_k^{\mathrm{w}} \rangle W^{-1} \hat{u}_k^{\mathrm{w}}, \tag{92}$$

where $\hat{u}_1^{\mathrm{w}}, \ldots, \hat{u}_r^{\mathrm{w}}$ are the top r left singular vectors of $Y^{\mathrm{w}} = WY$, and η_k are deterministic coefficients. We observe that the expression (92) may be evaluated for *any* vector Y_j, even when it is not one of the original $Y_1, \ldots, Y_n$, so long as we have access to the singular vectors $\hat{u}_k^{\mathrm{w}}$.

To formalize the problem, we suppose we have computed the sample vectors $\hat{u}_1^{\mathrm{w}}, \ldots, \hat{u}_r^{\mathrm{w}}$ based on n observed vectors $Y_1, \ldots, Y_n$, which we will call the *in-sample* observations. That is, the $\hat{u}_k^{\mathrm{w}}$ are the top left singular vectors of the whitened matrix $Y^{\mathrm{w}} = [Y_1^{\mathrm{w}}, \ldots, Y_n^{\mathrm{w}}]/\sqrt{n}$. We now receive a new observation $Y_0 = X_0 + \varepsilon_0$ from the same distribution, which we will refer to as an *out-of-sample* observation, and our goal is to predict the signal X_0.

We will consider predictors of the out-of-sample X_0 of the same form as (92):

$$\hat{X}_0 = \sum_{k=1}^{r} \eta_k^{\mathrm{o}} \langle WY_0, \hat{u}_k^{\mathrm{w}} \rangle W^{-1} \hat{u}_k^{\mathrm{w}}. \tag{93}$$

We wish to choose the coefficients η_k^{o} to minimize the AMSE, $\lim_{n\to\infty} \mathbb{E}\|\hat{X}_0 - X_0\|^2$.

Remark 15. We emphasize the difference between the in-sample prediction (92) and the out-of-sample prediction (93), beyond the different coefficients η_k and η_k^{o}. In (92), the vectors $u_1^{\mathrm{w}}, \ldots, u_r^{\mathrm{w}}$ are *dependent* on the in-sample observation Y_j, $1 \le j \le n$, because they are the top r left singular vectors of Y^{w}. However, in (93) they are *independent* of the out-of-sample observation Y_0, which is drawn independently from $Y_1, \ldots, Y_n$. As we will see, it is this difference that necessitates the different choice of coefficients η_k and η_k^{o} for the two problems.

In this section, we prove the following result comparing optimal out-of-sample prediction and in-sample prediction. Specifically, we derive the explicit formulas for the optimal out-of-sample coefficients η_k^{o} and the in-sample coefficients η_k; show that the coefficients are *not* equal; and show that the AMSE for both problems are nevertheless *identical*. Throughout this section, we assume the conditions and notation of Theorem 3.1.

Theorem 6.1. *Suppose $Y_1, \dots, Y_n$ are drawn iid from the spiked model, $Y_j = X_j + \varepsilon_j$, and $\hat{u}_1^{\mathrm{w}}, \dots, \hat{u}_r^{\mathrm{w}}$ are the top r left singular vectors of Y^{w}. Suppose $Y_0 = X_0 + \varepsilon_0$ is another sample from the same spiked model, drawn independently of $Y_1, \dots, Y_n$. Then the following results hold:*

1. *The optimal in-sample coefficients η_k are given by :*

$$\eta_k = \frac{(c_k^{\mathrm{w}})^2}{(c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \mu_\varepsilon \tau_k} \cdot \frac{\ell_k^{\mathrm{w}}}{\ell_k^{\mathrm{w}} + 1}. \tag{94}$$

2. *The optimal out-of-sample coefficients η_k^{o} are given by:*

$$\eta_k^{\mathrm{o}} = \frac{(c_k^{\mathrm{w}})^2}{(c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \mu_\varepsilon \tau_k} \cdot \frac{\ell_k^{\mathrm{w}}}{\ell_k^{\mathrm{w}} (c_k^{\mathrm{w}})^2 + 1}. \tag{95}$$

3. *The AMSEs for in-sample and out-of-sample prediction are identical, and equal to:*

$$AMSE = \sum_{k=1}^{r} \left(\frac{\ell_k^{\mathrm{w}}}{\tau_k} - \frac{(\ell_k^{\mathrm{w}})^2 (c_k^{\mathrm{w}})^4}{\ell_k^{\mathrm{w}} (c_k^{\mathrm{w}})^2 + 1} \frac{1}{\alpha_k \tau_k} \right), \tag{96}$$

where $\alpha_k = \left((c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \mu_\varepsilon \tau_k \right)^{-1}$.

Remark 16. To be clear, denoising each in-sample observation $Y_1, \dots, Y_n$ by applying (92) with η_k defined by (94) is *identical* to denoising $Y_1, \dots, Y_n$ by singular value shrinkage with whitening described in Algorithm 1. We derive this alternate form only to show that the coefficients η_k are different from the the optimal out-of-sample coefficients η_k^{o} to be used when Y_0 is independent from the $\hat{u}_k^{\mathrm{w}}$.

Remark 17. Theorem 6.1 extends the analogous result from [19], which was restricted to the standard spiked model with white noise.

The proof of Theorem 6.1 may be found in Appendix C. In Algorithm 3, we summarize the optimal out-of-sample prediction method, with the optimal coefficients derived in Theorem 6.1.

Algorithm 3 Optimal out-of-sample prediction

Input: Y_0; $\hat{u}_1^{\mathrm{w}}, \dots, \hat{u}_r^{\mathrm{w}}$; $\sigma_1^{\mathrm{w}}, \dots, \sigma_r^{\mathrm{w}}$
for all $k = 1, \dots, r$ **do**
  if $\sigma_k^{\mathrm{w}} > 1 + \sqrt{\gamma}$ **then**
    $\ell_k^{\mathrm{w}} = \left[(\sigma_k^{\mathrm{w}})^2 - 1 - \gamma + \sqrt{((\sigma_k^{\mathrm{w}})^2 - 1 - \gamma)^2 - 4\gamma} \right] / 2$
    $c_k^{\mathrm{w}} = \sqrt{\left(1 - \gamma/(\ell_k^{\mathrm{w}})^2\right) / \left(1 + \gamma/\ell_k^{\mathrm{w}}\right)}$
    $s_k^{\mathrm{w}} = \sqrt{1 - (c_k^{\mathrm{w}})^2}$
    $\mu_\varepsilon = \mathrm{tr}(\Sigma_\varepsilon)/p$
    $\tau_k = (c_k^{\mathrm{w}})^2 / \left[\|\Sigma_\varepsilon^{1/2} \hat{u}_k^{\mathrm{w}}\|^2 - (s_k^{\mathrm{w}})^2 \mu_\varepsilon \right]$
    $\alpha_k = 1 / \left((c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \mu_\varepsilon \tau_k \right)$
    $\eta_k^{\mathrm{o}} = \alpha_k \ell_k^{\mathrm{w}} (c_k^{\mathrm{w}})^2 / (\ell_k^{\mathrm{w}} (c_k^{\mathrm{w}})^2 + 1)$
  else if $\sigma_k^{\mathrm{w}} \le 1 + \sqrt{\gamma}$ **then**
    $\eta_k^{\mathrm{o}} = 0$
Output: $\hat{X}_0 = \sum_{k=1}^{r} \eta_k^{\mathrm{o}} \langle W Y_0, \hat{u}_k^{\mathrm{w}} \rangle W^{-1} \hat{u}_k^{\mathrm{w}}$

7 Subspace estimation and PCA

In this section, we focus on the task of *principal component analysis (PCA)*, or the estimation of the principal components $u_1, \dots, u_r$ of the signal X_j, and their span. Specifically, we assess the quality of the empirical PCs $\hat{u}_1, \dots, \hat{u}_r$ defined in (8). The reader may recall that these are constructed by whitening the observed vectors Y_j to produce Y_j^{w}; computing the top r left singular vectors of Y_j^{w}; and unwhitening and normalizing.

We first observe that in the classical regime $\gamma \to 0$, the angle between the subspaces $\text{span}\{\hat{u}_1, \dots, \hat{u}_r\}$ and $\text{span}\{u_1, \dots, u_r\}$ converges to 0 almost surely; we recall that the sine of the angle between subspaces $\mathcal{A}$ and $\mathcal{B}$ of $\mathbb{R}^p$ is defined by

$$\sin\Theta(\mathcal{A}, \mathcal{B}) = \|A_\perp^\top B\|_{\text{op}}, \tag{97}$$

where $A_\perp$ and B are matrices whose columns are orthonormal bases of $\mathcal{A}^\perp$ and $\mathcal{B}$, respectively.

Proposition 7.1. *Suppose $Y_1, \dots, Y_n$ are drawn from the spiked model, $Y_j = X_j + \varepsilon_j$. Let $\mathcal{U} = \text{span}\{u_1, \dots, u_r\}$ be the span of the population PCs, and $\hat{\mathcal{U}} = \text{span}\{\hat{u}_1, \dots, \hat{u}_r\}$ be the span of the empirical PCs. Then*

$$\lim_{n \to 0} \sin\Theta(\mathcal{U}, \hat{\mathcal{U}}) = 0, \tag{98}$$

where the limit holds almost surely as $n \to \infty$ and $p/n \to 0$.

The proof of Proposition 7.1 may be found in Appendix D.

Proposition 7.1 shows consistency of principal subspace estimation in the classical regime. We ask what happens in the high-dimensional setting $\gamma > 0$, where we typically do not expect to be able to have consistent estimation of the principal subspace. Our task here is to show that whitening will still improve estimation. To that end, in Section 7.1, we will show that under a uniform prior on the population PCs u_k, whitening improves estimation of the PCs. In Section 7.2, we will derive a bound on the error of estimating the principal subspace $\text{span}\{u_1, \dots, u_r\}$, under condition (19); we will show that the error rate matches the optimal rate of the estimator in [58]. Finally, in Section 7.3 we will complement these results by showing that under the uniform prior, whitening improves a natural signal-to-noise ratio.

7.1 Whitening improves subspace estimation for generic PCs

In this section, we consider the effect of whitening on estimating the PCs $u_1, \dots, u_r$. More precisely, we contrast two estimators of the u_k. On the one hand, we shall denote by $\hat{u}'_1, \dots \hat{u}'_r$ the left singular vectors of the raw data matrix Y, without applying any weighting matrix. On the other hand, we consider the vectors $\hat{u}_1, \dots, \hat{u}_r$ obtained by whitening, taking the top singular vectors of Y^{w}, unwhitening, and normalizing, as expressed by formula (8).

We claim that "generically", the vectors $\hat{u}_1, \dots, \hat{u}_r$ are superior estimators of $u_1, \dots, u_r$. By "generically", we mean when we impose a uniform prior over the population PCs $u_1, \dots, u_r$; that is, we assume the u_k are themselves random, drawn uniformly from the sphere in $\mathbb{R}^p$ and orthogonalized. This is precisely the "orthonormalized model" considered in [10].

We set $\tau = \lim_{p\to\infty} \text{tr}(\Sigma_\varepsilon^{-1})/p$, assuming this limit exists; and let $\varphi = \tau \cdot \mu_\varepsilon$. By Jensen's inequality, $\varphi \geq 1$, with strict inequality so long as Σ_ε is not a multiple of the identity.

Theorem 7.2. *Suppose Σ_ε has a finite number of distinct eigenvalues, each occurring with a fixed proportion as $p \to \infty$. Suppose too that $u_1, \dots, u_r$ are uniformly random orthonormal vectors in $\mathbb{R}^p$. Let $\hat{u}'_1, \dots, \hat{u}'_r$ be the left singular vectors of Y, and $\hat{u}_1, \dots, \hat{u}_r$ be the empirical PCs defined by (8). Then with probability approaching 1 as $n \to \infty$ and $p/n \to \gamma > 0$,*

$$|\langle \hat{u}'_k, u_k \rangle|^2 \leq R(\varphi) |\langle \hat{u}_k, u_k \rangle|^2, \quad 1 \leq k \leq r, \tag{99}$$

where R is decreasing, $R(1) = 1$, and $R(\varphi) < 1$ for $\varphi > 1$.

Furthermore, if $\hat{v}'_1, \dots, \hat{v}'_r$ are the right singular vectors of Y, and $\hat{v}_1, \dots, \hat{v}_r$ are the left singular vectors of Y^{w}, then

$$|\langle \hat{v}'_k, z_k \rangle|^2 \leq \tilde{R}(\varphi) |\langle \hat{v}_k, z_k \rangle|^2, \quad 1 \leq k \leq r, \tag{100}$$

with probability approaching 1 as $n \to \infty$ and $p/n \to \gamma > 0$, where $z_k = (z_{1k}, \dots, z_{nk})^\top / \sqrt{n}$, and where $\tilde{R}$ is decreasing, $\tilde{R}(1) = 1$, and $\tilde{R}(\varphi) < 1$ for $\varphi > 1$.

The proof of Theorem 7.2 may be found in Appendix D. It rests on a result from the recent paper [27], combined with the formula (47) for the asymptotic cosines between $\hat{u}_k$ and u_k.

Remark 18. The definition of $\tau = \text{tr}(\Sigma_\varepsilon^{-1})/p$ is consistent with our definition of $\tau_k = \lim_{p\to\infty} \|W^{-1} u_k^{\text{w}}\|^{-2}$ from (15). Indeed, since Theorem 7.2 assumes that $u_1, \dots, u_r$ are uniformly random unit vectors, the PCs u_k^{w} of X^{w} are asymptotically identical to $Wu_k / \|Wu_k\|$, since these vectors are almost surely orthogonal as $p \to \infty$. Consequently, for each $1 \leq k \leq r$ we have

$$\tau_k = \lim_{p\to\infty} \frac{1}{\|W^{-1} u_k^{\text{w}}\|^2} = \lim_{p\to\infty} \|Wu_k\|^2 \sim \frac{1}{p}\text{tr}(W^2) = \frac{1}{p}\text{tr}(\Sigma_p^{-1}) \sim \tau. \tag{101}$$

7.2 Minimax optimality of the empirical PCs

In this section, we consider the question of whether the empirical PCs $\hat{u}_1, \dots, \hat{u}_r$ can be significantly improved upon. In the recent paper [58], an estimator $\hat{\mathcal{U}}$ of the principal subspace $\mathcal{U} = \text{span}\{u_1, \dots, u_r\}$ is proposed that achieves the following error rate:

$$\mathbb{E}[\sin\Theta(\hat{\mathcal{U}}, \mathcal{U})] \leq \min\left\{ C\sqrt{\gamma}\left(\frac{\mu_\varepsilon^{1/2} + (r/p)^{1/2}\|\Sigma_\varepsilon\|_{\text{op}}^{1/2}}{\min_k \ell_k^{1/2}} + \frac{\mu_\varepsilon^{1/2}\|\Sigma_\varepsilon\|_{\text{op}}^{1/2}}{\min_k \ell_k} \right), 1 \right\}, \tag{102}$$

where C is a constant dependent on the *incoherence* of $u_1, \dots, u_r$, defined by $I(U) = \max_{1 \leq j \leq p} \|e_j^\top U\|^2$ where $U = [u_1, \dots, u_r] \in \mathbb{R}^{p \times r}$. Furthermore, the error rate (102) is shown to be *minimax optimal* over the class of models with PCs of bounded incoherence.

In this section, we show that when (19) holds, then the empirical PCs $\hat{u}_1, \dots, \hat{u}_r$ achieve the same error rate (102) almost surely in the limit $n \to \infty$, $p/n \to \gamma$. More precisely, we show the following:

Theorem 7.3. *Assume that the weighted orthogonality condition* (19) *holds. Suppose that* Σ_ε *is diagonal, and that there is a constant* C *so that*

$$|u_{jk}| \leq \frac{C}{\sqrt{p}} \tag{103}$$

for all $k = 1, \dots, r$ *and* $j = 1, \dots, p$. *Suppose* $Y_1, \dots, Y_n$ *are drawn iid from the spiked model. Let* $\hat{u}_1, \dots, \hat{u}_r$ *be the estimated PCs from equation* (8)*, and let* $\hat{\mathcal{U}} = \text{span}\{\hat{u}_1, \dots, \hat{u}_r\}$ *and* $\mathcal{U} = \text{span}\{u_1, \dots, u_r\}$.

Then almost surely in the limit $p/n \to \gamma$

$$\sin^2\Theta(\hat{\mathcal{U}}, \mathcal{U}) \leq \min\left\{ K\gamma\mu_\varepsilon\left(\frac{1}{\min_k \ell_k} + \frac{\|\Sigma_\varepsilon\|_{\text{op}}}{\min_k \ell_k^2} \right), 1 \right\}, \tag{104}$$

where K *is a constant depending only on* C *from* (103).

Remark 19. Theorem 7.3 shows that in the case $r = 1$, the estimate $\hat{u}$ obtained by whitening Y, computing the top left singular vector of Y^{w}, and then unwhitening and normalizing, is asymptotically minimax optimal. When $r > 1$, we require the extra condition (19) which does not appear in the minimax lower bound from [58].

The proof of Theorem 7.3 follows from the formula (47) for the cosines between u_k and $\hat{u}_k$ from Theorem 3.2. The details are found in Appendix D.

7.3 Whitening increases the operator norm SNR

In this section, we define a natural signal-to-noise ratio (SNR) for the spiked model, namely the ratio of operator norms between the signal and noise sample covariances. We show that under the generic model from Section 7.1 for the signal principal components u_k, the SNR increases after whitening.

We define the SNR by:

$$\text{SNR} = \frac{\|\hat{\Sigma}_x\|_{\text{op}}}{\|\hat{\Sigma}_\varepsilon\|_{\text{op}}} \tag{105}$$

where $\hat{\Sigma}_x = \frac{1}{n}\sum_{j=1}^n X_j X_j^\top$ and $\hat{\Sigma}_\varepsilon = \frac{1}{n}\sum_{j=1}^n \varepsilon_j \varepsilon_j^\top$ are the sample covariances of the signal and noise components, respectively (neither of which are observed).

After whitening, the observations change into:

$$Y_j^{\text{w}} = X_j^{\text{w}} + G_j, \tag{106}$$

and we define the new SNR to be:

$$\text{SNR}^{\text{w}} = \frac{\|\hat{\Sigma}_x^{\text{w}}\|_{\text{op}}}{\|\hat{\Sigma}_g\|_{\text{op}}} \tag{107}$$

where $\hat{\Sigma}_x^{\text{w}} = \frac{1}{n}\sum_{j=1}^n X_j^{\text{w}}(X_j^{\text{w}})^\top$ and $\hat{\Sigma}_g = \frac{1}{n}\sum_{j=1}^n G_j G_j^\top$.

As in Section 7.1, let $\tau = \lim_{p\to\infty} \text{tr}(\Sigma_\varepsilon^{-1})/p$ (assuming the limit exists), and define $\varphi = \tau \cdot \mu_\varepsilon$. Note that by Jensen's inequality, $\varphi \geq 1$, with strict inequality unless $\Sigma_\varepsilon = \nu I_p$. We will prove the following:

Proposition 7.4. *Suppose the population principal components $u_1, \dots, u_r$ are uniformly random orthonormal vectors in $\mathbb{R}^p$. Then in the limit $p/n \to \gamma > 0$,*

$$\mathrm{SNR}^{\mathrm{w}} \geq \varphi \mathrm{SNR}. \tag{108}$$

In other words, Proposition 7.4 states that for generic signals whitening increases the operator norm SNR by a factor of at least $\varphi \geq 1$. The proof may be found in Appendix D.

Remark 20. As explained in Remark 18, under the generic model assumed by Proposition 7.4, the notation τ is consistent with the definition of τ_k in (15).

Remark 21. Proposition 7.4 is similar in spirit to a result in [38], which essentially shows that the SNR defined by the nuclear norms, rather than operator norms, increases after whitening. However, in the $p \to \infty$ limit, defining the SNR using the ratio of nuclear norms is not as meaningful as using operator norms, because the ratio of nuclear norms always converges to 0 in the high-dimensional limit. Indeed, we have:

$$\|\hat{\Sigma}_x\|_* \to \sum_{k=1}^{r} \ell_k, \tag{109}$$

almost surely as $p, n \to \infty$. On the other hand,

$$\frac{1}{p}\|\hat{\Sigma}_\varepsilon\|_* \to \mu_\varepsilon. \tag{110}$$

In particular, $\|\hat{\Sigma}_\varepsilon\|_*$ grows like p, whereas $\|\hat{\Sigma}_x\|_*$ is bounded with p. When p is large, therefore, the norm of the noise swamps the norm of the signal. On the other hand, the operator norms of $\hat{\Sigma}_x$ and $\hat{\Sigma}_\varepsilon$ are both bounded, and may therefore be comparable in size.

8 Numerical results

In this section we report several numerical results that illustrate the performance of our predictor in the spiked model, as well as several beneficial properties of whitening. Code implementing the shrinkage with whitening algorithms will be made available online.

8.1 Comparison to the best linear predictor

In this experiment, we compared our predictor to the best linear predictor (BLP), defined in equation (82). The BLP is an oracle method, as it requires knowledge of the population covariance Σ_x, which is not accessible to us. However, Theorem 5.1 predicts that as $p/n \to 0$, the optimal shrinkage with whitening predictor will behave identically to the BLP.

In the same experiments, we also compare our method to OptShrink [44], the optimal singular value shrinker without any transformation. Theorem 5.3 predicts that as $p/n \to 0$, OptShrink will behave identically to a suboptimal linear filter.

In these these tests, we fixed a dimension equal to $p = 100$, and let n grow. Each signal was rank 3, with PCs chosen so that the first PC was a completely random unit vector, the second PC was set to zero on the first $p/2$ coordinates and random on the remaining coordinates, and the third PC was completely random on the first $p/2$ coordinates and zero on the remaining coordinates. The signal random variables z_{jk} were chosen to be Gaussian.

The noise covariance matrix Σ_ε was generated by taking equally spaced values between 1 and a specified condition number $\kappa > 1$, and then normalizing the resulting vector of eigenvalues to be a unit vector. This normalization was done so that in each test, the total energy of the noise remained constant.

Figure 3 plots the average prediction errors as a function of n for the three methods, for different condition numbers κ of the noise covariance Σ_ε. The errors are averaged over 500 runs of the experiment, with different draws of signal and noise. As expected, the errors for optimal shrinkage with whitening converge to those of the oracle BLP, while the errors for OptShrink appear to converge to a larger value, namely the error of the limiting suboptimal linear filter.

Remark 22. Unlike shrinkage with whitening, OptShrink does not make use of the noise covariance. Though access to the noise covariance would permit faster evaluation of the OptShrink algorithm using, for instance, the methods described in [37], we have found that this does not change the estimation accuracy of the method. Similarly, the BLP uses the true PCs of X_j, which are not used by either shrinkage method. The comparison between the methods must be understood in that context.

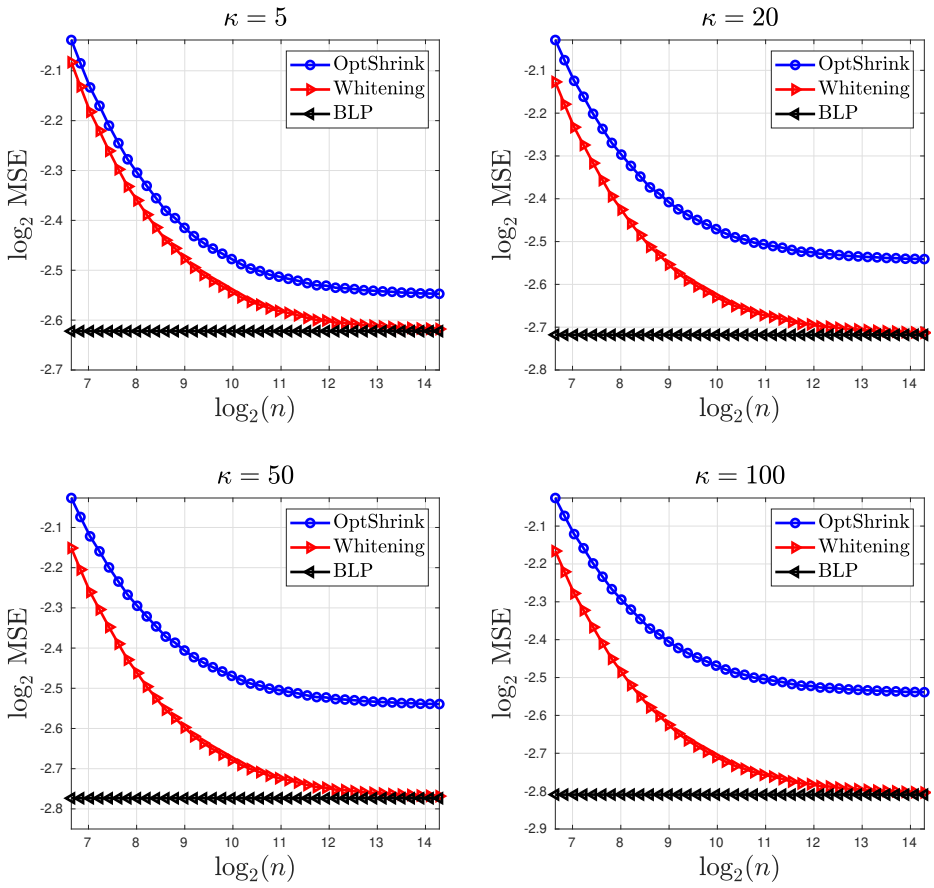

Figure 3: Prediction errors for the optimal whitened shrinker, the optimal unwhitened shrinker (OptShrink), and the best linear predictor (an oracle method).

8.2 Performance of singular value shrinkage

We examine the performance of optimal shrinkage with whitening for different values of γ and different condition numbers of the noise covariance. We compare to OptShrink [44] and the naive shrinker with whitening employed in [19], which uses singular values $\sqrt{\ell_k^{\mathrm{w}}} c_k^{\mathrm{w}} \tilde{c}_k^{\mathrm{w}}$; see Figures 1 and 2 and the associated text. This latter shrinker does not account for the change in angle between the singular vectors resulting from unwhitening.

In each run of the experiment, we fix the dimension $p = 1000$. We use a diagonal noise covariance with a specified condition number κ, whose entries are linearly spaced between $1/\kappa$ and 1, and increase with the index. We generate the orthonormal basis of PCs u_1, u_2, u_3 from the model described in Section 2.1.2, as follows: u_1 is a unifomly random unit vector; u_2 has Gaussian entries with linearly-spaced variances $a_1, \ldots, a_p$, where $a_p < a_{p-1} < \cdots < a_1$, $\sum_{i=1}^p a_i = 1$, and $a_1/a_p = 10$; and u_3 has Gaussian entries with linearly-spaced variances $b_1, \ldots, b_p$, where $b_1 < b_2 < \cdots < b_p$, $\sum_{i=1}^p b_i = 1$, and $b_p/b_1 = 10$. Gram-Schmidt is then performed on u_1, u_2, and u_3 to ensure they are orthonormal. For aspect ratio γ, the three signal singular values are $\gamma^{1/4} + i/2$, $i = 1, 2, 3$.

For different values of n, and hence of γ, we generate 50 draws of the data and record the average relative errors for each of the three methods. The results are plotted in Figure 4. As is apparent from the figures, both whitening methods typically outperform OptShrink. Furthermore, when n is large, both optimal shrinkage and naive shrinkage perform very similarly; this makes sense because both methods converge to the BLP as $n \to \infty$. By contrast, when γ is large, the benefits of using the optimal shrinker over the naive shrinker are more apparent.

Remark 23. As noted in Remark 22, we emphasize that unlike both whitening methods, OptShrink does not make use of the noise covariance, and the comparison between the methods must be understood in that context.

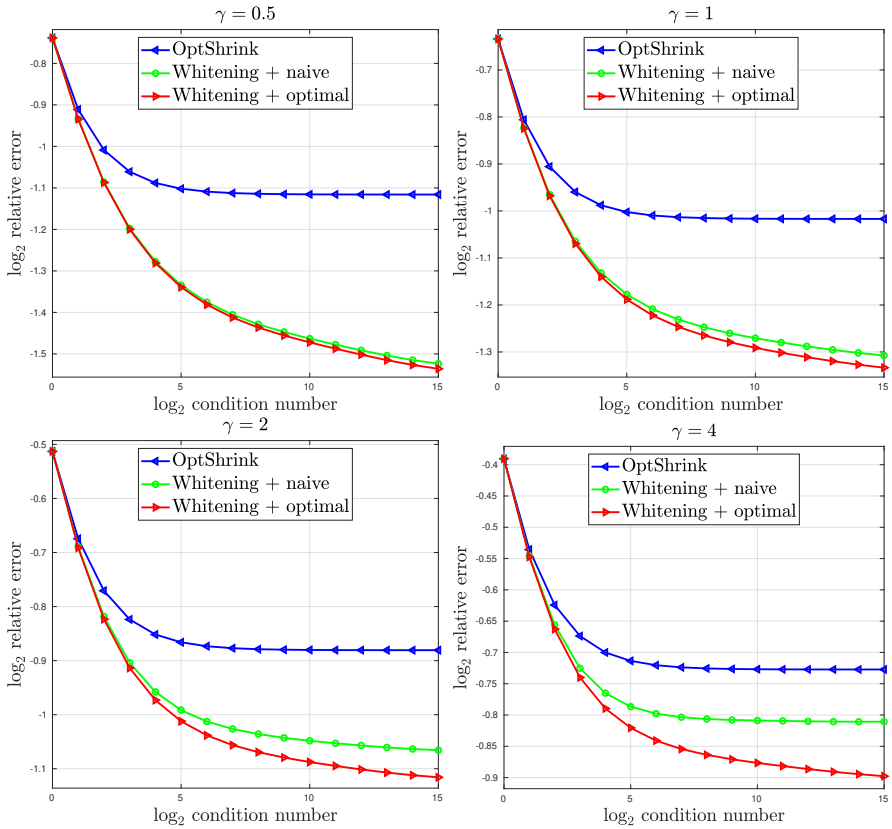

Figure 4: Comparison of whitening with optimal shrinkage; whitening with naive shrinkage; and OptShrink (no whitening), as a function of the noise covariance matrix's condition number κ.

8.3 Performance of eigenvalue shrinkage

We examine the performance of optimal eigenvalue shrinkage with whitening for different values of γ and different condition numbers of the noise covariance. We use nuclear norm loss, for which the optimal $\tilde{t}_k^2$ in Algorithm 2 is given by the formula

$$\tilde{t}_k^2 = \max\{\ell_k(2c_k^2 - 1), 0\}. \tag{111}$$

This formula is derived in [21].

We compare to two other methods. We consider optimal eigenvalue shrinkage without whitening, where the population eigenvalues and cosines between observed and population eigenvectors are estimated using the methods from [44]. We also consider the whitening and eigenvalue shrinkage procedure from [38], which shrinks the eigenvalues to the population values ℓ_k; this is an optimal procedure for operator norm loss [21], but suboptimal for nuclear norm loss.

As in Section 8.2, in each run of the experiment, we fix the dimension $p = 1000$. We use a diagonal noise covariance with a specified condition number κ, whose entries are linearly spaced between $1/\kappa$ and 1, and increase with the index. We generate the orthonormal basis of PCs u_1, u_2, u_3 from the model described in Section 2.1.2, as follows: u_1 is a unifomly random unit vector; u_2 has Gaussian entries with linearly-spaced variances $a_1, \ldots, a_p$, where $a_p < a_{p-1} < \cdots < a_1$, $\sum_{i=1}^p a_i = 1$, and $a_1/a_p = 10$; and u_3 has Gaussian entries with linearly-spaced variances $b_1, \ldots, b_p$, where $b_1 < b_2 < \cdots < b_p$, $\sum_{i=1}^p b_i = 1$, and $b_p/b_1 = 10$. Gram-Schmidt is then performed on u_1, u_2, and u_3 to ensure they are orthonormal. For aspect ratio γ, the three signal singular values are $\gamma^{1/4} + i$, $i = 1, 2, 3$.

For different values of n, and hence of γ, we generate 50 draws of the data and record the average relative errors $\|\hat{\Sigma}_x - \Sigma_x\|_*/\|\Sigma_x\|_*$ for each of the three methods. The results are plotted in Figure 5. As is apparent from the figures, optimal shrinkage with whitening outperforms the other two methods. For the smaller values of γ, optimal shrinkage without whitening outperforms the population shrinker with whitening when the condition number κ is small, since the benefits of whitening are not large; however, as κ grows, whitening with the suboptimal population shrinker begins to outperform. For larger γ, the cost of using the wrong shrinker outweigh the benefits of whitening, and the population shrinker with whitening is inferior to both other methods. This illustrates the importance of using a shrinker designed for the intended loss function.

8.4 Numerical comparison of the angles

In this section, we numerically illustrate Theorem 7.2 by examining the angles between the spanning vectors $\hat{u}_k$ (the empirical PCs) and $\hat{v}_k$ of $\hat{X}$ and, respectively, the population vectors u_k (the population PCs) and v_k. We show that these angles are smaller (or equivalently, their cosines are larger) than the corresponding angles between the population u_k and v_k and the singular vectors of the unwhitened data matrix Y.

Figure 6 plots the cosines as a function of the condition number κ of the noise matrix Σ_ε. In this experiment, we consider a rank 1 signal model for simplicity, with a uniformly random PC. We used dimension $p = 500$, and drew $n = 1000$ observations. For each condition number κ of Σ_ε, we generate Σ_ε as described in Section 8.1. For each test, we average the cosines over 50 runs of the experiment (drawing new signals and new noise each time). Both signal and noise are Gaussian. As we see, the cosines improve dramatically after whitening. As κ grows, i.e., the noise becomes more heteroscedastic, the improvement becomes more pronounced.

8.5 Estimating the noise covariance

In many applications, the true noise covariance may not be accessible. In this experiment, we consider the effect of estimating the noise covariance by the sample covariance from n' iid samples of pure noise, $\varepsilon_1, \ldots, \varepsilon_{n'}$, as n' grows.

We fix the dimension $p = 500$ and number of signal-plus-noise observations $n = 625$, and $r = 2$ signal singular values 3 and 5. We take the noise covariance to have condition number $\kappa = 500$, with eigenvalues equispaced between 1/100 and 1/5. The eigenvectors of the noise covariance are drawn uniformly at random.

For increasing values of $n' \geq p$, we draw n' iid realizations of the noise $\varepsilon_1, \ldots, \varepsilon_{n'}$, and form the

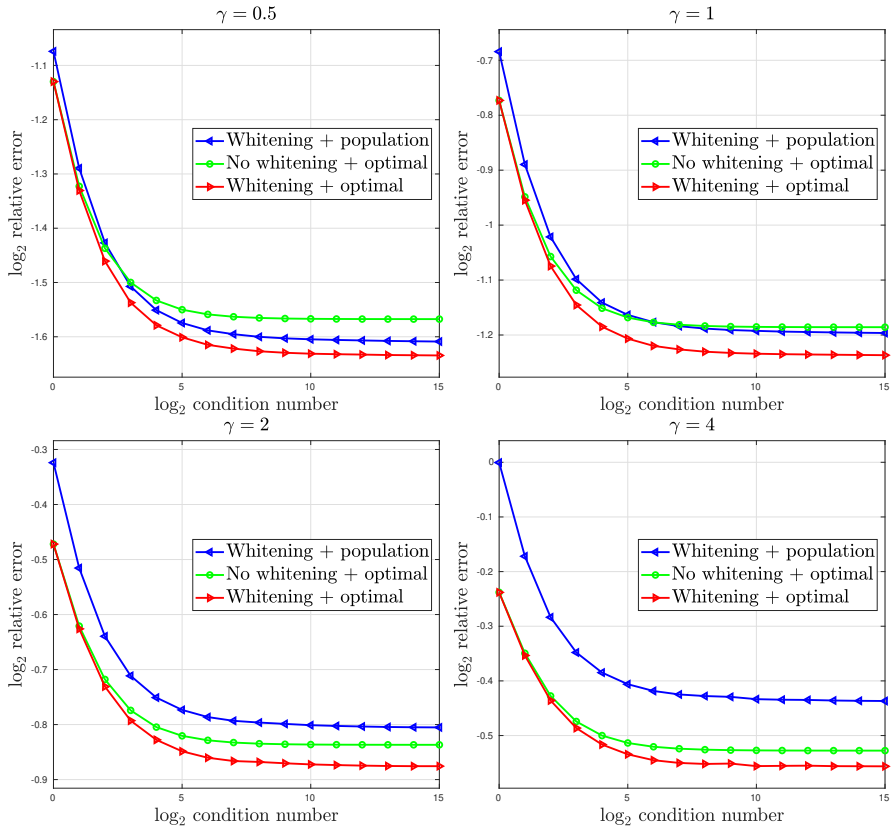

Figure 5: Comparison of whitening with optimal shrinkage; whitening with naive shrinkage; and OptShrink (no whitening), as a function of the noise covariance matrix's condition number κ.

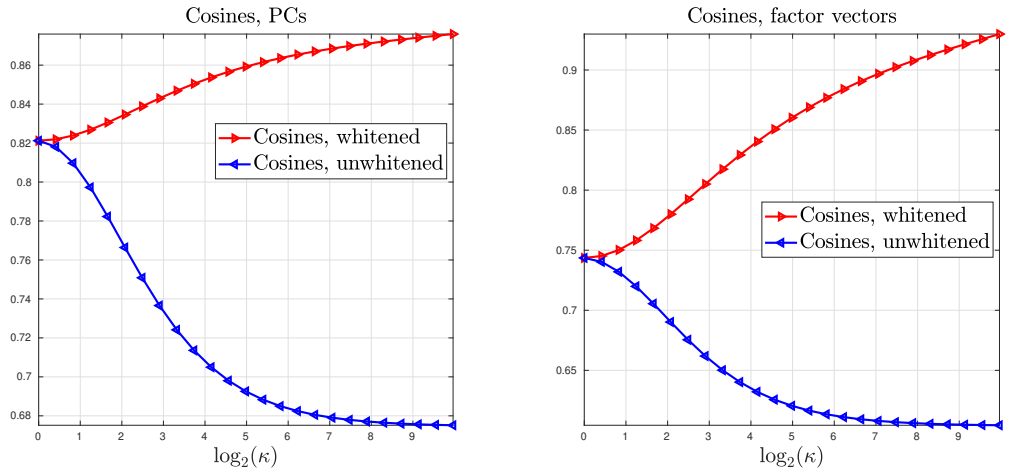

Figure 6: Comparison of the cosines between the empirical and population singular vectors, for the raw data and the whitened data, as a function of the noise covariance matrix's condition number κ.

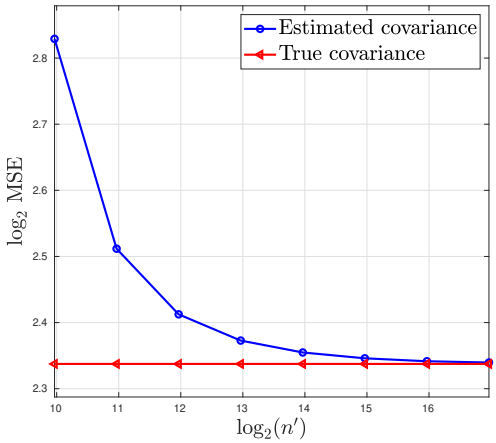

Figure 7: Comparison of the errors when using the true noise covariance Σ_ε and the sample noise covariance $\hat{\Sigma}_\varepsilon$ estimated from n' samples.

sample covariance:

$$\hat{\Sigma}_\varepsilon = \frac{1}{n'} \sum_{i=1}^{n'} \varepsilon_i \varepsilon_i^\top . \tag{112}$$

For each n', we perform Algorithm 1 using the sample covariance $\hat{\Sigma}_\varepsilon$. The experiment is repeated 2000 times for each value of n', and the errors averaged over these 2000 runs. Figure 7 plots the average error as a function of n'. We also apply Algorithm 1 using the true noise covariance Σ_ε, and plot the average error (which does not depend on n') in Figure 7 as well. The error when using the estimated covariance converges to the error when using the true covariance, indicating that Algorithm 1 is robust to estimation of the covariance.

8.6 Accuracy of error formulas and estimates

In this experiment, we test the accuracy of the error formula (96). There are three distinct quantities that we define. The first is the oracle AMSE, which we define from the known population parameters. The second is the estimated AMSE, which we will denote by $\widehat{\text{AMSE}}$; this is estimated using the observations $Y_1, \ldots, Y_n$ themselves. The third is the mean-squared error itself, $\|\hat{X} - X\|_{\text{F}}^2/n$. Of the three quantities, only $\widehat{\text{AMSE}}$ would be directly observed in practice. We define the discrepancy between AMSE and $\|\hat{X} - X\|_{\text{F}}^2/n$ as $|\text{AMSE} - \|\hat{X} - X\|_{\text{F}}^2/n|$, and the discrepancy between $\widehat{\text{AMSE}}$ and $\|\hat{X} - X\|_{\text{F}}^2$ as $|\widehat{\text{AMSE}} - \|\hat{X} - X\|_{\text{F}}^2/n|$.

Figure 8 plots the log discrepancies against $\log_2(p)$. We also include a table of the values themselves. In all experiments, we use the following parameters: the aspect ratio is $\gamma = 0.8$, the rank $r = 2$, the signal singular values are 3 and 2, u_1 is $\sqrt{2/p}$ on entries $1, \ldots, p/2$ and 0 elsewhere, u_2 is $\sqrt{2/p}$ on entries $p/2 + 1, \ldots, p$ and 0 elsewhere, and the noise covariance is diagonal with variances linearly spaced from $1/200$ to $3/2$, increasing with the coordinates.

We make two observations. First, the slope of each plot is approximately 0.5, indicating that the error formulas derived are accurate with error $O(n^{-1/2})$. This is precisely the rate we expect from [8]. Second, the discrepancies of AMSE and $\widehat{\text{AMSE}}$ are very close, and in fact the discrepancy of $\widehat{\text{AMSE}}$ is slightly smaller than that of AMSE. This indicates that the observed $\widehat{\text{AMSE}}$ provides a viable estimate for the actual error $\|\hat{X} - X\|_{\text{F}}^2/n$.

8.7 Comparing in-sample and out-of-sample prediction

In this next experiment, we compare the performance of in-sample and out-of-sample prediction, as described in Section 6. Optimal in-sample prediction is identical to performing optimal singular value shrinkage with noise whitening to the in-sample data $Y_1, \ldots, Y_n$. For out-of-sample prediction, we use the expression of the form (93) with the optimal coefficients η_k° from Proposition 6.1.

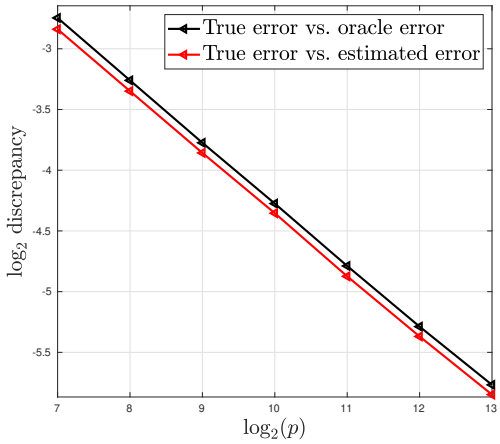

Figure 8: Logarithm of the discrepancies $|\text{AMSE} - \|\hat{X} - X\|_{\text{F}}^2/n|$ and $|\widehat{\text{AMSE}} - \|\hat{X} - X\|_{\text{F}}^2/n|$, versus $\log_2(p)$. AMSE is the oracle value of the error, and $\widehat{\text{AMSE}}$ is estimated from the data itself.

$\log_2(p)$	Discrepancy, AMSE	Discrepancy, $\widehat{\text{AMSE}}$
7	1.49e-01	1.40e-01
8	1.04e-01	9.82e-02
9	7.31e-02	6.90e-02
10	5.17e-02	4.89e-02
11	3.62e-02	3.41e-02
12	2.56e-02	2.42e-02
13	1.84e-02	1.74e-02

Table 3: Discrepancies $|\text{AMSE} - \|\hat{X} - X\|_{\text{F}}^2/n|$ and $|\widehat{\text{AMSE}} - \|\hat{X} - X\|_{\text{F}}^2/n|$. AMSE is the oracle value of the error, and $\widehat{\text{AMSE}}$ is estimated from the data itself.

We ran the following experiments. For a fixed dimension p, we generated a random value of $n > p$. We then chose three random PCs from the same model described in Section 8.1, and we generated pools of n in-sample and out-of-sample observations. We performed optimal shrinkage with whitening on the in-sample observations, and applied the out-of-sample prediction to the out-of-sample data using the vectors $\hat{u}_k^{\text{w}}$ computed from the in-sample data. We then computed the MSEs for the in-sample and out-of-sample data matrices. This whole procedure was repeated 2000 times.

Figure 9 shows scatterplots of the in-sample and out-of-sample predictions for $p = 50$ and $p = 500$. In both plots, we see that there is not a substantial difference between the in-sample and out-of-sample prediction errors, validating the asymptotic prediction made by Proposition 6.1. Even for the low-dimension of $p = 50$, there is very close agreement between the performances, and for $p = 500$ they perform nearly identically.

8.8 Signal detection and rank estimation

In this experiment, we show that whitening improves signal detection. We generated data from a rank 1 model, with a weak signal. We computed all the singular values of the original data matrix Y, and the whitened matrix Y^{w}. Figure 10 plots the the top 20 singular values for each matrix.

It is apparent from the comparison of these figures that the top singular value of the whitened matrix pops out from the bulk of noise singular values, making detection of the signal component very easy in this case. By contrast, the top singular value of the raw, unwhitened matrix Y^{w} does not stick out from the bulk. Proposition 7.4 would lead us to expect this type of behavior, since the signal matrix increases in strength relative to the noise matrix.

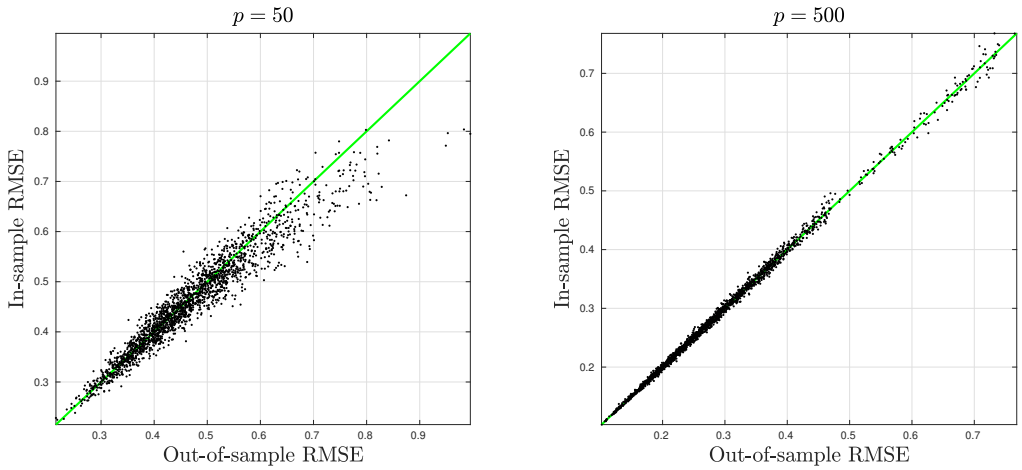

Figure 9: Comparison of in-sample and out-of-sample denoising for $p = 50$ and $p = 500$.

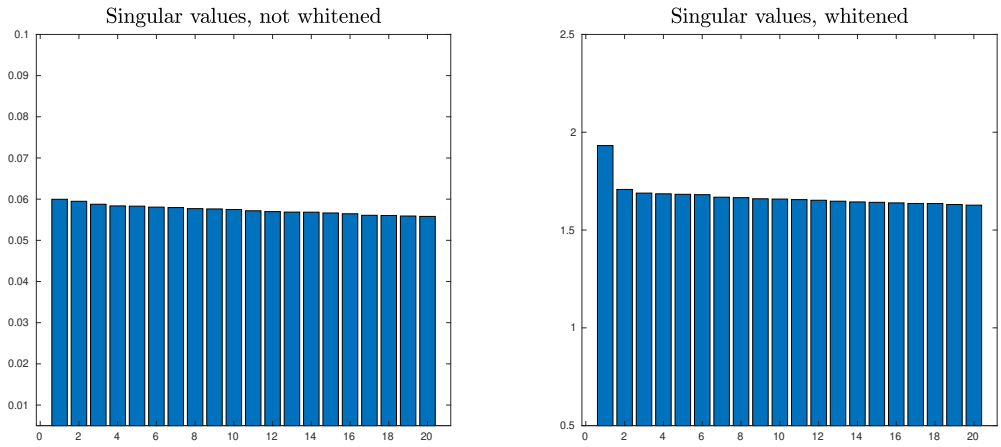

Figure 10: The top 20 empirical singular values of the raw data matrix Y and the whitened data matrix Y^{w}, for a rank 1 signal.

8.9 Non-gaussian noise

The theory we have derived relies on the orthogonal invariance of the noise matrix G. In this experiment, we study the agreement between the theoretically predicted values for c_k and $\tilde{c}_k$ and the observed values for finite n and p and non-Gaussian noise.

For different values of n we generated rank 1 signal matrices of size $n/2$-by-n, with top PC u having all entries equal to $1/\sqrt{1000}$, z_j Gaussian, and signal energy $\ell = 1$. We generated a noise matrix, where each entry has mean 0 and variance 1, drawn iid from a specified distribution. We then colored the noise matrix by multiplying it by $\Sigma_\varepsilon^{1/2} = \mathrm{diag}(\sqrt{\nu_1}, \ldots, \sqrt{\nu_p})$, where $\nu_1, \ldots, \nu_p$ are linearly spaced, $\nu_1 = 1/500$, and $\nu_p = 1$.

We considered four different distributions for the entries of G: the Gaussian distribution; the Rademacher distribution; and the Student t distributions with 10 and 3 degrees of freedom (normalized to have variance 1). For each distribution, we drew signal/noise pairs, and computed the absolute value of the cosines between the topmost left and right singular vectors of the observed matrix and the left and right singular vectors of the signal matrix. We then computed the average absolute difference (the discrepancy) between the observed cosines and the theoretically predicted values c and $\tilde{c}$ from Section 3. The errors are averaged over 20000 runs.

Table 4 contains the average discrepancies for c, and Table 5 contains the average errors for $\tilde{c}$, both for $n = 1000, 2000, 4000, 8000$. For the t distribution with 10 degrees of freedom and the Rademacher distribution, the discrepancies match those of the Gaussian to within the precision of the experiment. In particular, for these three noise distributions, the observed cosines appear to converge to the predicted

n	Gaussian	Rademacher	t, df=10	t, df=3
1000	8.173e-03	8.009e-03	8.147e-03	2.584e-01
2000	5.742e-03	5.794e-03	5.750e-03	3.610e-01
4000	4.069e-03	4.073e-03	4.071e-03	4.730e-01
8000	2.896e-03	2.933e-03	2.897e-03	5.866e-01

Table 4: Average discrepancies between c and $|\langle u, \hat{u} \rangle|$.

n	Gaussian	Rademacher	t, df=10	t, df=3
1000	3.627e-03	3.625e-03	3.650e-03	2.598e-01
2000	2.704e-03	2.707e-03	2.712e-03	3.708e-01
4000	1.951e-03	1.939e-03	1.952e-03	4.895e-01
8000	1.409e-03	1.388e-03	1.410e-03	6.112e-01

Table 5: Average discrepancies between $\tilde{c}$ and $|\langle v, \hat{v} \rangle|$.

asymptotic values at a rate of roughly $O(n^{-1/2})$. By contrast, for the t distribution with only 3 degrees of freedom, there is substantial discrepancy between the theoretical and observed cosines, and the discrepancies do not decrease with n (in fact, they grow).

These numerical results suggest that for noise distributions with sufficiently many finite moments, the distributions are approximately equal as those Gaussian noise, which in turn suggests that the limiting cosine values we have derived for Gaussian noise may hold for more general distributions.

9 Conclusions and future work

We have derived the optimal spectral shrinkers method for signal prediction and covariance estimation in the spiked model with heteroscedastic noise, where the data is whitened before shrinkage and unwhitened after shrinkage. We also showed the in that $\gamma \to 0$ regime, optimal singular value shrinkage with whitening converges to the best linear predictor, whereas optimal shrinkage without whitening converges to a suboptimal linear filter. We showed that under certain additional modeling assumptions, whitening improves the estimation of the signal's principal components, and achieves the optimal rate for subspace estimation when $r = 1$. We showed that the operator norm SNR of the observations increases after whitening. We also extended the analysis on out-of-sample prediction found in [19] to the whitening procedure.

There are a number of interesting directions for future research. First, we plan to revisit previous works that have employed similar shrinkage-plus-whitening procedures, but with the optimal shrinkers we have derived. It is of interest to determine how much of an improvement is achieved with the more principled choice we have presented.

As our current analysis is restricted to the setting of Gaussian noise, in future work we will try to extend the analysis to more general noise matrices. This likely requires a deeper understanding of the distribution of the projection of the empirical singular vectors onto the orthogonal complement of the population signal vectors in the setting of non-Gaussian noise.

While we have shown that whitening can improve subspace estimation generically, and matches the error rate (up to a constant) of [58], it is not clear if whitening is the optimal transformation for subspace estimation. In a different but closely related model to the one we have studied, where the noise variances differ across observations rather than across coordinates, it was found that certain weighting schemes can outperform whitening [28]. We note too that if the matrix Σ_ε is ill-conditioned, numerical instabilities may result from the whitening and unwhitening operations.

Finally, it is also of interest to better understand the procedure when the noise covariance Σ_ε is not known exactly, but must be estimated. This is a subject currently under investigation.

Acknowledgements

The authors would like to thank Edgar Dobriban, Matan Gavish, and Amit Singer for stimulating discussions related to this work. William Leeb acknowledges support from the Simons Foundation Collaboration

on Algorithms and Geometry, the NSF BIGDATA program IIS 1837992, and BSF award 2018230. Elad Romanov acknowledges support from Israeli Science Foundation grant number 1523/16.

References

[1] Joakim Andén and Amit Singer. Factor analysis for spectral estimation. In *Sampling Theory and Applications (SampTA), 2017 International Conference on*, pages 169–173. IEEE, 2017.

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

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

[7] 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.

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

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

[10] 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.

[11] Tamir Bendory, Alberto Bartesaghi, and Amit Singer. Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities. *IEEE Signal Processing Magazine*, 37(2):58–76, 2020.

[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] Peter J Bickel, Elizaveta Levina, et al. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[14] Timothy A. Brown. *Confirmatory factor analysis for applied research*. Guilford Publications, 2014.

[15] Andreas Buja and Nermin Eyuboglu. Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4):509–540, 1992.

[16] T. Tony Cai, Zhao Ren, and Harrison H. Zhou. Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probability Theory and Related Fields*, 156(1-2):101–143, 2013.

[17] 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.

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

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

[20] Edgar Dobriban and Art B. Owen. Deterministic parallel analysis: an improved method for selecting factors and principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):163–183, 2019.

[21] David L. Donoho, Matan Gavish, and Iain M Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of Statistics*, 46(6), 2018.

[22] David L. Donoho and Behrooz Ghorbani. Optimal covariance estimation for condition number loss in the spiked model. *arXiv preprint arXiv:1810.07403*, 2018.

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

[24] 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.

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

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

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

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

[29] John L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

[30] J. Edward Jackson. *A User's Guide to Principal Components*, volume 587. John Wiley & Sons, 2005.

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

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

[33] Julie Josse and François Husson. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6):1869–1879, 2012.

[34] Hamid Krim and Mats Viberg. Two decades of array signal processing research: the parametric approach. *IEEE signal processing magazine*, 13(4):67–94, 1996.

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

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

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

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

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

[40] Torben E. Lund, Kristoffer H. Madsen, Karam Sidaros, Wen-Lin Luo, and Thomas E. Nichols. Non-white noise in fMRI: does modelling have an impact? *Neuroimage*, 29(1):54–66, 2006.

[41] D. J. C. MacKay. Deconvolution. In *Information Theory, Inference and Learning Algorithms*, pages 550–551. Cambridge University Press, Camridge, UK, 2004.

[42] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[43] 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.

[44] 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.

[45] 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.

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

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

[48] 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.

[49] Jack W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55:331–339, 1995.

[50] Jack W. Silverstein and Zhidong Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54:175–192, 1995.

[51] Charles M. Stein. Some problems in multivariate analysis. Technical report, Stanford University Statistics Department, 1956.

[52] Charles M. Stein. Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics*, 74(5), 1986.

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

[54] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3):37–52, 1987.

[55] Luc Wouters, Hinrich W. Göhlmann, Luc Bijnens, Stefan U. Kass, Geert Molenberghs, and Paul J. Lewi. Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics*, 59(4):1131–1139, 2003.

[56] Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286, 2010.

[57] H. Henry Yue and Masayuki Tomoyasu. Weighted principal component analysis and its applications to improve FDC performance. In *Decision and Control, 43rd IEEE Conference on*, volume 4, pages 4262–4267. IEEE, 2004.

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

A Proof from Section 3

A.1 Proof of Theorem 3.1

We begin by recalling the result that describes the asymptotics of the spiked model with white noise. This result can be found in [46, 10]. We immediately obtain parts 1 and 4 of Theorem 3.1.

Theorem A.1. *If $p/n \to \gamma > 0$ as $n \to \infty$, the k^{th} largest singular value of Y^{w} converges almost surely to*

$$\sigma_k^{\mathrm{w}} = \begin{cases} \sqrt{(\ell_k^{\mathrm{w}} + 1)\left(1 + \frac{\gamma}{\ell_k^{\mathrm{w}}}\right)} & \text{if } \ell_k^{\mathrm{w}} > \sqrt{\gamma} \\ 1 + \sqrt{\gamma} & \text{otherwise} \end{cases}. \tag{113}$$

Furthermore, for $1 \le j, k \le r$:

$$\langle u_j^{\mathrm{w}}, \hat{u}_k^{\mathrm{w}} \rangle^2 \to \begin{cases} (c_k^{\mathrm{w}})^2, & \text{if } j = k \text{ and } \ell_k^{\mathrm{w}} > \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases} \tag{114}$$

and

$$\langle v_j^{\mathrm{w}}, \hat{v}_k^{\mathrm{w}} \rangle^2 \to \begin{cases} (\tilde{c}_k^{\mathrm{w}})^2, & \text{if } j = k \text{ and } \ell_k^{\mathrm{w}} > \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases} \tag{115}$$

where the limits hold almost surely as $p, n \to \infty$ and $p/n \to \gamma$.

We now turn to proving parts 2 and 3. Let $\mathcal{W} = \mathrm{span}\{u_1^{\mathrm{w}}, \ldots, u_r^{\mathrm{w}}\}$ be the r-dimensional subspace spanned by the whitened population PCs (the left singular vectors of X^{w}). For fixed n and p, write

$$\hat{u}_k^{\mathrm{w}} = \bar{c}_k^{\mathrm{w}} w_k^{\mathrm{w}} + \bar{s}_k^{\mathrm{w}} \tilde{u}_k^{\mathrm{w}}, \tag{116}$$

where $(\overline{c}_k^k)^2 + (\overline{s}_k^k)^2 = 1$, and $w_k^{\mathrm{w}} \in \mathcal{W}$, and $\tilde{u}_k^{\mathrm{w}} \perp \mathcal{W}$ are unit vectors. Because the whitened noise matrix is Gaussian, and hence orthogonally invariant, the vector $\tilde{u}_k^{\mathrm{w}}$ is uniformly distributed over the unit sphere in $\mathcal{W}^\perp$. Since the dimension of $\mathcal{W}$ is fixed, it follows immediately from Proposition 6.2 in [9] that for any unit vector $x \in \mathbb{R}^p$ independent of $\tilde{u}_k^{\mathrm{w}}$, the following limits hold almost surely:

$$\lim_{p\to\infty} (\tilde{u}_k^{\mathrm{w}})^\top x = 0, \tag{117}$$

and

$$\lim_{p\to\infty} \left\{ (\tilde{u}_k^{\mathrm{w}})^\top A \tilde{u}_k^{\mathrm{w}} - \mu_a \right\} = \lim_{p\to\infty} \left\{ (\tilde{u}_k^{\mathrm{w}})^\top A \tilde{u}_k^{\mathrm{w}} - \frac{1}{p}\mathrm{tr}(A) \right\} = 0. \tag{118}$$

From Theorem A.1, we know $|(w_k^{\mathrm{w}})^\top u_k^{\mathrm{w}}| \to 1$ and $(w_k^{\mathrm{w}})^\top u_j^{\mathrm{w}} \to 0$ almost surely when $j \neq k$; and $\overline{c}_k^{\mathrm{w}} \to c_k^{\mathrm{w}}$ almost surely. Consequently, we can write

$$\hat{u}_k^{\mathrm{w}} = c_k^{\mathrm{w}} u_k^{\mathrm{w}} + s_k^{\mathrm{w}} \tilde{u}_k^{\mathrm{w}} + \psi \tag{119}$$

where $\|\psi\| \to 0$ almost surely as $p \to \infty$. The inner product of ψ with any vectors of bounded norm will therefore also converge to 0. As a short-hand, we will write:

$$\hat{u}_k^{\mathrm{w}} \sim c_k^{\mathrm{w}} u_k^{\mathrm{w}} + s_k^{\mathrm{w}} \tilde{u}_k^{\mathrm{w}}, \tag{120}$$

to indicate that the norm of the difference of the two sides converges to 0 almost surely as $p \to \infty$.

From (120) we have:

$$A^{1/2}\hat{u}_k^{\mathrm{w}} \sim c_k^{\mathrm{w}} A^{1/2} u_k^{\mathrm{w}} + s_k^{\mathrm{w}} A^{1/2} \tilde{u}_k^{\mathrm{w}}. \tag{121}$$

Taking the squared norm of each side of (121) and using (117) and (118), we obtain:

$$\|A^{1/2}\hat{u}_k^{\mathrm{w}}\|^2 \sim (c_k^{\mathrm{w}})^2 \|A^{1/2} u_k^{\mathrm{w}}\|^2 + (s_k^{\mathrm{w}})^2 \|A^{1/2}\tilde{u}_k^{\mathrm{w}}\|^2 \sim \frac{(c_k^{\mathrm{w}})^2}{\tau_k^a} + (s_k^{\mathrm{w}})^2 \mu_a, \tag{122}$$

This completes the proof of part 2.

Part 3 is proved in the same fashion. Taking inner products with each side of (121), and using (117), we get

$$\langle A u_k^{\mathrm{w}}, \hat{u}_k^{\mathrm{w}} \rangle = \langle A^{1/2} u_k^{\mathrm{w}}, A^{1/2} \hat{u}_k^{\mathrm{w}} \rangle \sim \frac{c_k^{\mathrm{w}}}{\tau_k^a} + s_k^{\mathrm{w}} ((u_k^{\mathrm{w}})^\top A \tilde{u}_k^{\mathrm{w}}) \sim \frac{c_k^{\mathrm{w}}}{\tau_k^a}, \tag{123}$$

which is the desired result.

A.2 Proof of Theorem 3.2

We can decompose X as:

$$X = \sum_{k=1}^{r} \ell_k^{1/2} u_k z_k^\top / \sqrt{n}. \tag{124}$$

Since z_{jk} and $z_{jk'}$ are uncorrelated when $k \neq k'$, and both have variance 1, the vectors $z_k/\sqrt{n}$ are almost surely asymptotically orthonormal, i.e., $\lim_{n\to\infty} |\langle z_k, z_{k'} \rangle|/n = \delta_{kk'}$. It follows that the $z_k/\sqrt{n}$ are asymptotically equivalent to the right singular vectors v_k of X, that is,

$$\lim_{n\to\infty} \langle v_k, z_k \rangle^2 / n = 1 \tag{125}$$

almost surely; and the singular values of X are asymptotically equal to the $\ell_k^{1/2}$. That is, we can write:

$$X \sim \sum_{k=1}^{r} \ell_k^{1/2} u_k v_k^\top, \tag{126}$$

where $C \sim D$ indicates $\|C - D\|_{\mathrm{op}} \to 0$ as $p, n \to \infty$. Similarly, we can also write

$$X^{\mathrm{w}} \sim \sum_{k=1}^{r} (\ell_k^{\mathrm{w}})^{1/2} u_k^{\mathrm{w}} (v_k^{\mathrm{w}})^\top. \tag{127}$$

We can also decompose X^{w} by applying W to X:

$$X^{\mathrm{w}} = WX \sim \sum_{k=1}^{r} \ell_k^{1/2} W u_k v_k^\top = \sum_{k=1}^{r} (\ell_k \|W u_k\|^2)^{1/2} \overline{u}_k^{\mathrm{w}} v_k^\top. \tag{128}$$

The condition (19) immediately implies that $\overline{u}_j^{\mathrm{w}}$ and $\overline{u}_k^{\mathrm{w}}$ are asymptotically orthogonal whenever $j \neq k$. Comparing (127) and (128) then shows that almost surely,

$$\ell_k^{\mathrm{w}} \sim \ell_k \|W u_k\|^2, \tag{129}$$

$$\lim_{p\to\infty} \langle u_k^{\mathrm{w}}, \overline{u}_k^{\mathrm{w}} \rangle^2 = 1, \tag{130}$$

and

$$\lim_{n\to\infty} \langle v_k, v_k^{\mathrm{w}} \rangle^2 = 1. \tag{131}$$

From (130), $\langle u_k, \overline{u}_k \rangle^2 \sim 1$ follows immediately.

To prove the asymptotically equivalent formula for τ_k, we use (130):

$$\tau_k \sim \|W^{-1} u_k^{\mathrm{w}}\|^{-2} \sim \|W^{-1} \overline{u}_k^{\mathrm{w}}\|^{-2} \sim \|W^{-1} W u_k\|^{-2} \|W u_k\|^2 = \|W u_k\|^2. \tag{132}$$

To prove the formulas for the asymptotic cosine between u_j and $\hat{u}_k$ we take $A = W^{-1}$ in Theorem 3.1. When $j \neq k$, we have the formula

$$\hat{u}_k^{\mathrm{w}} \sim c_k^{\mathrm{w}} u_k^{\mathrm{w}} + s_k^{\mathrm{w}} \tilde{u}_k^{\mathrm{w}} \sim c_k^{\mathrm{w}} \frac{W u_k}{\sqrt{\tau_k}} + s_k^{\mathrm{w}} \tilde{u}_k^{\mathrm{w}} \tag{133}$$

and consequently

$$W^{-1} \hat{u}_k^{\mathrm{w}} \sim c_k^{\mathrm{w}} \frac{u_k}{\sqrt{\tau_k}} + s_k^{\mathrm{w}} W^{-1} \tilde{u}_k^{\mathrm{w}}. \tag{134}$$

We take inner products of each side with u_j. From the orthogonality of u_k and u_j, and using (117), we have:

$$\langle u_j, W^{-1} \hat{u}_k^{\mathrm{w}} \rangle \sim 0, \tag{135}$$

and consequently $\langle u_j, \hat{u}_k \rangle \sim 0$. When $j = k$, the formula for $\langle u_j, \hat{u}_k \rangle$ follows from Theorem 3.1.

Finally, we show that $\hat{u}_j$ and $\hat{u}_k$ are asymptotically orthogonal when $j \neq k$. We use the following lemma.

Lemma A.2. *Suppose $X = \sum_{k=1}^{r} \ell_k^{1/2} w_k v_k^\top$ is a p-by-n rank r matrix, and G is a matrix with iid Gaussian entries $g_{ij} \sim N(0, 1/n)$. Let $\hat{w}_1, \ldots, \hat{w}_m$ be the left singular vectors of $Y = X + G$, where $m = \min(p, n)$, and write*

$$\hat{w}_k \sim c_k w_k + s_k \tilde{w}_k \tag{136}$$

where $\tilde{w}_k$ is orthogonal to $w_1, \ldots, w_r$. Then for any sequence of matrices $A = A_p$ with bounded operator norms and any $1 \leq j \neq k \leq r$,

$$\lim_{p\to\infty} \tilde{w}_j^\top A \tilde{w}_k = 0 \tag{137}$$

almost surely.

Proof. First, we prove the cases where $A = I_p$; that is, we show $\tilde{w}_j$ and $\tilde{w}_k$ are asymptotically orthogonal whenever $1 \leq j \neq k \leq r$. Indeed, we have

$$\begin{aligned} s_j s_k \langle \tilde{w}_j, \tilde{w}_k \rangle &\sim \langle \hat{w}_j, \hat{w}_k \rangle + c_j c_k \langle w_j, w_k \rangle - c_j \langle w_j, \hat{w}_k \rangle - c_k \langle w_k, \hat{w}_j \rangle \\ &= -c_j \langle w_j, \hat{w}_k \rangle - c_k \langle w_k, \hat{w}_j \rangle. \end{aligned} \tag{138}$$

Since $\tilde{w}_j$ and $\tilde{w}_k$ are uniformly distributed on the subspace orthogonal to $w_1, \ldots, w_r$, the inner products $\langle w_j, \hat{w}_k \rangle$ and $\langle w_k, \hat{w}_j \rangle$ both converge to 0 almost surely as $p \to \infty$, proving the claim.

For general A, we note that the joint distribution of $\tilde{w}_j$ and $\tilde{w}_k$ is invariant to orthogonal transformations which leave fixed the r-dimensional subspace $\mathrm{span}\{w_1, \ldots, w_r\}$. The result then follows from Proposition 6.2 in [9], which implies that

$$\tilde{w}_j^\top A \tilde{w}_k^\top \sim \frac{1}{p} \mathrm{tr}(A) \tilde{w}_j^\top \tilde{w}_k \sim 0, \tag{139}$$

where we have used the asymptotic orthogonality of $\tilde{w}_j$ and $\tilde{w}_k$. □

Since $u_k \sim \overline{u}_k$ and u_j and u_k are orthogonal, taking inner products of each side of (133) with $W^{-1}\hat{u}_j^{\mathrm{w}}$ we get:

$$\langle W^{-1}\hat{u}_j^{\mathrm{w}}, W^{-1}\hat{u}_k^{\mathrm{w}}\rangle \sim s_j^{\mathrm{w}} s_k^{\mathrm{w}} \langle W^{-1}\tilde{u}_j^{\mathrm{w}}, W^{-1}\tilde{u}_k^{\mathrm{w}}\rangle = s_j^{\mathrm{w}} s_k^{\mathrm{w}} (\tilde{u}_j^{\mathrm{w}})^\top \Sigma_\varepsilon \tilde{u}_k^{\mathrm{w}}. \tag{140}$$

The result now follows from Lemma A.2.

B Proofs from Section 5

First, we establish the consistency of covariance estimation in the $\gamma = 0$ regime:

Proposition B.1. *If $p_n/n \to 0$ as $n \to \infty$, and the subgaussian norm of QY_j can be bounded by C independently of the dimension p, then the sample covariance matrix of $QY_1, \dots, QY_n$ converges to the population covariance $Q\Sigma_y Q$ in operator norm.*

Proof. We first quote the following result, stated as Corollary 5.50 in [53]:

Lemma B.2. *Let $Y_1, \dots, Y_n$ be iid mean zero subgaussian random vectors in $\mathbb{R}^p$ with covariance matrix Σ_y, and let $\epsilon \in (0,1)$ and $t \geq 1$. Then with probability at least $1 - 2\exp(-t^2 p)$,*

$$\text{If } n \geq C(t/\epsilon)^2 p, \text{ then } \|\hat{\Sigma}_y - \Sigma_y\| \leq \epsilon, \tag{141}$$

where $\hat{\Sigma}_y = \sum_{j=1}^n Y_j Y_j^\top / n$ is the sample covariance, and C is a constant.

We also state the well-known consequence of the Borel-Cantelli Lemma:

Lemma B.3. *Let $A_1, A_2, \dots$ be a sequence of random numbers, and let $\epsilon > 0$. Define:*

$$\mathcal{A}_n(\epsilon) = \{|A_n| > \epsilon\}. \tag{142}$$

If for every choice of $\epsilon > 0$ we have

$$\sum_{n=1}^{\infty} \mathbb{P}(\mathcal{A}_n(\epsilon)) < \infty, \tag{143}$$

then $A_n \to 0$ almost surely.

Now take $t = \epsilon\sqrt{n/Cp}$; then $n \geq C(t/\epsilon)^2 p$, and $t \geq 1$ for n sufficiently large. Consequently,

$$\mathbb{P}(\|\hat{\Sigma}_y - \Sigma_y\| > \epsilon) \leq 2\exp(-t^2 p) = 2\exp(-n\epsilon^2/C), \tag{144}$$

and so the series $\sum_{n\geq 1} \mathbb{P}(\|\hat{\Sigma}_y - \Sigma_y\| > \epsilon)$ converges, meaning $\|\hat{\Sigma}_y - \Sigma_y\| \to 0$ almost surely as $n \to \infty$.

We now need to check that the subgaussian norm of $Y_j = X_j + \varepsilon_j$ from the spiked model is bounded independently of the dimension p. But this is easy if the distribution of variances of ε_j is bounded, using, for example, Lemma 5.24 of [53]. □

An immediate corollary of Proposition B.1 is that the sample eigenvectors of $\hat{\Sigma}_y^q = Q\hat{\Sigma}_y Q$ are consistent estimators of the eigenvectors of $\Sigma_y^q = Q\Sigma_y Q$.

Corollary B.4. *Let $\Sigma_y^q = Q\Sigma_y Q$ be the population covariance of the random vector $Y_j^q = QY_j$, and let $\hat{\Sigma}_y^q = Q\hat{\Sigma}_y Q$ be the sample covariance of $Y_1^q, \dots, Y_n^q$. Let $u_1^q, \dots, u_r^q$ denote the top r eigenvectors of $\Sigma_y^q = Q\Sigma_y Q$, and $\hat{u}_1^1, \dots, \hat{u}_r^q$ the top r eigenvectors of $\hat{\Sigma}_y^q$.*

Then for $1 \leq k \leq r$,

$$\lim_{n\to\infty} |\langle \hat{u}_k^q, u_k^q\rangle| = 1, \tag{145}$$

where the limit holds almost surely as $n \to \infty$ and $p/n \to 0$.

We now turn to the proof of Theorem 5.2. First, we derive an expression for the BLP $\hat{X}_j^{\text{opt}}$. We have:

$$\begin{aligned}\hat{X}_j^{\text{opt}} &= \Sigma_x \left(\Sigma_x + \Sigma_\varepsilon\right)^{-1} Y_j \\ &= W^{-1} W\Sigma_x W \left(W\Sigma_x W + I\right)^{-1} W Y_j \\ &= W^{-1} \sum_{k=1}^{r} \frac{\ell_k^{\text{w}}}{\ell_k^{\text{w}} + 1} \langle W Y_j, u_k^{\text{w}} \rangle u_k^{\text{w}} \\ &= \sum_{k=1}^{r} \eta_k^{\text{opt}} \langle W Y_j, u_k^{\text{w}} \rangle W^{-1} u_k^{\text{w}}, \end{aligned} \tag{146}$$

where $W\Sigma_x W = \sum_{k=1}^r \ell_k^{\text{w}} u_k^{\text{w}} (u_k^{\text{w}})^\top$, and $\eta_k^{\text{opt}} = \ell_k^{\text{w}}/(\ell_k^{\text{w}} + 1)$.

Now, for any $s_1, \ldots, s_r$ satisfying

$$\lim_{\gamma \to 0} \frac{s_k}{\sigma_k^{\text{w}}} = \frac{\ell_k^{\text{w}}}{\ell_k^{\text{w}} + 1}. \tag{147}$$

we define the predictor $\hat{X}'$:

$$\hat{X}' = \sum_{k=1}^{r} s_k W^{-1} \hat{u}_k^{\text{w}} (\hat{v}_k^{\text{w}})^\top. \tag{148}$$

Following the same reasoning as in the proof of Lemma 5.4, we can write each column $\hat{X}'_j$ of $\sqrt{n}\hat{X}'$ as follows:

$$\hat{X}'_j = \sum_{k=1}^{r} (s_k/\sigma_k^{\text{w}}) \langle Y_j^{\text{w}}, \hat{u}_k^{\text{w}} \rangle W^{-1} \hat{u}_k^{\text{w}}. \tag{149}$$

Theorem 5.2 now follows from condition (147), formula (146), and Corollary B.4. Theorem 5.1 follows immediately, after observing that $\hat{X}$ has the same form as $\hat{X}'$ with $s_k = t_k$, and

$$\lim_{\gamma \to 0} \frac{t_k}{\sigma_k^{\text{w}}} = \lim_{\gamma \to 0} \frac{(\ell_k^{\text{w}})^{1/2} c_k^{\text{w}} \tilde{c}_k}{(c_k^{\text{w}})^2 + (s_k^{\text{w}})^2 \mu_\varepsilon \tau_k} \frac{1}{\sqrt{\ell_k^{\text{w}} + 1}} = \lim_{\gamma \to 0} \frac{(\ell_k^{\text{w}})^{1/2} \tilde{c}_k}{\sqrt{\ell_k^{\text{w}} + 1}} = \frac{\ell_k^{\text{w}}}{\ell_k^{\text{w}} + 1}. \tag{150}$$

Finally, we prove Theorem 5.3. By definition,

$$\hat{Y}_{Q,j} = \sum_{k=1}^{r} (s_k^q/\sigma_k^q) \langle Y_j^q, \hat{u}_k^q \rangle Q^{-1} \hat{u}_k^q \tag{151}$$

and

$$\hat{Y}_{Q,j}^{\text{lin}} = \sum_{k=1}^{r} \eta_k^q \langle Y_j^q, u_k^q \rangle Q^{-1} u_k^q. \tag{152}$$

The values s_k^q and η_k^q are each assumed to minimize the mean-squared error for their respective expressions. Consequently, since Corollary B.4 states that $\hat{u}_k^q \sim u_k^q$, we establish (88); (89) follows immediately from (146).

C Proof of Theorem 6.1

C.1 The optimal coefficients for in-sample prediction

Before deriving the optimal out-of-sample coefficients η_k^{o}, we will first derive the optimal in-sample coefficients η_k. That is, we will rewrite the optimal shrinkage with noise whitening in the form (92).

From Lemma 5.4, the in-sample coefficients η_k are the ratios of the optimal singular values t_k derived in Section 4.1 and the observed singular values of Y^{w}, denoted $\sigma_1^{\text{w}}, \ldots, \sigma_r^{\text{w}}$. From Theorem A.1, we know that

$$\sigma_k^{\text{w}} = \sqrt{(\ell_k^{\text{w}} + 1)\left(1 + \frac{\gamma}{\ell_k^{\text{w}}}\right)}, \tag{153}$$

and from Section 4.1 we know that

$$t_k = \frac{(\ell_k^{\mathrm{w}})^{1/2} c_k^{\mathrm{w}} \tilde{c}_k}{(c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \mu_\varepsilon \tau_k} = \alpha_k (\ell_k^{\mathrm{w}})^{1/2} c_k^{\mathrm{w}} \tilde{c}_k, \tag{154}$$

where $\alpha_k = \left((c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \mu_\varepsilon \tau_k\right)^{-1}$. Taking the ratio, and using formulas (39) and (40) for c_k^{w} and $\tilde{c}_k$, we obtain:

$$\eta_k = \frac{t_k}{\sigma_k^{\mathrm{w}}} = \alpha_k \frac{(\ell_k^{\mathrm{w}})^{1/2} c_k^{\mathrm{w}} \tilde{c}_k}{\sqrt{(\ell_k^{\mathrm{w}} + 1)\left(1 + \frac{\gamma}{\ell_k^{\mathrm{w}}}\right)}} = \alpha_k \frac{\ell_k^{\mathrm{w}} (c_k^{\mathrm{w}})^2}{\sqrt{(\ell_k^{\mathrm{w}} + 1)(\ell_k^{\mathrm{w}} + \gamma)}} \sqrt{\frac{(\ell_k^{\mathrm{w}})^2 + \gamma \ell_k^{\mathrm{w}}}{(\ell_k^{\mathrm{w}})^2 + \ell_k^{\mathrm{w}}}} = \alpha_k \frac{\ell_k^{\mathrm{w}} (c_k^{\mathrm{w}})^2}{\ell_k^{\mathrm{w}} + 1}. \tag{155}$$

That is, we have found the optimal in-sample coefficients to be:

$$\eta_k = \frac{1}{(c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \mu_\varepsilon \tau_k} \cdot \frac{\ell_k^{\mathrm{w}} (c_k^{\mathrm{w}})^2}{\ell_k^{\mathrm{w}} + 1}. \tag{156}$$

C.2 The optimal coefficients for out-of-sample prediction

In this section, we will derive the optimal out-of-sample coefficients η_k^{o}. We have a predictor of the form

$$\hat{X}_0 = \sum_{k=1}^{r} \eta_k^{\mathrm{o}} \langle W Y_0, \hat{u}_k^{\mathrm{w}} \rangle W^{-1} \hat{u}_k^{\mathrm{w}}, \tag{157}$$

where $\hat{u}_k^{\mathrm{w}}$ are the top left singular vectors of the in-sample observation matrix $Y^{\mathrm{w}} = W[Y_1, \ldots, Y_n]/\sqrt{n}$. We wish to choose the coefficients η_k^{o} that minimize the asymptotic mean squared error $\mathbb{E}\|X_0 - \hat{X}_0\|^2$. First, we can expand the MSE across the different principal components as follows:

$$\begin{aligned}\|X_0 - \hat{X}_0\|^2 =& \sum_{k=1}^{r} \|\ell_k^{1/2} z_{0k} u_k - \eta_k^{\mathrm{o}} \langle W Y_0, \hat{u}_k^{\mathrm{w}} \rangle W^{-1} \hat{u}_k^{\mathrm{w}}\|^2 \\ &+ \sum_{k \neq l} \langle \ell_k^{1/2} z_{0k} u_k - \eta_k^{\mathrm{o}} \langle W Y_0, \hat{u}_k^{\mathrm{w}} \rangle W^{-1} \hat{u}_k^{\mathrm{w}}, \ell_l^{1/2} z_{0l} u_l - \eta_l^{\mathrm{o}} \langle W Y_0, \hat{u}_l^{\mathrm{w}} \rangle W^{-1} \hat{u}_l^{\mathrm{w}} \rangle. \end{aligned} \tag{158}$$

After taking expectations, the cross-terms vanish and we are left with:

$$\mathbb{E}\|X_0 - \hat{X}_0\|^2 = \sum_{k=1}^{r} \mathbb{E}\|\ell_k^{1/2} z_{0k} u_k - \eta_k^{\mathrm{o}} \langle W Y_0, \hat{u}_k^{\mathrm{w}} \rangle W^{-1} \hat{u}_k^{\mathrm{w}}\|^2. \tag{159}$$

Since the sum separates across the η_k^{o}, we can minimize each summand individually. We write:

$$\begin{aligned}&\mathbb{E}\|\ell_k^{1/2} z_{0k} u_k - \eta_k^{\mathrm{o}} \langle W Y_0, \hat{u}_k^{\mathrm{w}} \rangle W^{-1} \hat{u}_k^{\mathrm{w}}\|^2 \\ =& \ell_k + (\eta_k^{\mathrm{o}})^2 \mathbb{E}\left[\langle W Y_0, \hat{u}_k^{\mathrm{w}} \rangle^2 \|W^{-1} \hat{u}_k^{\mathrm{w}}\|^2\right] - 2\ell_k^{1/2} \eta_k^{\mathrm{o}} \mathbb{E}\left[z_{0k} \langle W Y_0, \hat{u}_k^{\mathrm{w}} \rangle \langle u_k, W^{-1} \hat{u}_k^{\mathrm{w}} \rangle\right]. \end{aligned} \tag{160}$$

We first deal with the quadratic coefficient in η:

$$\begin{aligned}\langle W Y_0, \hat{u}_k^{\mathrm{w}} \rangle^2 \|W^{-1} \hat{u}_k^{\mathrm{w}}\|^2 &= \langle W X_0 + W\varepsilon_0, \hat{u}_k^{\mathrm{w}} \rangle^2 \|W^{-1} \hat{u}_k^{\mathrm{w}}\|^2 \\ &= \left(\langle W X_0, \hat{u}_k^{\mathrm{w}} \rangle^2 + \langle W\varepsilon_0, \hat{u}_k^{\mathrm{w}} \rangle^2 + \langle W X_0, \hat{u}_k^{\mathrm{w}} \rangle \langle W\varepsilon_0, \hat{u}_k^{\mathrm{w}} \rangle\right) \|W^{-1} \hat{u}_k^{\mathrm{w}}\|^2, \end{aligned} \tag{161}$$

and taking expectations, we get:

$$\mathbb{E}\left[\langle W Y_0, \hat{u}_k^{\mathrm{w}} \rangle^2 \|W^{-1} \hat{u}_k^{\mathrm{w}}\|^2\right] \sim \left(\mathbb{E}\left[\langle W X_0, \hat{u}_k^{\mathrm{w}} \rangle^2\right] + 1\right) \|W^{-1} \hat{u}_k^{\mathrm{w}}\|^2 \sim \left(\ell_k^{\mathrm{w}} (c_k^{\mathrm{w}})^2 + 1\right) \left(\frac{(c_k^{\mathrm{w}})^2}{\tau_k} + (s_k^{\mathrm{w}})^2 \mu_\varepsilon\right). \tag{162}$$

Now we turn to the linear coefficient in η:

$$\begin{aligned}\ell_k^{1/2} \mathbb{E}\left[z_{0k} \langle W Y_0, \hat{u}_k^{\mathrm{w}} \rangle \langle u_k, W^{-1} \hat{u}_k^{\mathrm{w}} \rangle\right] &= \ell_k^{1/2} \mathbb{E}\left[z_{0k} \left((\ell_k^{\mathrm{w}})^{1/2} z_{0k} c_k^{\mathrm{w}} + \langle W\varepsilon_0, \hat{u}_k^{\mathrm{w}} \rangle\right) \langle u_k, W^{-1} \hat{u}_k^{\mathrm{w}} \rangle\right] \\ &= \frac{\ell_k^{\mathrm{w}} c_k^{\mathrm{w}} \mathbb{E}\left[\langle u_k, W^{-1} \hat{u}_k^{\mathrm{w}} \rangle\right]}{\|W u_k\|} \\ &\sim \ell_k^{\mathrm{w}} (c_k^{\mathrm{w}})^2 \frac{1}{\tau_k}. \end{aligned} \tag{163}$$

Minimizing the quadratic for η_k°, we get:

$$\begin{aligned}\eta_k^{\circ} &= \left(\ell_k^{\mathrm{w}}(c_k^{\mathrm{w}})^2\frac{1}{\tau_k}\right) \Big/ \left(\left(\ell_k^{\mathrm{w}}(c_k^{\mathrm{w}})^2+1\right)\left(\frac{(c_k^{\mathrm{w}})^2}{\tau_k}+(s_k^{\mathrm{w}})^2\mu_\varepsilon\right)\right) \\ &= \frac{1}{(c_k^{\mathrm{w}})^2+(s_k^{\mathrm{w}})^2\mu_\varepsilon\tau_k}\cdot\frac{\ell_k^{\mathrm{w}}(c_k^{\mathrm{w}})^2}{\ell_k^{\mathrm{w}}(c_k^{\mathrm{w}})^2+1}. \end{aligned} \tag{164}$$

C.3 Equality of the AMSEs

Evaluating the out-of-sample error at the optimal out-of-sample coefficients η_k°, we find the optimal out-of-sample AMSE (where $\alpha_k = \left((c_k^{\mathrm{w}})^2+(s_k^{\mathrm{w}})^2\mu_\varepsilon\tau_k\right)^{-1}$):

$$\mathrm{AMSE} = \sum_{k=1}^{r}\left(\ell_k - \frac{(\ell_k^{\mathrm{w}})^2(c_k^{\mathrm{w}})^4}{\ell_k^{\mathrm{w}}(c_k^{\mathrm{w}})^2+1}\frac{1}{\alpha_k\tau_k}\right) = \sum_{k=1}^{r}\left(\frac{\ell_k^{\mathrm{w}}}{\tau_k} - \frac{(\ell_k^{\mathrm{w}})^2(c_k^{\mathrm{w}})^4}{\ell_k^{\mathrm{w}}(c_k^{\mathrm{w}})^2+1}\frac{1}{\alpha_k\tau_k}\right). \tag{165}$$

The AMSE of the in-sample predictor is:

$$\sum_{k=1}^{r}\ell_k(1-(c_k\tilde{c}_k)^2) = \sum_{k=1}^{r}\frac{\ell_k^{\mathrm{w}}}{\tau_k}\left(1-\frac{(c_k^{\mathrm{w}}\tilde{c}_k^{\mathrm{w}})^2}{\alpha_k}\right) = \sum_{k=1}^{r}\left(\frac{\ell_k^{\mathrm{w}}}{\tau_k} - \frac{\ell_k^{\mathrm{w}}(c_k^{\mathrm{w}}\tilde{c}_k^{\mathrm{w}})^2}{\alpha_k\tau_k}\right) \tag{166}$$

To show equality, we therefore need to show:

$$\ell_k^{\mathrm{w}}(c_k^{\mathrm{w}}\tilde{c}_k^{\mathrm{w}})^2 = \frac{(\ell_k^{\mathrm{w}})^2(c_k^{\mathrm{w}})^4}{\ell_k^{\mathrm{w}}(c_k^{\mathrm{w}})^2+1}. \tag{167}$$

But this follows from the equality of in-sample and out-of-sample AMSEs for the standard spiked model with isotropic noise, established in [19].

D Proofs from Section 7

D.1 Proof of Proposition 7.1

From Corollary B.4, $\hat{u}_k^{\mathrm{w}} \sim u_k^{\mathrm{w}}$, $1 \le k \le r$, in the sense that the angle between the vectors converges to 0. Consequently

$$\lim_{n\to 0}\Theta(\mathcal{U}^{\mathrm{w}},\hat{\mathcal{U}}^{\mathrm{w}}) = 0, \tag{168}$$

where $\mathcal{U}^{\mathrm{w}} = \mathrm{span}\{u_1^{\mathrm{w}},\dots,u_r^{\mathrm{w}}\}$ and $\hat{\mathcal{U}}^{\mathrm{w}} = \mathrm{span}\{\hat{u}_1^{\mathrm{w}},\dots,\hat{u}_r^{\mathrm{w}}\}$. Since W^{-1} has bounded operator norm and $\mathcal{U} = W^{-1}\mathcal{U}^{\mathrm{w}}$ and $\hat{\mathcal{U}} = W^{-1}\hat{\mathcal{U}}^{\mathrm{w}}$, the result follows immediately.

D.2 Proof of Theorem 7.2

Since the inner products between random unit vectors in $\mathbb{R}^p$ vanish as $p\to\infty$, we may assume that the u_k are drawn randomly with iid entries of variance $1/p$; the result will then follow for the orthonormalized vectors from the generic model. If $\Sigma_\varepsilon = \mathrm{diag}(\nu_1,\dots,\nu_p)$, then

$$\tau_k = \|\Sigma_\varepsilon^{-1/2}u_k\|^2 \sim \frac{1}{p}\sum_{j=1}^{p}\nu_j^{-1} = \tau. \tag{169}$$

We now define the n-by-p matrix $\tilde{Y} = Y^\top/\sqrt{\gamma}$, given by

$$\tilde{Y} = \sum_{k=1}^{r}\tilde{\ell}_k^{1/2}z_k u_k^\top + G^\top\Sigma_\varepsilon^{1/2}/\sqrt{p}, \tag{170}$$

where $\tilde{\ell}_k = \ell_k/\gamma$. Note that the noise matrix $G^\top\Sigma_\varepsilon^{1/2}$ has colored rows, not columns, and has been normalized by dividing by the square root of the number of its columns. Since the vectors u_k spanning the right singular subspace of $\tilde{Y}$ are assumed to be drawn uniformly from the unit sphere in $\mathbb{R}^p$, we may

apply Corollary 2 to Theorem 2 of [27] to the matrix $\tilde{Y}$. Defining $\tilde{\gamma} = 1/\gamma$ as the aspect ratio of $\tilde{Y}$, we have:

$$|\langle \hat{u}_k', u_k \rangle|^2 \leq \frac{1 - \tilde{\gamma}/(\tilde{\ell}_k/\mu_\varepsilon)^2}{1 + 1/(\tilde{\ell}_k/\mu_\varepsilon)} = \frac{1 - \gamma/(\ell_k^{\mathrm{w}}/\varphi)^2}{1 + \gamma/(\ell_k^{\mathrm{w}}/\varphi)} \equiv g(\ell_k^{\mathrm{w}}/\varphi), \tag{171}$$

where we have defined the function

$$g(\ell) = \frac{1 - \gamma/\ell^2}{1 + \gamma/\ell}. \tag{172}$$

On the other hand, the squared cosine $c_k^2 = |\langle \hat{u}_k, u_k \rangle|^2$ is equal to

$$c_k^2 = \frac{(c_k^{\mathrm{w}})^2}{(c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \varphi} = \frac{g(\ell_k^{\mathrm{w}})}{g(\ell_k^{\mathrm{w}}) + \varphi(1 - g(\ell_k^{\mathrm{w}}))}. \tag{173}$$

Our goal is to show that for all $\ell_k^{\mathrm{w}} > \sqrt{\gamma}$, and all $\varphi \geq 1$, that

$$g(\ell_k^{\mathrm{w}}/\varphi) \leq \frac{g(\ell_k^{\mathrm{w}})}{g(\ell_k^{\mathrm{w}}) + \varphi(1 - g(\ell_k^{\mathrm{w}}))}; \tag{174}$$

equivalently, we want to show that for all $\xi > 0$ and $\varphi > 1$,

$$g(\xi) \leq \frac{g(\xi\varphi)}{g(\xi\varphi) + \varphi(1 - g(\xi\varphi))}; \tag{175}$$

setting

$$G(\varphi) = \frac{g(\xi\varphi)}{g(\xi\varphi) + \varphi(1 - g(\xi\varphi))}, \tag{176}$$

this is equivalent to showing that $G(\varphi) \geq G(1)$ for all $\varphi \geq 1$. The derivative of G is equal to

$$\frac{d}{d\varphi} G(\varphi) = \frac{\gamma\xi^2\varphi^2 + 2\gamma^2\xi\varphi + \gamma^2}{(\xi^2\varphi^2 - \gamma + (\gamma\xi\varphi + \gamma)\varphi)^2} > 0, \tag{177}$$

which completes the first statement of the theorem.

The second statement concerning $\hat{v}_k$ is proved similarly. Again applying Corollary 2 to Theorem 2 of [27] to $\tilde{Y}$, we know that

$$|\langle \hat{v}_k', z_k \rangle|^2 \leq \frac{1 - \gamma/(\tilde{\ell}_k/\mu_\varepsilon)^2}{1 + \tilde{\gamma}/(\tilde{\ell}_k/\mu_\varepsilon)} = \frac{1 - \gamma/(\ell_k^{\mathrm{w}}/\varphi)^2}{1 + 1/(\ell_k^{\mathrm{w}}/\varphi)} \equiv h(\ell_k^{\mathrm{w}}/\varphi), \tag{178}$$

where we have defined the function

$$h(\ell) = \frac{1 - \gamma/\ell^2}{1 + 1/\ell}. \tag{179}$$

Since h is an increasing function of ℓ, and $|\langle \hat{v}_k, z_k \rangle|^2 = \tilde{c}_k^2 = h(\ell_k^{\mathrm{w}})$, the result follows.

D.3 Proof of Theorem 7.3

We begin the proof with some lemmas.

Lemma D.1. *Let $0 < B < 1$, and suppose q is the number of entries of u_k where $|u_{jk}| > B/\sqrt{p}$. Then*

$$q \geq p \cdot \frac{1 - B^2}{C^2 - B^2}, \tag{180}$$

where C is the incoherence parameter from (103)*.*

Proof. Let S_1 be the set of indices j on which $|u_{jk}| > B/\sqrt{p}$, and let S_2 be the set of indices j on which $|u_{jk}| \leq B/\sqrt{p}$. Because u_k is a unit vector, we then have

$$1 = \|u_k\|^2 = \sum_{j=1}^{p} u_{jk}^2 = \sum_{j \in S_1} u_{jk}^2 + \sum_{j \in S_2} u_{jk}^2 \leq (q/p)C^2 + (1 - q/p)B^2. \tag{181}$$

Rearranging, we find

$$\frac{q}{p} \geq \frac{1 - B^2}{C^2 - B^2}, \tag{182}$$

as claimed. □

Lemma D.2. *For each $1 \le k \le r$,*

$$\tau_k \ge \max\left\{\frac{\tilde{K}}{\mu_\varepsilon}, \frac{1}{\|\Sigma_\varepsilon\|_{\mathrm{op}}}\right\}, \tag{183}$$

where $\tilde{K}$ is a constant depending only on C from (103).

Proof. We will let $\nu_1, \dots, \nu_p$ denote the diagonal elements of Σ_ε. Take any number $0 < B < 1$, and let q be the number of indices where $|u_{jk}| > B/\sqrt{p}$. From Lemma D.1, $q/p \ge K_1$, a constant. Using the Cauchy-Schwarz inequality, we have:

$$\mu_\varepsilon \cdot \tau_k = \left(\sum_{j=1}^p \left(\frac{\sqrt{\nu_j}}{\sqrt{p}}\right)^2\right) \cdot \left(\sum_{j=1}^p \left(\frac{u_{jk}}{\sqrt{\nu_j}}\right)^2\right) \ge \left(\frac{1}{\sqrt{p}} \sum_{j=1}^p |u_{jk}|\right)^2 \ge \left(\frac{1}{\sqrt{p}}(K_1 p)\frac{B}{\sqrt{p}}\right)^2 = K_1^2 B^2. \tag{184}$$

This proves that $\tau_k \ge \tilde{K}/\mu_\varepsilon$.

Next, we observe that because $\sum_{j=1}^p u_{jk}^2 = 1$, we have

$$\tau_k = \sum_{j=1}^p \left(\frac{u_{jk}}{\sqrt{\nu_j}}\right)^2 \ge \min_{1 \le j \le p} \nu_j^{-1} = \left(\max_{1 \le j \le p} \nu_j\right)^{-1} = \frac{1}{\|\Sigma_\varepsilon\|_{\mathrm{op}}}, \tag{185}$$

completing the proof. □

We now turn to the proof of Theorem 7.3. We have

$$\|U_\perp^\top \hat{U}\|_{\mathrm{op}} = \|U_\perp U_\perp^\top \hat{U}\|_{\mathrm{op}} = \|\widetilde{U}\|_{\mathrm{op}} \tag{186}$$

where

$$\widetilde{U} = [\tilde{w}_1, \dots, \tilde{w}_r] \tag{187}$$

is the matrix whose columns are the projections $\tilde{w}_k$ of $\hat{u}_k$ onto the orthogonal complement of $\mathrm{span}\{u_1, \dots, u_r\}$. Then from Lemma A.2, we know that asymptotically $\tilde{w}_j \perp \tilde{w}_k$ if $j \ne k$; consequently,

$$\|\sin\Theta(\hat{U}, U)\|_{\mathrm{op}}^2 = \max_{1 \le k \le r} \|\tilde{w}_k\|^2 = \max_{1 \le k \le r} (1 - \langle \hat{u}_k, u_k \rangle^2) = \max_{1 \le k \le r} (1 - c_k^2). \tag{188}$$

From Theorem 3.2, for each $1 \le k \le r$, the squared sine between $\hat{u}_k$ and u_k is

$$1 - c_k^2 = 1 - \frac{(c_k^{\mathrm{w}})^2}{(c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \cdot \mu_\varepsilon \cdot \tau_k} = \frac{(s_k^{\mathrm{w}})^2 \cdot \mu_\varepsilon \cdot \tau_k}{(c_k^{\mathrm{w}})^2 + (s_k^{\mathrm{w}})^2 \cdot \mu_\varepsilon \cdot \tau_k}. \tag{189}$$

Since

$$(c_k^{\mathrm{w}})^2 = \frac{1 - \gamma/(\ell_k^{\mathrm{w}})^2}{1 + \gamma/\ell_k^{\mathrm{w}}} \tag{190}$$

and

$$(s_k^{\mathrm{w}})^2 = \frac{\gamma/\ell_k^{\mathrm{w}} + \gamma/(\ell_k^{\mathrm{w}})^2}{1 + \gamma/\ell_k^{\mathrm{w}}}, \tag{191}$$

we can simplify the expression by multiplying numerator and denominator by $(\ell_k^{\mathrm{w}})^2(1 + \gamma/\ell_k^{\mathrm{w}})$:

$$\begin{aligned} 1 - c_k^2 &= \frac{\gamma(\ell_k^{\mathrm{w}} + 1)\mu_\varepsilon \tau_k}{(\ell_k^{\mathrm{w}})^2 - \gamma + \gamma(\ell_k^{\mathrm{w}} + 1)\mu_\varepsilon \tau_k} \\ &= \frac{\gamma(\ell_k^{\mathrm{w}} + 1)\mu_\varepsilon \tau_k}{(\ell_k^{\mathrm{w}})^2} \cdot \frac{(\ell_k^{\mathrm{w}})^2}{(\ell_k^{\mathrm{w}})^2 - \gamma + \gamma(\ell_k^{\mathrm{w}} + 1)\mu_\varepsilon \tau_k}. \end{aligned} \tag{192}$$

Now, using Lemma D.2, there is a constant $0 < \tilde{K} < 1$ so that $\tau_k \mu_\varepsilon \ge \tilde{K}$. Consequently, since $\gamma < (\ell_k^{\mathrm{w}})^2$, we have:

$$\frac{(\ell_k^{\mathrm{w}})^2}{(\ell_k^{\mathrm{w}})^2 - \gamma + \gamma(\ell_k^{\mathrm{w}} + 1)\mu_\varepsilon \tau_k} \le \frac{(\ell_k^{\mathrm{w}})^2}{(\ell_k^{\mathrm{w}})^2 - (1 - \tilde{K})\gamma} \le \frac{(\ell_k^{\mathrm{w}})^2}{(\ell_k^{\mathrm{w}})^2 - (1 - \tilde{K})(\ell_k^{\mathrm{w}})^2} = \frac{1}{\tilde{K}}. \tag{193}$$

Combining equation (192) and inequality (193), the fact that $\ell_k^{\mathrm{w}} = \ell_k \cdot \tau_k$, and Lemma D.2, we obtain the bound:

$$\begin{aligned} 1 - c_k^2 &\leq \frac{1}{\tilde{K}} \left(\frac{\gamma(\ell_k^{\mathrm{w}} + 1)\mu_\varepsilon \tau_k}{(\ell_k^{\mathrm{w}})^2} \right) \\ &= \frac{1}{\tilde{K}} \left(\frac{\gamma \ell_k^{\mathrm{w}} \mu_\varepsilon \tau_k}{(\ell_k^{\mathrm{w}})^2} + \frac{\gamma \mu_\varepsilon \tau_k}{(\ell_k^{\mathrm{w}})^2} \right) \\ &= \frac{1}{\tilde{K}} \left(\frac{\gamma \mu_\varepsilon}{\ell_k} + \frac{\gamma \mu_\varepsilon}{\ell_k^2 \tau_k} \right) \\ &\leq \frac{1}{\tilde{K}} \left(\frac{\gamma \mu_\varepsilon}{\ell_k} + \frac{\gamma \mu_\varepsilon \|\Sigma_\varepsilon\|_{\mathrm{op}}}{\ell_k^2} \right). \end{aligned} \tag{194}$$

Taking the maximum over $1 \leq k \leq r$ proves the desired result.

D.4 Proof of Proposition 7.4

As in the proof of Theorem 7.2, since the inner products between random unit vectors in $\mathbb{R}^p$ vanish as $p \to \infty$, we may assume that the u_k are drawn randomly with iid entries of variance $1/p$; the result will then follow for the orthonormalized vectors from the generic model. We will use the fact that $\|\hat{\Sigma}_x\|_{\mathrm{op}} = \|X\|_{\mathrm{op}}^2$ and $\|\hat{\Sigma}_\varepsilon\|_{\mathrm{op}} = \|N\|_{\mathrm{op}}^2$. To show the increase in SNR after whitening, we will first derive a lower bound on the operator norm of the noise matrix N alone. Recall that $N = \Sigma_\varepsilon^{1/2} G$, where g_{ij} are iid $N(0, 1/n)$.

Take unit vectors c and d so that $Gd = \|G\|_{\mathrm{op}} c$. Then we have

$$\|N\|_{\mathrm{op}}^2 \geq \|\Sigma_\varepsilon^{1/2} G d\|^2 = \|G\|_{\mathrm{op}}^2 \|\Sigma_\varepsilon^{1/2} c\|^2 \tag{195}$$

Since the distribution of G is orthogonally invariant, the distribution of c is uniform over the unit sphere in $\mathbb{R}^n$. Consequently, $\|\Sigma_\varepsilon^{1/2} c\|^2 \sim \mathrm{tr}(\Sigma_\varepsilon)/p \sim \mu_\varepsilon$. Therefore,

$$\|N\|_{\mathrm{op}}^2 \gtrsim \mu_\varepsilon \|G\|_{\mathrm{op}}^2, \tag{196}$$

where "$\gtrsim$" indicates that the inequality holds almost surely in the large p, large n limit.

Next, from the assumption that the u_k are uniformly random, the parameters τ_k are all asymptotically given by:

$$\tau_k \sim \|\Sigma_\varepsilon^{-1/2} u_k\|^2 \sim \frac{\mathrm{tr}(\Sigma_\varepsilon^{-1})}{p} \sim \tau. \tag{197}$$

With this, we can show the improvement in SNR after whitening. We have:

$$\mathrm{SNR} \sim \frac{\ell_1}{\|N\|_{\mathrm{op}}^2} \lesssim \frac{\ell_1}{\mu_\varepsilon \|G\|_{\mathrm{op}}^2} \sim \frac{1}{\varphi} \frac{\ell_1 \tau}{\|G\|_{\mathrm{op}}^2} \sim \frac{1}{\varphi} \frac{\ell_1^{\mathrm{w}}}{\|G\|_{\mathrm{op}}^2} \sim \frac{\mathrm{SNR}^{\mathrm{w}}}{\varphi}. \tag{198}$$

This completes the proof.