Optimal prediction in the linearly transformed spiked model

Edgar Dobriban, William Leeb, and Amit Singer

July 12, 2018

Abstract

We consider the *linearly transformed spiked model*, where observations Y_i are noisy linear transforms of unobserved signals of interest X_i:

$$Y_i = A_i X_i + \varepsilon_i,$$

for $i = 1, \ldots, n$. The transform matrices A_i are also observed. We model X_i as random vectors lying on an unknown low-dimensional space. How should we predict the unobserved signals (regression coefficients) X_i?

The naive approach of performing regression for each observation separately is inaccurate due to the large noise. Instead, we develop optimal linear empirical Bayes methods for predicting X_i by "borrowing strength" across the different samples. Our methods are applicable to large datasets and rely on weak moment assumptions. The analysis is based on random matrix theory.

We discuss applications to signal processing, deconvolution, cryo-electron microscopy, and missing data in the high-noise regime. For missing data, we show in simulations that our methods are faster, more robust to noise and to unequal sampling than well-known matrix completion methods.

1   Introduction

In this paper we study the *linearly transformed spiked model*, where the observed data vectors Y_i are noisy linear transforms of unobserved signals of interest X_i:

$$Y_i = A_i X_i + \varepsilon_i, \ \ i = 1, \ldots, n.$$

We also observe the transform matrices A_i. A transform matrix reduces the dimension of the signal $X_i \in \mathbb{R}^p$ to a possibly observation-dependent dimension $q_i \leq p$, thus $A_i \in \mathbb{R}^{q_i \times p}$. Moreover, the signals are assumed to be random vectors lying on an unknown low-dimensional space, an assumption sometimes known as a spiked model (Johnstone, 2001).

Our main goal is to recover (estimate or predict) the unobserved signals X_i. The problem arises in many applications, some of which are discussed in the next section. Recovery is challenging due to the two different sources of information loss: First, the

1

transform matrices A_i reduce the dimension, since they are generally not invertible. It is crucial that the transform matrices differ between observations, as this allows us to reconstruct this lost information from different "snapshots" of X_i. Second, the observations are contaminated with additive noise ε_i. We study the regime where the size of the noise is much larger than the size of the signal. This necessitates methods that are not only numerically stable, but also reduce the noise significantly.

This setup can be viewed as a different linear regression problem for each sample $i = 1, \ldots, n$, with outcome vector Y_i and covariate matrix A_i. The goal is then to estimate the regression coefficients X_i. Since X_i are random, this is also a random effects model. Our specific setting, with low-rank X_i, is more commonly considered in spiked models, and we will call X_i the *signals*.

This paper assumes that the matrices $A_i^\top A_i \in \mathbb{R}^p$ are diagonal. Equivalently, we assume that the matrices $A_i^\top A_i$ all commute (and so can be jointly diagonalized). We will refer to this as the *commutative* model. This is mainly a technical assumption and we will see that it holds in many applications.

With large noise, predicting one X_i using one Y_i alone has low accuracy. Instead, our methods predict X_i by "borrowing strength" across the different samples. For this we model X_i as random vectors lying on an unknown low-dimensional space, which is reasonable in many applications. Thus our methods are a type of empirical Bayes methods (Efron, 2012).

Our methods are fast and applicable to big data, rely on weak distributional assumptions (only using moments), are robust to high levels of noise, and have certain statistical optimality results. Our analysis is based on recent insights from random matrix theory, a rapidly developing area of mathematics with many applications to statistics (e.g., Bai and Silverstein, 2009; Paul and Aue, 2014; Yao et al., 2015).

1.1 Motivation

We study the linearly transformed model motivated by its wide applicability to several important data analysis scenarios.

1.1.1 PCA and spiked model

In the well-known *spiked model* one observes data Y_i of the form $Y_i = X_i + \varepsilon_i$, where $X_i \in \mathbb{R}^p$ are unobserved signals lying on an unknown low dimensional space, and $\varepsilon_i \in \mathbb{R}^p$ is noise. With $A_i = I_p$ for all i, this is a special case of the commutative linearly transformed spiked model.

The spiked model is fundamental for understanding principal component analysis (PCA), and has been thoroughly studied under high-dimensional asymptotics. Its understanding will serve as a baseline in our study. Among the many references, see for instance Johnstone (2001); Baik et al. (2005); Baik and Silverstein (2006); Paul (2007); Nadakuditi and Edelman (2008); Nadler (2008); Bai and Ding (2012); Bai and Yao (2012); Benaych-Georges and Nadakuditi (2012); Onatski (2012); Onatski et al. (2013); Donoho et al. (2013); Onatski et al. (2014); Nadakuditi (2014); Gavish and Donoho (2014b); Johnstone and Onatski (2015); Hachem et al. (2015).

1.1.2 Noisy deconvolution in signal processing

The transformed spiked model is broadly relevant in signal acquisition and imaging. Measurement and imaging devices nearly never measure the "true" values of a signal. Rather, they measure a weighted average of the signal over a small window in time and/or space. Often, this local averaging can be modeled as the application of a convolution filter. For example, any time-invariant recording device in signal processing is modeled by a convolution (Mallat, 2008). Similarly, the blur induced by an imaging device can be modeled as convolution with a function, such as a Gaussian (Blackledge, 2006; Campisi and Egiazarian, 2016). In general, this filter will not be numerically invertible.

As is well-known, any convolution filter A_i is linear and diagonal in the Fourier basis; for example, see Stein and Shakarchi (2011). Consequently, $A_i^\top A_i$ is also diagonalized by the Fourier basis. Convolutions thus provide a rich source of examples of the linearly transformed spiked model.

1.1.3 Cryo-electron microscopy (cryo-EM)

Cryo - electron microscopy (cryo-EM) is an experimental method for mapping the structure of molecules. It allows imaging of heterogeneous samples, with mixtures or multiple conformations of molecules. This method has received a great deal of recent interest, and has recently led to the successful mapping of important molecules (e.g., Bai et al., 2015; Callaway, 2015).

Cryo-EM works by rapidly freezing a collection of molecules in a layer of thin ice, and firing an electron beam through the ice to produce two-dimensional images. The resulting observations can be modeled as $Y_i = A_i X_i + \varepsilon_i$, where X_i represents an unknown 3D molecule; A_i randomly rotates the molecule, projects it onto the xy-plane, and applies blur to the resulting image; and ε_i is noise (Katsevich et al., 2015). Since a low electron dose is used to avoid destroying the molecule, the images are typically very noisy.

When all the molecules in the batch are identical, i.e. $X_i = X$ for all i, the task of *ab-initio 3D reconstruction* is to recover the 3D molecule X from the noisy and blurred projections Y_i (Kam, 1980). Even more challenging is the problem of *heterogeneity*, in which several different molecules, or one molecule in different conformations, are observed together, without labels. The unseen molecules can usually be assumed to lie on some unknown low-dimensional space (Katsevich et al., 2015; Andén et al., 2015). Cryo-EM observations thus fit the linearly transformed spiked model.

The noisy deconvolution problem mentioned above is also encountered in cryo-EM. The operators A_i induce blur by convolution with a point-spread function (PSF), thus denoising leads to improved 3D reconstruction (Bhamre et al., 2016). The Fourier transform of the point-spread function is called the *contrast transfer function (CTF)*, and the problem of removing its effects from an image is known as *CTF correction*.

1.1.4 Missing data

Missing data can be modeled by *coordinate selection operators* A_i, such that $A_i(k, l) = 1$ if the k-th coordinate selected by A_i is l, and $A_i(k, l) = 0$ otherwise. Thus $A_i^\top A_i$ are diagonal with 0/1 entries indicating missing/observed coordinates. In the low-noise regime, missing data in matrices has recently been studied under the name of *matrix*

completion (e.g., Candès and Recht, 2009; Candès and Tao, 2010; Keshavan et al., 2009, 2010; Koltchinskii et al., 2011; Negahban and Wainwright, 2011; Recht, 2011; Rohde et al., 2011; Jain et al., 2013). As we discuss later, our methods perform well in the high-noise setting of this problem.

1.2 Our contributions

Our main contribution is to develop general methods predicting X_i in linearly transformed spiked models $Y_i = A_i X_i + \varepsilon_i$. We develop methods that are fast and applicable to big data, rely on weak moment assumptions, are robust to high levels of noise, and have certain optimality properties.

Our general approach is as follows: We model X_i as random vectors lying on an unknown low-dimensional space, $X_i = \sum_{k=1}^{r} \ell_k^{1/2} z_{ik} u_k$ for fixed unit vectors u_k and mean-zero scalar random variables z_{ik}, as usual in spiked models. In this model, the Best Linear Predictor (BLP), also known as the Best Linear Unbiased Predictor (BLUP), of X_i given Y_i is well known (Searle et al., 2009). (The more well known Best Linear Unbiased Estimator (BLUE) is defined for fixed-effects models where X_i are non-random parameters.) The BLP depends on the unknown population principal components u_k. In addition, it has a complicated form involving matrix inversion.

Our contributions are then:

1. We show that the BLP reduces to a simpler form in a certain natural high-dimensional model where $n, p \to \infty$ such that $p/n \to \gamma > 0$ (Sec. 7.8). In this simpler form, we can estimate the population principal components using the principal components (PCs) of the *backprojected data* $A_i^\top Y_i$ to obtain an Empirical BLP (EBLP) predictor (a type of moment-based empirical Bayes method), known up to some scaling coefficients. By an exchangeability argument, we show that the optimal scaling coefficients are the same as optimal singular value shrinkage coefficients for a certain novel random matrix model (Sec. 2.3).

2. We derive the asymptotically optimal singular value shrinkage coefficients (Sec. 3), by characterizing the spectrum of the backprojected data matrix (Sec. 3.1). This is our main technical contribution.

3. We derive a suitable "normalization" method to make our method fully implementable in practice (Sec. 2.4). This allows us to estimate the optimal shrinkage coefficients consistently, and to use well-known optimal shrinkage methods (Nadakuditi, 2014; Gavish and Donoho, 2014b). We also discuss how to estimate the rank (Sec. 3.4).

4. We also solve the out-of-sample prediction problem, where new Y_0, A_0 are observed, and X_0 is predicted using the existing data (Sec. 4).

5. We compare our methods to existing approaches for the special case of missing data problems via simulations (Sec. 5). These are reproducible with code provided on Github at `https://github.com/wleeb/opt-pred`.

4

2 Empirical linear prediction

2.1 The method

Our method is simple to state using elementary linear algebra. We give the steps here for convenience. In subsequent sections, we will explain each step, and prove the optimality of this procedure over a certain class of predictors. Our method has the following steps:

1. *Input*: Noisy linearly transformed observations Y_i, and transform matrices A_i, for $i = 1, \ldots, n$. Preliminary rank estimate r (see Sec. 3.4 for discussion).

2. Form backprojected data matrix $B = [A_1^\top Y_1, \ldots, A_n^\top Y_n]^\top$ and diagonal normalization matrix $\hat{M} = n^{-1/2} \sum_{i=1}^n A_i^\top A_i$. Form the normalized, backprojected data matrix $\tilde{B} = B\hat{M}^{-1}$.

3. *(Optional)* Multiply $\tilde{B}$ by a diagonal whitening matrix W, $\tilde{B} \leftarrow \tilde{B}W$. The definition of W is given in Sec. 3.3.1.

4. Compute the singular values σ_k and the top r singular vectors $\hat{u}_k, \hat{v}_k$ of the matrix $\tilde{B}$.

5. Compute $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_n)^\top = \sum_{k=1}^r \hat{\lambda}_k \hat{u}_k \hat{v}_k^\top$.

   Here $\hat{\lambda}_k$ are computed according to Sec. 3: $\hat{\lambda}_k = \hat{\ell}_k^{1/2} \hat{c}_k \hat{\tilde{c}}_k$, where $\hat{\ell}_k, \hat{c}_k, \hat{\tilde{c}}_k$ are estimated based on the formulas given in Theorem 3.1 by plug-in. Specifically, $\hat{\ell}_k = 1/\hat{D}(\sigma_k^2)$, $\hat{c}_k^2 = \hat{m}(\sigma_k^2)/[\hat{D}'(\sigma_k^2)\hat{\ell}_k]$, $\hat{\tilde{c}}_k^2 = \underline{\hat{m}}(\sigma_k^2)/[\hat{D}'(\sigma_k^2)\hat{\ell}_k]$, where $\hat{m}, \underline{\hat{m}}, \hat{D}, \hat{D}'$ are the plug-in estimators of the Stieltjes-transform-like functionals of the spectral distribution, using the bottom $\min(n,p) - r$ eigenvalues of the sample covariance matrix of the backprojected data. For instance, $\hat{m}$ is given in equation (3.2.1) (assuming $p \leq n$):

$$\hat{m}(x) = \frac{1}{p-r} \sum_{k=r+1}^p \frac{1}{\sigma_k^2 - x}.$$

6. If whitening was performed (Step 3), unwhiten the data, $\hat{X} \leftarrow \hat{X}W^{-1}$.

7. *Output*: Predictions $\hat{X}_i$ for X_i, for $i = 1, \ldots, n$.

The complexity of the method is dominated by computing the singular value spectrum of the backprojected matrix, which takes $O(\min(n,p)^2 \cdot \max(n,p))$ floating point operations. As we will show in Sec. 3.3, by choosing a certain whitening matrix W, the algorithm will only require computing the top r singular vectors and values of the backprojected data matrix, and so can typically be performed at an even lower cost using, for example, the Lanczos algorithm (Golub and Van Loan, 2012), especially when there is a low cost of applying the matrix $\tilde{B}$ to a vector.

2.2 Motivation I: from BLP to EBLP

We now explain the steps of our method. We will use the mean-squared error $\mathbb{E}\|\hat{X}_i - X_i\|^2$ to assess the quality of a predictor $\hat{X}_i$. Recall that we modeled the signals as $X_i = \sum_{k=1}^r \ell_k^{1/2} z_{ik} u_k$. It is well known in random effects models (e.g., Searle et al., 2009) that the best linear predictor, or BLP, of one signal X_i using Y_i, is:

$$\hat{X}_i^{BLP} = \Sigma_X A_i^\top (A_i \Sigma_X A_i^\top + \Sigma_\varepsilon)^{-1} Y_i. \tag{1}$$

Here, $\Sigma_X = \sum_{k=1}^{r} \ell_k u_k u_k^\top$ denotes the covariance matrix of one X_i, and Σ_ε is the covariance matrix of the noise ε_i. These are unknown parameters, so we need to estimate them in order to get a *bona fide* predictor. Moreover, though A_i are fixed parameters here, we will take them to be random later.

We are interested in the "high-dimensional" asymptotic regime, where the dimension p grows proportionally to the number of samples n; that is, $p = p(n)$ and $\lim_{n\to\infty} p(n)/n = \gamma > 0$. In this setting it is in general not possible to estimate the population covariance Σ_X consistently. Therefore, we focus our attention on alternate methods derived from the BLP.

The BLP involves the inverse of a matrix, which makes it hard to analyze. However, for certain *uniform models* (see Sec. 7.8 for a precise definition), we can show that the BLP is asymptotically equivalent to a simpler linear predictor not involving a matrix inverse:

$$\hat{X}_i^0 = \sum_{k=1}^{r} \eta_k^0 \langle A_i^\top Y_i, u_k \rangle u_k.$$

Here η_k^0 are certain constants given in Sec. 7.8. This simple form of the BLP guides our choice of predictor when the true PCs are not known. Let $\hat{u}_1, \ldots, \hat{u}_r$ be the empirical PCs; that is, the top eigenvectors of the sample covariance $\sum_{i=1}^{n} (A_i^\top Y_i)(A_i^\top Y_i)^\top/n$, or equivalently, the top left singular vectors of the matrix $[A_1^\top Y_1, \ldots, A_n^\top Y_n]^\top$. For coefficients $\eta = (\eta_1, \ldots, \eta_r)$, substituting $\hat{u}_k$ for u_k leads us to the following *empirical linear predictor*:

$$\hat{X}_i^\eta = \sum_{k=1}^{r} \eta_k \langle A_i^\top Y_i, \hat{u}_k \rangle \hat{u}_k.$$

Note that, since the empirical PCs $\hat{u}_k$ are used in place of the population PCs u_k, the coefficients η_k defining the BLP are no longer optimal, and must be adjusted downwards to account for the non-zero angle between u_k and $\hat{u}_k$. This phenomenon was studied in the context of the ordinary spiked model in Singer and Wu (2013).

2.3 Motivation II: Singular value shrinkage

Starting with BLP and replacing the unknown population PCs u_k with their empirical counterparts $\hat{u}_k$, we were lead to a predictor of the form $\hat{X}_i^\eta = \sum_{k=1}^{r} \eta_k \langle B_i, \hat{u}_k \rangle \hat{u}_k$, where $B_i = A_i^\top Y_i$ are the backprojected data. Now, the matrix $\hat{X}^\eta = [\hat{X}_1^\eta, \ldots, \hat{X}_n^\eta]^\top$ has the form

$$\hat{X}^\eta = \sum_{k=1}^{r} \eta_k \cdot B\hat{u}_k \hat{u}_k^\top = \sum_{k=1}^{r} \eta_k \sigma_k(B) \cdot \hat{v}_k \hat{u}_k^\top. \tag{2}$$

This has the same singular vectors as the matrix $B = [B_1, \ldots, B_n]^\top$ of backprojected data.

From now on, we will consider the A_i as random variables, which corresponds to an average-case analysis over their variability. Then observe that the predictors $\hat{X}_i^\eta$ are exchangeable random variables with respect to the randomness in A_i, ε_i, because they depend symmetrically on the data matrix B. Therefore, the prediction error

for a sample equals the average prediction error over all X_i, which is the normalized Frobenius norm for predicting the matrix $X = (X_1, \ldots, X_n)^\top$:

$$\mathbb{E}\|\hat{X}_i^\eta - X_i\|^2 = \frac{1}{n}\mathbb{E}\|\hat{X}^\eta - X\|_F^2.$$

Therefore, the empirical linear predictors are equivalent to performing singular value shrinkage of the matrix B to estimate X. That is, singular value shrinkage predictors are in one-to-one correspondence with the in-sample empirical linear predictors. Because singular value shrinkage is minimax optimal for matrix denoising problems with Gaussian white noise (Gavish and Donoho, 2014a), it is a natural choice of predictor in the more general setting we consider in this paper, where an optimal denoiser is not known.

2.4   The class of predictors: shrinkers of normalized, back-projected data

Motivated by the previous two sections, we are led to singular value shrinkage predictors of the matrix X. However, it turns out that rather than shrink the singular values of the matrix B of backprojected data $A_i^\top Y_i$, it is more natural to work instead with the matrix $\tilde{B}$ with rows $\tilde{B}_i = M^{-1}A_i^\top Y_i$, where $M = \mathbb{E}A_i^\top A_i$ is a diagonal normalization matrix. We will show later that we can use a sample estimate of M.

The heuristic to explain this is that we can write $A_i^\top A_i = M + E_i$, where E_i is a mean zero diagonal matrix. We will show in the proof of Thm. 3.1 that because the matrices $A_i^\top A_i$ commute, the matrix with rows $E_i X_i / \sqrt{n}$ has operator norm that vanishes in the high-dimensional limit $p/n \to \gamma$. Consequently, we can write:

$$B_i = A_i^\top Y_i = MX_i + A_i^\top \varepsilon_i + E_i X_i \sim \underbrace{MX_i}_{signal} + \underbrace{A_i^\top \varepsilon_i}_{noise}$$

Since X_i lies in an r-dimensional subspace, spanned by $u_1, \ldots, u_r$, MX_i also lies in the r-dimensional subspace spanned by $Mu_1, \ldots, Mu_r$. Furthermore, $A_i^\top \varepsilon_i$ is mean-zero and independent of MX_i. Consequently, $A_i^\top Y_i$ looks like a spiked model, with signal MX_i and noise $A_i^\top \varepsilon_i$.

Shrinkage of this matrix will produce a predictor of MX_i, not X_i itself. However, multiplying the data by M^{-1} fixes this problem: we obtain the approximation:

$$\tilde{B}_i = M^{-1}A_i^\top Y_i \sim X_i + \underbrace{M^{-1}A_i^\top \varepsilon_i}_{noise}.$$

After this normalization, the target signal of any shrinker becomes the true signal X_i itself.

Motivated by these considerations, we can finally state the class of problems we study. We consider predictors of the form:

$$\hat{X}_i^\eta = \sum_{k=1}^r \eta_k \langle \tilde{B}_i, \hat{u}_k \rangle \hat{u}_k$$

where $\tilde{B}_i = M^{-1}A_i^\top Y_i$, and we seek the AMSE-optimal coefficients η_k^* in the high-dimensional limit $p/n \to \gamma$; that is, our goal is to find the optimal coefficients η_k, minimizing the AMSE:

$$\eta^* = \arg\min_\eta \lim_{p,n\to\infty} \mathbb{E}\|\hat{X}_i^\eta - X_i\|^2.$$

We will show that the limit exists. The corresponding estimator $\hat{X}_i^{\eta^*}$ will be called the *empirical best linear predictor (EBLP)*. We will: (1) show that it is well-defined; (2) derive the optimal choice of η_k; (3) derive consistent estimators of the optimal η_k; and (4) derive consistently estimable formulas for the AMSE. As before, finding the optimal η_k is equivalent to performing optimal singular value shrinkage on the matrix $\tilde{B} = [\tilde{B}_1, \ldots, \tilde{B}_n]^\top$.

3 Derivation of the optimal coefficients

As described in Sec. 2, we wish to find the AMSE-optimal coefficients η_k for predictors of the form $\hat{X}_i^\eta = \sum_{k=1}^r \eta_k \langle \tilde{B}_i, \hat{u}_k \rangle \hat{u}_k$, where $\tilde{B}_i = M^{-1}A_i^\top Y_i$ is the normalized, back-projected data. Equivalently, we find the optimal singular values of the matrix with the same singular vectors as $\tilde{B} = [\tilde{B}_1, \ldots, \tilde{B}_n]^\top$.

Singular value shrinkage has been the subject of a lot of recent research. It is now well known that optimal singular value shrinkage depends on the asymptotic spectrum of the data matrix $\tilde{B}$ (e.g., Nadakuditi, 2014; Gavish and Donoho, 2017). We now fully characterize the spectrum, and use it to derive the optimal singular values. We then show that by estimating the optimal singular values by plug-in, we get the method described in Sec. 2.1.

3.1 The asymptotic spectral theory of the back-projected data

The main theorem characterizes the asymptotic spectral theory of the normalized back-projected data matrix $\tilde{B} = BM^{-1}$, and of the unnormalized version $B = [A_1^\top Y_1, \ldots, A_n^\top Y_n]^\top$. Our data are iid samples of the form $Y_i = A_i X_i + \varepsilon_i$.

We assume that the signals have the form $X_i = \sum_{k=1}^r \ell_k^{1/2} z_{ik} u_k$. Here u_k are deterministic signal directions with $\|u_k\| = 1$. We will assume that u_k are delocalized, so that $|u_k|_\infty \le C_p$ for some constants $C_p \to 0$ that we will specify later. The scalars z_{ik} are standardized independent random variables, specifying the variation in signal strength from sample to sample. For simplicity we assume that the deterministic spike strengths are different and sorted: $\ell_1 > \ell_2 > \ldots > \ell_r > 0$.

For a distribution H, let $F_{\gamma,H}$ denote the generalized Marchenko-Pastur distribution induced by H with aspect ratio γ (Marchenko and Pastur, 1967). Closely related to $F_{\gamma,H}$ is the so-called *companion distribution* $\underline{F}_{\gamma,H}(x) = \gamma F_{\gamma,H}(x) + (1-\gamma)\delta_0$. We will also need the Stieltjes transform $m_{\gamma,H}$ of $F_{\gamma,H}$, $m_{\gamma,H}(z) = \int (x-z)^{-1} dF_{\gamma,H}(x)$, and the Stieltjes transform $\underline{m}_{\gamma,H}$ of $\underline{F}_{\gamma,H}$. Based on these, one can define the D-transform of $F_{\gamma,H}$ by

$$D_{\gamma,H}(x) = x \cdot m_{\gamma,H}(x) \cdot \underline{m}_{\gamma,H}(x).$$

Up to the change of variables $x = y^2$, this agrees with the D-transform defined in Benaych-Georges and Nadakuditi (2012). Let $b^2 := b_H^2$ be the supremum of the support

of $F_{\gamma,H}$, and $D_{\gamma,H}(b_H^2) = \lim_{t\downarrow b} D_{\gamma,H}(t^2)$. It is easy to see that this limit is well defined, and is either finite or $+\infty$.

We will assume the following conditions:

1. **Commutativity condition.** The matrices $A_i^\top A_i$ commute with each other. Equivalently, they are jointly diagonal in some known basis. For simplicity of notation, we will assume without loss of generality that the $A_i^\top A_i$ are diagonal.

2. **Backprojected noise.** The vectors $\varepsilon_i^* = A_i^\top \varepsilon_i$ have independent entries of mean zero. If H_p is the distribution function of the variances of the entries of $M^{-1}\varepsilon_i^*$, then H_p is bounded away from zero; and $H_p \Rightarrow H$ almost surely, where H is a compactly supported distribution.

3. **Maximal noise variance.** The supremum of the support of H_p converges almost surely to the upper edge of the support of H.

4. **Noise moments.** $\mathbb{E}|\varepsilon_{ij}^*|^{6+\phi} < C$, $\mathbb{E}|E_{ij}|^{6+\phi} < C$ (recall that we defined $E_i = A_i^\top A_i - M$).

5. **Signal.** One of the following two assumptions holds for the signal directions u_k and signal coefficients z_{ij}:

    - **Polynomial moments and delocalization.** Suppose $\mathbb{E}|z_{ij}|^m \le C < \infty$ for some $m > 4$ and for all k

    $$\|u_k\|_\infty \cdot p^{(2+c)/m} \to_{a.s} 0$$

    for some $c > 0$.
    - **Exponential moments and logarithmic delocalization.** Suppose the z_{ij} are sub-gaussian in the sense that $\mathbb{E}\exp(t|z_{ij}|^2) \le C$ for some $t > 0$ and $C < \infty$, and that for all k

    $$\|u_k\|_\infty \cdot \sqrt{\log p} \to_{a.s} 0.$$

6. **Generic signal.** Let P be the diagonal matrix with $P_{jj} = \mathrm{Var}[M_j^{-1}\varepsilon_{ij}^*]$, where M_j are the diagonal entries of the diagonal matrix $M = \mathbb{E}A_i^\top A_i$. Then u_j are *generic* with respect to P, in the sense that there are some constants $\tau_k > 0$ such that:
    $$u_j^\top (P - zI_p)^{-1}u_k \to I(j = k) \cdot \tau_k \cdot m_H(z)$$
    for all $z \in \mathbb{C}^+$.

Before stating the main results, we make a few remarks on these assumptions. Assumption 1 holds for many applications, as discussed in Sec. 1.1. However, our analysis will go through if a weaker condition is placed on matrices $A_i^\top A_i$, namely that they are *diagonally dominant* in a known basis, in the sense that the off-diagonal elements are asymptotically negligible to the operator norm. Because it does not change anything essential in the analysis, for ease of exposition we will analyze the exact commutativity condition.

The part of Assumption 2 that the entries of $\varepsilon_i^* = A_i^\top \varepsilon_i$ are independent is easily checked for certain problems, such as missing data with independently selected coordinates. However, it may not always hold. For example, in the problem of CTF correction in cryo-EM (see Sec. 1.1), each A_i may be one of a discrete number of different CTFs; in this case, the assumption will not hold exactly. However, we have found
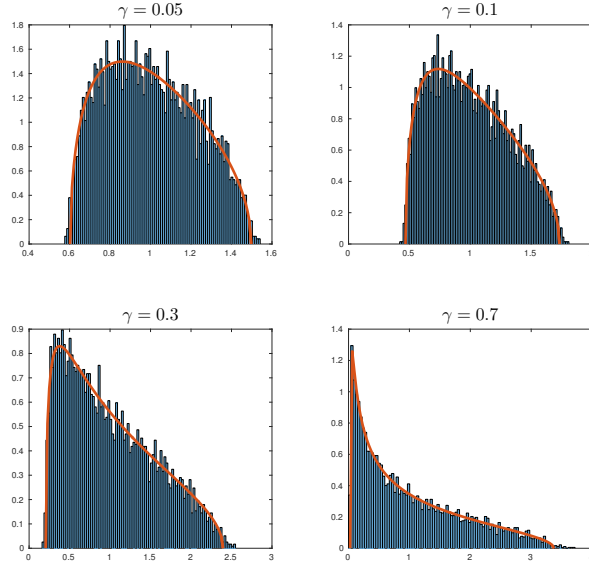
Figure 1: Histograms of empirical eigenvalues of whitened, backprojected noise using 30 CTFs, plotted against the Marchenko-Pastur density for different aspect ratios γ.

in practice that the Marchenko-Pastur law holds even in this regime. To illustrate this, in Fig. 1 we plot histograms of the sample covariance eigenvalues of simulated backprojected isotropic Gaussian noise using 30 different synthetic CTFs, generated using the ASPIRE software package (ASPIRE, 2017), for 30 defocus values between 0.5 and 3. We plot the coefficients of the backprojected noise in the first frequency block of a steerable basis with radial part the Bessel functions, as described in Bhamre et al. (2016) and Zhao et al. (2016). Because this frequency block only contains 49 coefficients, the histogram we plot is for 100 draws of the noise. We whiten the back-projected noise, so the population covariance is the identity. As is evident from the figure, there is a very tight agreement between the empirical distribution of eigenvalues and the Marchenko-Pastur laws.

Assumption 5 about the signals presents a tradeoff between the delocalization of the spike eigenvectors and the moments of the signal coefficients. If a weak polynomial moment assumption or order m holds for the signal coefficients z_{ij}, then it requires a delocalization at a polynomial rate $p^{-(2+c)/m}$ for the spike eigenvectors. In particular, this implies that at least a polynomial number of coefficients of u_k must be nonzero, so that u_k must be quite non-sparse. In contrast, if we assume a stronger sub-Gaussian moment condition for the noise, then only a logarithmic delocalization is required, which allows u_k to be quite sparse.

This assumption is similar to the incoherence condition from early works on matrix completion (e.g., Candès and Recht, 2009, etc.). Later works have shown that some form of recovery is possible even if we do not have incoherence (e.g., Koltchinskii et al., 2011). However, in our case, complete sparsity of order one (i.e., only a fixed number of nonzero coordinates) seems impossible to recover. Indeed, suppose the rank is one and $u = (1, 0, \ldots, 0)$. Then, all information about u and z is in the first coordinate. In our sampling model, we observe a fixed fraction q of the coordinates, and we can have $q < 1$. Thus, for the unobserved coordinates, there is no information about the z_i. Therefore, with the current random sampling mechanism, we think that accurate

estimation is not possible for fixed sparsity.

Assumption 6 generalizes the existing conditions for spiked models. In particular, it is easy to see that it holds when the vectors u_k are random with independent coordinates. Specifically, let x be a random vector with iid zero-mean entries with variance $1/p$. Then $\mathbb{E}x^\top(P - zI_p)^{-1}x = p^{-1}\text{tr}(P - zI_p)^{-1}$. Assumption 6 requires that this converges to $m_H(z)$, which follows from $H_p \Rightarrow H$. However, Assumption 6 is more general, as it does not require any kind of randomness in u_k.

Our main result in this section is the following.

Theorem 3.1 (Spectrum of transformed spiked models). *Under the above conditions, the eigenvalue distribution of $\tilde{B}^\top\tilde{B}/n$ converges to the general Marchenko-Pastur law $F_{\gamma,H}$ a.s. In addition, for $k \leq r$, the k-th largest eigenvalue of $\tilde{B}^\top\tilde{B}/n$ converges, $\lambda_k(\tilde{B}^\top\tilde{B})/n \to t_k^2$ a.s., where*

$$t_k^2 = \begin{cases} D_{\gamma,H}^{-1}(\frac{1}{\ell_k}) & \text{if } \ell_k > 1/D_{\gamma,H}(b_H^2), \\ b_H^2 & \text{otherwise.} \end{cases} \tag{3}$$

Moreover, let $\hat{u}_k$ be the right singular vector of $\tilde{B}$ corresponding to $\lambda_k(\tilde{B}^\top\tilde{B})$. Then $(u_j^\top\hat{u}_k)^2 \to c_{jk}^2$ a.s., where

$$c_{jk}^2 = \begin{cases} \frac{m_{\gamma,H}(t_k^2)}{D'_{\gamma,H}(t_k^2)\ell_k} & \text{if } j = k \text{ and } \ell_k > 1/D_{\gamma,H}(b_H^2), \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Finally, let $Z_j = n^{-1/2}(z_{1j}, \ldots, z_{nj})^\top$, and let $\hat{Z}_k$ be the k-th left singular vector of $\tilde{B}$. Then $(Z_j^\top\hat{Z}_k)^2 \to \tilde{c}_{jk}^2$ a.s., where

$$\tilde{c}_{jk}^2 = \begin{cases} \frac{m_{\gamma,H}(t_k^2)}{D'_{\gamma,H}(t_k^2)\ell_k} & \text{if } j = k \text{ and } \ell_k > 1/D_{\gamma,H}(b_H^2), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The proof is in Sec. 7.1. While the conclusion of this theorem is very similar to the results of Benaych-Georges and Nadakuditi (2012), our observation model $Y_i = A_iX_i + \varepsilon_i$ is entirely different from the one in that paper; we are addressing a different problem. Moreover, our technical assumptions are also more general and more realistic, and only require finite moments up to the sixth moment, unlike the more stringent conditions in previous work. In addition, we also have the result below, which differs from existing work.

For the un-normalized backprojected matrix B, a version of Thm. 3.1 applies *mutatis mutandis*. Specifically, we let H_p be the distribution of the variances of $A_i^\top\varepsilon_i$. We replace I_p with M in the assumptions when needed, so we let $\tau_k = \lim_{n\to\infty}\|Mu_k\|^2$, and $\nu_j = Mu_j/\|Mu_j\|$. Then the above result holds for B, with ℓ_k replaced by $\tau_k\ell_k$, and u_j replaced by ν_j. The proof is identical, and is also presented in Sec. 7.1.

3.2 Optimal singular value shrinkage

Theorem 3.1 describes precisely the limiting spectral theory of the matrix $\tilde{B}/\sqrt{n}$. Specifically, we derived formulas for the limiting cosines c_k and $\tilde{c}_k$ of the angles between the top r singular vectors of $\tilde{B}/\sqrt{n}$ and $X/\sqrt{n}$, and the relationship between the top singular values of these matrices.

11

It turns out, following the work of Gavish and Donoho (2017) and Nadakuditi (2014), that this information is sufficient to derive the optimal singular value shrinkage predictor of X. It is shown in Gavish and Donoho (2017) that $\lambda_i^* = \ell_k^{1/2} c_k \tilde{c}_k$, under the convention $c_k, \tilde{c}_k > 0$. Furthermore, the AMSE of this predictor is given by $\sum_{k=1}^{r} \ell_k(1 - c_k^2 \tilde{c}_k^2)$. We outline the derivation of these formulas in Sec. 7.11, though the reader may wish to refer to Gavish and Donoho (2017) for a more detailed description of the method, as well as extensions to other loss functions.

We next show how to derive consistent estimators of the angles and the limiting singular values of the observed matrix. Plugging these into the expression $\lambda_i^* = \ell_i^{1/2} c_i \tilde{c}_i$, we immediately obtain estimators of the optimal singular values λ_i^*. This will complete the proof that the algorithm given in Sec. 2.1 solves the problem posed in Sec. 2.4 and defines the EBLP.

3.2.1 Estimating ℓ_k, c_k and $\tilde{c}_k$

To evaluate the optimal λ_i^*, we estimate the values of ℓ_k, c_k, and $\tilde{c}_k$ using Thm. 3.1 whenever $\ell_k \geq b_H^2$ (that is, if the signal is strong enough). From (3) we have the formula $\ell_k = 1/D_{\gamma,H}(t_k^2)$ where t_k is the limiting singular value of the observed matrix $\tilde{B}/\sqrt{n}$. We also have the formulas (4) and (5) for c_k and the $\tilde{c}_k$.

We will estimate the Stieltjes transform $m_{\gamma,H}(z)$ by the *sample Stieltjes transform*, defined as:

$$\hat{m}_{\gamma,H}(z) = \frac{1}{p-r} \sum_{k=r+1}^{p} \frac{1}{\lambda_k - z},$$

where the sum is over the bottom $p - r$ eigenvalues λ_k of $\tilde{B}^\top \tilde{B}/n$. It is shown by Nadakuditi (2014) that $\hat{m}_{\gamma,H}$ is a consistent estimator of $m_{\gamma,H}$, and that using the corresponding plug-in estimators of $\underline{m}_{\gamma,H}$, $D_{\gamma,H}$ and $D'_{\gamma,H}$, we also obtain consistent estimators of ℓ_k, c_k, and $\tilde{c}_k$.

3.2.2 Using $\hat{M}$ in place of M

To make the procedure fully implementable, we must be able to estimate the mean matrix $M = \mathbb{E}A_i^\top A_i$. If M is estimated from the n iid matrices $A_i^\top A_i$ by the sample mean $\hat{M} = n^{-1} \sum_{i=1}^{n} A_i^\top A_i$, we show that multiplying by $\hat{M}^{-1}$ has asymptotically the same effect as multiplying by the true M^{-1}, assuming that the diagonal entries of M are bounded below. This justifies our use of $\hat{M}$.

Lemma 3.2. *Suppose that the entries M_i of M are bounded away from 0: $M_i \geq \delta$ for some $\delta > 0$, for all i. Let $\hat{M} = n^{-1} \sum_{i=1}^{n} A_i^\top A_i$. Then*

$$\lim_{p,n\to\infty} n^{-1/2} \|BM^{-1} - B\hat{M}^{-1}\|_{op} = 0.$$

See Sec. 7.10 for the proof. Note that the condition of this lemma are violated only when the entries of M can be arbitrarily small; but in this case, the information content in the data on the corresponding coordinates vanishes, so the problem itself is ill-conditioned. The condition is therefore reasonable in practice.

3.3 Prediction for weighted loss functions: whitening and big data

In certain applications there may be some directions that are more important than others, whose accurate prediction is more heavily prized. We can capture this by considering weighted Frobenius loss functions $\|\hat{X}_i - X_i\|_W^2 = \|W(\hat{X}_i - X_i)\|^2$, where W is a positive-definite matrix. Can we derive optimal shrinkers with respect to these weighted loss functions?

The weighted error can be written as $\|\hat{X}_i - X_i\|_W^2 = \|W(\hat{X}_i - X_i)\|^2 = \|\widehat{WX_i} - WX_i\|^2$. In other words, the problem of predicting X_i in the W-norm is identical to predicting WX_i in the usual Frobenius norm. Because the vectors WX_i lie in an r-dimensional subspace (spanned by $Wu_1, \ldots, Wu_r$), the same EBLP method we have derived for X_i can be applied to prediction of WX_i, assuming that the technical conditions we imposed for the original model hold for this transformed model. That is, we perform singular value shrinkage on the matrix of transformed observations $W\tilde{B}_i$.

To explore this further, recall that after applying the matrix M^{-1} to each vector $A_i^\top Y_i$, the data matrix behaves asymptotically like the matrix with columns $X_i + \tilde{\varepsilon}_i$, for some noise vectors $\tilde{\varepsilon}_i$ that are independent of the signal X_i. The observations $WM^{-1}A_i^\top Y_i$ are asymptotically equivalent to $WX_i + W\tilde{\varepsilon}_i$. If we choose W to be the square root of the inverse covariance of $\tilde{\varepsilon}_i$, then the effective noise term $W\tilde{\varepsilon}_i$ has a identity covariance; we call this transformation "whitening the effective noise".

One advantage of whitening is that there are closed formulas for the asymptotic spikes and cosines. This is because the Stieltjes transform of white noise has an explicit closed formula; see Bai and Silverstein (2009). To make sense of the formulas, we will assume that the low-rank model WX_i satisfies the assumptions we initially imposed on X_i; that is, we will assume:

$$WX_i = \sum_{k=1}^{r} \tilde{\ell}_k^{1/2} \tilde{z}_{ik} \tilde{u}_k \tag{6}$$

where the z_{ik} are iid and the $\tilde{u}_k$ are orthonormal. With this notation, the empirical eigenvalues of $W\tilde{B}^\top\tilde{B}W/n$ converge to

$$\lambda_k = \begin{cases} (\tilde{\ell}_k + 1)\left(1 + \frac{\gamma}{\tilde{\ell}_k}\right) & \text{if } \tilde{\ell}_k > \sqrt{\gamma}, \\ (1 + \sqrt{\gamma})^2 & \text{otherwise} \end{cases}$$

while the limit of the cosine of the angle between the kth empirical PC $\hat{u}_k$ and the kth population PC u_k is

$$c_k^2 = \begin{cases} \frac{1 - \gamma/\tilde{\ell}_k^2}{1 + \gamma/\tilde{\ell}_k} & \text{if } \tilde{\ell}_k > \sqrt{\gamma}, \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

and the limit of the cosine of the angle between the kth empirical left singular vector $\hat{v}_k$ and the kth left population singular vector v_k is

$$\tilde{c}_k^2 = \begin{cases} \frac{1 - \gamma/\tilde{\ell}_k^2}{1 + 1/\tilde{\ell}_k} & \text{if } \tilde{\ell}_k > \sqrt{\gamma}, \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

These formulas are derived in Benaych-Georges and Nadakuditi (2012); also see Paul (2007).

Following Sec. 3.2, the W-AMSE of the EBLP is $\sum_{k=1}^{r} \tilde{\ell}_k(1 - c_k^2 \tilde{c}_k^2)$. Since the parameters $\tilde{\ell}_k$, c_k and $\tilde{c}_k$ are estimable from the observations, the W-AMSE can be explicitly estimated.

Using these formulas makes evaluation of the optimal shrinkers faster, as we avoid estimating the Stieltjes transform from the bottom $p - r$ singular values of $\tilde{B}$. Using whitening, the entire method only requires computation of the top r singular vectors and values. Whitening thus enables us to scale our methods to extremely large datasets.

3.3.1 Estimating the whitening matrix W

In the observation model $Y_i = A_i^\top X_i + \varepsilon_i$, if the original noise term ε_i has identity covariance, that is $\Sigma_\varepsilon = I_p$, then it is straightforward to estimate the covariance of the "effective" noise vector $\tilde{\varepsilon}_i = M^{-1} A_i^\top \varepsilon_i$, and consequuently to estimate the whitening matrix $W = \Sigma_{\tilde{\varepsilon}}^{-1/2}$.

It is easy to see that $A_i^\top \varepsilon_i$ has covariance $M = \mathbb{E}[A_i^\top A_i]$, which is diagonal. Then the covariance of $\tilde{\varepsilon}_i$ is $M^{-1} M M^{-1} = M^{-1}$, and $W = M^{1/2}$. As in the proof of Lemma 3.2, W can be consistently estimated from the data by the sample mean $\sum_{i=1}^{n} (A_i^\top A_i)^{1/2}/n$.

3.4 Selecting the rank

Our method requires a preliminary rank estimate. Our results state roughly that, after backprojection, the linearly transformed spiked model becomes a spiked model. So we believe we may be able to adapt some popular methods for selecting the number of components in spiked models. There are many such methods, and it is not our goal to recommend a particular one. One popular method in applied work is a permutation method called parallel analysis (Buja and Eyuboglu, 1992; Dobriban, 2017), for which we have proposed improvements (Dobriban and Owen, 2017). For other methods, see Kritchman and Nadler (2008); Passemier and Yao (2012), and also Yao et al. (2015), Ch. 11, for a review.

If the method is strongly consistent, in the sense that the number of components is almost surely correctly estimated, then it is easy to see that the entire proof works. Specifically, the optimal singular value shrinkers can be obtained using the same orthonormalization method, and they can also be estimated consistently. Thus, for instance the methods from Passemier and Yao (2012); Dobriban and Owen (2017) are applicable if the spike strengths are sufficiently large.

4 Out-of-sample prediction

In Sec. 3, we derived the EBLP for predicting X_i from $Y_i = A_i X_i + \varepsilon_i$, $i = 1, \ldots, n$. We found the optimal coefficients η_k for the predictor $\sum_{k=1}^{r} \eta_k \langle \tilde{B}_i, \hat{u}_k \rangle \hat{u}_k$, where the $\hat{u}_k$ are the empirical PCs of the normalized back-projected data $\tilde{B}_i = \hat{M}^{-1} A_i^\top Y_i$.

Now suppose we are given another data point, call it $Y_0 = A_0 X_0 + \varepsilon_0$, drawn from the same model, but independent of $Y_1, \ldots, Y_n$, and we wish to predict X_0 from an expression of the form $\sum_{k=1}^{r} \eta_k \langle \tilde{B}_0, \hat{u}_k \rangle \hat{u}_k$.

At first glance, this problem appears identical to the one already solved. However, there is a subtle difference: the new data point is *independent of the empirical PCs* $\hat{u}_1, \ldots, \hat{u}_r$. It turns out that this independence forces us to use a *different* set of coefficients η_k to achieve optimal prediction.

We call this the problem of *out-of-sample prediction*, and the optimal predictor the *out-of-sample EBLP*. To be clear, we will refer to the problem of predicting $Y_1, \ldots, Y_n$ as *in-sample prediction*, and the optimal predictor as the *in-sample EBLP*. We call $(Y_1, A_1), \ldots, (Y_n, A_n)$ the *in-sample observations*, and (Y_0, A_0) the *out-of-sample observation*.

One might object that solving the out-of-sample problem is unnecessary, since we can always convert the out-of-sample problem into the in-sample problem. We could enlarge the in-sample data to include Y_0, and let $\hat{u}_k$ be the empirical PCs of this extended data set. While this is true, it is often not practical for several reasons. First, in on-line settings where a stream of data must be processed in real-time, recomputing the empirical PCs for each new observation may not be feasible. Second, if n is quite large, it may not be viable to store all of the in-sample data $Y_1, \ldots, Y_n$; the r vectors $\hat{u}_1, \ldots, \hat{u}_r$ require an order of magnitude less storage.

In this section we will first present the steps of the out-of-sample EBLP. Then we will provide a rigorous derivation. We will also show that the AMSEs for in-sample and out-of-sample EBLP with respect to squared W-norm loss are identical, where W is the inverse square root of the effective noise covariance. This is a rather surprising result that gives statistical justification for the use of out-of-sample EBLP, in addition to the computational considerations already described.

4.1 Out-of-sample EBLP

The out-of-sample denoising method can be stated simply, similarly to the in-sample algorithm in Sec. 2.1. We present the steps below.

1. *Input*: The top r in-sample empirical PCs $\hat{u}_1, \ldots, \hat{u}_r$. Estimates of the eigenvalues $\hat{\ell}_1, \ldots, \hat{\ell}_r$ and cosines $\hat{c}_1, \ldots, \hat{c}_r$. An estimate $\hat{\Sigma}_{\tilde{\varepsilon}}$ of the noise covariance $\Sigma_{\tilde{\varepsilon}}$ of the normalized backprojected noise vectors $\tilde{\varepsilon}_i = M^{-1} A_i^\top \varepsilon_i$. The diagonal matrix $\hat{M}^{-1}$ which is the inverse of an estimate of the covariance matrix of the noise ε_i, and an out-of-sample observation (Y_0, A_0).

2. Construct the vector $\tilde{B}_0 = \hat{M}^{-1} A_0^\top Y_0$.

3. Compute estimators of the out-of-sample coefficients $\eta_1, \ldots, \eta_r$. These are given by the formula $\hat{\eta}_k = \frac{\hat{\ell}_k \hat{c}_k^2}{\hat{\ell}_k \hat{c}_k^2 + \hat{d}_k}$, where $\hat{d}_k = \hat{u}_k^\top \hat{\Sigma}_{\tilde{\varepsilon}} \hat{u}_k$.

4. *Output*: Return the vector $\hat{X}_0 = \sum_{k=1}^r \hat{\eta}_k \langle \tilde{B}_0, \hat{u}_k \rangle \hat{u}_k$.

4.2 Deriving out-of-sample EBLP

We now derive the out-of-sample EBLP described in Sec. 4.1. Due to the independence between the (Y_0, A_0) and the empirical PCs $\hat{u}_k$, the derivation is much more straightforward than was the in-sample EBLP. Therefore, we present the entire calculation in the main body of the paper.

4.2.1 Covariance of $M^{-1}A_i^\top Y_i$

Let $\tilde{B}_i = M^{-1}A_i^\top Y_i = M^{-1}D_i X_i + M^{-1}A_i^\top \varepsilon_i$, with $X_i = \sum_{j=1}^r \ell_j^{1/2} z_{ij} u_j$ and $D_i = A_i^\top A_i$. Let $R_i = X_i + M^{-1}A_i^\top \varepsilon_i = X_i + \tilde{\varepsilon}_i$; so $\tilde{B}_i = R_i + E_i X_i$, with $E_i = I_p - M^{-1}A_i^\top A_i$.

Observe that

$$\mathrm{Cov}(\tilde{B}_i) = \mathrm{Cov}(R_i) + \mathrm{Cov}(E_i X_i) + \mathbb{E}R_i(E_i X_i)^\top + \mathbb{E}(E_i X_i)^\top R_i$$

and also that

$$\mathbb{E}R_i(E_i X_i)^\top = \mathbb{E}X_i X_i^\top E_i + \mathbb{E}\tilde{\varepsilon}_i X_i^\top E_i = 0$$

since $\mathbb{E}E_i = 0$ and $\mathbb{E}\varepsilon_i = 0$, and they are independent of X_i; similarly $\mathbb{E}(E_i X_i)^\top R_i = 0$ as well. Consequently,

$$\mathrm{Cov}(\tilde{B}_i) = \mathrm{Cov}(R_i) + \mathrm{Cov}(E_i X_i).$$

Let $c_j = \mathbb{E}E_{ij}^2$. Then

$$\mathbb{E}(E_i X_i)(E_i X_i)^\top = \sum_{j=1}^r \ell_j^{1/2} \begin{pmatrix} c_1 u_{j1}^2 & & & \\ & c_2 u_{j2}^2 & & \\ & & \ddots & \\ & & & c_p u_{jp}^2 \end{pmatrix}$$

which goes to zero in operator norm as $n, p \to \infty$, by the incoherence property of the u_k's, and because c_j are uniformly bounded under the assumptions of Theorem 3.1. Therefore $\|\Sigma_{\tilde{B}} - (\Sigma_X + \Sigma_{\tilde{\varepsilon}})\|_{op} \to 0$.

4.2.2 Out-of-sample coefficients and AMSE

We will compute the optimal (in sense of AMSE) coefficients for out-of-sample prediction. We have normalized, back-projected observations $\tilde{B}_i = M^{-1}D_i X_i + \tilde{\varepsilon}_i$, with $X_i = \sum_{j=1}^r \ell_j^{1/2} z_{ij} u_j$ and $\tilde{\varepsilon}_i = M^{-1}A_i^\top \varepsilon_i$.

We are looking for the coefficients $\eta_1, \ldots, \eta_r$ so that the estimator

$$\hat{X}_0^\eta = \sum_{j=1}^r \eta_j \langle \tilde{B}_0, \hat{u}_j \rangle \hat{u}_j \tag{9}$$

has minimal AMSE. Here, $\hat{u}_j$ are the empirical PCs based on the in-sample data $(Y_1, A_1), \ldots, (Y_n, A_1)$ (that is, the top r eigenvectors of $\sum_{j=1}^n \tilde{B}_i \tilde{B}_i^\top$), whereas (Y_0, A_0) is an out-of-sample datapoint.

It is easily shown that the contribution of η_k to the overall MSE is:

$$\ell_k + \eta_k^2 \mathbb{E}(\hat{u}_k^\top \tilde{B}_0)^2 - 2\eta_k \ell_k^{1/2} \mathbb{E}z_{0k}(\hat{u}_k^\top \tilde{B}_0)(\hat{u}_k^\top u_k).$$

It is also easy to see that the interaction terms obtained when expanding the MSE vanish.

To evaluate the quadratic coefficient above, first take the expectation over Y_0 and A_0 only, which gives:

$$\mathbb{E}_0(\hat{u}_k^\top \tilde{B}_0)^2 = \hat{u}_k^\top \Sigma_{\tilde{B}} \hat{u}_k \sim \hat{u}_k^\top \left(\sum_{j=1}^r \ell_j u_j u_j^\top + \Sigma_{\tilde{\varepsilon}} \right) \hat{u}_k$$

$$\sim \ell_k c_k^2 + \hat{u}_k^\top \Sigma_{\tilde{\varepsilon}} \hat{u}_k$$

Note that when the original noise ε_i is white (i.e. $\Sigma_\varepsilon = I_p$), we can estimate $d_k \equiv \hat{u}_k^\top \Sigma_{\tilde{\varepsilon}} \hat{u}_k$ using the approximation $\Sigma_{\tilde{\varepsilon}} \sim M^{-1}$, as in Sec. 3.3.1. Defining the estimator $\hat{d}_k = \hat{u}_k^\top M^{-1} \hat{u}_k$ (or $\hat{u}_k^\top \hat{M}^{-1} \hat{u}_k$, where $\hat{M} = \sum_{i=1}^n A_i^\top A_i / n$), we therefore have $|\hat{d}_k - d_k| \to 0$.

Now turn to the linear term. We have $\hat{u}_k^\top \tilde{B}_0 = \sum_{j=1}^r \ell_j^{1/2} z_{0j} \hat{u}_k^\top M^{-1} D_0 u_j + \hat{u}_k^\top \varepsilon_0$; using $\mathbb{E}[M^{-1} D_0] = I_p$ and using the almost sure convergence results, it follows after some simple calculation that $\ell_k^{1/2} \mathbb{E}[z_{0k} \hat{u}_k^\top \tilde{B}_0 \hat{u}_k^\top u_k] \to \ell_k c_k^2$. Consequently, the mean-squared error of the out-of-sample predictor (as a function of η_k) is asyptotically equivalent to:

$$\sum_{k=1}^r \left\{ \ell_k + \eta_k^2 (\ell_k c_k^2 + d_k) - 2\eta_k \ell_k c_k^2 \right\}.$$

This is minimized at $\eta_k^* = \frac{\ell_k c_k^2}{\ell_k c_k^2 + d_k}$ and the MSE is asymptotically equivalent to:

$$\sum_{k=1}^r \left(\ell_k - \frac{\ell_k^2 c_k^4}{\ell_k c_k^2 + d_k} \right).$$

This finishes the derivation of the optimal coefficients for out-of-sample prediction.

4.3 The whitened model

Following the approach described in Sec. 3.3, we can optimally predict X_0 using the W-loss, for any positive semi-definite matrix W. This is equivalent to performing optimal prediction of the signal WX_0 based on the observations $W\tilde{B}_0 = WM^{-1}D_0 X_0 + W\tilde{\varepsilon}_0$ in the usual Frobenius sense.

We can always transform the data so that the effective noise $W\tilde{\varepsilon} = WM^{-1}A_0^\top \tilde{\varepsilon}_0$ has identity covariance; that is, take $W = \Sigma_{\tilde{\varepsilon}}^{-1/2}$.

In this setting, the parameters $\hat{u}_k^\top W \Sigma_{\tilde{\varepsilon}}^{-1/2} W \hat{u}_k = \hat{u}_k^\top \hat{u}_k = 1$, and so $d_k = 1$. Consequently, the limiting AMSE is

$$\sum_{k=1}^r \left(\tilde{\ell}_k - \frac{\tilde{\ell}_k^2 c_k^4}{\tilde{\ell}_k c_k^2 + 1} \right) \tag{10}$$

where $\tilde{\ell}_k$ are the eigenvalues of the whitened model WX_i, assuming the model (6). Using the formulas (7) and (8) for c_k and $\tilde{c}_k$ as functions of $\tilde{\ell}_k$, it is straightforward to check that formula (10) is equal to $\sum_{k=1}^r \tilde{\ell}_k (1 - c_k^2 \tilde{c}_k^2)$, which is the in-sample AMSE with W-loss; we will show this in Sec. 7.12. That is, the AMSE for whitened observations are identical for in-sample and out-of-sample EBLP.

Thus, we state the following theorem:

17

Theorem 4.1 (Out-of-sample EBLP)**.** *Suppose our observations have the form $Y_i = A_i X_i + \varepsilon_i$, $i = 1, \ldots, n$, under the conditions of Thm. 3.1, and suppose in addition that (6) holds, with $W = \Sigma_{\tilde{\varepsilon}}^{-1/2}$ and $\tilde{\varepsilon}_i = M^{-1} A_i^\top \varepsilon_i$.*

Given an out-of-sample observation Y_0, A_0, consider a predictor of X_0 of the form (9). Then, for the optimal choice of η_k, the minimum asymptotic out-of-sample MSE achieved by this predictor in $\Sigma_{\tilde{\varepsilon}}^{-1/2}$-norm equals the corresponding expression for in-sample MSE.

Thus, asymptotically, out-of-sample denoising is not harder than in-sample denoising.

The remainder of the proof of Thm. 4.1 is contained in Sec. 7.12.

5 Matrix denoising and missing data

A well-studied problem to which our analysis applies is the problem of missing data, where coordinates are discarded from the observed vectors. Here the operators $D_i = A_i^\top A_i$ place zeros in the unobserved entries.

Without additive noise, recovering the matrix $X = [X_1, \ldots, X_n]^\top$ is known as *matrix completion*, and has been widely studied in statistics and signal processing. There are many methods with guarantees of exact recovery for certain classes of signals (Candès and Recht, 2009; Candès and Tao, 2010; Jain et al., 2013; Keshavan et al., 2010; Recht, 2011; Jain et al., 2013).

Many methods for matrix completion assume that the target matrix X is low-rank. This is the case for the linearly-transformed model as well, since the rows $X_i^\top$ of X all lie in the r-dimensional subspace spanned by $u_1, \ldots, u_r$. In the linearly-transformed model, the low-rank target matrix X is itself random, and the analysis we provide for the performance of EBLP is dependent on this random structure.

Our approach differs from most existing methods. Our methods have the following advantages:

1. **Speed.** Typical methods for matrix completion are based on solving optimization problems such as nuclear norm minimization (Candès and Recht, 2009; Candès and Tao, 2010). These require iterative algorithms, where an SVD is computed at each step. In contrast, when an upper bound on the rank of the target matrix is known a priori our methods require only one SVD, and are thus much faster. Some of the methods for rank estimation in the spiked model discussed in Sec. 3.4, such as Dobriban and Owen (2017) and Kritchman and Nadler (2008), require only one SVD as well; we believe that these methods can be adapted to the linearly-transformed spiked model, though this is outside the scope of the current paper.

2. **Robustness to high levels of noise.** Most matrix completion methods have guarantees of numerical stability: when the observed entries are accurate to a certain precision, the output will be accurate to almost the same precision. However, when the noise level swamps the signal, these stability guarantees are not informative. While many matrix completion methods can be made more robust by incorporating noise regularization, EBLP is designed to directly handle the high-noise regime. In Sec. 5.1, we show that our method is more robust to noise than regularized nuclear norm minimization.

3. **Applicability to uneven sampling.** While many matrix completion methods assume that the entries are observed with equal probability, other methods allow for uneven sampling across the rows and columns. Our method of EBLP allows for a different probability in each column of X. In Sec. 5.1.2 we compare our method to competing methods when the column sampling probabilities exhibit varying degrees of non-uniformity. In particular, we compare to the OptShrink method for noisy matrix completion (Nadakuditi, 2014), which is nearly identical to EBLP when the sampling is uniform, but is not designed for uneven sampling. We also compare to *weighted* nuclear norm minimization, designed to handle the uneven sampling.

4. **Precise performance guarantees.** Our shrinkage methods have precise asymptotic performance guarantees for their mean squared error. The errors can be estimated from the observations themselves.

In addition to these advantages, our method has the seeming shortcoming that unlike many algorithms for matrix completion, it never yields exact recovery. However, our methods lead to *consistent* estimators in the low-noise regime. In our model low noise corresponds to large spikes ℓ. It is easy to see that taking $\ell \to \infty$ we obtain an asymptotic MSE of $\mathbb{E}\|X_i - \hat{X}_i\|^2 = O(1)$, whereas $\mathbb{E}\|X_i\|^2 = \ell$. Thus the correlation $\mathrm{corr}(\hat{X}_i, X_i) \to 1$ in probability, and we get consistent estimators. Thus we still have good performance in low noise.

5.1 Simulations

In this section, we illustrate the finite-sample properties of our proposed EBLP with noise whitening. We compare this method to three other methods found in the literature. First is the OptSpace method of Keshavan et al. (2010). This algorithm is designed for uniform sampling of the matrix and relatively low noise levels, although a regularized version for larger noise has been proposed as well (Keshavan and Montanari, 2010). As we will see, OptSpace (without regularization) typically performs well in the low-noise regime, but breaks down when the noise is too high. We use the MATLAB code provided by Sewoong Oh on his website `http://swoh.web.engr.illinois.edu/software/optspace/code.html`. We note that, like EBLP, OptSpace makes use of a user-provided rank.

The second method is nuclear norm-regularized least squares (NNRLS), as described in Candès and Plan (2010). In the case of uniform sampling, we minimize the loss function $\frac{1}{2}\|X_\Omega - Y_\Omega\|^2 + w \cdot \|X\|_*$, where $\|\cdot\|_*$ denotes the nuclear norm and X_Ω denotes the vector of X's values on the set of observed entries Ω. Following the recommendation in Candès and Plan (2010) we take w to be the operator norm of the pure subsampled noise term; that is, $w = \|E_\Omega\|$, where E is the matrix of noise. With this choice of parameter, when the input data is indistinguishable from pure noise the estimator returned is the zero matrix. When the noise is white noise with variance σ^2, then $w = \sigma(\sqrt{p} + \sqrt{n})\sqrt{|\Omega|/(pn)}$ at noise variance σ^2. If the noise is colored, we determine w by simulation; we note that the Spectrode method of Dobriban (2015) might offer an alternative means of determining w. To solve the minimization, we use the accelerated gradient method of Ji and Ye (2009).

When the sampling probabilities differ across the columns of X, we compare to a weighted nuclear norm minimization. This minimizes the loss function $\frac{1}{2}\|X_\Omega - Y_\Omega\|^2 +$

$w \cdot \|XC_i\|_*$, where C is the diagonal matrix with entries $C_{ii} = \sqrt{p_i}$, and p_i is the probability that column i is sampled. Again, we choose w so that if there is no signal (i.e. $X = 0$), then the zero matrix is returned. This method has been widely studied (Srebro and Salakhutdinov, 2010; Negahban and Wainwright, 2011; Klopp, 2014; Chen et al., 2015).

The third method is OptShrink (Nadakuditi, 2014). OptShrink assumes the sampling of the matrix is uniform; when this is the case, the method is essentially identical to EBLP without whitening. However, for non-uniform sampling we find the EBLP outperforms OptShrink, especially as the noise level increases. In Sec. 5.1.3, we also compare EBLP with whitening to OptShrink (which does not perform whitening) with colored noise; we find that whitening improves performance as the overall noise level increases. When using EBLP and OptShrink with data that is not mean zero, we estimate the mean using the available-case estimator, and subtract it before shrinkage.

In Sec. 5.1.4, we compare in-sample and out-of-sample EBLP. We demonstrate a very good agreement between the RMSEs, as predicted by Thm. 4.1, especially at high sampling rates.

In Secs. 5.1.1, 5.1.2 and 5.1.3, we used the following experimental protocol. The signals X_i are drawn from a rank 10 model, with eigenvalues $1, 2, \ldots, 10$, and random mean. Except for Sec. 5.1.1, the PCs $u_1, \ldots, u_{10}$ were chosen to span a completely random 10-dimensional subspace of $\mathbb{R}^{300}$. We used the aspect ratio $\gamma = 0.8$, corresponding to a sample size of $n = 375$. The random variables z_{ik} were taken to be Gaussian, as was the additive noise. The matrices A_i are random coordinate selection operators, with each coordinate chosen with a given probability. When each entry of the matrix has probability δ of being selected, we will call δ the *sampling rate*.

We measure the accuracy of a predictor $\hat{X}$ of the matrix X using the root mean squared error, defined by $\|\hat{X} - X\|_F / \|X\|_F$. For each experiment, we plot the RMSEs of the different algorithms for forty runs of the experiment at increasing noise levels σ. The code for these experiments, as well as a suite of MATLAB codes for singular value shrinkage and EBLP, can be found online at `https://github.com/wleeb/opt-pred`.

5.1.1 Sparsity of the PCs

We compare the matrix completion algorithms when the PCs $u_1, \ldots, u_{10}$ have different amounts of sparsity. We say that a vector is m-sparse if only m coordinates are non-zero; we consider the cases where all the PCs are 10-sparse, $p/4$-sparse, $p/2$-sparse, and dense. We show the results in Fig. 3. Note that EBLP outperforms OptSpace and NNRLS at high noise levels, while it does worse than OptSpace at low noise levels in all sparsity regimes, and worse than both competing methods at low noise levels when the PCs are sparse.

5.1.2 Uneven sampling

In this experiment, each coordinate is assigned a different probability of being selected, where the probabilities range linearly from δ to $1 - \delta$ for $\delta \in (0, 1)$. In addition to NNRLS and OptSpace, we also compare EBLP to OptShrink (Nadakuditi, 2014), which assumes uniform sampling. With uniform sampling, the two procedures are nearly identical. However, EBLP performs better when the sampling is non-uniform.
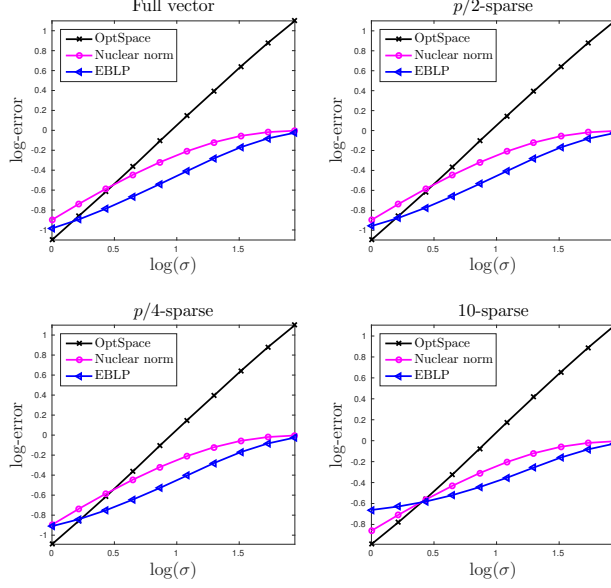
20

Figure 2: Log-RMSEs against log-noise for matrix completion. Each plot shows a different amount of sparsity in the PCs $u_1, \ldots, u_{10}$.
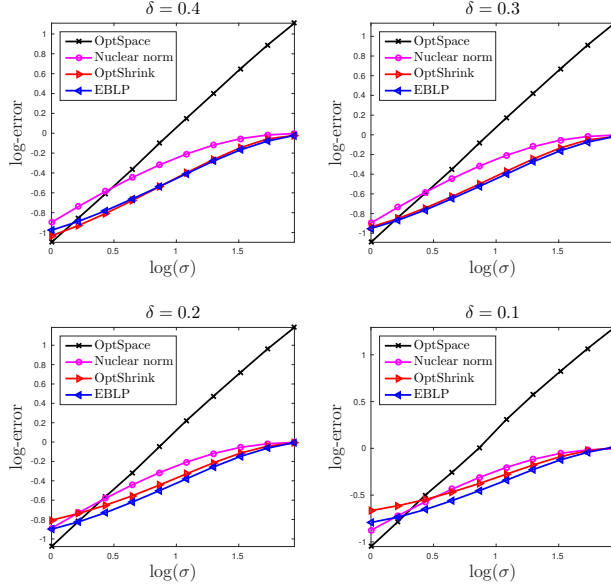

Figure 3: Log-RMSEs against log-noise for matrix completion. Each plot shows a different unevenness of sampling across the coordinates, with sampling probabilities ranging linearly from δ to $1 - \delta$.

5.1.3 Colored noise

We use colored noise whose covariance has condition number $\kappa > 1$. The noise covariance's eigenvalues increase linearly with the coordinates while having overall norm $p = 300$. In each experiment the noise is then multiplied by σ to increase the overall variance of the noise while maintaining the condition number. We subsample uniformly with probability 0.5. Again, we compare EBLP with whitening to NNRLS, OptSpace,

and OptShrink (which does not whiten). We observe that at high noise levels, EBLP with whitening outperforms OptShrink, while OptShrink performs better at low noise levels; and this effect increases with larger κ.
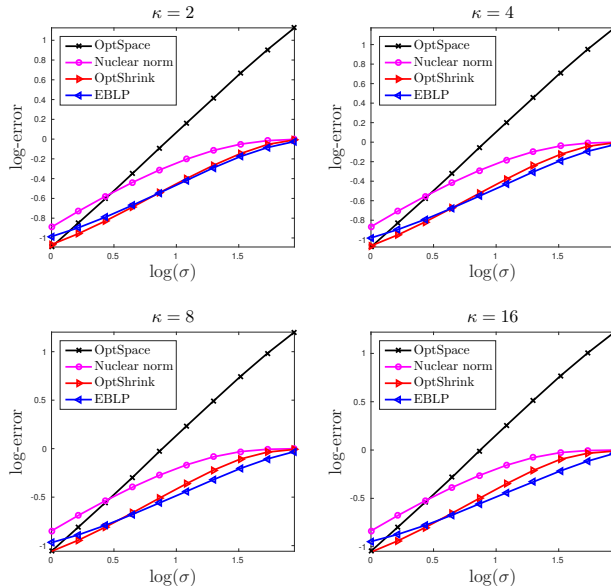

Figure 4: Log-RMSEs against log-noise for matrix completion. Each plot shows a different condition number κ of the noise covariance matrix, reflecting different amounts of heterogeneity in the noise.

5.1.4 In-sample vs. out-of-sample EBLP

In this experiment, we compare the performance of in-sample and out-of-sample EBLP. Thm. 4.1 predicts that asymptotically, the MSE of the two methods are identical. We illustrate this result in the finite-sample setting.

We fixed a dimension $p = 500$ and sampling rate δ, and generated random values of $n > p$ and $\ell > 0$. For each set of values, we randomly generated two rank 1 signal matrices of size n-by-p, X_{in} and X_{out}, added Gaussian noise, and subsampled these matrices uniformly at rate δ to obtain the backprojected observations $\tilde{B}_{in}$ and $\tilde{B}_{out}$. We apply the in-sample EBLP on $\tilde{B}_{in}$ to obtain $\hat{X}_{in}$, and using the singular vectors of $\tilde{B}_{in}$, we apply the out-of-sample EBLP to $\tilde{B}_{out}$ to obtain $\hat{X}_{out}$.

In Fig. 5, we show scatterplots of the RMSEs for the in-sample and out-of-sample data for each value of n and ℓ. We also plot the line $x = y$ for reference. The errors of in-sample and out-of-sample EBLP are very close to each other, though the finite sample effects are more prominent for small δ.

6 Conclusion

In this paper we considered the linearly transformed spiked model, and developed asymptotically optimal EBLP methods for predicting the unobserved signals in the commutative case of the model, under high-dimensional asympotics. For missing data,
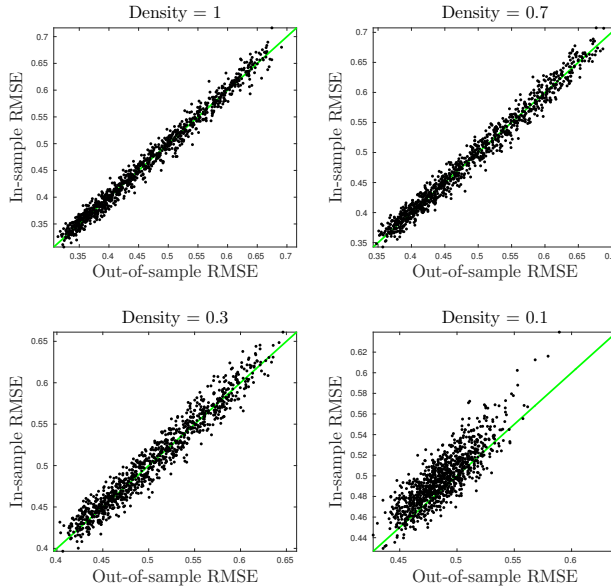
Figure 5: Scatterplots of the RMSEs of in-sample EBLP against out-of-sample EBLP for different sampling densities.

we showed in simulations that our methods are faster, more robust to noise and to unequal sampling than well-known matrix completion methods.

There are many exciting opportunities for future research. One problem is to extend our methods beyond the commutative case. This is challenging because the asymptotic spectrum of the backprojected matrix B becomes harder to characterize, and new proof methods are needed. Another problem is to understand the possible benefits of whitening. We saw that whitening enables fast optimal shrinkage, but understanding when it leads to improved denoising remains an open problem.

Acknowledgements

7 Proofs

7.1 Proof of Thm. 3.1

We present the proof of Thm. 3.1 for the backprojected matrix B. The proof for the normalized matrix $\tilde{B}$ is identical and omitted for brevity. The proof spans multiple sections, until Sec. 7.7. To illustrate the idea, we first prove the single-spiked case, when $r = 1$. The proof of the multispiked extension is provided in Sec. 7.6.

Since $A_i^\top A_i = M + E_i$, we have the following decomposition for the backprojected observations B_i:

$$B_i = A_i^\top Y_i = MX_i + A_i^\top \varepsilon_i + E_i X_i.$$

The first key observation is that after backprojection, we still have an *approximate* spiked model. The new signal component is MX_i, the new noise is $A_i^\top \varepsilon_i$. The error term $E_i X_i$ has an asymptotically negligible contribution in operator norm, as shown below. The proof is provided in Sec. 7.2.

Lemma 7.1. *Let E^* be the matrix formed by the vectors $n^{-1/2} E_i X_i$. Then the operator norm $\|E^*\| \to 0$ a.s.*

Since the claims of our theorem concern the limits of the spectral distribution, singular values, and angles between population and sample singular vectors, all of which are continuous functions with respect to the operator norm, it follows that we can equivalently prove the theorem for

$$B_i^* = \ell^{1/2} z_i \cdot Mu + A_i^\top \varepsilon_i.$$

Let us denote by $\nu = Mu/\xi^{1/2}$ the normalized backprojected signal, where $\xi = |Mu|^2 \to \tau$. We will extend the technique of Benaych-Georges and Nadakuditi (2012) to characterize the spiked eigenvalues in this model. We denote the normalized vector $Z = n^{-1/2}\tilde{Z}$, with $\tilde{Z} = (z_1, \ldots, z_n)^\top$, the normalized noise $N = n^{-1/2}\mathcal{E}^*$, where $\mathcal{E}^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)^\top = (A_1^\top \varepsilon_1, \ldots, A_n^\top \varepsilon_n)^\top$ and the normalized backprojected data matrix $\tilde{B}^* = n^{-1/2}B^*$, where $B^* = (B_1^*, \ldots, B_n^*)^\top$. Then, our model is

$$\tilde{B}^* = (\xi\ell)^{1/2} \cdot Z\nu^\top + N. \tag{11}$$

We will assume that $n, p \to \infty$ such that $p/n \to \gamma > 0$. For simplicity of notation, we will first assume that $n \leq p$, implying that $\gamma \geq 1$. It is easy to see that everything works when $n \geq p$.

By Lemma 4.1 of Benaych-Georges and Nadakuditi (2012), the singular values of $\tilde{B}^*$ that are not singular values of N are the positive reals t such that the 2-by-2 matrix

$$M_n(t) = \begin{bmatrix} t \cdot Z^\top(t^2 I_n - NN^\top)^{-1}Z & Z^\top(t^2 I_n - NN^\top)^{-1}N\nu \\ \nu^\top N^\top(t^2 I_n - NN^\top)^{-1}Z & t \cdot \nu^\top(t^2 I_p - N^\top N)^{-1}\nu \end{bmatrix} - \begin{bmatrix} 0 & (\xi\ell)^{-1/2} \\ (\xi\ell)^{-1/2} & 0 \end{bmatrix}$$

is not invertible, i.e., $\det[M_n(t)] = 0$. We will find almost sure limits of the entries of $M_n(t)$, to show that it converges to a deterministic matrix $M(t)$. Solving the equation $\det[M(t)] = 0$ will provide an equation for the almost sure limit of the spiked singular values of $\tilde{B}^*$. For this we will prove the following results:

Lemma 7.2 (The noise matrix). *The noise matrix N has the following properties:*

24

1. *The eigenvalue distribution of $N^\top N$ converges almost surely (a.s.) to the Marchenko-Pastur distribution $F_{\gamma,H}$ with aspect ratio $\gamma \geq 1$.*

2. *The top eigenvalue of $N^\top N$ converges a.s. to the upper edge b_H^2 of the support of $F_{\gamma,H}$.*

This is proved in Sec. 7.3. For brevity we write $b = b_H$. Since $\tilde{B}^*$ is a rank-one perturbation of N, it follows that the eigenvalue distribution of $\tilde{B}^{*\top}\tilde{B}^*$ also converges to the MP law $F_{\gamma,H}$. This proves the first claim of Thm 3.1.

Moreover, since $NN^\top$ has the same n eigenvalues as the nonzero eigenvalues of $N^\top N$, the two facts in Lemma 7.2 imply that when $t > b$, $n^{-1}\operatorname{tr}(t^2 I_n - NN^\top)^{-1} \to \int (t^2 - x)^{-1} d\underline{F}_{\gamma,H}(x) = -\underline{m}(t^2)$. Here $\underline{F}_{\gamma,H}(x) = \gamma F_{\gamma,H}(x) + (1 - \gamma)\delta_0$ and $\underline{m} = \underline{m}_{\gamma,H}$ is the Stieltjes transform of $\underline{F}_{\gamma,H}$. Clearly this convergence is uniform in t. As a special note, when t is a singular value of the random matrix N, we formally define $(t^2 I_p - N^\top N)^{-1} = 0$ and $(t^2 I_n - NN^\top)^{-1} = 0$. When $t > b$, the complement of this event happens a.s. In fact, from Lemma 7.2 it follows that $(t^2 I_p - N^\top N)^{-1}$ has a.s. bounded operator norm. Next we control the quadratic forms in the matrix M_n.

Lemma 7.3 (The quadratic forms). *When $t > b$, the quadratic forms in the matrix $M_n(t)$ have the following properties:*

1. $Z^\top(t^2 I_n - NN^\top)^{-1}Z - n^{-1}\operatorname{tr}(t^2 I_n - NN^\top)^{-1} \to 0$ *a.s.*

2. $Z^\top(t^2 I_n - NN^\top)^{-1}N\nu \to 0$ *a.s.*

3. $\nu^\top(t^2 I_p - N^\top N)^{-1}\nu \to -m(t^2)$ *a.s., where $m = m_{\gamma,H}$ is the Stieltjes transform of the Marchenko-Pastur distribution $F_{\gamma,H}$.*

Moreover the convergence of all three terms is uniform in $t > b + c$, for any $c > 0$.

This is proved in Sec. 7.4. The key technical innovation is the proof of the third part. Most results for controlling quadratic forms $x^\top A x$ are concentration bounds for random x. Here $x = \nu$ is fixed, and the matrix $A = (t^2 I_p - N^\top N)^{-1}$ is random instead. For this reason we adopt the "deterministic equivalents" technique of Bai et al. (2007) for quantities $x^\top(z I_p - N^\top N)^{-1}x$, with the key novelty that we can take the imaginary part of the complex argument to zero. The latter observation is nontrivial, and mirrors similar techniques used recently in universality proofs in random matrix theory (see e.g., the review by Erdős and Yau, 2012).

Lemmas 7.2 and 7.3 will imply that for $t > b$, the limit of $M_n(t)$ is

$$M(t) = \begin{bmatrix} -t \cdot \underline{m}(t^2) & -(\tau\ell)^{-1/2} \\ -(\tau\ell)^{-1/2} & -t \cdot m(t^2) \end{bmatrix}.$$

By the Weyl inequality, $\sigma_2(\tilde{B}^*) \leq \sigma_2((\xi\ell)^{1/2} \cdot Z\nu^\top) + \sigma_1(N) = \sigma_1(N)$. Since $\sigma_1(N) \to b$ a.s. by Lemma 7.2, we obtain that $\sigma_2(\tilde{B}^*) \to b$ a.s. Therefore for any $\varepsilon > 0$, a.s. only $\sigma_1(\tilde{B}^*)$ can be a singular value of $\tilde{B}^*$ in $(b+\varepsilon, \infty)$ that is not a singular value of N.

It is easy to check that $D(x) = x \cdot \underline{m}(x)m(x)$ is strictly decreasing on (b^2, ∞). Hence, denoting $h = \lim_{t\downarrow b} 1/D(t^2)$, for $\tau\ell > h$, the equation $D(t^2) = 1/(\tau\ell)$ has a unique solution $t \in (b, \infty)$. By Lemma A.1 of Benaych-Georges and Nadakuditi (2012), we conclude that for $\tau\ell > h$, $\sigma_1(\tilde{B}^*) \to t$ a.s., where t solves the equation $\det[M(t)] = 0$, or equivalently,

$$t^2 \cdot \underline{m}(t^2)m(t^2) = \frac{1}{\tau\ell}.$$

If $\tau\ell \leq h$, then we note that $\det[M_n(t)] \to \det[M(t)]$ uniformly on $t > b+\varepsilon$. Therefore, if $\det[M_n(t)]$ had a root $\sigma_1(\tilde{B}^*)$ in $(b+\varepsilon, \infty)$, $\det[M(t)]$ would also need to have a root there, which is a contradiction. Therefore, we conclude $\sigma_1(\tilde{B}^*) \leq b + \varepsilon$ a.s., for any $\varepsilon > 0$. Since $\sigma_1(\tilde{B}^*) \geq \sigma_2(\tilde{B}^*) \to b$, we conclude that $\sigma_1(\tilde{B}^*) \to b$ a.s., as desired. This finishes the spike convergence claim in Thm. 3.1.

Next, we turn to proving the convergence of the angles between the population and sample eigenvectors. Let $\hat{Z}$ and $\hat{u}$ be the singular vectors associated with the top singular value $\sigma_1(\tilde{B}^*)$ of $\tilde{B}^*$. Then, by Lemma 5.1 of Benaych-Georges and Nadakuditi (2012), if $\sigma_1(\tilde{B}^*)$ is not a singular value of X, then the vector $\eta = (\eta_1, \eta_2) = (\nu^\top \hat{u}, Z^\top \hat{Z})$ belongs to the kernel of the matrix $M_n(\sigma_1(\tilde{B}^*))$. By the above discussion, this 2-by-2 matrix is of course singular, so this provides one linear equation for the vector η (with $R = (t^2 I_n - NN^\top)^{-1}$)

$$t\eta_1 \cdot Z^\top R Z + \eta_2[Z^\top R N\nu - (\xi\ell)^{-1/2}] = 0.$$

By the same lemma cited above, it follows that we have the *norm identity* (with $t = \sigma_1(\tilde{B}^*)$)

$$t^2\eta_1^2 \cdot Z^\top R^2 Z + \eta_2^2 \cdot \nu^\top N^\top R^2 N\nu + 2t\eta_1\eta_2 \cdot Z^\top R^2 N\nu = (\xi\ell)^{-1}. \qquad (12)$$

This follows from taking the norm of the equation $t\eta_1 \cdot RZ + \eta_2 \cdot RN\nu = (\xi\ell)^{-1/2}\hat{Z}$ (see Lemma 5.1 in Benaych-Georges and Nadakuditi (2012)). We will find the limits of the quadratic forms below.

Lemma 7.4 (More quadratic forms). *The quadratic forms in the norm identity have the following properties:*

1. $Z^\top(t^2 I_n - NN^\top)^{-2}Z - n^{-1}\operatorname{tr}(t^2 I_n - NN^\top)^{-2} \to 0$ *a.s.*

2. $Z^\top(t^2 I_n - NN^\top)^{-2}N\nu \to 0$ *a.s.*

3. $\nu^\top N^\top(t^2 I_n - NN^\top)^{-2}N\nu \to m(t^2) + t^2 m'(t^2)$ *a.s., where m is the Stieltjes transform of the Marchenko-Pastur distribution $F_{\gamma,H}$.*

The proof is in Sec. 7.5. Again, the key novelty is the proof of the third claim. The standard concentration bounds do not apply, because u is non-random. Instead, we use an argument from complex analysis constructing a sequence of functions $f_n(t)$ such that their derivatives are $f_n'(t) = \nu^\top N^\top(t^2 I_n - NN^\top)^{-2}N\nu$, and deducing the convergence of $f_n'(t)$ from that of $f_n(t)$.

Lemma 7.4 implies that $n^{-1}\operatorname{tr}(t^2 I_n - NN^\top)^{-2} \to \int(t^2-x)^{-2}d\underline{F}_{\gamma,H}(x) = \underline{m}'(t^2)$ for $t > b$. Solving for η_1 in terms of η_2 from the first equation, plugging in to the second, and taking the limit as $n \to \infty$, we obtain that $\eta_2^2 \to c_2$, where

$$c_2\left(\frac{m'(t^2)}{\tau\ell\underline{m}(t^2)^2} + m(t^2) + t^2 m'(t^2)\right) = \frac{1}{\tau\ell}.$$

Using $D(x) = x \cdot m(x)\underline{m}(x)$, we find $c_2 = \underline{m}(t^2)/[D'(t^2)\tau\ell]$, where t solves (3). From the first equation, we then obtain $\eta_1^2 \to c_1$, where $c_1 = m(t^2)/[D'(t^2)\tau\ell]$, where t is as above. This finishes the proof of Thm. 3.1 in the single-spiked case. The proof of the multispiked case is a relatively simple extension of the previous argument, so we present it in Sec. 7.6.

7.2 Proof of Lemma 7.1

Since $X_i = \ell^{1/2} z_i u$, the (i,j)-th entry of E^* is $n^{-1/2} \ell^{1/2} z_i \cdot E_{ij} u(j)$. Now, denoting by $\odot$ elementwise products

$$\|E^*\| = \sup_{\|a\|=\|c\|=1} a^\top E^* c = n^{-1/2} \ell^{1/2} \sup_{\|a\|=\|c\|=1} (a \odot z)^\top E(c \odot u).$$

We have $\|a \odot z\| \leq \|a\| \max |z_i| = \max |z_i|$ and $\|c \odot u\| \leq \|c\| \max |u_i| = \max |u_i|$, hence

$$\|E^*\| \leq \ell^{1/2} \cdot \|n^{-1/2} E\| \max |z_i| \max |u_i|.$$

Below we will argue that $\|n^{-1/2} E\| \leq C$, a.s., so that the the operator norm is a.s. bounded. (The constant C can change from line to line.) Once we have established that, we have that, almost surely,

$$\|E^*\| \leq C \|z\|_\infty \|u\|_\infty.$$

Then, our main claim follows as long as

$$\|z\|_\infty \|u\|_\infty \to_{a.s.} 0. \tag{13}$$

This holds under several possible sets of assumptions:

1. **Polynomial moment assumption.** Suppose $\mathbb{E}|z_i|^m \leq C$ for some $m > 0$ and $C < \infty$. Then, since z has iid standardized entries, we can derive that

$$Pr(\max |z_i| \geq a) \leq \mathbb{E} \max |z_i|^m / a^m \leq nC/a^m.$$

To ensure that these probabilities are summable, we take m such that $n/a^m = C/n^{1+\phi'}$ for some small $\phi' > 0$. This is equivalent to $a = Cn^{(2+\phi')/m}$. It follows that $Pr(\max |z_i| \leq Cn^{(2+\phi')/m})$ a.s., for any $\phi' > 0$.

Therefore, the required delocalization condition on u is that

$$\|u\|_\infty \cdot n^{(2+\phi')/m} \to_{a.s} 0.$$

Therefore, it is enough to assume that

$$\|u\|_\infty \leq Cn^{-(2/m+c')},$$

for any small constant $c' > 0$. Since $\|u\| = 1$, we have that $\|u\|_\infty \geq n^{-1/2}$, so this is only feasible for $m > 4$. Since n is proportional to p, we can replace n by p in our assumption.

2. **Sub-gaussian assumption.** Suppose the z_i are sub-gaussian in the sense that $\mathbb{E} \exp(t|z_i|^2) \leq C$ for some $t > 0$ and $C < \infty$ (e.g., Vershynin, 2010). Then, since z has iid standardized entries, we can derive that

$$Pr(\max |z_i| \geq a) \leq \mathbb{E} \max \exp(t|z_i|^2) / \exp(ta^2) \leq nC / \exp(ta^2).$$

To ensure that these probabilities are summable, we take m such that $n/\exp(ta^2) = C/n^{1+\phi'}$ for some small $\phi' > 0$. This is equivalent to

$$a = \sqrt{\frac{(2 + \phi') \log n}{t}}.$$

Ignoring constants that do not depend on n, we can say that $a = C\sqrt{\log n}$. Since n is proportional to p, we can replace n by p in our assumption.

It follows that $Pr(\max |z_i| \leq C\sqrt{\log n})$ a.s., for some $C > 0$.

Therefore, the required delocalization condition on u is that

$$\|u\|_\infty \cdot \sqrt{\log n} \to_{a.s} 0.$$

It remains to show that $\|n^{-1/2}E\|$ is a.s. bounded. We will see below that for this it is enough that the $6 + \varepsilon$-th moment of all E_{ij} is uniformly bounded for some $\varepsilon > 0$. We will only give a proof sketch, as this consists of a small modification of Thm. 6.3 of Bai and Silverstein (2009). Their proof essentially goes through, except that one needs a slightly different, simpler, argument for the initial truncation step.

For this, we follow the same steps as the original proof of Cor. 6.6 from Bai and Silverstein (2009). The first step (on p. 128 of Bai and Silverstein (2009)) is a truncation, which relies on Thm. 5.8, a result controlling the operator norm of matrices with iid entries. The proof of this result is not provided in the book, since it is almost the same as Thm. 5.1. Therefore, we will show how to adapt the argument for Thm. 5.1 to our assumptions.

The first step of that proof (on p. 94) is a truncation of the entries of E at a threshold $cn^{1/2}$. Let $\tilde{E}_{ij} = E_{ij}I(|E_{ij}| \leq cn^{1/2})$, and let $\tilde{E}_{ij}$ the corresponding matrix. We need to show that $P(E \neq \tilde{E}, i.o.) = 0$. Let $A_n = \bigcup_{(i,j)\leq(n,p)} I(|E_{ij}| \geq cn^{1/2})$. Then $\{E \neq \tilde{E}\} = A_n$. By the Borel-Cantelli lemma, it is enough to show:

$$\sum_n P[A_n] = \sum_n P\left[\bigcup_{(i,j)\leq(n,p)} I\left(|E_{ij}| \geq cn^{1/2}\right)\right] < \infty.$$

We can bound

$$P[A_n] \leq np \cdot \max_{i,j} P\left(|E_{ij}| \geq cn^{1/2}\right) \leq np \cdot \max_{i,j} \mathbb{E}|E_{ij}|^k/[c^k n^{k/2}].$$

By taking the exponent $k = 6 + \varepsilon$, we see from the assumption $\mathbb{E}E_{ij}^{6+\varepsilon} < C$ that this bound is summable. Thus the first step of the proof of Thm. 5.1 adapts to our setting. Then, similarly to remark 5.7 on p.104, we obtain Thm. 5.8 under the present conditions. This shows that the first step of the proof of Cor. 6.6 goes through.

Continuing with the proof of Cor. 6.6 under the present conditions, we need on p. 129 that the conditions of Thm 5.9 be met. It is immediate to see that they are met. In the remaining part of the argument, as stated on p. 129, "no assumptions [beyond independence and moment assumptions] need to be made on the relationship between the E-s for different n." Since our assumptions ensure that after truncation we are in the same setting as that stated in Thm. 6.3, including the independence and moment assumptions, the rest of the argument applies verbatim. This finishes the proof of the sufficiency of finite $6 + \varepsilon$ moment for E. This also finishes the argument that $\|n^{-1/2}E\|$ is a.s. bounded, and thus that of the main lemma.

7.3 Proof of Lemma 7.2

Recall that $N = n^{-1/2}\mathcal{E}^*$, where $\mathcal{E}^*$ has rows $\varepsilon_i^* = A_i^\top \varepsilon_i$. According to our assumptions, the rows are iid, with independent entries having a distribution of variances H_p. Recall

28

that we assumed that $H_p \Rightarrow H$. Hence the eigenvalue distribution of $N^\top N$ converges to the Marchenko-Pastur distribution $F_{\gamma,H}$. This follows essentially by Thm. 4.3 of Bai and Silverstein (2009). While that result is stated for identically distributed random variables, it is easy to see that under our higher moment assumptions this is not needed, as the truncation steps on p. 70 of Bai and Silverstein (2009) go through; the situation is analogous to the modification of (Bai and Silverstein, 2009, Cor. 6.6) from Lemma 7.1.

Moreover, since $\mathbb{E}\varepsilon_{ij}^{*(6+\varepsilon)} < \infty$ and $\sup \operatorname{supp}(H_p) \to \sup \operatorname{supp}(H)$, the largest eigenvalue of $N^\top N$ converges a.s. to the upper edge b^2 of the support of $F_{\gamma,H}$, by the modification of (Bai and Silverstein, 2009, Cor. 6.6) presented in the proof of Lemma 7.1.

7.4 Proof of Lemma 7.3

Part 1: For $Z^\top(t^2 I_n - NN^\top)^{-1}Z$, note that Z has iid entries —with mean 0 and variance $1/n$—that are independent of N. We will use the following result:

Lemma 7.5 (Concentration of quadratic forms, consequence of Lemma B.26 in Bai and Silverstein (2009))**.** *Let $x \in \mathbb{R}^k$ be a random vector with i.i.d. entries and $\mathbb{E}[x] = 0$, for which $\mathbb{E}\left[(\sqrt{k}x_i)^2\right] = 1$ and $\sup_i \mathbb{E}\left[(\sqrt{k}x_i)^{4+\phi}\right] < C$ for some $\phi > 0$ and $C < \infty$. Moreover, let A_k be a sequence of random $k \times k$ symmetric matrices independent of x, with a.s. uniformly bounded eigenvalues. Then the quadratic forms $x^\top A_k x$ concentrate around their means: $x^\top A_k x - k^{-1} \operatorname{tr} A_k \to_{a.s.} 0$.*

We apply this lemma with $x = Z$, $k = p$ and $A_p = (t^2 I_n - NN^\top)^{-1}$. To get almost sure convergence, here it is required that z_i have finite $4 + \phi$-th moment. This shows the concentration of $Z^\top(t^2 I_n - NN^\top)^{-1}Z$.

Part 2: To show $Z^\top(t^2 I_n - NN^\top)^{-1}N\nu$ concentrates around 0, we note that $w = (t^2 I_n - NN^\top)^{-1}N\nu$ is a random vector independent of Z, with a.s. bounded norm. Hence, conditional on w:

$$Pr(|Z^\top w| \geq a|w) \leq a^{-4}\mathbb{E}|Z^\top w|^4 = a^{-4}[\sum_i \mathbb{E}Z_{ni}^4 w_i^4 + \sum_{i \neq j} \mathbb{E}Z_{ni}^2 \mathbb{E}Z_{nj}^2 w_i^2 w_j^2]$$

$$\leq a^{-4}\mathbb{E}Z_{n1}^4(\sum_i w_i^2)^2 = a^{-4}n^{-2}\mathbb{E}Z_1^4 \cdot \|w\|_2^4$$

For any C we can write

$$Pr(|Z^\top w| \geq a) \leq Pr(|Z^\top w| \geq a|\|w\| \leq C) + Pr(\|w\| > C).$$

For sufficiently large C, the second term, $Pr(\|w\| > C)$ is summable in n. By the above bound, the first term is summable for any C. Hence, by the Borel-Cantelli lemma, we obtain $|Z^\top w| \to 0$ a.s. This shows the required concentration.

Part 2: Finally we need to show that $\nu^\top(t^2 I_p - N^\top N)^{-1}\nu$ concentrates around a definite value. This is probably the most interesting part, because the vector ν is not random. Most results for controlling expressions of the above type are designed for random ν; however here the matrix N is random instead. For this reason we will adopt a different approach.

Under our assumption we have $\nu^\top (P - zI_p)^{-1}\nu \to m_H(z)$, for $z = t^2 + iv$ with $v > 0$ fixed, where P is the diagonal matrix with (j,j)-th entry $\mathrm{Var}\left[\varepsilon_{ij}^*\right]$. Therefore, Thm 1 of Bai et al. (2007) shows that $\nu^\top(zI_p - N^\top N)^{-1}\nu \to -m(z)$ a.s., where $m(z)$ is the Stieltjes transform of the Marchenko-Pastur distribution $F_{\gamma,H}$.

A close examination of their proofs reveals that their result holds when $v \to 0$ sufficiently slowly, for instance $v = n^{-\alpha}$ for $\alpha = 1/10$. The reason is that all bounds in the proof have the rate $n^{-k}v^{-l}$ for some small $k, l > 0$, and hence they converge to 0 for v of the above form.

For instance, the very first bounds in the proof of Thm 1 of Bai et al. (2007) are in Eq. (2.2) on page 1543. The first one states a bound of order $O(1/n^r)$. The inequalities leading up to it show that the bound is in fact $O(1/(n^r v^{2r}))$. Similarly, the second inequality, stated with a bound of order $O(1/n^{r/2})$ is in fact $O(1/(n^{r/2}v^r))$. These bounds go to zero when $v = n^{-\alpha}$ with small $\alpha > 0$. In a similar way, the remaining bounds in the theorem have the same property.

To get the convergence for real t^2 from the convergence for complex $z = t^2 + iv$, we note that

$$|\nu^\top(zI_p - N^\top N)^{-1}\nu - \nu^\top(t^2 I_p - N^\top N)^{-1}\nu| = v|\nu^\top(zI_p - N^\top N)^{-1}(t^2 I_p - N^\top N)^{-1}\nu| \leq$$
$$\leq v\|(t^2 I_p - N^\top N)^{-1}\|^2 \cdot u^\top u.$$

As discussed above, when $t > b$, the matrices $(t^2 I_p - N^\top N)^{-1}$ have a.s. bounded operator norm. Hence, we conclude that if $v \to 0$, then a.s.

$$\nu^\top(zI_p - N^\top N)^{-1}\nu - \nu^\top(t^2 I_p - N^\top N)^{-1}\nu \to 0.$$

Finally, $m(z) \to m(t^2)$ by the continuity of the Stieltjes transform for all $t^2 > 0$ (Bai and Silverstein, 2009). We conclude that $\nu^\top(t^2 I_p - N^\top N)^{-1}\nu \to -m(t^2)$ a.s. This finishes the analysis of the last quadratic form.

7.5   Proof of Lemma 7.4

Parts 1 and 2: The proofs of Part 1 and 2 are exactly analogous to those in Lemma 7.3. Indeed, the same arguments work despite the change from $(t^2 I_p - N^\top N)^{-1}$ to $(t^2 I_p - N^\top N)^{-2}$, because the only properties we used are its independence from Z, and its a.s. bounded operator norm. These also hold for $(t^2 I_p - N^\top N)^{-2}$, so the same proof works.

Part 3: We start with the identity $\nu^\top N^\top(t^2 I_n - NN^\top)^{-2}N\nu = -\nu^\top(t^2 I_p - N^\top N)^{-1}\nu + t^2\nu^\top(t^2 I_p - N^\top N)^{-2}\nu$. Since in Lemma 7.3 we have already established $\nu^\top(t^2 I_p - N^\top N)^{-1}\nu \to -m(t^2)$, we only need to show the convergence of $\nu^\top(t^2 I_p - N^\top N)^{-2}\nu$.

For this we will employ the following *derivative trick* (see e.g., Dobriban and Wager, 2015). We will construct a function with two properties: (1) its derivative is the quantity $\nu^\top(t^2 I_p - N^\top N)^{-2}\nu$ that we want, and (2) its limit is convenient to obtain. The following lemma will allow us to get our answer by interchanging the order of limits:

Lemma 7.6 (see Lemma 2.14 in Bai and Silverstein (2009)). *Let $f_1, f_2, \ldots$ be analytic on a domain D in the complex plane, satisfying $|f_n(z)| \leq M$ for every n and z in D.*

Suppose that there is an analytic function f on D such that $f_n(z) \to f(z)$ for all $z \in D$. Then it also holds that $f'_n(z) \to f'(z)$ for all $z \in D$.

Accordingly, consider the function $f_p(r) = -\nu^\top (rI_p - N^\top N)^{-1}\nu$. Its derivative is $f'_p(r) = \nu^\top (rI_p - N^\top N)^{-2}\nu$. Let $\mathcal{S} := \{x + iv : x > b + \varepsilon\}$ for a sufficiently small $\varepsilon > 0$, and let us work on the set of full measure where $\|N^\top N\| < b + \varepsilon/2$ eventually, and where $f_p(r) \to m(r)$. By inspection, f_p are analytic functions on $\mathcal{S}$ bounded as $|f_p| \leq 2/\varepsilon$. Hence, by Lemma 7.6, $f'_p(r) \to m'(r)$.

In conclusion, $\nu^\top N^\top (t^2 I_p - N^\top N)^{-2}N\nu \to m(t^2) + t^2 m'(t^2)$, finishing the proof.

7.6 Proof of Thm. 3.1 - Multispiked extension

Let us denote by $\nu_i = Mu_i/\xi_i^{1/2}$ the normalized backprojected signals, where $\xi_i = \|Mu_i\|^2 \to \tau_i$. For the proof we start as in Sec. 7.1, showing that we can equivalently work with $B_i^* = \sum_{k=1}^r (\xi_k \ell_k)^{1/2} z_{ik}\nu_k + \varepsilon_i^*$. Defining the $r \times r$ diagonal matrices L, Δ with diagonal entries ℓ_k, ξ_k (respectively), and the $n \times r$, $p \times r$ matrices $Z, \mathcal{V}$, with columns $Z_k = n^{-1/2}(z_{1k}, \ldots, z_{nk})^\top$ and ν_k respectively, we can thus work with

$$\tilde{B}^* = Z(\Delta L)^{1/2}\mathcal{V}^\top + N.$$

The matrix $M_n(t)$ is now $2r \times 2r$, and has the form

$$M_n(t) = \begin{bmatrix} t \cdot Z^\top (t^2 I_n - NN^\top)^{-1}Z & Z^\top (t^2 I_n - NN^\top)^{-1}N\mathcal{V} \\ \mathcal{V}^\top N^\top (t^2 I_n - NN^\top)^{-1}Z & t \cdot \mathcal{V}^\top (t^2 I_p - N^\top N)^{-1}\mathcal{V} \end{bmatrix} - \begin{bmatrix} 0_r & (\Delta L)^{-1/2} \\ (\Delta L)^{-1/2} & 0_r \end{bmatrix}.$$

It is easy to see that Lemma 7.2 still holds in this case. To find the limits of the entries of M_n, we need the following additional statement.

Lemma 7.7 (Multispiked quadratic forms)**.** *The quadratic forms in the multispiked case have the following properties for $t > b$:*

1. *$Z_k^\top R^\alpha Z_j \to 0$ a.s. for $\alpha = 1, 2$, if $k \neq j$.*
2. *$\nu_k^\top (t^2 I_p - N^\top N)^{-\alpha}\nu_j \to 0$ a.s. for $\alpha = 1, 2$, if $k \neq j$.*

This lemma is proved in Sec. 7.7, using similar techniques as those in Lemma 7.2. Defining the $r \times r$ diagonal matrices T with diagonal entries τ_k, we conclude that for $t > b$, $M_n(t) \to M(t)$ a.s., where now

$$M(t) = \begin{bmatrix} -t \cdot \underline{m}(t^2)I_r & -(TL)^{-1/2} \\ -(TL)^{-1/2} & -t \cdot m(t^2)I_r \end{bmatrix}.$$

As before, by Lemma A.1 of Benaych-Georges and Nadakuditi (2012), we get that for $\tau_k \ell_k > 1/D(b^2)$, $\sigma_k(\tilde{B}^*) \to t_k$ a.s., where $t_k^2 \cdot \underline{m}(t_k^2)m(t_k^2) = 1/(\tau_k \ell_k)$. This finishes the spike convergence proof.

To obtain the limit of the angles for $\hat{u}_k$ for a k such that $\ell_k > \tau_k D(b^2)$, consider the left singular vectors $\hat{Z}_k$ associated to $\sigma_k(\tilde{B}^*)$. Define the $2r$-vector

$$\alpha = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} (\Delta L)^{1/2}\mathcal{V}^\top \hat{u}_k \\ (\Delta L)^{1/2}Z^\top \hat{Z}_k \end{bmatrix}.$$

The vector α belongs to the kernel of $M_n(\sigma_k(\tilde{B}^*))$. As argued by Benaych-Georges and Nadakuditi (2012), the fact that the projection of α into the orthogonal complement

31

of $M(t_k)$ tends to zero, implies that $\alpha_j \to 0$ for all $j \notin \{k, k+r\}$. This proves that $\nu_j^\top \hat{u}_k \to 0$ for $j \neq k$, and the analogous claim for the left singular vectors.

The linear equation $M_n(\sigma_k(\tilde{B}^*))\alpha = 0$ in the k-th coordinate, where $k \leq r$, reads (with $t = \sigma_k(\tilde{B}^*)$):

$$t\alpha_k Z_k^\top R Z_k - \alpha_{r+k}(\xi_k \ell_k)^{-1/2} + \sum_{i \neq k} M_n(\sigma_k(\tilde{B}^*))_{ik}\alpha_k = 0.$$

Only the first two terms are non-negligible due to the behavior of M_n, so we obtain $t\alpha_k Z_k^\top R Z_k = \alpha_{r+k}(\xi_k \ell_k)^{-1/2} + o_p(1)$. Moreover taking the norm of the equation $\hat{Z}_k = R(tZ\beta_1 + N\mathcal{V}\beta_2)$ (see Lemma 5.1 in Benaych-Georges and Nadakuditi (2012)), we get

$$t^2 \sum_{i,j \leq r} \alpha_i \alpha_j Z_i^\top R^2 Z_j + \sum_{i,j \leq r} \alpha_{k+i}\alpha_{k+j}\nu_i^\top N^\top R^2 N\nu_j + \sum_{i,j \leq r} \alpha_i \alpha_{k+j} Z_i R^2 N\nu_j = 1.$$

From Lemma 7.7 and the discussion above, only the terms $\alpha_k^2 Z_k^\top R^2 Z_k$ and $\alpha_{r+k}^2 \nu_k^\top N^\top R^2 N\nu_k$ are non-negligible, so we obtain

$$t^2 \alpha_k^2 Z_k^\top R^2 Z_k + \alpha_{r+k}^2 \nu_k^\top N^\top R^2 N\nu_k = 1 + o_p(1).$$

Combining the two equations above,

$$\alpha_{r+k}^2 \left[\frac{Z_k^\top R^2 Z_k}{\xi_k \ell_k (Z_k^\top R Z_k)^2} + \nu_k^\top N^\top R^2 N\nu_k \right] = 1 + o_p(1).$$

Since this is the same equation as in the single-spiked case, we can take the limit in a completely analogous way. This finishes the proof.

7.7 Proof of Lemma 7.7

Part 1: The convergence $Z_k^\top R^\alpha Z_j \to 0$ a.s. for $\alpha = 1, 2$, if $k \neq j$, follows directly from the following well-known lemma, cited from Couillet and Debbah (2011):

Lemma 7.8 (Proposition 4.1 in Couillet and Debbah (2011)). *Let $x_n \in \mathbb{R}^n$ and $y_n \in \mathbb{R}^n$ be independent sequences of random vectors, such that for each n the coordinates of x_n and y_n are independent random variables. Moreover, suppose that the coordinates of x_n are identically distributed with mean 0, variance C/n for some $C > 0$ and fourth moment of order $1/n^2$. Suppose the same conditions hold for y_n, where the distribution of the coordinates of y_n can be different from those of x_n. Let A_n be a sequence of $n \times n$ random matrices such that $\|A_n\|$ is uniformly bounded. Then $x_n^\top A_n y_n \to_{a.s.} 0$.*

Part 2: To show $\nu_k^\top (t^2 I_p - N^\top N)^{-\alpha} \nu_j \to 0$ a.s. for $\alpha = 1, 2$, if $k \neq j$, the same technique cannot be used, because the vectors u_k are deterministic. However, it is straightforward to check that the method of Bai et al. (2007) that we adapted in proving Part 3 of Lemma 7.3 extends to proving $\nu_k^\top (t^2 I_p - N^\top N)^{-1} \nu_j \to 0$. Indeed, it is easy to see that all their bounds hold unchanged. In the final step, as a deterministic equivalent for $\nu_k^\top (t^2 I_p - N^\top N)^{-1} \nu_j \to 0$, one obtains $\nu_k^\top (t^2 I_p - P)^{-1} \nu_j$, where P is the diagonal matrix with (j,j)-th entry $\text{Var}\left[\varepsilon_{ij}^*\right]$, and this bilinear form tends to 0 by our assumption, showing $\nu_k^\top (t^2 I_p - N^\top N)^{-1} \nu_j \to 0$. Then $\nu_k^\top (t^2 I_p - N^\top N)^{-2} \nu_j \to 0$ follows from the derivative trick employed in Part 3 of Lemma 7.4. This finishes the proof.

7.8 The uniform model and the simple form of BLP

Here we introduce the *uniform model*, a special case of the linearly transformed model. In the uniform model, we have that $\mathbb{E}A_i^\top A_i = mI_p$, and moreover that $\mathbb{E}\varepsilon_i\varepsilon_i^\top = I_p$. This is useful for justifying the simpler form of the BLP that we are using:

Proposition 7.1. *In the uniform model, under the assumptions of Theorem 3.1, suppose also that the first eight moments of the entries of $A_i^\top A_i$ are uniformly bounded. Then the BLP $\hat{X}_i^{BLP}$ given in (1) is asymptotically equivalent to $\hat{X}_i^0 = \sum_k \ell_k/(1 + m\ell_k) \cdot u_k u_k^\top A_i^\top Y_i$, in the sense that $\mathbb{E}|\hat{X}_i^{BLP} - \hat{X}_i^0|^2 \to 0$.*

The proof is given below:

Recall that we observe $Y_i = A_i X_i + \varepsilon_i$. The BLP has the form $\hat{X}_i = \Sigma_X A_i^\top (A_i \Sigma_X A_i^\top + \Sigma_\varepsilon)^{-1} Y_i$. Now, $\Sigma_X = \sum_k \ell_k \cdot u_k u_k^\top$, while $\Sigma_\varepsilon = I_p$ in the uniform case. With $G = (A_i \sum_k \ell_k \cdot u_k u_k^\top A_i^\top + I_p)^{-1}$, we get that $\hat{X}_i = \sum_k \ell_k \cdot u_k u_k^\top A_i^\top G Y_i$.

7.9 Asymptotic BLP after backprojection.

Our goal is to show that $\hat{X}_i$ is equivalent to $\hat{X}_i^0 = \sum_k \ell_k/(1 + m\ell_k) \cdot u_k u_k^\top A_i^\top Y_i$, in the sense that $\mathbb{E}|\hat{X}_i - \hat{X}_i^0|^2 \to 0$.

Let us denote $v_k = A_i u_k$, and $G_k = (\sum_{j \neq k} \ell_j v_j v_j^\top + I)^{-1}$.

Then $\hat{X}_i = \sum_k \ell_k u_k v_k^\top G Y_i$ and $\hat{X}_i^0 = \sum_k \ell_k/(1 + m\ell_k) \cdot u_k v_k^\top Y_i$. Let us also define $m_k = v_k^\top G Y_i - 1/(1 + m\ell_k) \cdot v_k^\top Y_i$. Then, clearly,

$$\hat{X}_i - \hat{X}_i^0 = \sum_k m_k \cdot \ell_k u_k$$

Therefore it is enough to show that $\mathbb{E}m_k^2 \to 0$.

Using the rank one perturbation formula $u^\top (uu^\top + T)^{-1} = u^\top T^{-1}/(1 + u^\top T^{-1} u)$, we can write

$$v_k^\top G = v_k^\top \left[\sum_j \ell_j v_j v_j^\top + I \right]^{-1} = v_k^\top G_k / \left(1 + \ell_k v_k^\top G_k v_k \right).$$

In addition, by using the formula $(VV^\top + I)^{-1} = [I - V(V^\top V + I)^{-1} V^\top]$ for $V = [\ell_1^{1/2} v_1, \ldots, \ell_r^{1/2} v_r]$ (excluding v_k), we conclude that

$$v_k^\top G_k Y_i = v_k^\top \left[\sum_{j \neq k} \ell_j v_j v_j^\top + I \right]^{-1} Y_i = v_k^\top Y_i - v_k^\top V(V^\top V + I)^{-1} V^\top Y_i.$$

We thus have

$$m_k = \frac{v_k^\top G_k Y_i}{1 + \ell_k v_k^\top G_k v_k} - \frac{v_k^\top Y_i}{1 + m\ell_k}$$

$$= v_k^\top Y_i \left(\frac{1}{1 + \ell_k v_k^\top G_k v_k} - \frac{1}{1 + m \cdot \ell_k} \right) - \frac{v_k^\top V(V^\top V + I)^{-1} V^\top Y_i}{1 + \ell_k v_k^\top G_k v_k}.$$

Thus, it is enough to show

1. $\mathbb{E}(v_k^\top Y_i)^2 \cdot (v_k^\top G_k v_k - m)^2 \to 0$

2. $\mathbb{E}[v_k^\top V(V^\top V + I)^{-1} V^\top Y_i]^2 \to 0$

We prove these in turn below:

1. First, for (1):

   By using the formula $(VV^\top + I)^{-1} = [I - V(V^\top V + I)^{-1} V^\top]$, we see

   $$v_k^\top G_k v_k = v_k^\top \left[\sum_{j \neq k} \ell_j v_j v_j^\top + I \right]^{-1} v_k = v_k^\top v_k - v_k^\top V(V^\top V + I)^{-1} V^\top v_k.$$

   But $x^\top (V^\top V + I)^{-1} x \leq x^\top x$, because the eigenvalues of $(V^\top V + I)^{-1}$ are all at most unity. Thus, it is enough to show $\mathbb{E}(v_k^\top Y_i)^2 \cdot |V^\top v_k|^4 \to 0$ and $\mathbb{E}(v_k^\top Y_i)^2 \cdot (v_k^\top v_k - m)^2 \to 0$. For this, it is enough to show that $\mathbb{E}|V^\top v_k|^8 \to 0$, $\mathbb{E}(v_k^\top Y_i)^4 \to 0$, and $\mathbb{E}(v_k^\top v_k - m)^4 \to 0$.

   First we show $\mathbb{E}|V^\top v_k|^8 \to 0$. The entries of the $r-1$-dimensional vector $V^\top v_k$ are $\ell_j^{1/2} v_j^\top v_k = \ell_j^{1/2} u_j^\top D_i u_k$ for $j \neq k$. But we have that $\mathbb{E}(u_j^\top D_i u_k)^8 \to 0$. Indeed, since $D_i = mI + E_i$, and $u_k^\top u_j \to 0$, we only need to show that $\mathbb{E}(u_j^\top E_i u_k)^8 \to 0$. Since we assumed that E_{ij} has bounded 8th moments, this follows by expanding the moment. The details are omitted for brevity. Thus, under the assumptions made on E_i, $\mathbb{E}|V^\top v_k|^8 \to 0$.

   Second, we show that $\mathbb{E}(v_k^\top Y_i)^8 \to 0$. Indeed, $v_k^\top Y_i = u_k^\top A_i^\top (A_i \sum_j \ell_j^{1/2} z_{ij} u_j + \varepsilon_i)$. We can take the expectation over z_{ij} and ε_i, because they are independent from v_k. We obtain $\mathbb{E}(v_k^\top Y_i)^8 = \sum_j \ell_j^4 \mathbb{E}(u_k^\top D_i u_j)^8 + \mathbb{E}(u_k^\top D_i u_k)^4$. Similarly to above, each term converges to zero.

   Third, and finally, we show $\mathbb{E}(v_k^\top v_k - m)^4 \to 0$. Indeed, $v_k^\top v_k - m = u_k^\top D_i u_k - m \cdot u_k^\top u_k = u_k^\top (D_i - mI) u_k$, so the claim is equivalent to $\mathbb{E}(u_k^\top E_i u_k)^4 \to 0$, which follows as above. This finishes the proof.

2. By using $x^\top (V^\top V + I)^{-1} x \leq x^\top x$, it is enough to show $\mathbb{E}[v_k^\top VV^\top Y_i]^2 \to 0$. As above, this can be achieved by taking the expectation over z_{ij} and ε_i first, and then controlling $v_k^\top V$. The details are omitted for brevity.

This finishes the proof that BLP is equivalent to simple linear denoising.

7.10   Proof of Lemma 3.2

We will show that the operator norm of $\hat{M}^{-1} - M^{-1}$ converges to 0; since the operator norm of the data matrix B converges almost surely by the main theorem, the result follows. This is equivalent to showing that $\sup_i |M_i^{-1} - \hat{M}_i^{-1}| \to 0$, which, since the M_i are uniformly bounded away from 0, will follow if we show $\sup_i |M_i - \hat{M}_i| \to 0$ almost surely.

To show this, observe that by the Central Absolute Moment Inequality (see, for example, Mukhopadhyay, 2000) and the moment condition on D_{ij}, there is an absolute constant C such that

$$\mathbb{E}|\hat{M}_i - M_i|^{4+\phi} \leq Cn^{-(2+\phi/2)}$$

for all $i = 1, \ldots, p$. Therefore, for any $a > 0$,

$$\Pr(\sup_i |\hat{M}_i - M_i| \geq a) \leq \sum_{i=1}^{n} \Pr(|\hat{M}_i - M_i| \geq a)$$

$$\leq C \cdot n \cdot \frac{\mathbb{E}|\hat{M}_i - M_i|^{4+\phi}}{a^{4+\phi}}$$

$$\leq C a^{-(4+\phi)} n^{-(1+\phi/2)}.$$

Since this is summable, it follows that $\sup_i |\hat{M}_i - M_i| \to 0$ almost surely, as desired.

7.11 Derivation of optimal singular values and AMSE

The derivation of the optimal singualar values and the AMSE is a summary of what is found in Gavish and Donoho (2017). We provide it here for the reader's convenience.

Proposition 7.2. *There exist orthonormal bases $a_1, \ldots, a_p \in \mathbb{R}^p$ and $a'_1, \ldots, a'_n \in \mathbb{R}^n$ in which $n^{-1/2}X$ and $n^{-1/2}\hat{X}$ are jointly in block-diagonal form, with r 2-by-2 blocks. More precisely, there are orthogonal matrices A_1 and A_2 such that $A_1 \hat{X} A_2^\top = \oplus_{i=1}^{r} C_i$ and $A_1 X A_2^\top = \oplus_{i=1}^{r} D_i$, where C_i and D_i are 2-by-2 matrices given by*

$$C_i = \begin{pmatrix} \lambda_i c_i \tilde{c}_i & \lambda_i c_i \tilde{s}_i \\ \lambda_i \tilde{c}_i s_i & \lambda_i s_i \tilde{s}_i \end{pmatrix}, \qquad D_i = \begin{pmatrix} \ell_i^{1/2} & 0 \\ 0 & 0 \end{pmatrix}.$$

Proof. The proof, which is elementary linear algebra, is essentially contained in Gavish and Donoho (2017). $\qquad\square$

Since the rank r is fixed, and since the sines and cosines converge almost surely, it follows immediately that the quantity

$$Ł_\infty(\lambda_1, \ldots, \lambda_r) = \lim_{p,n \to \infty} n^{-1} \|\hat{X} - X\|_F^2$$

is well-defined, where $\hat{X}$ is the estimator such that $n^{-1}\hat{X}^\top \hat{X}$ has eigenvalues $\lambda_1, \ldots, \lambda_r$; and furthermore, since the squared Frobenius norm decomposes over blocks, we have:

$$Ł_\infty(\lambda_1, \ldots, \lambda_r) = \sum_{i=1}^{r} \|C_i - D_i\|_F^2$$

Consequently, the optimal λ_i is found by optimizing a single spike for a 2-by-2 block:

$$\lambda_i^* = \arg\min_\lambda \|C_i - D_i\|_F^2. \tag{14}$$

To solve for λ_i^* and find the AMSE, we write out the error explicitly:

$$\|C_i - D_i\|_F^2 = \lambda_i^2 - 2\ell_i^{1/2} c_i \tilde{c}_i + \ell_i$$

which is minimized at $\lambda_i^* = \ell_i^{1/2} c_i \tilde{c}_i$, and has minimum value $\ell_i(1 - c_i \tilde{c}_i)$. The total AMSE is therefore $\sum_{i=1}^{r} \ell_i(1 - c_i \tilde{c}_i)$.

7.12 Proof of Thm. 4.1

All that remains to show is that the out-of-sample W-AMSE is equal to the in-sample W-AMSE, when $W = \Sigma_\varepsilon^{-1/2}$. That is, we must show

$$\sum_{i=1}^{r} \left(\tilde{\ell}_i - \frac{\tilde{\ell}_i^2 c_i^4}{\tilde{\ell}_i c_i^2 + 1} \right) = \sum_{k=1}^{r} \tilde{\ell}_i (1 - c_i^2 \tilde{c}_i^2).$$

We will prove equality of the individual summands; denoting $\ell = \tilde{\ell}_i$, $c = c_i$ and $\tilde{c} = \tilde{c}_i$, this means showing $\ell - \frac{\ell^2 c_k^4}{\ell c^2 + 1} = \ell(1 - c^2 \tilde{c}^2)$. Straightforward algebraic manipulation shows that this is equivalent to showing $1/\tilde{c}^2 = 1 + 1/(\ell c^2)$. Substituting formulas (7) and (8), we have:

$$1/\tilde{c}^2 = \frac{1 + 1/\ell}{1 - \gamma/\ell^2} = \frac{\ell + 1}{\ell - \gamma/\ell} = 1 + \frac{1 + \gamma/\ell}{\ell - \gamma/\ell} = 1 + \frac{1}{\ell} \frac{1 + \gamma/\ell}{1 - \gamma/\ell^2} = 1 + 1/(\ell c^2)$$

as desired.

References

J. Andén, E. Katsevich, and A. Singer. Covariance estimation using conjugate gradient for 3D classification in cryo-EM. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 200–204. IEEE, 2015.

ASPIRE. Algorithms for Single Particle Reconstruction. `http://spr.math.princeton.edu/`, 2017.

X.-C. Bai, G. McMullan, and S. H. Scheres. How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences*, 40(1):49–57, 2015.

Z. Bai, B. Miao, and G. Pan. On asymptotics of eigenvectors of large sample covariance matrix. *The Annals of Probability*, 35(4):1532–1572, 2007.

Z. Bai and X. Ding. Estimation of spiked eigenvalues in spiked models. *Random Matrices: Theory and Applications*, 1(02):1150011, 2012.

Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, 2009.

Z. Bai and J. Yao. On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167–177, 2012.

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

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

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

T. Bhamre, T. Zhang, and A. Singer. Denoising and covariance estimation of single particle cryo-EM images. *Journal of Structural Biology*, 195(1):72–81, 2016.

J. M. Blackledge. *Digital Signal Processing: Mathematical and Computational Methods, Software Development and Applications*. Elsevier, 2006.

A. Buja and N. Eyuboglu. Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4):509–540, 1992.

E. Callaway. The revolution will not be crystallized. *Nature*, 525(7568):172, 2015.

P. Campisi and K. Egiazarian. *Blind image deconvolution: theory and applications.* CRC press, 2016.

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

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

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

Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Completing any low-rank matrix, provably. *The Journal of Machine Learning Research*, 16(1):2999–2034, 2015.

R. Couillet and M. Debbah. *Random Matrix Methods for Wireless Communications.* Cambridge University Press, 2011.

E. Dobriban. Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 04(04):1550019, 2015.

E. Dobriban. Factor selection by permutation. *arXiv preprint arXiv:1710.00479*, 2017.

E. Dobriban and A. B. Owen. Deterministic parallel analysis. *arXiv preprint arXiv:1711.04155*, 2017.

E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *arXiv preprint arXiv:1507.03003*, 2015.

D. L. Donoho, M. Gavish, and I. M. Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *arXiv preprint arXiv:1311.0851, to appear in AoS*, 2013.

B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

L. Erdős and H.-T. Yau. Universality of local spectral statistics of random matrices. *Bulletin of the American Mathematical Society*, 49(3):377–414, 2012.

M. Gavish and D. L. Donoho. Minimax risk of matrix denoising by singular value thresholding. *The Annals of Statistics*, 42(6):2413–2440, 2014a.

M. Gavish and D. L. Donoho. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152, 2017.

M. Gavish and D. L. Donoho. Optimal shrinkage of singular values. *arXiv preprint arXiv:1405.7511*, 2014b.

G. H. Golub and C. F. Van Loan. *Matrix Computations*, volume 3. JHU Press, 2012.

W. Hachem, A. Hardy, and J. Najim. A survey on the eigenvalues local behavior of large complex correlated wishart matrices. *ESAIM: Proceedings and Surveys*, 51: 150–174, 2015.

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

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

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

I. M. Johnstone and A. Onatski. Testing in high-dimensional spiked models. *arXiv*

preprint arXiv:1509.07269, 2015.

Z. Kam. The reconstruction of structure from electron micrographs of randomly oriented particles. *Journal of Theoretical Biology*, 82(1):15–39, 1980.

E. Katsevich, A. Katsevich, and A. Singer. Covariance matrix estimation for the cryo-EM heterogeneity problem. *SIAM Journal on Imaging Sciences*, 8(1):126–185, 2015.

R. H. Keshavan and A. Montanari. Regularization for matrix completion. In *Proceedings of International Symposium on Information Theory*, pages 1503–1507. IEEE, 2010.

R. H. Keshavan, S. Oh, and A. Montanari. Matrix completion from a few entries. In *2009 IEEE International Symposium on Information Theory*, pages 324–328. IEEE, 2009.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

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

V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, pages 2302–2329, 2011.

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

S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 2008.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.*, 114(4):507–536, 1967.

N. Mukhopadhyay. *Probability and Statistical Inference*. CRC Press, 2000.

R. R. 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.

R. R. Nadakuditi and A. Edelman. Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *Signal Processing, IEEE Transactions on*, 56(7):2625–2638, 2008.

B. Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817, 2008.

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

A. Onatski. Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258, 2012.

A. Onatski, M. J. Moreira, and M. Hallin. Asymptotic power of sphericity tests for high-dimensional data. *The Annals of Statistics*, 41(3):1204–1231, 2013.

A. Onatski, M. J. Moreira, and M. Hallin. Signal detection in high dimension: The multispiked case. *The Annals of Statistics*, 42(1):225–254, 2014.

D. Passemier and J.-F. Yao. On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices: Theory and Applications*, 1(01): 1150002, 2012.

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

D. Paul and A. Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

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

A. Rohde, A. B. Tsybakov, et al. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components*, volume 391. John Wiley & Sons, 2009.

A. Singer and H.-T. Wu. Two-dimensional tomography from noisy projections taken at unknown random directions. *SIAM Journal on Imaging Sciences*, 6(1):136–175, 2013.

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

E. M. Stein and R. Shakarchi. *Fourier Analysis: An Introduction*, volume 1. Princeton University Press, 2011.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

J. Yao, Z. Bai, and S. Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, 2015.

Z. Zhao, Y. Shkolnisky, and A. Singer. Fast steerable principal component analysis. *IEEE Transactions on Computational Imaging*, 2(1):1–12, 2016.