*The Annals of Statistics*
2020, Vol. 48, No. 1, 491–513
https://doi.org/10.1214/19-AOS1819


# OPTIMAL PREDICTION IN THE LINEARLY TRANSFORMED SPIKED MODEL

BY EDGAR DOBRIBAN[1], WILLIAM LEEB[2] AND AMIT SINGER[3]

[1]*Department of Statistics, The Wharton School, University of Pennsylvania, dobriban@wharton.upenn.edu*
[2]*School of Mathematics, University of Minnesota Twin Cities, wleeb@umn.edu*
[3]*Program in Applied and Computational Mathematics, Department of Mathematics, Princeton University, amits@math.princeton.edu*

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

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

for $i = 1, \dots, n$. The transform matrices $A_i$ are also observed. We model the unobserved signals (or regression coefficients) $X_i$ as vectors lying on an unknown low-dimensional space. Given only $Y_i$ and $A_i$ how should we predict or recover their values?

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

We show that this model has wide-ranging applications in signal processing, deconvolution, cryo-electron microscopy, and missing data with noise. For missing data, we show in simulations that our methods are 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 $X_i$:

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

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 $q_i \le p$, 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 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 $\varepsilon_i$. 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 $i = 1, \dots, n$, with outcome vector $Y_i$ and covariate matrix $A_i$. The goal is then to estimate the

Received September 2017; revised January 2019.
*MSC2010 subject classifications.* Primary 62H25; secondary 62H15, 45B05.
*Key words and phrases.* Principal component analysis, random matrix theory, shrinkage, high dimensional, spiked model, missing data, matrix completion.

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, Zheng and Bai (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, Ben Arous and Péché (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, Moreira and Hallin (2013), Donoho, Gavish and Johnstone (2018), Onatski, Moreira and Hallin (2014), Nadakuditi (2014), Gavish and Donoho (2017), Johnstone and Onatski (2015), Hachem, Hardy and Najim (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; see, for example, 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, McMullan and Scheres (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 $\varepsilon_i$ is noise (Katsevich, Katsevich and Singer (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, that is, $X_i = X$ 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, Katsevich and Singer (2015), Andén, Katsevich and Singer (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, Zhang and Singer (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, Montanari and Oh (2009, 2010), Koltchinskii, Lounici and Tsybakov (2011), Negahban and Wainwright (2011), Recht (2011), Rohde and Tsybakov (2011), Jain, Netrapalli and Sanghavi (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, Casella and McCulloch (2009)). (The more well known Best Linear Unbiased Estimator (BLUE) is defined for fixed-effects models where $X_i$ are nonrandom 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 $p/n \to \gamma > 0$ (Section .8 in the Supplementary Material (Dobriban, Leeb and Singer (2019))). 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 (Section 2.3).

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

3. We derive a suitable "normalization" method to make our method fully implementable in practice (Section 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 (2017)). We also discuss how to estimate the rank (Section 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 (Section 4).

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

## 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 $i = 1, \dots, n$. Preliminary rank estimate $r$ (see Section 3.4 for discussion).
2. Form backprojected data matrix $B = [A_1^\top Y_1, \dots, A_n^\top Y_n]^\top$ 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 Section 3.3.1.
4. Compute the singular values $\sigma_k$ and the top $r$ singular vectors $\hat{u}_k$, $\hat{v}_k$ of the matrix $\tilde{B}$.
5. Compute $\hat{X} = (\hat{X}_1, \dots, \hat{X}_n)^\top = \sum_{k=1}^r \hat{\lambda}_k \hat{u}_k \hat{v}_k^\top$.

Here $\hat{\lambda}_k$ are computed according to Section 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 (6) (assuming $p \le n$):

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

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 $i = 1, \dots, n$.

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 Section 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, Casella and McCulloch (2009)) that the best linear predictor, or BLP, of one signal $X_i$ using $Y_i$, is

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

Here, $\Sigma_X = \sum_{k=1}^r \ell_k u_k u_k^\top$ denotes the covariance matrix of one $X_i$, and $\Sigma_\varepsilon$ is the covariance matrix of the noise $\varepsilon_i$. 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 $\Sigma_X$ 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 Section .8 in the Supplementary Material (Dobriban, Leeb and Singer (2019)) 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 $\eta_k^0$ are certain constants given in Section .8 in the Supplementary Material (Dobriban, Leeb and Singer (2019)). 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 $\eta_k$ defining the BLP are no longer optimal, and must be adjusted downwards to account for the nonzero 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 $A_i, \varepsilon_i$, 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 (Donoho and Gavish (2014)), 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*, *backprojected 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 Theorem 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

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

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

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

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 $\eta_k^*$ in the high-dimensional limit $p/n \to \gamma$; that is, our goal is to find the optimal coefficients $\eta_k$, 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 $\eta_k$; (3) derive consistent estimators of the optimal $\eta_k$; and (4) derive consistently estimable formulas for the AMSE. As before, finding the optimal $\eta_k$ 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 Section 2, we wish to find the AMSE-optimal coefficients $\eta_k$ 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, backprojected data. Equivalently, we find the optimal singular values of the matrix with the same singular vectors as $\tilde{B} = [\tilde{B}_1, \dots, \tilde{B}_n]^\top$.

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 Section 2.1.

3.1. *The asymptotic spectral theory of the back-projected data.* The main theorem characterizes the asymptotic spectral theory of the normalized backprojected data matrix $\tilde{B} = BM^{-1}$, and of the unnormalized version $B = [A_1^\top Y_1, \dots, A_n^\top Y_n]^\top$. Our data are i.i.d. 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: $\ell_1 > \ell_2 > \cdots > \ell_r > 0$.

For a distribution $H$, let $F_{\gamma,H}$ denote the generalized Marchenko–Pastur distribution induced by $H$ with aspect ratio $\gamma$ (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 and independence*. The entries of the diagonal matrices $A_i^\top A_i$ are independent random variables. Moreover, recalling that we defined $E_i = A_i^\top A_i - M$, we have the bounded moment assumptions $\mathbb{E}|\varepsilon_{ij}^*|^{6+\phi} < C$, $\mathbb{E}|E_{ij}|^{6+\phi} < C$.

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 $k$

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

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 $k$

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

6. *Generic signal*. Let $P$ be the diagonal matrix with $P_{jj} = \text{Var}[M_j^{-1}\varepsilon_{ij}^*]$, 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 Section 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 Section 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 in practice that the Marchenko–Pastur law holds even in this regime. To illustrate this, in Figure 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, Zhang and Singer (2016) and Zhao, Shkolnisky and Singer (2016). Because this frequency block only contains 49 coefficients, the histogram we plot is for 100 draws of the noise. We whiten the backprojected 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 nonsparse. 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, Lounici and Tsybakov (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

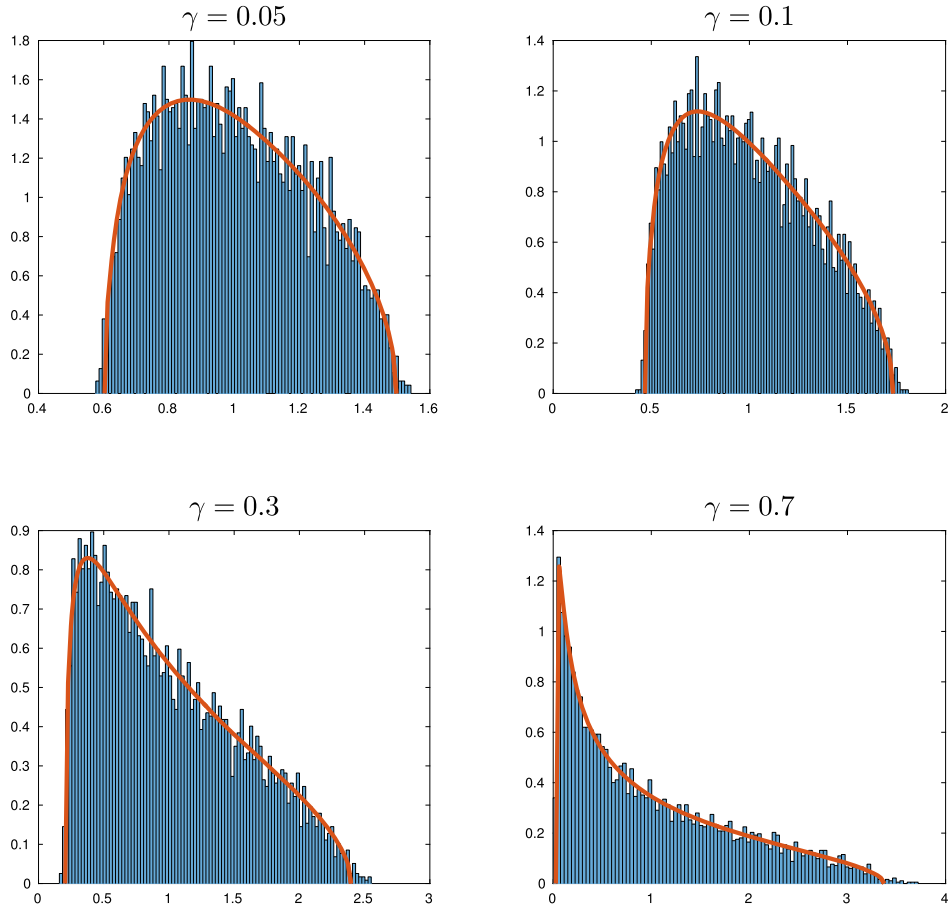


FIG. 1. *Histograms of empirical eigenvalues of whitened, backprojected noise using* 30 *CTFs, plotted against the Marchenko–Pastur density for different aspect ratios* $\gamma$.

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 $\mathbb{E}x^\top(P - zI_p)^{-1}x = p^{-1}\operatorname{tr}(P - zI_p)^{-1}$. 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* $k \le r$, *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*

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

*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*

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

*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*

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

The proof is in Section .1 in the Supplementary Material (Dobriban, Leeb and Singer (2019)). While the conclusion of this theorem is very similar to the results of Benaych-Georges and Nadakuditi (2012), our observation model $Y_i = A_i X_i + \varepsilon_i$ 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 Theorem 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 $v_j = Mu_j/\|Mu_j\|$. Then the above result holds for $B$, with $\ell_k$ replaced by $\tau_k \ell_k$, and $u_j$ replaced by $v_j$. The proof is identical, and is also presented in Section .1 in the Supplementary Material (Dobriban, Leeb and Singer (2019)).

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.

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 Section .11 in the Supplementary Material (Dobriban, Leeb and Singer (2019)), 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 $\lambda_i^*$. This will complete the proof that the algorithm given in Section 2.1 solves the problem posed in Section 2.4 and defines the EBLP.

3.2.1. *Estimating* $\ell_k$, $c_k$ *and* $\tilde{c}_k$. To evaluate the optimal $\lambda_i^*$, we estimate the values of $\ell_k$, $c_k$, and $\tilde{c}_k$ using Theorem 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

$$\hat{m}_{\gamma,H}(z) = \frac{1}{p-r}\sum_{k=r+1}^{p}\frac{1}{\lambda_k - z}, \tag{6}$$

where the sum is over the bottom $p-r$ eigenvalues $\lambda_k$ 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 $\ell_k$, $c_k$, 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 $n$ i.i.d. 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*

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

See Section .10 in the Supplementary Material (Dobriban, Leeb and Singer (2019)) 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

$$WX_i = \sum_{k=1}^{r} \tilde{\ell}_k^{1/2} \tilde{z}_{ik} \tilde{u}_k, \tag{7}$$

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

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

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

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

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

$$\tilde{c}_k^2 = \begin{cases} \dfrac{1 - \gamma/\tilde{\ell}_k^2}{1 + 1/\tilde{\ell}_k} & \text{if } \tilde{\ell}_k > \sqrt{\gamma}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

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

Following Section 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 $\varepsilon_i$ 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 consequently 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 (2019)). For other methods, see Kritchman

and Nadler (2008), Passemier and Yao (2012), and also Yao, Zheng and Bai ((2015), Chapter 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 (2019) are applicable if the spike strengths are sufficiently large.

**4. Out-of-sample prediction.** In Section 3, we derived the EBLP for predicting $X_i$ from $Y_i = A_i X_i + \varepsilon_i$, $i = 1, \dots, n$. We found the optimal coefficients $\eta_k$ 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 $Y_1, \dots, Y_n$, 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* $\hat{u}_1, \dots, \hat{u}_r$. It turns out that this independence forces us to use a *different* set of coefficients $\eta_k$ 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 $Y_1, \dots, Y_n$ as *in-sample prediction*, and the optimal predictor as the *in-sample EBLP*. We call $(Y_1, A_1), \dots, (Y_n, A_n)$ 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 $Y_1, \dots, Y_n$; the $r$ vectors $\hat{u}_1, \dots, \hat{u}_r$ 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 Section 2.1. We present the steps below.

1. *Input*: The top $r$ in-sample empirical PCs $\hat{u}_1, \dots, \hat{u}_r$. Estimates of the eigenvalues $\hat{\ell}_1, \dots, \hat{\ell}_r$ and cosines $\hat{c}_1, \dots, \hat{c}_r$. 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 $\varepsilon_i$, 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 $\eta_1, \dots, \eta_r$. 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 Section 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

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

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,

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

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, $\|\Sigma_{\tilde{B}} - (\Sigma_X + \Sigma_{\tilde{\varepsilon}})\|_{\mathrm{op}} \to 0$.

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

$$\hat{X}_0^\eta = \sum_{j=1}^r \eta_j \langle \tilde{B}_0, \hat{u}_j \rangle \hat{u}_j \tag{10}$$

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 $\eta_k$ 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 $\varepsilon_i$ 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 Section 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 $\eta_k$) is asymptotically equivalent to

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

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 Section 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

$$\sum_{k=1}^r \left(\tilde{\ell}_k - \frac{\tilde{\ell}_k^2 c_k^4}{\tilde{\ell}_k c_k^2 + 1}\right), \tag{11}$$

where $\tilde{\ell}_k$ are the eigenvalues of the whitened model $WX_i$, assuming the model (7). Using the formulas (8) and (9) for $c_k$ and $\tilde{c}_k$ as functions of $\tilde{\ell}_k$, it is straightforward to check that formula (11) 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 Section .12 in the Supplementary Material (Dobriban, Leeb and Singer (2019)). That is, the AMSE for whitened observations are identical for in-sample and out-of-sample EBLP.

Thus, we state the following theorem.

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 Theorem* 3.1, *and suppose in addition that* (7) *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* (10). *Then, for the optimal choice of* $\eta_k$, *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 Theorem 4.1 is contained in Section .12 in the Supplementary Material (Dobriban, Leeb and Singer (2019)).

**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, Netrapalli and Sanghavi (2013), Keshavan, Montanari and Oh (2010), Recht (2011)).

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 Section 3.4, such as Dobriban and Owen (2019) 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 Section 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 Section 5.1.2, we compare our method to competing methods when the column sampling probabilities exhibit varying degrees of nonuniformity. 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 $\ell$. 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 $\operatorname{corr}(\hat{X}_i, X_i) \to 1$ 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, Montanari and Oh (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 $X_\Omega$ denotes the vector of $X$'s values on the set of observed entries $\Omega$. 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 $\sigma^2$, then $w = \sigma(\sqrt{p} + \sqrt{n})\sqrt{|\Omega|/(pn)}$ at noise variance $\sigma^2$. 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 nonuniform sampling we find the EBLP outperforms OptShrink, especially as the noise level increases. In Section 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 Section 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 Sections 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 Section 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 $\delta$ of being selected, we will call $\delta$ 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 $\sigma$. 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.

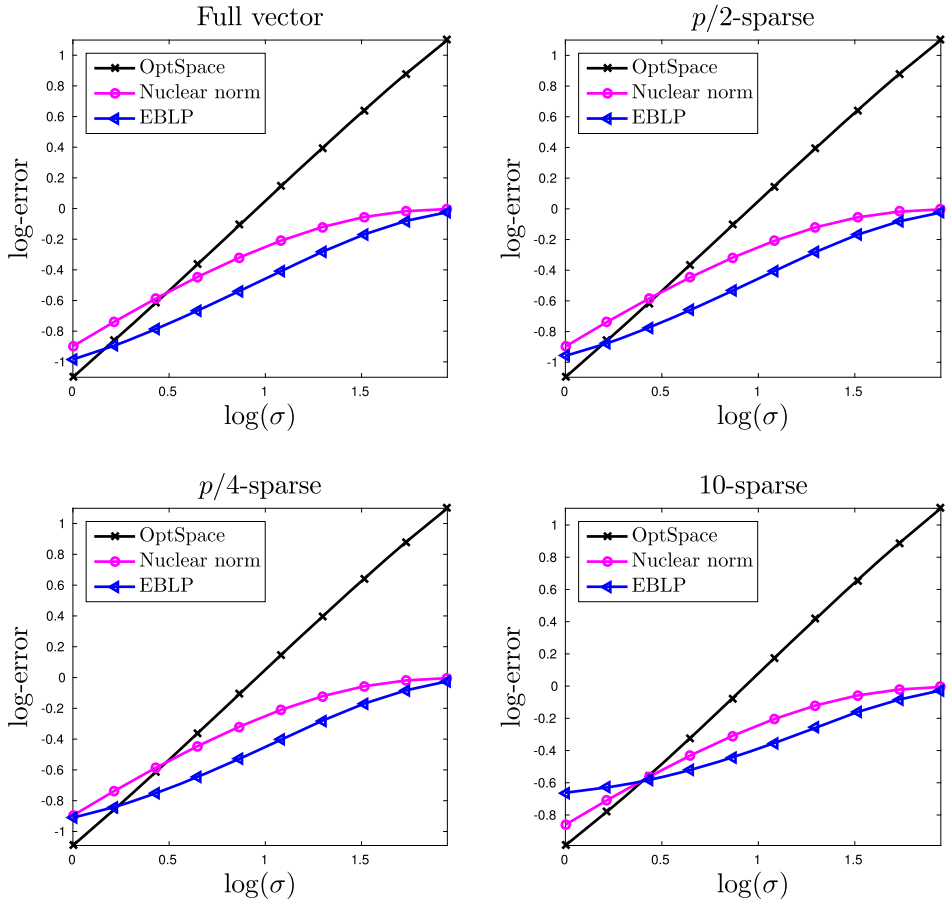


FIG. 2. *Log-RMSEs against log-noise for matrix completion. Each plot shows a different amount of sparsity in the PCs* $u_1, \ldots, u_{10}$.

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 nonzero; we consider the cases where all the PCs are 10-sparse, $p/4$-sparse, $p/2$-sparse, and dense. We show the results in Figure 2. 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 $\delta$ 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. We show the results in Figure 3. With uniform sampling, the two procedures are nearly identical. However, EBLP performs better when the sampling is nonuniform.

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 $\sigma$ 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,

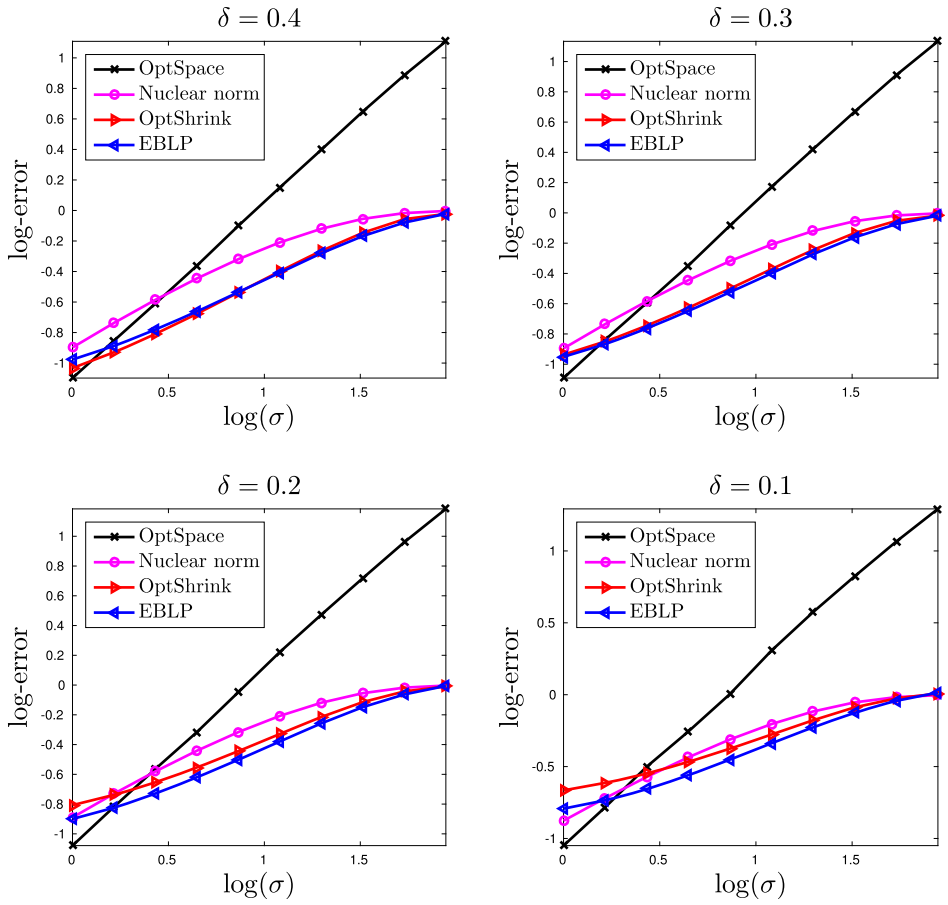


FIG. 3. *Log-RMSEs against log-noise for matrix completion. Each plot shows a different unevenness of sampling across the coordinates, with sampling probabilities ranging linearly from* $\delta$ *to* $1-\delta$.

and OptShrink (which does not whiten). We show the results in Figure 4. 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 $\kappa$.

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. Theorem 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 $\delta$, 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-$p$, $X_{\text{in}}$ and $X_{\text{out}}$, added Gaussian noise, and subsampled these matrices uniformly at rate $\delta$ to obtain the backprojected observations $\tilde{B}_{\text{in}}$ and $\tilde{B}_{\text{out}}$. We apply the in-sample EBLP on $\tilde{B}_{\text{in}}$ to obtain $\hat{X}_{\text{in}}$, and using the singular vectors of $\tilde{B}_{\text{in}}$, we apply the out-of-sample EBLP to $\tilde{B}_{\text{out}}$ to obtain $\hat{X}_{\text{out}}$.

In Figure 5, we show scatterplots of the RMSEs for the in-sample and out-of-sample data for each value of $n$ and $\ell$. 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 $\delta$.

**6. Conclusion.** In this paper, we considered the linearly transformed spiked model, and developed asymptotically optimal EBLP methods for predicting the unobserved signals in

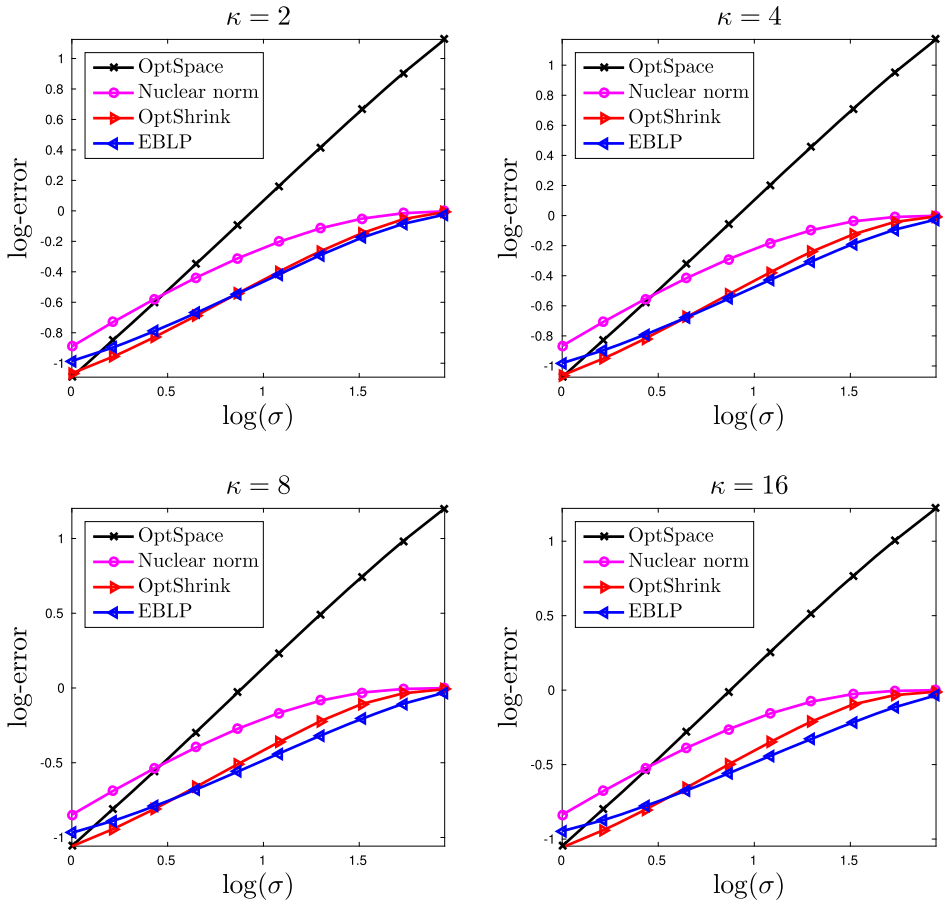


FIG. 4. *Log-RMSEs against log-noise for matrix completion. Each plot shows a different condition number $\kappa$ of the noise covariance matrix, reflecting different amounts of heterogeneity in the noise.*

the commutative case of the model, under high-dimensional asympotics. For missing data, 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.

**Acknowledgments.** The authors are grateful to the Associate Editor and the referees for detailed comments that have lead to improvements of this work. The authors thank Joakim Andén, Tejal Bhamre, Xiuyuan Cheng, David Donoho and Iain Johnstone for helpful discussions on this work. The authors are grateful to Matan Gavish for valuable suggestions on an earlier version of the manuscript.

This work was supported in part by award NSF BIGDATA IIS-1837992.

The first author was supported in part by NSF Grant DMS-1407813, and by an HHMI International Student Research Fellowship.

The second author was supported by the Simons Collaboration on Algorithms and Geometry.

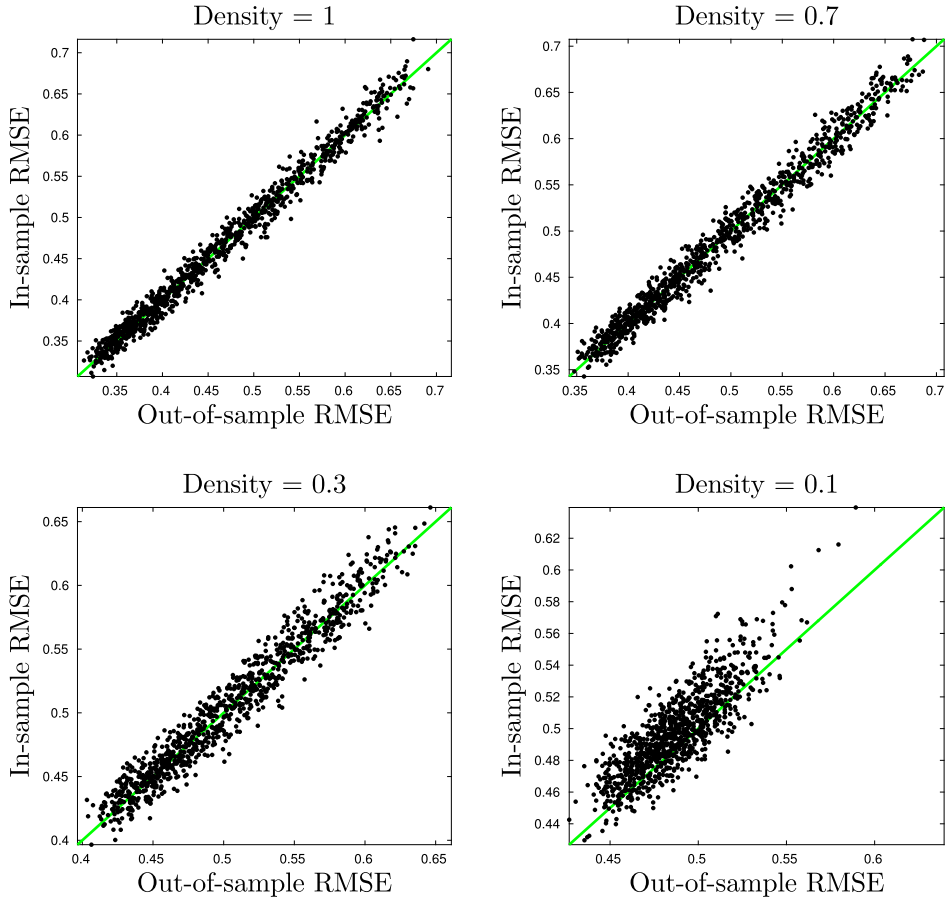


FIG. 5. *Scatterplots of the RMSEs of in-sample EBLP against out-of-sample EBLP for different sampling densities.*

The third author was supported in part by Award Number R01GM090200 from the NIGMS, FA9550-17-1-0291 from AFOSR, Simons Foundation Investigator Award and Simons Collaboration on Algorithms and Geometry, and the Moore Foundation Data-Driven Discovery Investigator Award.

## SUPPLEMENTARY MATERIAL

**Supplement to "Optimal prediction in the linearly transformed spiked model"** (DOI: 10.1214/19-AOS1819SUPP; .pdf). The supplementary material contains detailed proofs of certain results referred to in the main text.

## REFERENCES

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

ASPIRE (2017). Algorithms for single particle reconstruction. Available at http://spr.math.princeton.edu/.

BAI, Z. and DING, X. (2012). Estimation of spiked eigenvalues in spiked models. *Random Matrices Theory Appl.* **1** 1150011, 21. MR2934717 https://doi.org/10.1142/S2010326311500110

BAI, X.-C., MCMULLAN, G. and SCHERES, S. H. (2015). How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.* **40** 49–57.

BAI, Z. and SILVERSTEIN, J. W. (2009). *Spectral Analysis of Large Dimensional Random Matrices. Springer Series in Statistics*. Springer, New York. MR2567175 https://doi.org/10.1007/978-1-4419-0661-8

BAI, Z. and YAO, J. (2012). On sample eigenvalues in a generalized spiked population model. *J. Multivariate Anal.* **106** 167–177. MR2887686 https://doi.org/10.1016/j.jmva.2011.10.009

BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575 https://doi.org/10.1214/009117905000000233

BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. MR2279680 https://doi.org/10.1016/j.jmva.2005.08.003

BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *J. Multivariate Anal.* **111** 120–135. MR2944410 https://doi.org/10.1016/j.jmva.2012.04.019

BHAMRE, T., ZHANG, T. and SINGER, A. (2016). Denoising and covariance estimation of single particle cryo-EM images. *J. Struct. Biol.* **195** 72–81.

BLACKLEDGE, J. M. (2006). *Digital Signal Processing*: *Mathematical and Computational Methods*, *Software Development and Applications*. Elsevier, Amsterdam. MR2036815

BUJA, A. and EYUBOGLU, N. (1992). Remarks on parallel analysis. *Multivar. Behav. Res.* **27** 509–540.

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

CAMPISI, P. and EGIAZARIAN, K., eds. (2016). *Blind Image Deconvolution*: *Theory and Applications*. CRC Press, Boca Raton, FL. MR2404093 https://doi.org/10.1201/9781420007299

CANDÈS, E. J. and PLAN, Y. (2010). Matrix completion with noise. *Proc. IEEE* **98** 925–936.

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240 https://doi.org/10.1007/s10208-009-9045-5

CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56** 2053–2080. MR2723472 https://doi.org/10.1109/TIT.2010.2044061

CHEN, Y., BHOJANAPALLI, S., SANGHAVI, S. and WARD, R. (2015). Completing any low-rank matrix, provably. *J. Mach. Learn. Res.* **16** 2999–3034. MR3450532

DOBRIBAN, E. (2015). Efficient computation of limit spectra of sample covariance matrices. *Random Matrices Theory Appl.* **4** 1550019, 36. MR3418848 https://doi.org/10.1142/S2010326315500197

DOBRIBAN, E. (2017). Permutation methods for factor analysis and PCA. Preprint. Available at arXiv:1710.00479.

DOBRIBAN, E., LEEB, W. and SINGER, A. (2019). Supplement to "Optimal prediction in the linearly transformed spiked model." https://doi.org/10.1214/19-AOS1819SUPP.

DOBRIBAN, E. and OWEN, A. B. (2019). Deterministic parallel analysis: An improved method for selecting factors and principal components. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 163–183. MR3904784

DONOHO, D. and GAVISH, M. (2014). Minimax risk of matrix denoising by singular value thresholding. *Ann. Statist.* **42** 2413–2440. MR3269984 https://doi.org/10.1214/14-AOS1257

DONOHO, D., GAVISH, M. and JOHNSTONE, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.* **46** 1742–1778. MR3819116 https://doi.org/10.1214/17-AOS1601

EFRON, B. (2012). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758 https://doi.org/10.1017/CBO9780511761362

GAVISH, M. and DONOHO, D. L. (2017). Optimal shrinkage of singular values. *IEEE Trans. Inform. Theory* **63** 2137–2152. MR3626861 https://doi.org/10.1109/TIT.2017.2653801

GOLUB, G. H. and VAN LOAN, C. F. (2012). *Matrix Computations*, *Vol.* 3. *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR3024913

HACHEM, W., HARDY, A. and NAJIM, J. (2015). A survey on the eigenvalues local behavior of large complex correlated Wishart matrices. In *Modélisation Aléatoire et Statistique—Journées MAS* 2014. *ESAIM Proc. Surveys* **51** 150–174. EDP Sci., Les Ulis. MR3440796 https://doi.org/10.1051/proc/201551009

JAIN, P., NETRAPALLI, P. and SANGHAVI, S. (2013). Low-rank matrix completion using alternating minimization. In *STOC'*13*—Proceedings of the* 2013 *ACM Symposium on Theory of Computing* 665–674. ACM, New York. MR3210828 https://doi.org/10.1145/2488608.2488693

JI, S. and YE, J. (2009). An accelerated gradient method for trace norm minimization. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 457–464. ACM, New York.

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961 https://doi.org/10.1214/aos/1009210544

JOHNSTONE, I. M. and ONATSKI, A. (2015). Testing in high-dimensional spiked models. Preprint. Available at arXiv:1509.07269.

KAM, Z. (1980). The reconstruction of structure from electron micrographs of randomly oriented particles. *J. Theoret. Biol.* **82** 15–39.

KATSEVICH, E., KATSEVICH, A. and SINGER, A. (2015). Covariance matrix estimation for the cryo-EM heterogeneity problem. *SIAM J. Imaging Sci.* **8** 126–185. MR3302588 https://doi.org/10.1137/130935434

KESHAVAN, R. H. and MONTANARI, A. (2010). Regularization for matrix completion. In *Proceedings of International Symposium on Information Theory* 1503–1507. IEEE, New York.

KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from a few entries. *IEEE Trans. Inform. Theory* **56** 2980–2998. MR2683452 https://doi.org/10.1109/TIT.2010.2046205

KESHAVAN, R. H., OH, S. and MONTANARI, A. (2009). Matrix completion from a few entries. In 2009 *IEEE International Symposium on Information Theory* 324–328. IEEE, New York.

KLOPP, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20** 282–303. MR3160583 https://doi.org/10.3150/12-BEJ486

KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869 https://doi.org/10.1214/11-AOS894

KRITCHMAN, S. and NADLER, B. (2008). Determining the number of components in a factor model from limited noisy data. *Chemom. Intell. Lab. Syst.* **94** 19–32.

MALLAT, S. (2008). *A Wavelet Tour of Signal Processing*: *The Sparse Way*, *with Contributions from Gabriel Peyré*. Elsevier/Academic Press, Amsterdam. MR2479996

MARCHENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb.* **114** 507–536. MR0208649

NADAKUDITI, R. R. (2014). OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Trans. Inform. Theory* **60** 3002–3018. MR3200641 https://doi.org/10.1109/TIT.2014.2311661

NADAKUDITI, R. R. and EDELMAN, A. (2008). Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *IEEE Trans. Signal Process.* **56** 2625–2638. MR1500236 https://doi.org/10.1109/TSP.2008.917356

NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013 https://doi.org/10.1214/08-AOS618

NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. MR2816348 https://doi.org/10.1214/10-AOS850

ONATSKI, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J. Econometrics* **168** 244–258. MR2923766 https://doi.org/10.1016/j.jeconom.2012.01.034

ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2013). Asymptotic power of sphericity tests for high-dimensional data. *Ann. Statist.* **41** 1204–1231. MR3113808 https://doi.org/10.1214/13-AOS1100

ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2014). Signal detection in high dimension: The multispiked case. *Ann. Statist.* **42** 225–254. MR3189485 https://doi.org/10.1214/13-AOS1181

PASSEMIER, D. and YAO, J.-F. (2012). On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices Theory Appl.* **1** 1150002, 19. MR2930380 https://doi.org/10.1142/S201032631150002X

PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

PAUL, D. and AUE, A. (2014). Random matrix theory in statistics: A review. *J. Statist. Plann. Inference* **150** 1–29. MR3206718 https://doi.org/10.1016/j.jspi.2013.09.005

RECHT, B. (2011). A simpler approach to matrix completion. *J. Mach. Learn. Res.* **12** 3413–3430. MR2877360

ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342 https://doi.org/10.1214/10-AOS860

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (2009). *Variance Components*. *Wiley Series in Probability and Statistics* **391**. Wiley Interscience, Hoboken, NJ. MR2298115

SINGER, A. and WU, H.-T. (2013). Two-dimensional tomography from noisy projections taken at unknown random directions. *SIAM J. Imaging Sci.* **6** 136–175. MR3032950 https://doi.org/10.1137/090764657

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

STEIN, E. M. and SHAKARCHI, R. (2011). *Fourier Analysis*: *An Introduction*. *Princeton Lectures in Analysis* **1**. Princeton Univ. Press, Princeton, NJ. MR1970295

YAO, J., ZHENG, S. and BAI, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. *Cambridge Series in Statistical and Probabilistic Mathematics* **39**. Cambridge Univ. Press, New York. MR3468554 https://doi.org/10.1017/CBO9781107588080

ZHAO, Z., SHKOLNISKY, Y. and SINGER, A. (2016). Fast steerable principal component analysis. *IEEE Trans. Comput. Imag.* **2** 1–12. MR3472531 https://doi.org/10.1109/TCI.2016.2514700