The Annals of Applied Statistics
2018, Vol. 12, No. 4, 2121–2150
https://doi.org/10.1214/18-AOAS1146

*e*PCA: HIGH DIMENSIONAL EXPONENTIAL FAMILY PCA

BY LYDIA T. LIU[*,1], EDGAR DOBRIBAN[†,1,2] AND AMIT SINGER[‡,3]

University of California at Berkeley[*]*, University of Pennsylvania*[†] *and Princeton University*[‡]

Many applications involve large datasets with entries from exponential family distributions. Our main motivating application is photon-limited imaging, where we observe images with Poisson distributed pixels. We focus on X-ray Free Electron Lasers (XFEL), a quickly developing technology whose goal is to reconstruct molecular structure. In XFEL, estimating the principal components of the noiseless distribution is needed for denoising and for structure determination. However, the standard method, Principal Component Analysis (PCA), can be inefficient in non-Gaussian noise.

Motivated by this application, we develop *e*PCA (exponential family PCA), a new methodology for PCA on exponential families. *e*PCA is a fast method that can be used very generally for dimension reduction and denoising of large data matrices with exponential family entries.

We conduct a substantive XFEL data analysis using *e*PCA. We show that *e*PCA estimates the PCs of the distribution of images more accurately than PCA and alternatives. Importantly, it also leads to better denoising. We also provide theoretical justification for our estimator, including the convergence rate and the Marchenko–Pastur law in high dimensions. An open-source implementation is available.

1. Introduction. In many applications we have large collections of data vectors with entries sampled from exponential families (e.g. Poisson or Binomial). Our main motivating application is the important problem of molecular structure reconstruction using single-particle imaging. X-ray Free Electron Lasers (XFEL) are an experimental technique which leads to 2-D snapshots of imaged particles. The pixels have Poisson noise due to the small number of photons available in the short imaging time. To denoise the images and to reconstruct the 3-D structure, it is useful to estimate the covariance and principal components (PCs) of the pixels [see, e.g., Pande et al. (2015), Starodub et al. (2016)]. In addition to image processing, dimension reduction with non-Gaussian noise also arises in other settings such

Received September 2017; revised November 2017.

[1]The first two authors contributed equally to this work.

[2]Supported in part by NSF Grant DMS-1407813, and by an HHMI International Student Research Fellowship.

[3]Supported in part by Award Number R01GM090200 from the NIGMS, FA9550-12-1-0317 from AFOSR, Simons Foundation Investigator Award and Simons Collaboration on Algorithms and Geometry, and the Moore Foundation Data-Driven Discovery Investigator Award.

Key words and phrases. Denoising, shrinkage, XFEL imaging, random matrix theory.

as computational biology [Patterson, Price and Reich (2006)] and natural language processing [Deerwester et al. (1990)].

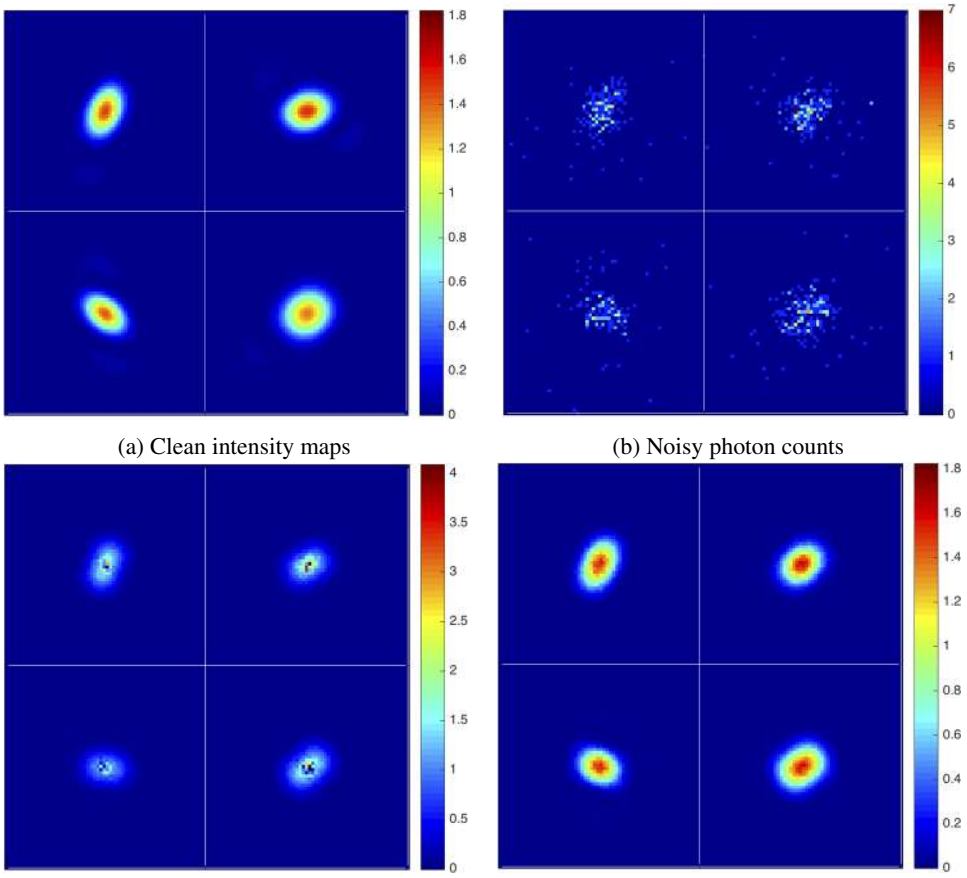
The standard method for dimension reduction and denoising of large datasets is Principal Component Analysis (PCA) [e.g., Anderson (2003), Jolliffe (2002)]. However, PCA is most naturally designed for Gaussian data, and there is no commonly agreed upon extension to exponential families [see, e.g., Jolliffe (2002), Section 14.4]. While there are some proposals, each of them has limitations, such as computational intractability for large datasets (see Section 2).

Motivated by single-particle imaging, we propose the new general method *e*PCA for PCA in exponential families. *e*PCA involves the eigendecomposition of a new covariance matrix estimator. Like usual PCA, it can be used for visualization and denoising of large data matrices. Moreover, *e*PCA has several appealing properties. First, it is a computationally efficient deterministic algorithm using basic linear algebra, making it as fast as usual PCA. This is in contrast to nonconvex likelihood methods, whose numerical solution has no convergence guarantees. Second, it is suitable for datasets with multiple types of variables (such as Poisson, binomial, and negative binomial). Third, it has substantial theoretical justification. We provide finite-sample convergence rates, and a precise high-dimensional analysis building on random matrix theory. Fourth, each step of *e*PCA is interpretable, which can be important to practitioners.

We conduct a substantive analysis of XFEL data using *e*PCA. We show that *e*PCA estimates the PCs more accurately than PCA and alternatives. Importantly, it also leads to better denoising, which is a crucial component in the overall XFEL pipeline. We perform extensive simulations with *e*PCA and show that in various metrics it outperforms usual PCA, PCA after standardization, and earlier PCA alternatives for exponential families (see Section 6.2). *e*PCA is publicly available in an open-source Matlab implementation from github.com/lydiatliu/epca/. That link also has software to reproduce our computational results.

To motivate our method, we now discuss in more detail the application to XFEL imaging, as well as potential applications in genomics.

1.1. *Single-particle imaging*: *XFEL.* XFEL is a rapidly developing and increasingly popular experimental method for understanding the three-dimensional structure of molecules [e.g., Bergmann, Yachandra and Yano (2017), Favre-Nicolin et al. (2015), Maia and Hajdu (2016)]. Single molecule XFEL imaging collects two-dimensional diffraction patterns of single particles at random orientations. A key advantage is that XFEL uses extremely short femtosecond X-ray pulses, during which the molecule does not change its structure. On the other hand, we only capture one diffraction pattern per particle and the particle orientations are unknown, so it is challenging to reconstruct the 3D structure at low signal-to-noise ratio. As illustrated in Figure 1, these images are very noisy due to the low number of photons that are typical for single particles [Pande et al. (2015)]. The count-noise at each detector is modeled with the Poisson distribution [see, e.g.,



(a) Clean intensity maps (b) Noisy photon counts

(c) Denoised (PCA) (d) Denoised (*e*PCA + EBLP)

FIG. 1. *XFEL diffraction pattern formation and denoising. See Section 6 for details.*

Martin et al. (2012)]. The standard deviation of a Poisson(λ) random variable is $\sqrt{\lambda}$, which is much larger than the mean λ when $\lambda \ll 1$. This explains why the images are noisier when the number of photons is low.

In order to reconstruct the 3-D structure of the particle, it may be possible to use Kam's method [Kam (1977, 1980), Saldin et al. (2009)]. A key requirement of Kam's method is to estimate the covariance matrix of the noiseless 2-D images, which is extremely difficult due to low photon counts. There has been much recent progress in using Kam's method for XFEL imaging [Starodub et al. (2016), Pande et al. (2014, 2015), Kurta et al. (2017)]. This motivates us to develop the *e*PCA method.

As a first example of the performance of *e*PCA, in Figure 1 we show the result of denoising synthetic XFEL using PCA and EBLP, a denoiser based on *e*PCA.

Clearly *e*PCA leads to better denoising than PCA. See Section 6 for details on the data generation and analysis.

1.2. *Genetic polymorphism data/SNPs*. In genomics, Single Nucleotide Polymorphism (SNP) data are the basis of thousands of Genome-Wide Association Studies (GWAS), which have recently led to hundreds of novel associations between common traits and genetic variants [e.g., Visscher et al. (2012)]. These types of studies are the backbone of modern medical genomics, and thus there is a great deal of interest in improving statistical inference in this area.

SNP data can be represented as an $n \times p$ matrix X with X_{ij} equal to the number of minor alleles (0, 1 or 2) of the jth SNP in the genome of the ith individual. The number of individuals n can be more than 10,000, while the number of SNPs p can be as large as 2.5 million. Binomial models are natural for such data, because each entry is a count. In addition to these genetic variants, a key health outcome (such as heart disease status) is also measured. The goal of a GWAS is to find genetic variants associated with the outcome.

PCA is already commonly used in the analysis of SNP data. One of the most common applications is to infer population structure and correct for population stratification [see, e.g., Patterson, Price and Reich (2006)]. For instance, in a dataset of European populations, there are subpopulations corresponding to the various countries. These subpopulations may not be known ahead of time. Since these subpopulations are systematically different, the inference of association between the outcome and the SNPs may be improved by correcting for the population structure. To correct for population structure, it is common to extract the PCs from the SNP data, and use them as covariates in the regression of the outcome on the SNPs. It is thus of interest to understand the proper way to estimate the covariance matrix and PCs. Our methods may lead to improved estimation and accuracy of the PCs, with improved downstream inferences in GWAS.

Another potential application area is RNA-sequencing, a new and rapidly developing experimental methodology in genomics that allows scientists to probe information at the single-cell level, to an unprecedented degree of granularity [see, e.g., Stegle, Teichmann and Marioni (2015) for a review]. In scRNA-seq, we observe read counts of many genes extracted from a large number of individual cells. Suppose X_{ij} is the number of reads mapped to gene j for sample cell i. A simple possible model assume that X_{ij} follows a negative binomial or a Poisson distribution. In the latter case, $X_{ij} \sim \text{Poisson}(\lambda_{ij})$ where λ_{ij} represents the rate at which reads map to gene j relative to other genes in sample cell i. Reads from single cells typically have many zeros, and therefore the Poisson or negative binomial models are more appropriate than a Gaussian approximation. Related negative binomial models are already in use [Anders and Huber (2010)].

1.3. *Our contributions*. We now briefly summarize our contributions:

1. We propose the new method *e*PCA, consisting of a new covariance estimator. We develop this in a sequence of steps (Sections 3 and 4): *diagonal debiasing* of the sample covariance matrix, *homogenization*, *shrinkage*, and *heterogenization*. We justify it by proving the standard Marchenko–Pastur law [Marčenko and Pastur (1967)] for the homogenized sample covariance matrix (Section 4.1.1), and by showing that homogenization improves the signal strength (Section 4.2.2).

An additional highly nontrivial eigenvalue shrinkage step—that we call *scaling*—is needed. We derive it by leveraging deep recent results from random matrix theory. This novel bias-correction step cannot be derived using classical low-dimensional asymptotics. We view this as a surprising theoretical discovery.

2. We apply *e*PCA to develop a new empirical Best Linear Predictor (EBLP) denoising method (Section 5), relying on random effects models [Searle, Casella and McCulloch (1992), Section 7.4].

3. We denoise synthetic XFEL data using *e*PCA, where it leads to more accurate PC estimates and better denoising than PCA (Section 6).

We also evaluate our covariance estimators in a simulation study, and show that they reduce the MSE for covariance, eigenvalue, and eigenvector estimation (Section 4.2.3), providing numerical justification for *e*PCA.

2. Related work. To give context for our method, we review related work. The reader interested in the methodology can skip directly to Section 3. We refer to Jolliffe (2002) for a detailed overview of PCA methodology, to Anderson (2003) for a more general overview of multivariate statistical analysis, and to Yao, Zheng and Bai (2015) for discussions of high-dimensional statistics, random matrix theory and PCA.

2.1. *Standardization and weighting in PCA*. In applying PCA, a key concern is whether or not to standardize the variables [e.g., Jolliffe (2002), Section 2.3]. Standardization ensures that results for different sets of random variables are more comparable, and also that PCs are less dominated by individual variables with large variances. Not standardizing makes statistical inference more convenient. In exploratory analyses, however, standardization is usually preferred. In our setting, the homogenization method (Section 4.1) has several advantages over standardization.

A more general class of methods is *weighted PCA*, where PCA is applied to rescaled random variables $w_j X(j)$, for some $w_j > 0$ [Jolliffe (2002), Section 2.3, Section 14.2] In general, choosing the weights can be nontrivial. Our homogenization step of *e*PCA (Section 4.1) is a particular weighting method, justified for data from exponential families. In addition to proposing it, we provide several theoretical justifications: the standard Marchenko–Pastur law, and the improvements in signal to noise ratio (SNR) (see Section 4.1).

2.2. *PCA in non-Gaussian distributions, GLLVMs.* There have been several approaches suggested for extending PCA to non-Gaussian distributions; see, for example, Jolliffe (2002), Section 14.4. One possibility is to use robust estimates of the covariance matrix [see Jolliffe (2002), Section 14.4, for references]. Another approach assumes that the natural parameter lies in a low-dimensional space [Collins, Dasgupta and Schapire (2001)], and then attempts to maximize the log-likelihood. This leads to a nonconvex optimization problem for which an alternating maximization method is proposed, without global convergence guarantees. More recently, Udell et al. (2014, 2016) described a similar generalization of PCA, while Li and Tao (2010) proposed another likelihood-based method, both without global convergence guarantees. Scalable methods include Josse and Wager (2016), albeit without precise performance guarantees in high dimensions.

Within factor analysis, generalized linear latent variable models (GLLVMs) model the relationship of an observed variable from a general distribution with unobserved latent variables [Bartholomew and Knott (1999), Huber, Ronchetti and Victoria-Feser (2004)]. These flexible likelihood-based methods enable careful modeling and statistical inference for parameters of interest in low-dimensional settings. However, estimation and inference are computationally challenging, and published examples have at most 10-20 dimensions [Huber, Ronchetti and Victoria-Feser (2004)]. In contrast our algorithm is as fast as PCA and we avoid any optimization problems. In addition, we have some understanding of the performance in high dimensions, by connecting to random matrix theory.

Finally, Chen and Storey (2015) recently proposed a method for the consistent estimation of low-dimensional latent structure in high-dimensional data. They take $p \to \infty$, while n is fixed, so this is an even more extremely high-dimensional setting. They have a diagonal bias-correction step similar to ours. However, we are able to leverage powerful eigenvalue shrinkage and rank selection methods, by working in a setting where probabilistic tools from random matrix theory apply.

2.3. *Denoising and covariance estimation by singular value shrinkage.* Recently, results from random matrix theory have been used for studying covariance estimation and PCA for Gaussian and rotationally invariant data [e.g., Nadakuditi (2014), Shabalin and Nobel (2013), Donoho, Gavish and Johnstone (2013)]. While the qualitative insights they identify—for example, the improvements due to eigenvalue shrinkage—are relevant to our setting, the specific results and methods do not apply directly.

The recent work of Bigot, Deledalle and Féral (2016) develops a generalized Stein's Unbiased Risk Estimation (SURE) approach for singular value shrinkage denoising of low-rank matrices in exponential families. However, their shrinkage formulas become numerically intractable for Frobenius norm beyond Gaussian errors, and they instead introduce a heuristic algorithm. Their work is geared toward higher signal-to-noise ratio settings.

2.4. *Image processing and denoising.* There are many approaches to denoising in image and signal processing, the majority designed for Gaussian noise [see, e.g., Starck, Murtagh and Fadili (2010)]. Most classical methods are designed for "single-image denoising," and do not share information across multiple images. Our setting is different, because we have many very noisy samples—for example, XFEL images.

Starck, Murtagh and Fadili (2010) Section 6.5 provides an overview of the classical methods for Poisson noise (but of course our method works for all exponential families). Popular approaches reduce to the Gaussian case by a wavelet transform such as a Haar transform [Nowak and Baraniuk (1999)]; by adaptive wavelet shrinkage; or by approximate variance stabilization such as the Anscombe transform. The latter is known to work well for Poisson signals with large parameters, due to approximate normality. However, the normal approximation breaks down for the Poisson with a small parameter, such as photon-limited XFEL [see, e.g., Starck, Murtagh and Fadili (2010), Section 6.6].

Other methods are based on singular value thresholding (SVT), with various approaches to handling non-Gaussian noise. For example, Furnival, Leary and Midgley (2017) performs SVT of the data matrix of image time-series in low noise, picking the regularization parameter to minimize the Poisson–Gaussian Unbiased Risk Estimator. We instead homogenize the data and propose a second-moment based denoising method. Alternatively, Cao and Xie (2014) frames denoising as a regularized maximum likelihood problem and uses SVT to optimize an approximation of the Poisson likelihood. Our approach avoids nonconvex likelihood optimization problems, and works beyond the Poisson distribution for all exponential families.

3. Covariance estimation.

*e*PCA is the eigendecomposition of a new covariance matrix estimator. To develop this estimator, we start with the sample covariance matrix and propose a sequence of improvements (see Table 1 and below). Our method is directly motivated by XFEL, so we explain the scientific background for every step along the way.

TABLE 1
Covariance estimators

Not.	Name	Formula	Def.	Motivation
S	Sample covariance	$S = n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^\top$	(4)	–
S_d	Diagonal debiasing	$S_d = S - \mathrm{diag}[V(\bar{Y})]$	(5)	Hierarchy
S_h	Homogenization	$S_h = D_n^{-\frac{1}{2}} S_d D_n^{-\frac{1}{2}}$	(6)	Heteroskedasticity
$S_{h,\eta}$	Shrinkage	$S_{h,\eta} = \eta(S_h)$	(7)	High dimensionality
S_{he}	Heterogenization	$S_{\mathrm{he}} = D_n^{\frac{1}{2}} S_{h,\eta} D_n^{\frac{1}{2}}$	(8)	Heteroskedasticity
S_s	Scaling	$S_s = \sum \hat{\alpha}_i \hat{v}_i \hat{v}_i^\top$, where $S_{\mathrm{he}} = \sum \hat{v}_i \hat{v}_i^\top$	(13)	Heteroskedasticity

We will work with observations Y from the canonical one-parameter exponential family with density

$$p_\theta(y) = \exp\big[\theta y - A(\theta)\big] \tag{1}$$

with respect to a σ-finite measure ν on $\mathbb{R}$ [see, e.g., Lehmann and Romano (2005)]. Here $\theta \in \mathbb{R}$ is the natural parameter of the family and $A(\theta) = \log \int \exp(\theta y) d\nu(y)$ is the log-partition function. We assume the distribution is well defined for all θ in an open set. The mean and variance of Y can be expressed as $\mathbb{E}Y = A'(\theta)$ and $\mathrm{Var}[Y] = A''(\theta)$, where we denote $g'(\theta) = dg(\theta)/d\theta$.

While our method works for all exponential families, we will use the Poisson distribution $y \sim \mathrm{Poisson}(x)$ as a running example. Here the carrier measure is the discrete measure with density $\nu(dy) = 1/y!$ with respect to the counting measure on the nonnegative integers, while $\theta = \log(x)$ and $A(\theta) = \exp(\theta)$.

3.1. *The observation model.* Let $Y \in \mathbb{R}^p$ be a random vector with some unknown distribution. We observe n i.i.d. noisy data vectors $Y_i \sim Y$. In the XFEL application, Y is the noisy image with the pixels as coordinates. We consider the following hierarchical model for Y. First, a latent vector—or hyperparameter—$\theta \in \mathbb{R}^p$ is drawn from a probability distribution D with mean μ_θ and covariance matrix Σ_θ. Conditional on θ, the coordinates of $Y = (Y(1), \ldots, Y(p))^\top$ are drawn independently from an exponential family $Y(j) \sim p_{\theta(j)}(y)$ defined in (1). Formally, denoting by $\dot{\sim}$ the mean and the covariance of a random vector:

$$\theta \dot{\sim} (\mu_\theta, \Sigma_\theta),$$

$$Y(j)|\theta(j) \sim p_{\theta(j)}(y), \qquad Y = \big(Y(1), \ldots, Y(p)\big)^\top.$$

Therefore, the mean of Y conditional on θ is

$$X := \mathbb{E}(Y|\theta) = \big(A'\big(\theta(1)\big), \ldots, A'\big(\theta(p)\big)\big)^\top = A'(\theta),$$

so the noisy data vector Y can be expressed as $Y = A'(\theta) + \tilde{\varepsilon}$, with $\mathbb{E}(\tilde{\varepsilon}|\theta) = 0$, while the marginal mean of Y is $\mathbb{E}Y = \mathbb{E}A'(\theta)$. Thus one can think of Y as a noisy realization of the clean vector $X = A'(\theta)$. However, the latent vector θ is also random and varies from sample to sample. In the XFEL application, $A'(\theta)$ are the unobserved noiseless images.

Our model is quite realistic in the XFEL application, where a small number of latent parameters determines the clean image, and the noise is added independently afterwards. Conditional independence given θ means that all latent effects that induce correlations do so through θ and not through some other mechanism. This is reasonable, as we can always capture much of the latent correlations in the "mean" structure by increasing the number of PCs. In addition, similar conditional independence is also common in empirical work such as bulk RNA-Seq analysis [e.g., Anders and Huber (2010)].

It is important that we model the *mean* $A'(\theta)$ of the exponential family as our clean signal, as opposed to the *natural parameter* θ. This is a reasonable assumption in many applications, where the means of the noisy signals lie on an approximately low-dimensional linear subspace. For instance, Basri and Jacobs (2003) found that the images of a single face under different lighting conditions inhabit an approximately nine-dimensional linear space. As mentioned in Section 2, this is a key modeling assumption distinguishing our approach from prior work like Collins, Dasgupta and Schapire (2001), and it enables a deterministic noniterative algorithm.

We thus have $Y = A'(\theta) + \operatorname{diag}[A''(\theta)]^{\frac{1}{2}}\varepsilon$, where the coordinates of ε are conditionally independent and standardized given θ. Therefore, the covariance of Y conditional on θ is

$$\operatorname{Cov}[Y|\theta] = \operatorname{diag}\big[A''\big(\theta(1)\big), \ldots, A''\big(\theta(p)\big)\big] = \operatorname{diag}\big[A''(\theta)\big].$$

The marginal covariance of Y is given by the law of total covariance:

$$\operatorname{Cov}[Y] = \operatorname{Cov}\big[\mathbb{E}(Y|\theta)\big] + \mathbb{E}\big[\operatorname{Cov}[Y|\theta]\big] = \operatorname{Cov}\big[A'(\theta)\big] + \mathbb{E}\operatorname{diag}\big[A''(\theta)\big]. \tag{2}$$

In particular, the coordinates of Y are independent only conditionally on θ, but not marginally. For the special case of Poisson observations $Y \sim \operatorname{Poisson}_p(X)$, where $X \in \mathbb{R}^p$ is random, we can write $Y = X + \operatorname{diag}(X)^{\frac{1}{2}}\varepsilon$. The natural parameter is the vector θ with $\theta(j) = \log X(j)$. Since $A'(\theta(j)) = A''(\theta(j)) = \exp(\theta(j)) = X(j)$, we see $\mathbb{E}Y = \mathbb{E}X$, and $\operatorname{Cov}[Y] = \operatorname{Cov}[X] + \mathbb{E}\operatorname{diag}[X]$.

3.2. *Diagonal debiasing.* We will propose several estimators of increasing sophistication to estimate the covariance matrix $\Sigma_x = \operatorname{Cov}[A'(\theta)]$ of the noiseless vectors $X_i = A'(\theta_i)$ (see Table 1). Clearly, due to the covariance equation (2), the sample covariance matrix of Y_i is biased for estimating the diagonal elements of Σ_x. Fortunately, this bias can be corrected. Indeed, we only need to subtract the noise variances $\mathbb{E}A''(\theta(j))$. We know that $\mathbb{E}Y(j) = \mathbb{E}A'(\theta(j))$, so it is natural to define associated estimators via the *variance map* of the exponential family, which takes a mean parameter $A'(\theta)$ into the associated variance parameter $A''(\theta)$. Formally,

$$V(m) = A''\big[\big(A'\big)^{-1}(m)\big]. \tag{3}$$

If the distribution of Y is nondegenerate, $A''(\theta) = \operatorname{Var}_\theta(Y) > 0$, so A' is increasing and invertible, and the variance map is well defined.

We define the sample covariance estimator

$$S = n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^\top, \tag{4}$$

where $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ is the sample mean. We estimate $\mathbb{E}A''(\theta)$ by $V(\bar{Y})$, and define the *diagonally debiased* covariance estimator

$$S_d = S - \operatorname{diag}\big[V(\bar{Y})\big]. \tag{5}$$

Continuing with the special case of a Poisson example, $A'(\theta) = A''(\theta) = \exp(\theta)$, so $V(m) = m$, and $S_d = S - \operatorname{diag}[\bar{Y}]$. In this example the estimator is unbiased, because V is linear. When V is nonlinear, the estimator can become slightly biased.

3.2.1. *The rate of convergence.* Our first theoretical result characterizes the finite-sample convergence rate of the diagonally debiased covariance estimator S_d, for any fixed n, p. This estimator is not a sample covariance matrix, which is inconsistent in our case when $n \to \infty$ and p is fixed. Thus it is necessary to study its convergence rate from first principles.

For this we need to make a few technical assumptions. First, we assume that the variance map V is Lipschitz with constant L. It is easy to check that this is true for the Gaussian, Poisson, and Binomial distributions. We also assume that the coordinates of the random vector θ are almost surely bounded, $\|\theta\|_\infty \le B$. Since A' is continuous and invertible, this is equivalent to the boundedness of $A'(\theta)$. This is reasonable in the areas that we are interested in—XFEL imaging does not have infinite energy, so we have an upper bound on the intensity of pixels. Finally we assume that $m_4 = \max_i \mathbb{E}[Y(i)^4] \ge C$ for some universal constant $C > 0$. This is reasonable, as it states that at least some entries of the random vector have nonvanishing magnitude.

Let $\lesssim$ denote inequality up to constants not depending on n and p. Let $\|\cdot\|_{\mathrm{Fr}}$ be the Frobenius norm and $\|\cdot\|$ be the operator norm. Our result, proved in Section A.1, is

THEOREM 3.1 (Rate of convergence of debiased covariance estimator). *The diagonally debiased covariance estimator S_d has the following rates of convergence. In the Frobenius norm, with $\mu := \mathbb{E}Y = \mathbb{E}X = \mathbb{E}A'(\theta)$,*

$$\mathbb{E}\big[\|S_d - \Sigma_x\|_{\mathrm{Fr}}\big] \lesssim \sqrt{\frac{p}{n}}\big[\sqrt{p}\cdot m_4 + \|\mu\|\big].$$

In operator norm, with the dimensional constant $C(p) = 4(1 + 2\lceil \log p\rceil)$:

$$\mathbb{E}\big[\|S_d - \Sigma_x\|\big] \lesssim \sqrt{C(p)}\frac{(\mathbb{E}\|Y\|^4)^{\frac{1}{2}} + (\log n)^3(\log p)^2}{\sqrt{n}} + \sqrt{\frac{p}{n}}\left[1 + \sqrt{\frac{p}{n}} + \|\mu\|\right].$$

The two error rates are both of interest, and complement each other. The Frobenius norm rate captures the deviation across all entries of the covariance matrix. The operator norm rate is typically faster than the Frobenius norm rate. For instance, in XFEL it is reasonable to assume that the total intensity across all detectors is fixed as the resolution increases. This leads to a fixed value for $\mathbb{E}\|Y\|^4$ that

does not grow with n. The operator norm rate can be as fast as $(p/n)^{\frac{1}{2}}$ while the Frobenius norm rate is $p/n^{\frac{1}{2}}$.

Our operator norm concentration result implies that S_d is positive semidefinite with high probability, as long as the bound is sufficiently small. While we always see in simulations that S_d is positive semidefinite, this property can be guaranteed if need be, by computing the eigendecomposition of S_d and setting any negative eigenvalues to zero.

Our proof of Theorem 3.1 exploits that exponential family random variables are sub-exponential, so we can use corresponding moment bounds. We also rely on operator-norm bounds for random matrices from Tropp (2016) and on moment bounds from Boucheron et al. (2005).

We point out here that later we will consider a different set of asymptotics, to understand the effects of high dimension. In that setting, S_d is typically of much larger rank than Σ_x, and the rest of the *e*PCA algorithm exploits this structure to improve S_d. The analysis in this section is informative when the dimension is fixed and small.

4. Homogenization and shrinkage.

4.1. *Homogenization.* In the previous sections, we showed that the diagonally debiased sample covariance matrix converges at a rate $O(pn^{-\frac{1}{2}})$. Next we propose a shrinkage method to improve this estimator in the high-dimensional regime where $n, p \to \infty$ and $p/n \to \gamma > 0$. As a preliminary step, it is helpful to homogenize the empirical covariance matrix and remove the effects of heteroskedasticity. This allows us to get closer to the *standard spiked model* [Johnstone (2001)] where the noise has the same variance for all features. In that setting covariance estimation via eigenvalue shrinkage has been thoroughly studied [Donoho, Gavish and Johnstone (2013)].

The vector of noise variances affecting the different components is $\mathbb{E}[A''(\theta)]$. For a given signal $Y = A'(\theta) + \operatorname{diag}[A''(\theta)]^{\frac{1}{2}}\varepsilon$, homogenization transforms it to $Y_h = \operatorname{diag}[A''(\theta)]^{-\frac{1}{2}}A'(\theta) + \varepsilon$. The covariance is transformed from $\operatorname{Cov}[Y]$ to $\operatorname{diag}[A''(\theta)]^{-\frac{1}{2}}\operatorname{Cov}[Y]\operatorname{diag}[A''(\theta)]^{-\frac{1}{2}}$. Since the diagonal correction $D_n = \operatorname{diag}[V(\bar{Y})]$ estimates $\mathbb{E}\operatorname{diag}[A''(\theta)]$, we define the *homogenized* covariance estimator by

$$S_h = D_n^{-\frac{1}{2}} S_d D_n^{-\frac{1}{2}} = D_n^{-\frac{1}{2}} S D_n^{-\frac{1}{2}} - I_p. \tag{6}$$

For the special case of Poisson observations, every entry of the noisy vector has to be divided by square root of the corresponding entry of the sample mean, so $S_h = \operatorname{diag}[\bar{Y}]^{-\frac{1}{2}} S \operatorname{diag}[\bar{Y}]^{-\frac{1}{2}} - I_p$.

Homogenization is different from *standardization*, the classical method for removing heteroskedasticity. To standardize, each feature—for example, pixel—is

divided by its empirical standard deviation [e.g., Jolliffe (2002), Section 2.3]. This ensures that all features have the same norm. The sample covariance matrix becomes a sample correlation matrix. In our case it turns out that this procedure "over-corrects." The overall variance $\mathrm{Var}[Y(i)]$ of each feature is the sum of the signal variance $\mathrm{Var}[A'(\theta(i))]$ and the noise variance $\mathbb{E}[A''(\theta(i))]$. Homogenization divides by the estimated noise standard errors, while standardization divides by the *overall* standard error due to the signal and noise.

Therefore, in our setting homogenization is more justified than standardization. Moreover, the standard Marchenko–Pastur law holds for the homogenized estimator (Theorem 4.2 in the next section). This also suggests that the top "noise" eigenvalue has a well-understood Tracy–Widom distribution asymptotically [Johnstone (2001)], which can be used to devise tests of significance. Another justification is that standardization improves the signal strength for "delocalized" eigenvectors (Section 4.2.2). We discuss these in detail below.

4.1.1. *Marchenko–Pastur law.* A key advantage of homogenization is that the homogenized estimator has a well-understood asymptotic behavior. In contrast, the unhomogenized estimator has a more complicated behavior. In this section, we show both of the above claims. We show that the limit spectra of our covariance matrix estimators are characterized by the Marchenko–Pastur (MP) law [Marčenko and Pastur (1967)], proving the general MP law for the sample covariance S, and the standard MP law for the homogenized covariance S_h.

For simplicity, we consider the case is when $\theta \in \mathbb{R}^p$ is fixed. This can be thought of as the "null" case, where all mean signals are the same. Then we can write $Y_i = A'(\theta) + \mathrm{diag}[A''(\theta)]^{\frac{1}{2}}\varepsilon_i$, where ε_i have independent standardized entries. Therefore, letting $\mathcal{Y}$ be the $n \times p$ matrix whose rows are $Y_i^\top$, we have $\mathcal{Y} = \vec{1}A'(\theta)^\top + \mathcal{E}\,\mathrm{diag}[A''(\theta)]^{\frac{1}{2}}$, where $\vec{1} = (1, 1, \ldots, 1)^\top$ is the vector of all ones, and $\mathcal{E}$ is an $n \times p$ matrix of independent standardized random variables.

Let H_p be the uniform distribution on the p scalars $A''(\theta(i))$, $i = 1, \ldots, p$. We assume that $A''(\theta(i)) > c$ for some universal constant $c > 0$. In the special case of the Poisson example, this means that the individual rates $x(i)$ are bounded away from 0. The reason for this assumption is to avoid the very sparse regime, where only a few nonzero entries per row are observed. In that case, the MP law is not expected to hold.

Consider the high-dimensional asymptotic limit when $n, p \to \infty$ so that $p/n \to \gamma > 0$. Suppose moreover that H_p converges weakly to some limit distribution, that is, $H_p \Rightarrow H$. Since $\mathrm{diag}[A''(\theta)]$ can be viewed as the population covariance matrix of the noise, H is the limit population spectral distribution (PSD). Since $\mathcal{E}$ has independent standardized entries with bounded moments, it follows that the distribution of the p eigenvalues of $n^{-1}\mathcal{Y}^\top\mathcal{Y}$ converges almost surely to the general Marchenko–Pastur distribution $F_{\gamma,H}$ [Bai and Silverstein (2010), Theorem 4.3].

Now the sample covariance matrix S is a rank-one perturbation of $n^{-1}\mathcal{Y}^\top\mathcal{Y}$. Therefore its eigenvalue distribution also converges to the MP law. We state this for comparison with the next result.

PROPOSITION 4.1 (Marchenko–Pastur law for sample covariance matrix). *The eigenvalue distribution of S converges almost surely to the general Marchenko–Pastur distribution $F_{\gamma,H}$.*

Since the general MP law has a complicated implicit description that needs to be studied numerically [see, e.g., Dobriban (2015)], it is useful to work with the homogenized covariance matrix S_h. Indeed, we establish that the standard Marchenko–Pastur law characterizes its limit spectrum. The standard Marchenko–Pastur distribution has a closed-form density, and there are many useful tools already available for low-rank covariance estimation [e.g., Shabalin and Nobel (2013), Donoho, Gavish and Johnstone (2013)].

THEOREM 4.2 (Marchenko–Pastur law for homogenized covariance matrix). *The eigenvalue distribution of $S_h + I_p$ converges almost surely to the standard Marchenko–Pastur distribution with aspect ratio γ.*

In the proof presented in Appendix A.1.3 [Liu, Dobriban and Singer (2018)], we deduce this from the Marchenko–Pastur law for the error matrix $n^{-\frac{1}{2}}\mathcal{E}$, for which standard results from Bai and Silverstein (2010) apply. The emergence of the standard MP law motivates the shrinkage method presented next.

4.2. *Eigenvalue shrinkage.* Since the early work of Stein [Stein (1956)] it is known that the estimation error of the sample covariance can be decreased by eigenvalue shrinkage. Therefore, we will apply an eigenvalue shrinkage method to the homogenized covariance matrix S_h. Let $\eta(\cdot)$ be a generic matrix shrinker, defined for symmetric matrices M with eigendecomposition $M = U\Lambda U^\top$ as $\eta(M) = U\eta(\Lambda)U^\top$. Here $\eta(\Lambda)$ is defined by applying the scalar shrinker η—typically a nonlinear function—elementwise on the diagonal of the diagonal matrix Λ. Then our *homogenized and shrunken* estimators will have the form

$$S_{h,\eta} = \eta(S_h) = \eta\big(D_n^{-\frac{1}{2}} S_d D_n^{-\frac{1}{2}}\big). \tag{7}$$

We are interested in settings where the clean signals lie on a low-dimensional subspace. We then expect the true covariance matrix Σ_x of the clean signals to be of low rank. However, based on Theorem 4.2, even in the case when $\Sigma_x = 0$, the empirical homogenized covariance matrix is of full rank, and its eigenvalues have an asymptotic MP distribution. We are thus interested in shrinkers η that set all noise eigenvalues to zero, specifically $\eta(x) = 0$ for x within the support of the

shifted MP distribution $x \in [(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2] - 1$. An example is operator norm shrinkage [Donoho, Gavish and Johnstone (2013)].

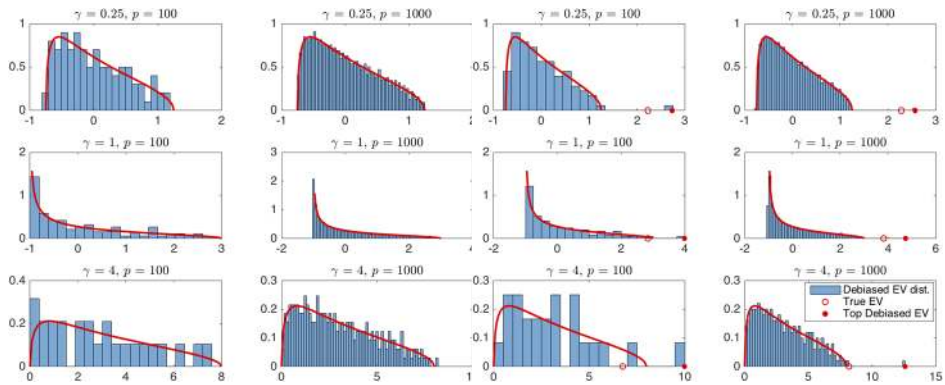
However, homogenization by $D_n \neq I_p$ also changes the direction of the eigenvectors. Therefore, to improve the accuracy of subspace estimates after eigenvalue shrinkage, we *heterogenize*, multiplying back by the estimated standard errors. We define the *heterogenized* covariance estimator as

$$S_{\text{he}} = D_n^{\frac{1}{2}} \cdot S_{h,\eta} \cdot D_n^{\frac{1}{2}}. \tag{8}$$

Heterogenization is a nonlinear operation that changes both the eigenvectors and eigenvalues. While it improves the estimates of the eigenvectors (PCs), it turns out that it introduces a bias in the eigenvalues. Therefore, we will need a final *scaling* step to correct this bias (Section 4.2.3).

To understand homogenization empirically, we perform two simulations. First, we generate nonnegative i.i.d $\{X_i\}_{1 \le i \le n}$ lying in a low-dimensional space of dimension r: we pick r vectors $v_1, \ldots, v_r \in \mathbb{R}^p$ whose coordinates are i.i.d uniformly distributed in $[0, 1]$, and normalize each to have an L1 norm of unity. For each i, sample r coefficients $a_{i1}, \ldots, a_{ir}$ independently from the uniform distribution on $[0, 1]$. Define $X_i = a_{i1}v_1 + \cdots + a_{ir}v_r$. Note that X_i are nonnegative, reside in a hyperplane spanned by $v_1, \ldots, v_r$, and the mean and covariance of X_i can be found easily in terms of $v_1, \ldots, v_r$. The coefficients $a_{i1}, \ldots, a_{ir}$ are also normalized so that $a_{i1} + \cdots + a_{ir} = A$, where $A = 25(1+\sqrt{\gamma})^2$ is a constant relating to signal strength, chosen empirically to push the top eigenvalue outside of the bulk. Finally we sample $Y_i \sim \text{Poisson}_p(X_i)$ independently.

We display a Monte Carlo instance of the eigenvalue histogram of S_h on Figure 2. When $r = 1$, the covariance matrix has rank 0 and the standard MP



(a) Rank 0 covariance matrix (b) Rank 1 covariance matrix (Spiked)

FIG. 2. *Empirical distribution of eigenvalues of homogenized sample covariance S_h for different values of $\gamma = p/n$, with the corresponding shifted Marchenko–Pastur density overlaid as a red curve. Data simulated according to 4.2. In the legend for (b), "Top Debiased EV" refers top eigenvalue of S_h, while "True EV" refers to the top eigenvalue of $D_n^{-\frac{1}{2}} \Sigma_x D_n^{-\frac{1}{2}}$, which we want to estimate.*

TABLE 2
Spiked models: *Summary of the original and homogenized spiked model*

Model	**Original**	**Homogenized**
Latent Signal	$X_i = u + z_i v$	$D^{-\frac{1}{2}} X_i = D^{-\frac{1}{2}} u + z_i D^{-\frac{1}{2}} v$
Marginal Covariance	$\mathrm{Cov}[Y] = vv^\top + D$	$\mathrm{Cov}[Y_h] = D^{-\frac{1}{2}} vv^\top D^{-\frac{1}{2}} + I_p$
Eigenvector	$v_{\mathrm{norm}} = v/\|v\|$	$w = D^{-\frac{1}{2}} v/\|D^{-\frac{1}{2}} v\|$
Spike	$t = v^\top v$	$\ell = v^\top D^{-1} v$
SNR	$\frac{v^\top v}{\mathrm{tr}\, D}$	$\frac{v^\top D^{-1} v}{p}$

distribution—shifted by -1—is a good match [Figure 2(a)]. This is in accordance with Theorem 4.2. When $r = 2$, the covariance matrix has rank 1 and the standard MP distribution still matches the bulk of the noise eigenvalues [Figure 2(b)]. Moreover, we observe the same qualitative behavior as in the classical spiked model, where the top *empirical eigenvalue* overshoots the *population eigenvalue*. Next we study this phenomenon more precisely.

4.2.1. *The spiked model*: *Colored and homogenized*. To develop a method for estimating the eigenvalue after homogenization and heterogenization, we study a generalization of the spiked model [Johnstone (2001)] appropriate for our setting. Specifically, based on the covariance structure of the noisy signal, equation (2), we model the mean parameter $X = A'(\theta)$ of the exponential family—the clean observation—as a low rank vector. For simplicity, we will present the results in the rank one case (summarized in Table 2), but they generalize directly to higher rank.

Suppose that the ith clean observation has the form $X_i = A'(\theta_i) = u + z_i v$, where u, v are deterministic p-dimensional vectors, and z_i are i.i.d. standardized random variables. In the Poisson case where $Y_i \sim \mathrm{Poisson}_p(X_i)$, this assumes that the latent mean vectors are $X_i = u + z_i v$. The vector u is the global mean of the clean images, while v denotes the direction in which they vary.

For X_i to be a valid mean parameter, we need the additional condition that $u(j) + z_i|v(j)| \in A'(\Theta)$, for all i, j, where Θ is the natural parameter space of the exponential family, and $f(S)$ denotes the forward map of the set S under the function f. For instance, in the Poisson case, we need that $X_i(j) \geq 0$ for all i, j. If we take z_i to be uniform random variables on $[-\sqrt{3}, \sqrt{3}]$, so that their variance is unity, then a sufficient condition is that $u(j) \geq \sqrt{3}|v(j)|$ for all j.

Using our formula for the marginal covariance of the noisy observations, $\mathrm{Cov}[Y] = \mathrm{Cov}[X] + \mathbb{E}\,\mathrm{diag}[V(X)]$, and defining $D = \mathbb{E}\,\mathrm{diag}[V(X)]$, we obtain

$$\mathrm{Cov}[Y] = vv^\top + D. \tag{9}$$

For instance, in the Poisson case we have $\mathrm{Cov}[Y] = vv^\top + \mathrm{diag}[u]$.

We homogenize the observations dividing by the elements of $D^{\frac{1}{2}}$. The elements of D are expected values of variances. They are thus positive, except for coordinates that can be discarded because they have no variability. The homogenized observations are $Y_h = D^{-\frac{1}{2}} Y$, and their population covariance matrix is

$$\operatorname{Cov}[Y_h] = D^{-\frac{1}{2}} v v^\top D^{-\frac{1}{2}} + I_p. \tag{10}$$

We now compare this with the usual *standard spiked model* [Johnstone (2001)] where the observations Y_h are Gaussian and have covariance matrix $\operatorname{Cov}[Y_h] = \ell w w^\top + I_p$, where $\ell \geq 0$ and the vector w has unit norm. The top eigenvalue is called the "spike." This model has been thoroughly studied in probability theory and statistics. In particular, the Baik–Ben Arous–Péché (BBP) phase transition (PT) [Baik, Ben Arous and Péché (2005)] shows that when $n, p \to \infty$ such that $p/n \to \gamma > 0$, the top eigenvalue of the sample covariance matrix asymptotically separates from the Marchenko–Pastur bulk if the population spike $\ell > \sqrt{\gamma}$. Otherwise, the top sample eigenvalue does not separate from the MP bulk. This was shown first for complex Gaussian observations, then generalized to other distributions [see, e.g., Yao, Zheng and Bai (2015)].

Heuristically, comparing with (10), we surmise that a spiked model with $\ell = v^\top D^{-1} v$ and $w = D^{-\frac{1}{2}} v / \| D^{-\frac{1}{2}} v \|$ is a good approximation in our case. In particular the BBP phase transition should happen approximately when $v^\top D^{-1} v = \sqrt{\gamma}$. In the Poisson case the condition is $v^\top \operatorname{diag}[u]^{-1} v = \sqrt{\gamma}$. Next we provide numerical evidence for this surmise, and develop its consequences.

4.2.2. *Homogenization improves SNR.* In this section we justify our homogenization method theoretically, showing that it can improve the signal-to-noise ratio. This was observed empirically in previous work on covariance estimation in a related setting, but a theoretical explanation is lacking [Bhamre, Zhang and Singer (2016)].

As usual, we define the SNR of a "signal + noise" vector observation $y = s + n$ as the ratio of the trace of the covariances of s and of n. In the unhomogenized model from equation (9)

$$\mathrm{SNR} = \frac{\operatorname{tr} \operatorname{Cov}[X]}{\operatorname{tr} \mathbb{E} \operatorname{diag}[V(X)]} = \frac{\operatorname{tr} v v^\top}{\operatorname{tr} D} = \frac{v^\top v}{\operatorname{tr} D}.$$

In particular, the SNR is of order $O(1/p)$ in the typical case when the vector v has norm of unit order. In the homogenized model from equation (10), the SNR equals $v^\top D^{-1} v / p$.

Suppose now that v is approximately *delocalized* in the sense that $p \cdot v^\top D^{-1} v \approx \operatorname{tr} D^{-1} \cdot v^\top v$. This holds for instance if the entries of v are i.i.d. centered random variables with the same variance σ^2. In that case, $\mathbb{E} v^\top D^{-1} v = \sigma^2 \operatorname{tr} D^{-1}$ and $\mathbb{E} v^\top v = \sigma^2 p$, and under higher moment assumptions it is easy to show the concentration of these quantities around their means, showing delocalization as above. If

v is delocalized, then we obtain that the SNR in the homogenized model is higher than in the original model. Indeed, this follows because D is diagonal, so by the Cauchy–Schwarz inequality

$$\frac{v^\top D^{-1} v}{p} \approx \frac{\operatorname{tr} D^{-1} \cdot v^\top v}{p^2} = \frac{\sum_{i=1}^p D_i^{-1} \cdot v^\top v}{p^2} \geq \frac{v^\top v}{\sum_{i=1}^p D_i} = \frac{v^\top v}{\operatorname{tr} D}.$$

Moreover, we can define the *improvement* (or *amplification*) in SNR as

$$\mathcal{I} = \frac{\operatorname{tr} D}{p} \cdot \frac{v^\top D^{-1} v}{v^\top v}. \tag{11}$$

The above heuristic can be formalized as follows:

PROPOSITION 4.3. *Suppose the signal eigenvector v is delocalized in the sense that for some $\varepsilon > 0$,*

$$\frac{v^\top D^{-1} v}{v^\top v} \geq (1 - \varepsilon) \frac{\operatorname{tr}[D^{-1}]}{p}.$$

Let moreover β be the following measure of heteroskedasticity:

$$\beta = \frac{\sum_{i=1}^p D_i \cdot \sum_{i=1}^p D_i^{-1}}{p^2} \geq 1.$$

Then the SNR is improved by homogenization, by a ratio $\mathcal{I} \geq (1 - \varepsilon)\beta$.

If β is large and $\varepsilon > 0$ is small, the SNR can improve substantially.

4.2.3. *Eigenvalue shrinkage and scaling*. We now continue with our overall goal of estimating the covariance matrix $\operatorname{Cov}[X] = vv^\top$ of X. This has one nonzero eigenvalue $t = \|v\|^2$ and corresponding eigenvector $v_{\text{norm}} = v/\|v\|$. We use the top eigenvector of the heterogenized covariance matrix S_{he} as an estimator of v_{norm}. To estimate t, a first thought is to use the top empirical eigenvalue of S_{he}, but as we show next, this naive estimator is biased. To correct the bias, we will leverage recent results from random matrix theory. The need for this step can be understood by leveraging deep insights from that area, specifically precise asymptotic results on the inconsistency of eigenvectors, presented below.

For data with independent coordinates and equal variances, the cumulative work of many authors [e.g., Baik, Ben Arous and Péché (2005), Paul (2007), Baik and Silverstein (2006), Benaych-Georges and Nadakuditi (2011) etc.] shows that if the population spike ℓ is above the BBP phase transition—that is, $\ell > \sqrt{\gamma}$—then the top sample spike pops out from the Marchenko–Pastur distribution of the "noise" eigenvalues. The top eigenvalue will converge to the value given by *the spike forward map*:

$$\lambda(\ell; \gamma) = \begin{cases} (1 + \ell)\left(1 + \dfrac{\gamma}{\ell}\right) & \text{if } \ell > \gamma^{\frac{1}{2}}, \\ \left(1 + \gamma^{\frac{1}{2}}\right)^2 & \text{otherwise.} \end{cases}$$

We conjecture that the BBP phase transition also applies to our case, and describes the behavior of the spikes after homogenization. We have verified this in numerical simulations in certain cases (data not shown due to space limitations). Therefore, as in previous work, we propose to estimate ℓ consistently by inverting the spike forward map [see, e.g., Lee, Zou and Wright (2010), Donoho, Gavish and Johnstone (2013)], that is, defining $\hat{\ell} = \lambda^{-1}(\lambda_{\max}(S_h))$. Donoho, Gavish and Johnstone (2013) provided an asymptotic optimality result for this estimator of the spike in operator norm loss.

Once we have a good estimator $\hat{\ell}$ of $\ell = v^\top D^{-1} v$, a first thought is to estimate $t = v^\top v$ as the top eigenvalue of the heterogenized covariance matrix S_{he}. However, this estimator is biased. The estimation accuracy is affected in a significant way by the inconsistency of the empirical eigenvector $\hat{w}$ of S_h as an estimator of the true eigenvector $w = D^{-\frac{1}{2}} v / \|D^{-\frac{1}{2}} v\|$. We can quantify this heuristically based on results for Gaussian data. In the Gaussian standard spiked model the empirical and true eigenvectors have an asymptotically deterministic angle: $(w^\top \hat{w})^2 \to c^2(\ell; \gamma)$ almost surely, where $c(\ell; \gamma)$ is the *cosine forward map* given by [e.g., Paul (2007), Benaych-Georges and Nadakuditi (2011) etc.]:

$$c(\ell; \gamma)^2 = \begin{cases} \dfrac{1 - \gamma/\ell^2}{1 + \gamma/\ell} & \text{if } \ell > \gamma^{\frac{1}{2}}, \\ 0 & \text{otherwise.} \end{cases}$$

Heuristically, in finite samples we can write $\hat{w} \approx cw + s\varepsilon$, where $s = s(\ell; \gamma) \geq 0$ is the sine defined by $s^2 = 1 - c^2$, and ε is white noise with approximate norm $\|\varepsilon\| = 1$. Then, since $w^\top D w = v^\top v / v^\top D^{-1} v = t/\ell$, and $\varepsilon^\top D \varepsilon \approx \operatorname{tr}(D)/d$, we have

$$\begin{aligned} \|\hat{v}\|^2 &\approx \ell \cdot \hat{w}^\top D \hat{w} \approx \ell \cdot (cw + s\varepsilon)^\top D (cw + s\varepsilon) \\ &\approx \ell \cdot \left(c^2 w^\top D w + s^2 \varepsilon^\top D \varepsilon\right) \approx t c^2 + \ell s^2 \operatorname{tr}(D)/p. \end{aligned}$$

Comparing this to $\|v\|^2 = t = tc^2 + ts^2$, we find that the bias is

$$\|\hat{v}\|^2 - t \approx s^2 \left(v^\top D^{-1} v \cdot \operatorname{tr}(D)/p - v^\top v\right) = s^2 t \cdot (\mathcal{I} - 1) \geq 0.$$

This suggests that $\|\hat{v}\|^2$ is an upward biased estimator of $t = \|v\|^2$. Interestingly, the bias is closely related to the improvement $\mathcal{I}$ in SNR. Moreover, this calculation makes it clear that the bias comes from the inconsistency of the eigenvectors in high dimensions. Thus, the random matrix theoretic results characterizing the inconsistency are crucial in this step.

To correct the bias, we propose an estimator of the form $\hat{t}(\alpha) = \alpha \|\hat{v}\|^2$ for which $\alpha \|\hat{v}\|^2 \approx \|v\|^2$. We have $\|\hat{v}\|^2 \approx t \cdot [1 + s^2(\mathcal{I} - 1)]$, suggesting that we define $\alpha = [1 + s^2(\mathcal{I} - 1)]^{-1}$. This quantity is an unknown population parameter, and it depends on s^2 and $\mathcal{I}$. We can estimate s^2 in the usual way by $\hat{s}^2 = s^2(\hat{\ell}; \gamma)$. Since $\mathcal{I}$ itself depends on the parameter t we are trying to estimate, we plug in the same

estimator $\hat{t}(\alpha) = \alpha\|\hat{v}\|^2$, leading to the following estimator of $\mathcal{I}$ (where we also define τ for future use):

$$\hat{\mathcal{I}}(\alpha) = \frac{\operatorname{tr} D_n}{p} \cdot \frac{\hat{\ell}}{\hat{t}(\alpha)} = \frac{\operatorname{tr} D_n}{p} \cdot \frac{\hat{\ell}}{\alpha\|\hat{v}\|^2} = \frac{\tau}{\alpha}.$$

Since $\alpha = [1+s^2(\mathcal{I}-1)]^{-1}$, it is reasonable to require that the fixed-point equation $\hat{\alpha} = [1+\hat{s}^2(\hat{\mathcal{I}}(\hat{\alpha})-1)]^{-1}$ holds.

We can equivalently rewrite the fixed-point equation as $1/\hat{\alpha} = \hat{c}^2 + \hat{s}^2\hat{\mathcal{I}}(\hat{\alpha}) = \hat{c}^2 + \hat{s}^2\tau/\hat{\alpha}$. Or, when $\hat{c}^2 > 0$,

$$\hat{\alpha} = \frac{1-\hat{s}^2\tau}{\hat{c}^2}. \tag{12}$$

When $\hat{c}^2 = 0$, that is, when $\hat{\ell} \le \sqrt{\gamma}$, the equation reads $1/\hat{\alpha} = \tau/\hat{\alpha}$. If $\tau = 1$, this has solution $\alpha = 1$, else it has no solution. Therefore, when $\hat{c}^2 = 0$, we define $\hat{\alpha} = 1$. We finally define $\hat{t}(\hat{\alpha}) = \hat{\alpha}\|\hat{v}\|^2$. The implication is that we ought to rescale the estimated magnitude of the signal subspace corresponding to v by $\hat{\alpha}$.

In the multispiked case, suppose $X_j = u + \sum_{i=1}^r z_{ij} v_i$. Then the marginal covariance of Y is $\operatorname{Cov}[Y] = \sum_{i=1}^r v_i v_i^\top + D$. Suppose that the v_i are sorted in the order of decreasing norm. Suppose moreover that the heterogenized sample covariance S_{he} has the form $S_{\mathrm{he}} = \sum_{i=1}^r \hat{v}_i \hat{v}_i^\top = \sum_{i=1}^r \hat{\lambda}_i \hat{u}_i \hat{u}_i^\top$, where $\hat{u}_i$ are orthonormal, and the $\hat{\lambda}_i \ge 0$ are sorted in decreasing order. Based on our above discussion, we define the *scaled* covariance matrix as

$$S_s = \sum_{i=1}^r \hat{\alpha}_i \hat{v}_i \hat{v}_i^\top, \tag{13}$$

where $\hat{\alpha}_i$ is defined in (12), with $\hat{s}^2 = \hat{s}_i^2 = s^2(\hat{\ell}_i; \gamma)$. This concludes our methodology for covariance estimation. We use the terminology *e*PCA for the eigendecomposition of the covariance matrix estimator (13). Both the eigenvalues and the eigenvectors of this estimator are different from those of the sample covariance matrix.

*e*PCA is summarized in Algorithm 1. Clearly, *e*PCA is applicable when the variables $x(i)$ have known nonidentical distributions, which the modification that homogenization should be done by the mean-variance map of the distribution of each particular coordinate. As discussed at the beginning of Section 4.2, we assume here that we have a guess r for the number of PCs. In exploratory analyses, one can often try several choices for r. While there are many formal methods for choosing the rank r [see, e.g., Jolliffe (2002)], it is beyond our scope to investigate them in detail here.

Algorithm 1: Covariance matrix estimation and *e*PCA

Input: Data $Y = [Y_1, \dots, Y_n]^\top \in \mathbb{R}^{n \times p}$; Desired rank $r \le p$;
Mean-variance map V of exponential family, as defined in (3).
Output: Covariance estimator $S_s \in \mathbb{R}^{p \times p}$ of noiseless vectors; *e*PCA: eigendecomposition of S_s.

1 Compute the sample mean $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$
2 Compute the sample covariance matrix $S = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^\top$
3 Compute the variance estimates $D_n = \operatorname{diag}[V(\bar{Y})]$
4 Homogenize and diagonally debias the covariance matrix $S_h = D_n^{-\frac{1}{2}} S D_n^{-\frac{1}{2}} - I_p$
5 Compute the eigendecomposition $S_h = \hat{W} \Lambda \hat{W}^\top$
6 Shrink the eigenvalues $S_{h,\eta} = \hat{W} \eta(\Lambda_r) \hat{W}^\top = \sum_{i=1}^r \hat{\ell}_i \hat{w}_i \hat{w}_i^\top$ of top r eigenvalues $\Lambda_r = \operatorname{diag}(\lambda_1, \dots, \lambda_r)$.
7 Compute the scaling coefficients $\hat{\alpha}_i = [1 - s^2(\hat{\ell}_i; \gamma)\tau_i]/c^2(\hat{\ell}_i; \gamma)$ [as in (12)]
8 Heterogenize the covariance matrix $S_{\mathrm{he}} = D_n^{\frac{1}{2}} S_{h,\eta} D_n^{\frac{1}{2}}$
9 Scale the covariance matrix $S_s = \sum \hat{\alpha}_i \hat{v}_i \hat{v}_i^\top$, where the eigendecomposition of S_{he} is $S_{\mathrm{he}} = \sum \hat{v}_i \hat{v}_i^\top$

4.2.4. *Simulations with ePCA.* We report the results of a simulation study with *e*PCA. As an example, we simulate data Y_i from a Poisson model. We remind the reader that our algorithm works for all exponential families. We let $Y_i \sim \mathrm{Poisson}_p(X_i)$, where the mean parameters are $X_i = u + z_i \ell^{\frac{1}{2}} v$, the z_i are i.i.d. unit variance random variables uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$, and $u \in \mathbb{R}^p$ has entries $u(i)$ sorted in increasing order on a uniform grid on $[1, 3]$, while $v \in \mathbb{R}^p$ has entries $v(i)$ sorted in increasing order on a uniform grid on $[-1, 1]$, standardized so that $\|v\|^2 = 1$. We take the dimension $p = 500$, and $\gamma = \frac{1}{2}$, so $n = 1000$. The phase transition occurs when the spike is $\ell = \sqrt{\gamma}/v^\top \operatorname{diag}[u]^{-1} v \approx 1.2$. We vary the spike strength ℓ on a uniform grid of size 20 on $[0, 3]$. We generate $n_M = 100$ independent Monte Carlo trials, and compute the mean of the heterogenized spike estimator $\hat{t} = \|\hat{v}\|^2$ and the *e*PCA—or scaled—estimator $\hat{t}(\hat{\alpha}) = \hat{\alpha}\|\hat{v}\|^2$.

The results displayed in Figure 3 (left) show that the *e*PCA/scaled estimator (top eigenvalue of S_s) reduces the bias of the heterogenized estimator (top eigenvalue of S_{he}) especially for large spikes. Both are much better than the debiased estimator (top eigenvalue of S_d). Below the phase transition (vertical line), both estimators have the same approximate value.

We can also define an estimator of the improvement in SNR $\mathcal{I}$, as $\hat{\mathcal{I}}(\hat{\alpha})$. The mean of this estimator over the same simulation is displayed in Figure 3 (middle). We observe that it is approximately unity below the PT. This makes sense, because the spike is below the PT both before and after homogenization. The improvement

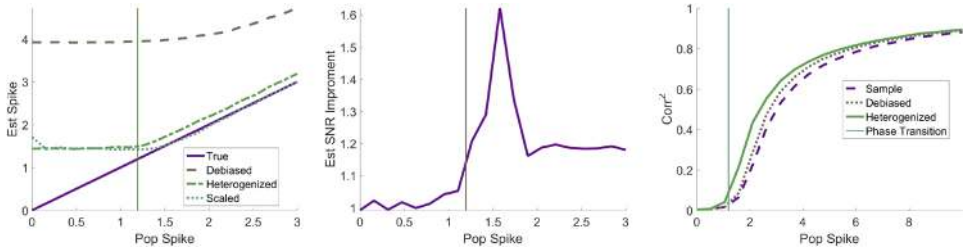


FIG. 3. *Simulation with ePCA. Left*: *Spike estimation*; *true*, *debiased*, *heterogenized*, *and scaled* (*ePCA*). *Middle*: *Estimated improvement in SNR due to homogenization. Right*: *Squared correlation between v and leading eigenvector of various covariance estimates*; *sample*, *debiased*, *heterogenized* (*ePCA*). *Plotted against the spike.*

in SNR has a "jump" just above the PT, because the spike pops out from the bulk after homogenization. This is where homogenization helps the most. However, $\hat{\mathcal{I}}$ is not "infinitely large" because the signal is detectable in the unhomogenized spectrum, except it is spread across all eigenvalues [see, e.g., Dobriban (2017)]. Finally, $\hat{\mathcal{I}}(\hat{\alpha})$ drops to a lower value, still above unity, and stabilizes. We find this an illuminating way to quantify the improvement due to homogenization.

Finally, we also display the mean of the squared correlation between the true and empirical eigenvectors of various covariance estimators in Figure 3 (right). The predicted PT matches the empirical PT. The *e*PCA eigenvector—top eigenvector of S_s—in this case agrees with the eigenvector of the heterogenized covariance matrix S_{he}, because both are of rank one. *e*PCA has the highest correlation, and the improvement is significant just above the PT.

4.3. *Homogenization agrees with HWE normalization.* It is of special interest that for Binomial(2) data, and specifically for biallelic genetic markers such as Single Nucleotide Polymorphisms, our homogenization method recovers exactly the well-known normalization assuming Hardy–Weinberg equilibrium (HWE). In these datasets the entries X_{ij} are counts ranging from 0 to 2 denoting the number of copies of the variant allele of biallelic marker j in the genome of individual i. The HWE normalization divides the entries of SNP j by $\sqrt{2\hat{p}_j(1-\hat{p}_j)}$, where $\hat{p}_j = (2n)^{-1}\sum_i X_{ij}$ is the estimated allele frequency of variant j [e.g., Patterson, Price and Reich (2006), page 2075]. It is easy to see that this is exactly the same as our homogenization method assuming that the individual data points X_{ij} are Binomial(2)-distributed.

Previously, the HWE normalization was motivated by a connection to genetic drift, and by the empirical observation that it improves results on observational and simulated data [Patterson, Price and Reich (2006), p. 2075]. Our theoretical results justify HWE normalization. In particular, our Theorem 4.2 suggests that the Marchenko–Pastur is an accurate null distributions after homogenization.

Numerical results also suggest that the approximations to both the MP law and the Tracy–Widom distribution for the top eigenvalue are more accurate than after standardization (data not shown for space reasons). In addition, our result on the improved SNR (Proposition 4.3) suggests that "signal" becomes easier to identify after homogenization.

However, in practice we often see similar results with homogenization and standardization. In many SNP datasets, the variants not approximately in HWE—that is, the variants for which a goodness of fit test to a Binomial(2) distribution is rejected—are removed as part of data quality control. Therefore, most remaining SNPs have an empirical distribution well fit by a Binomial(2). In such cases standardization and homogenization lead to similar results.

5. Denoising. As an application of *e*PCA, we develop a method to denoise the observed data. Formally the goal of denoising is to predict the noiseless signal vectors $X_i = A'(\theta_i)$. Our model is a random effects model [see, e.g., Searle, Casella and McCulloch (1992)], hence we predict X_i using the Best Linear Predictor—or BLP [Searle, Casella and McCulloch (1992), Section 7.4]. Let $\tilde{\mathbb{E}}(X|Y) = BY + C$ denote the minimum MSE linear predictor of the random vector X using Y, where B is a deterministic matrix, and C is a deterministic vector. This is known under various names, including the *Wiener filter*; see Section 1.3. We will refer to it as the BLP, which is the common terminology in random effects models. It is well known [e.g., Searle, Casella and McCulloch (1992), Section 7.4] that

$$B = \Sigma_x\big[\operatorname{diag}\big[\mathbb{E}A''(\theta)\big] + \Sigma_x\big]^{-1} \quad \text{and}$$

$$C = \operatorname{diag}\big[\mathbb{E}A''(\theta)\big]\big[\operatorname{diag}\big[\mathbb{E}A''(\theta)\big] + \Sigma_x\big]^{-1}\mathbb{E}A'(\theta).$$

The BLP depends on the unknown parameters Σ_x, $\operatorname{diag}[\mathbb{E}A''(\theta)]$, and $\mathbb{E}[A'(\theta)]$. The standard strategy, known as *Empirical BLP* or EBLP [e.g., Searle, Casella and McCulloch (1992)] is to estimate these unknown parameters using the entire dataset, and denoise the vectors Y_i by plug-in:

$$\hat{X}_i = \hat{\Sigma}_x\big[\operatorname{diag}\big[\hat{\mathbb{E}}A''(\theta)\big] + \hat{\Sigma}_x\big]^{-1}Y_i + \operatorname{diag}\big[\hat{\mathbb{E}}A''(\theta)\big]\big[\operatorname{diag}\big[\hat{\mathbb{E}}A''(\theta)\big] + \hat{\Sigma}_x\big]^{-1}\bar{Y}.$$

We will use *e*PCA, that is, the scaled covariance matrix S_s proposed in (13) to estimate Σ_x. As before in Section 3.2, we will use the sample mean $\bar{Y}$ to estimate $\mathbb{E}[A'(\theta)]$, and $V(\bar{Y})$ to estimate the noise variances $\mathbb{E}A''(\theta)$. However, in principle different estimators could be used.

For the special case of the Poisson distribution, we have

$$\hat{X}_i = S_s\big(\operatorname{diag}[\bar{Y}] + S_s\big)^{-1}\hat{Y}_i + \operatorname{diag}[\bar{Y}]\big(\operatorname{diag}[\bar{Y}] + S_s\big)^{-1}\bar{Y}.$$

In some examples there are coordinates where $\bar{Y}(j) = 0$. In our XFEL application this corresponds to pixels where no photon was observed during the entire experiment. This causes a problem because the matrix $\widehat{\Sigma} = \operatorname{diag}[\bar{Y}] + S_s$ may no

longer be invertible: S_s is of low rank, while diag[$\bar{Y}$] is also not of full rank. To avoid this problem, we implement a ridge-regularized covariance estimator $\widehat{\Sigma}_\varepsilon = (1-\varepsilon)\widehat{\Sigma} + \varepsilon \cdot \tilde{m} I_p$ as in Ledoit and Wolf (2004), where $\tilde{m} = \operatorname{tr} \widehat{\Sigma}/p$ and $\varepsilon > 0$ is a small constant. Note that $\operatorname{tr} \widehat{\Sigma}_\varepsilon = \operatorname{tr} \widehat{\Sigma}$. The ridge-regularized estimator $\widehat{\Sigma}_\varepsilon$ has a small bias, but is invertible. In our default implementation we take $\varepsilon = 0.1$. Similar results are achieved in our XFEL application for ε in the range of 0.05–0.2. In new applications we suggest that the user try this range of ε and choose one based on empirical performance. The same method can be implemented for any exponential family. Another potential solution to the invertibility problem—not pursued here—is to discard the pixels with $\bar{Y}(j) = 0$.

6. Experiments. We conduct a substantive XFEL data analysis example using *e*PCA, and compare with PCA. We generate $n_0 = 70{,}000$ noiseless XFEL diffraction intensity maps of a lysozyme (Protein Data Bank 1AKI) with physical realism using the state of the art methods [Hantke, Ekeberg and Maia (2016)]. We rescale the average pixel intensity to 0.04 such that shot noise dominates, as suggested in prior work [e.g., Schwander et al. (2012)]. To sample an arbitrary number n of noisy diffraction patterns, we sample an 64-pixel-by-64-pixel intensity map at random, and then sample the photon count of each detector pixel from a Poisson distribution whose mean is the pixel intensity. The resulting images are also 64 pixels by 64 pixels, so $p = 4096$. Figure 1 illustrates the intensity maps and the resulting noisy diffraction patterns.

6.1. *Covariance estimation.* To evaluate performance on covariance estimation, we vary the sample size n in the range $3 \le \log_{10}(n) \le 5$. We fix the rank of each estimator to be 10, though other choices lead to similar results. The diagonally debiased, heterogenized, and scaled covariance estimates S_d, S_{he}, S_s each improve on the sample covariance S (Figure 4) in MSE. The largest improvement is due to diagonal debiasing, but scaling leads to the smallest MSE.

Figure 5 summarizes the error of eigenvalue estimation. The *e*PCA eigenvalues are indeed much closer to the true eigenvalues than the eigenvalues of the debiased or sample covariance matrices S_d or S. The estimation error for *e*PCA eigenvalues is small regardless of sample size.

We visualize the eigenvectors (or eigenimages) for XFEL diffraction patterns in Figure 6. The *e*PCA eigenvectors—those of the heterogenized matrix S_{he}—accurately estimate two more eigenimages with small eigenvalues than alternative methods. This shows that *e*PCA significantly improves on PCA for covariance estimation in XFEL data.

The *e*PCA/heterogenized eigenvectors 1 to 2 in Figure 6 appear misaligned with the corresponding true eigenvectors. A likely explanation is that the top eigenvectors have similar eigenvalues, leading to some reordering and rotation in the estimated eigenvectors. Therefore, we also report the error of estimating the overall

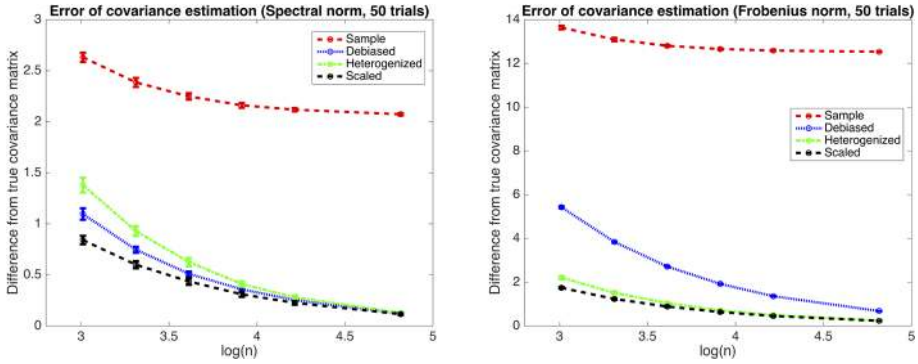


FIG. 4. *Error of covariance matrix estimation, measured as the spectral norm (left) and Frobenius norm (right) of the difference between each covariance estimate (Sample, Debiased, Heterogenized, Scaled) and the true covariance matrix.*

low-rank subspace, for rank $r = 10$, measured as the estimation MSE of the projection matrix $U_r U_r^T$. Other values of r lead to comparable results. Figure 7 clearly shows that the *e*PCA/heterogenized covariance matrix best estimates the low-rank subspace inhabited by the clean data.

6.2. *Denoising.* Finally, we report the results of denoising the XFEL patterns. We compare "PCA denoising" or "vanilla projection," that is, orthogonal projection onto sample or *e*PCA/heterogenized eigenimages; and EBLP denoising. PCA denoising results in clear artifacts, while the reconstructions after EBLP denoising are always the closest to the clean images (Figure 8). In EBLP denoising, our

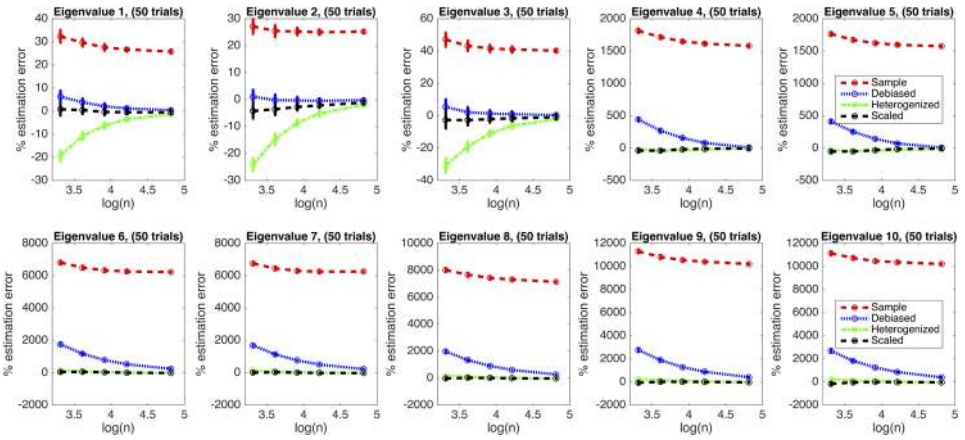


FIG. 5. *Error of eigenvalue estimation for the top five eigenvalues, measured as percentage error relative to the true eigenvalue, for XFEL data. We plot the mean and standard deviation (as error bars) over* 50 *Monte Carlo trials.*

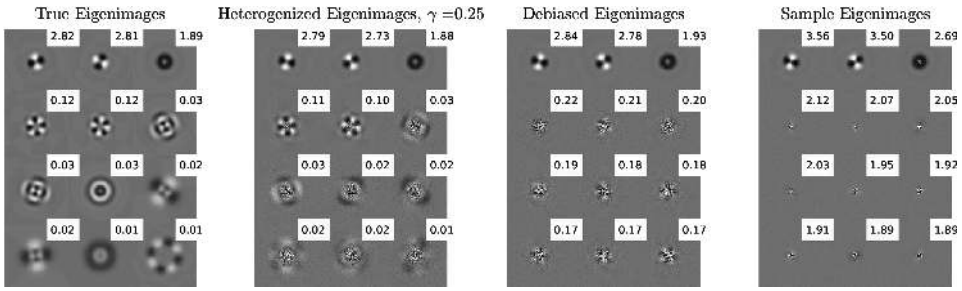


FIG. 6. *XFEL Eigenimages for* $\gamma = \frac{1}{4}$, *ordered by eigenvalue.*

scaled covariance matrix leads to much better results than the sample covariance matrix. EBLP also does better when measured by reconstruction mean squared error,

$$\text{MSE} := (pn)^{-1} \sum_{i=1}^{n} \|\hat{X}_i - X_i\|^2.$$

We also compare *e*PCA to the exponential family PCA method based on alternating minimization proposed by Collins, Dasgupta and Schapire (2001) in Figure 9. *e*PCA is faster and recovers the images with higher accuracy, as measured by MSE (see the caption of Figure 9). Our experiments with variance stabilizing transforms, such as the Anscombe [Anscombe (1948)] and Freeman–Tukey transforms [Freeman and Tukey (1950)], all gave denoising results significantly worse than standard PCA (results not shown due to space limitations). This may be be-

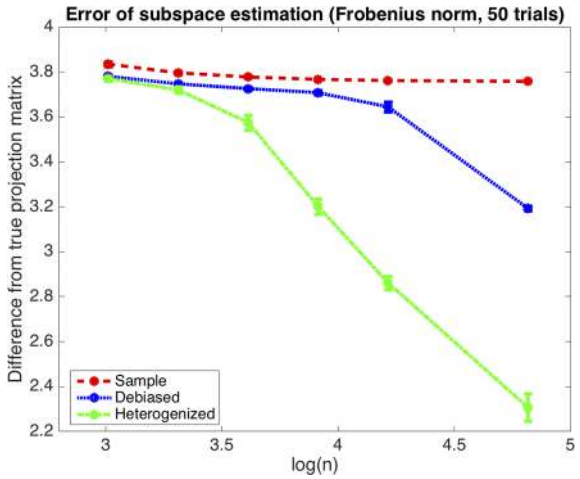


FIG. 7. *Subspace estimation error for XFEL data. We plot the mean and standard deviation* (*as error bars*) *over* 50 *Monte Carlo trials.*

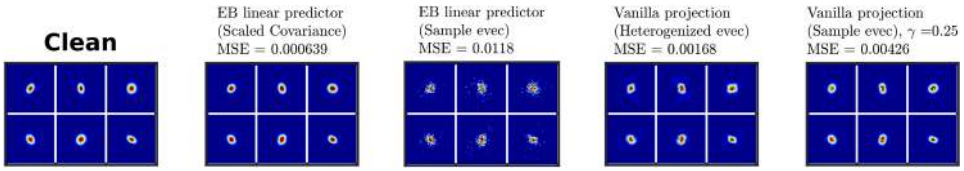


FIG. 8. *Sampled reconstructions using the XFEL dataset* ($n = 16{,}384$; $p = 4096$), *fixing the rank of covariance estimates at* $r = 10$. *Color scale of each reconstruction clipped to match that of clean images*.

cause the known inverse transforms [e.g., Mäkitalo and Foi (2011)] are ineffective in the photon-limited regime.

In conclusion, *e*PCA is accurate for the covariance estimation and PCA in XFEL data analysis. It also works well for denoising XFEL diffraction patterns.

7. Future work. In the context of XFEL imaging, each diffraction pattern is equally likely to appear in any possible in-plane rotation. As a result, the covariance matrix commutes with rotations and is block diagonalized in any basis made of outer products of radial functions and angular Fourier modes, such as the Fourier–Bessel basis [Zhao, Shkolnisky and Singer (2016)]. Indeed, the eigenimages in Figure 5 clearly show angular oscillation. Incorporating "steerability" into our methodology, that is, including the block diagonal structure into the estimation framework would lead to more accurate covariance estimation, as it effectively reduces the dimension p.

Furthermore, it would be valuable to prove rigorously the results about the spiked model for exponential families. Our results in Section 4.1.1 only cover the null case, but it would be useful to know rigorously the behavior of the signal eigenvalues in nonnull cases. In addition, it could be useful to have a principled method to choose the rank. Moreover, it would also lead to a better theoretical understanding of our scaled estimator S_s.

Finally, it would be important to go beyond the one-parameter exponential family considered in this paper. For instance, a lot of nominally Poisson problems have

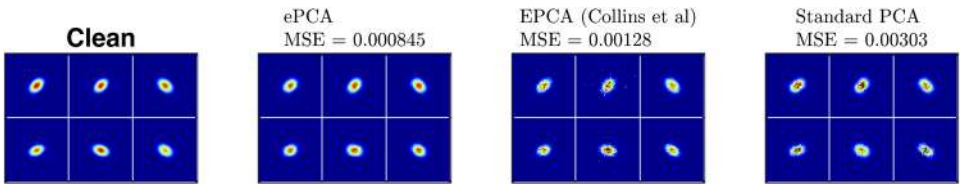


FIG. 9. *Comparing various methods' sampled reconstructions of the XFEL dataset* ($n = 1000$; $p = 4096$), *fixing the rank estimate for each method to* $r = 8$. *For reference, the MSE for noisy images is* 0.0401. *We also note that ePCA took* 13.9 *seconds, while Collins, Dasgupta and Schapire (2001)'s exponential family PCA took* 10,900 *seconds, or three hours, to finish running on a* 2.7 *GHz Intel Core i5 processor.*

overdispersion. Our model allows some overdispersion, because the parameter θ can be random. To handle overdispersion more generally, it would be of interest to extend *e*PCA to the setting where the mean-variance map V is unknown but estimable. This would allow us to handle overdispersed data with low-rank structure more flexibly, for example, when the variance is an unknown function of the mean [Anders and Huber (2010)].

Acknowledgments. The authors are grateful to Yuval Kluger and Art Owen for helpful comments on an earlier version of the manuscript. They wish to thank Joey Arthur, Nick Patterson, Patrick Perry, Peter Schwander, Joel Tropp, Ramon van Handel, Jingshu Wang, Teng Zhang, and Jane Zhao for valuable discussions. They thank Julia Fukuyama, Susan Holmes, Lan Huong Nguyen, and Kris Sankaran for valuable feedback on *e*PCA software and applications. They thank Filipe Maia, Max Hantke, and Benjamin Rose for help with software.

SUPPLEMENTARY MATERIAL

Appendix (DOI: 10.1214/18-AOAS1146SUPP; .pdf). Technical derivations and proofs.

REFERENCES

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*. **11** R106.

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

ANSCOMBE, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35** 246–254. MR0028556

BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. Springer, New York. MR2567175

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

BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal*. **97** 1382–1408. MR2279680

BARTHOLOMEW, D. J. and KNOTT, M. (1999). *Latent Variable Models and Factor Analysis*, 2nd ed. *Kendall's Library of Statistics* **7**. Edward Arnold, London. MR1711686

BASRI, R. and JACOBS, D. W. (2003). Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Mach. Intell*. **25** 218–233.

BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math*. **227** 494–521. MR2782201

BERGMANN, U., YACHANDRA, V. and YANO, J., eds. (2017). *X-Ray Free Electron Lasers*. The Royal Society of Chemistry, Croydon.

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

BIGOT, J., DELEDALLE, C. and FÉRAL, D. (2016). Generalized SURE for optimal shrinkage of singular values in low-rank matrix denoising. Preprint. Available at arXiv:1605.07412.

BOUCHERON, S., BOUSQUET, O., LUGOSI, G. and MASSART, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** 514–560. MR2123200

CAO, Y. and XIE, Y. (2014). Low-rank matrix recovery in Poisson noise. In *Signal and Information Processing* (*GlobalSIP*), 2014 *IEEE Global Conference on* 384–388. IEEE, New York.

CHEN, X. and STOREY, J. D. (2015). Consistent estimation of low-dimensional latent structure in high-dimensional data. Preprint. Available at arXiv:1510.03497.

COLLINS, M., DASGUPTA, S. and SCHAPIRE, R. (2001). A generalization of principal component analysis to the exponential family. *Advances in Neural Information Processing Systems* (*NIPS*).

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. and HARSHMAN, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* **41** 391–407.

DOBRIBAN, E. (2015). Efficient computation of limit spectra of sample covariance matrices. *Random Matrices Theory Appl*. **4** 1550019, 36. MR3418848

DOBRIBAN, E. (2017). Sharp detection in PCA under correlations: All eigenvalues matter. *Ann. Statist*. **45** 1810–1833. MR3670197

DONOHO, D., GAVISH, M. and JOHNSTONE, I. (2013). Optimal shrinkage of eigenvalues in the spiked covariance model. Preprint. Available at arXiv:1311.0851.

FAVRE-NICOLIN, V., BARUCHEL, J., RENEVIER, H., EYMERY, J. and BORBÉLY, A. (2015). XTOP: High-resolution X-ray diffraction and imaging. *Journal of Applied Crystallography* **48** 620–620.

FREEMAN, M. F. and TUKEY, J. W. (1950). Transformations related to the angular and the square root. *Ann. Math. Stat*. **21** 607–611. MR0038028

FURNIVAL, T., LEARY, R. K. and MIDGLEY, P. A. (2017). Denoising time-resolved microscopy image sequences with singular value thresholding. *Ultramicroscopy* **178** 112–124.

HANTKE, M. F., EKEBERG, T. and MAIA, F. R. N. C. (2016). Condor: A simulation tool for Flash X-Ray imaging. *Journal of Applied Crystallography* **49** 1356–1362.

HUBER, P., RONCHETTI, E. and VICTORIA-FESER, M.-P. (2004). Estimation of generalized linear latent variable models. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **66** 893–908. MR2102471

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist*. **29** 295–327. MR1863961

JOLLIFFE, I. T. (2002). *Principal Component Analysis*, 2nd ed. Springer, New York. MR2036084

JOSSE, J. and WAGER, S. (2016). Bootstrap-based regularization for low-rank matrix estimation. *J. Mach. Learn. Res*. **17** 1–29. MR3555015

KAM, Z. (1977). Determination of macromolecular structure in solution by spatial correlation of scattering fluctuations. *Macromolecules* **10** 927–934.

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

KURTA, R. P., DONATELLI, J. J., YOON, C. H. et al. (2017). Correlations in scattered X-Ray laser pulses reveal nanoscale structural features of viruses. *Phys. Rev. Lett*. **119** 158102.

LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal*. **88** 365–411. MR2026339

LEE, S., ZOU, F. and WRIGHT, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Ann. Statist*. **38** 3605–3629. MR2766862

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

LI, J. and TAO, D. (2010). Simple exponential family PCA. In *AISTATS* 453–460.

LIU, L. T, DOBRIBAN, E. and SINGER, A. (2018). Supplement to "*e*PCA: High dimensional exponential family PCA." DOI:10.1214/18-AOAS1146SUPP.

MAIA, F. R. N. C. and HAJDU, J. (2016). The trickle before the torrent-diffraction data from X-ray lasers. *Sci. Data* **3** 160059.

MÄKITALO, M. and FOI, A. (2011). Optimal inversion of the Anscombe transformation in low-count Poisson image denoising. *IEEE Trans. Image Process*. **20** 99–109. MR2767607

MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb.* **72** 507–536. MR0208649

MARTIN, A. V., WANG, F., LOH, N. D., EKEBERG, T. et al. (2012). Noise-robust coherent diffractive imaging with a single diffraction pattern. *Opt. Express* **20** 16650–16661.

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

NOWAK, R. D. and BARANIUK, R. G. (1999). Wavelet-domain filtering for photon imaging systems. *IEEE Trans. Image Process.* **8** 666–678.

PANDE, K., SCHWANDER, P., SCHMIDT, M. and SALDIN, D. (2014). Deducing fast electron density changes in randomly orientated uncrystallized biomolecules in a pump–probe experiment. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **369** 20130332.

PANDE, K., SCHMIDT, M., SCHWANDER, P. and SALDIN, D. K. (2015). Simulations on time-resolved structure determination of uncrystallized biomolecules in the presence of shot noise. *Struct. Dyn.* **2** 024103.

PATTERSON, N., PRICE, A. L. and REICH, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* **2** e190.

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

SALDIN, D. K., SHNEERSON, V. L., FUNG, R. and OURMAZD, A. (2009). Structure of isolated biomolecules obtained from ultrashort x-ray pulses: Exploiting the symmetry of random orientations. *J. Phys., Condens. Matter* **21** 134014.

SCHWANDER, P., GIANNAKIS, D., YOON, C. H. and OURMAZD, A. (2012). The symmetries of image formation by scattering. II. Applications. *Opt. Express* **20** 12827–12849.

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. Wiley, New York. MR1190470

SHABALIN, A. A. and NOBEL, A. B. (2013). Reconstruction of a low-rank matrix in the presence of Gaussian noise. *J. Multivariate Anal.* **118** 67–76. MR3054091

STARCK, J.-L., MURTAGH, F. and FADILI, J. M. (2010). *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*. Cambridge Univ. Press, Cambridge. MR2643260

STARODUB, D., AQUILA, A., BAJT, S. et al. (2012). Single-particle structure determination by correlations of snapshot X-ray diffraction patterns. *Nat. Commun.* **3**.

STEGLE, O., TEICHMANN, S. A. and MARIONI, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16** 133–145.

STEIN, C. (1956). Some problems in multivariate analysis. Technical report, Dept. Statistics, Stanford Univ., Stanford, CA.

TROPP, J. A. (2016). The expected norm of a sum of independent random matrices: An elementary approach. In *High Dimensional Probability VII. Progress in Probability* **71** 173–202. Springer, Cham. MR3565264

UDELL, M., HORN, C., ZADEH, R. and BOYD, S. (2014). Generalized low rank models. In *NIPS Workshop on Distributed Machine Learning and Matrix Computations*.

UDELL, M., HORN, C., ZADEH, R. and BOYD, S. (2016). Generalized low rank models. *Found. Trends Mach. Learn.* **9** 1–118.

VISSCHER, P. M., BROWN, M. A., MCCARTHY, M. I. and YANG, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.* **90** 7–24.

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

ZHAO, Z., SHKOLNISKY, Y. and SINGER, A. (2016). Fast steerable principal component analysis. *IEEE Trans. Comput. Imaging* **2** 1–12. MR3472531

L. T. LIU
DEPARTMENT OF ELECTRICAL ENGINEERING
AND COMPUTER SCIENCES
UNIVERSITY OF CALIFORNIA AT BERKELEY
396 SODA HALL #1776
BERKELEY, CALIFORNIA 94720-1776
USA
E-MAIL: lydiatliu@berkeley.edu

E. DOBRIBAN
WHARTON STATISTICS DEPARTMENT
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: dobriban@wharton.upenn.edu

A. SINGER
DEPARTMENT OF MATHEMATICS,
AND PROGRAM IN APPLIED
AND COMPUTATIONAL MATHEMATICS
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: amits@math.princeton.edu