

ePCA: High Dimensional Exponential Family PCA

Lydia T. Liu^{*†} Edgar Dobriban[‡] Amit Singer[§]

March 8, 2017

Abstract

Many applications, such as photon-limited imaging and genomics, involve large datasets with noisy entries from exponential family distributions. It is of interest to estimate the covariance structure and principal components of the noiseless distribution. Principal Component Analysis (PCA), the standard method for this setting, can be inefficient when the noise is non-Gaussian.

We develop *ePCA* (exponential family PCA), a new methodology for PCA on exponential family distributions. *ePCA* can be used for dimensionality reduction and denoising of large data matrices. *ePCA* involves the eigendecomposition of a new covariance matrix estimator, constructed in a simple and deterministic way using moment calculations, shrinkage, and random matrix theory.

We provide several theoretical justifications for our estimator, including the finite-sample convergence rate, and the Marchenko-Pastur law in high dimensions. *ePCA* compares favorably to PCA and various PCA alternatives for exponential families, in simulations as well as in XFEL and SNP data analysis. An open-source implementation is [available](#).

1 Introduction

In many applications we have large collections of data vectors with entries sampled from exponential families (such as Poisson or Binomial). This setting arises in image processing, computational biology, and natural language processing, among others. It is often of interest to reduce the dimensionality and understand the structure of the data.

The standard method for dimension reduction and denoising of large datasets is Principal Component Analysis (PCA) (e.g., [Jolliffe, 2002](#); [Anderson, 2003](#)). However, PCA is most naturally designed for Gaussian data, and there is no commonly agreed upon extension to non-Gaussian settings such as exponential families (see e.g., [Jolliffe, 2002](#), Sec. 14.4). While there are several proposals for extending PCA to non-Gaussian distributions, each of them has certain limitations, such as computational intractability for large datasets (see Sec. 2 for a detailed discussion).

We propose the new method *ePCA* for PCA of data from exponential families. *ePCA* 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, *ePCA* has several appealing properties. First, it is a computationally efficient deterministic algorithm that comprises a small number of basic linear algebraic operations, making it as fast as usual PCA and scalable to “big” datasets. This is in contrast to typical likelihood approaches involving iterative methods such as alternating least squares, the EM algorithm etc., without convergence guarantees. Second, it is a flexible method 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 *ePCA* is interpretable, which can be important to practitioners.

*The first two authors contributed equally to this work.

[†]e-mail: ltliu@princeton.edu. 1980 Frist Center, Princeton University, Princeton, NJ, 08544

[‡]e-mail: dobriban@stanford.edu. Department of Statistics, Stanford University, Stanford, CA, 94305

[§]e-mail: amits@math.princeton.edu. Department of Mathematics, and Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ, 08544

We perform extensive simulations with *ePCA* and show that in several metrics it outperforms usual PCA, PCA after standardization, and PCA alternatives for exponential families (see Sec. 6.1.2). We apply *ePCA* to simulated X-ray Free Electron Laser (XFEL) data, where it leads to better denoising—visually and in MSE—than PCA. We also apply *ePCA* to a dataset from the Human Genome Diversity Project (HGDP) measuring Single Nucleotide Polymorphisms, where it leads to a clearer structure than PCA.

ePCA 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 a few potential application areas.

1.1 Denoising XFEL diffraction patterns

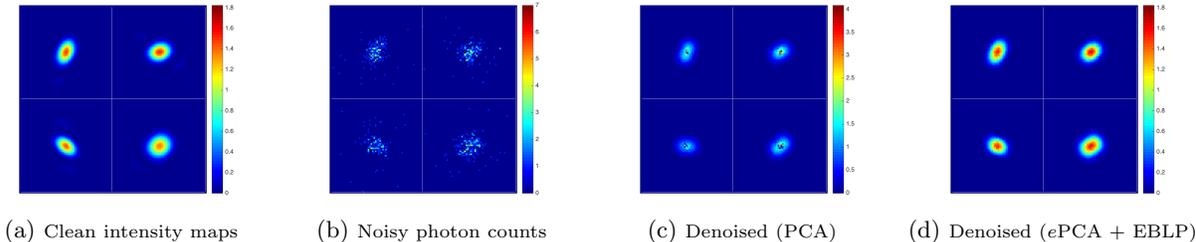


Figure 1.1: XFEL diffraction pattern formation model and denoising. See section 6.1 for details.

X-ray free electron lasers (XFEL) are an increasingly popular experimental technique to understand the three-dimensional structure of molecules (e.g., Favre-Nicolin et al., 2015; Maia and Hajdu, 2016). XFEL imaging leads to two-dimensional diffraction patterns of single particles. A key advantage of XFEL is that it uses extremely short femtosecond X-ray pulses, during which the molecule does not change its structure. As illustrated in Figure 1.1, these images are very noisy due to the low number of photons, and the count-noise at each detector follows an approximately Poisson distribution. Further, we only capture one diffraction pattern per particle, and the particle orientations are unknown.

In order to reconstruct the 3-D structure of the particle, one approach is to use expectation-maximization (EM) (e.g., Scheres et al., 2007; Loh and Elser, 2009). Alternatively, assuming that the orientations are uniformly distributed over the special orthogonal group $SO(3)$, Kam’s method (Kam, 1977, 1980) could provide a way to estimate the 3-D structure without optimizing likelihood via EM (see e.g., Saldin et al., 2009). A requirement is to estimate the covariance matrix of the noiseless 2-D images. This motivates us to develop the *ePCA* method for covariance estimation and PCA of Poisson data, and more generally for exponential families. To illustrate the improvement in covariance estimation of *ePCA* over PCA, in Figure 1.1 we show the result of denoising simulated XFEL using different estimated covariance matrices, where EBLP refers to the denoiser we develop in section 5 for use in conjunction with *ePCA*.

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).

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 j -th SNP in the genome of the i -th individual. The number of individuals n can be more than 10,000, while the number of SNPs can be as large as 2.5 million. Binomial models are natural for such data. PCA is commonly used to infer population structure from SNP data, with a wide range of applications, including correcting for confounding in GWAS (see e.g., Patterson et al. (2006)). It is thus of interest to understand the proper way to estimate the covariance matrix and PCs.

Among other potential application areas, we point out RNA-sequencing, where negative binomial models are routinely in use (Anders and Huber, 2010).

1.3 Our contributions

We now briefly summarize our contributions:

1. We propose the new method *ePCA* for PCA of exponential family data. *ePCA* is based on a new covariance estimator that we develop in a sequence of steps (Sec. 3 and 4). We start with a *diagonal debiasing* of the sample covariance matrix (Sec 3.2), and characterize the finite-sample convergence rate to the population covariance matrix (Sec. 3.2.1).
2. To improve performance for high-dimensional data, we propose a method of *homogenization, shrinkage, and heterogenization* of the debiased covariance matrix (Sec. 4). Homogenization is a form of variable weighting, different from the usual method of standardizing the features to have unit variance. We justify it by proving the standard Marchenko-Pastur law (Marchenko and Pastur, 1967) for the homogenized sample covariance matrix (Sec. 4.1.1), and by showing that homogenization improves the signal strength (Sec. 4.2.2). An additional eigenvalue shrinkage step—that we call *scaling*—is needed beyond the well-understood shrinkage methods for homoskedastic Gaussian distributions (e.g., Donoho et al., 2013). This leads to our final covariance estimator, and *ePCA* consists of its the eigendecomposition. We evaluate our covariance estimators in a simulation study, and show that they reduce the MSE for covariance, eigenvalue, and eigenvector estimation (Sec. 4.2.3).
3. For biallelic genetic markers such as Single Nucleotide Polymorphisms (SNPs), homogenization agrees with the widely used normalization assuming Hardy-Weinberg equilibrium (HWE) (Sec. 4.3). This provides perhaps the first theoretical justification for HWE normalization.
4. We apply *ePCA* to develop a new denoising method (Sec. 5), a form of empirical Best Linear Predictor (EBLP) from random effects models (Searle et al., 2009, Sec. 7.4), where we use our covariance estimator to estimate parameters in the BLP denoiser. In areas such as electrical engineering and signal processing, the BLP is known as the “Wiener filter” or the “Linear Minimum Mean Squared Estimator (LMMSE)” (e.g., Kay, 1993, Ch. 12).
5. We apply *ePCA* denoising to simulated XFEL data where it leads to better denoising than PCA (Sec. 6.1) We also apply *ePCA* to a SNP dataset from the Human Genome Diversity Project (HGDP) (Li et al., 2008), where it leads to a clearer structure in the PC scores than PCA (Sec. 6.4b).

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 including PCA, and to Yao et al. (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, Sec. 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 (Sec. 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, Sec. 2.3., Sec. 14.2.) In general, choosing the weights can be nontrivial. Our homogenization step of *ePCA* (Sec. 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 Sec. 4.1).

2.2 PCA in non-Gaussian distributions, GLLVMs

There have been several approaches suggested for extending PCA to non-Gaussian distributions, see. e.g., Jolliffe (2002), Sec. 14.4. One possibility is to use robust estimates of the covariance matrix (see Jolliffe, 2002, Sec. 14.4, for references). Another approach assumes that the natural parameter lies in a low dimensional space (Collins et al., 2001), and then attempts to maximize the log-likelihood. This leads to a non-convex 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 (Knott and Bartholomew, 1999; Huber et al., 2004). These flexible likelihood-based methods enable careful modelling 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 et al., 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.

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., Shabalin and Nobel, 2013; Donoho et al., 2013; Nadakuditi, 2014). While the qualitative insights they identify—e.g., 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 et al. (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 towards 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 et al., 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—e.g., XFEL images.

Starck et al. (2010) Sec. 6.5. provides an overview of the classical methods for Poisson noise. 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 et al., 2010, Sec. 6.6).

Other methods are based on singular value thresholding (SVT), with various approaches to handling non-Gaussian noise. For example, Furnival et al. (2016) 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.

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).

Table 1: Covariance estimators

Notation	Name	Formula	Defined in	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	$S_d = S - \text{diag}[V(\bar{Y})]$	(5)	Hierarchy
S_h	Homogenization	$S_h = D_n^{-1/2} S_d D_n^{-1/2}$	(6)	Heteroskedasticity
$S_{h,\eta}$	Shrinkage	$S_{h,\eta} = \eta(S_h)$	(7)	High dimensionality
S_{he}	Heterogenization	$S_{he} = D_n^{1/2} S_{h,\eta} D_n^{1/2}$	(8)	Heteroskedasticity
S_s	Scaling	$S_s = \sum \hat{\alpha}_i \hat{v}_i \hat{v}_i^\top$, where $S_{he} = \sum \hat{v}_i \hat{v}_i^\top$	(13)	Heteroskedasticity

Algorithm 1: Covariance matrix estimation and ePCA

Input: Data $Y = [Y_1, \dots, Y_n]^\top \in \mathbb{R}^{n \times p}$; Desired rank $r \leq p$;

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; ePCA: 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 $D_n = \text{diag}[V(\bar{Y})]$
 - 4 Homogenize and diagonally debias the covariance matrix $S_h = D_n^{-1/2} S D_n^{-1/2} - I_p$
 - 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 $\Lambda_r = \text{diag}(\lambda_1, \dots, \lambda_r)$.
 - 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 $S_{he} = D_n^{1/2} S_{h,\eta} D_n^{1/2}$
 - 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 $S_{he} = \sum \hat{v}_i \hat{v}_i^\top$
-

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

$$p_\theta(y) = \exp[\theta y - A(\theta)] \quad (1)$$

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 $\text{Var}[Y] = A''(\theta)$, where we denote $g'(\theta) = dg(\theta)/d\theta$.

Our running example will be the Poisson distribution $y \sim \text{Poisson}(x)$. Here the carrier measure is the discrete measure with density $\nu(dy) = 1/y!$ with respect to the counting measure on the non-negative 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 $Y = (Y(1), \dots, Y(p))^\top$ are drawn independently from an exponential family $Y(j) \sim p_{\theta(j)}(y)$ defined in (1). Formally, denoting by $\tilde{\sim}$ the mean and the covariance of a random vector:

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

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

Therefore, the mean of Y conditional on θ is

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

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.

The assumption of conditional independence given θ may sound restrictive. It means that all latent effects that induce correlations do so through θ and not through some other mechanism. However, we can always capture some 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 . One reason is that this enables a simple and deterministic algorithm, in contrast to the typical likelihood methods. Another reason is that in many applications, it is reasonable to assume that the means of noisy signals have a “low-complexity” structure, such as lying on a 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 9-dimensional linear space. As mentioned in [Sec. 2](#), this is a key modelling assumption distinguishing our approach from prior work like [Collins et al. \(2001\)](#).

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

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

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

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

For Poisson observations $Y \sim \text{Poisson}_p(X)$, where $X \in \mathbb{R}^p$ is random, we can write $Y = X + \text{diag}(X)^{1/2}\varepsilon$. 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 $\text{Cov}[Y] = \text{Cov}[X] + \mathbb{E} \text{diag}[X]$.

3.2 Diagonal debiasing

We will propose several estimators of increasing sophistication to estimate the covariance matrix $\Sigma_x = \text{Cov}[A'(\theta)]$ 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,

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

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

We define the sample covariance estimator

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

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

$$S_d = S - \text{diag}[V(\bar{Y})]. \quad (5)$$

Continuing with our Poisson example, $A'(\theta) = A''(\theta) = \exp(\theta)$, so $V(m) = m$, and $S_d = S - \text{diag}[\bar{Y}]$. In this example the estimator is unbiased, because V is linear. When V is non-linear, 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 $n \rightarrow \infty$ 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 and Poisson distributions. We also assume that the coordinates of the random vector θ are almost surely bounded, $\|\theta\|_\infty \leq B$. 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 $m_4 = \max_i \mathbb{E}[Y(i)^4] \geq C$ for some universal constant $C > 0$. This is reasonable, as it states that at least some entries of the random vector have non-vanishing magnitude.

Let \lesssim denote inequality up to constants not depending on n and p . Let $\|\cdot\|_{\text{Fr}}$ be the Frobenius norm and $\|\cdot\|$ be the operator norm. Our result, proved in Sec. 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 Frobenius norm, with $\mu := \mathbb{E}Y = \mathbb{E}X = \mathbb{E}A'(\theta)$:

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

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

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

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 $(p/n)^{1/2}$ while the Frobenius norm rate is $p/n^{1/2}$.

Our proof of Thm. 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. \(2005\)](#).

4 Homogenization and shrinkage

4.1 Homogenization

In the previous sections, we showed that the diagonally debiased sample covariance matrix converges at a rate $O(pn^{-1/2})$. Next we propose a shrinkage method to improve this estimator in the high dimensional regime where $n, p \rightarrow \infty$ and $p/n \rightarrow \gamma > 0$. 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 et al., 2013](#)).

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

$$S_h = D_n^{-1/2}S_dD_n^{-1/2} = D_n^{-1/2}SD_n^{-1/2} - I_p. \quad (6)$$

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

Homogenization is different from *standardization*, the classical method for removing heteroskedasticity. To standardize, each feature—e.g., pixel—is divided by its empirical standard deviation (e.g., Jolliffe, 2002, Sec. 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 $\text{Var}[Y(i)]$ of each feature is the sum of the signal variance $\text{Var}[A'(\theta(i))]$ 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 (Thm. 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 (Sec. 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 simple 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 (Marchenko 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 $Y_i = A'(\theta) + \text{diag}[A''(\theta)]^{1/2} \varepsilon_i$, 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 $\mathcal{Y} = \bar{1} A'(\theta)^\top + \mathcal{E} \text{diag}[A''(\theta)]^{1/2}$, where $\bar{1} = (1, 1, \dots, 1)^\top$ 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 $A''(\theta(i))$, $i = 1, \dots, p$. We assume that $A''(\theta(i)) > c$ for some universal constant $c > 0$. In 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 $n, p \rightarrow \infty$ so that $p/n \rightarrow \gamma > 0$. Suppose moreover that H_p converges weakly to some limit distribution, i.e., $H_p \Rightarrow H$. Since $\text{diag}[A''(\theta)]$ 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, 2009, Thm. 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 simple closed-form density, and there are many useful tools already available for low-rank covariance estimation (e.g., Shabalin and Nobel, 2013; Donoho et al., 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 .

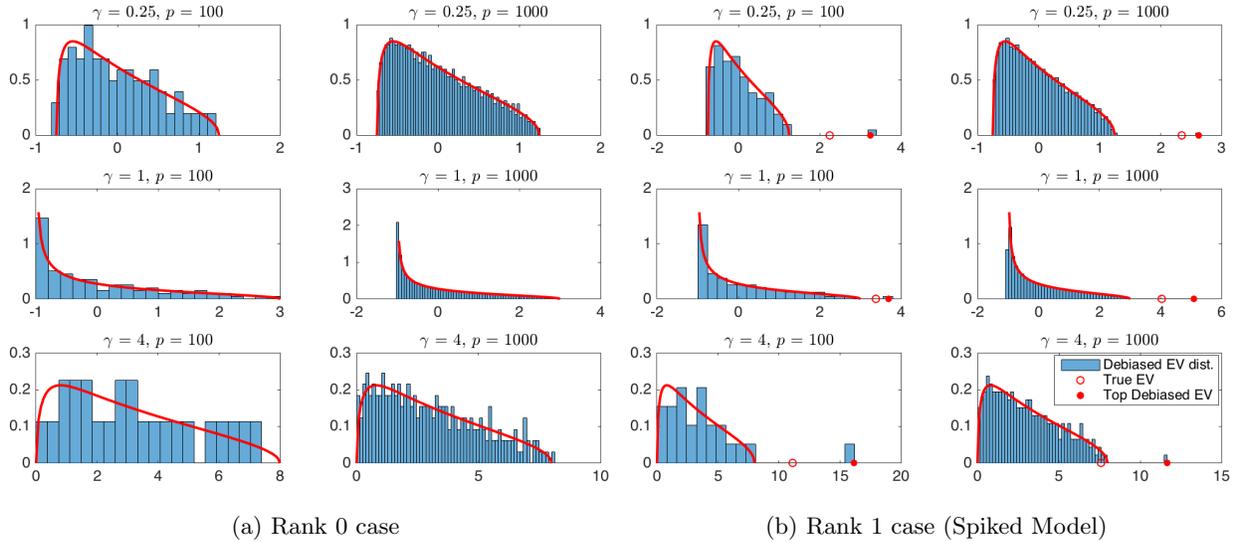


Figure 4.1: Empirical distribution of eigenvalues of homogenized sample covariance S_h for different values of $\gamma = p/n$, with the corresponding shifted Marchenko-Pastur density overlaid as a red curve. Data simulated according to 4.2. In the legend for (b), ‘Top Debiased EV’ refers top eigenvalue of S_h , while ‘True EV’ refers to the top eigenvalue of $D_n^{-1/2} \Sigma_x D_n^{-1/2}$, which we want to estimate.

In the proof presented in Appendix A.1.3, we deduce this from the Marchenko-Pastur law for the error matrix $n^{-1/2} \mathcal{E}$, for which standard results from Bai and Silverstein (2009) 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 shrunk* estimators will have the form

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

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 Thm. 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 et al., 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:

$$S_{he} = D_n^{1/2} \cdot S_{h,\eta} \cdot D_n^{1/2}. \quad (8)$$

Heterogenization is a non-linear operation that changes both the eigenvectors and eigenvalues. While it

Table 2: Spiked models: Summary of the original and homogenized spiked model.

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

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 (Sec. 4.2.3).

To understand homogenization empirically, we perform two simulations. First, we generate non-negative i.i.d $\{X_i\}_{1 \leq i \leq n}$ lying in a low-dimensional space of dimension r : we pick r vectors $v_1, \dots, v_r \in \mathbb{R}^p$ 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}, \dots, a_{ir} independently from the uniform distribution on $[0, 1]$. Define $X_i = a_{i1} v_1 + \dots + a_{ir} v_r$. Note that X_i are non-negative, reside in a hyperplane spanned by v_1, \dots, v_r , and the mean and covariance of X_i can be found easily in terms of v_1, \dots, v_r . The coefficients a_{i1}, \dots, a_{ir} are also normalized so that $a_{i1} + \dots + a_{ir} = A$, 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 4.1. When $r = 0$, the standard MP distribution—shifted by -1 —is a good match (Fig. 4.1a). This is in accordance with Thm. 4.2. When $r = 1$, the standard MP distribution still matches the bulk of the noise eigenvalues (Fig. 4.1b). Moreover, we observe the same qualitative behaviour 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, Eq. (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, but they generalize directly to higher rank.

Suppose that the i -th 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 $Y_i \sim \text{Poisson}_p(X_i)$, 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, $\text{Cov}[Y] = \text{Cov}[X] + \mathbb{E} \text{diag}[V(X)]$, and defining $D = \mathbb{E} \text{diag}[V(X)]$, we obtain

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

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

We homogenize the observations dividing by the elements of $D^{1/2}$. The elements of D are expected values of variances. They are thus positive, except for coordinates that that can be discarded because they have

no variability. The homogenized observations are $Y_h = D^{-1/2}Y$, and their population covariance matrix is

$$\text{Cov}[Y_h] = D^{-1/2}vv^\top D^{-1/2} + I_p. \quad (10)$$

We now compare this with the usual *standard spiked model* (Johnstone, 2001) where the observations Y_h are Gaussian and have covariance matrix $\text{Cov}[Y_h] = \ell ww^\top + I_p$, 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 et al., 2005) shows that when $n, p \rightarrow \infty$ such that $p/n \rightarrow \gamma > 0$, 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 et al., 2015).

Heuristically, comparing with (10), we surmise that a spiked model with $\ell = v^\top D^{-1}v$ and $w = D^{-1/2}v/\|D^{-1/2}v\|$ 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 $v^\top \text{diag}[u]^{-1}v = \sqrt{\gamma}$. 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 et al., 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 Eq. (9)

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

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 Eq. (10), the SNR equals $v^\top D^{-1}v/p$.

Suppose now that v is approximately *delocalized* in the sense that $p \cdot v^\top D^{-1}v \approx \text{tr } D^{-1} \cdot v^\top v$. 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, $\mathbb{E}v^\top D^{-1}v = \sigma^2 \text{tr } D^{-1}$ 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

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

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

$$\mathcal{I} = \frac{\text{tr } D}{p} \cdot \frac{v^\top D^{-1}v}{v^\top v}. \quad (11)$$

The above heuristic can be formalized as follows:

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

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

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 $\text{Cov}[X] = vv^\top$ of X . This has one nonzero eigenvalue $t = \|v\|^2$ and corresponding eigenvector $v_{norm} = v/\|v\|$. 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.

For data with independent coordinates and equal variances, the cumulative work of many authors (e.g., Baik et al., 2005; Baik and Silverstein, 2006; Paul, 2007; Benaych-Georges and Nadakuditi, 2011, etc) shows that if the population spike ℓ is above the BBP phase transition—i.e., $\ell > \sqrt{\gamma}$ —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*:

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

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 et al., 2010; Donoho et al., 2013), i.e., defining $\hat{\ell} = \lambda^{-1}(\lambda_{\max}(S_h))$. Donoho et al. (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 $w = D^{-1/2}v/\|D^{-1/2}v\|$. 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: $(w^\top \hat{w})^2 \rightarrow c^2(\ell; \gamma)$ almost surely, where $c(\ell; \gamma)$ is the *cosine forward map* given by (e.g., Paul, 2007; Benaych-Georges and Nadakuditi, 2011, etc):

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

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 $w^\top Dw = v^\top v/v^\top D^{-1}v = t/\ell$, and $\varepsilon^\top D\varepsilon \approx \text{tr}(D)/d$, we have

$$\|\hat{v}\|^2 \approx \ell \cdot \hat{w}^\top D\hat{w} \approx \ell \cdot (cw + s\varepsilon)^\top D(cw + s\varepsilon) \approx \ell \cdot (c^2 w^\top Dw + s^2 \varepsilon^\top D\varepsilon) \approx tc^2 + \ell s^2 \text{tr}(D)/p.$$

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

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

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.

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):

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

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 $\hat{c}^2 > 0$,

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

When $\hat{c}^2 = 0$, i.e., when $\hat{\ell} \leq \sqrt{\gamma}$, 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 multispike case, suppose $X_j = u + \sum_{i=1}^r z_{ij} v_i$. Then the marginal covariance of Y is $\text{Cov}[Y] = \sum_{i=1}^r v_i v_i^\top + D$. 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 $S_{he} = \sum_{i=1}^r \hat{v}_i \hat{v}_i^\top = \sum_{i=1}^r \hat{\lambda}_i \hat{u}_i \hat{u}_i^\top$, where \hat{u}_i are orthonormal, and the $\hat{\lambda}_i \geq 0$ are sorted in decreasing order. Based on our above discussion, we define the *scaled* covariance matrix as

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

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 *ePCA* 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.

ePCA is summarized in Alg. 1. Clearly, *ePCA* is applicable when the variables $x(i)$ have known non-identical 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 Sec. 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.

4.2.4 Simulations with *ePCA*

We report the results of a simulation study with *ePCA*. We simulate data Y_i from the Poisson model $Y_i \sim \text{Poisson}_p(X_i)$, where the mean parameters are $X_i = u + z_i \ell^{1/2} v$, 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 $\gamma = 1/2$, so $n = 1000$. The phase transition occurs when the spike is $\ell = \sqrt{\gamma}/v^\top \text{diag}[u]^{-1} v \approx 1.2$. 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 *ePCA*—or scaled—estimator $\hat{t}(\hat{\alpha}) = \hat{\alpha} \|\hat{v}\|^2$.

The results displayed in Fig. 4.2 (left) show that the *ePCA*/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 Fig. 4.2 (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 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, 2016). 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 4.2 (right). The predicted PT matches the empirical PT. The *ePCA* eigenvector—top eigenvector of S_s —in this case agrees with the eigenvector of the heterogenized covariance

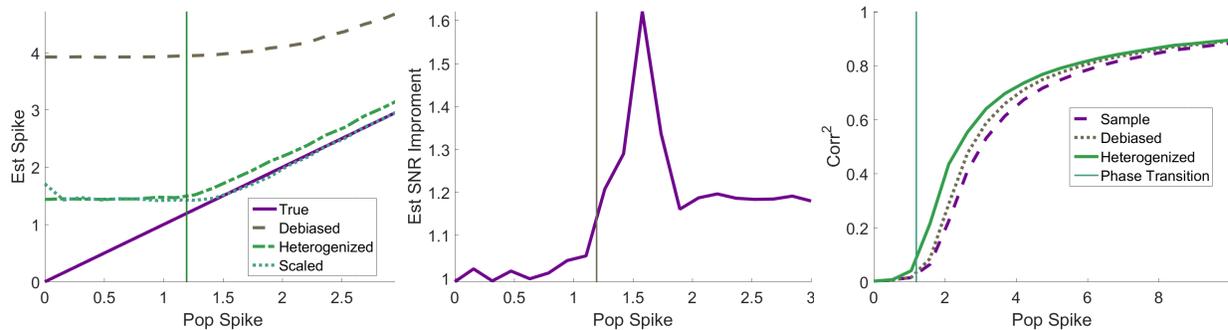


Figure 4.2: Simulation with $ePCA$. Left: Spike estimation; true, debiased, heterogenized, and scaled ($ePCA$). Middle: Estimated improvement in SNR due to homogenization. Right: Squared correlation between v and leading eigenvector of various covariance estimates; sample, debiased, heterogenized ($ePCA$). Plotted against the spike.

matrix S_{he} , because both are of rank one. $ePCA$ 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 et al., 2006, p. 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 et al., 2006, p. 2075). Our theoretical results justify HWE normalization. In particular, our Thm. 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 (Prop. 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—i.e., 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 $ePCA$, 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 et al., 2009), hence we predict X_i using the Best Linear Predictor—or BLP (Searle et al., 2009, Sec. 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 Sec. 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 et al., 2009, Sec. 7.4) that

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

The BLP depends on the unknown parameters Σ_x , $\text{diag}[\mathbb{E}A''(\theta)]$, and $\mathbb{E}[A'(\theta)]$. The standard strategy, known as *Empirical BLP* or EBLP (e.g., Searle et al., 2009) is to estimate these unknown parameters using the entire dataset, and denoise the vectors Y_i by plug-in:

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

We will use *ePCA*, i.e., the scaled covariance matrix S_s proposed in (13) to estimate Σ_x . As before in Sec. 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 Poisson distribution, we have

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

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 $\hat{\Sigma} = \text{diag}[\bar{Y}] + S_s$ may no longer be invertible: S_s is of low rank, while $\text{diag}[\bar{Y}]$ is also not of full rank. To avoid this problem, we implement a ridge-regularized covariance estimator $\hat{\Sigma}_\varepsilon = (1 - \varepsilon)\hat{\Sigma} + \varepsilon \cdot \tilde{m}I_p$ as in Ledoit and Wolf (2004), where $\tilde{m} = \text{tr} \hat{\Sigma} / p$ and $\varepsilon > 0$ is a small constant. Note that $\text{tr} \hat{\Sigma}_\varepsilon = \text{tr} \hat{\Sigma}$. The ridge-regularized estimator $\hat{\Sigma}_\varepsilon$ 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 apply *ePCA* to a simulated XFEL dataset, and an empirical genetics dataset, comparing with PCA.

6.1 XFEL images

We simulate $n_0 = 70,000$ noiseless XFEL diffraction intensity maps of a lysozyme (Protein Data Bank 1AKI) with Condor (Hantke et al., 2016). We rescale the average pixel intensity to 0.04 such that shot noise dominates, following previous work (e.g., Schwander et al., 2012). To sample an arbitrary number n of noisy diffraction patterns, we sample an 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 images are 64 pixels by 64 pixels, so $p = 4096$. Figure 1.1 illustrates the intensity maps and the resulting noisy diffraction patterns.

6.1.1 Covariance estimation

For covariance estimation, we vary the sample size n in the range $3 \leq \log_{10}(n) \leq 5$. 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 (Fig. 6.1) in MSE. The largest improvement is due to diagonal debiasing, but scaling leads to the smallest MSE.

Figure 6.2 summarizes the error of eigenvalue estimation. The *ePCA* 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 *ePCA* eigenvalues is small regardless of sample size.

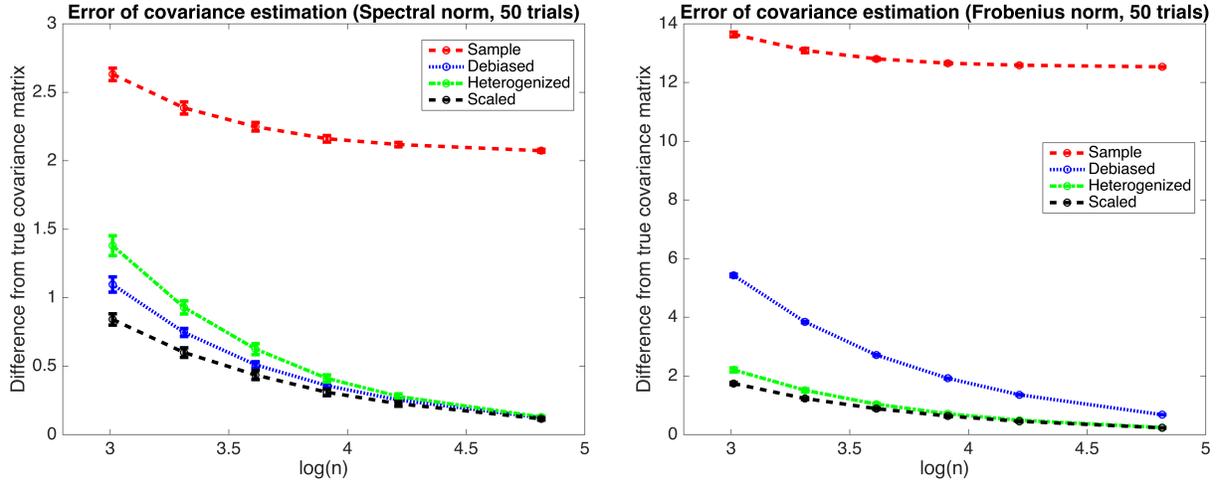


Figure 6.1: Error of covariance matrix estimation, measured as the spectral norm (left) and Frobenius norm (right) of the difference between each covariance estimate (Sample, Debiased, Heterogenized, Scaled) and the true covariance matrix.

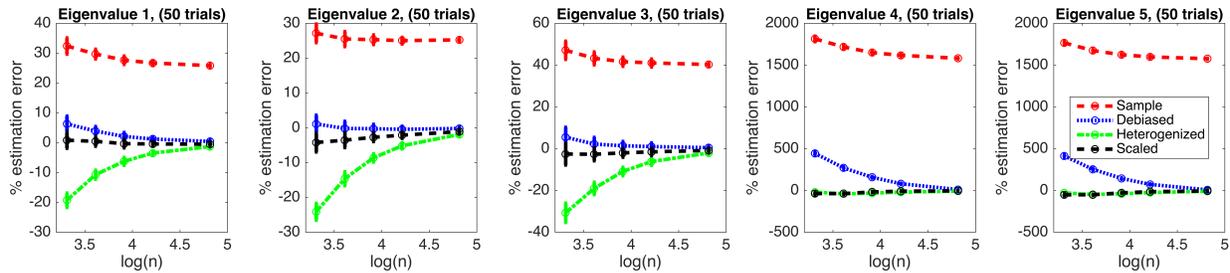


Figure 6.2: Error of eigenvalue estimation for the top 5 eigenvalues, measured as percentage error relative to the true eigenvalue, for XFEL data. We plot the mean and standard deviation (as error bars) over 50 Monte Carlo trials.

We visualize the eigenvectors (or eigenimages) for XFEL diffraction patterns in Figure 6.3. 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.

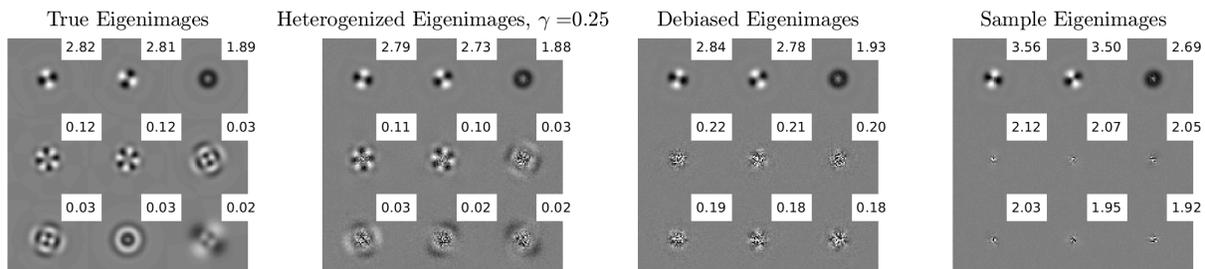


Figure 6.3: XFEL Eigenimages for $\gamma = 1/4$, ordered by eigenvalue

The e PCA/heterogenized eigenvectors 1 to 2 in Figure 6.3 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 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 6.4a clearly shows that the e PCA/heterogenized covariance matrix best estimates the low-rank subspace inhabited by the clean data.

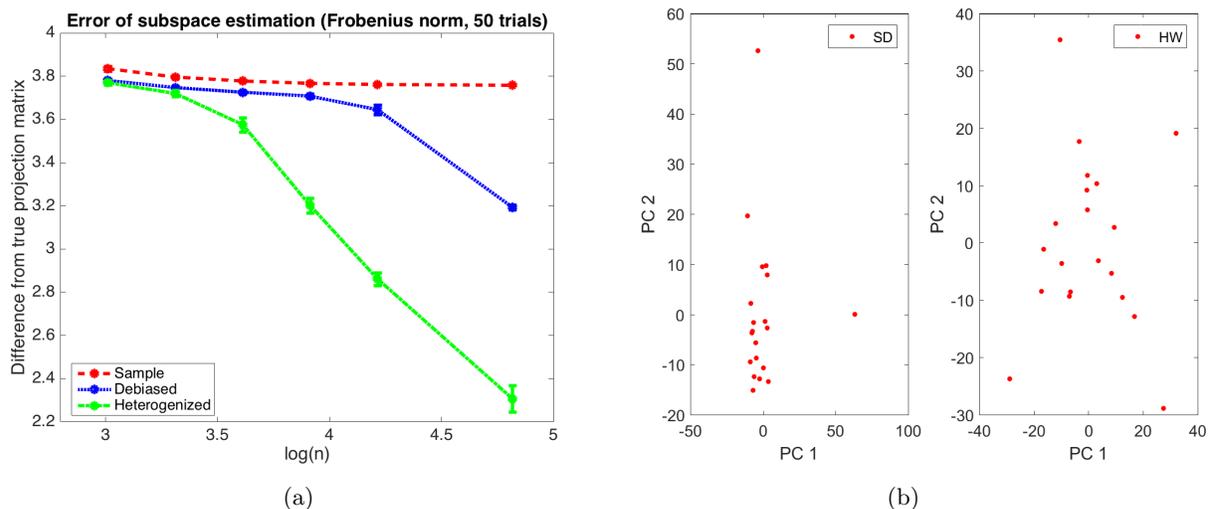


Figure 6.4: a) Subspace estimation error for XFEL data. We plot the mean and standard deviation (as error bars) over 50 Monte Carlo trials. b) HGDP dataset: PC scores of 20 CEU samples after standardization (SD, left) and homogenization/HWE normalization (HW, right).

6.1.2 Denoising

Finally, we report the results of denoising the XFEL patterns. We compare “PCA denoising” or “vanilla projection”, i.e., 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 (Fig. 6.5). In EBLP denoising, our scaled covariance matrix leads to much better results than the sample covariance matrix. EBLP also does better when measured by reconstruction mean squared error, $MSE := (pn)^{-1} \sum_{i=1}^n \|\hat{X}_i - X_i\|^2$.

We also compare e PCA to the exponential family PCA method based on alternating minimization proposed by Collins et al. (2001) in Figure 6.6. e PCA is faster and recovers the images with higher accuracy, as measured by MSE (see the caption of Figure 6.6). 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 because the known inverse transforms (e.g., Makitalo and Foi, 2011) are ineffective in the photon-limited regime.

6.2 HGDP dataset

We also apply e PCA to a subset of the Human Genome Diversity Project (HGDP) dataset (Li et al., 2008), which contains Single Nucleotide Polymorphism (SNP) markers obtained from human samples. We obtained a homogeneous random set of $n = 20$ Caucasian samples from the CEU cohort, typed on $p = 120,631$ SNPs. We removed SNPs that showed no variability, with $p' = 107,026$ SNPs remaining. For each SNP we imputed

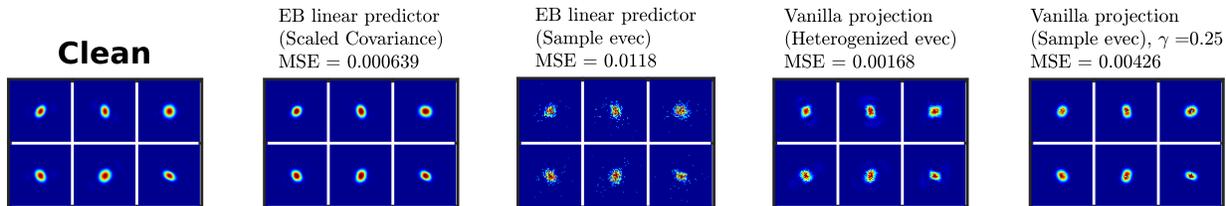


Figure 6.5: Sampled reconstructions using the XFEL dataset ($n = 16,384$; $p = 4096$), fixing the rank of covariance estimates at $r = 10$. Color scale of each reconstruction clipped to match that of clean images.

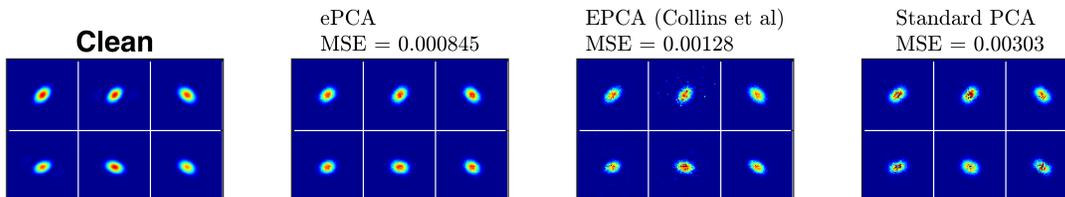


Figure 6.6: Comparing various methods' sampled reconstructions of the XFEL dataset ($n = 1000$; $p = 4096$), fixing the rank estimate for each method to $r = 8$. For reference, the MSE for noisy images is 0.0401. We also note that *ePCA* took 13.9 seconds, while [Collins et al. \(2001\)](#)'s exponential family PCA took 10900 seconds, or 3 hours, to finish running on a 2.7 GHz Intel Core i5 processor.

missing data as the mean of the available samples. We then computed the PC scores starting from two covariance matrices: (1) the one obtained after usual standardization of each feature to have unit norm, and (2) S_h obtained by using our homogenization method, which in this case agrees with HWE normalization as defined in e.g., [Patterson et al. \(2006\)](#) (see Sec. 4.3). In Fig. 6.4b we see that homogenization/HWE normalization apparently leads to a clearer structure in the PC scores than standardization. Two samples on the standardized PC scores appear to be extreme outliers, but our data is a homogeneous random sample and we do not expect outliers. This suggests that standardization is more sensitive to outliers or artifacts. These results are in line with the existing empirical observations about the superiority of HWE normalization ([Patterson et al., 2006](#)).

Acknowledgements

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, Kris Sankaran, Joel Tropp, Ramon van Handel, Jingshu Wang, Teng Zhang, and Jane Zhao for valuable discussions. They thank Filipe Maia, Max Hantke, and Benjamin Rose for help with software. They thank Patrick Perry for pointing out GLLVMs. A. S. was partially supported 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. E. D. was partially supported by NSF grant DMS-1407813, and by an HHMI International Student Research Fellowship.

References

- S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):1, 2010.
T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 2003.
F. J. Anscombe. The transformation of poisson, binomial and negative-binomial data. *Biometrika*, 35(3-4):246, 1948.

- Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, 2009.
- J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.
- J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697, 2005.
- R. Basri and D. W. Jacobs. Lambertian Reflectance and Linear Subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.
- F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.
- T. Bhamre, T. Zhang, and A. Singer. Denoising and covariance estimation of single particle cryo-EM images. *Journal of Structural Biology*, 195(1):72–81, 2016.
- J. Bigot, C. Deledalle, and D. Féral. Generalized SURE for optimal shrinkage of singular values in low-rank matrix denoising. *arXiv preprint arXiv:1605.07412*, 2016.
- S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Annals of Probability*, 33(2):514–560, 2005.
- Y. Cao and Y. Xie. Low-rank matrix recovery in Poisson noise. In *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, pages 384–388. IEEE, 2014.
- M. Collins, S. Dasgupta, and R. Schapire. A generalization of principal component analysis to the exponential family. *Advances in Neural Information Processing Systems (NIPS)*, 2001.
- E. Dobriban. Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 04(04):1550019, 2015.
- E. Dobriban. Sharp detection in PCA under correlations: all eigenvalues matter. *arXiv preprint arXiv:1602.06896*, to appear in *The Annals of Statistics*, 2016.
- D. Donoho, M. Gavish, and I. Johnstone. Optimal shrinkage of eigenvalues in the Spiked Covariance Model. *arXiv preprint arXiv:1311.0851*, 0906812:1–35, 2013.
- V. Favre-Nicolin, J. Baruchel, H. Renevier, J. Eymery, and A. Borbély. XTOP: high-resolution X-ray diffraction and imaging. *Journal of Applied Crystallography*, 48(3):620–620, 2015.
- M. F. Freeman and J. W. Tukey. Transformations Related to the Angular and the Square Root. *Ann. Math. Statist.*, 21(4):607–611, 12 1950.
- T. Furnival, R. K. Leary, and P. A. Midgley. Denoising time-resolved microscopy image sequences with singular value thresholding. *Ultramicroscopy*, 2016. ISSN 0304-3991.
- M. F. Hantke, T. Ekeberg, and F. R. N. C. Maia. Condor: A simulation tool for flash x-ray imaging. *Journal of Applied Crystallography*, 49(4):1356–1362, 2016.
- P. Huber, E. Ronchetti, and M.-P. Victoria-Feser. Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4):893–908, 2004.
- I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327, 2001.
- I. Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2002.
- J. Josse and S. Wager. Bootstrap-based regularization for low-rank matrix estimation. *Journal of Machine Learning Research*, 17(124):1–29, 2016.
- Z. Kam. The reconstruction of structure from electron micrographs of randomly oriented particles. *Journal of Theoretical Biology*, 82(1):15–39, 1980.
- Z. Kam. Determination of macromolecular structure in solution by spatial correlation of scattering fluctuations. *Macromolecules*, 10(5):927–934, 1977.
- S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*, volume 3. Prentice Hall, 1993.
- M. Knott and D. J. Bartholomew. *Latent variable models and factor analysis*. Edward Arnold, 1999.
- O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.
- S. Lee, F. Zou, and F. A. Wright. Convergence and prediction of principal component scores in high-dimensional settings. *Annals of Statistics*, 38(6):3605–3629, 2010.
- E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2005.
- J. Li and D. Tao. Simple exponential family PCA. In *AISTATS*, pages 453–460, 2010.

- J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–1104, 2008.
- N.-T. D. Loh and V. Elser. Reconstruction algorithm for single-particle diffraction imaging experiments. *Phys. Rev. E*, 80:026705, Aug 2009.
- F. R. Maia and J. Hajdu. The trickle before the torrent—diffraction data from X-ray lasers. *Scientific Data*, 3, 2016.
- M. Makitalo and A. Foi. Optimal inversion of the anscombe transformation in low-count Poisson image denoising. *IEEE Transactions on Image Processing*, 20(1):99–109, Jan 2011. ISSN 1057-7149.
- V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.*, 114(4):507–536, 1967.
- R. R. Nadakuditi. Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Transactions on Information Theory*, 60(5):3002–3018, 2014.
- R. D. Nowak and R. G. Baraniuk. Wavelet-domain filtering for photon imaging systems. *IEEE Transactions on Image Processing*, 8(5):666–678, 1999.
- N. Patterson, A. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190, 2006.
- D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.
- D. K. Saldin, V. L. Shneerson, R. Fung, and A. Ourmazd. Structure of isolated biomolecules obtained from ultrashort x-ray pulses: exploiting the symmetry of random orientations. *Journal of Physics: Condensed Matter*, 21(13), 2009.
- S. H. Scheres, H. Gao, M. Valle, G. T. Herman, P. P. Eggermont, J. Frank, and J.-M. Carazo. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature Methods*, 4(1):27–29, 2007.
- P. Schwander, D. Giannakis, C. H. Yoon, and A. Ourmazd. The symmetries of image formation by scattering. II. Applications. *Opt. Express*, 20(12):12827–12849, Jun 2012. doi: 10.1364/OE.20.012827.
- S. R. Searle, G. Casella, and C. E. McCulloch. *Variance components*. John Wiley & Sons, 2009.
- A. A. Shabalyn and A. B. Nobel. Reconstruction of a low-rank matrix in the presence of gaussian noise. *Journal of Multivariate Analysis*, 118:67–76, 2013.
- J.-L. Starck, F. Murtagh, and J. M. Fadili. *Sparse image and signal processing: wavelets, curvelets, morphological diversity*. Cambridge university press, 2010.
- C. Stein. Some problems in multivariate analysis. *Technical Report, Dept of Statistics, Stanford University*, 1956.
- J. A. Tropp. The Expected Norm of a Sum of Independent Random Matrices: An Elementary Approach. In *High-Dimensional Probability VII*, Progress in Probability 71. Birkhaeuser, 2016.
- M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized Low Rank Models. In *NIPS Workshop on Distributed Machine Learning and Matrix Computations*, 2014.
- M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized Low Rank Models. *Foundations and Trends in Machine Learning*, 9(1):1–118, 2016.
- R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.
- P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.
- J. Yao, Z. Bai, and S. Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, 2015.

A Appendix

A.1 Proof of Theorem 3.1

Let $\mu = \mathbb{E}Y = \mathbb{E}A'(\theta)$ and $B_0 = \mathbb{E}YY^\top = \text{Cov}[Y] + \mu\mu^\top = \Sigma_x + \text{diag}[\mathbb{E}A''(\theta)] + \mu\mu^\top$. Let $\|\cdot\|_a$ denote a generic matrix norm, such as the operator norm or the Frobenius norm. By the triangle inequality and the Cauchy-Schwarz inequality

$$\mathbb{E}[\|S_d - \Sigma_x\|_a] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Y_i Y_i^\top - \bar{Y}\bar{Y}^\top - \text{diag}[V(\bar{Y})] - \Sigma_x\right\|_a\right]$$

$$\begin{aligned}
&\leq \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n Y_i Y_i^\top - B_0 \right\|_a \right] + \mathbb{E} [\|\bar{Y} \bar{Y}^\top - \mu \mu^\top\|_a] + \mathbb{E} [\|\text{diag}[V(\bar{Y})] - \text{diag}[\mathbb{E}A''(\theta)]\|_a] \\
&\leq \left[\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n Y_i Y_i^\top - B_0 \right\|_a^2 \right]^{1/2} + \mathbb{E} [\|\bar{Y} \bar{Y}^\top - \mu \mu^\top\|_a] + \mathbb{E} [\|\text{diag}[V(\bar{Y})] - \text{diag}[\mathbb{E}A''(\theta)]\|_a]
\end{aligned}$$

We now consider the Frobenius and operator norms separately. For the Frobenius norm, using

$$\mathbb{E} [\|\text{diag}[V(\bar{Y})] - \text{diag}[\mathbb{E}A''(\theta)]\|_{\text{Fr}}] = \mathbb{E} [\|V(\bar{Y}) - \mathbb{E}A''(\theta)\|]$$

and Propositions A.3, A.4, and A.6, we find

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

Now, given that $m_4 = \max_i \mathbb{E}Y(i)^4$ is at least $O(1)$, the second and the last term is of smaller order than the first one. This leads to the bound $\mathbb{E}[\|S_d - \Sigma_x\|_{\text{Fr}}] \lesssim \sqrt{\frac{p}{n}} [\sqrt{p} \cdot m_4 + \|\mu\|]$.

For the operator norm, using $\mathbb{E} [\|\text{diag}[V(\bar{Y})] - \text{diag}[\mathbb{E}A''(\theta)]\|] \leq \mathbb{E} [\|V(\bar{Y}) - \mathbb{E}A''(\theta)\|]$ and Propositions A.3, A.4, and A.7, we find

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

This finishes the proof.

A.1.1 Sup-exponential properties

In this section we establish the sub-exponential property of our random variables. This is needed in the next sections in proving the rates of convergence.

Proposition A.1. A random variable $Y \sim p_\theta(y)$ from the exponential family is sub-exponential.

Proof. The moment generating function of Y is $\mathbb{E}[\exp(tY)] = \exp(A(\theta + t) - A(\theta))$. Since A is differentiable on an open neighborhood of θ , clearly $\mathbb{E}[\exp(tY)] \leq e$ for small t . Therefore, by the moment generating function characterization of sub-exponential random variables given in (5.16) of Vershynin (2012), Y is sub-exponential. \square

In the following proposition, we allow that the prior parameter θ is random, while requiring that it is bounded.

Proposition A.2. Let $Y \sim p_\theta(y)$. If θ is random and supported on a compact interval, then Y is sub-exponential.

Proof. By the characterization of sub-exponential random variables in (5.16) of Vershynin (2012), it is enough to show that $\mathbb{E}[\exp(A(\theta + t) - A(\theta))] \leq e$ for small t . Suppose θ is supported on $[a, b]$. Since $A(\theta + t)$ is continuously differentiable in a neighborhood of θ , we have $|A(\theta + t) - A(\theta)| \leq Ct|A'(\theta)| \leq Ct \sup_{\theta \in [a, b]} |A'(\theta)|$ for some $C > 0$, and for all t . Hence $\mathbb{E}[\exp(A(\theta + t) - A(\theta))] \leq \mathbb{E}[\exp(tC \sup_{\theta \in [a, b]} |A'(\theta)|)] \leq e$. The last inequality holds for sufficiently small t . \square

A.1.2 Auxiliary rates

Using the sub-exponential properties, we now prove the rates of convergence needed in the proof of Thm. 3.1 presented in Sec. A.1. Let $K(i) = \sup_{q \geq 1} q^{-1} (\mathbb{E}Y(i)^q)^{1/q}$ be the sub-exponential norm of the i -th coordinate of Y (see e.g., Vershynin, 2012, Sec 5.2.4). By assumption, these norms are uniformly bounded, so that $K(i) \leq K < \infty$ for some universal constant K .

Proposition A.3. $\mathbb{E}[\|V(\bar{Y}) - \mathbb{E}A''(\theta)\|] \lesssim \frac{\sqrt{p}}{\sqrt{n}}$ up to universal constant factors.

Proof. By the Cauchy-Schwarz inequality, $[\mathbb{E}\|V(\bar{Y}) - \mathbb{E}A''(\theta)\|]^2 \leq \mathbb{E}[\|V(\bar{Y}) - \mathbb{E}A''(\theta)\|^2]$. Since the latter quantity decomposes into d mean squared error terms, it is enough to show that each of them is bounded by C/n up to universal constant factors. Now,

$$\mathbb{E}[V(\bar{Y}(i)) - \mathbb{E}A''(\theta(i))]^2 \leq 2\mathbb{E}[V(\bar{Y}(i)) - V(\mathbb{E}\bar{Y}(i))]^2 + 2\mathbb{E}[V(\mathbb{E}\bar{Y}(i)) - \mathbb{E}A''(\theta(i))]^2.$$

For the first term, by the Lipschitz property of V , and by the definition of K , we have

$$\mathbb{E}[V(\bar{Y}(i)) - V(\mathbb{E}\bar{Y}(i))]^2 \leq L^2\mathbb{E}[\bar{Y}(i) - \mathbb{E}\bar{Y}(i)]^2 = n^{-1}L^2\text{Var}Y(i) \leq n^{-1}L^2\mathbb{E}Y(i)^2 \leq n^{-1}cL^2K^2.$$

For the second term, notice that $A''(\theta(i)) = V(\mathbb{E}[\bar{Y}(i)|\theta(i)])$. Denoting for convenience $Z = \bar{Y}(i)$, $\alpha = \theta(i)$, this reads $A''(\alpha) = V(\mathbb{E}[Z|\alpha])$, and thus $T := V(\mathbb{E}\alpha) - \mathbb{E}V(\mathbb{E}[Z|\alpha]) = \mathbb{E}\{V(\mathbb{E}\alpha) - V(\mathbb{E}[Z|\alpha])\}$. Hence, by the Cauchy-Schwarz inequality and by the Lipschitz property of V , $\mathbb{E}T^2 \leq \mathbb{E}\{V(\mathbb{E}\alpha) - V(\mathbb{E}[Z|\alpha])\}^2 \leq L^2\mathbb{E}(\mathbb{E}\alpha - \mathbb{E}[Z|\alpha])^2$. Finally, the term $\mathbb{E}(\mathbb{E}\alpha - \mathbb{E}[Z|\alpha])^2 = \text{Var}(\bar{Y}(i)|\theta(i)) = n^{-1}\text{Var}(Y(i)|\theta(i)) \leq n^{-1}cL^2K^2$ since $Y(i)$ is sub-exponential with norm at most K . Putting together all bounds, we obtain $\mathbb{E}[V(\bar{Y}(i)) - \mathbb{E}A''(\theta(i))]^2 \lesssim n^{-1}$ up to universal constant factors. By the remark in the beginning of the argument, this finishes the proof. \square

Proposition A.4. $\mathbb{E}[\|\mu\mu^\top - \bar{Y}\bar{Y}^\top\|_a] \lesssim \frac{p}{n} + \frac{\|\mu\|\sqrt{p}}{\sqrt{n}}$ up to universal constant factors, where $\|\cdot\|_a$ denotes the Frobenius norm or the operator norm.

Proof. Clearly $\|ab^\top\|_a = \|a\|\|b\|$. Then

$$\begin{aligned} \|aa^\top - bb^\top\|_a &= \|-a(b-a)^\top - (b-a)a^\top - (b-a)(b-a)^\top\|_a \\ &\leq \|(b-a)(b-a)^\top\|_a + \|a(b-a)^\top\|_a + \|(b-a)a^\top\|_a \text{ by the triangle inequality} \\ &= \|b-a\|^2 + 2\|a\|\|b-a\|. \end{aligned}$$

Using this, by Proposition A.5, $\mathbb{E}[\|\mu\mu^\top - \bar{Y}\bar{Y}^\top\|_a] \leq \mathbb{E}[\|\mu - \bar{Y}\|^2] + 2\mathbb{E}[\|\mu\|\|\mu - \bar{Y}\|] \lesssim \frac{p}{n} + \frac{\|\mu\|\sqrt{p}}{\sqrt{n}}$. \square

Proposition A.5. We have $\mathbb{E}[\|\bar{Y} - \mu\|^2] \lesssim \frac{p}{n}$ and $\mathbb{E}[\|\bar{Y} - \mu\|] \lesssim \frac{\sqrt{p}}{\sqrt{n}}$ up to universal constant factors.

Proof. By the Cauchy-Schwarz inequality, $\mathbb{E}[\|\bar{Y} - \mu\|^2] \leq \mathbb{E}[\|\bar{Y} - \mu\|]^2$. Then by the definition of the sub-exponential norm K , we have $\mathbb{E}[Y(i) - \mathbb{E}Y(i)]^2 \leq \mathbb{E}[Y(i)]^2 \leq cK^2$. Hence $\mathbb{E}[\|\bar{Y} - \mu\|^2] = n^{-1}\sum_{i=1}^p \mathbb{E}(Y(i) - \mathbb{E}Y(i))^2 \leq n^{-1}cpK^2$. This finishes the proof. \square

Proposition A.6 (Bounding the deviation of the second moment estimator for Y : Frobenius norm). Let $T_i = \frac{1}{n}(Y_i Y_i^\top - B)$ and $V_n = \sum_{i=1}^n T_i$. Then $\mathbb{E}[\|V_n\|_{\text{Fr}}^2] \lesssim \frac{p^2}{n}m_4$.

Proof. Since the Y_i are independent and identically distributed, and $\mathbb{E}T_i = 0$, we have

$$\mathbb{E}[\|V_n\|_{\text{Fr}}^2] = \mathbb{E}\left[\left\|\sum_{i=1}^n T_i\right\|_{\text{Fr}}^2\right] = n\mathbb{E}\|T_1\|_{\text{Fr}}^2 = \frac{1}{n}\mathbb{E}(\|Y_1 Y_1^\top\|_{\text{Fr}}^2 + \|B\|_{\text{Fr}}^2 - 2\text{Tr}(Y_1 Y_1^\top B)) = \frac{1}{n}(\mathbb{E}(\|Y_1\|^2)^2 - \text{Tr}(B^2)).$$

Now we can bound $\mathbb{E}(\|Y_1\|^2)^2 \leq p^2 \max_i \mathbb{E}Y_1(i)^4 \lesssim p^2 m_4$, proving the desired claim. \square

Proposition A.7 (Bounding the deviation of the second moment estimator for Y : Operator norm). Let $T_i = \frac{1}{n}(Y_i Y_i^\top - B)$ and $V_n = \sum_{i=1}^n T_i$. Then

$$\begin{aligned} \mathbb{E}[\|V_n\|^2]^{1/2} &\leq \sqrt{C(p)} \|\mathbb{E}[V_n^2]\|^{1/2} + \sqrt{C(p)} \cdot \left(\mathbb{E}\left[\max_i \|T_i\|^2\right]\right)^{1/2} \\ &\lesssim \sqrt{C(p)} \frac{(\mathbb{E}\|Y\|^4)^{1/2} + (\log n)^3 (\log p)^2}{\sqrt{n}}. \end{aligned}$$

Proof. The first inequality follows directly from Theorem 5.1 in [Tropp \(2016\)](#). Now we find an explicit expression for the right hand side. For the first term, since the Y_i are independent and identically distributed, and the T_i are centered, $\mathbb{E}V_n^2 = \mathbb{E}(\sum_{i=1}^n T_i)^2 = n\mathbb{E}T_1^2 = \frac{1}{n}(\mathbb{E}\|Y_1\|^2 Y_1 Y_1^\top - B^2)$. Since $\mathbb{E}V_n^2$ and B^2 are positive semi-definite, so $\|\mathbb{E}\|Y_1\|^2 Y_1 Y_1^\top - B^2\| \leq \|\mathbb{E}\|Y_1\|^2 Y_1 Y_1^\top\|$, we have $\|\mathbb{E}[V_n^2]\| \leq \frac{1}{n}\|\mathbb{E}(\|Y_1\|^2 Y_1 Y_1^\top)\|$.

Now for any fixed vector u with $\|u\| = 1$, $u^\top \mathbb{E}(\|Y_1\|^2 Y_1 Y_1^\top)u = \mathbb{E}\|Y_1\|^2 (u^\top Y_1)^2 \leq \mathbb{E}(\|Y_1\|^2)^2$. This gives the first term, $\mathbb{E}\|Y\|^4$.

For the second term, by the triangle inequality and $(a+b)^2 \leq 2(a^2 + b^2)$,

$$\mathbb{E} \left[\max_i \|T_i\|^2 \right] = \frac{1}{n} \mathbb{E} \left[\max_i \|Y_i Y_i^\top - A\|^2 \right] \leq \frac{2}{n} (\mathbb{E} \max_i \|Y_i Y_i^\top\|^2 + \|B\|^2).$$

When taking square roots as required by the theorem statement, the second term in this inequality can be bounded by $\|B\| \leq \text{Tr}(B) = \mathbb{E}\|Y\|^2 \leq (\mathbb{E}\|Y\|^4)^{1/2}$. For the first term in the bound, defining $Q_i = \sum_{j=1}^d Y_i(j)^2$, we have $\|Y_i Y_i^\top\|^2 = Q_i^2$, so for $m \geq 2$

$$\mathbb{E} \left[(\max_i Q_i)^2 \right] \leq \mathbb{E} \left[\left(\sum_{i=1}^n Q_i^m \right)^{2/m} \right] \leq \left(\sum_{i=1}^n \mathbb{E}[Q_i^m] \right)^{2/m} = (n\mathbb{E}[Q_1^m])^{2/m},$$

where the second inequality follows from Jensen's inequality. Choosing $m = \log n$, and then applying [Lemma A.8](#) the last term can be upper bounded by

$$e^2 \left((\mathbb{E}[Q_1^m])^{1/m} \right)^2 \lesssim [\mathbb{E}Q_1 + (\log n)^3 (\log p)^2]^2 \lesssim [\text{Tr}(B) + (\log n)^3 (\log p)^2]^2.$$

Finally, we use $\text{Tr}(B) = \mathbb{E}\|Y\|^2 \leq (\mathbb{E}\|Y\|^4)^{1/2}$ again. Putting these together leads to the result. \square

Lemma A.8. Let $Y(1), \dots, Y(p)$ be independent random variables distributed according to an exponential family $Y(j) \sim p_{\theta(j)}$ for deterministic $\theta(j) \in \mathbb{R}$. Let $Q = \sum_{j=1}^p Y(j)^2$. Define $\kappa := \frac{\sqrt{e}}{2(\sqrt{e}-1)} < 1.27$ and let η be any value in $(0, 1)$. Then for any $m \geq 1$

$$(\mathbb{E}[Q^m])^{1/m} \leq (1 + \eta)\mathbb{E}[Q] + C \frac{\kappa}{2} (1 + 1/\eta) K m^3 (\log p)^2$$

where C is a small constant.

Proof. By Theorem 8 in [Boucheron et al. \(2005\)](#), we get the following Rosenthal-type bound:

$$(\mathbb{E}[Q^m])^{1/m} \leq (1 + \eta)\mathbb{E}[Q] + \frac{\kappa}{2} m (1 + 1/\eta) \left(\mathbb{E} \left[\left(\max_{j \leq p} (Y(j)^2) \right)^m \right] \right)^{1/m}$$

We proceed to bound the second term on the right hand side:

$$\mathbb{E} \left[\left(\max_{j \leq p} (Y(j)^2) \right)^m \right]^{1/m} = \mathbb{E} \left[\max_{j \leq p} Y(j)^{2m} \right]^{1/m} \leq \mathbb{E} \left[\left(\sum_{j \leq p} Y(j)^{2m \log p} \right)^{1/\log p} \right]^{1/m} \leq \left(\sum_{j \leq p} \mathbb{E} [Y(j)^{2m \log p}] \right)^{1/m \log p}$$

where the last claim follows from Jensen's inequality. This can be further bounded as

$$p^{1/m \log p} \left(\max_{j \leq p} \mathbb{E} [Y(j)^{2m \log p}] \right)^{1/m \log p} \lesssim K (m \log p)^2.$$

On the last line, we have used the moments characterization of sub-exponentiality. \square

A.1.3 Proof of Theorem 4.2

It is enough to study the singular values of $\mathcal{Y}_w = n^{-1/2}\mathcal{Y}_c D_n^{-1/2}$, where $\mathcal{Y}_c = \mathcal{Y} - \bar{\mathbf{Y}}\bar{\mathbf{Y}}^\top$ is the centered data matrix, because $S_h + I_p = \mathcal{Y}_w^\top \mathcal{Y}_w$. Our strategy to show that \mathcal{Y}_w is well approximated by the noise matrix $n^{-1/2}\mathcal{E}$, which is more convenient to study directly.

Indeed, since $n^{-1/2}\mathcal{E}$ has independent entries of mean 0, variance $1/n$, and fourth moment of order $1/n^2$, the distribution of the squares of its singular values converges almost surely to the standard Marchenko-Pastur distribution with aspect ratio γ . Moreover, its operator norm converges to $1 + \gamma^{1/2}$ a.s. (Bai and Silverstein, 2009).

This implies that the same two properties hold for the auxiliary matrix $\mathcal{Y}_a = n^{-1/2}(\mathcal{Y} - \bar{\mathbf{1}}A'(\theta)^\top)D_n^{-1/2}$. Indeed, we can bound the operator norm of the difference $E = n^{-1/2}\mathcal{E} - \mathcal{Y}_a$ as

$$\|E\| = \|n^{-1/2}\mathcal{E}(I_p - \text{diag}[A''(\theta)]^{1/2}D_n^{-1/2})\| \leq \|n^{-1/2}\mathcal{E}\| \|I_p - \text{diag}[A''(\theta)]^{1/2}D_n^{-1/2}\|.$$

Now $\|n^{-1/2}\mathcal{E}\| \rightarrow 1 + \gamma^{1/2}$ a.s., and $\|I_p - \text{diag}[A''(\theta)]^{1/2}D_n^{-1/2}\| \rightarrow 0$ a.s. by Lemma A.9 presented below. This shows that the spectral distribution and operator norm of \mathcal{Y}_a converge as required.

Finally, the difference of the homogenized data matrix and the auxiliary matrix has rank one: $\mathcal{Y}_w - \mathcal{Y}_a = n^{-1/2}\bar{\mathbf{1}}(A'(\theta)^\top - \bar{\mathbf{Y}}^\top)D_n^{-1/2}$. Therefore \mathcal{Y}_w has the same Marchenko-Pastur limiting spectrum as \mathcal{Y}_a . This finishes the proof of Theorem 4.2.

Lemma A.9 (Convergence of empirical homogenization matrix). We have $\|I_p - \text{diag}[A''(\theta)]^{1/2}D_n^{-1/2}\| \rightarrow 0$ a.s.

Proof. Since $|1 - x^{1/2}| \leq |1 - x|$ for all $x \geq 0$, it is enough to show that

$$\max_i \left| 1 - \frac{A''(\theta(i))}{V(\bar{Y}(i))} \right| = \max_i \left| \frac{A''(\theta(i))}{V(\bar{Y}(i))} \right| \left| 1 - \frac{V(\bar{Y}(i))}{A''(\theta(i))} \right| \rightarrow 0.$$

From the expression on the right, we see that it is enough to show that $\max_i |1 - V(\bar{Y}(i))/A''(\theta(i))| \rightarrow 0$ almost surely. To use the Borel-Cantelli lemma, we show how to bound the probability of $V(\bar{Y}(i))/A''(\theta(i)) - 1 \geq \varepsilon$; the other direction is analogous. Since we assumed V is Lipschitz continuous with a uniform Lipschitz constant L , denoting $\delta = \varepsilon/L$, it is enough to bound the probability that $\bar{Y}(i) - A'(\theta(i)) \geq \delta A''(\theta(i))$. We can write

$$\mathbb{P}\{\bar{Y}(i) - A'(\theta(i)) \geq \delta A''(\theta(i))\} = \mathbb{P}\left\{\frac{\sum_{j=1}^n Y_j(i)}{n} \geq A'(\theta(i)) + \delta A''(\theta(i))\right\} \leq \mathbb{E} \exp\left\{t \sum_{j=1}^n Y_j(i) - nt[A'(\theta(i)) + \delta A''(\theta(i))]\right\}.$$

The moment generating function of $Y_j(i)$ is $\exp[A(\theta(i) + t) - A(\theta(i))]$, so the last quantity equals $\exp[n\{A(\theta(i) + t) - A(\theta(i)) - tA'(\theta(i)) - t\delta A''(\theta(i))\}]$. For t small enough (depending on A'' on a neighborhood of $\theta(i)$), this is less than $\exp[-n\delta A''(\theta(i))/2]$. Since we assumed that $A''(\theta(i)) > c$ for some universal constant $c > 0$, we get the bound $\exp[-n\delta c/2]$.

We get a similar upper bound for the probability of deviation in the other direction. We conclude that for some constants $C, c' > 0$, $\sum_n \Pr(\max_i |1 - V(\bar{Y}(i))/A''(\theta(i))| > \varepsilon) \leq C \sum_n n \exp[-c'n] < \infty$. hence by the Borel-Cantelli lemma, $\max_i |1 - V(\bar{Y}(i))/A''(\theta(i))| \rightarrow 0$ almost surely. This finishes the proof. \square