*e*PCA: High Dimensional Exponential Family Principal Component Analysis

Lydia T. Liu '17

ORFE/PACM, Princeton University

April 20, 2017

Joint work with Edgar Dobriban and Amit Singer

Overview

Principal Component Analysis (PCA)

- **PCA**: useful tool for dimensionality reduction and data analysis
- Data X: $n \times p$ matrix (n samples from p-dimensional population)
- PC's: linear combinations of features explaining the most variance
- eigenvectors of sample covariance matrix $\hat{\Sigma}_n = \frac{1}{n} X^\top X$
- corresponding eigenvalue λ_i is variance of PC_i

PCA in classical vs. high-dimensional regimes

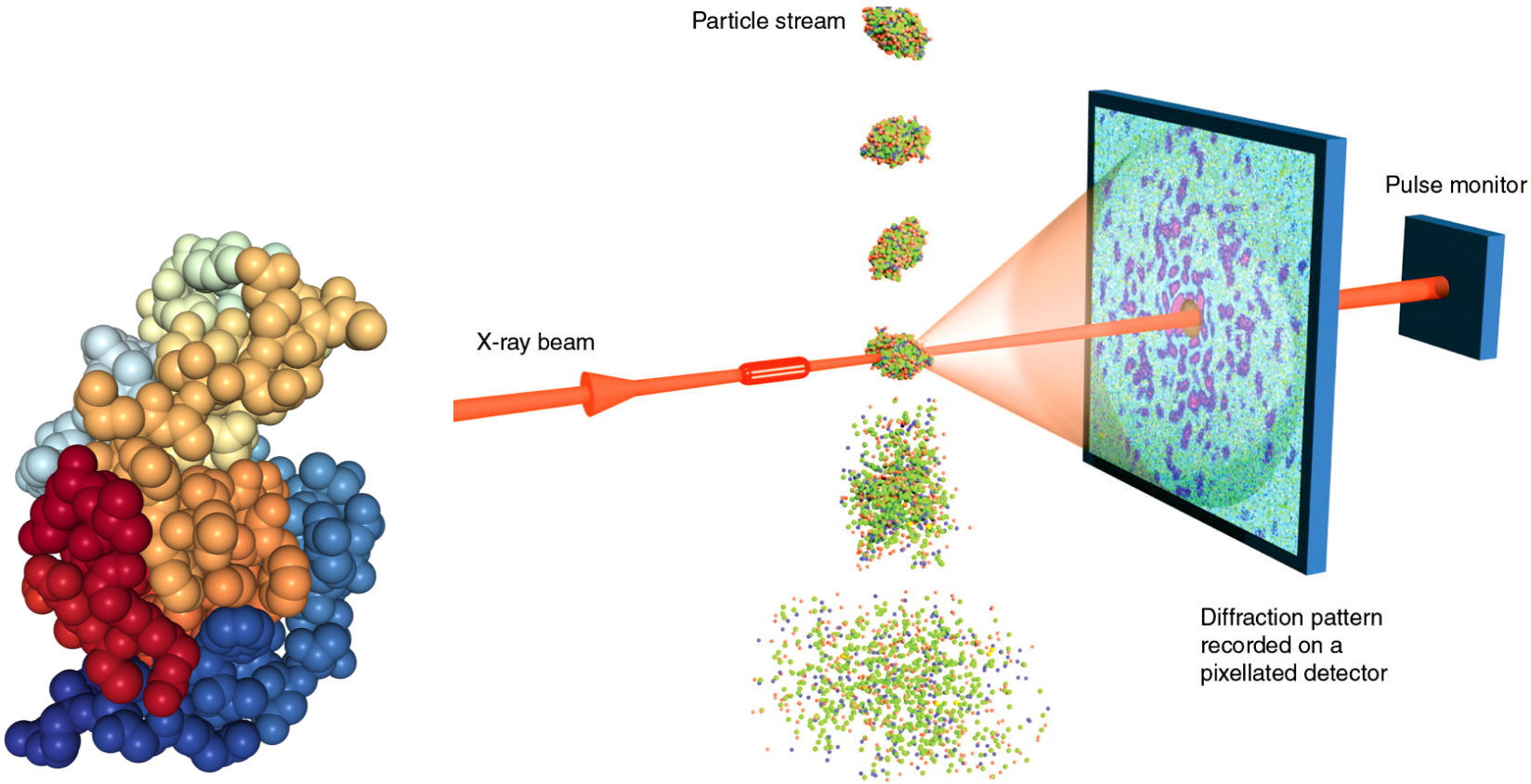
Classical setting: p fixed, $n \to \infty$

- $\hat{\Sigma}_n \to \Sigma$ by the law of large numbers

High dimensional setting: $p/n = \gamma \in (0, \infty), p \to \infty, n \to \infty$

- Classical asymptotics don't apply. Original PCA is inconsistent!
- Limiting spectrum: Marchenko-Pastur (MP) distribution (Marchenko and Pastur 1967)
- Acute need to **shrink** empirical eigenvalues (Donoho et al. 2013)

Motivating example for ePCA: XFEL

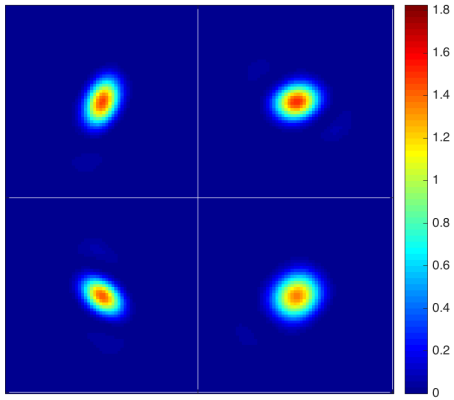


(a) 3D structure of a lysozyme

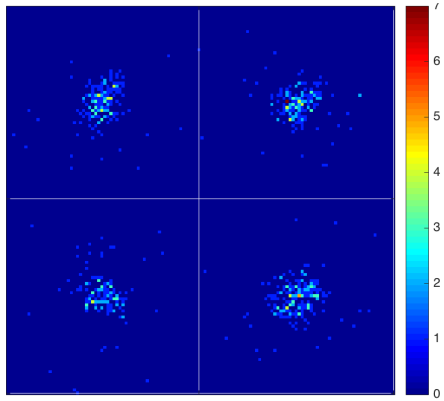
(b) XFEL imaging process (Gaffney and Chapman 2007)

Figure: X-ray free electron laser (XFEL) molecular imaging

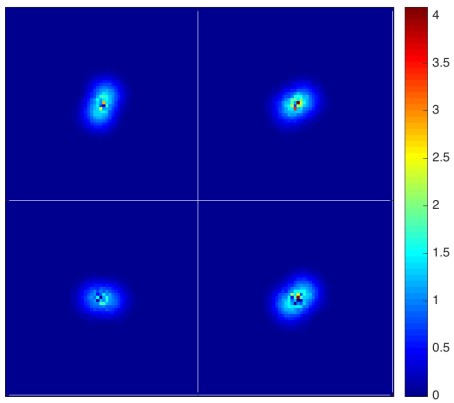
Demo of *e*PCA on XFEL images



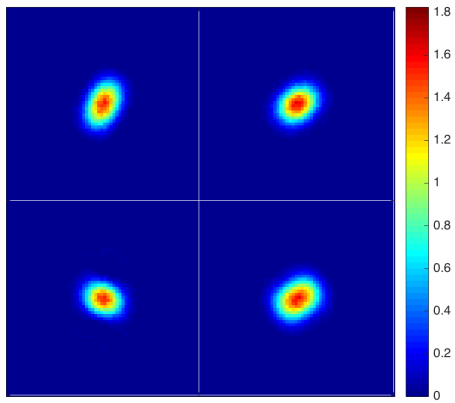
(a) Clean intensity maps



(b) Noisy photon counts



(c) Denoised (PCA)



(d) Denoised (ePCA)

Figure: XFEL diffraction images ($n = 70,000, p = 65,536$)

PCA for the exponential family

In many applications X_{ij}'s have an exponential family distribution

- SNPs: Binomial
- RNA-seq: Negative Binomial
- XFEL/photon-limited imaging: low-intensity Poisson

Single-parameter exponential family distributions

Has density of the form $f_\theta(y) = \exp \theta y - A(\theta)$.

No commonly agreed upon version of PCA for non-Gaussian data (Jolliffe 2002)

A new method: *e*PCA

Deterministic 4-step algorithm using moments and shrinkage

Advantages compared to previous proposals

- Likelihood/generalized linear latent variable models (Collins et al. 2001; Knott and Bartholomew 1999; Udell et al. 2016)
  - lack of global convergence guarantees

- Gaussianizing transforms: wavelet, Anscombe (Anscombe 1948; Starck et al. 2010)
  - unsuitable for low-intensity

Problem formulation

Sampling model for Poisson

- Each p-dim latent vector is drawn i.i.d. from distribution D

$$(X(1), \cdots, X(p))^\top = X \sim D(\mu, \mathbf{\Sigma})$$

— e.g., the noiseless image. μ and $\mathbf{\Sigma}$ are the mean and covariance of D.

- Observations $Y_i \sim Y \in \mathbb{R}^p$ — e.g., the noisy image
- Model for Y: draw latent $X \in \mathbb{R}^p$, then

$$Y = (Y(1), \cdots, Y(p))^\top \text{ where } Y(j) \sim \mathit{Poisson}(X(j))$$

Goal: Recover information about the original distribution D, i.e. $\mathbf{\Sigma}$. Recover X.

Problem formulation

Sampling model for Exponential family

- One-parameter exponential family with density

$$f_\theta(y) = \exp \theta y - A(\theta)$$

  Natural parameter θ, $\mathbb{E}[y] = A'(\theta), \mathrm{Var}[y] = A''(\theta)$
- Observations $Y_i \sim Y \in \mathbb{R}^p$
- Model for Y: draw latent $\boldsymbol{\theta} \in \mathbb{R}^p$, then

$$Y(j) \mid \boldsymbol{\theta}(j) \sim f_{\boldsymbol{\theta}(j)}(y), \;\; Y = (Y(1), \cdots, Y(p))^\top$$

Goal: Recover $\Sigma_x := \mathrm{Cov}(A'(\boldsymbol{\theta}))$ and $X := \mathbb{E}(Y|\boldsymbol{\theta}) = A'(\boldsymbol{\theta})$

Mean parameter is low-rank

Mean $X = \mathbb{E}(Y|\boldsymbol{\theta})$ has unknown low-dim structure

- as opposed to natural parameter $\boldsymbol{\theta}$
- can leverage Random Matrix Theory $\Rightarrow$ simple method
- reasonable for image data (Basri and Jacobs 2003)

Our sampling model is realistic

Two applications

	1. XFEL images	2. Single cell RNA-seq
Sample latent vector X	Random 2D intensity map due to random 3D orientation of molecule	Random expression rate due to biological variation of cells
Sample $Y \sim Poisson(X)$	Noisy 2D image Y due to low photon count	Noisy read counts Y due to technical variation

Summary of *e*PCA

*e*PCA can be seen as a sequence of improved covariance estimators

Table: Covariance estimators

Name	Formula	Motivation
Sample covariance	$S = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^\top$	-
Diagonal debiasing	$S_d = S - \text{diag}[V(\bar{Y})]$	Hierarchy
Homogenization	$S_h = D_n^{-1/2} S_d D_n^{-1/2}$	Heteroskedasticity
Shrinkage	$S_{h,\eta} = \eta(S_h)$	High dimensionality
Heterogenization	$S_{he} = D_n^{1/2} S_{h,\eta} D_n^{1/2}$	Heteroskedasticity
Scaling	$S_s = \sum \hat{\alpha}_i \hat{v}_i \hat{v}_i^\top$ $(S_{he} = \sum \hat{v}_i \hat{v}_i^\top)$	Heteroskedasticity

Diagonally debiasing the sample covariance

Poisson case:

- Sample Mean $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- Sample Covariance $S = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \overline{Y}_n\right)\left(Y_i - \overline{Y}_n\right)^T$
- But it's inconsistent! $\mathbb{E}[S] = \mathbf{\Sigma} + \mathrm{diag}[\mu]$
- $D_n := \mathrm{diag}\left(\overline{Y}_n\right)$
- Diagonally debiased covariance $S_d = S - D_n$

Diagonally debiasing the sample covariance

Exponential family:

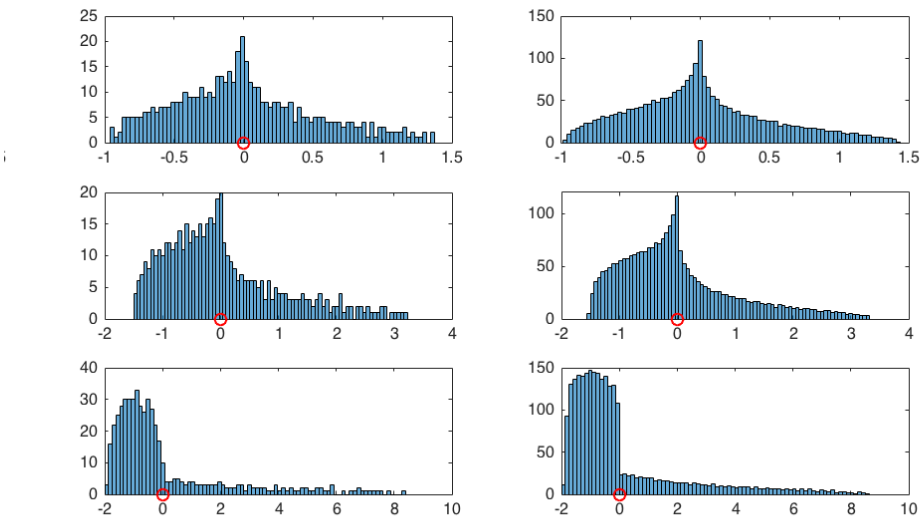
- $\mathrm{Cov}[Y] = \mathrm{Cov}[A'(\theta)] + \mathbb{E}\mathrm{diag}[A''(\theta)]$ by law of total variance
- Mean-variance map: $V(y) = A''[(A')^{-1}(y)]$
- $D_n := \mathrm{diag}[V(\overline{Y}_n)]$ estimates $\mathrm{diag}[A''(\theta)]$
- $S_d = S - D_n$

Theorem (Convergence of the debiased covariance (Liu et al. 2016))

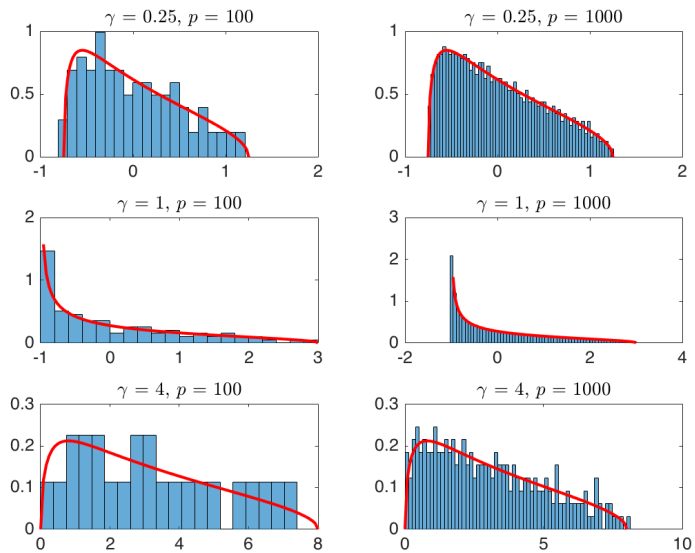
$$\mathbb{E}[\|S_d - \mathbf{\Sigma}_x\|_F] \lesssim \sqrt{\frac{p}{n}}\,[\sqrt{p} \cdot m_4 + \|\mu\|]$$

Homogenization

- Motivation: Remove the effects of heteroskedasticity $\Rightarrow$ closer to standard spiked model (Johnstone 2001) + MP law
- $S_h = D_n^{-1/2} S_d D_n^{-1/2} = D_n^{-1/2} S D_n^{-1/2} - I_p$
- Different from Standardization (dividing each measurement by its empirical standard deviation)!



(a) Spectrum before homogenization



(b) Spectrum after homogenization with red MP density

Marchenko-Pastur Law for *e*PCA

Theorem (Marchenko-Pastur Law (Liu et al. 2016))

- *The eigenvalue distribution of S converges a.s. to general MP $F_{\gamma,H}$.*
- *The eigenvalue distribution of $S_h + I_p$ converges a.s. to the standard MP with aspect ratio γ.*

Importance of Homogenization and the MP law

- Use optimal eigenvalue shrinkers for covariance estimation (Donoho et al. 2013; Lee et al. 2010)
- Improves signal strength (Liu et al. 2016)
- Matches Hardy-Weinberg equilibrium normalization (Patterson et al. 2006)

Eigenvalue shrinkage

- Reduce noise by shrinkage
- $\eta(\cdot)$ scalar shrinker, applied elementwise to eigenvalues:

$$M = V\Lambda V^\top;\ \eta(M) = V\eta(\Lambda)V^\top;\ S_{h,\eta} = \eta(S_h)$$

- Optimal shrinkage functions (Donoho et al. 2013)

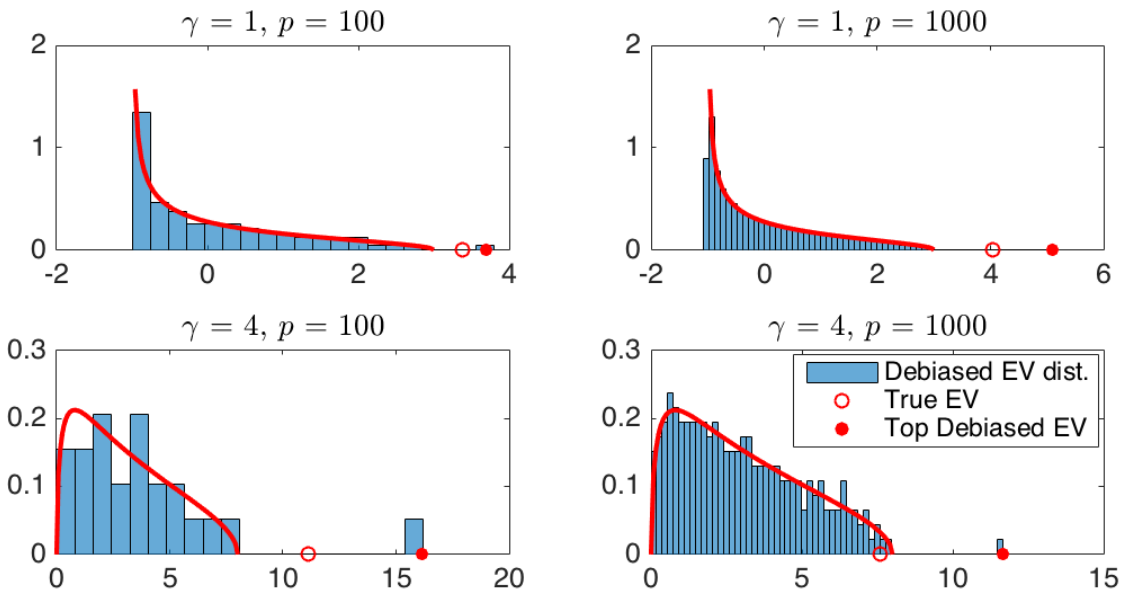


Figure: Need to Shrink! Spiked model spectrum after homogenization

Heterogenization

- Homogenized covariance matrix S_h doesn't estimate the original covariance matrix $\mathbf{\Sigma}$! $\Rightarrow$ need to heterogenize by multiplying back the estimated standard errors

- Improves eigenvector estimates

- $S_{he} = D_n^{1/2} \cdot S_{h,\eta} \cdot D_n^{1/2}$

Scaling

Heterogenization induces a bias in the eigenvalues (Liu et al. 2016)

- Need for final Scaling step to correct the bias
- $S_s = \sum \hat{\alpha}_i \hat{v}_i \hat{v}_i^\top$ ($S_{he} = \sum \hat{v}_i \hat{v}_i^\top$)
- Scaling function based on assuming a Gaussian spiked model (Baik et al. 2005; Johnstone 2001)
  - we conjecture the literature about phase transition also applies to our model

Denoising by EBLP

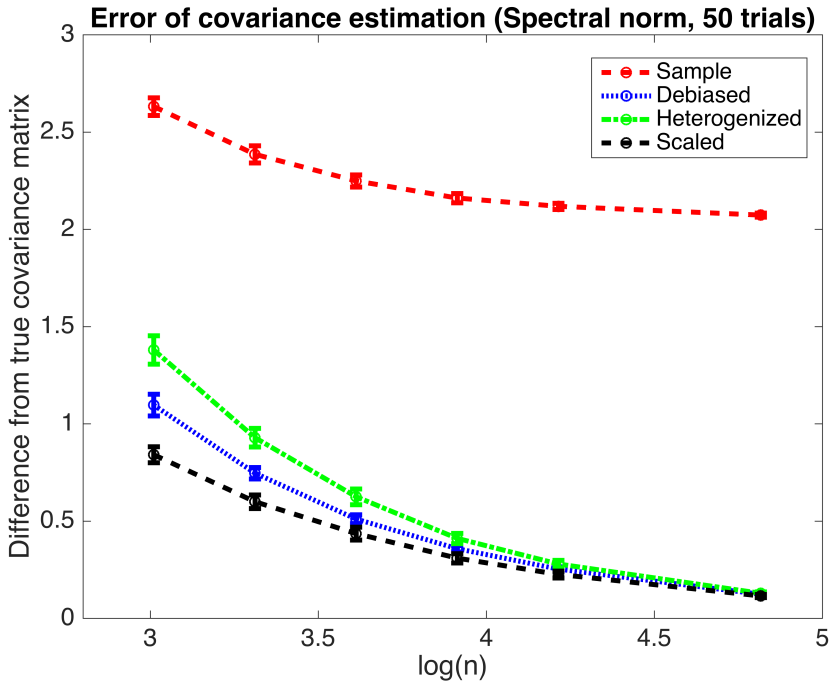
- Use standard strategy *Empirical Best Linear Predictor* (EBLP) from random effects model (Searle et al. 2009). First we estimate $\mathbb{E}[A''(\theta)]$ and Σ_x using *e*PCA. Then, estimate

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

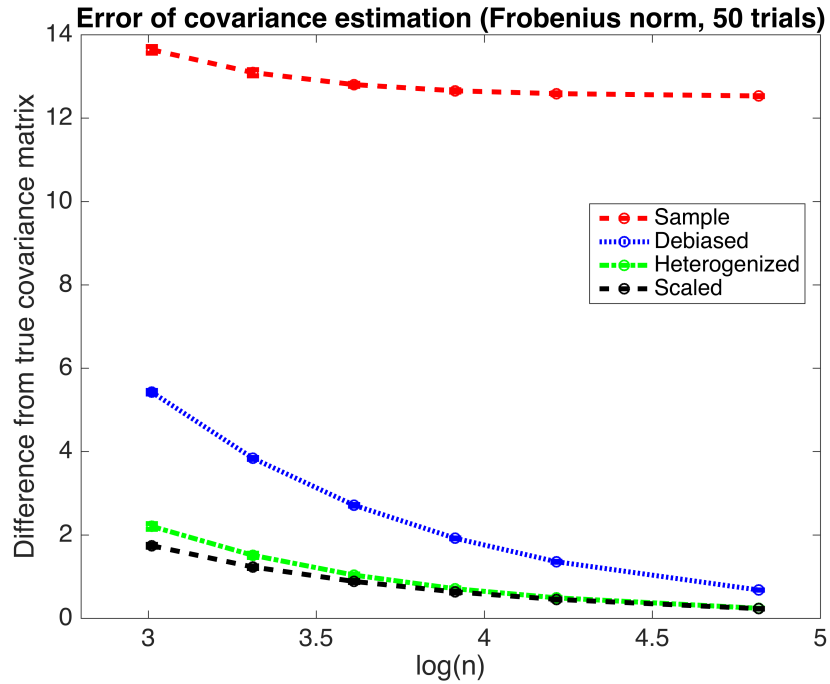
- For the Poisson distribution,

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

Covariance estimation



(a) Spectral norm



(b) Frobenius norm

Figure: Error of covariance matrix estimation: norm of the difference between each successive covariance estimate (Sample, +Debiased, +Heterogenized, +Scaled) and the true covariance $\mathbf{\Sigma}_x$.

Eigenvalues

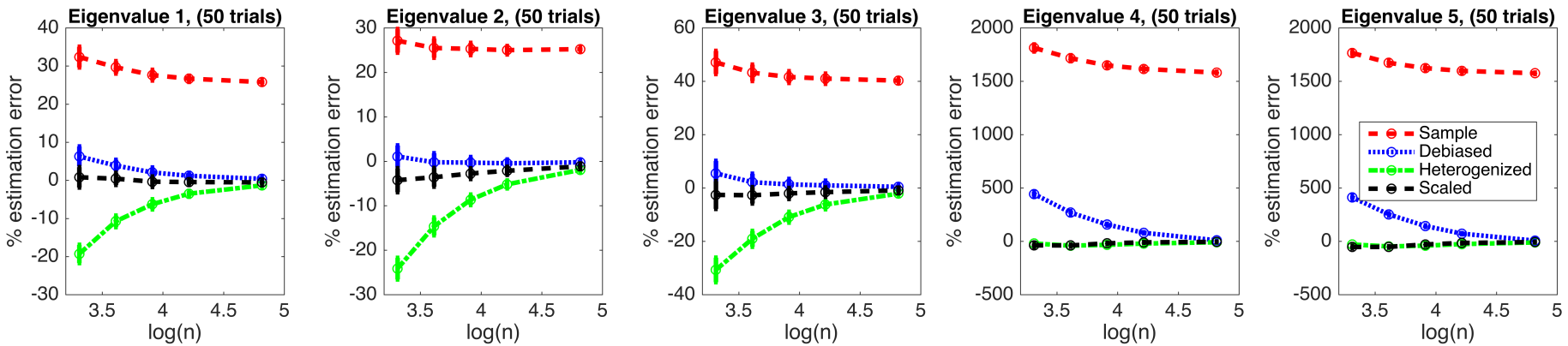


Figure: Error of eigenvalue estimation for the top 5 eigenvalues, measured as percentage error relative to the true eigenvalue

Eigenvectors

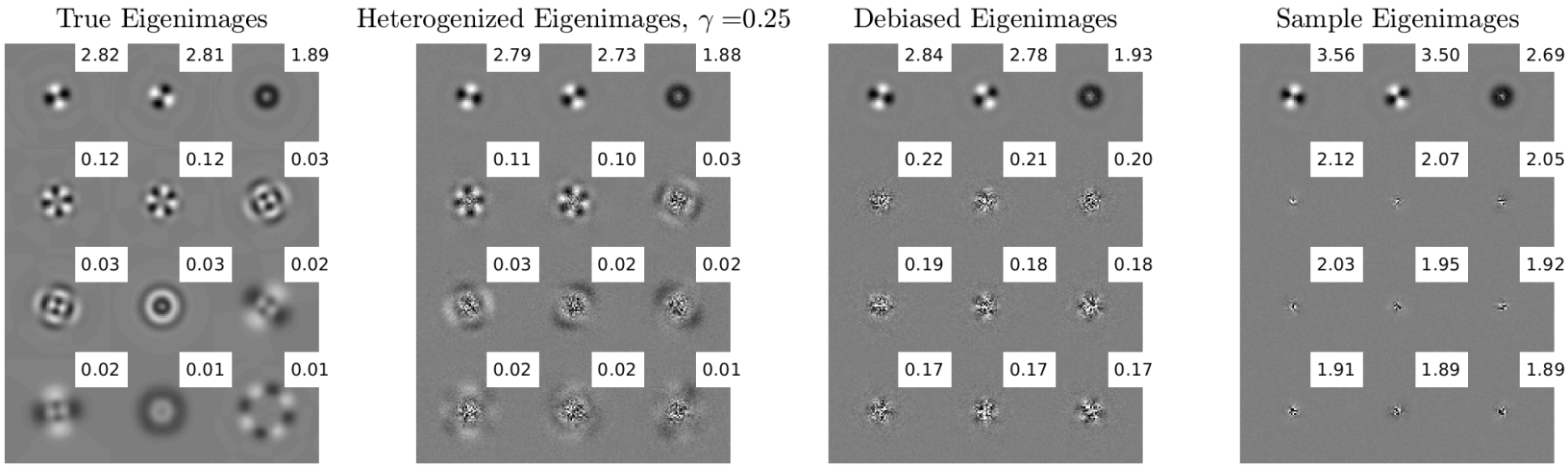


Figure: XFEL Eigenimages for $\gamma = 1/4$, ordered by eigenvalue

Denoising comparison

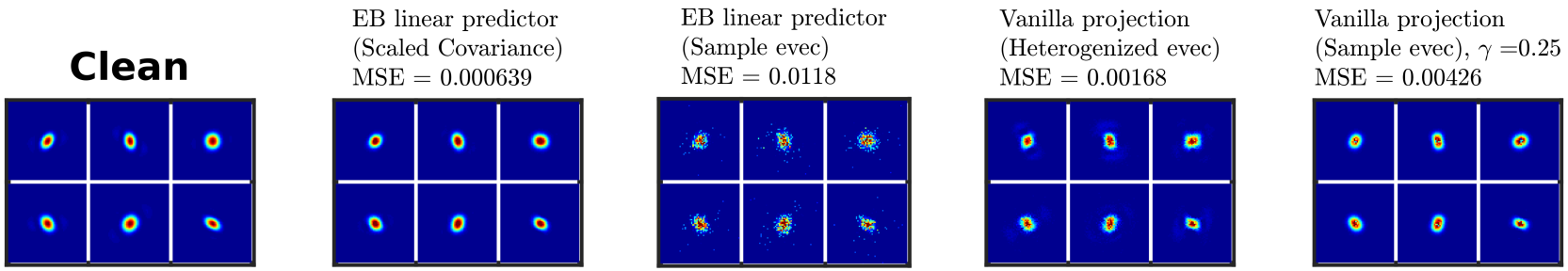


Figure: Sampled reconstructions using the XFEL dataset

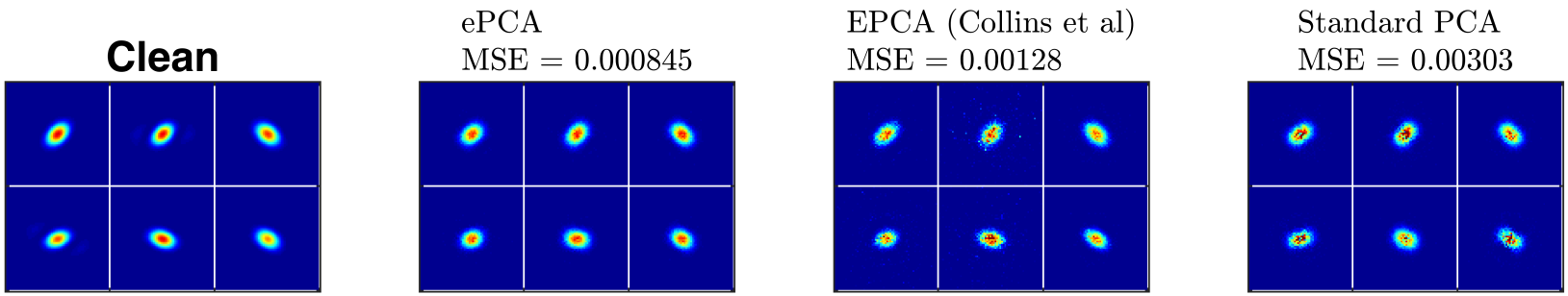


Figure: Comparing various methods' sampled reconstructions of the XFEL dataset *e*PCA took 13.9 seconds, while (Collins et al. 2001)'s exponential family PCA took 10900 seconds, or 3 hours

Denoising results

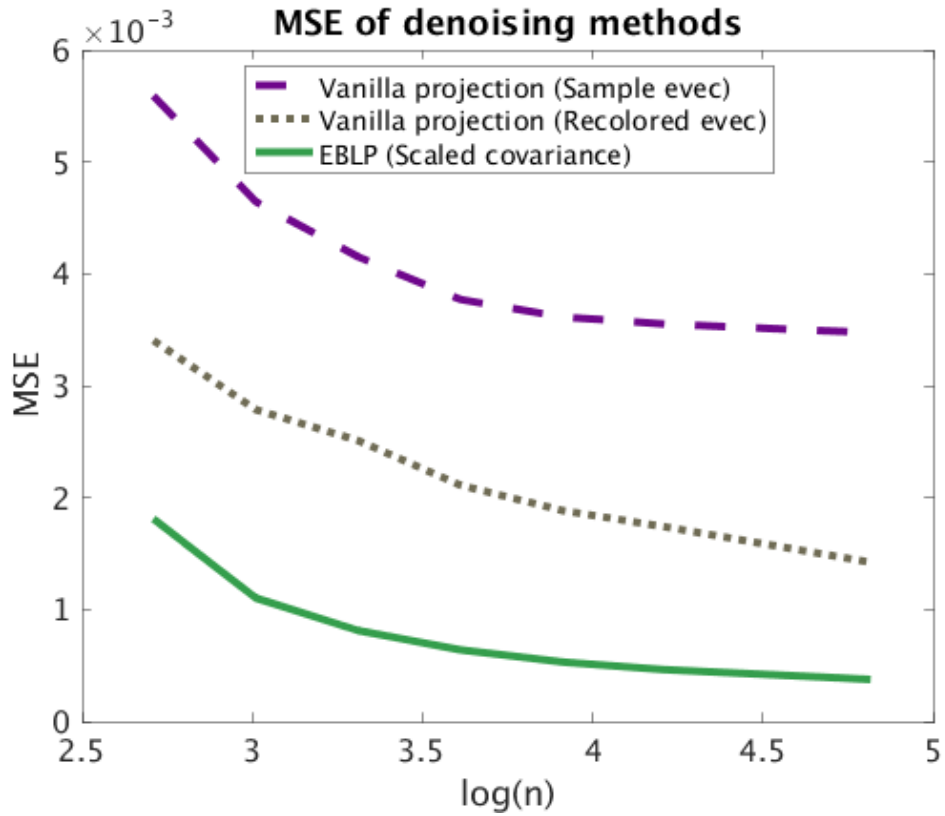


Figure: MSE against log10 sample size. Mean over 50 Monte Carlo trials. Purple: PCA. Grey: ePCA (projection only). Green: ePCA +EBLP

Software

- *e*PCA is publicly available in an open-source Matlab implementation from `github.com/lydiatliu/epca/`
- Link to *e*PCA main function
- e.g. `[cov,~,~,~,~,eigval,eigvec,~,~,~] = exp_fam_pca(data,'poisson')`

Summary

1. New method *e*PCA for PCA of exponential family data, based on a new covariance estimator.

2. *Homogenization*, *shrinkage*, *heterogenization*, and *scaling* of the debiased covariance matrix improve performance for high-dimensional data. Each step has theoretical justifications.

3. Applied ePCA to simulated XFEL data $\Rightarrow$ reduce the MSE for covariance, eigenvalue, eigenvector estimation.

4. Used *e*PCA to develop new denoising method, a form of empirical Best Linear Predictor (EBLP) from random effects models. Demonstrate on simulated XFEL data.

References I

F. J. Anscombe. "THE TRANSFORMATION OF POISSON, BINOMIAL AND NEGATIVE-BINOMIAL DATA". In: *Biometrika* 35.3-4 (1948), p. 246.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". In: *Annals of Probability* 33.5 (2005), pp. 1643–1697.

Ronen Basri and David W Jacobs. "Lambertian Reflectance and Linear Subspaces". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.2 (2003), pp. 218–233.

Michael Collins, S Dasgupta, and Re Schapire. "A generalization of principal component analysis to the exponential family". In: *Advances in Neural Information Processing Systems (NIPS)* (2001).

References II

DI Donoho, M Gavish, and Im Johnstone. "Optimal shrinkage of eigenvalues in the Spiked Covariance Model". In: *arXiv preprint arXiv:1311.0851* 0906812 (2013), pp. 1–35. arXiv: arXiv:1311.0851v1.

K. J. Gaffney and H. N. Chapman. "Imaging Atomic Structure and Dynamics with Ultrafast X-ray Scattering". In: *Science* 316.5830 (2007), pp. 1444–1448. ISSN: 0036-8075. DOI: 10.1126/science.1135923. eprint: http://science.sciencemag.org/content/316/5830/1444.full.pdf. URL: http://science.sciencemag.org/content/316/5830/1444.

Iain M Johnstone. "On the distribution of the largest eigenvalue in principal components analysis". In: *Annals of Statistics* 29.2 (2001), pp. 295–327.

References III

Ian Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2002.

Martin Knott and David J Bartholomew. *Latent variable models and factor analysis*. Edward Arnold, 1999.

Seunggeun Lee, Fei Zou, and Fred A Wright. "Convergence and prediction of principal component scores in high-dimensional settings". In: *Annals of Statistics* 38.6 (2010), pp. 3605–3629.

Vladimir A Marchenko and Leonid A Pastur. "Distribution of eigenvalues for some sets of random matrices". In: *Mat. Sb.* 114.4 (1967), pp. 507–536.

N Patterson, AL Price, and D Reich. "Population structure and eigenanalysis". In: *PLoS Genet* 2.12 (2006), e190.

References IV

Shayle R Searle, George Casella, and Charles E McCulloch. *Variance components*. John Wiley & Sons, 2009.

Andrey A Shabalin and Andrew B Nobel. "Reconstruction of a low-rank matrix in the presence of Gaussian noise". In: *Journal of Multivariate Analysis* 118 (2013), pp. 67–76.

Jean-Luc Starck, Fionn Murtagh, and Jalal M Fadili. *Sparse image and signal processing: wavelets, curvelets, morphological diversity*. Cambridge university press, 2010.

Madeleine Udell et al. "Generalized Low Rank Models". In: *Foundations and Trends in Machine Learning* 9.1 (2016), pp. 1–118.

References V

L. T. Liu, E. Dobriban, and A. Singer. "*e*PCA: High Dimensional Exponential Family PCA". In: *ArXiv e-prints* (Nov. 2016). eprint: 1611.05550.