*e*PCA: Exponential Family PCA

Edgar Dobriban

Stanford Statistics

Joint work with Lydia T. Liu and Amit Singer

February 21, 2017

Overview

Overview of my work on PCA

Concrete problem: exponential family PCA

- Motivating example: XFEL
- The *e*PCA method
- XFEL illustration

Other work

Overview

Overview of my work on PCA

Concrete problem: exponential family PCA

- Motivating example: XFEL
- The *e*PCA method
- XFEL illustration

Other work

Principal component analysis (PCA)

- PCA: useful tool in statistics and data science
- Data M: $n \times p$ matrix - n samples from p-dimensional population
- PCs: linear combinations of features explaining the most variance
- eigenvectors of sample covariance matrix $\widehat{\Sigma} = \frac{1}{n} M^\top M$
- corresponding eigenvalue λ_i is variance of PC i

PCA examples

Singular value decomposition for genome-wide expression data processing and modeling

Orly Alter*†, Patrick O. Brown‡, and David Botstein*

Departments of *Genetics and ‡Biochemistry, Stanford University, Stanford, CA 94305

Contributed by David Botstein, June 15, 2000

We describe the use of singular value decomposition in transforming genome-wide expression data from genes × arrays space to reduced diagonalized "eigengenes" × "eigenarrays" space, where the eigengenes (or eigenarrays) are unique orthonormal superpositions of the genes (or arrays). Normalizing the data by filtering out the eigengenes (and eigenarrays) that are inferred to represent noise or experimental artifacts enables meaningful comparison of the expression of different genes across different arrays in differ-

of any chosen subset of eigengenes (or eigenarrays). Upon comparing two or more similar experiments, with a regulator being overactive or underactive in one but normally expressed in the others, the expression pattern of one of the significant eigengenes may be correlated with the expression patterns of this regulator and its targets. This eigengene, therefore, can be associated with the observed genome-wide effect of the regulator. The expression pattern of the corresponding eigenarray is

OPEN ACCESS Freely available online

PLoS GENETICS

Population Structure and Eigenanalysis

Nick Patterson[1*], Alkes L. Price[1,2], David Reich[1,2]

1 Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America, 2 Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America

Current methods for inferring population structure from genetic data do not provide formal significance tests for population differentiation. We discuss an approach to studying population structure (principal components analysis) that was first applied to genetic data by Cavalli-Sforza and colleagues. We place the method on a solid statistical footing, using results from modern statistics to develop formal significance tests. We also uncover a general "phase change" phenomenon about the ability to detect structure in genetic data, which emerges from the statistical theory we use, and has an important implication for the ability to discover structure in genetic data: for a fixed but large dataset size, divergence between two populations (as measured, for example, by a statistic like F_{ST}) below a threshold is essentially undetectable, but a little above threshold, detection will be easy. This means that we can predict the dataset size needed to detect structure.

Citation: Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2(12): e190. doi:10.1371/journal.pgen.0020190

NeuroImage 142 (2016) 394–406

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Denoising of diffusion MRI using random matrix theory

Jelle Veraart[a,b,*], Dmitry S. Novikov[b], Daan Christiaens[c], Benjamin Ades-aron[b], Jan Sijbers[a], Els Fieremans[b]

[a] iMinds Vision Lab (Dept. of Physics), University of Antwerp, Antwerp, Belgium
[b] Center for Biomedical Imaging, Department of Radiology, New York University School of Medicine, NY, USA
[c] ESAT/PSI, Department of Electrical Engineering, KU Leuven, Leuven, Belgium

ARTICLE INFO

Article history:
Received 31 March 2016
Accepted 9 August 2016
Available online 11 August 2016

Keywords:
Marchenko-Pastur distribution
Precision
Accuracy
PCA

ABSTRACT

We introduce and evaluate a post-processing technique for fast denoising of diffusion-weighted MR images. By exploiting the intrinsic redundancy in diffusion MRI using universal properties of the eigenspectrum of random covariance matrices, we remove noise-only principal components, thereby enabling signal-to-noise ratio enhancements. This yields parameter maps of improved quality for visual, quantitative, and statistical interpretation. By studying statistics of residuals, we demonstrate that the technique suppresses local signal fluctuations that solely originate from thermal noise rather than from other sources such as anatomical detail. Furthermore, we achieve improved precision in the estimation of diffusion parameters and fiber orientations in the human brain without compromising the accuracy and spatial resolution.

Journal of the Royal Statistical Society SERIES B Statistical Methodology

J. R. Statist. Soc. B (2013)
75, *Part* 4, *pp.* 603–680

Large covariance estimation by thresholding principal orthogonal complements

Jianqing Fan,
Princeton University, USA

Yuan Liao
University of Maryland, College Park, USA

and Martina Mincheva
Princeton University, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, February 13th, 2013*, Professor G. A. Young *in the Chair*]

Summary. The paper deals with the estimation of a high dimensional covariance with a conditional sparsity structure and fast diverging eigenvalues. By assuming a sparse error covariance matrix in an approximate factor model, we allow for the presence of some cross-sectional correlation even after taking out common but unobservable factors. We introduce the principal orthogonal complement thresholding method 'POET' to explore such an approximate factor structure with sparsity. The POET-estimator includes the sample covariance matrix, the factor-based covariance matrix, the thresholding estimator and the adaptive thresholding estimator as specific examples. We provide mathematical insights when the factor analysis is approximately the same as the principal component analysis for high dimensional data. The rates of convergence of the sparse residual covariance matrix and the conditional sparse covariance matrix are studied under various norms. It is shown that the effect of estimating the unknown factors vanishes as the dimensionality increases. The uniform rates of convergence for the unobserved factors and their factor loadings are derived. The asymptotic results are also verified by extensive simulation studies. Finally, a real data application on portfolio allocation is presented.

PCA in statistics: classical vs big data

- Classical statistics: n large, p small
  - Well understood (Anderson, 2003).
- "Big data": n, p large: classical statistics misleading

Estimator	Low dim.	High dim.
$\widehat{\Sigma} = \frac{1}{n} M^\top M$	unbiased	unbiased
Eigenvalue λ_i	consistent	inconsistent
Eigenvector (PC)	consistent	inconsistent

My work on PCA

- computational characterization of eigenvalue bias (Dobriban, 2015a)
- optimal hypothesis testing/detection (Dobriban, 2016b)
- estimation in exponential families (Liu et al., 2016)
- prediction of missing, noisy data (Dobriban et al., 2016)

Range of my work: spans Data Science

- interdisciplinary applications (Fortney et al., 2015; Liu et al., 2016)
- software development (Dobriban, 2015b; Dobriban and Fortney, 2015)
- statistical methodology (Liu et al., 2016; Dobriban, 2016a)
- statistical theory (Dobriban, 2016b; Dobriban and Wager, 2015; Dobriban et al., 2016)
- probability (Dobriban et al., 2016)
- computational mathematics (Dobriban, 2015a)

Mathematics of PCA

- "Big data" PCA characterized by random matrix theory (RMT)
  - fast developing, challenging area of probability
  - potential to greatly improve multivariate data analysis

PCA+RMT: Bias of eigenvalues

- Eigenvalues have dramatic bias
  - Simple simulation shows this (next)
  - Theory since Marchenko and Pastur (1967)
  - Underlies everything in the field
- With Spectrode (Dobriban, 2015a) can compute it

PCA+RMT: Bias of eigenvalues

- $M = Z_{n\times p}\Sigma^{1/2}$
  - $Z_{n\times p}$ has iid standardized entries
  - Σ unobserved $p \times p$ covariance matrix: $\text{Cov}\,[m_i, m_i] = \Sigma$
- High dimension: $n, p \to \infty$, $p/n \to \gamma > 0$
- $H_p = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i(\Sigma)} \to H$
- Eigenvalues of $\widehat{\Sigma}$: Marchenko-Pastur (MP) distribution (Marchenko and Pastur, 1967)
  - $F_p = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i(\widehat{\Sigma})} \to F_{\gamma,H}$
- Deep & Fundamental

Standard MP distribution

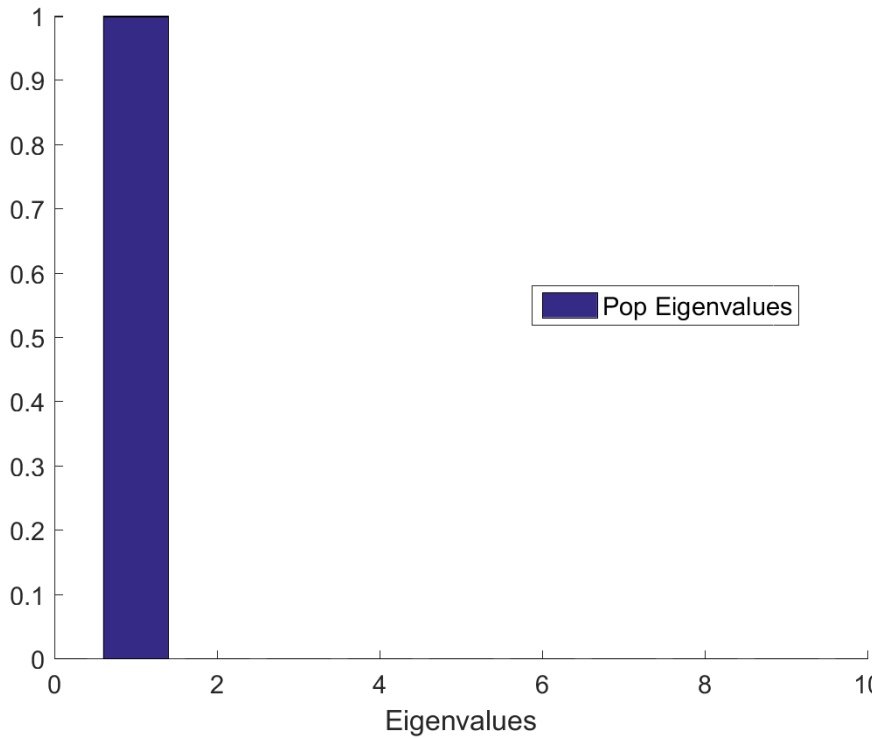


Figure: Eigenvalues H_p of an identity covariance matrix $\Sigma = I_p$.

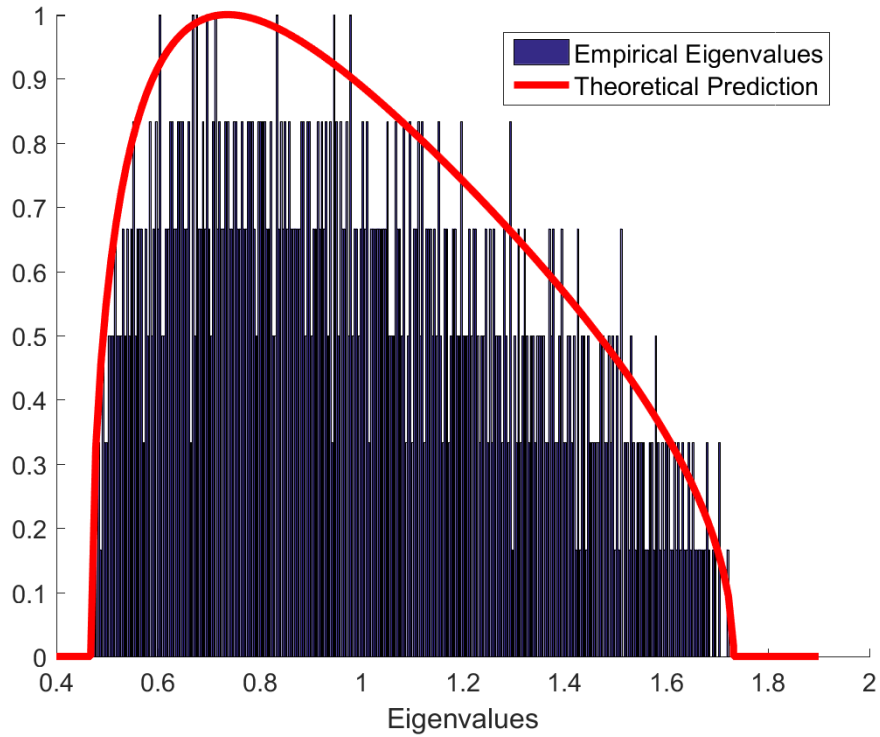


Figure: Eigenvalues F_p of $\widehat{\Sigma} = \frac{1}{n} M^\top M$ (blue), and standard MP distribution $F_{\gamma,H}$ (red). $\gamma = 1/10$.

General MP distribution $F_{\gamma,H}$

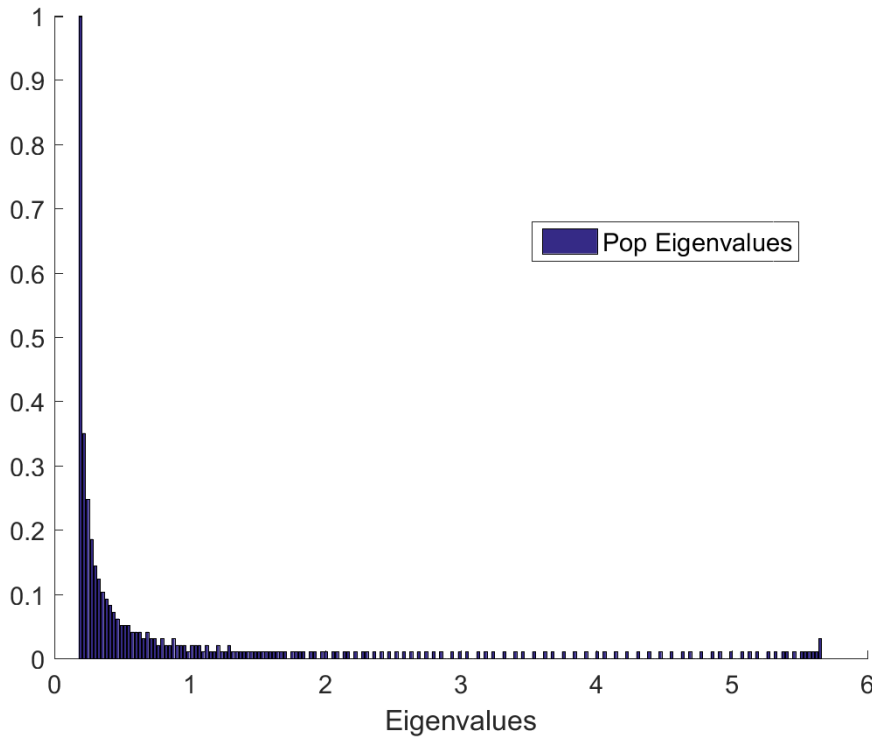


Figure: Eigenvalues H_p of an AR-1 covariance matrix Σ with $\Sigma_{ij} = \rho^{|i-j|}$ ($p = 400$; $\rho = 0.7$). $H = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i}$

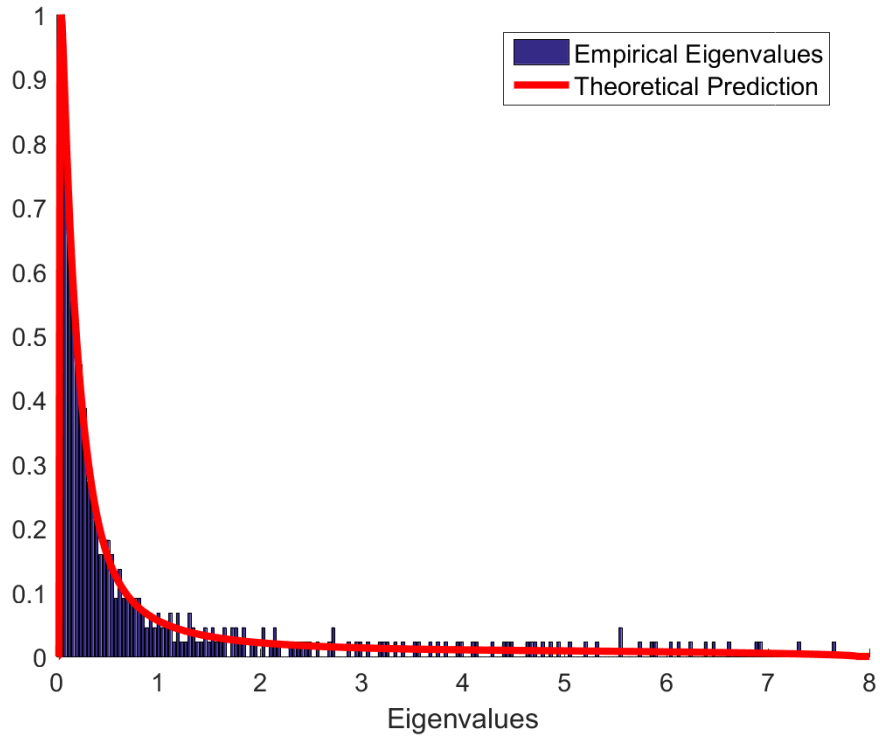


Figure: Eigenvalues F_p of a sample covariance matrix $\widehat{\Sigma}$ with $n = 800$ samples. Density of $F_{\gamma,H}$ computed with Spectrode (Dobriban, 2015a).

MP distribution

- Will see it later
- Optimal testing: determines form of test
- *e*PCA: Will bias-correct—shrink—eigenvalues

Optimal testing

- Eigenvalue bias: engulfs signal
- Popular top eigenvalue test has trivial power
- Linear statistic $\mathrm{tr}\{\varphi(\widehat{\Sigma})\}$ has more power

Optimal testing under local alternatives

$$H_{p,0} : H_p = H, \text{ vs}$$
$$H_{p,1} : H_p = \left(1 - \frac{h}{p}\right) H + \frac{h}{p} G.$$

- Under $H_{p,0}$, $\mathrm{tr}(\varphi(\widehat{\Sigma})) - c_p \Rightarrow \mathcal{N}(0, \sigma_\varphi^2)$
- Under $H_{p,1}$, $\mathrm{tr}(\varphi(\widehat{\Sigma})) - c_p \Rightarrow \mathcal{N}(\mu_\varphi, \sigma_\varphi^2)$.

$$\mu_\varphi = -h\langle \varphi', \Delta \rangle \text{ and } \sigma_\varphi^2 = \langle \varphi', K\varphi' \rangle.$$

Theorem (Dobriban (2016b))

1. *If* $\Delta \in \mathrm{Im}(K)$, *the optimal* φ *obey a Fredholm integral equation:*

$$K(\varphi') = -\eta\Delta.$$

2. *If* $\Delta \notin \mathrm{Im}(K)$, *the power is unity.*

Weak derivative of the Marchenko-Pastur map

Theorem (Dobriban (2016b))

The forward Machenko-Pastur map $H \to F_{\gamma,H}$ has a well-defined weak derivative for compactly supported H, G

$$\frac{F_{\gamma,(1-\varepsilon)H+\varepsilon G} - F_{\gamma,H}}{\varepsilon} \Rightarrow \delta F_\gamma(H, G) = \Delta.$$

The limit δF_γ is a compactly supported signed measure with finite total variation, and has zero total mass: $\delta F_\gamma(\mathbb{R}) = 0$. Furthermore,

1. *The companion Stieltjes transform $s(z)$ of $\delta F_\gamma(H, G)$ is*

$$s(z) = -\gamma\, v'(z) \int \frac{t}{1 + tv(z)} d\nu(t), \tag{1}$$

   where $\nu = G - H$, and $v(z)$ is the companion Stieltjes transform of $F_{\gamma,H}$.
2. *The weak derivative is affine in the second argument: $\delta F_\gamma(H, aP + bQ) = a\,\delta F_\gamma(H, P) + b\,\delta F_\gamma(H, Q)$.*

Overview

Overview of my work on PCA

Concrete problem: exponential family PCA

- Motivating example: XFEL
- The *e*PCA method
- XFEL illustration

Other work

PCA in exponential families (Liu et al., 2016)

- In many applications Y_{ij} have an exponential family distribution
  - SNPs: Binomial
  - RNA-seq: Negative Binomial
  - photon-limited XFEL: low-intensity Poisson
- No commonly agreed upon version of PCA for non-Gaussian data (Jolliffe, 2002)

X-ray free electron lasers (XFEL)

- Increasingly popular method to find 3D structure of molecules (e.g., Favre-Nicolin et al., 2015; Maia and Hajdu, 2016)
- 2-D diffraction patterns of single particles

XFEL

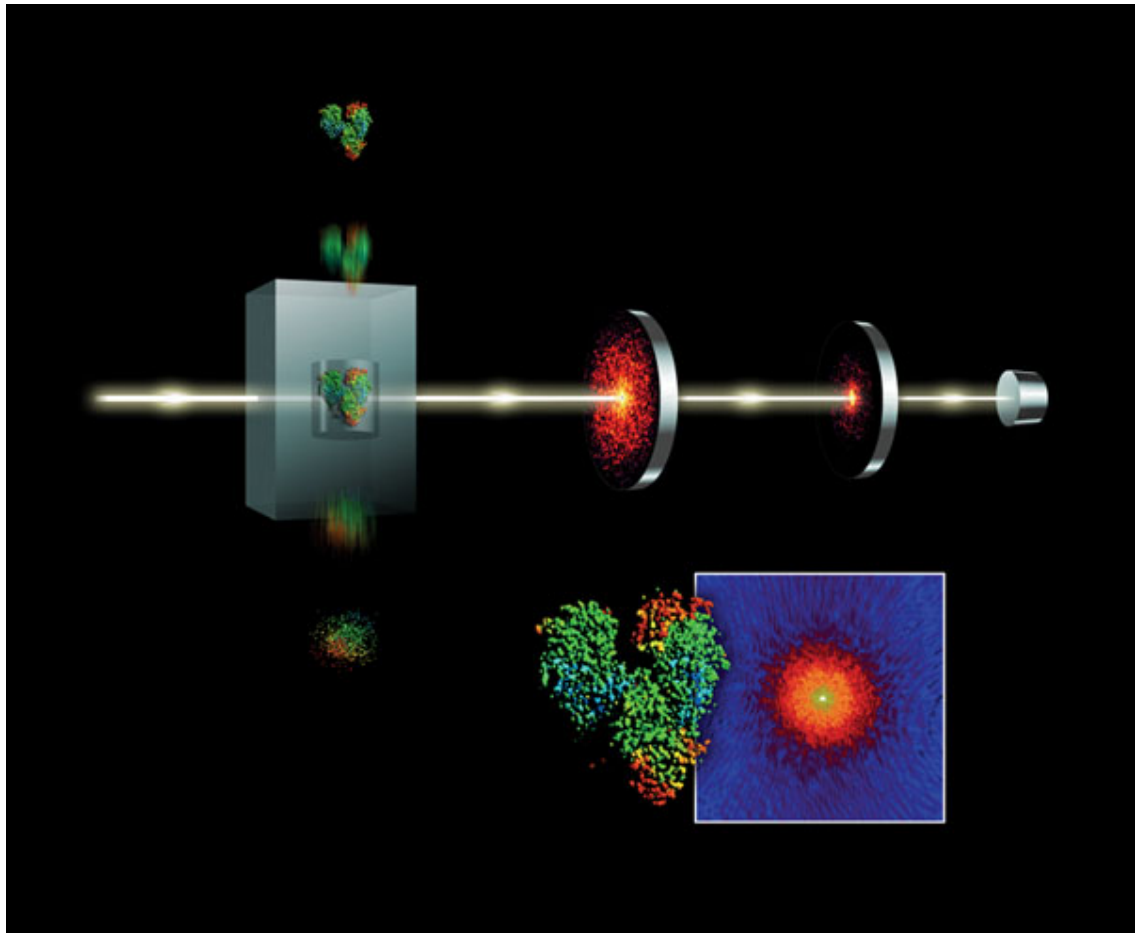


Figure: Schematic of XFEL imaging process, courtesy of SLAC

XFEL

- Short femtosecond X-ray pulses — molecule does not change structure
- Low number of photons — Poisson count-noise
- Eventual goal: structure reconstruction
- Crucial step: PCA/denoising (Kam, 1980)

XFEL example

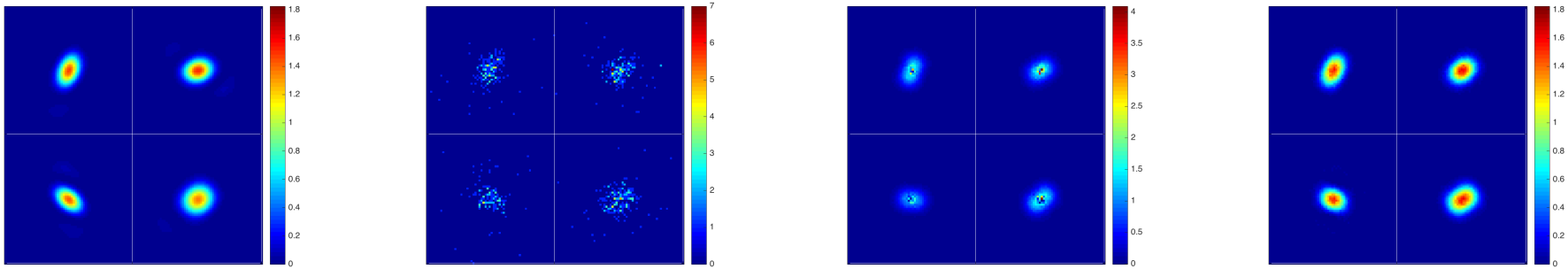


Figure: XFEL diffraction images. From left to right: Clean intensity maps. Noisy photon counts. Denoised (PCA). Denoised (*e*PCA).

Overview

Overview of my work on PCA

Concrete problem: exponential family PCA

- Motivating example: XFEL
- The *e*PCA method
- XFEL illustration

Other work

The new *e*PCA method

- Deterministic four-step alg. using moments and shrinkage
- Advantages compared to previous proposals
  - likelihood/generalized linear latent variable models (Knott and Bartholomew, 1999; Collins et al., 2001; Udell et al., 2016)
    - non-convex heuristics
  - Gaussian-izing transform: wavelet, Anscombe (Jolliffe, 2002; Starck et al., 2010)
    - unsuitable for low-intensity

ePCA

- Eigendecomposition of a new covariance estimator
- Start with sample covariance, do algebra, shrinkage

Name	Formula
Sample covariance	$S = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^\top$
Whitening	$S_w = D_n^{-1/2} S D_n^{-1/2}$
Shrinkage	$S_{w,\eta} = \eta(S_w)$
Recoloring	$S_r = D_n^{1/2} S_{w,\eta} D_n^{1/2}$
Scaling	$S_s = \sum \hat{\alpha}_i \hat{v}_i \hat{v}_i^\top$, where $S_r = \sum \hat{v}_i \hat{v}_i^\top$

*e*PCA: Sampling model

- One-parameter exponential family with density (wrt some measure)

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

- $\mathbb{E}[y] = A'(\theta)$, $\mathrm{Var}\,[y] = A''(\theta)$
- Observations $Y_i \sim Y \in \mathbb{R}^p$ — e.g., the noisy image
- Model for Y: draw latent $\theta \in \mathbb{R}^p$, then

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

*e*PCA: The mean model

- Mean $X := \mathbb{E}(Y|\theta) = A'(\theta)$ has unknown low-dim structure
  - as opposed to natural parameter θ
  - can leverage RMT $\rightarrow$ simple method
  - reasonable for image data (Basri and Jacobs, 2003)

ePCA: The covariance

- Covariance of Y conditional on θ:

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

- $Y = A'(\theta) + \mathrm{diag}[A''(\theta)]^{1/2}\varepsilon$, where $\varepsilon(i)|\theta$ independent standardized
- Marginal covariance of Y

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

ePCA Step 1/4: Whitening

- Remove heteroskedastic noise variances $D = \text{diag}[A''(\theta)]$

$$Y = A'(\theta) + D^{1/2}\varepsilon \quad \to \quad D^{-1/2}Y = D^{-1/2}A'(\theta) + \varepsilon$$

- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $S = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^\top$
- Mean-variance map: $V(m) = A''[(A')^{-1}(m)]$
- $D_n = \text{diag}[V(\bar{Y})]$ estimates D
- Whitening: $S_w = D_n^{-1/2} S D_n^{-1/2}$

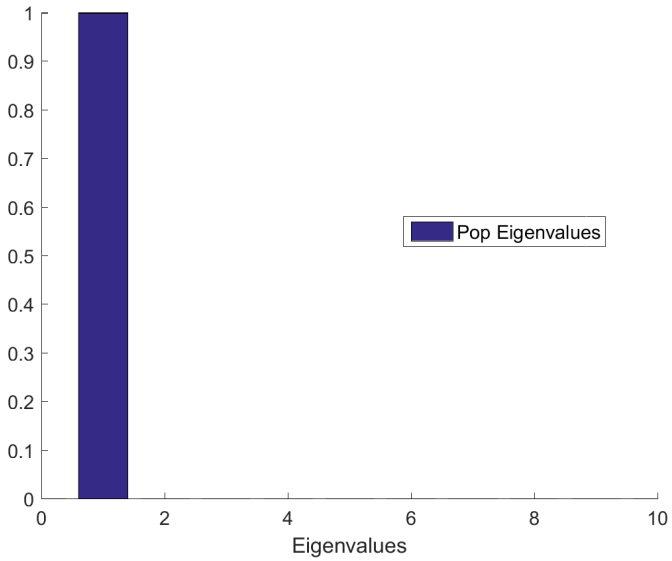
ePCA: Marchenko-Pastur (MP) law

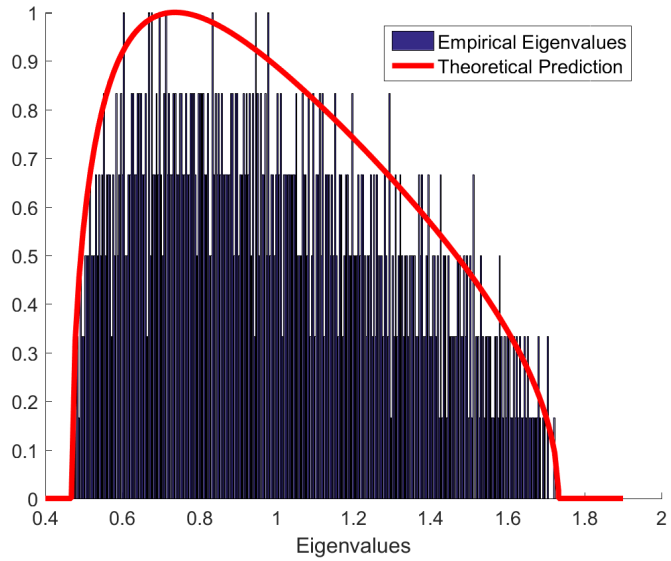
- $\theta \in \mathbb{R}^p$ fixed sequence of latent vectors
- $A''(\theta(i)) > c$ for some $c > 0$
- $n, p \to \infty$ so that $p/n \to \gamma > 0$
- $H_p = \frac{1}{p} \sum_{i=1}^{p} \delta_{A''(\theta(i))}$, $H_p \Rightarrow H$

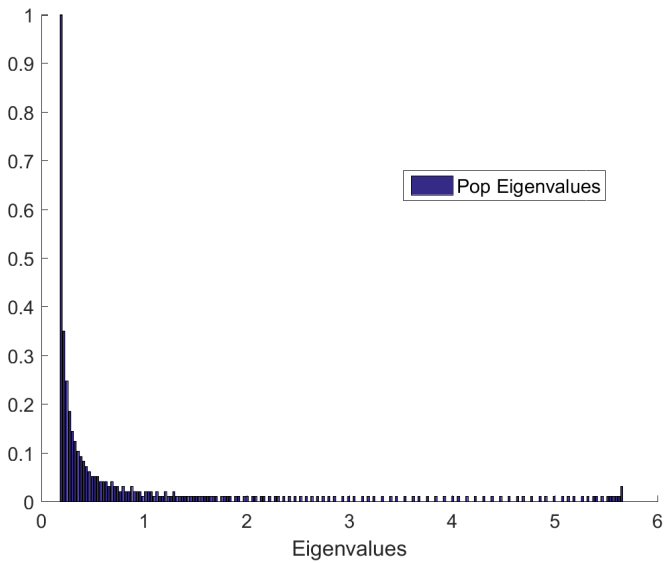
Theorem (MP law, LDS'16)

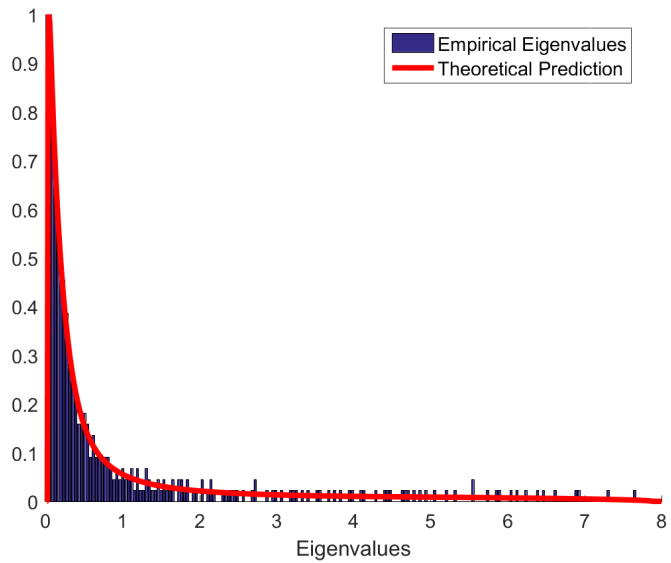
1. *The eigenvalue distribution of S converges a.s. to the general MP distribution $F_{\gamma,H}$.*
2. *The eigenvalue distribution of S_w converges a.s. to the standard MP distribution.*

MP distributions: Standard & General









*e*PCA: MP law importance/implications

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

*e*PCA Step 2/4: Eigenvalue shrinkage

- Reduce noise by shrinkage
- $\eta(\cdot)$ scalar shrinker, applied elementwise to eigenvalues
- Shrink $C = U \cdot \Lambda \cdot U^\top \to \eta(C) = U \cdot \eta(\Lambda) \cdot U^\top$:

$$S_{w,\eta} = \eta(S_w) = \eta(D_n^{-1/2} S D_n^{-1/2}).$$

- MP law "universality": use optimal shrinkers

*e*PCA Step 3/4: Recoloring

- Recolor to improve PCs:

$$S_r = D_n^{1/2} \cdot S_{w,\eta} \cdot D_n^{1/2} = D_n^{1/2} \cdot \eta(D_n^{-1/2} S D_n^{-1/2}) \cdot D_n^{1/2}.$$

- Perhaps surprisingly, induces bias in "signal" eigenvalues

*e*PCA Step 4/4: Scaling

- After whitening assume formulas for Gaussian "spiked model" (Johnstone, 2001; Baik et al., 2005; Baik and Silverstein, 2006)
  - signal eigenvalues: $\lambda_i \to \lambda(\ell_i; \gamma)$
  - signal PCs: $(v_i^\top \hat{v}_i)^2 \to c^2(\ell_i; \gamma)$
- Scale $S_r = \sum_{i=1}^k \hat{\lambda}_i \hat{u}_i \hat{u}_i^\top \to S_s = \sum_{i=1}^k \alpha_i \hat{\lambda}_i \hat{u}_i \hat{u}_i^\top$,
  - $$\hat{\alpha}_i = \frac{1 - \hat{s}_i^2 \tau_i}{\hat{c}_i^2}$$
  - $\hat{\ell}_i = \lambda^{-1}(\lambda_i(S_w); \gamma)$, $\tau_i = \frac{\operatorname{tr} D_n}{p} \cdot \frac{\hat{\ell}_i}{\|\hat{v}_i\|^2}$
  - $\hat{c}_i^2 = c^2(\hat{\ell}_i; \gamma)$, $\hat{s}_i^2 = 1 - \hat{c}_i^2$

Spike descriptors λ, c^2

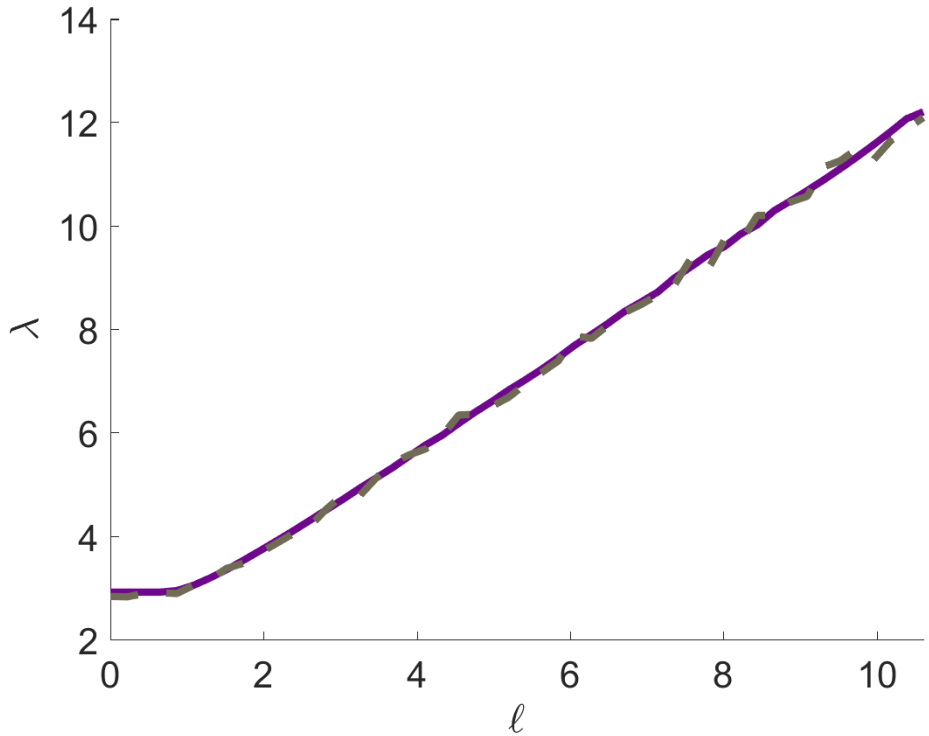


Figure: Spike forward map $\ell \to \lambda(\ell; \gamma)$. $\gamma = 1/2$. $n = 200$. 10 MC simulations.

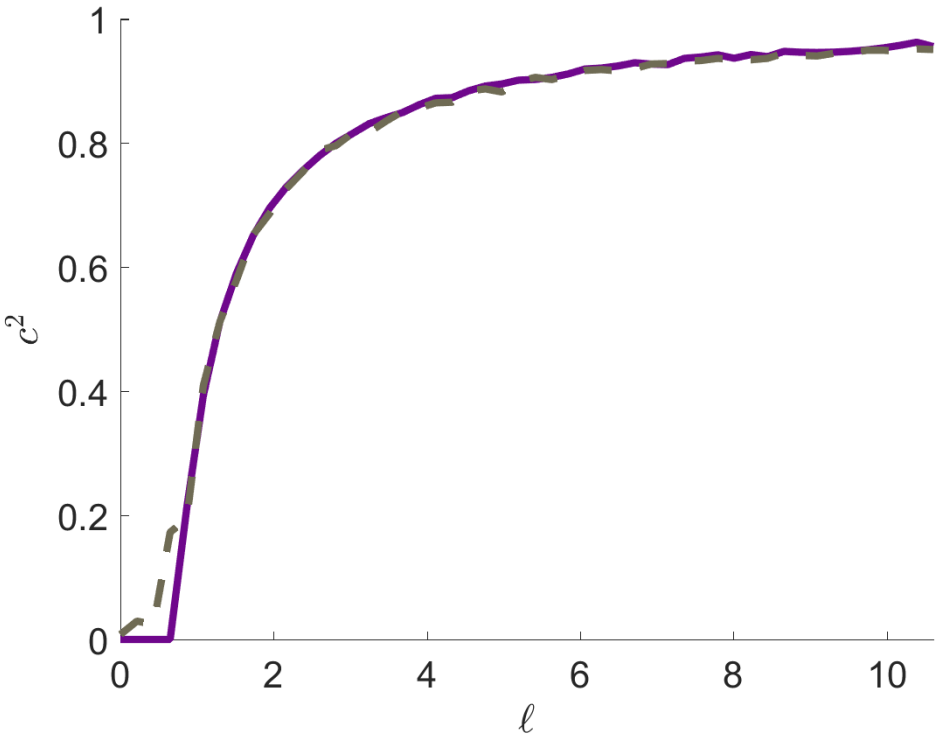


Figure: Cosine forward map $\ell \to \lambda(\ell; \gamma)$

ePCA summary

Name	Formula
Sample covariance	$S = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^\top$
Whitening	$S_w = D_n^{-1/2} S D_n^{-1/2}$
Shrinkage	$S_{w,\eta} = \eta(S_w)$
Recoloring	$S_r = D_n^{1/2} S_{w,\eta} D_n^{1/2}$
Scaling	$S_s = \sum \hat{\alpha}_i \hat{v}_i \hat{v}_i^\top$, where $S_r = \sum \hat{v}_i \hat{v}_i^\top$

Overview

Overview of my work on PCA

Concrete problem: exponential family PCA

- Motivating example: XFEL
- The *e*PCA method
- XFEL illustration

Other work

XFEL data analysis

- physically realistic simulation with Condor (Hantke et al., 2016)
- $n = 20,000$ diffraction maps of lysozyme (Protein Data Bank 1AKI)
- 64×64 pixels, so $p = 4096$
- sample maps at random, then sample pixel photon count from a Poisson distribution whose mean is the pixel intensity.

Lysozyme structure

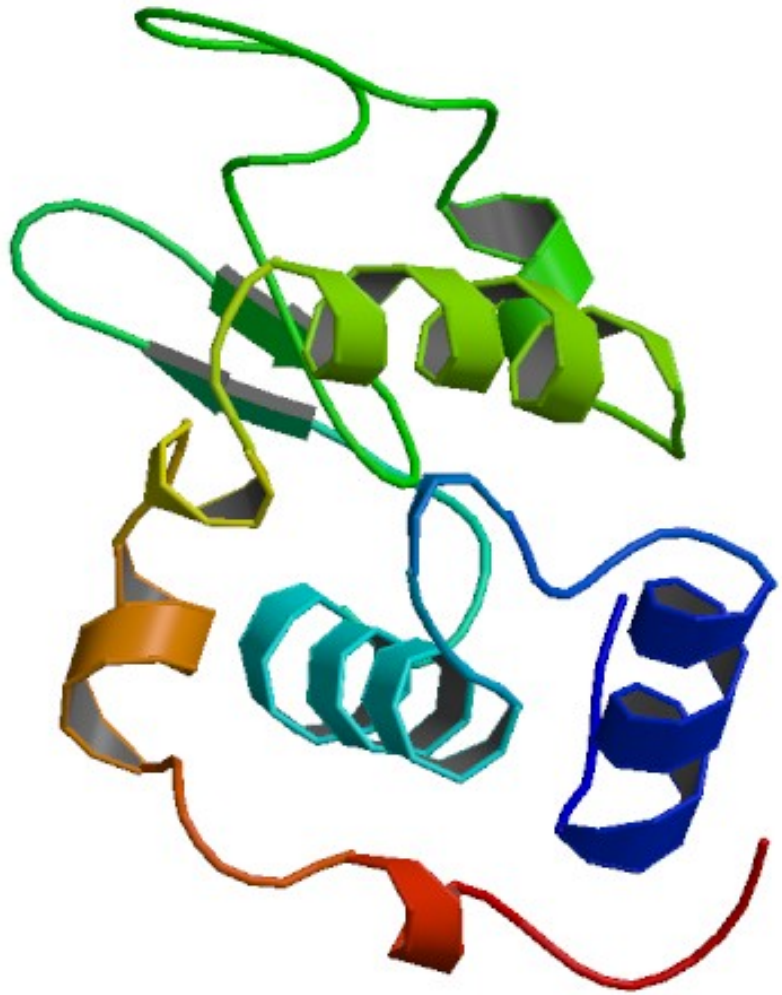


Figure: Structure of Lysozyme (PDB 1AKI), courtesy of PDB

*e*PCA Eigenimages

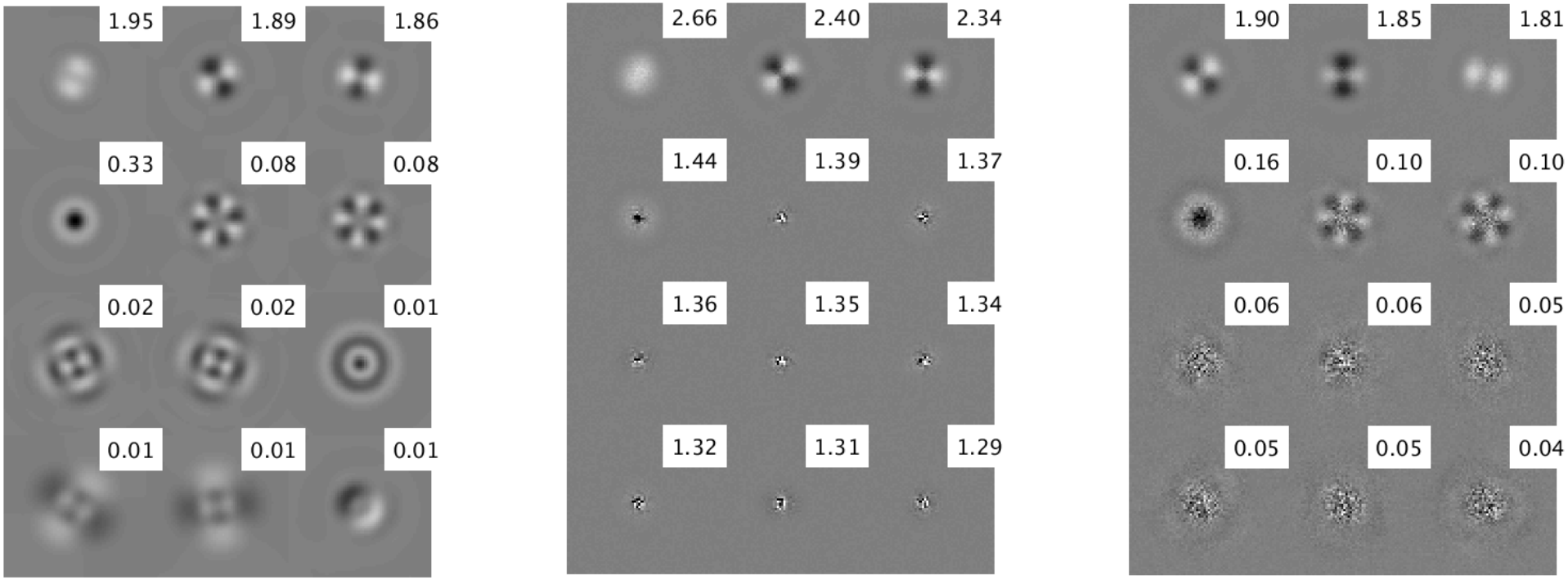


Figure: XFEL Eigenimages ordered by eigenvalue. From left to right: Clean eigenimages. PCA. *e*PCA.

XFEL images and *e*PCA Denoising

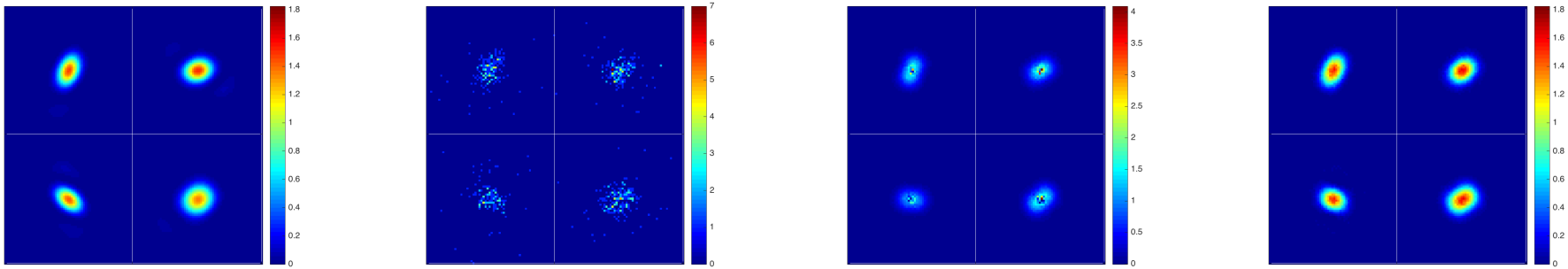


Figure: XFEL diffraction images. From left to right: Clean intensity maps. Noisy photon counts. Denoised (PCA). Denoised (*e*PCA).

Denoising MSE

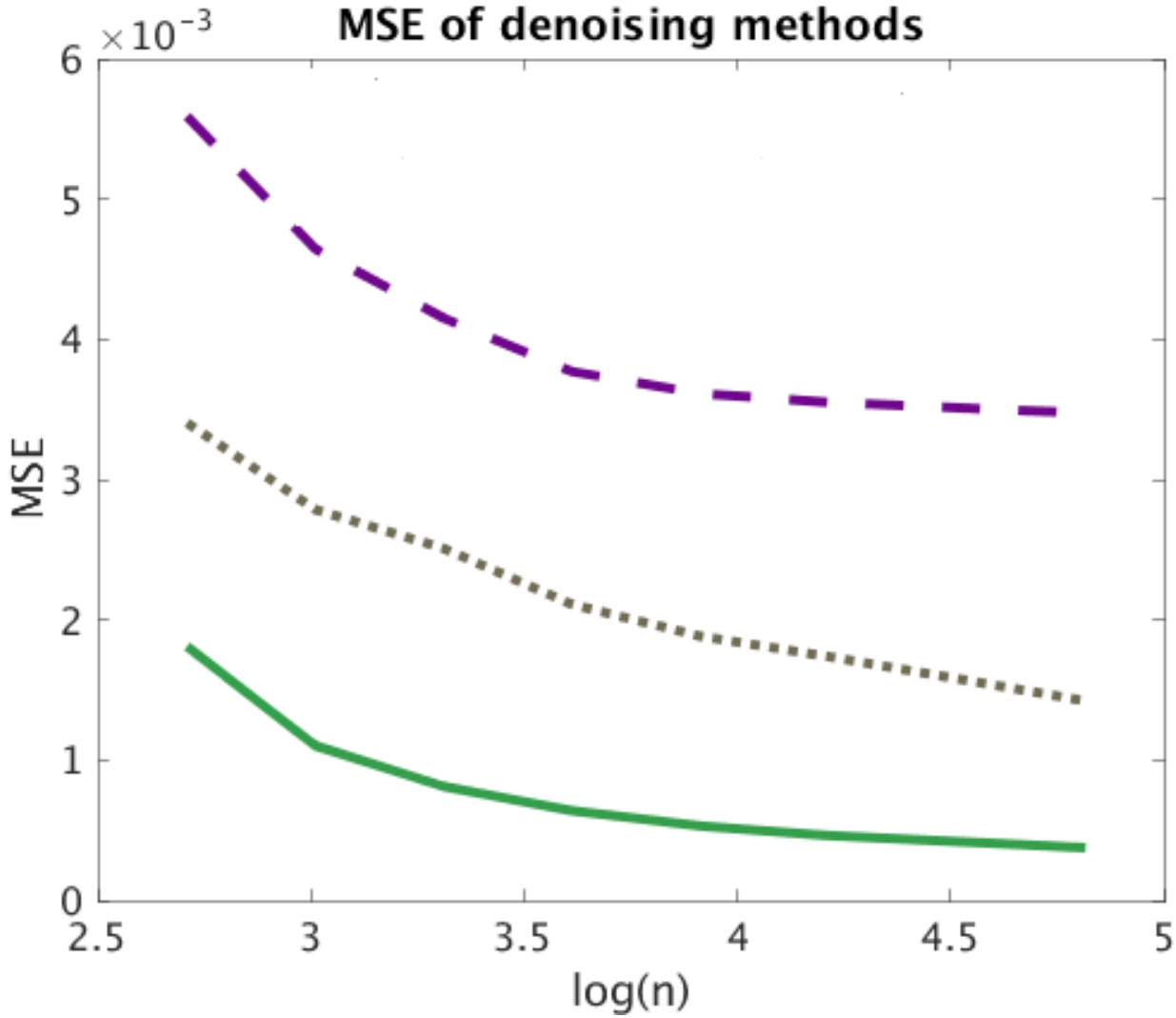


Figure: MSE against $\log_{10}$ sample size. Mean over 50 Monte Carlo trials. Purple: PCA (projection). Grey: *e*PCA (projection). Green: *e*PCA (EBLP)

*e*PCA

- extension of PCA to exponential family data
- simple deterministic non-iterative algorithm
- with theoretical justification
- PCA $\rightarrow$ *e*PCA
  - similar to linear models $\rightarrow$ GLMs
  - previous approaches claim same (Collins et al., 2001). guarantees don't scale

Overview

Overview of my work on PCA

Concrete problem: exponential family PCA

- Motivating example: XFEL
- The *e*PCA method
- XFEL illustration

Other work

Optimal hypothesis tests for PCs

- How to test for the presence of significant PCs?
- Classical statistics: use top eigenvalue based test (Anderson, 1960's)
- Big data: bias in eigenvalues engulfs signal (BBP'05)
- For special model, LRT is different (Onatski et al., 2013, 2014)
- Optimal tests for general local alternatives model

Local alternatives model

- Recall $X = Z_{n\times p}\Sigma^{1/2}$ and $H_p = \frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i(\Sigma)}$
- Local alternatives model:

$$H_{p,0} : H_p = \left(1 - \frac{h}{p}\right) H + \frac{h}{p} G_0, \text{ vs}$$

$$H_{p,1} : H_p = \left(1 - \frac{h}{p}\right) H + \frac{h}{p} G_1.$$

Optimal tests in local alternatives model

- Given (H, h, G_0, G_1, γ) we derive a linear spectral statistic $T = \mathrm{tr}\{\varphi(\widehat{\Sigma})\}$.
- Gives the asymptotically best test for $H_{p,0}$ against $H_{p,1}$

Mean-variance problem

- There are mean and variance parameters $\mu_\varphi, \sigma_\varphi^2$ s.t. for some c_p
  - Under $H_{p,0}$, $\mathrm{tr}(\varphi(\widehat{\Sigma})) - c_p \Rightarrow \mathcal{N}(0, \sigma_\varphi^2)$
  - Under $H_{p,1}$, $\mathrm{tr}(\varphi(\widehat{\Sigma})) - c_p \Rightarrow \mathcal{N}(\mu_\varphi, \sigma_\varphi^2)$.
- With $\langle f, g \rangle = \int_{\mathcal{I}} f(x) g(x) dx$

$$\mu_\varphi = -h\langle \varphi', \Delta \rangle \quad \text{and}$$
$$\sigma_\varphi^2 = \langle \varphi', K\varphi' \rangle.$$

- Find optimal LSS φ, maximizing the efficacy

$$\max_\varphi \frac{\mu_\varphi}{\sigma_\varphi}$$

Main result: Finding the optimal LSS

Theorem (Dobriban (2016b))

Two cases for testing (H, G_0) *vs* (H, G_1) *in the local alternatives model:*

1. *If* $\Delta \in \mathrm{Im}(K)$*, the optimal linear spectral statistics* φ *are given by a Fredholm integral equation:*

$$K(\varphi') = -\eta\Delta,$$

   where $\eta > 0$ *is any constant.*

2. *On the other hand, if* $\Delta \notin \mathrm{Im}(K)$*, then the maximal efficacy is* $+\infty$*.*

Computing the optimal LSS

- Solve $Kg = -\Delta$, i.e., $\int k(x,y)g(y)dy = -\Delta(x)$,

$$k(x,y) = \frac{1}{2\pi^2} \log\left(1 + 4\frac{\Im(\underline{s}(x))\,\Im(\underline{s}(y))}{|\underline{s}(x) - \underline{s}(y)|^2}\right)$$

  and $\underline{s}(x)$ is Stieltjes transform of $(1-\gamma)F_{\gamma,H} + \gamma\delta_0$
- Use SPECTRODE to compute $\underline{s}(x)$, k, Δ. Discretize to linear equation

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley New York.

Bai, Z. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *The Annals of Probability*, 32(1A):553–605.

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

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

Basri, R. and Jacobs, D. W. (2003). Lambertian Reflectance and Linear Subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233.

Collins, M., Dasgupta, S., and Schapire, R. (2001). A generalization of principal component analysis to the exponential family. *Nips*, (1).

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

Dobriban, E. (2015b). *EigenEdge: Computing with Eigenvalue Distributions of Large Random Matrices of the Covariance Type*. Matlab package.

Dobriban, E. (2016a). A general convex framework for multiple testing with prior information. *arXiv preprint arXiv:1603.05334*.

Dobriban, E. (2016b). Sharp detection in PCA under correlations: all eigenvalues matter. *to appear in The Annals of Statistics*.

Dobriban, E. and Fortney, K. (2015). *pweight: P-Value Weighting*. R package version 0.0.1.

Dobriban, E., Leeb, W., and Singer, A. (2016). PCA from noisy, linearly reduced data: the diagonal case. *arXiv preprint arXiv:1611.10333*.

Dobriban, E. and Wager, S. (2015). High-dimensional asymptotics of prediction: Ridge

regression and classification. *arXiv preprint arXiv:1507.03003, under minor revision at The Annals of Statistics*.

Donoho, D., Gavish, M., and Johnstone, I. (2013). Optimal shrinkage of eigenvalues in the Spiked Covariance Model. *arXiv preprint arXiv:1311.0851*, 0906812:1–35.

Favre-Nicolin, V., Baruchel, J., Renevier, H., Eymery, J., and Borbély, A. (2015). XTOP: high-resolution X-ray diffraction and imaging. *Journal of Applied Crystallography*, 48(3):620–620.

Fortney, K., Dobriban, E., Garagnani, P., Pirazzini, C., Monti, D., Mari, D., Atzmon, G., Barzilai, N., Franceschi, C., Owen, A. B., et al. (2015). Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLoS Genet*, 11(12):e1005728.

Hantke, M. F., Ekeberg, T., and Maia, F. R. N. C. (2016). Condor: A simulation tool for flash x-ray imaging. *Journal of Applied Crystallography*, 49(4):1356–1362.

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

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

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

Knott, M. and Bartholomew, D. J. (1999). *Latent variable models and factor analysis*. Edward Arnold.

Lee, S., Zou, F., and Wright, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Annals of Statistics*, 38(6):3605–3629.

Liu, L. T., Dobriban, E., and Singer, A. (2016). *e*PCA: High dimensional exponential family PCA. *arXiv preprint arXiv:1611.05550*.

Maia, F. R. and Hajdu, J. (2016). The trickle before the torrentdiffraction data from X-ray lasers. *Scientific Data*, 3.

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

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

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

Patterson, N., Price, A., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190.

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

Starck, J.-L., Murtagh, F., and Fadili, J. M. (2010). *Sparse image and signal processing: wavelets, curvelets, morphological diversity*. Cambridge university press.

Udell, M., Horn, C., Zadeh, R., and Boyd, S. (2016). Generalized Low Rank Models. *Foundations and Trends in Machine Learning*, 9(1):1–118.

*e*PCA application: Denoising

- Predict X using Y
- Best Linear Predictor (BLP): $\tilde{\mathbb{E}}(X|Y) = BY + C$

$$B = \Sigma_x \left[\mathbb{E}D + \Sigma_x\right]^{-1}$$
$$C = \mathbb{E}D \left[\mathbb{E}D + \Sigma_x\right]^{-1} \mathbb{E}Y.$$

- Empirical BLP (EBLP) via *e*PCA:

$$\hat{X}_i = S_s \left[D_n + S_s\right]^{-1} Y_i + D_n \left[D_n + S_s\right]^{-1} \bar{Y}.$$

Optimal LSS example: AR-1

- population covariance matrix

$$\Sigma = \begin{bmatrix} t & 0^\top \\ 0 & M \end{bmatrix}$$

- Spike t
- $M_{ij} = \rho^{|i-j|}$
- $H = spec(M)$
- Test "pure AR" vs "spiked AR"

$$H_{p,0} : H_p = \left(1 - \frac{1}{p}\right) H + \frac{1}{p}\delta_1, \text{ vs}$$
$$H_{p,1} : H_p = \left(1 - \frac{1}{p}\right) H + \frac{1}{p}\delta_t.$$

Optimal LSS example: AR-1

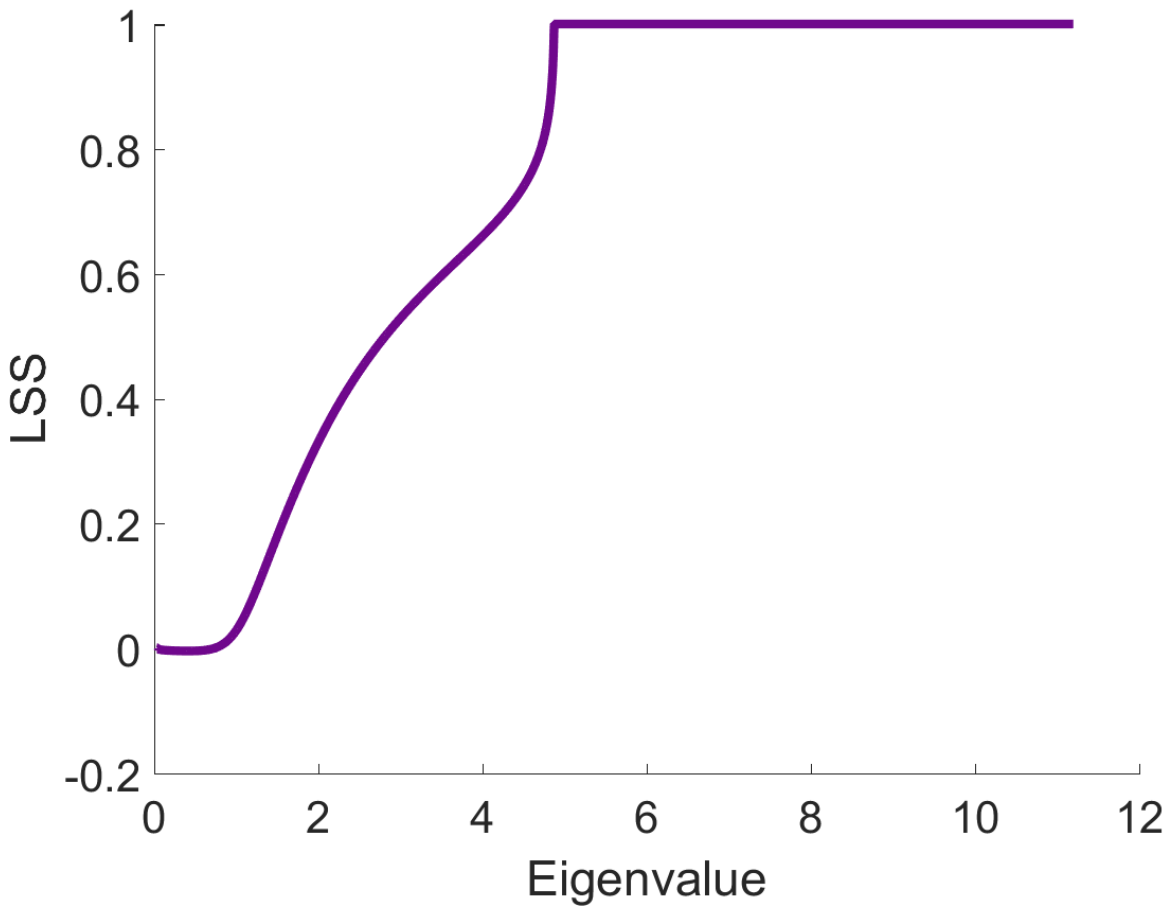


Figure: New optimal LSS $\varphi(x)$ in AR-1 model: $\gamma = 0.5, \rho = 0.5,\ t = 3.5$ below PT.

Probability background - CLT for LSS

Theorem. [Bai and Silverstein (2004) CLT]: Let $X = Z_{n\times p}\Sigma^{1/2}$ and Z_{ij} iid real standardized with $\mathbb{E}Z_{ij}^4 = 3$. If $H_p \Rightarrow H$, for φ analytic on a compact interval $\mathcal{I}$ including all supports of ESDs, we have

$$\mathrm{tr}(\varphi(\widehat{\Sigma})) - p\int_{\mathcal{I}} \varphi(x) dF_{\gamma,H_p}(x) \Rightarrow \mathcal{N}(m_\varphi, \sigma_\varphi^2)$$

- $\sigma_\varphi^2 = \int_{\mathcal{I}\times\mathcal{I}} \varphi'(x)\varphi'(y)k(x,y)dx\,dy = \langle \varphi', K\varphi'\rangle$, where k is a covariance kernel, and K is the associated operator
- $\underline{s}(x)$ is the limit Stieltjes transform of $(1-\gamma)F_{\gamma,H} + \gamma\delta_0$ as $z \to x \in \mathbb{R}$

$$k(x,y) = \frac{1}{2\pi^2}\log\left(1 + 4\frac{\Im(\underline{s}(x))\,\Im(\underline{s}(y))}{|\underline{s}(x) - \underline{s}(y)|^2}\right)$$

Possible future work

- PCA
  - *e*PCA: Spiked model universality
  - *e*PCA for scRNA-seq (Y. Kluger)
  - Shrinkage of PC scores (N. Patterson)
  - Inference in fast PCA (N. Patterson)
  - Selecting number of factors by permutation (M. MacKay)
- XFEL: Molecular reconstruction (A. Singer)
- replicable P-value weighting (A. Owen)
- more general Spectrode: space-time, heavy-tailed, MANOVA,...
- "Poisson regularized" Stein's covariance estimator
- open to other problems...