Optimal detection of principal components in high-dimensional data

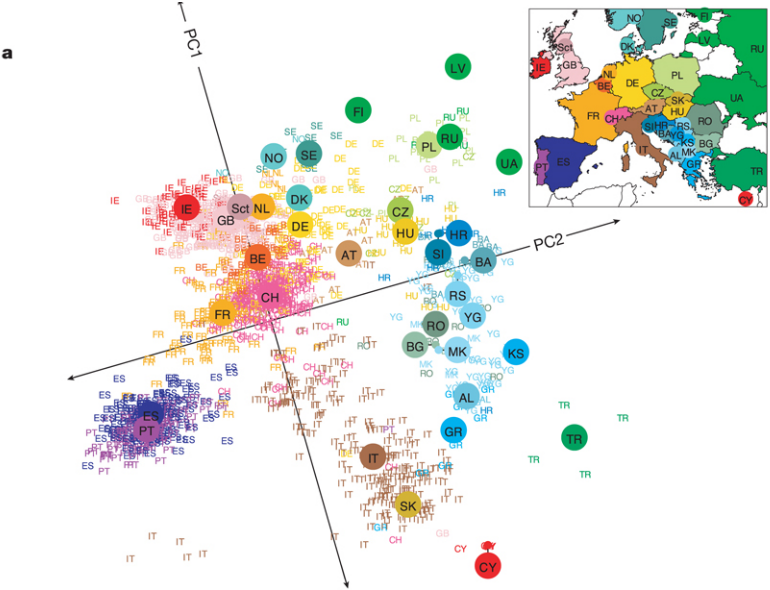
Edgar Dobriban

Statistics, Stanford

PCA

- Principal component analysis (PCA) is a widely used method for dimension reduction
- X an $n \times p$ matrix.
  - n samples from centered p-dimensional population
  - n individuals, p features: genetic markers, phenotypes (age, height...)
- PCs: linear combinations Xu_i of features that explain a lot of variance
- u_i - eigenvectors of sample covariance matrix $\widehat{\Sigma} = n^{-1}X^\top X$
- corresponding eigenvalue λ_i is variance of PC

Genes mirror geography within Europe – Novembre et al. (2008)



PCA in practice

- how to choose number of components?
- “scree plot” : eigenvalues in decreasing order
- look for the elbow - separated eigenvalue
- in high-dimension, this may miss “weak” PCs

PCA Review: bulk eigenvalues

- $X = Z_{n\times p}\Sigma^{1/2}$
  - $Z_{n\times p}$ has iid standardized entries
  - Σ unobserved $p \times p$ covariance matrix: $\mathrm{Cov}\left[x_i, x_i\right] = \Sigma$
- high dimension: $n, p \to \infty$, $p/n \to \gamma > 0$ (wlog $p/n = \gamma$)
- $l_1 \geq l_2 \geq \ldots \geq l_p$ eigenvalues of Σ. Distribution $H_p = p^{-1}\sum_i \delta_{l_i}$
- $H_p \Rightarrow H$
- sample eigenvalues λ_i of $\widehat{\Sigma} = n^{-1}X^\top X$ not consistent estimates of their population counterparts l_i: $\lambda_i \nrightarrow l_i$
- “bulk” distribution of λ_i converges to a different limit: the Marchenko-Pastur map $F_{\gamma,H}$ (Marchenko and Pastur, 1967)

MP map example $H \to F_{\gamma,H}$: white noise $\Sigma = I_p$

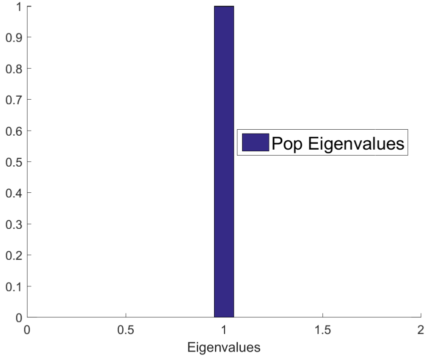


Figure : Eigenvalues $H = \delta_1$ of an identity covariance matrix $\Sigma = I_p$.

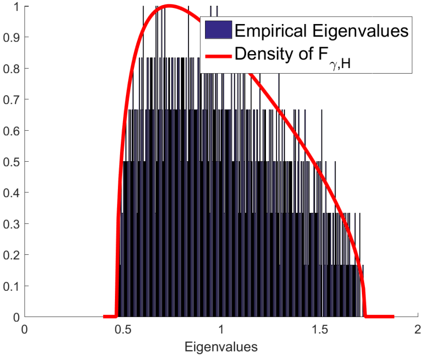


Figure : Marchenko-Pastur density: $g(x) = \sqrt{(\gamma_+ - x)(x - \gamma_-)}/\ (2\pi\gamma x)$, $x \in [\gamma_-, \gamma_+]$, $\gamma_\pm = (1 \pm \sqrt{\gamma})^2$, $\gamma = 1/2$.

MP map example: Autoregressive model, order 1, (AR-1)

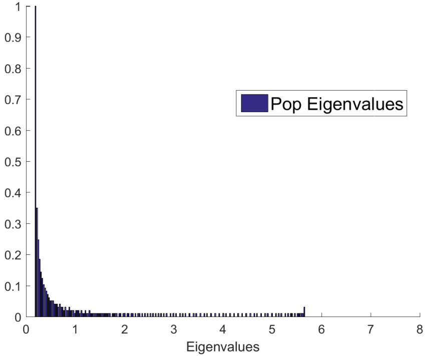


Figure : Eigenvalues H of an AR-1 covariance matrix Σ with $\Sigma_{ij} = \rho^{|i-j|}$ ($p = 400$; $\rho = 0.7$).

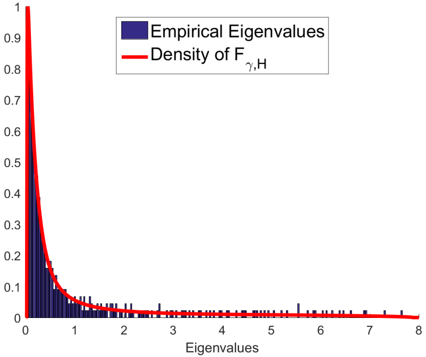


Figure : Eigenvalues of a sample covariance matrix $\widehat{\Sigma}$ with $n = 800$ samples.

PCA Review: spiked covariance model with one spike

- (noise) distribution of $l_2, ..., l_p$ converges to H
- (signal) spike $t = l_1$ fixed
- top sample eigenvalue λ_1 "pushed upward" from l_1, converges a.s. to deterministic limit
- BBP phase transition (Baik et al., 2005; Benaych-Georges and Nadakuditi, 2011):
  - above the phase transition (PT): if l_1 large enough, λ_1 separates from the MP map "bulk"
  - below the PT: else λ_1 does not separate

Spiked model: AR-1 example

- population covariance matrix

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

- spike t
- M is a $p \times p$ AR(1) covariance matrix $M_{ij} = \rho^{|i-j|}$. $\rho = 0.5$
- sample $n = 500$ Gaussian variates of $p = 250$ dimensions
- $mean(trace(M)) = 1$
- null: $t = 1$, alternative: larger t

AR-1 Example - Above PT

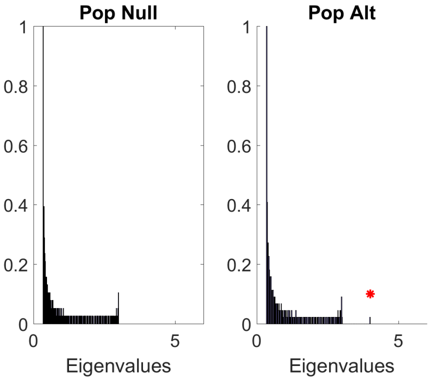


Figure : Eigenvalues of Σ. Null: $t = 1$. Alternative: $t = 4$.

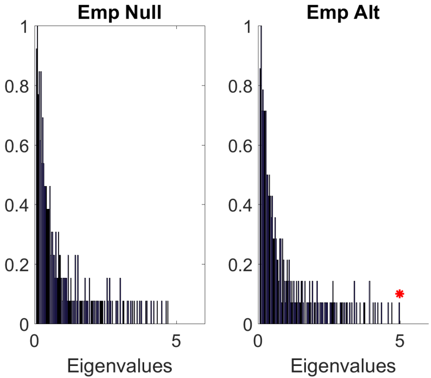


Figure : Eigenvalues of $\widehat{\Sigma}$. Null and alternative.

AR-1 Example - Below PT

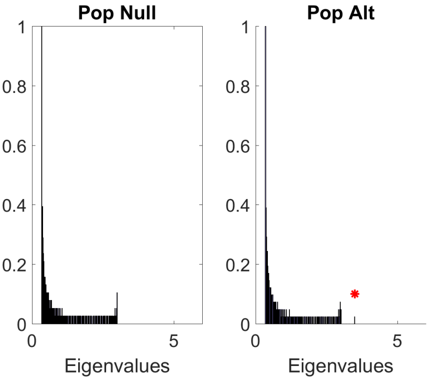


Figure : Eigenvalues of Σ. Null: $t = 1$. Alternative: $t = 3.5$.

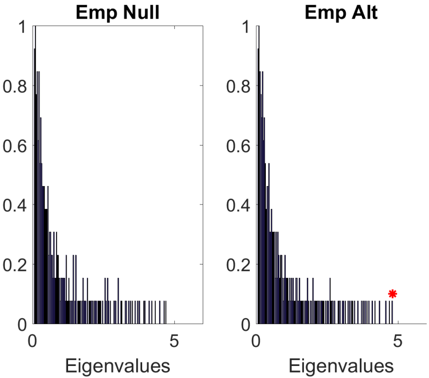


Figure : Eigenvalues of $\widehat{\Sigma}$. Null and alternative.

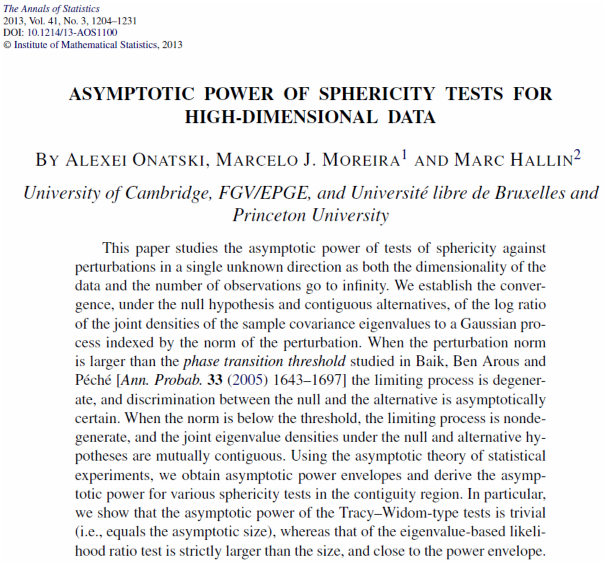
Statistical implications

- Below PT, top eigenvalue test based on λ_1 has trivial power
- Can we detect PCs below the phase transition?
- Onatski et al. (2013, 2014) (OMH) consider the real Gaussian standard spiked model of Johnstone (2001)

$$H_0 : \Sigma_p = I_p, \text{ vs}$$

$$H_1 : \Sigma_p = I_p + \sum_{j=1}^{r} (l_j - 1) v_j v_j^\top, v_j \text{ unknown orthonormal}$$

Onatski, Moreira, Hallin (2013)



The Annals of Statistics
2013, Vol. 41, No. 3, 1204–1231
DOI: 10.1214/13-AOS1100

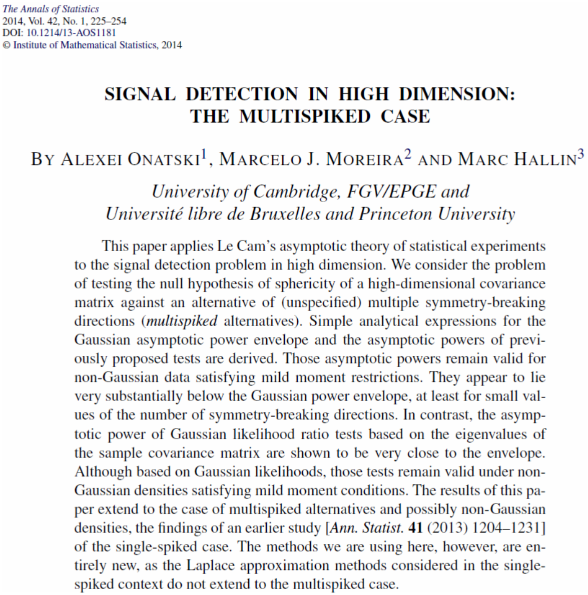
ASYMPTOTIC POWER OF SPHERICITY TESTS FOR HIGH-DIMENSIONAL DATA

BY ALEXEI ONATSKI, MARCELO J. MOREIRA[1] AND MARC HALLIN[2]

University of Cambridge, FGV/EPGE, and Université libre de Bruxelles and Princeton University

This paper studies the asymptotic power of tests of sphericity against perturbations in a single unknown direction as both the dimensionality of the data and the number of observations go to infinity. We establish the convergence, under the null hypothesis and contiguous alternatives, of the log ratio of the joint densities of the sample covariance eigenvalues to a Gaussian process indexed by the norm of the perturbation. When the perturbation norm is larger than the *phase transition threshold* studied in Baik, Ben Arous and Péché [*Ann. Probab.* **33** (2005) 1643–1697] the limiting process is degenerate, and discrimination between the null and the alternative is asymptotically certain. When the norm is below the threshold, the limiting process is nondegenerate, and the joint eigenvalue densities under the null and alternative hypotheses are mutually contiguous. Using the asymptotic theory of statistical experiments, we obtain asymptotic power envelopes and derive the asymptotic power for various sphericity tests in the contiguity region. In particular, we show that the asymptotic power of the Tracy–Widom-type tests is trivial (i.e., equals the asymptotic size), whereas that of the eigenvalue-based likelihood ratio test is strictly larger than the size, and close to the power envelope.

Onatski, Moreira, Hallin (2014)



The Annals of Statistics
2014, Vol. 42, No. 1, 225–254
DOI: 10.1214/13-AOS1181

SIGNAL DETECTION IN HIGH DIMENSION: THE MULTISPIKED CASE

BY ALEXEI ONATSKI[1], MARCELO J. MOREIRA[2] AND MARC HALLIN[3]

University of Cambridge, FGV/EPGE and
Université libre de Bruxelles and Princeton University

This paper applies Le Cam's asymptotic theory of statistical experiments to the signal detection problem in high dimension. We consider the problem of testing the null hypothesis of sphericity of a high-dimensional covariance matrix against an alternative of (unspecified) multiple symmetry-breaking directions (*multispiked* alternatives). Simple analytical expressions for the Gaussian asymptotic power envelope and the asymptotic powers of previously proposed tests are derived. Those asymptotic powers remain valid for non-Gaussian data satisfying mild moment restrictions. They appear to lie very substantially below the Gaussian power envelope, at least for small values of the number of symmetry-breaking directions. In contrast, the asymptotic power of Gaussian likelihood ratio tests based on the eigenvalues of the sample covariance matrix are shown to be very close to the envelope. Although based on Gaussian likelihoods, those tests remain valid under non-Gaussian densities satisfying mild moment conditions. The results of this paper extend to the case of multispiked alternatives and possibly non-Gaussian densities, the findings of an earlier study [*Ann. Statist.* **41** (2013) 1204–1231] of the single-spiked case. The methods we are using here, however, are entirely new, as the Laplace approximation methods considered in the single-spiked context do not extend to the multispiked case.

Results of OMH

- log-likelihood ratio test (LRT):

$$L_{n,p}(l_1, \ldots, l_r; \lambda_1, \ldots, \lambda_p) = \log \left[\frac{p_{n,p}(\lambda_1, \ldots, \lambda_p; l_1, \ldots, l_r)}{p_{n,p}(\lambda_1, \ldots, \lambda_p; 1, \ldots, 1)} \right]$$

- for $r = 1$, $L_{n,p}$ for $H_0 : l_1 = 1$ vs $H_1 : l_1 = t$ is equivalent to a linear spectral statistic (LSS)

$$L_{n,p}(t; \lambda_1, \ldots, \lambda_p) = \mathrm{tr}(\varphi(\widehat{\Sigma})) + c_p + o_P(1)$$

- φ explicit
- using CLT for LSS, find the optimal detection power achievable by any test

Results

Supplement

Our results

- OMH is one very specific example where we can calculate optimal tests
- In this talk, we find optimal tests for local alternatives generally
- Our construction is mathematically precise and clean. It can derive OMH results plus much else
- Before this work, we had very little information about optimal tests. Now we have a great deal

Local alternatives model

- Local alternatives model: bulk H perturbed by spikes G_0 vs G_1:

$$H_{p,0}: H_p = \left(1-\frac{h}{p}\right) H + \frac{h}{p} G_0, \text{ vs}$$
$$H_{p,1}: H_p = \left(1-\frac{h}{p}\right) H + \frac{h}{p} G_1.$$

- e.g., standard spiked model: $H = G_0 = \delta_1,\ G_1 = \delta_t$
- allows correlations, flexible nonparametric data modelling

Optimal tests in local alternatives model

- Given (H, h, G_0, G_1, γ) we derive a function φ, the score function of 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 constants c_p
    - under $H_{p,0}$, $\operatorname{tr}(\varphi(\widehat{\Sigma})) - c_p \Rightarrow \mathcal{N}(0, \sigma_\varphi^2)$
    - under $H_{p,1}$, $\operatorname{tr}(\varphi(\widehat{\Sigma})) - c_p \Rightarrow \mathcal{N}(\mu_\varphi, \sigma_\varphi^2)$.
- 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 (D.,2016)

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

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

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

   where $\eta > 0$ is any constant.

2. *On the other hand, if $\Delta \notin \mathrm{Im}(K)$, then the maximal efficacy is $+\infty$. If moreover $\Delta \notin \overline{\mathrm{Im}}(K)$, the optimal LSS are all functions φ with $K(\varphi') = 0$ and $\langle \varphi', \Delta \rangle < 0$.*

Recovering the OMH LSS

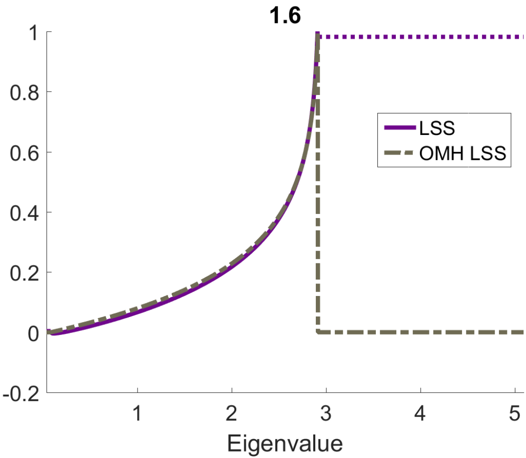


Figure : The optimal LSS and the OMH LSS in the standard spiked model. $H = G_0 = \delta_1,\ G_1 = \delta_t$. $t = 1.6$ and $\gamma = 1/2$.

New 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

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

Optimal LSS example: AR-1

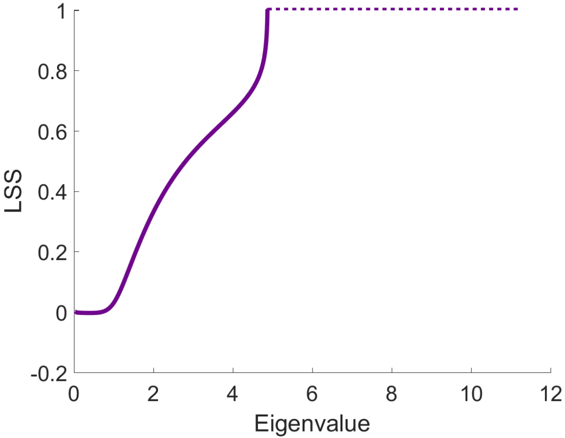


Figure : Our new optimal LSS $\varphi(x)$ in AR-1 example with $\gamma = 0.5, \rho = 0.5,\ t = 3.5$ below PT.

Example: detection power in AR-1

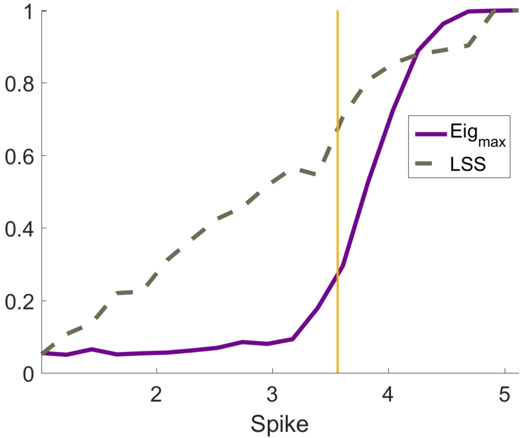


Figure : Detection power of LSS and top-eigenvalue test as a function of spike t. Vertical line: Location of PT. $\gamma = 0.5, \rho = 0.5, n = 500$

Computation

- need to solve the optimal LSS equation $K(\varphi') = -\eta\Delta$, where K is integral operator induced by kernel $k(x, y)$
- need efficient algorithm for computing $k(x, y)$
- depends on Marchenko-Pastur forward map $F_{\gamma,H}$ (Bai and Silverstein, 2004)
- Key problem: How to compute the Marchenko-Pastur forward map?
- surprisingly, not well studied.

Spectrode computes of MP map $F_{\gamma,H}$

- in Dobriban (2015) developed a fast ODE-based method Spectrode

Spectrode: Input and Output
Input:
$H \leftarrow$ population spectrum
$\gamma \leftarrow$ aspect ratio
Output:
$\hat{f}(x) \leftarrow$ density of MP map $F_{\gamma,H}$
$\hat{l}_k, \hat{u}_k \leftarrow$ endpoints of intervals in the support

SPECTRODE: Autoregressive model

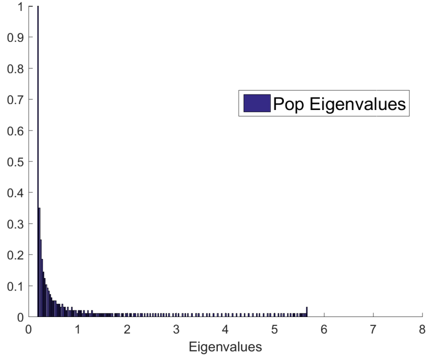


Figure : Eigenvalues H of an AR-1 covariance matrix Σ with $\Sigma_{ij} = \rho^{|i-j|}$ ($p = 400$; $\rho = 0.7$).

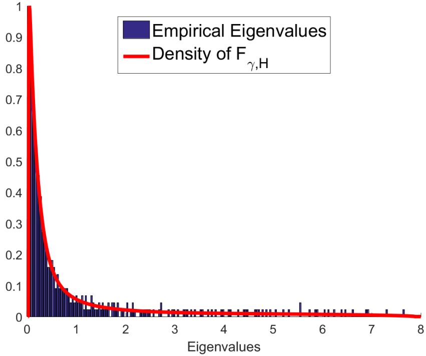


Figure : Eigenvalues of a sample covariance matrix $\widehat{\Sigma}$ with $n = 800$ samples. Density computed with SPECTRODE

Spectrode: "Comb" model

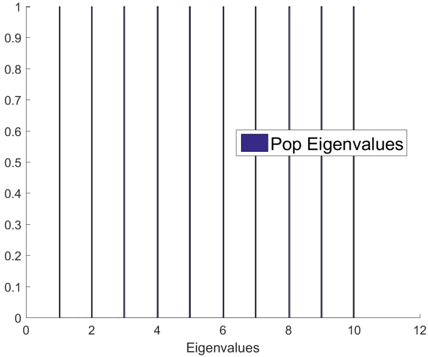


Figure : "Comb" model
$H = 10^{-1}\sum_{i=1}^{10}\delta_{l_i}$, with $l_i = i$.

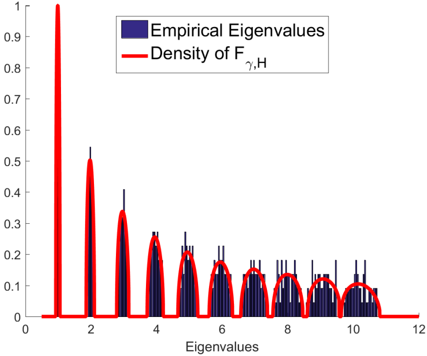


Figure : Eigenvalues of $\widehat{\Sigma}$ with $n = 800$ samples, and density. Density computed with Spectrode

Spectrode: a "universal" MP calculator

- Useful for a variety of problems, see Dobriban (2015):
  - examples of limit spectra, teaching
  - quantiles, moments and contour integrals of the MP map
  - principal component analysis (here)
  - covariance matrix estimation
  - bootstrap
- fast and accurate
- Matlab and R software at github.com/dobriban
  - with documentation and examples

Summary

- Optimal testing for principal components in high dimensions
  - Go below the phase transition using linear spectral statistics
- Enabled by SPECTRODE—new computational tool for RMT
- Thanks
  - Support: NSF, HHMI
  - Discussion: David Donoho, Iain Johnstone

Results

Supplement

Idea behind Spectrode

1. Stieltjes transform $x \to \underline{s}(x) = \mathbb{E}\frac{1}{\lambda - x}$ increasing for $x \in \mathbb{R}$ outside of the support of $F_{\gamma,H}$.
   - ST has increasing inverse there (Silverstein and Choi, 1995)
2. MP/Silverstein equation defines inverse ST, for $z \in \mathbb{C}^+$

$$z = -\frac{1}{\underline{s}(z)} + \gamma \int \frac{t}{1 + t\underline{s}(z)} dH(t).$$

   - Use this for $z = x + i\varepsilon$, small ε, to find the edges of the support
3. Run ODE derived from Silv eq to find smooth density within support
   - Starting point using fixed-pont algoritm

Proof idea

Goal: get an expression for the mean

- Start with the CLT of Bai & Silverstein (2004)
- In the local alternatives model:

$$\text{under } H_0 : \operatorname{tr}(\varphi(\widehat{\Sigma})) - p\int_{\mathcal{I}} \varphi(x)dF_{\gamma,H_{p,0}} \Rightarrow \mathcal{N}(m_\varphi, \sigma_\varphi^2), \text{ while}$$
$$\text{under } H_1 : \operatorname{tr}(\varphi(\widehat{\Sigma})) - p\int_{\mathcal{I}} \varphi(x)dF_{\gamma,H_{p,1}} \Rightarrow \mathcal{N}(m_\varphi, \sigma_\varphi^2).$$

- In the limit, test $\mathcal{N}(0, \sigma_\varphi^2)$ vs $\mathcal{N}(\mu_\varphi, \mu_\varphi^2)$, where

$$\mu_\varphi = \lim_{p\to\infty} \int_{\mathcal{I}} \varphi(x)d\left[p(F_{\gamma,H_{p,1}} - F_{\gamma,H_{p,0}})\right],$$

Key new object: Weak derivative of MP map

- The weak derivative of MP map F_γ is the signed measure

$$\delta\mathcal{F}_\gamma(H, G) = \lim_{\varepsilon \to 0} \frac{F_{\gamma,(1-\varepsilon)H+\varepsilon G} - F_{\gamma,H}}{\varepsilon}.$$

- so $p[F_{\gamma,H_{p,1}} - F_{\gamma,H_{p,0}}] \Rightarrow h \cdot \Delta$, where $\Delta = \delta\mathcal{F}_\gamma(H, G_1) - \delta\mathcal{F}_\gamma(H, G_0)$
- integrate by parts

$$\mu_\varphi = h \cdot \langle \varphi, d\Delta \rangle = -h \cdot \langle \varphi', \Delta \rangle.$$

Key new object: Weak derivative of MP map

- nice properties:
  - has a density within support of F_γ
  - spike is above PT iff isolated point mass in $\delta\mathcal{F}_\gamma$ — a new perspective on phase transitions in spiked models

Weak derivative: Example

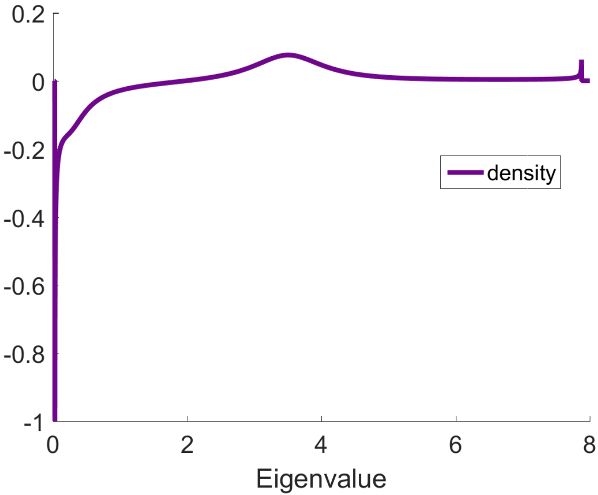


Figure : Density of weak derivative $\delta\mathcal{F}_{\gamma}(H, G)$ in AR-1 example $H = spec(\Sigma)$, $G = \delta_{3.5}$ below PT;

References I

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

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.

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.

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

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

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

J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

References II

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

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

J. W. Silverstein and S.-I. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *J. Multivariate Anal.*, 54(2):295–309, 1995.