Random Matrices for Big Data Signal Processing and Machine Learning

(ICASSP'2017, New Orleans)

Romain COUILLET and Hafiz TIOMOKO ALI

CentraleSupélec, France

March, 2017

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

Context

Baseline scenario: $x_1, \ldots, x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) i.i.d. with $E[x_1] = 0$, $E[x_1 x_1^*] = C_N$:

Context

Baseline scenario: $x_1, \ldots, x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) i.i.d. with $E[x_1] = 0$, $E[x_1 x_1^*] = C_N$:

- If $x_1 \sim \mathcal{N}(0, C_N)$, ML estimator for C_N is the sample covariance matrix (SCM)

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^* .$$

Context

Baseline scenario: $x_1, \ldots, x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) i.i.d. with $E[x_1] = 0$, $E[x_1 x_1^*] = C_N$:

- If $x_1 \sim \mathcal{N}(0, C_N)$, ML estimator for C_N is the sample covariance matrix (SCM)

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^*.$$

- If $n \to \infty$, then, strong law of large numbers

$$\hat{C}_N \xrightarrow{\text{a.s.}} C_N.$$

  or equivalently, in spectral norm

$$\left\| \hat{C}_N - C_N \right\| \xrightarrow{\text{a.s.}} 0.$$

Context

Baseline scenario: $x_1, \ldots, x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) i.i.d. with $E[x_1] = 0$, $E[x_1 x_1^*] = C_N$:

- If $x_1 \sim \mathcal{N}(0, C_N)$, ML estimator for C_N is the sample covariance matrix (SCM)

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^*.$$

- If $n \to \infty$, then, strong law of large numbers

$$\hat{C}_N \xrightarrow{\text{a.s.}} C_N.$$

  or equivalently, in spectral norm

$$\left\| \hat{C}_N - C_N \right\| \xrightarrow{\text{a.s.}} 0.$$

Random Matrix Regime

- No longer valid if $N, n \to \infty$ with $N/n \to c \in (0, \infty)$,

$$\left\| \hat{C}_N - C_N \right\| \not\to 0.$$

Context

Baseline scenario: $x_1,\dots,x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) i.i.d. with $E[x_1]=0$, $E[x_1x_1^*]=C_N$:

- If $x_1 \sim \mathcal{N}(0,C_N)$, ML estimator for C_N is the sample covariance matrix (SCM)

$$\hat{C}_N = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^*.$$

- If $n \to \infty$, then, strong law of large numbers

$$\hat{C}_N \xrightarrow{\text{a.s.}} C_N.$$

or equivalently, in spectral norm

$$\left\|\hat{C}_N - C_N\right\| \xrightarrow{\text{a.s.}} 0.$$

Random Matrix Regime

- No longer valid if $N,n\to\infty$ with $N/n \to c \in (0,\infty)$,

$$\left\|\hat{C}_N - C_N\right\| \not\to 0.$$

- For practical N,n with $N \simeq n$, leads to dramatically wrong conclusions

Context

Baseline scenario: $x_1, \ldots, x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) i.i.d. with $E[x_1] = 0$, $E[x_1 x_1^*] = C_N$:

- If $x_1 \sim \mathcal{N}(0, C_N)$, ML estimator for C_N is the sample covariance matrix (SCM)

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^* .$$

- If $n \to \infty$, then, strong law of large numbers

$$\hat{C}_N \xrightarrow{\text{a.s.}} C_N .$$

  or equivalently, in spectral norm

$$\left\| \hat{C}_N - C_N \right\| \xrightarrow{\text{a.s.}} 0.$$

Random Matrix Regime

- No longer valid if $N, n \to \infty$ with $N/n \to c \in (0, \infty)$,

$$\left\| \hat{C}_N - C_N \right\| \not\to 0.$$

- For practical N, n with $N \simeq n$, leads to dramatically wrong conclusions
- Even for $N = n/100$.

The Large Dimensional Fallacies

Setting: $x_i \in \mathbb{C}^N$ i.i.d., $x_1 \sim \mathcal{CN}(0, I_N)$

The Large Dimensional Fallacies

Setting: $x_i \in \mathbb{C}^N$ i.i.d., $x_1 \sim \mathcal{CN}(0, I_N)$

- assume $N = N(n)$ such that $N/n \to c > 1$

The Large Dimensional Fallacies

Setting: $x_i \in \mathbb{C}^N$ i.i.d., $x_1 \sim \mathcal{CN}(0, I_N)$

- assume $N = N(n)$ such that $N/n \to c > 1$
- then, joint point-wise convergence

$$\max_{1 \leq i,j \leq N} \left| \left[\hat{C}_N - I_N \right]_{ij} \right| = \max_{1 \leq i,j \leq N} \left| \frac{1}{n} X_{j,\cdot} X_{i,\cdot}^* - \boldsymbol{\delta}_{ij} \right| \xrightarrow{\text{a.s.}} 0.$$

The Large Dimensional Fallacies

Setting: $x_i \in \mathbb{C}^N$ i.i.d., $x_1 \sim \mathcal{CN}(0, I_N)$

- assume $N = N(n)$ such that $N/n \to c > 1$
- then, joint point-wise convergence

$$\max_{1 \leq i,j \leq N} \left| \left[\hat{C}_N - I_N \right]_{ij} \right| = \max_{1 \leq i,j \leq N} \left| \frac{1}{n} X_{j,\cdot} X_{i,\cdot}^* - \boldsymbol{\delta}_{ij} \right| \xrightarrow{\text{a.s.}} 0.$$

- however, eigenvalue mismatch

$$0 = \lambda_1(\hat{C}_N) = \ldots = \lambda_{N-n}(\hat{C}_N) \leq \lambda_{N-n+1}(\hat{C}_N) \leq \ldots \leq \lambda_N(\hat{C}_N)$$
$$1 = \lambda_1(I_N) = \ldots = \lambda_{N-n}(I_N) = \lambda_{N-n+1}(\hat{C}_N) = \ldots = \lambda_N(I_N)$$

The Large Dimensional Fallacies

Setting: $x_i \in \mathbb{C}^N$ i.i.d., $x_1 \sim \mathcal{CN}(0, I_N)$

- assume $N = N(n)$ such that $N/n \to c > 1$
- then, joint point-wise convergence

$$\max_{1\leq i,j\leq N} \left| \left[\hat{C}_N - I_N \right]_{ij} \right| = \max_{1\leq i,j\leq N} \left| \frac{1}{n} X_{j,\cdot} X_{i,\cdot}^* - \boldsymbol{\delta}_{ij} \right| \xrightarrow{\text{a.s.}} 0.$$

- however, eigenvalue mismatch

$$\begin{aligned} 0 = \lambda_1(\hat{C}_N) = \ldots = \lambda_{N-n}(\hat{C}_N) \leq \lambda_{N-n+1}(\hat{C}_N) \leq \ldots \leq \lambda_N(\hat{C}_N) \\ 1 = \lambda_1(I_N) = \ldots = \lambda_{N-n}(I_N) = \lambda_{N-n+1}(\hat{C}_N) = \ldots = \lambda_N(I_N) \end{aligned}$$

$\Rightarrow$ no convergence in spectral norm.

The Marčenko–Pastur law

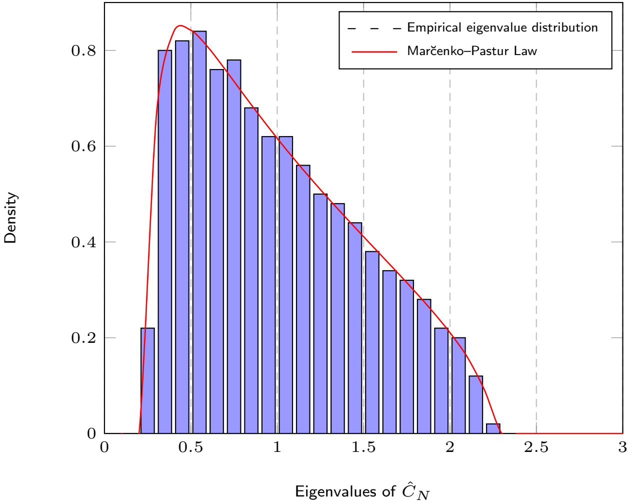


Figure: Histogram of the eigenvalues of $\hat{C}_N$ for $N = 500$, $n = 2000$, $C_N = I_N$.

The Marčenko–Pastur law

Definition (Empirical Spectral Density)

Empirical spectral density (e.s.d.) μ_N of Hermitian matrix $A_N \in \mathbb{C}^{N\times N}$ is

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\delta}_{\lambda_i(A_N)}.$$

The Marčenko–Pastur law

Definition (Empirical Spectral Density)

Empirical spectral density (e.s.d.) μ_N of Hermitian matrix $A_N \in \mathbb{C}^{N \times N}$ is

$$\mu_N = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\delta}_{\lambda_i(A_N)}.$$

Theorem (Marčenko–Pastur Law **[Marčenko,Pastur'67]**)

$X_N \in \mathbb{C}^{N \times n}$ with i.i.d. zero mean, unit variance entries.
As $N, n \to \infty$ with $N/n \to c \in (0, \infty)$, e.s.d. μ_N of $\frac{1}{n} X_N X_N^*$ satisfies

$$\mu_N \xrightarrow{\text{a.s.}} \mu_c$$

weakly, where

- $\mu_c(\{0\}) = \max\{0, 1 - c^{-1}\}$

The Marčenko–Pastur law

Definition (Empirical Spectral Density)

Empirical spectral density (e.s.d.) μ_N of Hermitian matrix $A_N \in \mathbb{C}^{N \times N}$ is

$$\mu_N = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\delta}_{\lambda_i(A_N)}.$$

Theorem (Marčenko–Pastur Law **[Marčenko,Pastur'67]**)

$X_N \in \mathbb{C}^{N \times n}$ with i.i.d. zero mean, unit variance entries.
As $N, n \to \infty$ with $N/n \to c \in (0, \infty)$, e.s.d. μ_N of $\frac{1}{n} X_N X_N^*$ satisfies

$$\mu_N \xrightarrow{\text{a.s.}} \mu_c$$

weakly, where

- $\mu_c(\{0\}) = \max\{0, 1 - c^{-1}\}$
- *on $(0, \infty)$, μ_c has continuous density f_c supported on $[(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]$*

$$f_c(x) = \frac{1}{2\pi c x} \sqrt{(x - (1 - \sqrt{c})^2)((1 + \sqrt{c})^2 - x)}.$$

The Marčenko–Pastur law

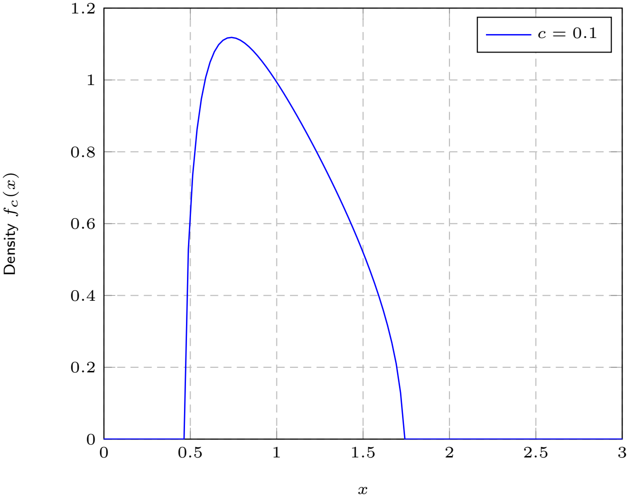


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{N\to\infty} N/n$.

The Marčenko–Pastur law

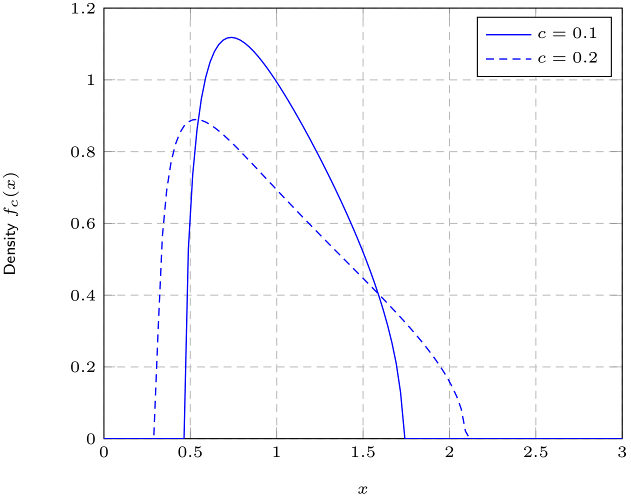


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{N\to\infty} N/n$.

The Marčenko–Pastur law

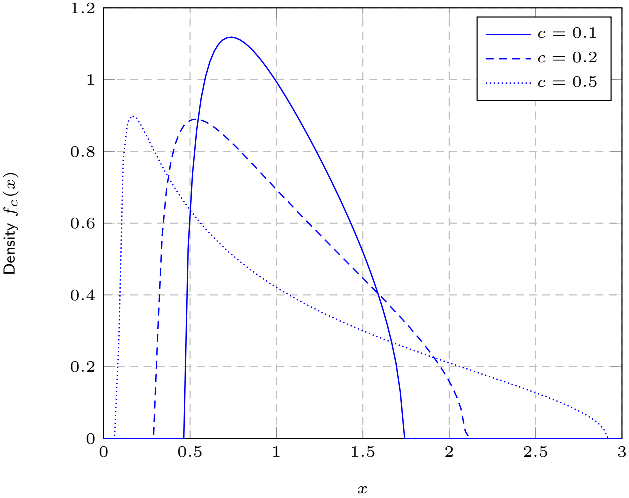


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{N \to \infty} N/n$.

Outline

- Basics of Random Matrix Theory
    - Motivation: Large Sample Covariance Matrices
    - The Stieltjes Transform Method
    - Spiked Models
    - Other Common Random Matrix Models
- Applications
    - Random Matrices and Robust Estimation
    - Spectral Clustering Methods and Random Matrices
    - Community Detection on Graphs
    - Kernel Spectral Clustering
    - Kernel Spectral Clustering: Subspace Clustering
    - Semi-supervised Learning
    - Support Vector Machines
    - Neural Networks: Extreme Learning Machines
- Perspectives

The Stieltjes transform

Definition (Stieltjes Transform)

For μ real probability measure of support $\operatorname{supp}(\mu)$, Stieltjes transform m_μ defined, for $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$, as

$$m_\mu(z) = \int \frac{1}{t-z} \mu(dt).$$

The Stieltjes transform

Definition (Stieltjes Transform)

For μ real probability measure of support $\operatorname{supp}(\mu)$, Stieltjes transform m_μ defined, for $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$, as

$$m_\mu(z) = \int \frac{1}{t-z} \mu(dt).$$

Property (Inverse Stieltjes Transform)

For $a < b$ continuity points of μ,

$$\mu([a,b]) = \lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \int_a^b \Im[m_\mu(x + \imath\varepsilon)]dx$$

The Stieltjes transform

Definition (Stieltjes Transform)

For μ real probability measure of support $\operatorname{supp}(\mu)$, Stieltjes transform m_μ defined, for $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$, as

$$m_\mu(z) = \int \frac{1}{t-z} \mu(dt).$$

Property (Inverse Stieltjes Transform)

For $a < b$ continuity points of μ,

$$\mu([a,b]) = \lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \int_a^b \Im[m_\mu(x + \imath\varepsilon)] dx$$

Besides, if μ has a density f at x,

$$f(x) = \lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \Im[m_\mu(x + \imath\varepsilon)].$$

The Stieltjes transform

Property (Relation to e.s.d.)

If μ e.s.d. of Hermitian $A \in \mathbb{C}^{N\times N}$, (i.e., $\mu = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\delta}_{\lambda_i(A)}$)

$$m_\mu(z) = \frac{1}{N}\operatorname{tr}\,(A - zI_N)^{-1}$$

The Stieltjes transform

Property (Relation to e.s.d.)

If μ e.s.d. of Hermitian $A \in \mathbb{C}^{N\times N}$, (i.e., $\mu = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\delta}_{\lambda_i(A)}$)

$$m_\mu(z) = \frac{1}{N}\mathrm{tr}\ (A - zI_N)^{-1}$$

Proof:

$$\begin{aligned} m_\mu(z) &= \int \frac{\mu(dt)}{t-z} = \frac{1}{N}\sum_{i=1}^N \frac{1}{\lambda_i(A)-z} = \frac{1}{N}\mathrm{tr}\ (\mathrm{diag}\{\lambda_i(A)\} - zI_N)^{-1} \\ &= \frac{1}{N}\mathrm{tr}\ (A - zI_N)^{-1}. \end{aligned}$$

The Stieltjes transform

Property (Stieltjes transform of Gram matrices)

For $X \in \mathbb{C}^{N \times n}$, and

- μ e.s.d. of XX^*
- $\tilde{\mu}$ e.s.d. of X^*X

Then

$$m_\mu(z) = \frac{n}{N} m_{\tilde{\mu}}(z) - \frac{N-n}{N} \frac{1}{z}.$$

The Stieltjes transform

Property (Stieltjes transform of Gram matrices)

For $X \in \mathbb{C}^{N\times n}$, and

- μ e.s.d. of XX^*
- $\tilde{\mu}$ e.s.d. of X^*X

Then

$$m_\mu(z) = \frac{n}{N} m_{\tilde{\mu}}(z) - \frac{N-n}{N}\frac{1}{z}.$$

Proof:

$$m_\mu(z) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\lambda_i(XX^*) - z} = \frac{1}{N}\sum_{i=1}^{n} \frac{1}{\lambda_i(X^*X) - z} + \frac{1}{N}(N-n)\frac{1}{0-z}.$$

The Stieltjes transform

Three fundamental lemmas in all proofs.

Lemma (Resolvent Identity)

For $A, B \in \mathbb{C}^{N\times N}$ invertible,

$$A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}.$$

The Stieltjes transform

Three fundamental lemmas in all proofs.

Lemma (Resolvent Identity)

For $A, B \in \mathbb{C}^{N \times N}$ invertible,

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}.$$

Corollary

For $t \in \mathbb{C}$, $x \in \mathbb{C}^N$, $A \in \mathbb{C}^{N \times N}$, with A and $A + txx^*$ invertible,

$$(A + txx^*)^{-1}x = \frac{A^{-1}x}{1 + tx^*A^{-1}x}.$$

The Stieltjes transform

Three fundamental lemmas in all proofs.

Lemma (Rank-one perturbation)

For $A, B \in \mathbb{C}^{N\times N}$ Hermitian nonnegative definite, e.s.d. μ of A, $t > 0$, $x \in \mathbb{C}^N$, $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$,

$$\left| \frac{1}{N} \operatorname{tr} B \left(A + txx^* - zI_N \right)^{-1} - \frac{1}{N} \operatorname{tr} B \left(A - zI_N \right)^{-1} \right| \leq \frac{1}{N} \frac{\|B\|}{\operatorname{dist}(z, \operatorname{supp}(\mu))}$$

The Stieltjes transform

Three fundamental lemmas in all proofs.

Lemma (Rank-one perturbation)

For $A, B \in \mathbb{C}^{N\times N}$ Hermitian nonnegative definite, e.s.d. μ of A, $t > 0$, $x \in \mathbb{C}^N$, $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$,

$$\left| \frac{1}{N} \operatorname{tr} B \left(A + txx^* - zI_N\right)^{-1} - \frac{1}{N} \operatorname{tr} B \left(A - zI_N\right)^{-1} \right| \leq \frac{1}{N} \frac{\|B\|}{\operatorname{dist}(z, \operatorname{supp}(\mu))}$$

In particular, as $N \to \infty$, if $\limsup_N \|B\| < \infty$,

$$\frac{1}{N} \operatorname{tr} B \left(A + txx^* - zI_N\right)^{-1} - \frac{1}{N} \operatorname{tr} B \left(A - zI_N\right)^{-1} \to 0.$$

The Stieltjes transform

Three fundamental lemmas in all proofs.

Lemma (Trace Lemma)

For

- $x \in \mathbb{C}^N$ with i.i.d. entries with zero mean, unit variance, finite $2p$ order moment,
- $A \in \mathbb{C}^{N \times N}$ deterministic (or independent of x),

then

$$E\left[\left|\frac{1}{N}x^* Ax - \frac{1}{N}\operatorname{tr} A\right|^p\right] \leq K\frac{\|A\|^p}{N^{p/2}}.$$

Three fundamental lemmas in all proofs.

Lemma (Trace Lemma)

For

- $x \in \mathbb{C}^N$ with i.i.d. entries with zero mean, unit variance, finite $2p$ order moment,
- $A \in \mathbb{C}^{N \times N}$ deterministic (or independent of x),

then

$$E\left[\left|\frac{1}{N}x^* Ax - \frac{1}{N}\operatorname{tr} A\right|^p\right] \leq K\frac{\|A\|^p}{N^{p/2}}.$$

In particular, if $\limsup_N \|A\| < \infty$, and x has entries with finite eighth-order moment,

$$\frac{1}{N}x^* Ax - \frac{1}{N}\operatorname{tr} A \xrightarrow{\text{a.s.}} 0$$

(by Markov inequality and Borel Cantelli lemma).

Proof of the Marčenko–Pastur law

Theorem (Marčenko–Pastur Law **[Marčenko,Pastur'67]**)

$X_N \in \mathbb{C}^{N \times n}$ with i.i.d. zero mean, unit variance entries.
As $N, n \to \infty$ with $N/n \to c \in (0, \infty)$, e.s.d. μ_N of $\frac{1}{n} X_N X_N^*$ satisfies

$$\mu_N \xrightarrow{\text{a.s.}} \mu_c$$

weakly, where

- $\mu_c(\{0\}) = \max\{0, 1 - c^{-1}\}$

Proof of the Marčenko–Pastur law

Theorem (Marčenko–Pastur Law **[Marčenko,Pastur'67]**)

$X_N \in \mathbb{C}^{N\times n}$ with i.i.d. zero mean, unit variance entries.
As $N, n \to \infty$ with $N/n \to c \in (0,\infty)$, e.s.d. μ_N of $\frac{1}{n}X_N X_N^*$ satisfies

$$\mu_N \xrightarrow{\text{a.s.}} \mu_c$$

weakly, where

- $\mu_c(\{0\}) = \max\{0, 1 - c^{-1}\}$
- *on $(0,\infty)$, μ_c has continuous density f_c supported on $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$*

$$f_c(x) = \frac{1}{2\pi c x}\sqrt{(x-(1-\sqrt{c})^2)((1+\sqrt{c})^2-x)}.$$

Proof of the Marčenko–Pastur law

Stieltjes transform approach.

Proof of the Marčenko–Pastur law

Stieltjes transform approach.

Proof

- With μ_N e.s.d. of $\frac{1}{n} X_N X_N^*$,

$$m_{\mu_N}(z) = \frac{1}{N} \operatorname{tr} \left(\frac{1}{n} X_N X_N^* - z I_N \right)^{-1} = \frac{1}{N} \sum_{i=1}^{N} \left[\left(\frac{1}{n} X_N X_N^* - z I_N \right)^{-1} \right]_{ii}.$$

Proof of the Marčenko–Pastur law

Stieltjes transform approach.

Proof

- With μ_N e.s.d. of $\frac{1}{n}X_NX_N^*$,

$$m_{\mu_N}(z) = \frac{1}{N}\mathrm{tr}\left(\frac{1}{n}X_NX_N^* - zI_N\right)^{-1} = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{1}{n}X_NX_N^* - zI_N\right)^{-1}\right]_{ii}.$$

- Write

$$X_N = \begin{bmatrix} y^* \\ Y_{N-1} \end{bmatrix} \in \mathbb{C}^{N\times n}$$

Proof of the Marčenko–Pastur law

Stieltjes transform approach.

Proof

- With μ_N e.s.d. of $\frac{1}{n}X_N X_N^*$,

$$m_{\mu_N}(z) = \frac{1}{N}\operatorname{tr}\left(\frac{1}{n}X_N X_N^* - zI_N\right)^{-1} = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{1}{n}X_N X_N^* - zI_N\right)^{-1}\right]_{ii}.$$

- Write

$$X_N = \begin{bmatrix} y^* \\ Y_{N-1} \end{bmatrix} \in \mathbb{C}^{N\times n}$$

so that, for $\Im[z] > 0$,

$$\left(\frac{1}{n}X_N X_N^* - zI_N\right)^{-1} = \begin{pmatrix} \frac{1}{n}y^*y - z & \frac{1}{n}y^*Y_{N-1} \\ \frac{1}{n}Y_{N-1}y & \frac{1}{n}Y_{N-1}Y_{N-1}^* - zI_{N-1} \end{pmatrix}^{-1}.$$

Proof of the Marčenko–Pastur law

Proof (continued)

- From block matrix inverse formula

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A-BD^{-1}C)^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1} \\ -(A-BD^{-1}C)^{-1}CA^{-1} & (D-CA^{-1}B)^{-1} \end{pmatrix}$$

we have

$$\left[\left(\frac{1}{n}X_N X_N^* - zI_N\right)^{-1}\right]_{11} = \frac{1}{-z - z\frac{1}{n}y^*(\frac{1}{n}Y_{N-1}^*Y_{N-1} - zI_n)^{-1}y}.$$

Proof of the Marčenko–Pastur law

Proof (continued)

- From block matrix inverse formula

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(A - BD^{-1}C)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}$$

  we have

$$\left[\left(\frac{1}{n}X_N X_N^* - zI_N\right)^{-1}\right]_{11} = \frac{1}{-z - z\frac{1}{n}y^*(\frac{1}{n}Y_{N-1}^* Y_{N-1} - zI_n)^{-1}y}.$$

- By Trace Lemma, as $N, n \to \infty$

$$\left[\left(\frac{1}{n}X_N X_N^* - zI_N\right)^{-1}\right]_{11} - \frac{1}{-z - z\frac{1}{n}\mathrm{tr}\,(\frac{1}{n}Y_{N-1}^* Y_{N-1} - zI_n)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

Proof of the Marčenko–Pastur law

Proof (continued)

- By Rank-1 Perturbation Lemma ($X_N^* X_N = Y_{N-1}^* Y_{N-1} + yy^*$), as $N, n \to \infty$

$$\left[\left(\frac{1}{n} X_N X_N^* - z I_N \right)^{-1} \right]_{11} - \frac{1}{-z - z \frac{1}{n} \operatorname{tr} \left(\frac{1}{n} X_N^* X_N - z I_n \right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

Proof of the Marčenko–Pastur law

Proof (continued)

- By Rank-1 Perturbation Lemma ($X_N^* X_N = Y_{N-1}^* Y_{N-1} + yy^*$), as $N, n \to \infty$

$$\left[\left(\frac{1}{n} X_N X_N^* - zI_N\right)^{-1}\right]_{11} - \frac{1}{-z - z\frac{1}{n}\operatorname{tr}\left(\frac{1}{n}X_N^* X_N - zI_n\right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

- Since $\frac{1}{n}\operatorname{tr}\left(\frac{1}{n}X_N^* X_N - zI_n\right)^{-1} = \frac{1}{n}\operatorname{tr}\left(\frac{1}{n}X_N X_N^* - zI_N\right)^{-1} - \frac{n-N}{n}\frac{1}{z}$,

$$\left[\left(\frac{1}{n} X_N X_N^* - zI_N\right)^{-1}\right]_{11} - \frac{1}{1 - \frac{N}{n} - z - z\frac{1}{n}\operatorname{tr}\left(\frac{1}{n}X_N X_N^* - zI_N\right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

Proof of the Marčenko–Pastur law

Proof (continued)

- By Rank-1 Perturbation Lemma ($X_N^* X_N = Y_{N-1}^* Y_{N-1} + yy^*$), as $N, n \to \infty$

$$\left[\left(\frac{1}{n} X_N X_N^* - z I_N\right)^{-1}\right]_{11} - \frac{1}{-z - z\frac{1}{n}\operatorname{tr}\left(\frac{1}{n} X_N^* X_N - z I_n\right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

- Since $\frac{1}{n}\operatorname{tr}\left(\frac{1}{n} X_N^* X_N - z I_n\right)^{-1} = \frac{1}{n}\operatorname{tr}\left(\frac{1}{n} X_N X_N^* - z I_N\right)^{-1} - \frac{n-N}{n}\frac{1}{z}$,

$$\left[\left(\frac{1}{n} X_N X_N^* - z I_N\right)^{-1}\right]_{11} - \frac{1}{1 - \frac{N}{n} - z - z\frac{1}{n}\operatorname{tr}\left(\frac{1}{n} X_N X_N^* - z I_N\right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

- Repeating for entries $(2,2), \ldots, (N,N)$, and averaging, we get (for $\Im[z] > 0$)

$$m_{\mu_N}(z) - \frac{1}{1 - \frac{N}{n} - z - z\frac{N}{n} m_{\mu_N}(z)} \xrightarrow{\text{a.s.}} 0.$$

Proof of the Marčenko–Pastur law

Proof (continued)

- Then $m_{\mu_N}(z) \xrightarrow{\text{a.s.}} m(z)$ solution to

$$m(z) = \frac{1}{1 - c - z - czm(z)}$$

Proof of the Marčenko–Pastur law

Proof (continued)

- Then $m_{\mu_N}(z) \xrightarrow{\text{a.s.}} m(z)$ solution to

$$m(z) = \frac{1}{1 - c - z - czm(z)}$$

i.e., (with branch of $\sqrt{f(z)}$ such that $m(z) \to 0$ as $|z| \to \infty$)

$$m(z) = \frac{1-c}{2cz} - \frac{1}{2c} + \frac{\sqrt{\left(z - (1+\sqrt{c})^2\right)\left(z - (1-\sqrt{c})^2\right)}}{2cz}.$$

Proof of the Marčenko–Pastur law

Proof (continued)

- Then $m_{\mu_N}(z) \xrightarrow{\text{a.s.}} m(z)$ solution to

$$m(z) = \frac{1}{1 - c - z - czm(z)}$$

i.e., (with branch of $\sqrt{f(z)}$ such that $m(z) \to 0$ as $|z| \to \infty$)

$$m(z) = \frac{1-c}{2cz} - \frac{1}{2c} + \frac{\sqrt{\left(z - (1+\sqrt{c})^2\right)\left(z - (1-\sqrt{c})^2\right)}}{2cz}.$$

- Finally, by inverse Stieltjes Transform, for $x > 0$,

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \Im[m(x + \imath\varepsilon)] = \frac{\sqrt{\left((1+\sqrt{c})^2 - x\right)\left(x - (1-\sqrt{c})^2\right)}}{2\pi c x} 1_{\{x \in [(1-\sqrt{c})^2, (1+\sqrt{c})^2]\}}.$$

And for $x = 0$,

$$\lim_{\varepsilon \downarrow 0} \imath\varepsilon \Im[m(\imath\varepsilon)] = \left(1 - c^{-1}\right) 1_{\{c>1\}}.$$

Sample Covariance Matrices

Theorem (Sample Covariance Matrix Model **[Silverstein,Bai'95]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N \in \mathbb{C}^{N \times n}$*, with*

- $C_N \in \mathbb{C}^{N \times N}$ *nonnegative definite with e.s.d.* $\nu_N \to \nu$ *weakly,*
- $X_N \in \mathbb{C}^{N \times n}$ *has i.i.d. entries of zero mean and unit variance.*

As $N, n \to \infty$*,* $N/n \to c \in (0, \infty)$*,* $\tilde{\mu}_N$ *e.s.d. of* $\frac{1}{n} Y_N^* Y_N \in \mathbb{C}^{n \times n}$ *satisfies*

$$\tilde{\mu}_N \xrightarrow{\text{a.s.}} \tilde{\mu}$$

weakly, with $m_{\tilde{\mu}}(z)$*,* $\Im[z] > 0$*, unique solution with* $\Im[m_{\tilde{\mu}}(z)] > 0$ *of*

$$m_{\tilde{\mu}}(z) = \left(-z + c \int \frac{t}{1 + t m_{\tilde{\mu}}(z)} \nu(dt) \right)^{-1}.$$

Sample Covariance Matrices

Theorem (Sample Covariance Matrix Model **[Silverstein,Bai'95]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N \in \mathbb{C}^{N \times n}$*, with*

- $C_N \in \mathbb{C}^{N \times N}$ *nonnegative definite with e.s.d.* $\nu_N \to \nu$ *weakly,*
- $X_N \in \mathbb{C}^{N \times n}$ *has i.i.d. entries of zero mean and unit variance.*

As $N, n \to \infty$, $N/n \to c \in (0, \infty)$, $\tilde{\mu}_N$ *e.s.d. of* $\frac{1}{n} Y_N^* Y_N \in \mathbb{C}^{n \times n}$ *satisfies*

$$\tilde{\mu}_N \xrightarrow{\text{a.s.}} \tilde{\mu}$$

weakly, with $m_{\tilde{\mu}}(z)$, $\Im[z] > 0$, *unique solution with* $\Im[m_{\tilde{\mu}}(z)] > 0$ *of*

$$m_{\tilde{\mu}}(z) = \left(-z + c \int \frac{t}{1 + t m_{\tilde{\mu}}(z)} \nu(dt) \right)^{-1}.$$

Moreover, $\tilde{\mu}$ *is continuous on* $\mathbb{R}^+$ *and real analytic wherever positive.*

Sample Covariance Matrices

Theorem (Sample Covariance Matrix Model **[Silverstein,Bai'95]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N \in \mathbb{C}^{N\times n}$*, with*

- $C_N \in \mathbb{C}^{N\times N}$ *nonnegative definite with e.s.d.* $\nu_N \to \nu$ *weakly,*
- $X_N \in \mathbb{C}^{N\times n}$ *has i.i.d. entries of zero mean and unit variance.*

As $N, n \to \infty$*,* $N/n \to c \in (0,\infty)$*,* $\tilde{\mu}_N$ *e.s.d. of* $\frac{1}{n}Y_N^* Y_N \in \mathbb{C}^{n\times n}$ *satisfies*

$$\tilde{\mu}_N \xrightarrow{\text{a.s.}} \tilde{\mu}$$

weakly, with $m_{\tilde{\mu}}(z)$*,* $\Im[z] > 0$*, unique solution with* $\Im[m_{\tilde{\mu}}(z)] > 0$ *of*

$$m_{\tilde{\mu}}(z) = \left(-z + c \int \frac{t}{1 + t m_{\tilde{\mu}}(z)} \nu(dt) \right)^{-1}.$$

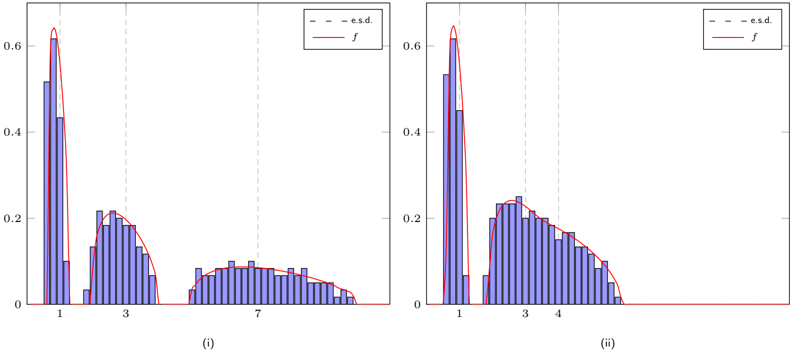
Moreover, $\tilde{\mu}$ *is continuous on* $\mathbb{R}^+$ *and real analytic wherever positive.*

Immediate corollary: For μ_N e.s.d. of $\frac{1}{n}Y_N Y_N^* = \frac{1}{n}\sum_{i=1}^n C_N^{\frac{1}{2}} x_i x_i^* C_N^{\frac{1}{2}}$,

$$\mu_N \xrightarrow{\text{a.s.}} \mu$$

weakly, with $\tilde{\mu} = c\mu + (1-c)\boldsymbol{\delta}_0$.

Sample Covariance Matrices



(i)

(ii)

Figure: Histogram of the eigenvalues of $\frac{1}{n}Y_N Y_N^*$, $n = 3000$, $N = 300$, with C_N diagonal with evenly weighted masses in (i) 1, 3, 7, (ii) 1, 3, 4.

Further Models and Deterministic Equivalents

Theorem (Doubly-correlated i.i.d. matrices)

Let $B_N = C_N^{\frac{1}{2}} X_N T_N X_N^* C_N^{\frac{1}{2}}$*, with e.s.d.* μ_N*,* $X_k \in \mathbb{C}^{N \times n}$ *with i.i.d. entries of zero mean, variance* $1/n$*,* C_N *Hermitian nonnegative definite,* T_N *diagonal nonnegative,* $\limsup_N \max(\|C_N\|, \|T_N\|) < \infty$*. Denote* $c = N/n$*. Then, as* $N, n \to \infty$ *with bounded ratio* c*, for* $z \in \mathbb{C} \setminus \mathbb{R}^-$*,*

$$m_{\mu_N}(z) - m_N(z) \xrightarrow{\text{a.s.}} 0, \quad m_N(z) = \frac{1}{N} \mathit{tr} \left(-zI_N + \bar{e}_N(z) C_N\right)^{-1}$$

with $\bar{e}(z)$ *unique solution in* $\{z \in \mathbb{C}^+, \bar{e}_N(z) \in \mathbb{C}^+\}$ *or* $\{z \in \mathbb{R}^-, \bar{e}_N(z) \in \mathbb{R}^+\}$ *of*

$$e_N(z) = \frac{1}{N} \mathit{tr} C_N \left(-zI_N + \bar{e}_N(z) C_N\right)^{-1}$$

$$\bar{e}_N(z) = \frac{1}{n} \mathit{tr} T_N \left(I_n + c e_N(z) T_N\right)^{-1}.$$

Other Refined Sample Covariance Models

Side note on other models.
Similar results for multiple matrix models:

Other Refined Sample Covariance Models

Side note on other models.
Similar results for multiple matrix models:

- **Information-plus-noise**: $Y_N = A_N + X_N$, A_N deterministic
- **Variance profile**: $Y_N = P_N \odot X_N$ (entry-wise product)
- **Per-column covariance**: $Y_N = [y_1, \ldots, y_n]$, $y_i = C_{N,i}^{\frac{1}{2}} x_i$
- etc.

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

No Eigenvalue Outside the Support

Theorem (No Eigenvalue Outside the Support **[Silverstein,Bai'98]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N \in \mathbb{C}^{N \times n}$, *with*

- $C_N \in \mathbb{C}^{N \times N}$ *nonnegative definite with e.s.d.* $\nu_N \to \nu$ *weakly,*

No Eigenvalue Outside the Support

Theorem (No Eigenvalue Outside the Support **[Silverstein,Bai'98]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N \in \mathbb{C}^{N\times n}$, *with*

- $C_N \in \mathbb{C}^{N\times N}$ *nonnegative definite with e.s.d.* $\nu_N \to \nu$ *weakly,*
- $E[|X_N|_{ij}^4] < \infty$,

No Eigenvalue Outside the Support

Theorem (No Eigenvalue Outside the Support **[Silverstein,Bai'98]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N \in \mathbb{C}^{N \times n}$*, with*

- $C_N \in \mathbb{C}^{N \times N}$ *nonnegative definite with e.s.d.* $\nu_N \to \nu$ *weakly,*
- $E[|X_N|_{ij}^4] < \infty$,
- $X_N \in \mathbb{C}^{N \times n}$ *has i.i.d. entries of zero mean and unit variance,*

No Eigenvalue Outside the Support

Theorem (No Eigenvalue Outside the Support **[Silverstein,Bai'98]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N \in \mathbb{C}^{N \times n}$*, with*

- $C_N \in \mathbb{C}^{N \times N}$ *nonnegative definite with e.s.d.* $\nu_N \to \nu$ *weakly,*
- $E[|X_N|_{ij}^4] < \infty$*,*
- $X_N \in \mathbb{C}^{N \times n}$ *has i.i.d. entries of zero mean and unit variance,*
- $\max_i \operatorname{dist}(\lambda_i(C_N), \operatorname{supp}(\nu)) \to 0$*.*

No Eigenvalue Outside the Support

Theorem (No Eigenvalue Outside the Support **[Silverstein,Bai'98]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N \in \mathbb{C}^{N\times n}$*, with*

- $C_N \in \mathbb{C}^{N\times N}$ *nonnegative definite with e.s.d.* $\nu_N \to \nu$ *weakly,*
- $E[|X_N|_{ij}^4] < \infty$*,*
- $X_N \in \mathbb{C}^{N\times n}$ *has i.i.d. entries of zero mean and unit variance,*
- $\max_i \operatorname{dist}(\lambda_i(C_N), \operatorname{supp}(\nu)) \to 0$*.*

Let $\tilde{\mu}$ *be the limiting e.s.d. of* $\frac{1}{n} Y_N^* Y_N$ *as before. Let* $[a,b] \subset \mathbb{R}^* \setminus \operatorname{supp}(\tilde{\nu})$*. Then,*

$$\left\{ \lambda_i \left(\frac{1}{n} Y_N^* Y_N \right) \right\}_{i=1}^{n} \cap [a,b] = \emptyset$$

for all large n*, almost surely.*

No Eigenvalue Outside the Support

Theorem (No Eigenvalue Outside the Support **[Silverstein,Bai'98]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N \in \mathbb{C}^{N \times n}$*, with*

- $C_N \in \mathbb{C}^{N \times N}$ *nonnegative definite with e.s.d.* $\nu_N \to \nu$ *weakly,*
- $E[|X_N|_{ij}^4] < \infty$*,*
- $X_N \in \mathbb{C}^{N \times n}$ *has i.i.d. entries of zero mean and unit variance,*
- $\max_i \operatorname{dist}(\lambda_i(C_N), \operatorname{supp}(\nu)) \to 0$*.*

Let $\tilde{\mu}$ *be the limiting e.s.d. of* $\frac{1}{n} Y_N^* Y_N$ *as before. Let* $[a, b] \subset \mathbb{R}^* \setminus \operatorname{supp}(\tilde{\nu})$*. Then,*

$$\left\{ \lambda_i \left(\frac{1}{n} Y_N^* Y_N \right) \right\}_{i=1}^{n} \cap [a, b] = \emptyset$$

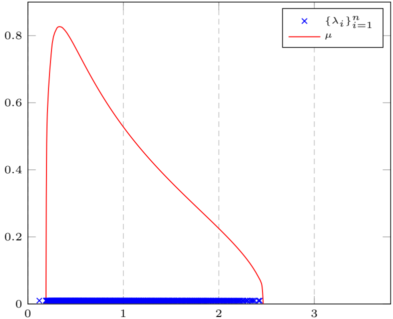
for all large n*, almost surely.*

In practice: This means that eigenvalues of $\frac{1}{n} Y_N^* Y_N$ cannot be bound at macroscopic distance from the bulk, for N, n large.

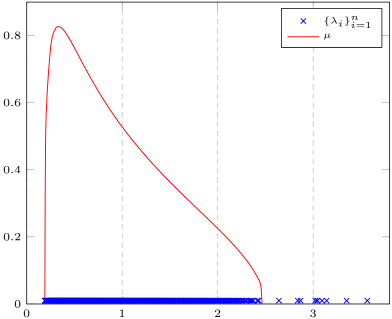
Spiked Models

Breaking the rules. If we break

- **Rule 1**: Infinitely many eigenvalues may wander away from $\mathrm{supp}(\mu)$.



$E[X_{ij}^4] < \infty$



$E[X_{ij}^4] = \infty$

Spiked Models

If we break:

- **Rule 2**: C_N may create isolated eigenvalues in $\frac{1}{n}Y_N Y_N^*$, called spikes.

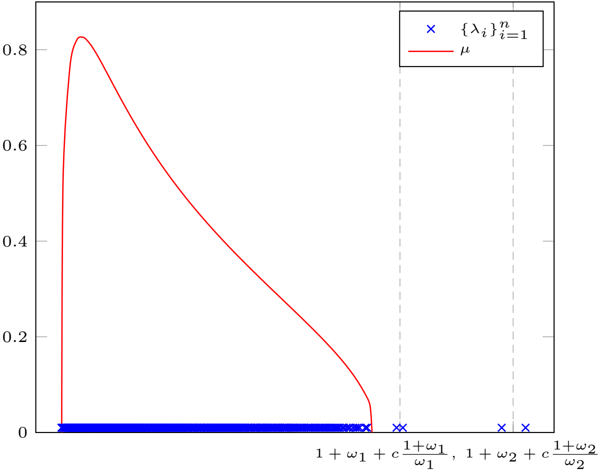


Figure: Eigenvalues of $\frac{1}{n}Y_N Y_N^*$, $C_N = \text{diag}(\underbrace{1,\ldots,1}_{N-4},2,2,3,3)$, $N=500$, $n=1500$.

Spiked Models

Theorem (Eigenvalues **[Baik,Silverstein'06]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N$, with

- X_N *with i.i.d. zero mean, unit variance, $E[|X_N|_{ij}^4] < \infty$.*
- $C_N = I_N + P$, $P = U\Omega U^*$, *where, for K fixed,*

$$\Omega = \textit{diag}(\omega_1, \ldots, \omega_K) \in \mathbb{R}^{K\times K}, \text{ with } \omega_1 \geq \ldots \geq \omega_K > 0.$$

Spiked Models

Theorem (Eigenvalues **[Baik,Silverstein'06]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N$, *with*

- X_N *with i.i.d. zero mean, unit variance,* $E[|X_N|_{ij}^4] < \infty$.
- $C_N = I_N + P$, $P = U\Omega U^*$, *where, for* K *fixed,*

$$\Omega = \textit{diag}\,(\omega_1, \ldots, \omega_K) \in \mathbb{R}^{K \times K}, \text{ with } \omega_1 \geq \ldots \geq \omega_K > 0.$$

Then, as $N, n \to \infty$, $N/n \to c \in (0, \infty)$, *denoting* $\lambda_i = \lambda_i(\frac{1}{n} Y_N Y_N^*)$,

- *if* $\omega_m > \sqrt{c}$,

$$\lambda_m \xrightarrow{\text{a.s.}} 1 + \omega_m + c\frac{1 + \omega_m}{\omega_m} > (1 + \sqrt{c})^2$$

Spiked Models

Theorem (Eigenvalues **[Baik,Silverstein'06]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N$, *with*

- X_N *with i.i.d. zero mean, unit variance,* $E[|X_N|_{ij}^4] < \infty$.
- $C_N = I_N + P$, $P = U\Omega U^*$, *where, for* K *fixed,*

$$\Omega = \mathit{diag}(\omega_1, \ldots, \omega_K) \in \mathbb{R}^{K\times K}, \text{ with } \omega_1 \geq \ldots \geq \omega_K > 0.$$

Then, as $N, n \to \infty$, $N/n \to c \in (0,\infty)$, *denoting* $\lambda_i = \lambda_i(\frac{1}{n}Y_N Y_N^*)$,

- *if* $\omega_m > \sqrt{c}$,

$$\lambda_m \xrightarrow{\text{a.s.}} 1 + \omega_m + c\frac{1+\omega_m}{\omega_m} > (1+\sqrt{c})^2$$

- *if* $\omega_m \in (0, \sqrt{c}]$,

$$\lambda_m \xrightarrow{\text{a.s.}} (1+\sqrt{c})^2$$

Spiked Models

Proof

- **Two ingredients**: Algebraic calculus + trace lemma

Spiked Models

Proof

- **Two ingredients**: Algebraic calculus + trace lemma
- **Find eigenvalues away from eigenvalues of $\frac{1}{n}X_N X_N^*$**:

$$\begin{aligned} 0 &= \det\left(\frac{1}{n}Y_N Y_N^* - \lambda I_N\right) \\ &= \det(C_N)\det\left(\frac{1}{n}X_N X_N^* - \lambda C_N^{-1}\right) \\ &= \det\left(\frac{1}{n}X_N X_N^* - \lambda I_N + \lambda(I_N - C_N^{-1})\right) \\ &= \det\left(\frac{1}{n}X_N X_N^* - \lambda I_N\right)\det\left(I_N + \lambda(I_N - C_N^{-1})\left(\frac{1}{n}X_N X_N^* - \lambda I_N\right)^{-1}\right). \end{aligned}$$

Spiked Models

Proof

- **Two ingredients**: Algebraic calculus + trace lemma
- **Find eigenvalues away from eigenvalues of** $\frac{1}{n}X_N X_N^*$:

$$\begin{aligned}
0 &= \det\left(\frac{1}{n}Y_N Y_N^* - \lambda I_N\right) \\
&= \det(C_N)\det\left(\frac{1}{n}X_N X_N^* - \lambda C_N^{-1}\right) \\
&= \det\left(\frac{1}{n}X_N X_N^* - \lambda I_N + \lambda(I_N - C_N^{-1})\right) \\
&= \det\left(\frac{1}{n}X_N X_N^* - \lambda I_N\right)\det\left(I_N + \lambda(I_N - C_N^{-1})\left(\frac{1}{n}X_N X_N^* - \lambda I_N\right)^{-1}\right).
\end{aligned}$$

- **Use low rank property**:

$$I_N - C_N^{-1} = I_N - (I_N + U\Omega U^*)^{-1} = U(I_K + \Omega^{-1})^{-1}U^*,\ \Omega \in \mathbb{C}^{K\times K}.$$

Hence

$$0 = \det\left(\frac{1}{n}X_N X_N^* - \lambda I_N\right)\det\left(I_N + \lambda U(I_K + \Omega^{-1})^{-1}U^*\left(\frac{1}{n}X_N X_N^* - \lambda I_N\right)^{-1}\right)$$

Spiked Models

Proof

- **Two ingredients**: Algebraic calculus + trace lemma
- **Find eigenvalues away from eigenvalues of $\frac{1}{n}X_NX_N^*$**:

$$\begin{aligned}
0 &= \det\left(\frac{1}{n}Y_NY_N^* - \lambda I_N\right) \\
&= \det(C_N)\det\left(\frac{1}{n}X_NX_N^* - \lambda C_N^{-1}\right) \\
&= \det\left(\frac{1}{n}X_NX_N^* - \lambda I_N + \lambda(I_N - C_N^{-1})\right) \\
&= \det\left(\frac{1}{n}X_NX_N^* - \lambda I_N\right)\det\left(I_N + \lambda(I_N - C_N^{-1})\left(\frac{1}{n}X_NX_N^* - \lambda I_N\right)^{-1}\right).
\end{aligned}$$

- **Use low rank property**:

$$I_N - C_N^{-1} = I_N - (I_N + U\Omega U^*)^{-1} = U(I_K + \Omega^{-1})^{-1}U^*,\ \Omega \in \mathbb{C}^{K\times K}.$$

Hence

$$0 = \det\left(\frac{1}{n}X_NX_N^* - \lambda I_N\right)\det\left(I_N + \lambda U(I_K + \Omega^{-1})^{-1}U^*\left(\frac{1}{n}X_NX_N^* - \lambda I_N\right)^{-1}\right)$$

Proof (2)

- **Sylverster's identity** $(\det(I+AB)=\det(I+BA))$,

$$0=\det\left(\frac{1}{n}X_NX_N^*-\lambda I_N\right)\det\left(I_K+\lambda(I_K+\Omega^{-1})^{-1}U^*\left(\frac{1}{n}X_NX_N^*-\lambda I_N\right)^{-1}U\right)$$

Spiked Models

Proof (2)

- **Sylverster's identity** $(\det(I+AB)=\det(I+BA))$,

$$0=\det\left(\frac{1}{n}X_NX_N^*-\lambda I_N\right)\det\left(I_K+\lambda(I_K+\Omega^{-1})^{-1}U^*\left(\frac{1}{n}X_NX_N^*-\lambda I_N\right)^{-1}U\right)$$

- **No eigenvalue outside the support [Bai,Sil'98]**: $\det(\frac{1}{n}X_NX_N^*-\lambda I_N)$ has no zero beyond $(1+\sqrt{c})^2$ for all large n a.s.

Spiked Models

Proof (2)

- **Sylverster's identity** $(\det(I + AB) = \det(I + BA))$,

$$0 = \det\left(\frac{1}{n}X_N X_N^* - \lambda I_N\right)\det\left(I_K + \lambda(I_K + \Omega^{-1})^{-1}U^*\left(\frac{1}{n}X_N X_N^* - \lambda I_N\right)^{-1}U\right)$$

- **No eigenvalue outside the support [Bai,Sil'98]**: $\det(\frac{1}{n}X_N X_N^* - \lambda I_N)$ has no zero beyond $(1 + \sqrt{c})^2$ for all large n a.s.
- **Extension of Trace Lemma**: for each $z \in \mathbb{C} \setminus \mathrm{supp}(\mu)$,

$$U^*\left(\frac{1}{n}X_N X_N^* - zI_N\right)^{-1}U \xrightarrow{\text{a.s.}} m_\mu(z)I_K.$$

(X_N being "almost-unitarily invariant", U can be seen as formed of random "i.i.d.-like" vectors)

Proof (2)

- **Sylverster's identity** $(\det(I + AB) = \det(I + BA))$,

$$0 = \det\left(\frac{1}{n}X_N X_N^* - \lambda I_N\right)\det\left(I_K + \lambda(I_K + \Omega^{-1})^{-1}U^*\left(\frac{1}{n}X_N X_N^* - \lambda I_N\right)^{-1}U\right)$$

- **No eigenvalue outside the support [Bai,Sil'98]**: $\det(\frac{1}{n}X_N X_N^* - \lambda I_N)$ has no zero beyond $(1+\sqrt{c})^2$ for all large n a.s.
- **Extension of Trace Lemma**: for each $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$,

$$U^*\left(\frac{1}{n}X_N X_N^* - zI_N\right)^{-1}U \xrightarrow{\text{a.s.}} m_\mu(z)I_K.$$

(X_N being "almost-unitarily invariant", U can be seen as formed of random "i.i.d.-like" vectors)

- As a result, for all large n a.s.,

$$\begin{aligned}0 &= \det\left(I_K + \lambda(I_K + \Omega^{-1})^{-1}U^*(\frac{1}{n}X_N X_N^* - \lambda I_N)^{-1}U\right)\\ &\simeq \prod_{m=1}^{M}\left(1 + \frac{\lambda}{1+\omega_m^{-1}}m_\mu(\lambda)\right)^{k_m} = \prod_{m=1}^{M}\left(1 + \frac{\lambda\omega_m}{1+\omega_m}m_\mu(\lambda)\right)^{k_m}\end{aligned}$$

Proof (3)

- **Limiting solutions:** zeros (with multiplicity) of

$$1 + \frac{\lambda \omega_m}{1 + \omega_m} m_\mu(\lambda) = 0.$$

Spiked Models

Proof (3)

- **Limiting solutions:** zeros (with multiplicity) of

$$1 + \frac{\lambda \omega_m}{1+\omega_m} m_\mu(\lambda) = 0.$$

- Using Marčenko–Pastur law properties ($m_\mu(z) = (1 - c - z - czm_\mu(z))^{-1}$),

$$\lambda \in \left\{ 1 + \omega_m + c\frac{1+\omega_m}{\omega_m} \right\}_{m=1}^{M}.$$

Spiked Models

Theorem (Eigenvectors **[Paul'07]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N$*, with*

- X_N *with i.i.d. zero mean, unit variance, finite fourth order moment entries*
- $C_N = I_N + P$, $P = \sum_{i=1}^{K} \omega_i u_i u_i^*$, $\omega_1 > \ldots > \omega_M > 0$.

Spiked Models

Theorem (Eigenvectors **[Paul'07]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N$*, with*

- X_N *with i.i.d. zero mean, unit variance, finite fourth order moment entries*
- $C_N = I_N + P$*,* $P = \sum_{i=1}^K \omega_i u_i u_i^*$*,* $\omega_1 > \ldots > \omega_M > 0$*.*

Then, as $N, n \to \infty$*,* $N/n \to c \in (0, \infty)$*, for* $a, b \in \mathbb{C}^N$ *deterministic and* $\hat{u}_i$ *eigenvector of* $\lambda_i(\frac{1}{n} Y_N Y_N^*)$*,*

$$a^* \hat{u}_i \hat{u}_i^* b - \frac{1 - c\omega_i^{-2}}{1 + c\omega_i^{-1}} a^* u_i u_i^* b \cdot 1_{\omega_i > \sqrt{c}} \xrightarrow{\text{a.s.}} 0$$

In particular,

$$|\hat{u}_i^* u_i|^2 \xrightarrow{\text{a.s.}} \frac{1 - c\omega_i^{-2}}{1 + c\omega_i^{-1}} \cdot 1_{\omega_i > \sqrt{c}}.$$

Spiked Models

Theorem (Eigenvectors [Paul'07])

Let $Y_N = C_N^{\frac{1}{2}} X_N$, *with*

- X_N *with i.i.d. zero mean, unit variance, finite fourth order moment entries*
- $C_N = I_N + P$, $P = \sum_{i=1}^K \omega_i u_i u_i^*$, $\omega_1 > \ldots > \omega_M > 0$.

Then, as $N, n \to \infty$, $N/n \to c \in (0, \infty)$, *for* $a, b \in \mathbb{C}^N$ *deterministic and* $\hat{u}_i$ *eigenvector of* $\lambda_i(\frac{1}{n} Y_N Y_N^*)$,

$$a^* \hat{u}_i \hat{u}_i^* b - \frac{1 - c\omega_i^{-2}}{1 + c\omega_i^{-1}} a^* u_i u_i^* b \cdot 1_{\omega_i > \sqrt{c}} \xrightarrow{\text{a.s.}} 0$$

In particular,

$$|\hat{u}_i^* u_i|^2 \xrightarrow{\text{a.s.}} \frac{1 - c\omega_i^{-2}}{1 + c\omega_i^{-1}} \cdot 1_{\omega_i > \sqrt{c}}.$$

Proof: Based on Cauchy integral + similar ingredients as eigenvalue proof

$$a^* \hat{u}_i \hat{u}_i^* b = \frac{1}{2\pi\imath} \oint_{\mathcal{C}_i} a^* \left(\frac{1}{n} Y_N Y_N^* - z I_N \right)^{-1} b \, dz$$

for $\mathcal{C}_m$ contour circling around λ_i only.

Spiked Models

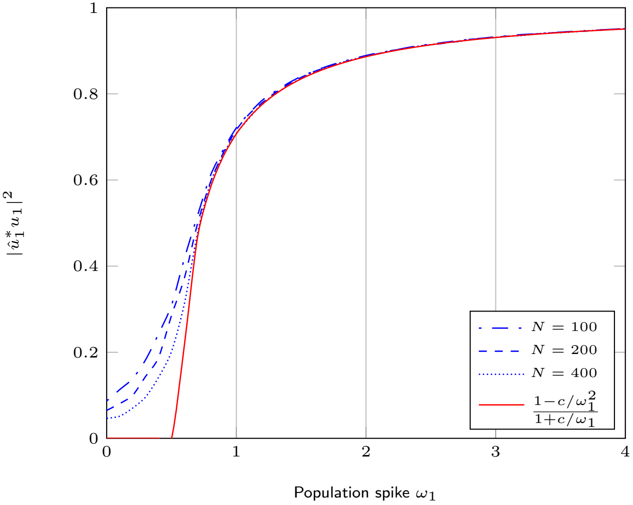


Figure: Simulated versus limiting $|\hat{u}_1^* u_1|^2$ for $Y_N = C_N^{\frac{1}{2}} X_N$, $C_N = I_N + \omega_1 u_1 u_1^*$, $N/n = 1/3$, varying ω_1.

Tracy–Widom Theorem

Theorem (Phase Transition **[Baik,BenArous,Péché'05]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N$*, with*

- X_N *with i.i.d. complex Gaussian zero mean, unit variance entries,*
- $C_N = I_N + P$*,* $P = \sum_{i=1}^K \omega_i u_i u_i^*$*,* $\omega_1 > \ldots > \omega_K > 0$ *(*$K \geq 0$*).*

Tracy–Widom Theorem

Theorem (Phase Transition **[Baik,BenArous,Péché'05]**)

Let $Y_N = C_N^{\frac{1}{2}} X_N$, *with*

- X_N *with i.i.d. complex Gaussian zero mean, unit variance entries,*
- $C_N = I_N + P$, $P = \sum_{i=1}^K \omega_i u_i u_i^*$, $\omega_1 > \ldots > \omega_K > 0$ *(*$K \geq 0$*).*

Then, as $N, n \to \infty$, $N/n \to c < 1$,

- *If* $\omega_1 < \sqrt{c}$ *(or* $K = 0$*),*

$$N^{\frac{2}{3}} \frac{\lambda_1 - (1+\sqrt{c})^2}{(1+\sqrt{c})^{\frac{4}{3}} c^{\frac{1}{2}}} \xrightarrow{\mathcal{L}} T_2, \text{ (complex Tracy–Widom law)}$$

Tracy–Widom Theorem

Theorem (Phase Transition [Baik,BenArous,Péché'05])

Let $Y_N = C_N^{\frac{1}{2}} X_N$, with

- *X_N with i.i.d. complex Gaussian zero mean, unit variance entries,*
- *$C_N = I_N + P$, $P = \sum_{i=1}^K \omega_i u_i u_i^*$, $\omega_1 > \ldots > \omega_K > 0$ ($K \geq 0$).*

Then, as $N, n \to \infty$, $N/n \to c < 1$,

- *If $\omega_1 < \sqrt{c}$ (or $K = 0$),*

$$N^{\frac{2}{3}} \frac{\lambda_1 - (1+\sqrt{c})^2}{(1+\sqrt{c})^{\frac{4}{3}} c^{\frac{1}{2}}} \xrightarrow{\mathcal{L}} T_2, \quad \text{(complex Tracy–Widom law)}$$

- *If $\omega_1 > \sqrt{c}$,*

$$\left(\frac{(1+\omega_1)^2}{c} - \frac{(1+\omega_1)^2}{\omega_1^2}\right)^{\frac{1}{2}} N^{\frac{1}{2}} \left[\lambda_1 - \left(1 + \omega_1 + c\frac{1+\omega_1}{\omega_1}\right)\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Tracy–Widom Theorem

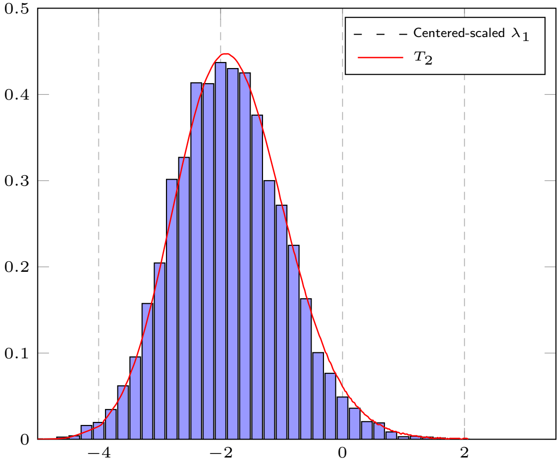


Figure: Distribution of $N^{\frac{2}{3}} c^{-\frac{1}{2}} (1+\sqrt{c})^{-\frac{4}{3}} \left[\lambda_1(\frac{1}{n} X_N X_N^*) - (1+\sqrt{c})^2\right]$ versus Tracy–Widom (T_2), $N = 500$, $n = 1500$.

Other Spiked Models

Similar results for multiple matrix models:

- **Additive spiked model**: $Y_N = \frac{1}{n}XX^* + P$, P deterministic and low rank
- $Y_N = \frac{1}{n}X^*(I+P)X$
- $Y_N = \frac{1}{n}(X+P)^*(X+P)$
- $Y_N = \frac{1}{n}TX^*(I+P)XT$
- etc.

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

The Semi-circle law

Theorem

Let $X_N \in \mathbb{C}^{N\times N}$ Hermitian with e.s.d. μ_N such that $\frac{1}{\sqrt{N}}[X_N]_{i>j}$ are i.i.d. with zero mean and unit variance. Then, as $N \to \infty$,

$$\mu_N \xrightarrow{\text{a.s.}} \mu$$

with $\mu(dt) = \frac{1}{2\pi}\sqrt{(4-t^2)^+}dt$. In particular, m_μ satisfies

$$m_\mu(z) = \frac{1}{-z - m_\mu(z)}.$$

The Semi-circle law

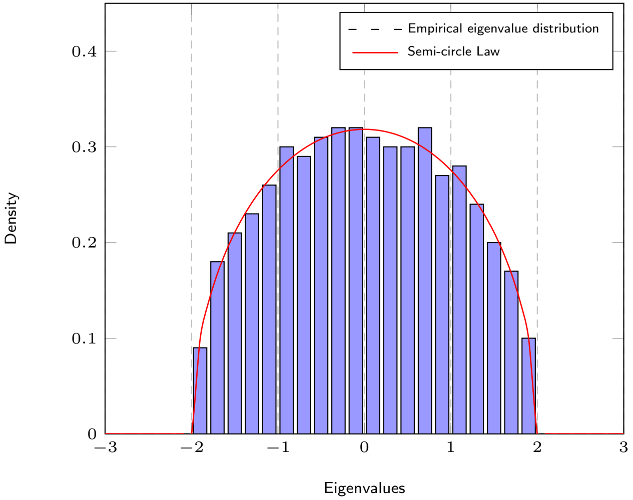


Figure: Histogram of the eigenvalues of Wigner matrices and the semi-circle law, for $N = 500$

The Circular law

Theorem

Let $X_N \in \mathbb{C}^{N\times N}$ *with e.s.d.* μ_N *be such that* $\frac{1}{\sqrt{N}}[X_N]_{ij}$ *are i.i.d. entries with zero mean and unit variance. Then, as* $N \to \infty$*,*

$$\mu_N \xrightarrow{\text{a.s.}} \mu$$

with μ *a complex-supported measure with* $\mu(dz) = \frac{1}{2\pi}\delta_{|z|\leq 1}dz$*.*

The Circular law

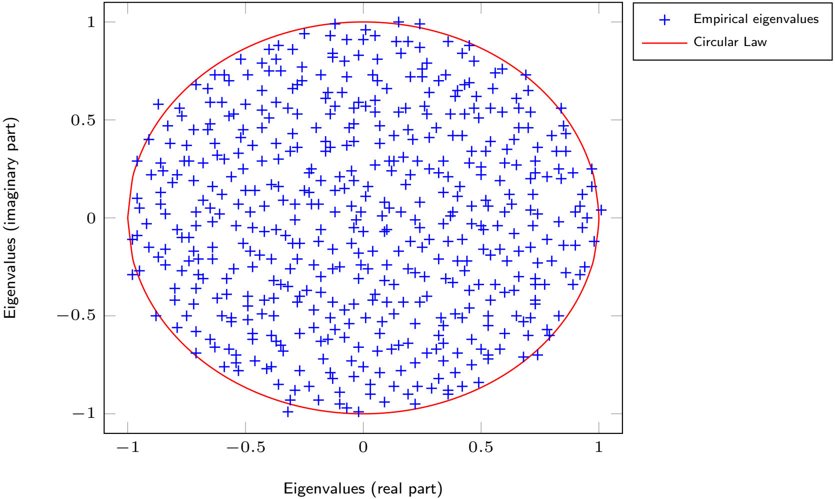


Figure: Eigenvalues of $\mathbf{X}_N$ with i.i.d. standard Gaussian entries, for $N = 500$.

Bibliographical references: Maths Book and Tutorial References I

From most accessible to least:

- Couillet, R., & Debbah, M. (2011). Random matrix methods for wireless communications. Cambridge University Press.
- Tao, T. (2012). Topics in random matrix theory (Vol. 132). Providence, RI: American Mathematical Society.
- Bai, Z., & Silverstein, J. W. (2010). Spectral analysis of large dimensional random matrices (Vol. 20). New York: Springer.
- Pastur, L. A., Shcherbina, M., & Shcherbina, M. (2011). Eigenvalue distribution of large random matrices (Vol. 171). Providence, RI: American Mathematical Society.
- Anderson, G. W., Guionnet, A., & Zeitouni, O. (2010). An introduction to random matrices (Vol. 118). Cambridge university press.

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

Context

Baseline scenario: $x_1, \ldots, x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) i.i.d. with $E[x_1] = 0$, $E[x_1 x_1^*] = C_N$:

Context

Baseline scenario: $x_1, \ldots, x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) i.i.d. with $E[x_1] = 0$, $E[x_1 x_1^*] = C_N$:

- If $x_1 \sim \mathcal{N}(0, C_N)$, ML estimator for C_N is sample covariance matrix (SCM)

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^*.$$

Context

Baseline scenario: $x_1, \ldots, x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) i.i.d. with $E[x_1] = 0$, $E[x_1 x_1^*] = C_N$:

- If $x_1 \sim \mathcal{N}(0, C_N)$, ML estimator for C_N is sample covariance matrix (SCM)

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^*.$$

- **[Huber'67]** If $x_1 \sim (1-\varepsilon)\mathcal{N}(0, C_N) + \varepsilon G$, G unknown, robust estimator ($n > N$)

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^{n} \max\left\{\ell_1, \frac{\ell_2}{\frac{1}{N} x_i^* \hat{C}_N^{-1} x_i}\right\} x_i x_i^* \text{ for some } \ell_1, \ell_2 > 0.$$

Context

Baseline scenario: $x_1,\ldots,x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) i.i.d. with $E[x_1]=0$, $E[x_1x_1^*]=C_N$:

- If $x_1 \sim \mathcal{N}(0,C_N)$, ML estimator for C_N is sample covariance matrix (SCM)

$$\hat{C}_N = \frac{1}{n}\sum_{i=1}^n x_i x_i^*.$$

- **[Huber'67]** If $x_1 \sim (1-\varepsilon)\mathcal{N}(0,C_N)+\varepsilon G$, G unknown, robust estimator ($n>N$)

$$\hat{C}_N = \frac{1}{n}\sum_{i=1}^n \max\left\{\ell_1, \frac{\ell_2}{\frac{1}{N}x_i^*\hat{C}_N^{-1}x_i}\right\} x_i x_i^* \text{ for some } \ell_1,\ell_2>0.$$

- **[Maronna'76]** If x_1 elliptical (and $n>N$), ML estimator for C_N given by

$$\hat{C}_N = \frac{1}{n}\sum_{i=1}^n u\left(\frac{1}{N}x_i^*\hat{C}_N^{-1}x_i\right) x_i x_i^* \text{ for some non-increasing } u.$$

Context

Baseline scenario: $x_1,\ldots,x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) i.i.d. with $E[x_1]=0$, $E[x_1x_1^*]=C_N$:

- If $x_1 \sim \mathcal{N}(0,C_N)$, ML estimator for C_N is sample covariance matrix (SCM)

$$\hat{C}_N = \frac{1}{n}\sum_{i=1}^n x_i x_i^*.$$

- **[Huber'67]** If $x_1 \sim (1-\varepsilon)\mathcal{N}(0,C_N)+\varepsilon G$, G unknown, robust estimator ($n>N$)

$$\hat{C}_N = \frac{1}{n}\sum_{i=1}^n \max\left\{\ell_1, \frac{\ell_2}{\frac{1}{N}x_i^*\hat{C}_N^{-1}x_i}\right\} x_i x_i^* \text{ for some } \ell_1,\ell_2>0.$$

- **[Maronna'76]** If x_1 elliptical (and $n>N$), ML estimator for C_N given by

$$\hat{C}_N = \frac{1}{n}\sum_{i=1}^n u\left(\frac{1}{N}x_i^*\hat{C}_N^{-1}x_i\right) x_i x_i^* \text{ for some non-increasing } u.$$

- **[Pascal'13; Chen'11]** If $N>n$, x_1 elliptical or with outliers, shrinkage extensions

$$\hat{C}_N(\rho) = (1-\rho)\frac{1}{n}\sum_{i=1}^n \frac{x_i x_i^*}{\frac{1}{N}x_i^*\hat{C}_N^{-1}(\rho)x_i} + \rho I_N$$

$$\check{C}_N(\rho) = \frac{\check{B}_N(\rho)}{\frac{1}{N}\operatorname{tr}\check{B}_N(\rho)},\ \check{B}_N(\rho) = (1-\rho)\frac{1}{n}\sum_{i=1}^n \frac{x_i x_i^*}{\frac{1}{N}x_i^*\check{C}_N^{-1}(\rho)x_i} + \rho I_N$$

Context

Results only known for N fixed and $n \to \infty$:

- not appropriate in settings of interest today (BigData, array processing, MIMO)

Context

Results only known for N fixed and $n \to \infty$:

- not appropriate in settings of interest today (BigData, array processing, MIMO)

We study such $\hat{C}_N$ in the regime

$$N, n \to \infty, \ N/n \to c \in (0, \infty).$$

Context

Results only known for N fixed and $n \to \infty$:

- not appropriate in settings of interest today (BigData, array processing, MIMO)

We study such $\hat{C}_N$ in the regime

$$N, n \to \infty, \ N/n \to c \in (0, \infty).$$

- Math interest:
  - limiting eigenvalue distribution of $\hat{C}_N$
  - limiting values and fluctuations of functionals $f(\hat{C}_N)$

Context

Results only known for N fixed and $n \to \infty$:

- not appropriate in settings of interest today (BigData, array processing, MIMO)

We study such $\hat{C}_N$ in the regime

$$N, n \to \infty, \;\; N/n \to c \in (0, \infty).$$

- Math interest:
  - limiting eigenvalue distribution of $\hat{C}_N$
  - limiting values and fluctuations of functionals $f(\hat{C}_N)$
- Application interest:
  - comparison between SCM and robust estimators
  - performance of robust/non-robust estimation methods
  - improvement thereof (by proper parametrization)

Model Description

Definition (Maronna's Estimator)

For $x_1, \ldots, x_n \in \mathbb{C}^N$ with $n > N$, $\hat{C}_N$ is the solution (upon existence and uniqueness) of

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^{n} u\left(\frac{1}{N} x_i^* \hat{C}_N^{-1} x_i\right) x_i x_i^*$$

Model Description

Definition (Maronna's Estimator)

For $x_1, \ldots, x_n \in \mathbb{C}^N$ with $n > N$, $\hat{C}_N$ is the solution (upon existence and uniqueness) of

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^{n} u\left(\frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^*$$

where $u : [0, \infty) \to (0, \infty)$ is

- non-increasing
- such that $\phi(x) \triangleq x u(x)$ increasing of supremum ϕ_∞ with

$$1 < \phi_\infty < c^{-1}, \ c \in (0,1).$$

The Results in a Nutshell

For various models of the x_i's,

- First order convergence:

$$\left\| \hat{C}_N - \hat{S}_N \right\| \xrightarrow{\text{a.s.}} 0$$

for some tractable random matrices $\hat{S}_N$.
⇒ We only discuss this result here.

The Results in a Nutshell

For various models of the x_i's,

- First order convergence:

$$\left\| \hat{C}_N - \hat{S}_N \right\| \xrightarrow{\text{a.s.}} 0$$

for some tractable random matrices $\hat{S}_N$.
$\Rightarrow$ We only discuss this result here.

- Second order results:

$$N^{1-\varepsilon}\left(a^*\hat{C}_N^k b - a^*\hat{S}_N^k b\right) \xrightarrow{\text{a.s.}} 0$$

allowing transfer of CLT results.

The Results in a Nutshell

For various models of the x_i's,

- First order convergence:

$$\left\| \hat{C}_N - \hat{S}_N \right\| \xrightarrow{\text{a.s.}} 0$$

for some tractable random matrices $\hat{S}_N$.
⇒ We only discuss this result here.

- Second order results:

$$N^{1-\varepsilon} \left(a^* \hat{C}_N^k b - a^* \hat{S}_N^k b \right) \xrightarrow{\text{a.s.}} 0$$

allowing transfer of CLT results.

- Applications:
  - improved robust covariance matrix estimation
  - improved robust tests / estimators
  - specific examples in statistics at large, array processing, statistical finance, etc.

(Elliptical) scenario

Theorem (Large dimensional behavior, elliptical case)

For $x_i = \sqrt{\tau_i} w_i$, τ_i impulsive (random or not), w_i unitarily invariant, $\|w_i\| = N$,

$$\left\| \hat{C}_N - \hat{S}_N \right\| \xrightarrow{\text{a.s.}} 0$$

with, for some v related to u ($v = u \circ g^{-1}$, $g(x) = x(1 - c\phi(x))^{-1}$),

$$\hat{C}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} u\left(\frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^*, \quad \hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} v(\tau_i \gamma_N) x_i x_i^*$$

and γ_N unique solution of

$$1 = \frac{1}{n} \sum_{j=1}^{n} \frac{\gamma v(\tau_i \gamma)}{1 + c\gamma v(\tau_i \gamma)}.$$

(Elliptical) scenario

Theorem (Large dimensional behavior, elliptical case)

For $x_i = \sqrt{\tau_i} w_i$, τ_i impulsive (random or not), w_i unitarily invariant, $\|w_i\| = N$,

$$\left\| \hat{C}_N - \hat{S}_N \right\| \xrightarrow{\text{a.s.}} 0$$

with, for some v related to u ($v = u \circ g^{-1}$, $g(x) = x(1 - c\phi(x))^{-1}$),

$$\hat{C}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} u\left(\frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^*, \quad \hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} v(\tau_i \gamma_N) x_i x_i^*$$

and γ_N unique solution of

$$1 = \frac{1}{n} \sum_{j=1}^{n} \frac{\gamma v(\tau_i \gamma)}{1 + c\gamma v(\tau_i \gamma)}.$$

Corollaries

- **Spectral measure**: $\mu_N^{\hat{C}_N} - \mu_N^{\hat{S}_N} \xrightarrow{\mathcal{L}} 0$ a.s. ($\mu_N^X \triangleq \frac{1}{n} \sum_{i=1}^n \boldsymbol{\delta}_{\lambda_i(X)}$)

(Elliptical) scenario

Theorem (Large dimensional behavior, elliptical case)

For $x_i = \sqrt{\tau_i} w_i$, τ_i impulsive (random or not), w_i unitarily invariant, $\|w_i\| = N$,

$$\left\| \hat{C}_N - \hat{S}_N \right\| \xrightarrow{\text{a.s.}} 0$$

with, for some v related to u ($v = u \circ g^{-1}$, $g(x) = x(1 - c\phi(x))^{-1}$),

$$\hat{C}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} u \left(\frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^*, \quad \hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} v(\tau_i \gamma_N) x_i x_i^*$$

and γ_N unique solution of

$$1 = \frac{1}{n} \sum_{j=1}^{n} \frac{\gamma v(\tau_i \gamma)}{1 + c\gamma v(\tau_i \gamma)}.$$

Corollaries

- **Spectral measure**: $\mu_N^{\hat{C}_N} - \mu_N^{\hat{S}_N} \xrightarrow{\mathcal{L}} 0$ a.s. ($\mu_N^X \triangleq \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\delta}_{\lambda_i(X)}$)
- **Local convergence**: $\max_{1 \leq i \leq N} |\lambda_i(\hat{C}_N) - \lambda_i(\hat{S}_N)| \xrightarrow{\text{a.s.}} 0.$

(Elliptical) scenario

Theorem (Large dimensional behavior, elliptical case)

For $x_i = \sqrt{\tau_i} w_i$, τ_i impulsive (random or not), w_i unitarily invariant, $\|w_i\| = N$,

$$\left\| \hat{C}_N - \hat{S}_N \right\| \xrightarrow{\text{a.s.}} 0$$

with, for some v related to u ($v = u \circ g^{-1}$, $g(x) = x(1 - c\phi(x))^{-1}$),

$$\hat{C}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} u \left(\frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^*, \quad \hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} v(\tau_i \gamma_N) x_i x_i^*$$

and γ_N unique solution of

$$1 = \frac{1}{n} \sum_{j=1}^{n} \frac{\gamma v(\tau_i \gamma)}{1 + c\gamma v(\tau_i \gamma)}.$$

Corollaries

- **Spectral measure**: $\mu_N^{\hat{C}_N} - \mu_N^{\hat{S}_N} \xrightarrow{\mathcal{L}} 0$ a.s. ($\mu_N^X \triangleq \frac{1}{n} \sum_{i=1}^n \boldsymbol{\delta}_{\lambda_i(X)}$)
- **Local convergence**: $\max_{1 \leq i \leq N} |\lambda_i(\hat{C}_N) - \lambda_i(\hat{S}_N)| \xrightarrow{\text{a.s.}} 0.$
- **Norm boundedness**: $\limsup_N \|\hat{C}_N\| < \infty$

$\longrightarrow$ Bounded spectrum (unlike SCM!)

Large dimensional behavior

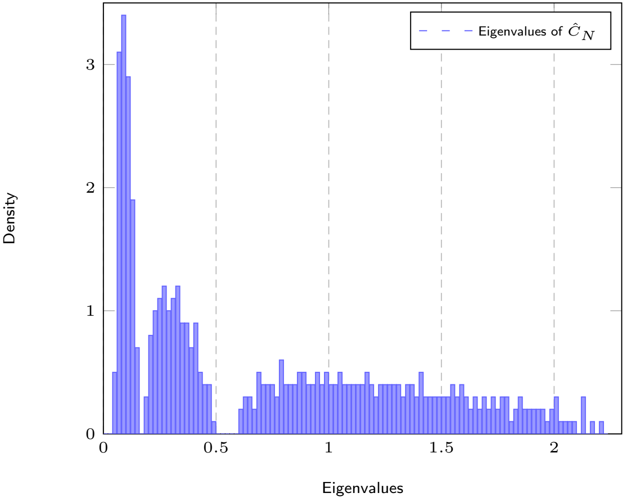


Figure: $n = 2500$, $N = 500$, $C_N = \text{diag}(I_{125}, 3I_{125}, 10I_{250})$, $\tau_i \sim \Gamma(.5, 2)$ i.i.d.

Large dimensional behavior

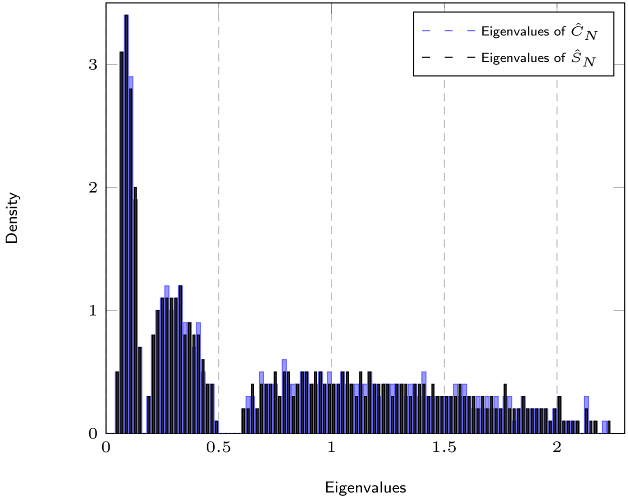


Figure: $n = 2500$, $N = 500$, $C_N = \text{diag}(I_{125}, 3I_{125}, 10I_{250})$, $\tau_i \sim \Gamma(.5, 2)$ i.i.d.

Large dimensional behavior

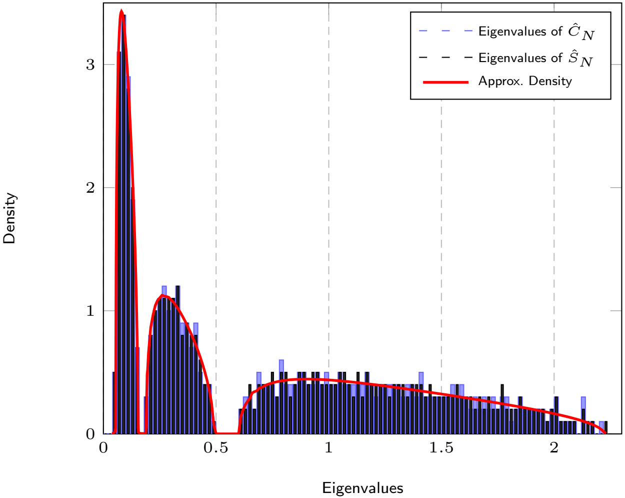


Figure: $n = 2500$, $N = 500$, $C_N = \text{diag}(I_{125}, 3I_{125}, 10I_{250})$, $\tau_i \sim \Gamma(.5, 2)$ i.i.d.

Elements of Proof

Definition (v and ψ)

Letting $g(x) = x(1 - c\phi(x))^{-1}$ (on $\mathbb{R}_+$),

$$\begin{aligned} v(x) &\triangleq (u \circ g^{-1})(x) && \text{non-increasing} \\ \psi(x) &\triangleq xv(x) && \text{increasing and bounded by } \psi_\infty. \end{aligned}$$

Elements of Proof

Definition (v and ψ)

Letting $g(x) = x(1 - c\phi(x))^{-1}$ (on $\mathbb{R}_+$),

$$v(x) \triangleq (u \circ g^{-1})(x) \quad \text{non-increasing}$$
$$\psi(x) \triangleq xv(x) \qquad \text{increasing and bounded by } \psi_\infty.$$

Lemma (Rewriting $\hat{C}_N$)

It holds (with $C_N = I_N$) that

$$\hat{C}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} \tau_i v\left(\tau_i d_i\right) w_i w_i^*$$

with $(d_1, \ldots, d_n) \in \mathbb{R}_+^n$ a.s. unique solution to

$$d_i = \frac{1}{N} w_i^* \hat{C}_{(i)}^{-1} w_i = \frac{1}{N} w_i^* \left(\frac{1}{n} \sum_{j \neq i} \tau_j v(\tau_j d_j) w_j w_j^* \right)^{-1} w_i, \; i = 1, \ldots, n.$$

Elements of Proof

Remark (Quadratic Form close to Trace)

Random matrix insight: $(\frac{1}{n}\sum_{j\neq i}\tau_j v(\tau_j d_j) w_j w_j^*)^{-1}$ “almost independent” of w_i, so

Elements of Proof

Remark (Quadratic Form close to Trace)

Random matrix insight: $(\frac{1}{n}\sum_{j\neq i}\tau_j v(\tau_j d_j) w_j w_j^*)^{-1}$ "almost independent" of w_i, so

$$d_i = \frac{1}{N} w_i^* \left(\frac{1}{n} \sum_{j\neq i} \tau_j v(\tau_j d_j) w_j w_j^* \right)^{-1} w_i \simeq \frac{1}{N} \operatorname{tr} \left(\frac{1}{n} \sum_{j\neq i} \tau_j v(\tau_j d_j) w_j w_j^* \right)^{-1} \simeq \gamma_N$$

for some deterministic sequence $(\gamma_N)_{N=1}^{\infty}$, irrespective of i.

Elements of Proof

Remark (Quadratic Form close to Trace)

Random matrix insight: $(\frac{1}{n}\sum_{j\neq i}\tau_j v(\tau_j d_j) w_j w_j^*)^{-1}$ "almost independent" of w_i, so

$$d_i = \frac{1}{N} w_i^* \left(\frac{1}{n} \sum_{j\neq i} \tau_j v(\tau_j d_j) w_j w_j^* \right)^{-1} w_i \simeq \frac{1}{N} \mathrm{tr} \left(\frac{1}{n} \sum_{j\neq i} \tau_j v(\tau_j d_j) w_j w_j^* \right)^{-1} \simeq \gamma_N$$

for some deterministic sequence $(\gamma_N)_{N=1}^{\infty}$, irrespective of i.

Lemma (Key Lemma)

Letting $e_i \triangleq \frac{v(\tau_i d_i)}{v(\tau_i \gamma_N)}$ with γ_N unique solution to

$$1 = \frac{1}{n} \sum_{k=1}^{n} \frac{\psi(\tau_i \gamma_N)}{1 + c\psi(\tau_i \gamma_N)}$$

we have

$$\max_{1\leq i\leq n} |e_i - 1| \xrightarrow{\text{a.s.}} 0.$$

Proof of the Key Lemma: $\max_i |e_i - 1| \xrightarrow{\text{a.s.}} 0$, $e_i = \frac{v(\tau_i d_i)}{v(\tau_i \gamma_N)}$

Property (Quadratic form and γ_N)

$$\max_{1\leq i\leq n}\left|\frac{1}{N}w_i^*\left(\frac{1}{n}\sum_{j\neq i}\tau_j v(\tau_j\gamma_N)w_jw_j^*\right)^{-1}w_i-\gamma_N\right|\xrightarrow{\text{a.s.}}0.$$

Proof of the Key Lemma: $\max_i |e_i - 1| \xrightarrow{\text{a.s.}} 0$, $e_i = \frac{v(\tau_i d_i)}{v(\tau_i \gamma_N)}$

Property (Quadratic form and γ_N)

$$\max_{1\leq i\leq n} \left| \frac{1}{N} w_i^* \left(\frac{1}{n} \sum_{j\neq i} \tau_j v(\tau_j \gamma_N) w_j w_j^* \right)^{-1} w_i - \gamma_N \right| \xrightarrow{\text{a.s.}} 0.$$

Proof of the Property

- Uniformity easy (moments of all orders for $[w_i]_j$).
- By a “quadratic form similar to trace” approach, we get

$$\max_{1\leq i\leq n} \left| \frac{1}{N} w_i^* \left(\frac{1}{n} \sum_{j\neq i} \tau_j v(\tau_j \gamma_N) w_j w_j^* \right)^{-1} w_i - m(0) \right| \xrightarrow{\text{a.s.}} 0$$

  with $m(0)$ unique positive solution to **[MarPas’67; BaiSil’95]**

$$m(0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\tau_i v(\tau_i \gamma_N)}{1 + c\tau_i v(\tau_i \gamma_N) m(0)}.$$

- γ_N precisely solves this equation, thus $m(0) = \gamma_N$.

Proof of the Key Lemma: $\max_i |e_i - 1| \xrightarrow{\text{a.s.}} 0$, $e_i = \frac{v(\tau_i d_i)}{v(\tau_i \gamma_N)}$

Substitution Trick (case $\tau_i \in [a,b] \subset (0,\infty)$)

Up to relabelling $e_1 \leq \ldots \leq e_n$, use

$$v(\tau_n \gamma_N) e_n = v(\tau_n d_n) = v\left(\tau_n \frac{1}{N} w_n^* \left(\frac{1}{n}\sum_{i<n} \tau_i \underbrace{v(\tau_i d_i)}_{=v(\tau_i\gamma_N)e_i} w_i w_i^*\right)^{-1} w_n\right)$$

$$\leq v\left(\tau_n e_n^{-1} \frac{1}{N} w_n^* \left(\frac{1}{n}\sum_{i<n} \tau_i v(\tau_i \gamma_N) w_i w_i^*\right)^{-1} w_n\right)$$

$$\leq v\left(\tau_n e_n^{-1}(\gamma_N - \varepsilon_n)\right) \text{ a.s., } \varepsilon_n \to 0 \text{ (slow).}$$

Proof of the Key Lemma: $\max_i |e_i - 1| \xrightarrow{\text{a.s.}} 0$, $e_i = \frac{v(\tau_i d_i)}{v(\tau_i \gamma_N)}$

Substitution Trick (case $\tau_i \in [a,b] \subset (0,\infty)$)

Up to relabelling $e_1 \leq \ldots \leq e_n$, use

$$v(\tau_n \gamma_N) e_n = v(\tau_n d_n) = v\left(\tau_n \frac{1}{N} w_n^* \left(\frac{1}{n} \sum_{i<n} \tau_i \underbrace{v(\tau_i d_i)}_{= v(\tau_i \gamma_N) e_i} w_i w_i^* \right)^{-1} w_n \right)$$

$$\leq v\left(\tau_n e_n^{-1} \frac{1}{N} w_n^* \left(\frac{1}{n} \sum_{i<n} \tau_i v(\tau_i \gamma_N) w_i w_i^* \right)^{-1} w_n \right)$$

$$\leq v\left(\tau_n e_n^{-1} (\gamma_N - \varepsilon_n) \right) \text{ a.s., } \varepsilon_n \to 0 \text{ (slow).}$$

Use properties of ψ to get

$$\psi\left(\tau_n \gamma_N\right) \leq \psi\left(\tau_n e_n^{-1} \gamma_N\right) \left(1 - \varepsilon_n \gamma_N^{-1}\right)^{-1}$$

Proof of the Key Lemma: $\max_i |e_i - 1| \xrightarrow{\text{a.s.}} 0$, $e_i = \frac{v(\tau_i d_i)}{v(\tau_i \gamma_N)}$

Substitution Trick (case $\tau_i \in [a,b] \subset (0,\infty)$)

Up to relabelling $e_1 \leq \ldots \leq e_n$, use

$$v(\tau_n\gamma_N)e_n = v(\tau_n d_n) = v\left(\tau_n \frac{1}{N} w_n^* \left(\frac{1}{n}\sum_{i<n} \tau_i \underbrace{v(\tau_i d_i)}_{=v(\tau_i\gamma_N)e_i} w_i w_i^*\right)^{-1} w_n\right)$$

$$\leq v\left(\tau_n e_n^{-1} \frac{1}{N} w_n^* \left(\frac{1}{n}\sum_{i<n} \tau_i v(\tau_i\gamma_N) w_i w_i^*\right)^{-1} w_n\right)$$

$$\leq v\left(\tau_n e_n^{-1}(\gamma_N - \varepsilon_n)\right) \text{ a.s., } \varepsilon_n \to 0 \text{ (slow).}$$

Use properties of ψ to get

$$\psi\left(\tau_n\gamma_N\right) \leq \psi\left(\tau_n e_n^{-1}\gamma_N\right)\left(1 - \varepsilon_n\gamma_N^{-1}\right)^{-1}$$

Conclusion: If $e_n > 1 + \ell$ i.o., as $\tau_n \in [a,b]$, on subsequence $\left\{\begin{array}{l}\tau_n \to \tau_0 > 0 \\ \gamma_N \to \gamma_0 > 0\end{array}\right.$,

$$\psi(\tau_0\gamma_0) \leq \psi\left(\frac{\tau_0\gamma_0}{1+\ell}\right), \text{ a contradiction.}$$

Outlier Data

Theorem (Outlier Rejection)

Observation set

$$X = \left[x_1, \ldots, x_{(1-\varepsilon_n)n}, a_1, \ldots, a_{\varepsilon_n n}\right]$$

where $x_i \sim \mathcal{CN}(0, C_N)$ *and* $a_1, \ldots, a_{\varepsilon_n n} \in \mathbb{C}^N$ *deterministic outliers. Then,*

$$\left\| \hat{C}_N - \hat{S}_N \right\| \xrightarrow{\text{a.s.}} 0$$

where

$$\hat{S}_N \triangleq v\left(\gamma_N\right) \frac{1}{n} \sum_{i=1}^{(1-\varepsilon_n)n} x_i x_i^* + \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} v\left(\alpha_{i,n}\right) a_i a_i^*$$

with γ_N *and* $\alpha_{1,n}, \ldots, \alpha_{\varepsilon_n n,n}$ *unique positive solutions to*

$$\gamma_N = \frac{1}{N} tr C_N \left(\frac{(1-\varepsilon)v(\gamma_N)}{1+cv(\gamma_N)\gamma_N} C_N + \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} v\left(\alpha_{i,n}\right) a_i a_i^* \right)^{-1}$$

$$\alpha_{i,n} = \frac{1}{N} a_i^* \left(\frac{(1-\varepsilon)v(\gamma_N)}{1+cv(\gamma_N)\gamma_N} C_N + \frac{1}{n} \sum_{j\neq i}^{\varepsilon_n n} v\left(\alpha_{j,n}\right) a_j a_j^* \right)^{-1} a_i, \;\; i = 1, \ldots, \varepsilon_n n.$$

Outlier Data

- For $\varepsilon_n n = 1$,

$$\hat{S}_N = v\left(\frac{\phi^{-1}(1)}{1-c}\right)\frac{1}{n}\sum_{i=1}^{n-1} x_i x_i^* + \left(v\left(\frac{\phi^{-1}(1)}{1-c}\frac{1}{N}a_1^* C_N^{-1} a_1\right) + o(1)\right) a_1 a_1^*$$

Outlier rejection relies on $\frac{1}{N}a_1^* C_N^{-1} a_1 \lessgtr 1$.

Outlier Data

- For $\varepsilon_n n = 1$,

$$\hat{S}_N = v\left(\frac{\phi^{-1}(1)}{1-c}\right)\frac{1}{n}\sum_{i=1}^{n-1} x_i x_i^* + \left(v\left(\frac{\phi^{-1}(1)}{1-c}\frac{1}{N}a_1^* C_N^{-1} a_1\right) + o(1)\right) a_1 a_1^*$$

  Outlier rejection relies on $\frac{1}{N}a_1^* C_N^{-1} a_1 \lessgtr 1$.

- For $a_i \sim \mathcal{CN}(0, D_N)$, $\varepsilon_n \to \varepsilon \geq 0$,

$$\hat{S}_N = v(\gamma_n)\frac{1}{n}\sum_{i=1}^{(1-\varepsilon_n)n} x_i x_i^* + v(\alpha_n)\frac{1}{n}\sum_{i=1}^{\varepsilon_n n} a_i a_i^*$$

$$\gamma_n = \frac{1}{N}\mathrm{tr}\, C_N\left(\frac{(1-\varepsilon)v(\gamma_n)}{1+cv(\gamma_n)\gamma_n}C_N + \frac{\varepsilon v(\alpha_n)}{1+cv(\alpha_n)\alpha_n}D_N\right)^{-1}$$

$$\alpha_n = \frac{1}{N}\mathrm{tr}\, D_N\left(\frac{(1-\varepsilon)v(\gamma_n)}{1+cv(\gamma_n)\gamma_n}C_N + \frac{\varepsilon v(\alpha_n)}{1+cv(\alpha_n)\alpha_n}D_N\right)^{-1}.$$

Outlier Data

- For $\varepsilon_n n = 1$,

$$\hat{S}_N = v\left(\frac{\phi^{-1}(1)}{1-c}\right)\frac{1}{n}\sum_{i=1}^{n-1} x_i x_i^* + \left(v\left(\frac{\phi^{-1}(1)}{1-c}\frac{1}{N}a_1^* C_N^{-1} a_1\right) + o(1)\right) a_1 a_1^*$$

  Outlier rejection relies on $\frac{1}{N}a_1^* C_N^{-1} a_1 \lessgtr 1$.

- For $a_i \sim \mathcal{CN}(0, D_N)$, $\varepsilon_n \to \varepsilon \geq 0$,

$$\hat{S}_N = v(\gamma_n)\frac{1}{n}\sum_{i=1}^{(1-\varepsilon_n)n} x_i x_i^* + v(\alpha_n)\frac{1}{n}\sum_{i=1}^{\varepsilon_n n} a_i a_i^*$$

$$\gamma_n = \frac{1}{N}\operatorname{tr} C_N\left(\frac{(1-\varepsilon)v(\gamma_n)}{1+cv(\gamma_n)\gamma_n}C_N + \frac{\varepsilon v(\alpha_n)}{1+cv(\alpha_n)\alpha_n}D_N\right)^{-1}$$

$$\alpha_n = \frac{1}{N}\operatorname{tr} D_N\left(\frac{(1-\varepsilon)v(\gamma_n)}{1+cv(\gamma_n)\gamma_n}C_N + \frac{\varepsilon v(\alpha_n)}{1+cv(\alpha_n)\alpha_n}D_N\right)^{-1}.$$

  For $\varepsilon_n \to 0$,

$$\hat{S}_N = v\left(\frac{\phi^{-1}(1)}{1-c}\right)\frac{1}{n}\sum_{i=1}^{(1-\varepsilon_n)n} x_i x_i^* + \frac{1}{n}\sum_{i=1}^{\varepsilon_n n} v\left(\frac{\phi^{-1}(1)}{1-c}\frac{1}{N}\operatorname{tr} D_N C_N^{-1}\right) a_i a_i^*$$

  Outlier rejection relies on $\frac{1}{N}\operatorname{tr} D_N C_N^{-1} \lessgtr 1$.

Outlier Data

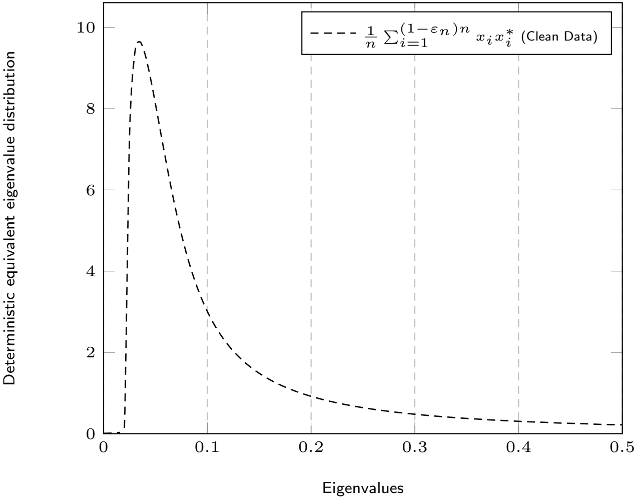


Figure: Limiting eigenvalue distributions. $[C_N]_{ij} = .9^{|i-j|}$, $D_N = I_N$, $\varepsilon = .05$.

Outlier Data

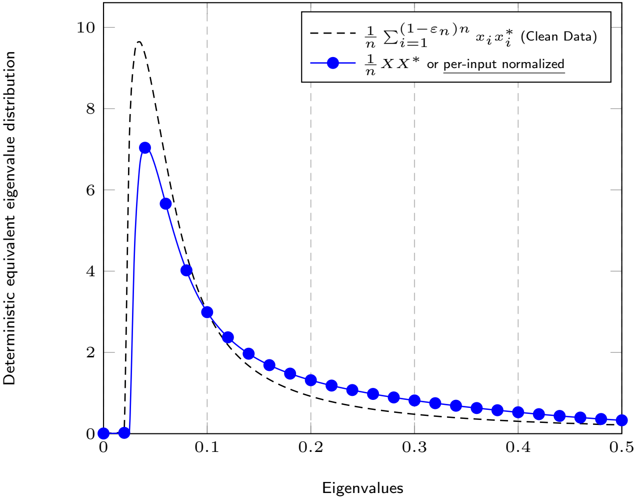


Figure: Limiting eigenvalue distributions. $[C_N]_{ij} = .9^{|i-j|}$, $D_N = I_N$, $\varepsilon = .05$.

Outlier Data

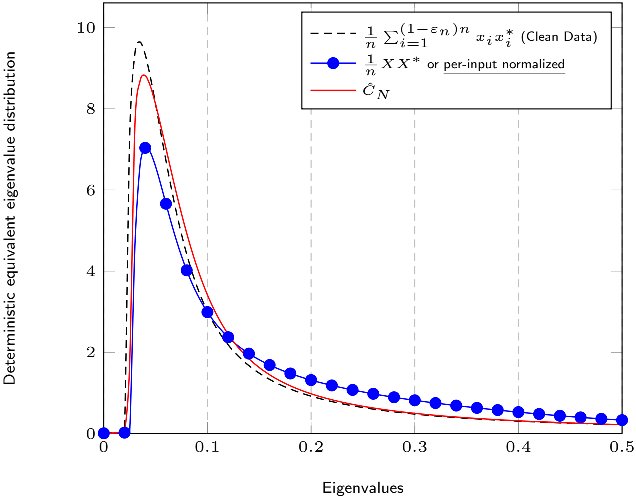
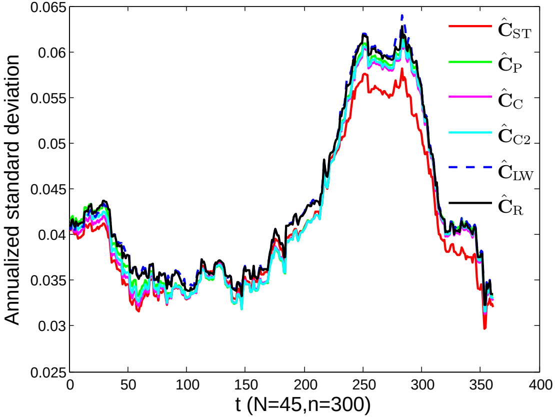


Figure: Limiting eigenvalue distributions. $[C_N]_{ij} = .9^{|i-j|}$, $D_N = I_N$, $\varepsilon = .05$.

Example of application to statistical finance

- Robust matrix-optimized portfolio allocation $\hat{\mathbf{C}}_{\mathrm{ST}}$



Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

Reminder on Spectral Clustering Methods

Context: Two-step classification of n objects based on similarity $A \in \mathbb{R}^{n\times n}$:

1. extraction of eigenvectors $U = [u_1, \ldots, u_\ell]$ with “dominant” eigenvalues

Reminder on Spectral Clustering Methods

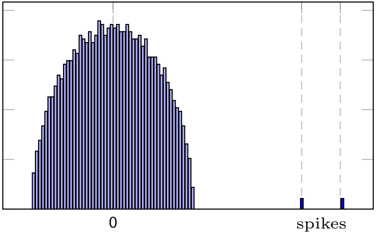
Context: Two-step classification of n objects based on similarity $A \in \mathbb{R}^{n \times n}$:

1. extraction of eigenvectors $U = [u_1, \ldots, u_\ell]$ with “dominant” eigenvalues
2. classification of vectors $U_{1,\cdot}, \ldots, U_{n,\cdot} \in \mathbb{R}^\ell$ using k-means/EM.

Reminder on Spectral Clustering Methods

Context: Two-step classification of n objects based on similarity $A \in \mathbb{R}^{n \times n}$:

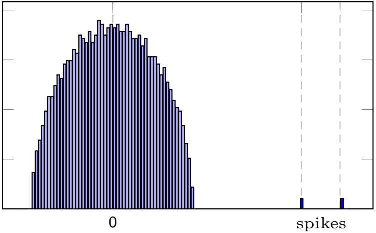
1. extraction of eigenvectors $U = [u_1, \ldots, u_\ell]$ with “dominant” eigenvalues
2. classification of vectors $U_{1,\cdot}, \ldots, U_{n,\cdot} \in \mathbb{R}^\ell$ using k-means/EM.



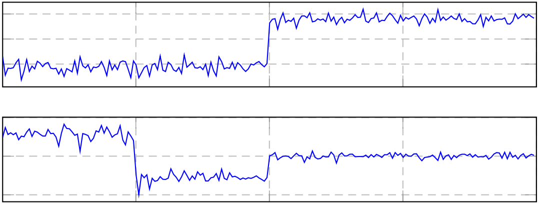
Reminder on Spectral Clustering Methods

Context: Two-step classification of n objects based on similarity $A \in \mathbb{R}^{n\times n}$:

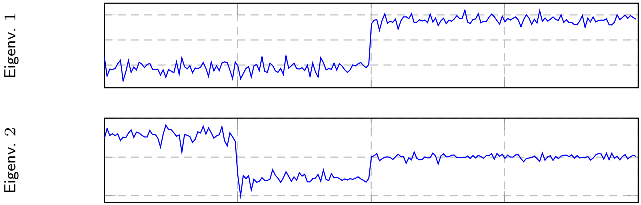
1. extraction of eigenvectors $U = [u_1, \ldots, u_\ell]$ with "dominant" eigenvalues
2. classification of vectors $U_{1,\cdot}, \ldots, U_{n,\cdot} \in \mathbb{R}^\ell$ using k-means/EM.



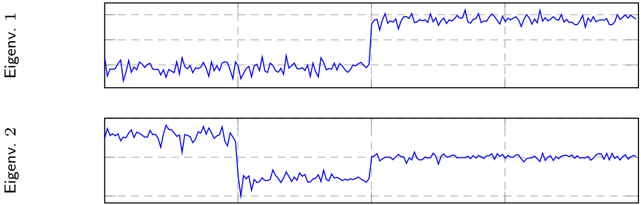
$\Downarrow$ **Eigenvectors** $\Downarrow$
(in practice, shuffled!!)



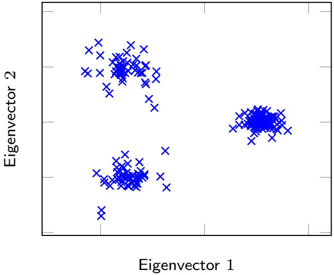
Reminder on Spectral Clustering Methods



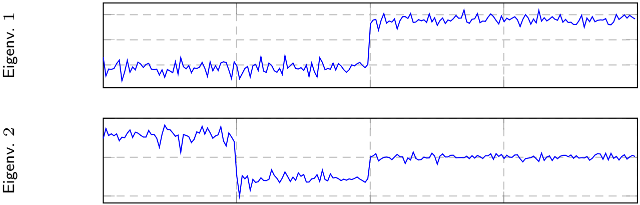
Reminder on Spectral Clustering Methods



⇓ **ℓ-dimensional representation** ⇓
(shuffling no longer matters!)

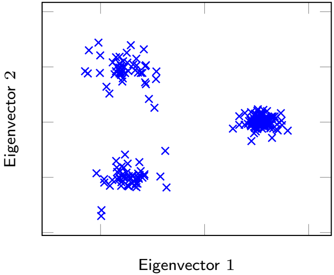


Reminder on Spectral Clustering Methods



⇓ **ℓ-dimensional representation** ⇓

(shuffling no longer matters!)



⇓

EM or k-means clustering.

The Random Matrix Approach

A two-step method:

1. If A_n is not a “standard” random matrix, retrieve $\tilde{A}_n$ such that

$$\left\|A_n - \tilde{A}_n\right\| \xrightarrow{\text{a.s.}} 0$$

   in operator norm as $n \to \infty$.

The Random Matrix Approach

A two-step method:

1. If A_n is not a "standard" random matrix, retrieve $\tilde{A}_n$ such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as $n \to \infty$.

$\Rightarrow$ Transfers crucial properties from A_n to $\tilde{A}_n$:

The Random Matrix Approach

A two-step method:

1. If A_n is not a “standard” random matrix, retrieve $\tilde{A}_n$ such that

$$\left\|A_n - \tilde{A}_n\right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as $n \to \infty$.

$\Rightarrow$ Transfers crucial properties from A_n to $\tilde{A}_n$:

- limiting eigenvalue distribution

The Random Matrix Approach

A two-step method:

1. If A_n is not a “standard” random matrix, retrieve $\tilde{A}_n$ such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as $n \to \infty$.

$\Rightarrow$ Transfers crucial properties from A_n to $\tilde{A}_n$:

- limiting eigenvalue distribution
- spikes

The Random Matrix Approach

A two-step method:

1. If A_n is not a "standard" random matrix, retrieve $\tilde{A}_n$ such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as $n \to \infty$.

$\Rightarrow$ Transfers crucial properties from A_n to $\tilde{A}_n$:

- limiting eigenvalue distribution
- spikes
- eigenvectors of isolated eigenvalues.

The Random Matrix Approach

A two-step method:

1. If A_n is not a “standard” random matrix, retrieve $\tilde{A}_n$ such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

   in operator norm as $n \to \infty$.

   $\Rightarrow$ Transfers crucial properties from A_n to $\tilde{A}_n$:

   - limiting eigenvalue distribution
   - spikes
   - eigenvectors of isolated eigenvalues.

2. From $\tilde{A}_n$, perform spiked model analysis:

The Random Matrix Approach

A two-step method:

1. If A_n is not a "standard" random matrix, retrieve $\tilde{A}_n$ such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

   in operator norm as $n \to \infty$.

   $\Rightarrow$ Transfers crucial properties from A_n to $\tilde{A}_n$:

   - limiting eigenvalue distribution
   - spikes
   - eigenvectors of isolated eigenvalues.

2. From $\tilde{A}_n$, perform spiked model analysis:

   - exhibit phase transition phenomenon

The Random Matrix Approach

A two-step method:

1. If A_n is not a "standard" random matrix, retrieve $\tilde{A}_n$ such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as $n \to \infty$.

$\Rightarrow$ Transfers crucial properties from A_n to $\tilde{A}_n$:

- limiting eigenvalue distribution
- spikes
- eigenvectors of isolated eigenvalues.

2. From $\tilde{A}_n$, perform spiked model analysis:

- exhibit phase transition phenomenon
- "read" the content of isolated eigenvectors of $\tilde{A}_n$.

The Random Matrix Approach

The Spike Analysis:
For "noisy plateaus"-looking isolated eigenvectors $u_1,\ldots,u_\ell$ of $\tilde{A}_n$, write

$$u_i = \sum_{a=1}^{k} \alpha_i^a \frac{j_a}{\sqrt{n_a}} + \sigma_i^a w_i^a$$

with $j_a \in \mathbb{R}^n$ canonical vector of class $\mathcal{C}_a$, w_i^a noise orthogonal to j_a,

The Random Matrix Approach

The Spike Analysis:
For "noisy plateaus"-looking isolated eigenvectors $u_1, \ldots, u_\ell$ of $\tilde{A}_n$, write

$$u_i = \sum_{a=1}^{k} \alpha_i^a \frac{j_a}{\sqrt{n_a}} + \sigma_i^a w_i^a$$

with $j_a \in \mathbb{R}^n$ canonical vector of class $\mathcal{C}_a$, w_i^a noise orthogonal to j_a, and evaluate

$$\alpha_i^a = \frac{1}{\sqrt{n_a}} u_i^\mathsf{T} j_a$$
$$(\sigma_i^a)^2 = \left\| u_i - \alpha_i^a \frac{j_a}{\sqrt{n_a}} \right\|^2.$$

The Random Matrix Approach

The Spike Analysis:
For "noisy plateaus"-looking isolated eigenvectors $u_1, \ldots, u_\ell$ of $\tilde{A}_n$, write

$$u_i = \sum_{a=1}^{k} \alpha_i^a \frac{j_a}{\sqrt{n_a}} + \sigma_i^a w_i^a$$

with $j_a \in \mathbb{R}^n$ canonical vector of class $\mathcal{C}_a$, w_i^a noise orthogonal to j_a, and evaluate

$$\alpha_i^a = \frac{1}{\sqrt{n_a}} u_i^{\mathsf{T}} j_a$$

$$(\sigma_i^a)^2 = \left\| u_i - \alpha_i^a \frac{j_a}{\sqrt{n_a}} \right\|^2 .$$

$\Longrightarrow$ Can be done using complex analysis calculus, e.g.

$$\begin{aligned} (\alpha_i^a)^2 &= \frac{1}{n_a} j_a^{\mathsf{T}} u_i u_i^{\mathsf{T}} j_a \\ &= \frac{1}{2\pi\imath} \oint_{\gamma_a} \frac{1}{n_a} j_a^{\mathsf{T}} \left(\tilde{A}_n - zI_n \right)^{-1} j_a dz. \end{aligned}$$

Outline

Basics of Random Matrix Theory

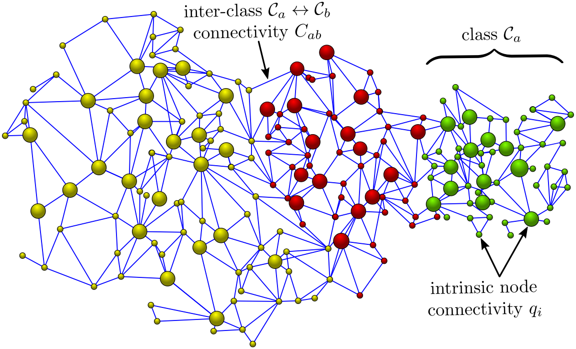
- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

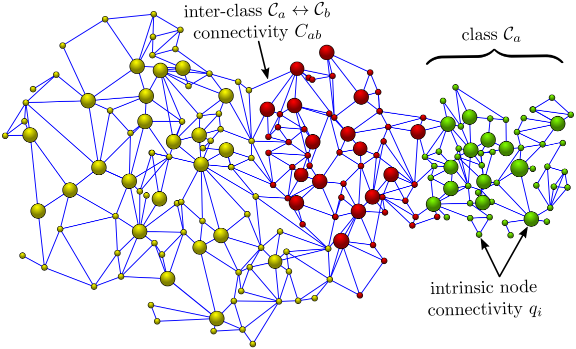
System Setting



Assume n-node, m-edges **undirected** graph G, with

- "intrinsic" average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.

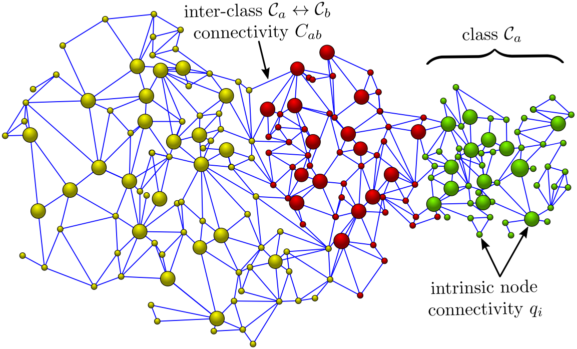
System Setting



Assume n-node, m-edges **undirected** graph G, with

- "intrinsic" average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.
- k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ independent of $\{q_i\}$ of (large) sizes $n_1, \ldots, n_k$, with preferential attachment C_{ab} between $\mathcal{C}_a$ and $\mathcal{C}_b$

System Setting



inter-class $\mathcal{C}_a \leftrightarrow \mathcal{C}_b$ connectivity C_{ab}

class $\mathcal{C}_a$

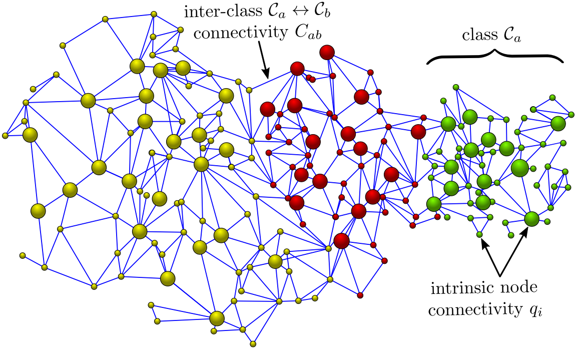
intrinsic node connectivity q_i

Assume n-node, m-edges **undirected** graph G, with

- "intrinsic" average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.
- k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ independent of $\{q_i\}$ of (large) sizes $n_1, \ldots, n_k$, with preferential attachment C_{ab} between $\mathcal{C}_a$ and $\mathcal{C}_b$
- induces edge probability for node $i \in \mathcal{C}_a$, $j \in \mathcal{C}_b$,

$$P(i \sim j) = q_i q_j C_{ab}.$$

System Setting



Assume n-node, m-edges **undirected** graph G, with

- "intrinsic" average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.
- k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ independent of $\{q_i\}$ of (large) sizes $n_1, \ldots, n_k$, with preferential attachment C_{ab} between $\mathcal{C}_a$ and $\mathcal{C}_b$
- induces edge probability for node $i \in \mathcal{C}_a$, $j \in \mathcal{C}_b$,

$$P(i \sim j) = q_i q_j C_{ab}.$$

- adjacency matrix A with $A_{ij} \sim \text{Bernoulli}(q_i q_j C_{ab})$.

Objective

Study of spectral methods:

- standard methods based on adjacency A, modularity $A - \frac{dd^{\mathsf{T}}}{2m}$, normalized adjacency $D^{-1}AD^{-1}$, etc. (adapted to **dense nets**)
- refined methods based on Bethe Hessian $(r^2-1)I_n - rA + D$ (adapted to **sparse nets**!)

Objective

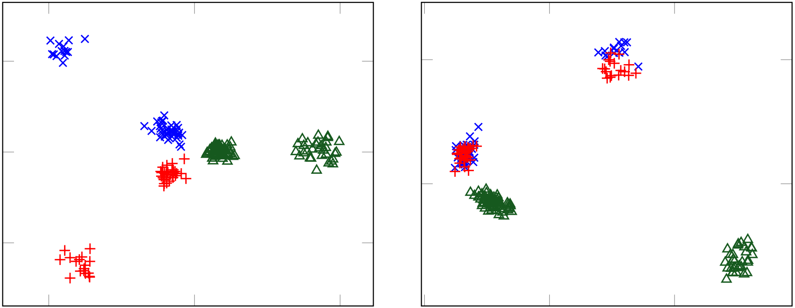
Study of spectral methods:

- standard methods based on adjacency A, modularity $A - \frac{dd^{\mathsf{T}}}{2m}$, normalized adjacency $D^{-1}AD^{-1}$, etc. (adapted to **dense nets**)
- refined methods based on Bethe Hessian $(r^2-1)I_n - rA + D$ (adapted to **sparse nets**!)

Improvement to realistic graphs:

- observation of failure of standard methods above
- improvement by new methods.

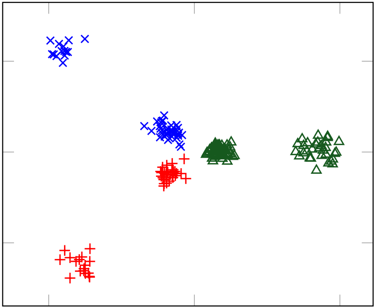
Limitations of Adjacency/Modularity Approach



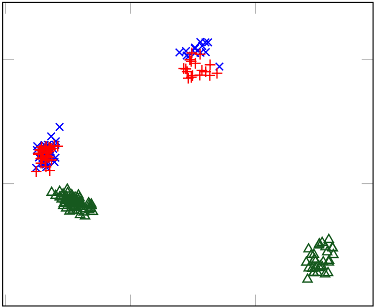
(Modularity)

(Bethe Hessian)

Limitations of Adjacency/Modularity Approach



(Modularity)



(Bethe Hessian)

Scenario: 3 classes with μ bi-modal (e.g., $\mu = \frac{3}{4}\delta_{0.1} + \frac{1}{4}\delta_{0.5}$)

→ Leading eigenvectors of A (or modularity $A - \frac{dd^{\mathsf{T}}}{2m}$) biased by q_i distribution.
→ Similar behavior for Bethe Hessian.

Regularized Modularity Approach

Connectivity Model: $P(i \sim j) = q_i q_j C_{ab}$ for $i \in \mathcal{C}_a$, $j \in \mathcal{C}_b$.

Dense Regime Assumptions: Non trivial regime when, as $n \to \infty$,

$$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}},\ M_{ab} = O(1).$$

Regularized Modularity Approach

Connectivity Model: $P(i \sim j) = q_i q_j C_{ab}$ for $i \in \mathcal{C}_a$, $j \in \mathcal{C}_b$.

Dense Regime Assumptions: Non trivial regime when, as $n \to \infty$,

$$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}, \; M_{ab} = O(1).$$

$\Rightarrow$ Community information is **weak but highly REDUNDANT!**

Regularized Modularity Approach

Connectivity Model: $P(i \sim j) = q_i q_j C_{ab}$ for $i \in \mathcal{C}_a$, $j \in \mathcal{C}_b$.

Dense Regime Assumptions: Non trivial regime when, as $n \to \infty$,

$$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}, \;\; M_{ab} = O(1).$$

$\Rightarrow$ Community information is **weak but highly REDUNDANT!**

Considered Matrix:
For $\alpha \in [0,1]$, (and with $D = \operatorname{diag}(A1_n) = \operatorname{diag}(d)$ the degree matrix), $m = \frac{1}{2} d^\mathsf{T} 1$ the number of edges

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{2m} \right] D^{-\alpha}.$$

Regularized Modularity Approach

Connectivity Model: $P(i \sim j) = q_i q_j C_{ab}$ for $i \in \mathcal{C}_a$, $j \in \mathcal{C}_b$.

Dense Regime Assumptions: Non trivial regime when, as $n \to \infty$,

$$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}},\ M_{ab} = O(1).$$

$\Rightarrow$ Community information is **weak but highly REDUNDANT!**

Considered Matrix:
For $\alpha \in [0,1]$, (and with $D = \operatorname{diag}(A1_n) = \operatorname{diag}(d)$ the degree matrix), $m = \frac{1}{2}d^\mathsf{T}1$ the number of edges

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{2m} \right] D^{-\alpha}.$$

Our results in a nutshell:

- we find optimal $\alpha_{\rm opt}$ having best phase transition.

Regularized Modularity Approach

Connectivity Model: $P(i \sim j) = q_i q_j C_{ab}$ for $i \in \mathcal{C}_a$, $j \in \mathcal{C}_b$.

Dense Regime Assumptions: Non trivial regime when, as $n \to \infty$,

$$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}},\ M_{ab} = O(1).$$

$\Rightarrow$ Community information is **weak but highly REDUNDANT!**

Considered Matrix:
For $\alpha \in [0,1]$, (and with $D = \operatorname{diag}(A1_n) = \operatorname{diag}(d)$ the degree matrix), $m = \frac{1}{2}d^{\mathsf{T}}1$ the number of edges

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[A - \frac{dd^{\mathsf{T}}}{2m}\right] D^{-\alpha}.$$

Our results in a nutshell:

- we find optimal $\alpha_{\rm opt}$ having best phase transition.
- we find consistent estimator $\hat{\alpha}_{\rm opt}$ from A alone.

Regularized Modularity Approach

Connectivity Model: $P(i \sim j) = q_i q_j C_{ab}$ for $i \in \mathcal{C}_a$, $j \in \mathcal{C}_b$.

Dense Regime Assumptions: Non trivial regime when, as $n \to \infty$,

$$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}, \; M_{ab} = O(1).$$

$\Rightarrow$ Community information is **weak but highly REDUNDANT!**

Considered Matrix:
For $\alpha \in [0,1]$, (and with $D = \operatorname{diag}(A1_n) = \operatorname{diag}(d)$ the degree matrix), $m = \frac{1}{2} d^{\mathsf{T}} 1$ the number of edges

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[A - \frac{dd^{\mathsf{T}}}{2m} \right] D^{-\alpha}.$$

Our results in a nutshell:

- we find optimal $\alpha_{\rm opt}$ having best phase transition.
- we find consistent estimator $\hat{\alpha}_{\rm opt}$ from A alone.
- we claim optimal eigenvector regularization $D^{\alpha-1}u$, u eigenvector of L_α.

Asymptotic Equivalence

Theorem (Limiting Random Matrix Equivalent)

For each $\alpha \in [0,1]$, *as* $n \to \infty$, $\|L_\alpha - \tilde{L}_\alpha\| \to 0$ *almost surely, where*

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{2m} \right] D^{-\alpha}$$

$$\tilde{L}_\alpha = \frac{1}{\sqrt{n}} D_q^{-\alpha} X D_q^{-\alpha} + U \Lambda U^\mathsf{T}$$

with $D_q = \operatorname{diag}(\{q_i\})$, X *zero-mean random matrix,*

$$U = \begin{bmatrix} D_q^{1-\alpha} \frac{J}{\sqrt{n}} & D_q^{-\alpha} X 1_n \end{bmatrix}, \quad \textit{rank } k+1$$

$$\Lambda = \begin{bmatrix} (I_k - 1_k c^\mathsf{T}) M (I_k - c 1_k^\mathsf{T}) & -1_k \\ 1_k^\mathsf{T} & 0 \end{bmatrix}$$

and $J = [j_1, \ldots, j_k]$, $j_a = [0, \ldots, 0, 1_{n_a}^\mathsf{T}, 0, \ldots, 0]^\mathsf{T} \in \mathbb{R}^n$ *canonical vector of class* $\mathcal{C}_a$.

Asymptotic Equivalence

Theorem (Limiting Random Matrix Equivalent)

For each $\alpha \in [0,1]$, as $n \to \infty$, $\|L_\alpha - \tilde{L}_\alpha\| \to 0$ almost surely, where

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{2m} \right] D^{-\alpha}$$

$$\tilde{L}_\alpha = \frac{1}{\sqrt{n}} D_q^{-\alpha} X D_q^{-\alpha} + U \Lambda U^\mathsf{T}$$

with $D_q = \operatorname{diag}(\{q_i\})$, X zero-mean random matrix,

$$U = \begin{bmatrix} D_q^{1-\alpha} \frac{J}{\sqrt{n}} & D_q^{-\alpha} X 1_n \end{bmatrix}, \quad \textit{rank } k+1$$

$$\Lambda = \begin{bmatrix} (I_k - 1_k c^\mathsf{T}) M (I_k - c 1_k^\mathsf{T}) & -1_k \\ 1_k^\mathsf{T} & 0 \end{bmatrix}$$

and $J = [j_1, \ldots, j_k]$, $j_a = [0, \ldots, 0, 1_{n_a}^\mathsf{T}, 0, \ldots, 0]^\mathsf{T} \in \mathbb{R}^n$ canonical vector of class $\mathcal{C}_a$.

Consequences:

- isolated eigenvalues beyond phase transition ↔ $\lambda(M) >$ "spectrum edge"
  ⇒ **optimal choice $\alpha_{\rm opt}$ of α** from study of noise spectrum.

Asymptotic Equivalence

Theorem (Limiting Random Matrix Equivalent)

For each $\alpha \in [0,1]$, *as* $n \to \infty$, $\|L_\alpha - \tilde{L}_\alpha\| \to 0$ *almost surely, where*

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{2m} \right] D^{-\alpha}$$

$$\tilde{L}_\alpha = \frac{1}{\sqrt{n}} D_q^{-\alpha} X D_q^{-\alpha} + U \Lambda U^\mathsf{T}$$

with $D_q = \operatorname{diag}(\{q_i\})$, X *zero-mean random matrix,*

$$U = \begin{bmatrix} D_q^{1-\alpha} \frac{J}{\sqrt{n}} & D_q^{-\alpha} X 1_n \end{bmatrix}, \quad \textit{rank } k+1$$

$$\Lambda = \begin{bmatrix} (I_k - 1_k c^\mathsf{T}) M (I_k - c 1_k^\mathsf{T}) & -1_k \\ 1_k^\mathsf{T} & 0 \end{bmatrix}$$

and $J = [j_1, \ldots, j_k]$, $j_a = [0, \ldots, 0, 1_{n_a}^\mathsf{T}, 0, \ldots, 0]^\mathsf{T} \in \mathbb{R}^n$ *canonical vector of class* $\mathcal{C}_a$.

Consequences:

- isolated eigenvalues beyond phase transition ↔ $\lambda(M) >$ "spectrum edge"
  ⇒ **optimal choice** $\alpha_{\rm opt}$ **of** α from study of noise spectrum.
- **eigenvectors correlated to** $D_q^{1-\alpha} J$
  ⇒ **Natural regularization by** $D^{\alpha-1}$**!**

Eigenvalue Spectrum

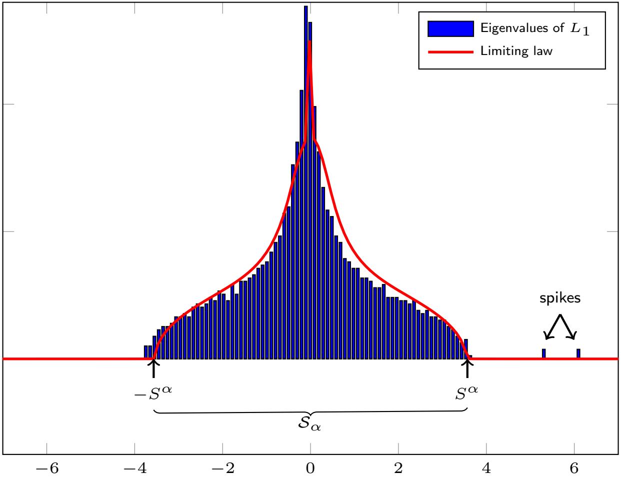


Figure: Eigenvalues of L_1, $K = 3$, $n = 2000$, $c_1 = 0.3, c_2 = 0.3, c_3 = 0.4$, $\mu = \frac{1}{2}\delta_{q_{(1)}} + \frac{1}{2}\delta_{q_{(2)}}$, $q_{(1)} = 0.4$, $q_{(2)} = 0.9$, M defined by $M_{ii} = 12$, $M_{ij} = -4, i \neq j$.

Phase Transition

Theorem (Phase Transition)

For $\alpha \in [0,1]$, isolated eigenvalue $\lambda_i(L_\alpha)$ if $|\lambda_i(\bar{M})| > \tau^\alpha$, $\bar{M} = (\mathcal{D}(c) - cc^\mathsf{T})M$,

$$\tau^\alpha = \lim_{x \downarrow S_+^\alpha} -\frac{1}{e_2^\alpha(x)}, \text{ phase transition threshold}$$

with $[S_-^\alpha, S_+^\alpha]$ limiting eigenvalue support of L_α and $e_2^\alpha(x)$ ($|x| > S_+^\alpha$) solution of

$$e_1^\alpha(x) = \int \frac{q^{1-2\alpha}}{-x - q^{1-2\alpha} e_1^\alpha(x) + q^{2-2\alpha} e_2^\alpha(x)} \mu(dq)$$

$$e_2^\alpha(x) = \int \frac{q^{2-2\alpha}}{-x - q^{1-2\alpha} e_1^\alpha(x) + q^{2-2\alpha} e_2^\alpha(x)} \mu(dq).$$

In this case, $-\frac{1}{e_2^\alpha(\lambda_i(L_\alpha))} = \lambda_i(\bar{M})$.

Phase Transition

Theorem (Phase Transition)

For $\alpha \in [0,1]$, isolated eigenvalue $\lambda_i(L_\alpha)$ if $|\lambda_i(\bar{M})| > \tau^\alpha$, $\bar{M} = (\mathcal{D}(c) - cc^\mathsf{T})M$,

$$\tau^\alpha = \lim_{x \downarrow S_+^\alpha} -\frac{1}{e_2^\alpha(x)}, \text{ phase transition threshold}$$

with $[S_-^\alpha, S_+^\alpha]$ limiting eigenvalue support of L_α and $e_2^\alpha(x)$ ($|x| > S_+^\alpha$) solution of

$$e_1^\alpha(x) = \int \frac{q^{1-2\alpha}}{-x - q^{1-2\alpha} e_1^\alpha(x) + q^{2-2\alpha} e_2^\alpha(x)} \mu(dq)$$

$$e_2^\alpha(x) = \int \frac{q^{2-2\alpha}}{-x - q^{1-2\alpha} e_1^\alpha(x) + q^{2-2\alpha} e_2^\alpha(x)} \mu(dq).$$

In this case, $-\frac{1}{e_2^\alpha(\lambda_i(L_\alpha))} = \lambda_i(\bar{M})$.

Clustering still possible when $\lambda_i(\bar{M}) = (\min_\alpha \tau_\alpha) + \varepsilon$.

- **"Optimal"** $\alpha = \alpha_{\rm opt}$:

$$\alpha_{\rm opt} = \operatorname{argmin}_{\alpha \in [0,1]} \{\tau_\alpha\}.$$

Phase Transition

Theorem (Phase Transition)

For $\alpha \in [0,1]$, isolated eigenvalue $\lambda_i(L_\alpha)$ if $|\lambda_i(\bar{M})| > \tau^\alpha$, $\bar{M} = (\mathcal{D}(c) - cc^\mathsf{T})M$,

$$\tau^\alpha = \lim_{x \downarrow S_+^\alpha} -\frac{1}{e_2^\alpha(x)}, \text{ phase transition threshold}$$

with $[S_-^\alpha, S_+^\alpha]$ limiting eigenvalue support of L_α and $e_2^\alpha(x)$ ($|x| > S_+^\alpha$) solution of

$$e_1^\alpha(x) = \int \frac{q^{1-2\alpha}}{-x - q^{1-2\alpha} e_1^\alpha(x) + q^{2-2\alpha} e_2^\alpha(x)} \mu(dq)$$

$$e_2^\alpha(x) = \int \frac{q^{2-2\alpha}}{-x - q^{1-2\alpha} e_1^\alpha(x) + q^{2-2\alpha} e_2^\alpha(x)} \mu(dq).$$

In this case, $-\frac{1}{e_2^\alpha(\lambda_i(L_\alpha))} = \lambda_i(\bar{M})$.

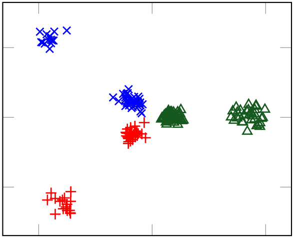
Clustering still possible when $\lambda_i(\bar{M}) = (\min_\alpha \tau_\alpha) + \varepsilon$.

- **"Optimal"** $\alpha = \alpha_{\text{opt}}$:

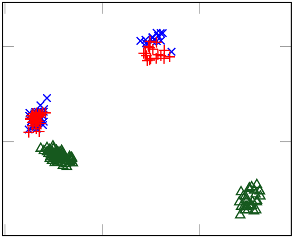
$$\alpha_{\text{opt}} = \operatorname{argmin}_{\alpha \in [0,1]} \{\tau_\alpha\}.$$

- From $\max_i \left| \frac{d_i}{\sqrt{d^\mathsf{T} 1_n}} - q_i \right| \xrightarrow{\text{a.s.}} 0$, we obtain consistent estimator $\hat{\alpha}_{\text{opt}}$ of α_{opt}.

Simulated Performance Results (2 masses of q_i)

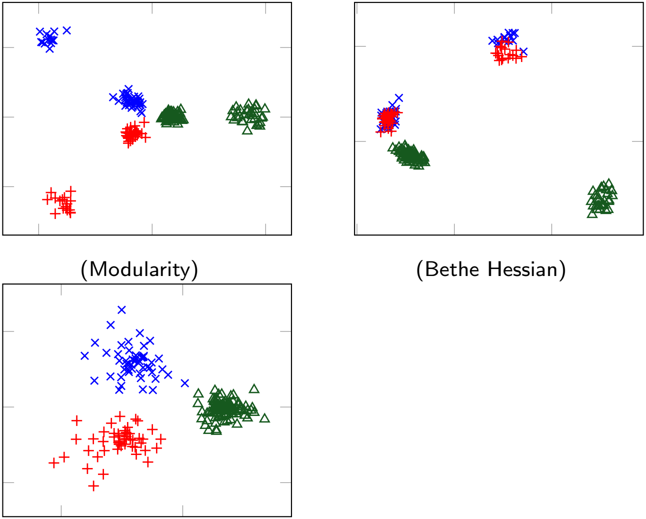


(Modularity)



(Bethe Hessian)

Simulated Performance Results (2 masses of q_i)



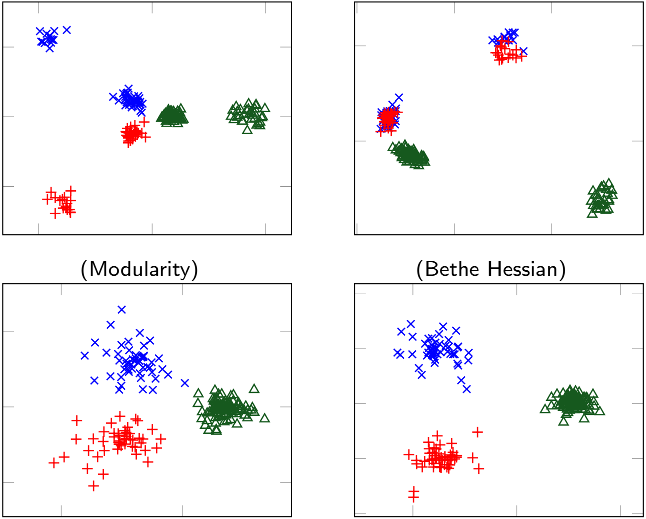
(Modularity)

(Bethe Hessian)

(Algo with $\alpha = 1$)

Figure: Two dominant eigenvectors (x-y axes) for $n = 2000$, $K = 3$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$, $q_{(1)} = 0.1$, $q_{(2)} = 0.5$, $c_1 = c_2 = \frac{1}{4}$, $c_3 = \frac{1}{2}$, $M = 100I_3$.

Simulated Performance Results (2 masses of q_i)



(Modularity)

(Bethe Hessian)

(Algo with $\alpha = 1$)

(Algo with α_{opt})

Figure: Two dominant eigenvectors (x-y axes) for $n = 2000$, $K = 3$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$, $q_{(1)} = 0.1$, $q_{(2)} = 0.5$, $c_1 = c_2 = \frac{1}{4}$, $c_3 = \frac{1}{2}$, $M = 100 I_3$.

Simulated Performance Results (2 masses for q_i)

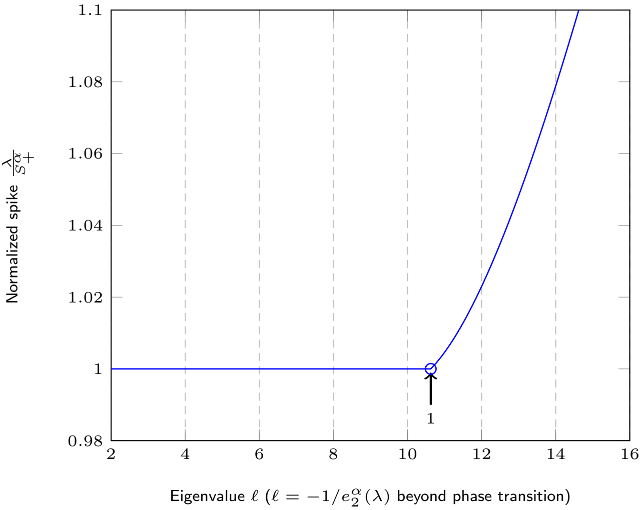


Figure: Largest eigenvalue λ of L_α as a function of the largest eigenvalue ℓ of $(\mathcal{D}(c) - cc^\mathsf{T})M$, for $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, for $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \alpha_{\text{opt}}\}$ (indicated below the graph). Here, $\alpha_{\text{opt}} = 0.07$. Circles indicate phase transition. Beyond phase transition, $\ell = -1/e_2^\alpha(\lambda)$.

Simulated Performance Results (2 masses for q_i)

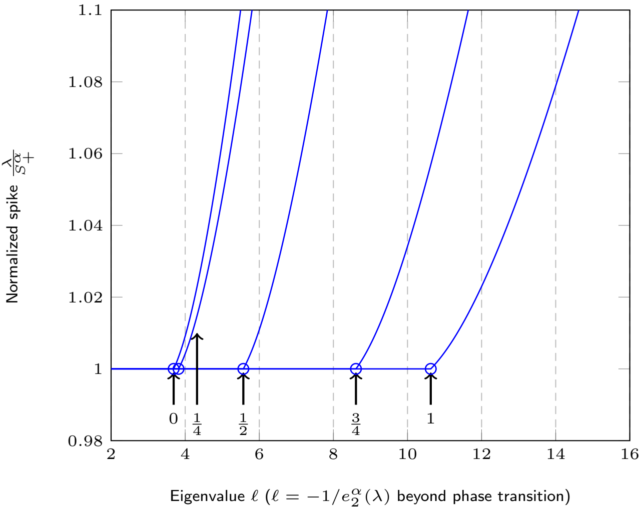


Figure: Largest eigenvalue λ of L_α as a function of the largest eigenvalue ℓ of $(\mathcal{D}(c) - cc^\mathsf{T})M$, for $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, for $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \alpha_{\text{opt}}\}$ (indicated below the graph). Here, $\alpha_{\text{opt}} = 0.07$. Circles indicate phase transition. Beyond phase transition, $\ell = -1/e_2^\alpha(\lambda)$.

Simulated Performance Results (2 masses for q_i)

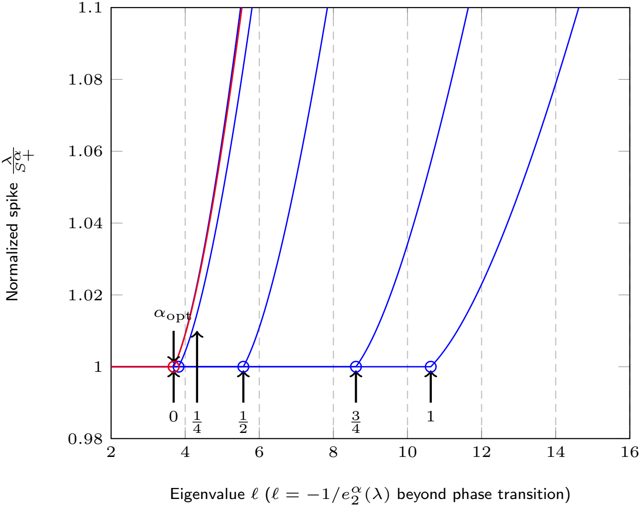


Figure: Largest eigenvalue λ of L_α as a function of the largest eigenvalue ℓ of $(\mathcal{D}(c) - cc^\mathsf{T})M$, for $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, for $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \alpha_{\rm opt}\}$ (indicated below the graph). Here, $\alpha_{\rm opt} = 0.07$. Circles indicate phase transition. Beyond phase transition, $\ell = -1/e_2^\alpha(\lambda)$.

Simulated Performance Results (2 masses for q_i)

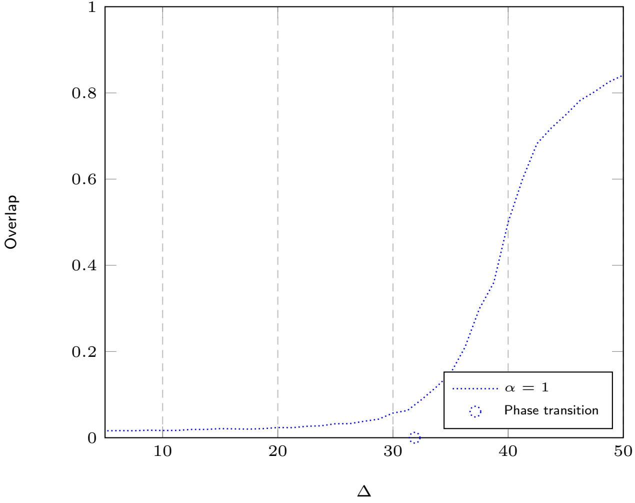


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\rm opt} = 0.07$.

Simulated Performance Results (2 masses for q_i)

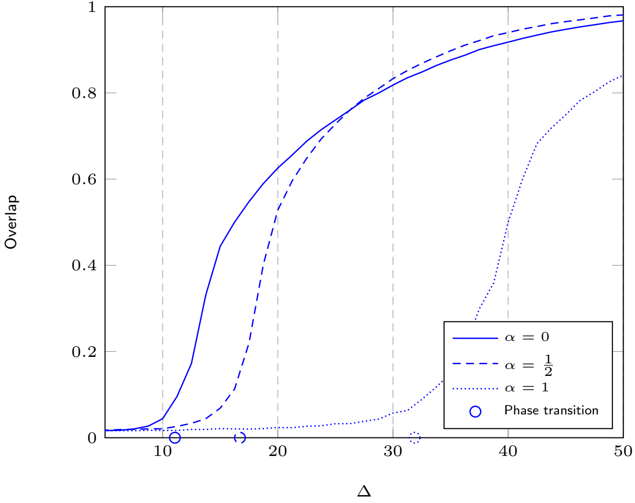


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\rm opt} = 0.07$.

Simulated Performance Results (2 masses for q_i)

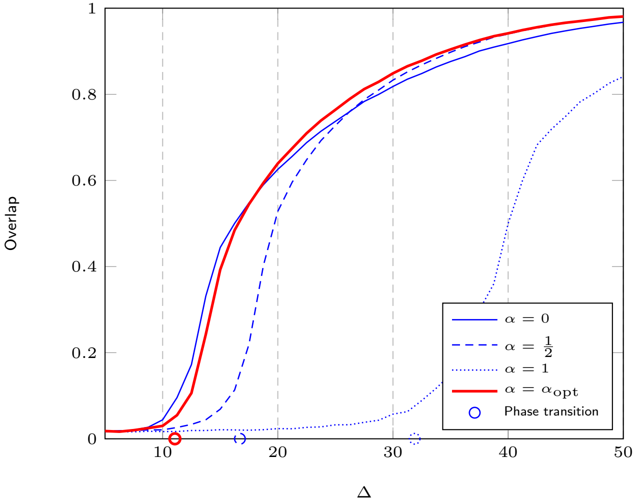


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\text{opt}} = 0.07$.

Simulated Performance Results (2 masses for q_i)

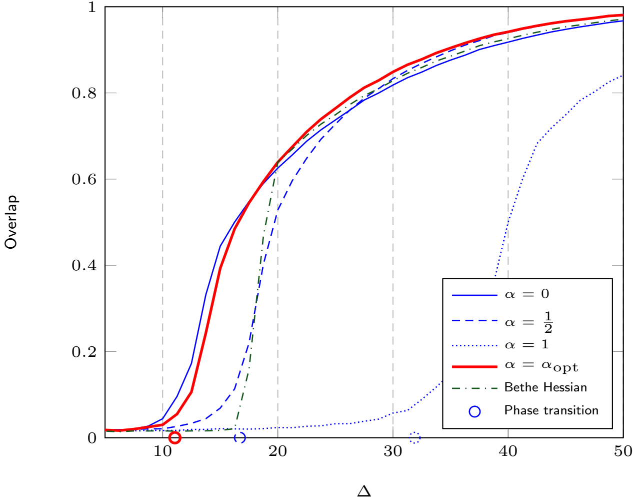


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\text{opt}} = 0.07$.

Simulated Performance Results (2 masses for q_i)

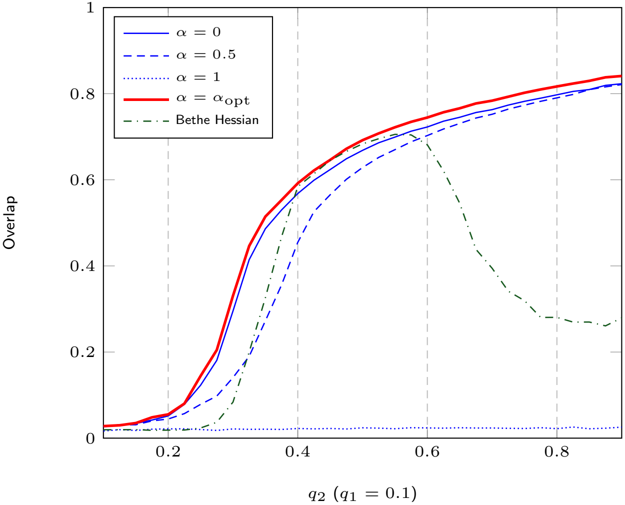


Figure: Overlap performance for $n = 3000$, $K = 3$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} \in [0.1, 0.9]$, $M = 10(2I_3 - 1_3 1_3^{\mathsf{T}})$, $c_i = \frac{1}{3}$.

Theoretical Performance

Analysis of eigenvectors reveals:

- eigenvectors are “noisy staircase vectors”

Theoretical Performance

Analysis of eigenvectors reveals:

- eigenvectors are “noisy staircase vectors”
- conjectured Gaussian fluctuations of eigenvector entries

Theoretical Performance

Analysis of eigenvectors reveals:

- eigenvectors are “noisy staircase vectors”
- conjectured Gaussian fluctuations of eigenvector entries
- for $q_i = q_0$ (homogeneous case), same variance for all entries
- in non-homogeneous case, we can compute “average variance per class”
  $\Rightarrow$ Heuristic asymptotic performance upper-bound using EM.

Theoretical Performance Results (uniform distribution for q_i)

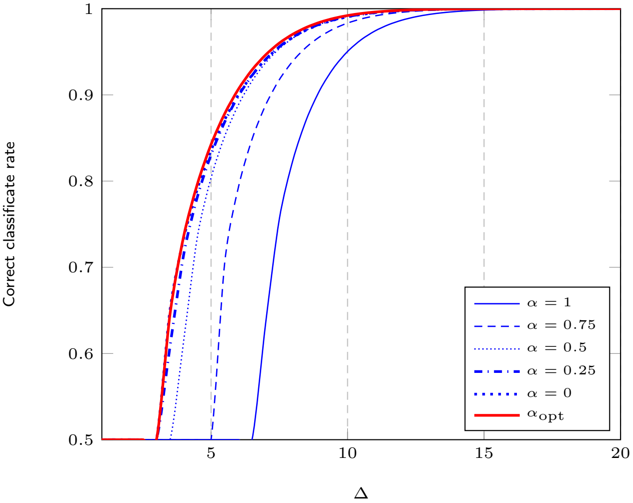


Figure: Theoretical probability of correct recovery for $n = 2000$, $K = 2$, $c_1 = 0.6$, $c_2 = 0.4$, μ uniformly distributed in $[0.2, 0.8]$, $M = \Delta I_2$, for $\Delta \in [0, 20]$.

Some Takeaway messages

Main findings:

Some Takeaway messages

Main findings:

- Degree heterogeneity breaks community structures in eigenvectors.
  ⇒ Compensation by $D^{\alpha-1}$ normalization of eigenvectors.

Some Takeaway messages

Main findings:

- Degree heterogeneity breaks community structures in eigenvectors.
  $\Rightarrow$ Compensation by $D^{\alpha-1}$ normalization of eigenvectors.
- Classical debate over "best normalization" of adjacency (or modularity) matrix A not trivial to solve.
  $\Rightarrow$ With heterogeneous degrees, we found a good on-line method.

Some Takeaway messages

Main findings:

- Degree heterogeneity breaks community structures in eigenvectors.
  $\Rightarrow$ Compensation by $D^{\alpha-1}$ normalization of eigenvectors.
- Classical debate over "best normalization" of adjacency (or modularity) matrix A not trivial to solve.
  $\Rightarrow$ With heterogeneous degrees, we found a good on-line method.
- Simulations support good performances even for "rather sparse" settings.

Some Takeaway messages

Main findings:

- Degree heterogeneity breaks community structures in eigenvectors.
  $\Rightarrow$ Compensation by $D^{\alpha-1}$ normalization of eigenvectors.
- Classical debate over "best normalization" of adjacency (or modularity) matrix A not trivial to solve.
  $\Rightarrow$ With heterogeneous degrees, we found a good on-line method.
- Simulations support good performances even for "rather sparse" settings.

But strong limitations:

Some Takeaway messages

Main findings:

- Degree heterogeneity breaks community structures in eigenvectors.
  $\Rightarrow$ Compensation by $D^{\alpha-1}$ normalization of eigenvectors.
- Classical debate over "best normalization" of adjacency (or modularity) matrix A not trivial to solve.
  $\Rightarrow$ With heterogeneous degrees, we found a good on-line method.
- Simulations support good performances even for "rather sparse" settings.

But strong limitations:

- Key assumption: $C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}$.
  $\Rightarrow$ Everything collapses if different regime.

Some Takeaway messages

Main findings:

- Degree heterogeneity breaks community structures in eigenvectors.
  ⇒ Compensation by $D^{\alpha-1}$ normalization of eigenvectors.
- Classical debate over "best normalization" of adjacency (or modularity) matrix A not trivial to solve.
  ⇒ With heterogeneous degrees, we found a good on-line method.
- Simulations support good performances even for "rather sparse" settings.

But strong limitations:

- Key assumption: $C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}$.
  ⇒ Everything collapses if different regime.
- Simulations on small networks in fact give ridiculous arbitrary results.

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

Kernel Spectral Clustering

Problem Statement

- Dataset $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: "cluster" data in k similarity classes $\mathcal{S}_1, \ldots, \mathcal{S}_k$.

Kernel Spectral Clustering

Problem Statement

- Dataset $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: "cluster" data in k similarity classes $\mathcal{S}_1, \ldots, \mathcal{S}_k$.

- Typical metric to optimize:

$$\text{(RatioCut)} \;\; \operatorname{argmin}_{\mathcal{S}_1 \cup \ldots \cup \mathcal{S}_k = \{1,\ldots,n\}} \sum_{i=1}^{k} \sum_{\substack{j \in \mathcal{S}_i \\ \bar{j} \notin \mathcal{S}_i}} \frac{\kappa(x_j, x_{\bar{j}})}{|\mathcal{S}_i|}$$

for some similarity kernel $\kappa(x, y) \geq 0$ (large if x similar to y).

Kernel Spectral Clustering

Problem Statement

- Dataset $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: "cluster" data in k similarity classes $\mathcal{S}_1, \ldots, \mathcal{S}_k$.

- Typical metric to optimize:

$$(\text{RatioCut}) \;\; \operatorname{argmin}_{\mathcal{S}_1 \cup \ldots \cup \mathcal{S}_k = \{1,\ldots,n\}} \sum_{i=1}^{k} \sum_{\substack{j \in \mathcal{S}_i \\ \bar{j} \notin \mathcal{S}_i}} \frac{\kappa(x_j, x_{\bar{j}})}{|\mathcal{S}_i|}$$

for some similarity kernel $\kappa(x, y) \geq 0$ (large if x similar to y).

- Can be shown equivalent to

$$(\text{RatioCut}) \;\; \operatorname{argmin}_{M \in \mathcal{M}} \operatorname{tr} M^{\mathsf{T}}(D - K)M$$

where $\mathcal{M} \subset \mathbb{R}^{n \times k} \cap \left\{M; \; M_{ij} \in \{0, |\mathcal{S}_j|^{-\frac{1}{2}}\}\right\}$ (in particular, $M^{\mathsf{T}}M = I_k$) and

$$K = \{\kappa(x_i, x_j)\}_{i,j=1}^{n}, \; D_{ii} = \sum_{j=1}^{n} K_{ij}.$$

Kernel Spectral Clustering

Problem Statement

- Dataset $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: "cluster" data in k similarity classes $\mathcal{S}_1, \ldots, \mathcal{S}_k$.

- Typical metric to optimize:

$$(\text{RatioCut}) \;\; \operatorname{argmin}_{\mathcal{S}_1 \cup \ldots \cup \mathcal{S}_k = \{1,\ldots,n\}} \sum_{i=1}^{k} \sum_{\substack{j \in \mathcal{S}_i \\ \bar{j} \notin \mathcal{S}_i}} \frac{\kappa(x_j, x_{\bar{j}})}{|\mathcal{S}_i|}$$

for some similarity kernel $\kappa(x,y) \geq 0$ (large if x similar to y).

- Can be shown equivalent to

$$(\text{RatioCut}) \;\; \operatorname{argmin}_{M \in \mathcal{M}} \operatorname{tr} M^{\mathsf{T}}(D - K)M$$

where $\mathcal{M} \subset \mathbb{R}^{n \times k} \cap \left\{ M;\; M_{ij} \in \{0, |\mathcal{S}_j|^{-\frac{1}{2}}\} \right\}$ (in particular, $M^{\mathsf{T}} M = I_k$) and

$$K = \{\kappa(x_i, x_j)\}_{i,j=1}^{n}, \;\; D_{ii} = \sum_{j=1}^{n} K_{ij}.$$

- But integer problem! Usually NP-complete.

Kernel Spectral Clustering

Towards kernel spectral clustering

- Kernel spectral clustering: discrete-to-continuous relaxations of such metrics

$$(\mathrm{RatioCut})\ \ \operatorname{argmin}_{M,\ M^{\mathsf{T}}M=I_K} \operatorname{tr} M^{\mathsf{T}}(D-K)M$$

  i.e., eigenvector problem:

  1. find eigenvectors of smallest eigenvalues
  2. retrieve classes from eigenvector components

Kernel Spectral Clustering

Towards kernel spectral clustering

- Kernel spectral clustering: discrete-to-continuous relaxations of such metrics

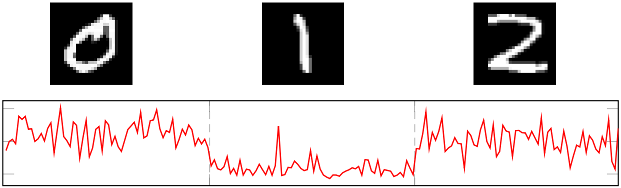
$$(\mathrm{RatioCut})\ \ \mathrm{argmin}_{M,\ M^{\mathsf{T}}M=I_K}\, \mathrm{tr}\, M^{\mathsf{T}}(D-K)M$$

  i.e., eigenvector problem:

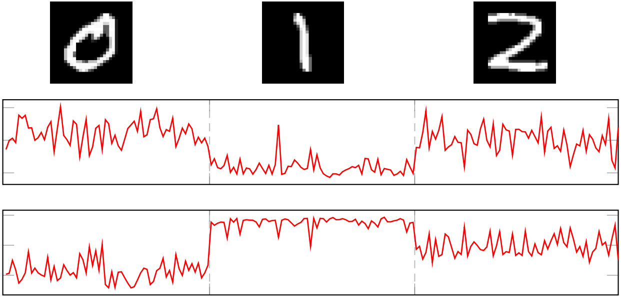
  1. find eigenvectors of smallest eigenvalues
  2. retrieve classes from eigenvector components

- Refinements:
  - working on K, $D-K$, $I_n-D^{-1}K$, $I_n-D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$, etc.
  - several steps algorithms: Ng–Jordan–Weiss, Shi–Malik, etc.

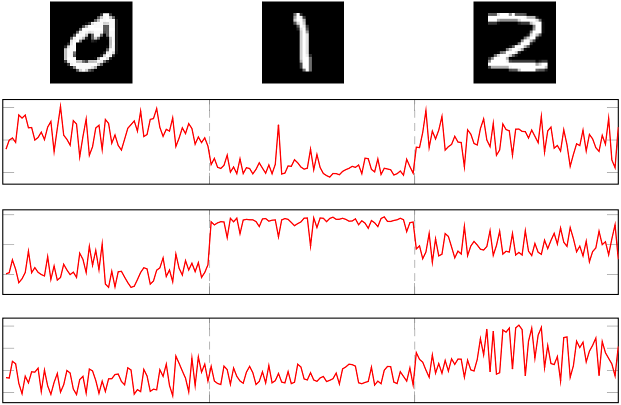
Kernel Spectral Clustering



Kernel Spectral Clustering



Kernel Spectral Clustering



Kernel Spectral Clustering

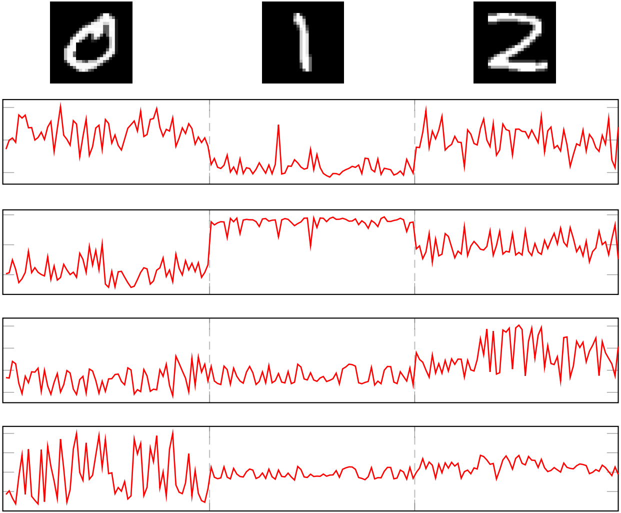


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data.

Methodology and objectives

Current state:

- Algorithms derived from ad-hoc procedures (e.g., relaxation).
- Little understanding of performance, even for Gaussian mixtures!
- Let alone when both p and n are large (BigData setting)

Methodology and objectives

Current state:

- Algorithms derived from ad-hoc procedures (e.g., relaxation).
- Little understanding of performance, even for Gaussian mixtures!
- Let alone when both p and n are large (BigData setting)

Objectives and Roadmap:

- Develop mathematical analysis framework for BigData kernel spectral clustering ($p, n \to \infty$)

Methodology and objectives

Current state:

- Algorithms derived from ad-hoc procedures (e.g., relaxation).
- Little understanding of performance, even for Gaussian mixtures!
- Let alone when both p and n are large (BigData setting)

Objectives and Roadmap:

- Develop mathematical analysis framework for BigData kernel spectral clustering ($p, n \to \infty$)
- Understand:
  1. Phase transition effects (i.e., when is clustering possible?)
  2. Content of each eigenvector
  3. Influence of kernel function
  4. Performance comparison of clustering algorithms

Methodology and objectives

Current state:

- Algorithms derived from ad-hoc procedures (e.g., relaxation).
- Little understanding of performance, even for Gaussian mixtures!
- Let alone when both p and n are large (BigData setting)

Objectives and Roadmap:

- Develop mathematical analysis framework for BigData kernel spectral clustering ($p, n \to \infty$)
- Understand:
  1. Phase transition effects (i.e., when is clustering possible?)
  2. Content of each eigenvector
  3. Influence of kernel function
  4. Performance comparison of clustering algorithms

Methodology:

- Use statistical assumptions (Gaussian mixture)
- Benefit from doubly-infinite independence and random matrix tools

Model and Assumptions

Gaussian mixture model:

- $x_1, \ldots, x_n \in \mathbb{R}^p$,
- k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$,
- $x_1, \ldots, x_{n_1} \in \mathcal{C}_1, \ldots, x_{n-n_k+1}, \ldots, x_n \in \mathcal{C}_k$,
- $\mathcal{C}_a = \{x \mid x \sim \mathcal{N}(\mu_a, C_a)\}$.

Then, for $x_i \in \mathcal{C}_a$, with $w_i \sim N(0, C_a)$,

$$x_i = \mu_a + w_i.$$

Model and Assumptions

Gaussian mixture model:

- $x_1, \ldots, x_n \in \mathbb{R}^p$,
- k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$,
- $x_1, \ldots, x_{n_1} \in \mathcal{C}_1, \ldots, x_{n-n_k+1}, \ldots, x_n \in \mathcal{C}_k$,
- $\mathcal{C}_a = \{x \mid x \sim \mathcal{N}(\mu_a, C_a)\}$.

Then, for $x_i \in \mathcal{C}_a$, with $w_i \sim N(0, C_a)$,

$$x_i = \mu_a + w_i.$$

Assumption (Convergence Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$,
2. **Class scaling**: $\frac{n_a}{n} \to c_a \in (0, 1)$,
3. **Mean scaling**: *with* $\mu^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} \mu_a$ *and* $\mu_a^\circ \triangleq \mu_a - \mu^\circ$, *then*

$$\|\mu_a^\circ\| = O(1)$$

4. **Covariance scaling**: *with* $C^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} C_a$ *and* $C_a^\circ \triangleq C_a - C^\circ$, *then*

$$\|C_a\| = O(1), \quad \frac{1}{\sqrt{p}} \mathit{tr} C_a^\circ = O(1) \Rightarrow \mathit{tr} C_a^\circ C_b^\circ = O(p)$$

Model and Assumptions

Kernel Matrix:

- Kernel matrix of interest:

$$K = \left\{ f\left(\frac{1}{p} \|x_i - x_j\|^2 \right) \right\}_{i,j=1}^{n}$$

for some sufficiently smooth nonnegative f.

Model and Assumptions

Kernel Matrix:

- Kernel matrix of interest:

$$K = \left\{ f\left(\frac{1}{p} \|x_i - x_j\|^2 \right) \right\}_{i,j=1}^{n}$$

for some sufficiently smooth nonnegative f.

- We study the normalized Laplacian:

$$L = nD^{-\frac{1}{2}} K D^{-\frac{1}{2}}$$

with $d = K1_n$, $D = \text{diag}(d)$.

Model and Assumptions

Difficulty: L is a very intractable random matrix

- non-linear f
- non-trivial dependence between entries of L

Model and Assumptions

Difficulty: L is a very intractable random matrix

- non-linear f
- non-trivial dependence between entries of L

Strategy:

1. Find random equivalent $\hat{L}$ (i.e., $\|L - \hat{L}\| \xrightarrow{\text{a.s.}} 0$ as $n, p \to \infty$) based on:
   - concentration: $K_{ij} \to \text{constant}$ as $n, p \to \infty$ (for all $i \neq j$)
   - Taylor expansion around limit point

Model and Assumptions

Difficulty: L is a very intractable random matrix

- non-linear f
- non-trivial dependence between entries of L

Strategy:

1. Find random equivalent $\hat{L}$ (i.e., $\|L - \hat{L}\| \xrightarrow{\text{a.s.}} 0$ as $n, p \to \infty$) based on:
    - concentration: $K_{ij} \to \text{constant}$ as $n, p \to \infty$ (for all $i \neq j$)
    - Taylor expansion around limit point
2. Apply spiked random matrix approach to study:
    - existence of isolated eigenvalues in $\hat{L}$: phase transition

Model and Assumptions

Difficulty: L is a very intractable random matrix

- non-linear f
- non-trivial dependence between entries of L

Strategy:

1. Find random equivalent $\hat{L}$ (i.e., $\|L-\hat{L}\| \xrightarrow{\text{a.s.}} 0$ as $n,p\to\infty$) based on:
   - concentration: $K_{ij}\to\text{constant}$ as $n,p\to\infty$ (for all $i\neq j$)
   - Taylor expansion around limit point
2. Apply spiked random matrix approach to study:
   - existence of isolated eigenvalues in $\hat{L}$: phase transition
   - eigenvector projections on canonical class-basis

Random Matrix Equivalent

Results on K:

- **Key Remark:** Under our assumptions, uniformly on $i, j \in \{1, \ldots, n\}$,

$$\frac{1}{p}\left\|x_i - x_j\right\|^2 \xrightarrow{\text{a.s.}} \tau$$

for some common limit τ.

Random Matrix Equivalent

Results on K:

- **Key Remark:** Under our assumptions, uniformly on $i, j \in \{1, \ldots, n\}$,

$$\frac{1}{p}\|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$$

  for some common limit τ.

- large dimensional approximation for K:

$$K = \underbrace{f(\tau)1_n 1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)} + \underbrace{\sqrt{n}A_1}_{\text{low rank, } O_{\|\cdot\|}(\sqrt{n})} + \underbrace{A_2}_{\text{informative terms, } O_{\|\cdot\|}(1)}$$

Random Matrix Equivalent

Results on K:

- **Key Remark:** Under our assumptions, uniformly on $i,j \in \{1,\ldots,n\}$,

$$\frac{1}{p}\|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$$

  for some common limit τ.
- large dimensional approximation for K:

$$K = \underbrace{f(\tau)1_n1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)} + \underbrace{\sqrt{n}A_1}_{\text{low rank, } O_{\|\cdot\|}(\sqrt{n})} + \underbrace{A_2}_{\text{informative terms, } O_{\|\cdot\|}(1)}$$

- difficult to handle (3 orders to manipulate!)

Random Matrix Equivalent

Results on K:

- **Key Remark:** Under our assumptions, uniformly on $i,j \in \{1,\ldots,n\}$,

$$\frac{1}{p}\|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$$

  for some common limit τ.
- large dimensional approximation for K:

$$K = \underbrace{f(\tau)1_n1_n^{\mathsf{T}}}_{O_{\|\cdot\|}(n)} + \underbrace{\sqrt{n}A_1}_{\text{low rank, } O_{\|\cdot\|}(\sqrt{n})} + \underbrace{A_2}_{\text{informative terms, } O_{\|\cdot\|}(1)}$$

- difficult to handle (3 orders to manipulate!)

Observation: Spectrum of $L = nD^{-\frac{1}{2}}KD^{-\frac{1}{2}}$:

- Dominant eigenvalue n with eigenvector $D^{\frac{1}{2}}1_n$
- All other eigenvalues of order $O(1)$.

Random Matrix Equivalent

Results on K:

- **Key Remark:** Under our assumptions, uniformly on $i,j \in \{1,\ldots,n\}$,

$$\frac{1}{p}\|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$$

  for some common limit τ.

- large dimensional approximation for K:

$$K = \underbrace{f(\tau)1_n1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)} + \underbrace{\sqrt{n}A_1}_{\text{low rank, } O_{\|\cdot\|}(\sqrt{n})} + \underbrace{A_2}_{\text{informative terms, } O_{\|\cdot\|}(1)}$$

- difficult to handle (3 orders to manipulate!)

Observation: Spectrum of $L = nD^{-\frac{1}{2}}KD^{-\frac{1}{2}}$:

- Dominant eigenvalue n with eigenvector $D^{\frac{1}{2}}1_n$
- All other eigenvalues of order $O(1)$.

$\Rightarrow$ Naturally leads to study:

- Projected normalized Laplacian (or "modularity"-type Laplacian):

$$L' = nD^{-\frac{1}{2}}KD^{-\frac{1}{2}} - n\frac{D^{\frac{1}{2}}1_n1_n^\mathsf{T}D^{\frac{1}{2}}}{1_n^\mathsf{T}D1_n} = nD^{-\frac{1}{2}}\left(K - \frac{dd^\mathsf{T}}{1^\mathsf{T}d}\right)D^{-\frac{1}{2}}.$$

- Dominant (normalized) eigenvector $\frac{D^{\frac{1}{2}}1_n}{\sqrt{1_n^\mathsf{T}D1_n}}$.

Random Matrix Equivalent

Theorem (Random Matrix Equivalent)

As $n, p \to \infty$, in operator norm, $\left\| L' - \hat{L}' \right\| \xrightarrow{\text{a.s.}} 0$, where

$$\hat{L}' = -2\frac{f'(\tau)}{f(\tau)}\left[\frac{1}{p}PW^\mathsf{T}WP + UBU^\mathsf{T}\right] + \alpha(\tau)I_n$$

and $\tau = \frac{2}{p}\mathrm{tr}\,C^\circ$, $W = [w_1, \ldots, w_n] \in \mathbb{R}^{p\times n}$ ($x_i = \mu_a + w_i$), $P = I_n - \frac{1}{n}1_n1_n^\mathsf{T}$,

$$U = \left[\frac{1}{\sqrt{p}}J, \Phi, \psi\right] \in \mathbb{R}^{n\times(2k+4)}$$

$$B = \begin{bmatrix} B_{11} & I_k - 1_kc^\mathsf{T} & \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t \\ I_k - c1_k^\mathsf{T} & 0_{k\times k} & 0_{k\times 1} \\ \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t^\mathsf{T} & 0_{1\times k} & \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \end{bmatrix} \in \mathbb{R}^{(2k+4)\times(2k+4)}$$

$$B_{11} = M^\mathsf{T}M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^\mathsf{T} - \frac{f''(\tau)}{f'(\tau)}T + \frac{p}{n}\frac{f(\tau)\alpha(\tau)}{2f'(\tau)}1_k1_k^\mathsf{T} \in \mathbb{R}^{k\times k}.$$

Random Matrix Equivalent

Theorem (Random Matrix Equivalent)

As $n, p \to \infty$, in operator norm, $\left\| L' - \hat{L}' \right\| \xrightarrow{\text{a.s.}} 0$, where

$$\hat{L}' = -2\frac{f'(\tau)}{f(\tau)}\left[\frac{1}{p}PW^{\mathsf{T}}WP + UBU^{\mathsf{T}}\right] + \alpha(\tau)I_n$$

and $\tau = \frac{2}{p}\text{tr}C^{\circ}$, $W = [w_1, \ldots, w_n] \in \mathbb{R}^{p\times n}$ ($x_i = \mu_a + w_i$), $P = I_n - \frac{1}{n}1_n 1_n^{\mathsf{T}}$,

$$U = \left[\frac{1}{\sqrt{p}}J, \Phi, \psi\right] \in \mathbb{R}^{n\times(2k+4)}$$

$$B = \begin{bmatrix} B_{11} & I_k - 1_k c^{\mathsf{T}} & \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t \\ I_k - c1_k^{\mathsf{T}} & 0_{k\times k} & 0_{k\times 1} \\ \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t^{\mathsf{T}} & 0_{1\times k} & \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \end{bmatrix} \in \mathbb{R}^{(2k+4)\times(2k+4)}$$

$$B_{11} = M^{\mathsf{T}}M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^{\mathsf{T}} - \frac{f''(\tau)}{f'(\tau)}T + \frac{p}{n}\frac{f(\tau)\alpha(\tau)}{2f'(\tau)}1_k 1_k^{\mathsf{T}} \in \mathbb{R}^{k\times k}.$$

$\frac{1}{\sqrt{p}}J = [j_1, \ldots, j_k] \in \mathbb{R}^{n\times k}$, j_a canonical vector of class $\mathcal{C}_a$.

Random Matrix Equivalent

Theorem (Random Matrix Equivalent)

As $n, p \to \infty$, in operator norm, $\left\| L' - \hat{L}' \right\| \xrightarrow{\text{a.s.}} 0$, where

$$\hat{L}' = -2\frac{f'(\tau)}{f(\tau)}\left[\frac{1}{p}PW^{\mathsf{T}}WP + UBU^{\mathsf{T}}\right] + \alpha(\tau)I_n$$

and $\tau = \frac{2}{p}\operatorname{tr} C^{\circ}$, $W = [w_1, \ldots, w_n] \in \mathbb{R}^{p\times n}$ ($x_i = \mu_a + w_i$), $P = I_n - \frac{1}{n}1_n 1_n^{\mathsf{T}}$,

$$U = \left[\frac{1}{\sqrt{p}}J, \Phi, \psi\right] \in \mathbb{R}^{n\times(2k+4)}$$

$$B = \begin{bmatrix} B_{11} & I_k - 1_k c^{\mathsf{T}} & \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t \\ I_k - c1_k^{\mathsf{T}} & 0_{k\times k} & 0_{k\times 1} \\ \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t^{\mathsf{T}} & 0_{1\times k} & \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \end{bmatrix} \in \mathbb{R}^{(2k+4)\times(2k+4)}$$

$$B_{11} = M^{\mathsf{T}}M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^{\mathsf{T}} - \frac{f''(\tau)}{f'(\tau)}T + \frac{p}{n}\frac{f(\tau)\alpha(\tau)}{2f'(\tau)}1_k 1_k^{\mathsf{T}} \in \mathbb{R}^{k\times k}.$$

$M = [\mu_1^{\circ}, \ldots, \mu_k^{\circ}] \in \mathbb{R}^{n\times k}$, $\mu_a^{\circ} = \mu_a - \sum_{b=1}^{k}\frac{n_b}{n}\mu_b$.

Random Matrix Equivalent

Theorem (Random Matrix Equivalent)

As $n, p \to \infty$, *in operator norm,* $\left\| L' - \hat{L}' \right\| \xrightarrow{\text{a.s.}} 0$, *where*

$$\hat{L}' = -2\frac{f'(\tau)}{f(\tau)}\left[\frac{1}{p}PW^{\mathsf{T}}WP + UBU^{\mathsf{T}}\right] + \alpha(\tau)I_n$$

and $\tau = \frac{2}{p}\operatorname{tr}C^{\circ}$, $W = [w_1, \ldots, w_n] \in \mathbb{R}^{p\times n}$ $(x_i = \mu_a + w_i)$, $P = I_n - \frac{1}{n}1_n1_n^{\mathsf{T}}$,

$$U = \left[\frac{1}{\sqrt{p}}J, \Phi, \psi\right] \in \mathbb{R}^{n\times(2k+4)}$$

$$B = \begin{bmatrix} B_{11} & I_k - 1_kc^{\mathsf{T}} & \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t \\ I_k - c1_k^{\mathsf{T}} & 0_{k\times k} & 0_{k\times 1} \\ \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t^{\mathsf{T}} & 0_{1\times k} & \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \end{bmatrix} \in \mathbb{R}^{(2k+4)\times(2k+4)}$$

$$B_{11} = M^{\mathsf{T}}M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^{\mathsf{T}} - \frac{f''(\tau)}{f'(\tau)}T + \frac{p}{n}\frac{f(\tau)\alpha(\tau)}{2f'(\tau)}1_k1_k^{\mathsf{T}} \in \mathbb{R}^{k\times k}.$$

$t = \left[\frac{1}{\sqrt{p}}\operatorname{tr}C_1^{\circ}, \ldots, \frac{1}{\sqrt{p}}\operatorname{tr}C_k^{\circ}\right] \in \mathbb{R}^k$, $C_a^{\circ} = C_a - \sum_{b=1}^{k}\frac{n_b}{n}C_b$.

Random Matrix Equivalent

Theorem (Random Matrix Equivalent)

As $n, p \to \infty$*, in operator norm,* $\left\| L' - \hat{L}' \right\| \xrightarrow{\text{a.s.}} 0$*, where*

$$\hat{L}' = -2\frac{f'(\tau)}{f(\tau)}\left[\frac{1}{p}PW^{\mathsf{T}}WP + UBU^{\mathsf{T}}\right] + \alpha(\tau)I_n$$

and $\tau = \frac{2}{p}\operatorname{tr} C^{\circ}$, $W = [w_1, \ldots, w_n] \in \mathbb{R}^{p\times n}$ *(*$x_i = \mu_a + w_i$*),* $P = I_n - \frac{1}{n}1_n 1_n^{\mathsf{T}}$,

$$U = \left[\frac{1}{\sqrt{p}}J, \Phi, \psi\right] \in \mathbb{R}^{n\times(2k+4)}$$

$$B = \begin{bmatrix} B_{11} & I_k - 1_k c^{\mathsf{T}} & \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t \\ I_k - c1_k^{\mathsf{T}} & 0_{k\times k} & 0_{k\times 1} \\ \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t^{\mathsf{T}} & 0_{1\times k} & \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \end{bmatrix} \in \mathbb{R}^{(2k+4)\times(2k+4)}$$

$$B_{11} = M^{\mathsf{T}}M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^{\mathsf{T}} - \frac{f''(\tau)}{f'(\tau)}T + \frac{p}{n}\frac{f(\tau)\alpha(\tau)}{2f'(\tau)}1_k 1_k^{\mathsf{T}} \in \mathbb{R}^{k\times k}.$$

$T = \left\{\frac{1}{p}\operatorname{tr} C_a^{\circ} C_b^{\circ}\right\}_{a,b=1}^{k} \in \mathbb{R}^{k\times k}$, $C_a^{\circ} = C_a - \sum_{b=1}^{k}\frac{n_b}{n}C_b$.

Random Matrix Equivalent

Some consequences:

- $\hat{L}'$ is a spiked model: UBU^{T} seen as low rank perturbation of $\frac{1}{p}PW^{\mathsf{T}}WP$

Random Matrix Equivalent

Some consequences:

- $\hat{L}'$ is a spiked model: UBU^{T} seen as low rank perturbation of $\frac{1}{p}PW^{\mathsf{T}}WP$
- If $f'(\tau) = 0$,
  - L' asymptotically deterministic!
  - only t and T can be discriminated upon
- If $f''(\tau) = 0$, (e.g., $f(x) = x$) T unused
- If $\frac{5f'(\tau)}{8f(\tau)} = \frac{f''(\tau)}{2f'(\tau)}$, t (seemingly) unused

Random Matrix Equivalent

Some consequences:

- $\hat{L}'$ is a spiked model: UBU^{T} seen as low rank perturbation of $\frac{1}{p}PW^{\mathsf{T}}WP$
- If $f'(\tau) = 0$,
  - L' asymptotically deterministic!
  - only t and T can be discriminated upon
- If $f''(\tau) = 0$, (e.g., $f(x) = x$) T unused
- If $\frac{5f'(\tau)}{8f(\tau)} = \frac{f''(\tau)}{2f'(\tau)}$, t (seemingly) unused

Further analysis:

- Determine separability condition for eigenvalues
- Evaluate eigenvalue positions when separable
- Evaluate eigenvector projection to canonical basis $j_1, \ldots, j_k$
- Evaluate fluctuation of eigenvectors.

Isolated eigenvalues: Gaussian inputs

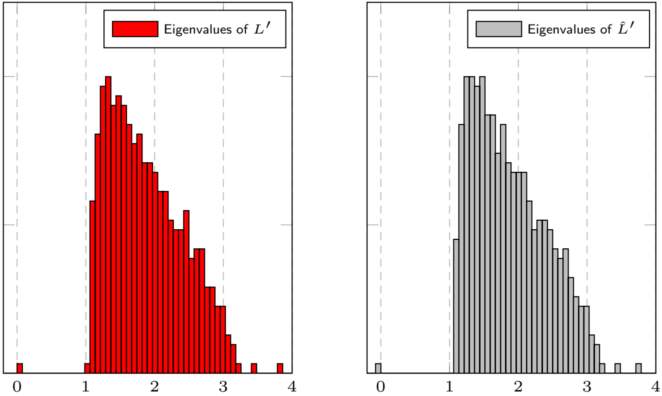


Figure: Eigenvalues of L' and $\hat{L}'$, $k = 3$, $p = 2048$, $n = 512$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $[\mu_a]_j = 4\boldsymbol{\delta}_{aj}$, $C_a = (1 + 2(a-1)/\sqrt{p})I_p$, $f(x) = \exp(-x/2)$.

Theoretical Findings versus MNIST

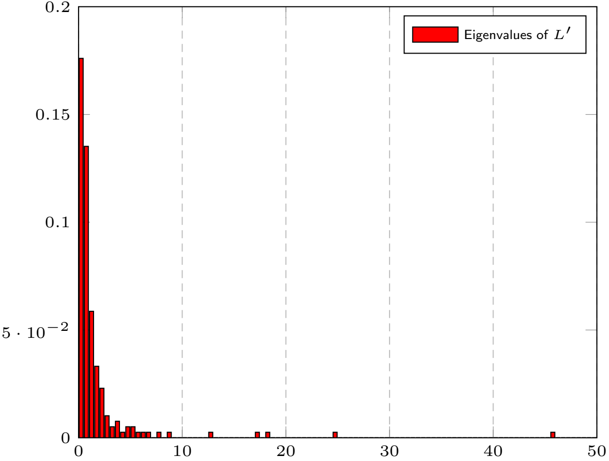


Figure: Eigenvalues of L' (red) and (equivalent Gaussian model) $\hat{L}'$ (white), MNIST data, $p = 784$, $n = 192$.

Theoretical Findings versus MNIST

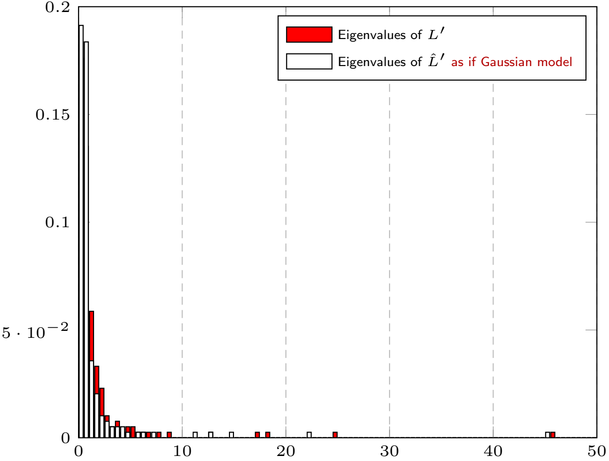


Figure: Eigenvalues of L' (red) and (equivalent Gaussian model) $\hat{L}'$ (white), MNIST data, $p = 784$, $n = 192$.

Theoretical Findings versus MNIST

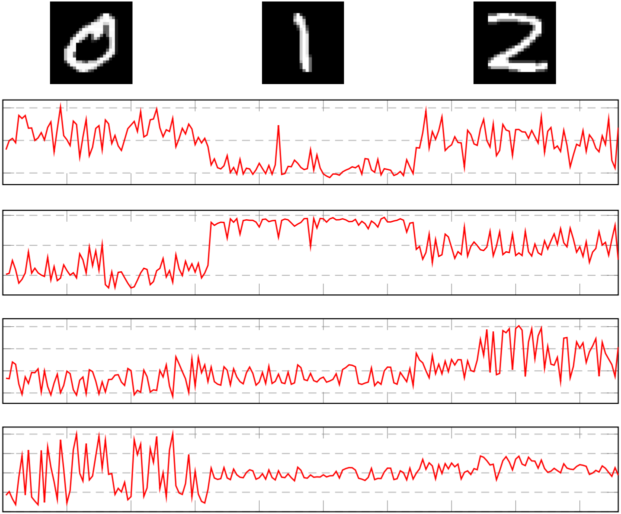


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data (**red**) and theoretical findings (**blue**).

Theoretical Findings versus MNIST

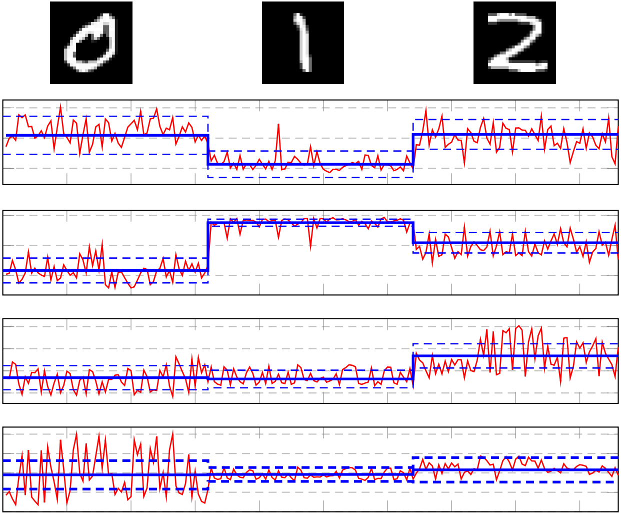


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data (**red**) and theoretical findings (**blue**).

Theoretical Findings versus MNIST

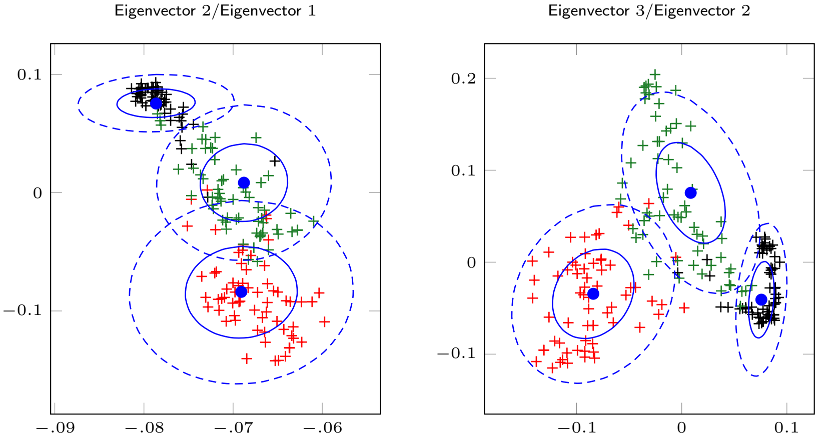


Figure: 2D representation of eigenvectors of L, for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in **blue**. Class 1 in **red**, Class 2 in **black**, Class 3 in **green**.

Further Results and Some Takeaway messages

General surprising findings:

Further Results and Some Takeaway messages

General surprising findings:

- “Good kernel functions” f need not be decreasing.

Further Results and Some Takeaway messages

General surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.

Further Results and Some Takeaway messages

General surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.
- Clustering possible despite $\|x_i - x_j\|^2 \to \tau$, i.e., no first order data difference
  $\Rightarrow$ **Breaks original intuitions and problem layout!**

Further Results and Some Takeaway messages

General surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.
- Clustering possible despite $\|x_i - x_j\|^2 \to \tau$, i.e., no first order data difference
  $\Rightarrow$ **Breaks original intuitions and problem layout!**

Further surprises. . . :

Further Results and Some Takeaway messages

General surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.
- Clustering possible despite $\|x_i - x_j\|^2 \to \tau$, i.e., no first order data difference
  $\Rightarrow$ **Breaks original intuitions and problem layout!**

Further surprises. . . :

- For $C_1 = \ldots = C_K = I_p$, **kernel choice is irrelevant**! (as long as $f'(\tau) \neq 0$)

Further Results and Some Takeaway messages

General surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.
- Clustering possible despite $\|x_i - x_j\|^2 \to \tau$, i.e., no first order data difference
  $\Rightarrow$ **Breaks original intuitions and problem layout!**

Further surprises. . . :

- For $C_1 = \ldots = C_K = I_p$, **kernel choice is irrelevant**! (as long as $f'(\tau) \neq 0$)
- For $\mu_1 = \ldots = \mu_K = 0$ and $C_a = (1 + \gamma_a p^{-\frac{1}{2}})I_p$, **only ONE isolated eigenvector**!

Further Results and Some Takeaway messages

General surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.
- Clustering possible despite $\|x_i - x_j\|^2 \to \tau$, i.e., no first order data difference $\Rightarrow$ **Breaks original intuitions and problem layout!**

Further surprises. . . :

- For $C_1 = \ldots = C_K = I_p$, **kernel choice is irrelevant**! (as long as $f'(\tau) \neq 0$)
- For $\mu_1 = \ldots = \mu_K = 0$ and $C_a = (1 + \gamma_a p^{-\frac{1}{2}}) I_p$, **only ONE isolated eigenvector**!
- It is possible to observe irrelevant eigenvectors! (that contain only noise)

Further Results and Some Takeaway messages

General surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.
- Clustering possible despite $\|x_i - x_j\|^2 \to \tau$, i.e., no first order data difference $\Rightarrow$ **Breaks original intuitions and problem layout!**

Further surprises. . . :

- For $C_1 = \ldots = C_K = I_p$, **kernel choice is irrelevant**! (as long as $f'(\tau) \neq 0$)
- For $\mu_1 = \ldots = \mu_K = 0$ and $C_a = (1 + \gamma_a p^{-\frac{1}{2}})I_p$, **only ONE isolated eigenvector**!
- It is possible to observe irrelevant eigenvectors! (that contain only noise)

Validity of the Results:

Further Results and Some Takeaway messages

General surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.
- Clustering possible despite $\|x_i - x_j\|^2 \to \tau$, i.e., no first order data difference
  $\Rightarrow$ **Breaks original intuitions and problem layout!**

Further surprises...:

- For $C_1 = \ldots = C_K = I_p$, **kernel choice is irrelevant**! (as long as $f'(\tau) \neq 0$)
- For $\mu_1 = \ldots = \mu_K = 0$ and $C_a = (1 + \gamma_a p^{-\frac{1}{2}})I_p$, **only ONE isolated eigenvector**!
- It is possible to observe irrelevant eigenvectors! (that contain only noise)

Validity of the Results:

- Needs a concentration of measure assumption: $\|x_i - x_j\|^2 \to \tau$.
- Invalid for heavy-tailed distributions (where $\|x_i\| = \|\sqrt{\tau_i} z_i\|$ needs not converge).

Further Results and Some Takeaway messages

General surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.
- Clustering possible despite $\|x_i - x_j\|^2 \to \tau$, i.e., no first order data difference
  $\Rightarrow$ **Breaks original intuitions and problem layout!**

Further surprises. . . :

- For $C_1 = \ldots = C_K = I_p$, **kernel choice is irrelevant**! (as long as $f'(\tau) \neq 0$)
- For $\mu_1 = \ldots = \mu_K = 0$ and $C_a = (1 + \gamma_a p^{-\frac{1}{2}})I_p$, **only ONE isolated eigenvector**!
- It is possible to observe irrelevant eigenvectors! (that contain only noise)

Validity of the Results:

- Needs a concentration of measure assumption: $\|x_i - x_j\|^2 \to \tau$.
- Invalid for heavy-tailed distributions (where $\|x_i\| = \|\sqrt{\tau_i} z_i\|$ needs not converge).
- Suprising fit between theory and practice: are images like Gaussian vectors?
  - kernels extract primarily first order properties (means, covariances)
  - without image processing (rotations, scale invariance), good enough features.

Last word: the suprising case $f'(\tau) = 0$...

Reminder:

Theorem (Random Matrix Equivalent)

As $n, p \to \infty$, in operator norm, $\left\| L' - \hat{L}' \right\| \xrightarrow{\text{a.s.}} 0$, *where*

$$\hat{L}' = -2\frac{f'(\tau)}{f(\tau)}\frac{1}{p}PW^\mathsf{T}WP - 2\frac{f'(\tau)}{f(\tau)}UBU^\mathsf{T} + \alpha(\tau)I_n$$

and $\tau = \frac{2}{p}\text{tr}\,C^\circ$, $W = [w_1, \ldots, w_n] \in \mathbb{R}^{p\times n}$ *(*$x_i = \mu_a + w_i$*),* $P = I_n - \frac{1}{n}1_n 1_n^\mathsf{T}$,

$$U = \left[\frac{1}{\sqrt{p}}J, *\right], \; B = \begin{bmatrix} B_{11} & * \\ * & * \end{bmatrix}$$

$$B_{11} = M^\mathsf{T}M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^\mathsf{T} - \frac{f''(\tau)}{f'(\tau)}T + \frac{p}{n}\frac{f(\tau)\alpha(\tau)}{2f'(\tau)}1_k 1_k^\mathsf{T}.$$

Last word: the suprising case $f'(\tau) = 0$...

Reminder:

Theorem (Random Matrix Equivalent)

As $n, p \to \infty$, in operator norm, $\left\| L' - \hat{L}' \right\| \xrightarrow{\text{a.s.}} 0$, where

$$\hat{L}' = -2\frac{f'(\tau)}{f(\tau)}\frac{1}{p}PW^{\mathsf{T}}WP - 2\frac{f'(\tau)}{f(\tau)}UBU^{\mathsf{T}} + \alpha(\tau)I_n$$

and $\tau = \frac{2}{p}\mathrm{tr}\, C^{\circ}$, $W = [w_1, \ldots, w_n] \in \mathbb{R}^{p \times n}$ ($x_i = \mu_a + w_i$), $P = I_n - \frac{1}{n}1_n 1_n^{\mathsf{T}}$,

$$U = \left[\frac{1}{\sqrt{p}}J, *\right], \; B = \begin{bmatrix} B_{11} & * \\ * & * \end{bmatrix}$$

$$B_{11} = M^{\mathsf{T}}M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^{\mathsf{T}} - \frac{f''(\tau)}{f'(\tau)}T + \frac{p}{n}\frac{f(\tau)\alpha(\tau)}{2f'(\tau)}1_k 1_k^{\mathsf{T}}.$$

When $f'(\tau) \to 0$**,**

- Means M disappears $\Rightarrow$ Impossible classification from means.

Last word: the suprising case $f'(\tau) = 0$...

Reminder:

Theorem (Random Matrix Equivalent)

As $n, p \to \infty$, in operator norm, $\left\| L' - \hat{L}' \right\| \xrightarrow{\text{a.s.}} 0$, where

$$\hat{L}' = -2\frac{f'(\tau)}{f(\tau)}\frac{1}{p}PW^\mathsf{T}WP - 2\frac{f'(\tau)}{f(\tau)}UBU^\mathsf{T} + \alpha(\tau)I_n$$

and $\tau = \frac{2}{p}\mathrm{tr}\, C^\circ$, $W = [w_1, \ldots, w_n] \in \mathbb{R}^{p \times n}$ ($x_i = \mu_a + w_i$), $P = I_n - \frac{1}{n}1_n 1_n^\mathsf{T}$,

$$U = \left[\frac{1}{\sqrt{p}}J, *\right], \; B = \begin{bmatrix} B_{11} & * \\ * & * \end{bmatrix}$$

$$B_{11} = M^\mathsf{T}M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^\mathsf{T} - \frac{f''(\tau)}{f'(\tau)}T + \frac{p}{n}\frac{f(\tau)\alpha(\tau)}{2f'(\tau)}1_k 1_k^\mathsf{T}.$$

When $f'(\tau) \to 0$,

- Means M disappears ⇒ Impossible classification from means.
- **More importantly:** $PWW^\mathsf{T}P$ disappears
  ⇒ Asymptotic deterministic matrix equivalent!
  ⇒ Perfect asymptotic clustering in theory!

Outline

Basics of Random Matrix Theory
- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications
- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

Position of the Problem

Problem: Cluster large data $x_1, \ldots, x_n \in \mathbb{R}^p$ based on "spanned subspaces".

Position of the Problem

Problem: Cluster large data $x_1, \ldots, x_n \in \mathbb{R}^p$ based on "spanned subspaces".

Method:

- Still assume $x_1, \ldots, x_n$ belong to k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$.
- Zero-mean Gaussian model for the data: for $x_i \in \mathcal{C}_k$,

$$x_i \sim \mathcal{N}(0, C_k).$$

Position of the Problem

Problem: Cluster large data $x_1,\ldots,x_n \in \mathbb{R}^p$ based on "spanned subspaces".

Method:

- Still assume $x_1,\ldots,x_n$ belong to k classes $\mathcal{C}_1,\ldots,\mathcal{C}_k$.
- Zero-mean Gaussian model for the data: for $x_i \in \mathcal{C}_k$,

$$x_i \sim \mathcal{N}(0, C_k).$$

- Performance of $L = nD^{-\frac{1}{2}}KD^{-\frac{1}{2}} - n\frac{D^{\frac{1}{2}}1_n1_n^\mathsf{T}D^{\frac{1}{2}}}{1_n^\mathsf{T}D1_n}$, with

$$K = \left\{ f\left(\|\bar{x}_i - \bar{x}_j\|^2 \right) \right\}_{1\leq i,j\leq n}, \quad \bar{x} = \frac{x}{\|x\|}$$

in the regime $n, p \to \infty$.

Model and Reminders

Assumption 1 [Classes]. Vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ i.i.d. from k-class Gaussian mixture, with $x_i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(0, C_k)$ (sorted by class for simplicity).

Model and Reminders

Assumption 1 [Classes]. Vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ i.i.d. from k-class Gaussian mixture, with $x_i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(0, C_k)$ (sorted by class for simplicity).

Assumption 2a [Growth Rates]. As $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \operatorname{tr} C_a = 1$ and $\operatorname{tr} C_a^\circ C_b^\circ = O(p)$, with $C_a^\circ = C_a - C^\circ$, $C^\circ = \sum_{b=1}^k c_b C_b$.

Model and Reminders

Assumption 1 [Classes]. Vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ i.i.d. from k-class Gaussian mixture, with $x_i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(0, C_k)$ (sorted by class for simplicity).

Assumption 2a [Growth Rates]. As $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \operatorname{tr} C_a = 1$ and $\operatorname{tr} C_a^\circ C_b^\circ = O(p)$, with $C_a^\circ = C_a - C^\circ$, $C^\circ = \sum_{b=1}^k c_b C_b$.

Theorem (Corollary of Previous Section)

Let f smooth with $f'(2) \neq 0$. Then, under Assumptions 2a,

$$L = nD^{-\frac{1}{2}} K D^{-\frac{1}{2}} - n \frac{D^{\frac{1}{2}} 1_n 1_n^\mathsf{T} D^{\frac{1}{2}}}{1_n^\mathsf{T} D 1_n}, \text{ with } K = \left\{ f\left(\|\bar{x}_i - \bar{x}_j\|^2 \right) \right\}_{i,j=1}^n \quad (\bar{x} = x / \|x\|)$$

exhibits phase transition phenomenon

Model and Reminders

Assumption 1 [Classes]. Vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ i.i.d. from k-class Gaussian mixture, with $x_i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(0, C_k)$ (sorted by class for simplicity).

Assumption 2a [Growth Rates]. As $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \operatorname{tr} C_a = 1$ and $\operatorname{tr} C_a^\circ C_b^\circ = O(p)$, with $C_a^\circ = C_a - C^\circ$, $C^\circ = \sum_{b=1}^k c_b C_b$.

Theorem (Corollary of Previous Section)

Let f smooth with $f'(2) \neq 0$. Then, under Assumptions 2a,

$$L = nD^{-\frac{1}{2}} K D^{-\frac{1}{2}} - n \frac{D^{\frac{1}{2}} 1_n 1_n^\mathsf{T} D^{\frac{1}{2}}}{1_n^\mathsf{T} D 1_n}, \text{ with } K = \left\{ f\left(\|\bar{x}_i - \bar{x}_j\|^2 \right) \right\}_{i,j=1}^n \quad (\bar{x} = x / \|x\|)$$

exhibits phase transition phenomenon, i.e., leading eigenvectors of L asymptotically contain structural information about $\mathcal{C}_1, \ldots, \mathcal{C}_k$ if and only if

$$T = \left\{ \frac{1}{p} \operatorname{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k$$

has sufficiently large eigenvalues.

The case $f'(2) = 0$

Assumption 2b [Growth Rates]. As $n \to \infty$, for each $a \in \{1, \dots, k\}$,

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p}\operatorname{tr} C_a = 1$ and ~~$\operatorname{tr} C_a^\circ C_b^\circ = O(p)$~~, with $C_a^\circ = C_a - C^\circ$, $C^\circ = \sum_{b=1}^k c_b C_b$.

The case $f'(2) = 0$

Assumption 2b [Growth Rates]. As $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p}\operatorname{tr} C_a = 1$ and $\operatorname{tr} C_a^\circ C_b^\circ = O(\sqrt{p})$, with $C_a^\circ = C_a - C^\circ$, $C^\circ = \sum_{b=1}^k c_b C_b$.

(in this regime, previous kernels clearly fail)

The case $f'(2) = 0$

Assumption 2b [Growth Rates]. As $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p}\operatorname{tr} C_a = 1$ and $\operatorname{tr} C_a^\circ C_b^\circ = O(\sqrt{p})$, with $C_a^\circ = C_a - C^\circ$, $C^\circ = \sum_{b=1}^k c_b C_b$.

(in this regime, previous kernels clearly fail)

Theorem (Random Equivalent for $f'(2) = 0$)

Let f be smooth with $f'(2) = 0$ and

$$\mathcal{L} \equiv \sqrt{p}\frac{f(2)}{2f''(2)}\left[L - \frac{f(0) - f(2)}{f(2)}P\right], \quad P = I_n - \frac{1}{n}1_n 1_n^\mathsf{T}.$$

Then, under Assumptions 2b,

$$\mathcal{L} = P\Phi P + \left\{\frac{1}{\sqrt{p}}\mathit{tr}\,(C_a^\circ C_b^\circ)\frac{1_{n_a}1_{n_b}^\mathsf{T}}{p}\right\}_{a,b=1}^k + o_{\|\cdot\|}(1)$$

where $\Phi_{ij} = \boldsymbol{\delta}_{i \neq j}\sqrt{p}\left[(x_i^\mathsf{T} x_j)^2 - E[(x_i^\mathsf{T} x_j)^2]\right]$.

The case $f'(2) = 0$

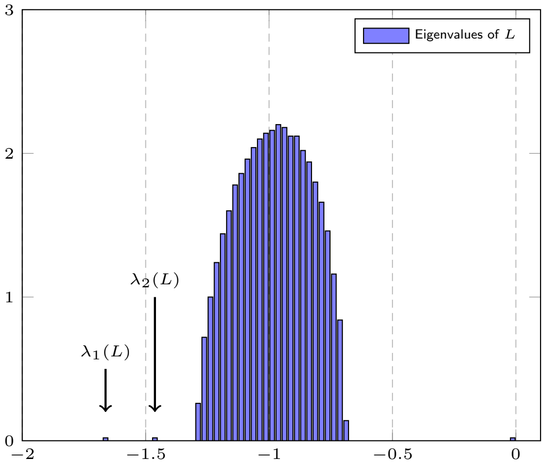
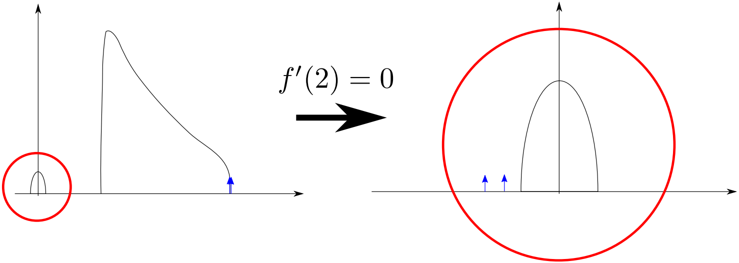


Figure: Eigenvalues of L, $p = 1000$, $n = 2000$, $k = 3$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $C_i \propto I_p + (p/8)^{-\frac{5}{4}} W_i W_i^{\mathsf{T}}$, $W_i \in \mathbb{R}^{p\times(p/8)}$ of i.i.d. $\mathcal{N}(0,1)$ entries, $f(t) = \exp(-(t-2)^2)$.

⇒ No longer a Marcenko–Pastur like bulk, but rather a semi-circle bulk!

The case $f'(2) = 0$



$f'(2) = 0$

The case $f'(2) = 0$

Roadmap. We now need to:

- study the spectrum of Φ

The case $f'(2) = 0$

Roadmap. We now need to:

- study the spectrum of Φ
- study the isolated eigenvalues of $\mathcal{L}$ (and the phase transition)

The case $f'(2) = 0$

Roadmap. We now need to:

- study the spectrum of Φ
- study the isolated eigenvalues of $\mathcal{L}$ (and the phase transition)
- retrieve information from the eigenvectors.

The case $f'(2) = 0$

Roadmap. We now need to:

- study the spectrum of Φ
- study the isolated eigenvalues of $\mathcal{L}$ (and the phase transition)
- retrieve information from the eigenvectors.

Theorem (Semi-circle law for Φ)

Let $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\delta}_{\lambda_i(\mathcal{L})}$. Then, under Assumption 2b,

$$\mu_n \xrightarrow{\text{a.s.}} \mu$$

with μ the semi-circle distribution

$$\mu(dt) = \frac{1}{2\pi c_0 \omega^2}\sqrt{(4c_0\omega^2 - t^2)^+}dt, \quad \omega = \lim_{p\to\infty} \sqrt{2}\frac{1}{p}\mathit{tr}(C^\circ)^2.$$

The case $f'(2) = 0$

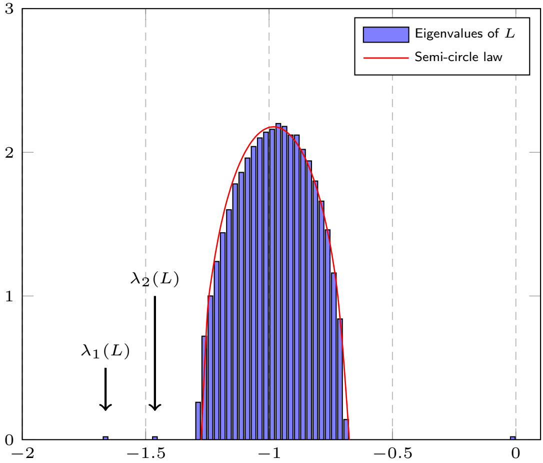


Figure: Eigenvalues of L, $p = 1000$, $n = 2000$, $k = 3$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $C_i \propto I_p + (p/8)^{-\frac{5}{4}} W_i W_i^\mathsf{T}$, $W_i \in \mathbb{R}^{p\times(p/8)}$ of i.i.d. $\mathcal{N}(0,1)$ entries, $f(t) = \exp(-(t-2)^2)$.

The case $f'(2) = 0$

Denote now

$$\mathcal{T} \equiv \lim_{p\to\infty} \left\{ \frac{\sqrt{c_a c_b}}{\sqrt{p}} \operatorname{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^{k}.$$

The case $f'(2) = 0$

Denote now

$$\mathcal{T} \equiv \lim_{p\to\infty} \left\{ \frac{\sqrt{c_a c_b}}{\sqrt{p}} \operatorname{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^{k}.$$

Theorem (Isolated Eigenvalues)

Let $\nu_1 \geq \ldots \geq \nu_k$ eigenvalues of $\mathcal{T}$. Then, if $\sqrt{c_0}|\nu_i| > \omega$, $\mathcal{L}$ has an isolated eigenvalue λ_i satisfying

$$\lambda_i \xrightarrow{\text{a.s.}} \rho_i \equiv c_0 \nu_i + \frac{\omega^2}{\nu_i}.$$

The case $f'(2) = 0$

Theorem (Isolated Eigenvectors)

For each isolated eigenpair (λ_i, u_i) *of* $\mathcal{L}$ *corresponding to* (ν_i, v_i) *of* $\mathcal{T}$, *write*

$$u_i = \sum_{a=1}^{k} \alpha_i^a \frac{j_a}{\sqrt{n_a}} + \sigma_i^a w_i^a$$

with $j_a = [0_{n_1}^\mathsf{T}, \ldots, 1_{n_a}^\mathsf{T}, \ldots, 0_{n_k}^\mathsf{T}]^\mathsf{T}$, $(w_i^a)^\mathsf{T} j_a = 0$, $\operatorname{supp}(w_i^a) = \operatorname{supp}(j_a)$, $\|w_i^a\| = 1$. *Then, under Assumptions 1–2b,*

$$\alpha_i^a \alpha_i^b \xrightarrow{\text{a.s.}} \left(1 - \frac{1}{c_0} \frac{\omega^2}{\nu_i^2}\right) [v_i v_i^\mathsf{T}]_{ab}$$

$$(\sigma_i^a)^2 \xrightarrow{\text{a.s.}} \frac{c_a}{c_0} \frac{\omega^2}{\nu_i^2}$$

and the fluctuations of u_i, u_j, $i \neq j$, *are asymptotically uncorrelated.*

The case $f'(2) = 0$

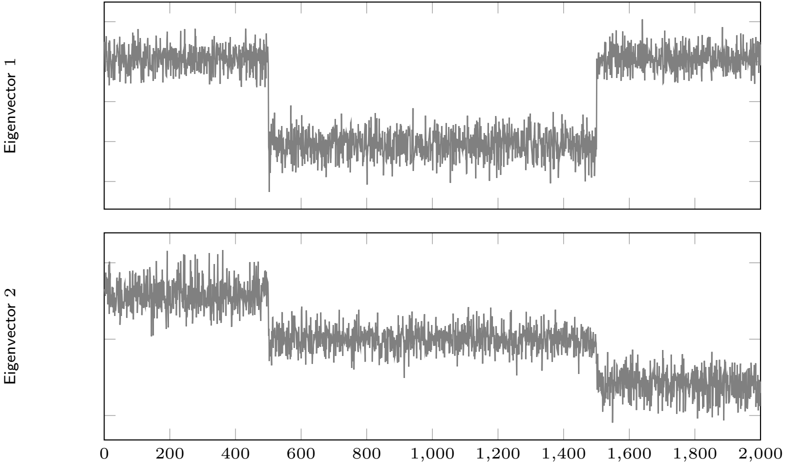


Figure: Leading two eigenvectors of $\mathcal{L}$ (or equivalently of L) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$.

The case $f'(2) = 0$

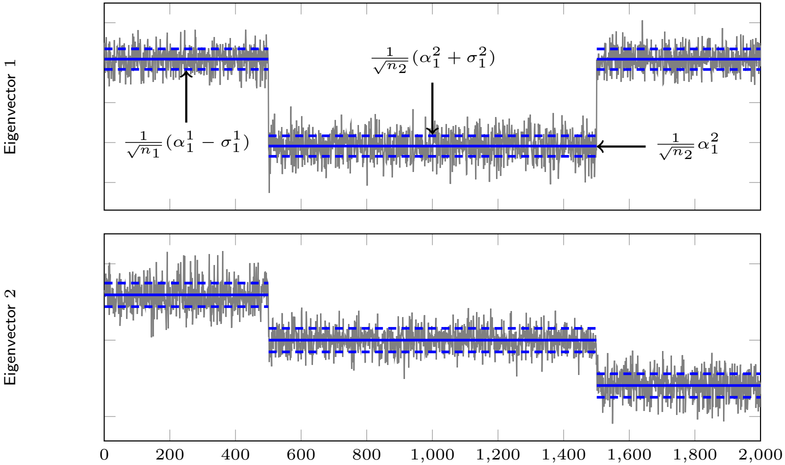


Figure: Leading two eigenvectors of $\mathcal{L}$ (or equivalently of L) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$.

The case $f'(2) = 0$

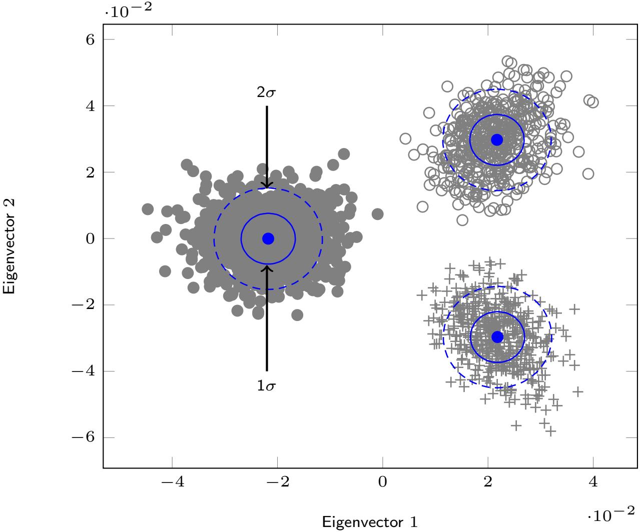


Figure: Leading two eigenvectors of $\mathcal{L}$ (or equivalently of L) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$.

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

Problem Statement

Context: Similar to clustering:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ in k classes, but with labelled and unlabelled data.

Problem Statement

Context: Similar to clustering:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ in k classes, but with labelled and unlabelled data.
- Problem statement: $(d_i = [K1_n]_i)$

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^{k} \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2$$

such that $F_{ia} = \delta_{\{x_i \in \mathcal{C}_a\}}$, for all labelled x_i.

Problem Statement

Context: Similar to clustering:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ in k classes, but with labelled and unlabelled data.
- Problem statement: $(d_i = [K1_n]_i)$

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^{k} \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2$$

  such that $F_{ia} = \delta_{\{x_i \in \mathcal{C}_a\}}$, for all labelled x_i.

- **Solution**: denoting $F^{(u)} \in \mathbb{R}^{n_u \times k}$, $F^{(l)} \in \mathbb{R}^{n_l \times k}$ the restriction to unlabelled/labelled data,

$$F^{(u)} = \left(I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}$$

  where we naturally decompose

$$K = \begin{bmatrix} K_{(l,l)} & K_{(l,u)} \\ K_{(u,l)} & K_{(u,u)} \end{bmatrix}$$

$$D = \begin{bmatrix} D_{(l)} & 0 \\ 0 & D^{(u)} \end{bmatrix} = \operatorname{diag}\{K1_n\}.$$

Problem Statement

Using $F^{(u)}$:

- From $F^{(u)}$, classification algorithm:

$$\text{Classify } x_i \text{ in } \mathcal{C}_a \quad \Leftrightarrow \quad F_{ia} = \max_{b\in\{1,\ldots,k\}} \{F_{ib}\}\,.$$

Problem Statement

Using $F^{(u)}$:

- From $F^{(u)}$, classification algorithm:

$$\text{Classify } x_i \text{ in } \mathcal{C}_a \quad \Leftrightarrow \quad F_{ia} = \max_{b \in \{1,\ldots,k\}} \{F_{ib}\}.$$

Objectives: For $x_i \sim \mathcal{N}(\mu_a, C_a)$, and as $n, p \to \infty$, ($n_u, n_l \to \infty$ or $n_u \to \infty$, $n_l = O(1)$)

Problem Statement

Using $F^{(u)}$:

- From $F^{(u)}$, classification algorithm:

$$\text{Classify } x_i \text{ in } \mathcal{C}_a \quad \Leftrightarrow \quad F_{ia} = \max_{b\in\{1,\ldots,k\}} \{F_{ib}\}\,.$$

Objectives: For $x_i \sim \mathcal{N}(\mu_a, C_a)$, and as $n, p \to \infty$, ($n_u, n_l \to \infty$ or $n_u \to \infty$, $n_l = O(1)$)

- Tractable approximation (in norm) for the vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, \ldots, k$
- Joint asymptotic behavior of $[F^{(u)}]_{i,\cdot}$
  $\Rightarrow$ From which classification probability is retrieved.

Problem Statement

Using $F^{(u)}$:

- From $F^{(u)}$, classification algorithm:

$$\text{Classify } x_i \text{ in } \mathcal{C}_a \quad \Leftrightarrow \quad F_{ia} = \max_{b \in \{1,\ldots,k\}} \{F_{ib}\}.$$

Objectives: For $x_i \sim \mathcal{N}(\mu_a, C_a)$, and as $n, p \to \infty$, $(n_u, n_l \to \infty$ or $n_u \to \infty$, $n_l = O(1))$

- Tractable approximation (in norm) for the vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, \ldots, k$
- Joint asymptotic behavior of $[F^{(u)}]_{i,\cdot}$
  $\Rightarrow$ From which classification probability is retrieved.
- Understanding the impact of α
  $\Rightarrow$ Finding optimal α choice online?

MNIST Data Example

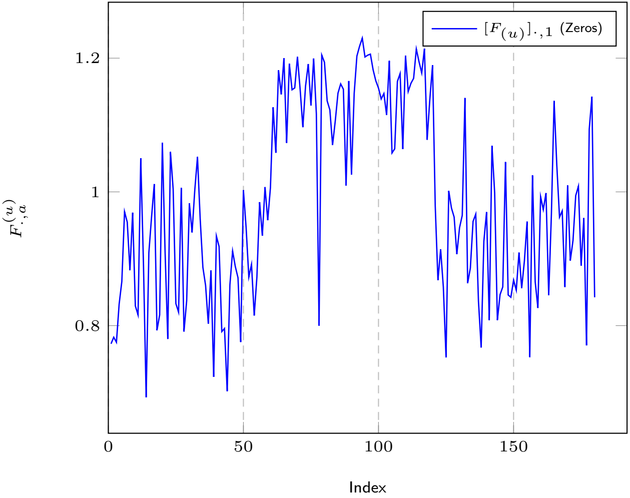


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

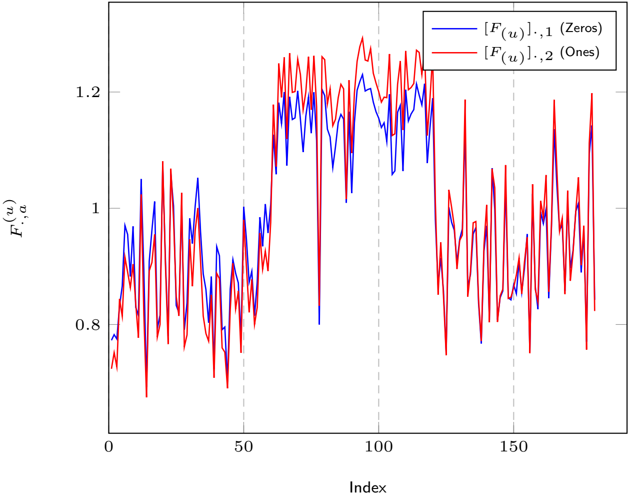


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

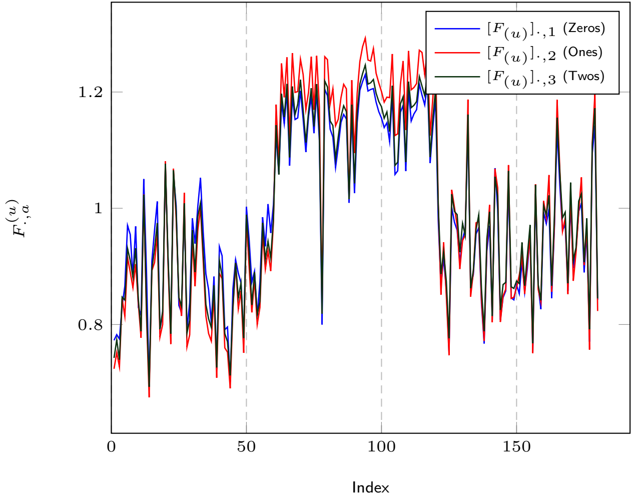


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

Comments

Not at all what we expect!:

Comments

Not at all what we expect!:

- Intuitively, $[F^{(u)}]_{i,a}$ should be close to 1 if $x_i \in \mathcal{C}_a$ or 0 if $x_i \notin \mathcal{C}_a$ (from cost function $K_{ij}(F_{i,a} - F_{j,a})^2$)

Comments

Not at all what we expect!:

- Intuitively, $[F^{(u)}]_{i,a}$ should be close to 1 if $x_i \in \mathcal{C}_a$ or 0 if $x_i \notin \mathcal{C}_a$ (from cost function $K_{ij}(F_{i,a} - F_{j,a})^2$)
- Here, strong class-wise biases

Comments

Not at all what we expect!:

- Intuitively, $[F^{(u)}]_{i,a}$ should be close to 1 if $x_i \in \mathcal{C}_a$ or 0 if $x_i \notin \mathcal{C}_a$ (from cost function $K_{ij}(F_{i,a} - F_{j,a})^2$)
- Here, strong class-wise biases
- **But, more surprisingly, it still works very well !**

Comments

Not at all what we expect!:

- Intuitively, $[F^{(u)}]_{i,a}$ should be close to 1 if $x_i \in \mathcal{C}_a$ or 0 if $x_i \notin \mathcal{C}_a$ (from cost function $K_{ij}(F_{i,a} - F_{j,a})^2$)
- Here, strong class-wise biases
- **But, more surprisingly, it still works very well !**

We need to understand why...

MNIST Data Example

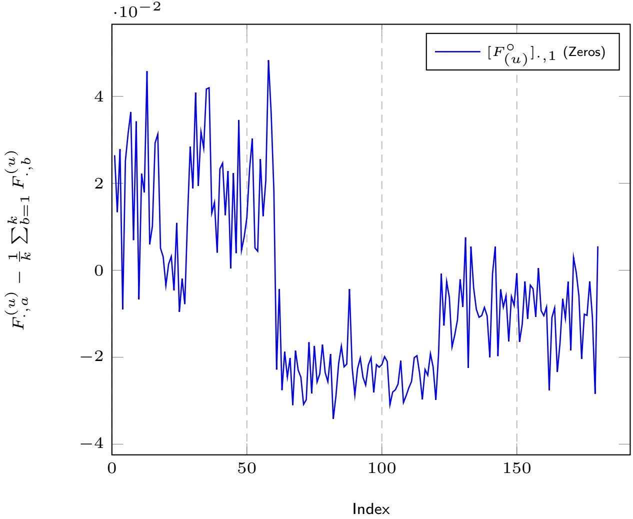


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k1_k^{\mathsf{T}}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

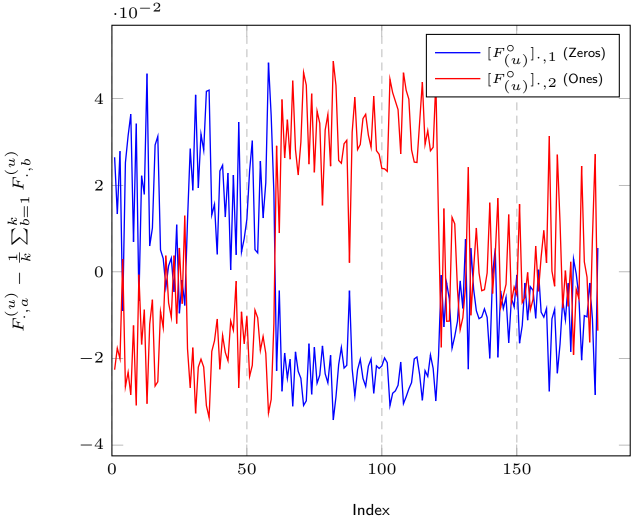


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k 1_k^{\mathsf{T}}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

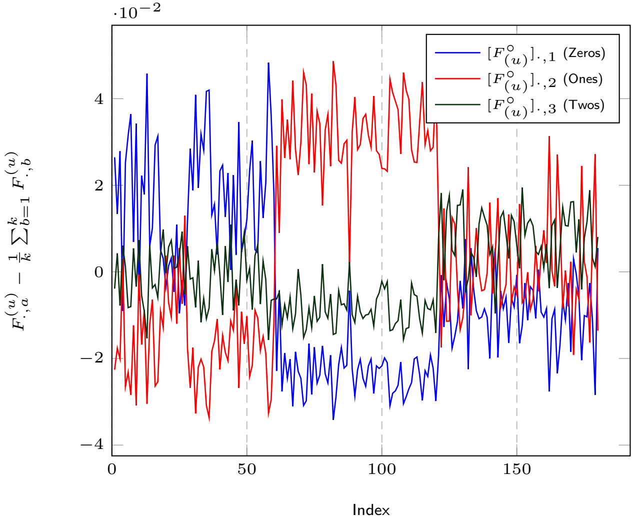


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k 1_k^{\mathsf{T}}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

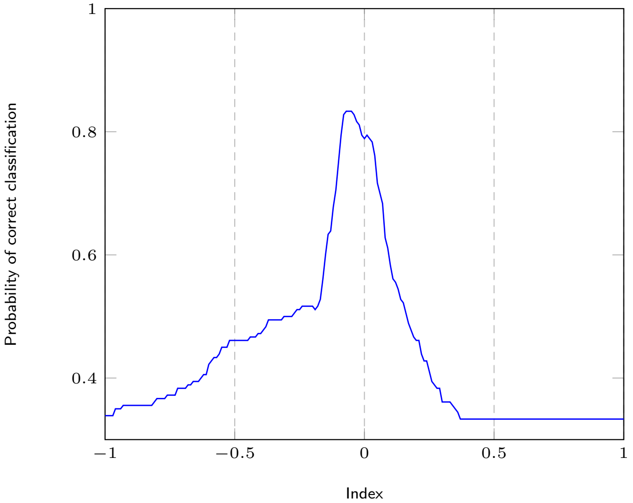


Figure: Performance as a function of α, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

Theoretical Findings

Method: We assume $n_l/n \to c_l \in (0,1)$ ("numerous" labelled data setting)

- Recall that we aim at characterizing

$$F^{(u)} = \left(I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}$$

  - A priori difficulty linked to resolvent of involved random matrix!
  - Painstaking product of complex matrices.

Theoretical Findings

Method: We assume $n_l/n \to c_l \in (0,1)$ ("numerous" labelled data setting)

- Recall that we aim at characterizing

$$F^{(u)} = \left(I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}$$

  - A priori difficulty linked to resolvent of involved random matrix!
  - Painstaking product of complex matrices.

- Using Taylor expansion of K as $n, p \to \infty$, we get

$$K_{(u,u)} = f(\tau) 1_{n_u} 1_{n_u}^{\mathsf{T}} + O_{\|\cdot\|}(n^{-\frac{1}{2}})$$
$$D_{(u)} = n f(\tau) I_{n_u} + O(n^{\frac{1}{2}})$$

and similarly for $K_{(u,l)}$, $D_{(l)}$.

Theoretical Findings

Method: We assume $n_l/n \to c_l \in (0,1)$ ("numerous" labelled data setting)

- Recall that we aim at characterizing

$$F^{(u)} = \left(I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1}\right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}$$

  - A priori difficulty linked to resolvent of involved random matrix!
  - Painstaking product of complex matrices.

- Using Taylor expansion of K as $n, p \to \infty$, we get

$$K_{(u,u)} = f(\tau) 1_{n_u} 1_{n_u}^{\mathsf{T}} + O_{\|\cdot\|}(n^{-\frac{1}{2}})$$
$$D_{(u)} = n f(\tau) I_{n_u} + O(n^{\frac{1}{2}})$$

  and similarly for $K_{(u,l)}$, $D_{(l)}$.

- So that

$$\left(I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1}\right)^{-1} = \left(I_{n_u} - \frac{1_{n_u} 1_{n_u}^{\mathsf{T}}}{n} + O_{\|\cdot\|}(n^{-\frac{1}{2}})\right)^{-1}$$

  which can be easily Taylor expanded!

Main Results

Results:

- In the first order,

$$F^{(u)}_{\cdot,a} = C\frac{n_{l,a}}{n}\left[v + \alpha\frac{t_a 1_{n_u}}{\sqrt{n}}\right] + \underbrace{O(n^{-1})}_{\text{Information is here!}}$$

where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}}\operatorname{tr} C^{\circ}_a$.

Main Results

Results:

- In the first order,

$$F_{\cdot,a}^{(u)} = C\frac{n_{l,a}}{n}\left[v + \alpha\frac{t_a 1_{n_u}}{\sqrt{n}}\right] + \underbrace{O(n^{-1})}_{\text{Information is here!}}$$

where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}}\operatorname{tr} C_a^\circ$.

- Many consequences:

Main Results

Results:

- In the first order,

$$F_{\cdot,a}^{(u)} = C\frac{n_{l,a}}{n}\left[v + \alpha\frac{t_a 1_{n_u}}{\sqrt{n}}\right] + \underbrace{O(n^{-1})}_{\text{Information is here!}}$$

  where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}}\operatorname{tr} C_a^\circ$.

- Many consequences:
  - Random non-informative bias linked to v

Main Results

Results:

- In the first order,

$$F^{(u)}_{\cdot,a} = C\frac{n_{l,a}}{n}\left[v + \alpha\frac{t_a 1_{n_u}}{\sqrt{n}}\right] + \underbrace{O(n^{-1})}_{\text{Information is here!}}$$

  where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}}\operatorname{tr} C_a^\circ$.

- Many consequences:
  - Random non-informative bias linked to v
  - Strong Impact of $n_{l,a}$!
    $\Rightarrow$ All $n_{l,a}$ must be equal **OR** $F^{(u)}$ need be scaled!

Main Results

Results:

- In the first order,

$$F_{\cdot,a}^{(u)} = C\frac{n_{l,a}}{n}\left[v + \alpha\frac{t_a 1_{n_u}}{\sqrt{n}}\right] + \underbrace{O(n^{-1})}_{\text{Information is here!}}$$

  where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}}\operatorname{tr} C_a^\circ$.

- Many consequences:
  - Random non-informative bias linked to v
  - Strong Impact of $n_{l,a}$!
    $\Rightarrow$ All $n_{l,a}$ must be equal **OR** $F^{(u)}$ need be scaled!
  - Additional per-class bias $\alpha t_a 1_{n_u}$: no information here
    $\Rightarrow$ **Forces the choice**

$$\alpha = 0 + \frac{\beta}{\sqrt{p}}.$$

Main Results

Results:

- In the first order,

$$F_{\cdot,a}^{(u)} = C\frac{n_{l,a}}{n}\left[v + \alpha\frac{t_a 1_{n_u}}{\sqrt{n}}\right] + \underbrace{O(n^{-1})}_{\text{Information is here!}}$$

  where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}}\operatorname{tr} C_a^\circ$.

- Many consequences:
  - Random non-informative bias linked to v
  - Strong Impact of $n_{l,a}$!
    $\Rightarrow$ All $n_{l,a}$ must be equal **OR** $F^{(u)}$ need be scaled!
  - Additional per-class bias $\alpha t_a 1_{n_u}$: no information here
    $\Rightarrow$ **Forces the choice**

$$\alpha = 0 + \frac{\beta}{\sqrt{p}}.$$

  - Relevant information hidden in smaller order terms!

Main Results

As a consequence of the remarks above, we take

$$\alpha = \frac{\beta}{\sqrt{p}}$$

and define

$$\hat{F}_{i,a}^{(u)} = \frac{np}{n_{l,a}} F_{ia}^{(u)}.$$

Main Results

As a consequence of the remarks above, we take

$$\alpha = \frac{\beta}{\sqrt{p}}$$

and define

$$\hat{F}^{(u)}_{i,a} = \frac{np}{n_{l,a}} F^{(u)}_{ia}.$$

Theorem

For $x_i \in \mathcal{C}_b$ unlabelled, we have

$$\hat{F}_{i,\cdot} - G_b \to 0,\ G_b \sim \mathcal{N}(m_b, \Sigma_b)$$

where $m_b \in \mathbb{R}^k$, $\Sigma_b \in \mathbb{R}^{k\times k}$ given by

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)}\tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)}\tilde{t}_a\tilde{t}_b + \frac{2f''(\tau)}{f(\tau)}\tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2}t_a t_b + \beta\frac{n}{n_l}\frac{f'(\tau)}{f(\tau)}t_a + B_b$$

$$(\Sigma_b)_{a_1a_2} = \frac{2\mathrm{tr}\,C_b^2}{p}\left(\frac{f'(\tau)^2}{f(\tau)^2} - \frac{f''(\tau)}{f(\tau)}\right)^2 t_{a_1}t_{a_2} + \frac{4f'(\tau)^2}{f(\tau)^2}\left([M^{\mathsf{T}}C_bM]_{a_1a_2} + \frac{\delta_{a_1}^{a_2}p}{n_{l,a_1}}T_{ba_1}\right)$$

with t, T, M as before, $\tilde{X}_a = X_a - \sum_{d=1}^k \frac{n_{l,d}}{n_l}X_d^\circ$ and B_b bias independent of a.

Main Results

Corollary (Asymptotic Classification Error)

For $k = 2$ classes and $a \neq b$,

$$P(\hat{F}_{i,a} > \hat{F}_{ib} \mid x_i \in \mathcal{C}_b) - Q\left(\frac{(m_b)_b - (m_b)_a}{\sqrt{[1,-1]\Sigma_b[1,-1]^{\mathsf{T}}}}\right) \to 0.$$

Main Results

Corollary (Asymptotic Classification Error)

For $k = 2$ classes and $a \neq b$,

$$P(\hat{F}_{i,a} > \hat{F}_{ib} \mid x_i \in \mathcal{C}_b) - Q\left(\frac{(m_b)_b - (m_b)_a}{\sqrt{[1,-1]\Sigma_b[1,-1]^\mathsf{T}}}\right) \to 0.$$

Some consequences:

- non obvious choices of appropriate kernels
- non obvious choice of optimal β (induces a possibly beneficial bias)
- importance of n_l versus n_u.

MNIST Data Example

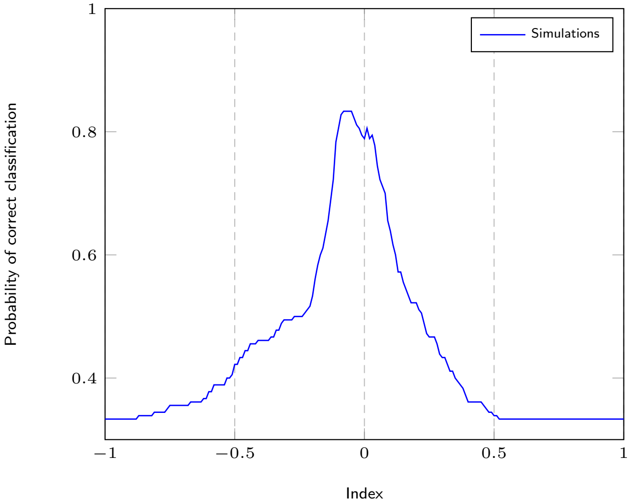


Figure: Performance as a function of α, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

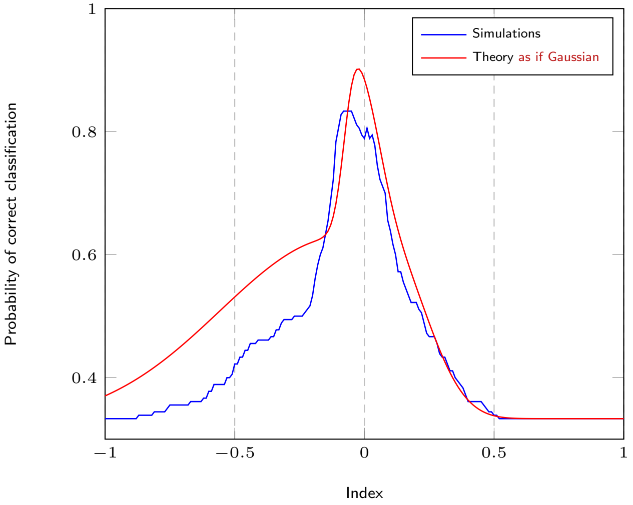


Figure: Performance as a function of α, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

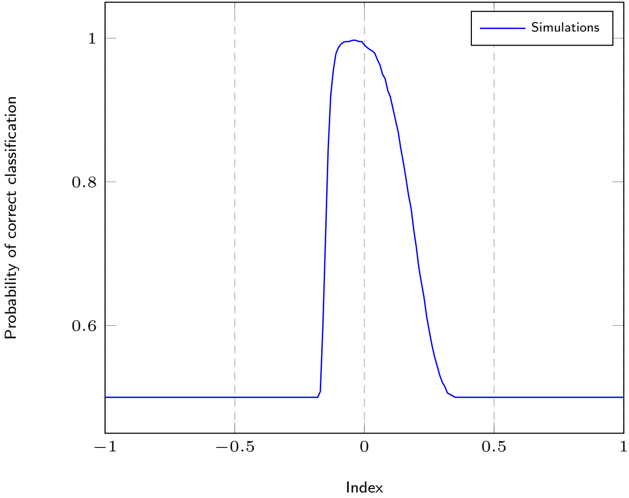


Figure: Performance as a function of α, for 2-class MNIST data (zeros, ones), $n = 1568$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

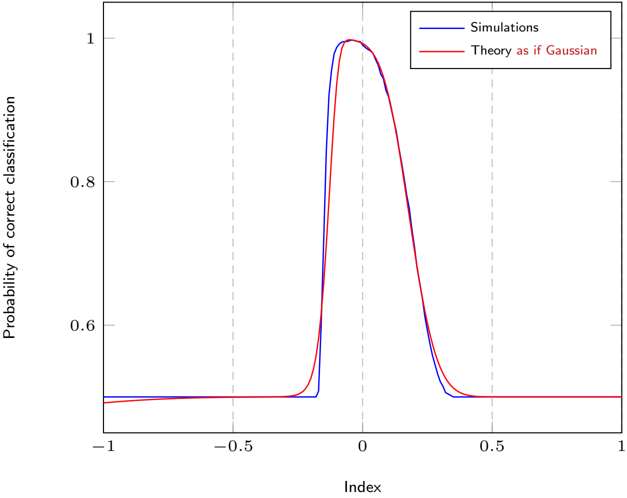


Figure: Performance as a function of α, for 2-class MNIST data (zeros, ones), $n = 1568$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

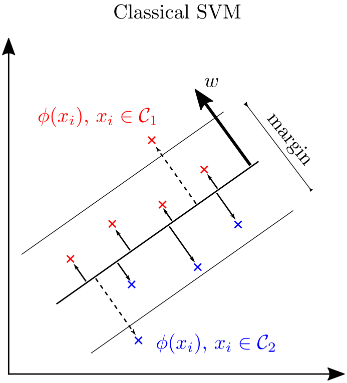
Outline

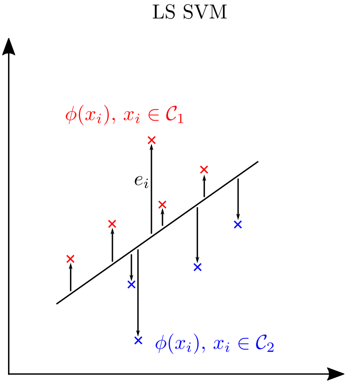
- Basics of Random Matrix Theory
  - Motivation: Large Sample Covariance Matrices
  - The Stieltjes Transform Method
  - Spiked Models
  - Other Common Random Matrix Models

- Applications
  - Random Matrices and Robust Estimation
  - Spectral Clustering Methods and Random Matrices
  - Community Detection on Graphs
  - Kernel Spectral Clustering
  - Kernel Spectral Clustering: Subspace Clustering
  - Semi-supervised Learning
  - Support Vector Machines
  - Neural Networks: Extreme Learning Machines

- Perspectives

Problem Statement


Classical SVM
w
$\phi(x_i),\ x_i \in \mathcal{C}_1$
margin
$\phi(x_i),\ x_i \in \mathcal{C}_2$


LS SVM
$\phi(x_i),\ x_i \in \mathcal{C}_1$
e_i
$\phi(x_i),\ x_i \in \mathcal{C}_2$

Problem Statement

Context: All data are labelled, we classify the next incoming one:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ in $k = 2$ classes.

Problem Statement

Context: All data are labelled, we classify the next incoming one:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ in $k = 2$ classes.
- For kernel $K(x, y) = \phi(x)^\mathsf{T}\phi(y)$, $\phi(x) \in \mathbb{R}^q$, find hyperplane directed by (w, b) to "isolate each class".

$$(w, b) = \operatorname{argmin}_{w \in \mathbb{R}^{q-1}} \|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} c(x_i; w, b)$$

for a certain cost function $c(x; w, b)$.

Problem Statement

Context: All data are labelled, we classify the next incoming one:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ in $k = 2$ classes.
- For kernel $K(x,y) = \phi(x)^\mathsf{T}\phi(y)$, $\phi(x) \in \mathbb{R}^q$, find hyperplane directed by (w, b) to "isolate each class".

$$(w,b) = \operatorname{argmin}_{w \in \mathbb{R}^{q-1}} \|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} c(x_i; w, b)$$

for a certain cost function $c(x; w, b)$.

Solutions:

- **Classical SVM**:

$$c(x_i; w, b) = \imath_{\{y_i(w^\mathsf{T}\phi(x_i)+b) \geq 1\}}$$

with $y_i = \pm 1$ depending on class.
⇒ Solved by quadratic programming methods.
⇒ Analysis requires joint RMT + convex optimization tools (very interesting but left for later...).

Problem Statement

Context: All data are labelled, we classify the next incoming one:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ in $k = 2$ classes.
- For kernel $K(x,y) = \phi(x)^\mathsf{T}\phi(y)$, $\phi(x) \in \mathbb{R}^q$, find hyperplane directed by (w, b) to "isolate each class".

$$(w, b) = \operatorname{argmin}_{w \in \mathbb{R}^{q-1}} \|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} c(x_i; w, b)$$

for a certain cost function $c(x; w, b)$.

Solutions:

- **Classical SVM**:

$$c(x_i; w, b) = \imath_{\{y_i(w^\mathsf{T}\phi(x_i)+b)\geq 1\}}$$

with $y_i = \pm 1$ depending on class.
⇒ Solved by quadratic programming methods.
⇒ Analysis requires joint RMT + convex optimization tools (very interesting but left for later...).

- **LS SVM**:

$$c(x_i; w, b) = \gamma e_i^2 \equiv \gamma(y_i - w^\mathsf{T}\phi(x_i) - b)^2.$$

⇒ Explicit solution (but not sparse!).

LS SVM

Since $w = \sum_{i=1}^{n} \alpha_i \phi(x_i)$, for new datum x, decision based on (sign of)

$$g(x) = \alpha^\mathsf{T} K(\cdot, x) + b$$

with $K(x_i, x_j) = f\left(\frac{1}{p}\|x_i - x_j\|^2\right)$ (Mercer Conditions)
and where $\alpha \in \mathbb{R}^n$ and b given by

$$\alpha = Q\left(I_n - \frac{1_n 1_n^\mathsf{T} Q}{1_n^\mathsf{T} Q 1_n}\right) y$$

$$b = \frac{1_n^\mathsf{T} Q y}{1_n^\mathsf{T} Q 1_n}$$

where $Q = (K + \frac{n}{\gamma} I_n)^{-1}$, $y = [y_i]_{i=1}^n$, $\gamma > 0$ some parameter to set.

LS SVM

Since $w = \sum_{i=1}^n \alpha_i \phi(x_i)$, for new datum x, decision based on (sign of)

$$g(x) = \alpha^{\mathsf{T}} K(\cdot, x) + b$$

with $K(x_i, x_j) = f\left(\frac{1}{p}\|x_i - x_j\|^2\right)$ (Mercer Conditions)
and where $\alpha \in \mathbb{R}^n$ and b given by

$$\alpha = Q\left(I_n - \frac{1_n 1_n^{\mathsf{T}} Q}{1_n^{\mathsf{T}} Q 1_n}\right) y$$

$$b = \frac{1_n^{\mathsf{T}} Q y}{1_n^{\mathsf{T}} Q 1_n}$$

where $Q = (K + \frac{n}{\gamma} I_n)^{-1}$, $y = [y_i]_{i=1}^n$, $\gamma > 0$ some parameter to set.

Objectives:

- Study behavior of $g(x)$

LS SVM

Since $w = \sum_{i=1}^{n} \alpha_i \phi(x_i)$, for new datum x, decision based on (sign of)

$$g(x) = \alpha^{\mathsf{T}} K(\cdot, x) + b$$

with $K(x_i, x_j) = f\left(\frac{1}{p}\|x_i - x_j\|^2\right)$ (Mercer Conditions)
and where $\alpha \in \mathbb{R}^n$ and b given by

$$\alpha = Q\left(I_n - \frac{1_n 1_n^{\mathsf{T}} Q}{1_n^{\mathsf{T}} Q 1_n}\right) y$$

$$b = \frac{1_n^{\mathsf{T}} Q y}{1_n^{\mathsf{T}} Q 1_n}$$

where $Q = (K + \frac{n}{\gamma} I_n)^{-1}$, $y = [y_i]_{i=1}^n$, $\gamma > 0$ some parameter to set.

Objectives:

- Study behavior of $g(x)$
- For $x \in \mathcal{C}_a$, determine probability of success.

LS SVM

Since $w = \sum_{i=1}^n \alpha_i \phi(x_i)$, for new datum x, decision based on (sign of)

$$g(x) = \alpha^\mathsf{T} K(\cdot, x) + b$$

with $K(x_i, x_j) = f\left(\frac{1}{p}\|x_i - x_j\|^2\right)$ (Mercer Conditions)
and where $\alpha \in \mathbb{R}^n$ and b given by

$$\alpha = Q\left(I_n - \frac{1_n 1_n^\mathsf{T} Q}{1_n^\mathsf{T} Q 1_n}\right) y$$

$$b = \frac{1_n^\mathsf{T} Q y}{1_n^\mathsf{T} Q 1_n}$$

where $Q = (K + \frac{n}{\gamma} I_n)^{-1}$, $y = [y_i]_{i=1}^n$, $\gamma > 0$ some parameter to set.

Objectives:

- Study behavior of $g(x)$
- For $x \in \mathcal{C}_a$, determine probability of success.
- Optimize the parameter γ and the kernel K.

Results

As before, $x_i \sim \mathcal{N}(\mu_a, C_a)$, $a = 1, \ldots, k$, with identical growth conditions, here for $k = 2$.

Results

As before, $x_i \sim \mathcal{N}(\mu_a, C_a)$, $a = 1, \ldots, k$, with identical growth conditions, here for $k = 2$.

Results: As $n, p \to \infty$,

- in the first order

$$g(x) = \frac{n_2 - n_1}{n} + \frac{0}{\sqrt{p}} + \underbrace{\frac{G(x)}{p}}_{\text{Relevant terms here!}}$$

Results

As before, $x_i \sim \mathcal{N}(\mu_a, C_a)$, $a = 1, \ldots, k$, with identical growth conditions, here for $k = 2$.

Results: As $n, p \to \infty$,

- in the first order

$$g(x) = \frac{n_2 - n_1}{n} + \frac{0}{\sqrt{p}} + \underbrace{\frac{G(x)}{p}}_{\text{Relevant terms here!}}$$

- asymptotic Gaussian behavior of $G(x)$:

Theorem

For $x \in \mathcal{C}_b$, $G(x) - G_b \to 0$, $G_b \sim \mathcal{N}(m_b, \sigma_b^2)$, where

$$m_b = \begin{cases} -2c_2 \cdot c_1 c_2 \gamma \mathcal{D}, & b = 1 \\ +2c_1 \cdot c_1 c_2 \gamma \mathcal{D}, & b = 2 \end{cases}$$

$$\mathcal{D} = -2f'(\tau)\|\mu_2 - \mu_1\|^2 + \frac{f''(\tau)}{p}\left(\mathit{tr}\,(C_2 - C_1)\right)^2 + \frac{2f''(\tau)}{p}\mathit{tr}\left((C_2 - C_1)^2\right)$$

$$\sigma_b^2 = 8\gamma^2 c_1^2 c_2^2 \left[\frac{(f''(\tau))^2}{p^2}\left(\mathit{tr}\,(C_2 - C_1)\right)^2 \mathit{tr}C_b^2 + 2\left(f'(\tau)\right)^2 (\mu_2 - \mu_1)^{\mathsf{T}} C_b (\mu_2 - \mu_1) + \frac{2\,(f'(\tau))^2}{n}\left(\frac{\mathit{tr}C_1 C_b}{c_1} + \frac{\mathit{tr}C_2 C_b}{c_2}\right)\right]$$

Results

Consequences:

- Strong class-size bias
  ⇒ Proper threshold must depend on $n_2 - n_1$.

Results

Consequences:

- Strong class-size bias
  $\Rightarrow$ Proper threshold must depend on $n_2 - n_1$.
- Natural cancellation of $O(n^{-\frac{1}{2}})$ terms.
  $\Rightarrow$ Similar effect as observed in (properly normalized) kernel spectral clustering.
- Choice of γ asymptotically irrelevant.

Results

Consequences:

- Strong class-size bias
  $\Rightarrow$ Proper threshold must depend on $n_2 - n_1$.
- Natural cancellation of $O(n^{-\frac{1}{2}})$ terms.
  $\Rightarrow$ Similar effect as observed in (properly normalized) kernel spectral clustering.
- Choice of γ asymptotically irrelevant.
- Need to choose $f'(\tau) < 0$ and $f''(\tau) > 0$ (not the case for clustering or SSL!)

Theory and simulations of $g(x)$

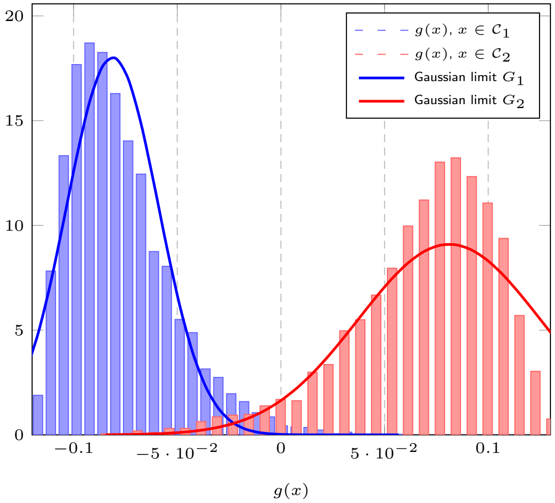


Figure: Values of $g(x)$ for MNIST data (1's and 7's), $n = 256$, $p = 784$, standard Gaussian kernel.

Classification performance

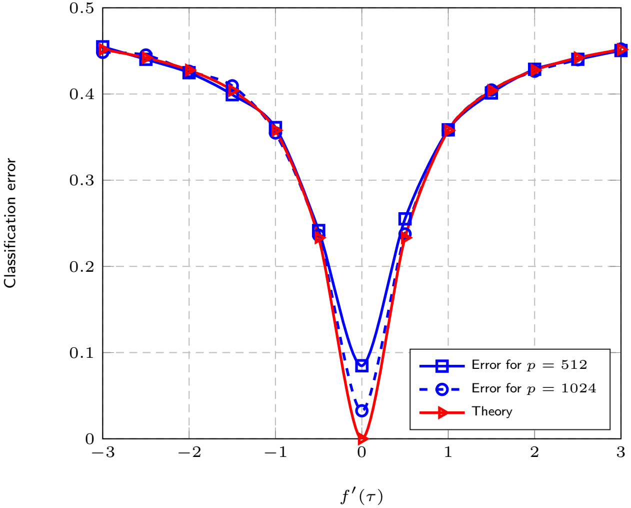


Figure: Performance of LS-SVM, $c_0 = 1/4, c_1 = c_2 = 1/2, \gamma = 1$, polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x \in \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$.

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

Problem Statement

General plan for the study of neural networks:

- Objective is to study performance of neural networks:

Problem Statement

General plan for the study of neural networks:

- Objective is to study performance of neural networks:
  - linear or not (linear is easy but not interesting, non-linear is hard)

Problem Statement

General plan for the study of neural networks:

- Objective is to study performance of neural networks:
  - linear or not (linear is easy but not interesting, non-linear is hard)
  - from shallow to deep

Problem Statement

General plan for the study of neural networks:

- Objective is to study performance of neural networks:
  - linear or not (linear is easy but not interesting, non-linear is hard)
  - from shallow to deep
  - recurrent or not (dynamic systems, stability considerations)

Problem Statement

General plan for the study of neural networks:

- Objective is to study performance of neural networks:
  - linear or not (linear is easy but not interesting, non-linear is hard)
  - from shallow to deep
  - recurrent or not (dynamic systems, stability considerations)
  - back-propagated or not (LS regression versus gradient descent approaches)

Problem Statement

General plan for the study of neural networks:

- Objective is to study performance of neural networks:
  - linear or not (linear is easy but not interesting, non-linear is hard)
  - from shallow to deep
  - recurrent or not (dynamic systems, stability considerations)
  - back-propagated or not (LS regression versus gradient descent approaches)

- Starting point: simple networks

Problem Statement

General plan for the study of neural networks:

- Objective is to study performance of neural networks:
  - linear or not (linear is easy but not interesting, non-linear is hard)
  - from shallow to deep
  - recurrent or not (dynamic systems, stability considerations)
  - back-propagated or not (LS regression versus gradient descent approaches)

- Starting point: simple networks
  - Extreme learning machines: single layer, randomly connected input, LS regressed output.

Problem Statement

General plan for the study of neural networks:

- Objective is to study performance of neural networks:
  - linear or not (linear is easy but not interesting, non-linear is hard)
  - from shallow to deep
  - recurrent or not (dynamic systems, stability considerations)
  - back-propagated or not (LS regression versus gradient descent approaches)

- Starting point: simple networks
  - Extreme learning machines: single layer, randomly connected input, LS regressed output.
  - Echo-state networks: single **interconnected** layer, randomly connected input, LS regressed output.

Problem Statement

General plan for the study of neural networks:

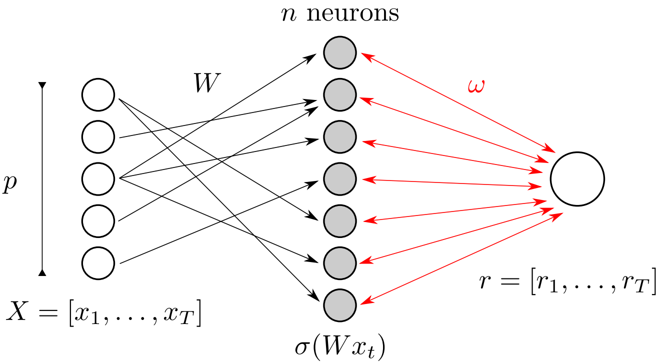
- Objective is to study performance of neural networks:
  - linear or not (linear is easy but not interesting, non-linear is hard)
  - from shallow to deep
  - recurrent or not (dynamic systems, stability considerations)
  - back-propagated or not (LS regression versus gradient descent approaches)

- Starting point: simple networks
  - Extreme learning machines: single layer, randomly connected input, LS regressed output.
  - Echo-state networks: single **interconnected** layer, randomly connected input, LS regressed output.
  - Deeper structures: back-propagation of error.

Extreme Learning Machines

Context: for a learning period T

- input vectors $x_1, \ldots, x_T \in \mathbb{R}^p$, output scalars (or binary values) $r_1, \ldots, r_T \in \mathbb{R}$
- n-neuron layer, randomly connected input $W \in \mathbb{R}^{n \times p}$
- ridge-regressed output $\omega \in \mathbb{R}^n$
- non-linear activation function σ.



Extreme Learning Machines

Objectives: evaluate training and testing MSE performance as $n, p, T \to \infty$

Extreme Learning Machines

Objectives: evaluate training and testing MSE performance as $n, p, T \to \infty$

- Training MSE:

$$E_\gamma(X, r) = \frac{1}{T}\|r - \omega^\mathsf{T}\Sigma\|^2$$

with

$$\Sigma = [\sigma(Wx_1), \ldots, \sigma(Wx_T)]$$

$$\omega = \frac{1}{T}\Sigma\left(\frac{1}{T}\Sigma^\mathsf{T}\Sigma + \gamma I_T\right)^{-1} r.$$

Extreme Learning Machines

Objectives: evaluate training and testing MSE performance as $n,p,T\to\infty$

- Training MSE:

$$E_\gamma(X,r)=\frac{1}{T}\|r-\omega^{\mathsf{T}}\Sigma\|^2$$

  with

$$\Sigma=[\sigma(Wx_1),\ldots,\sigma(Wx_T)]$$
$$\omega=\frac{1}{T}\Sigma\left(\frac{1}{T}\Sigma^{\mathsf{T}}\Sigma+\gamma I_T\right)^{-1}r.$$

- Testing MSE: upon new pair $(\hat{X},\hat{r})$ of length $\hat{T}$,

$$\hat{E}_\gamma(X,r;\hat{X},\hat{r})=\frac{1}{\hat{T}}\|\hat{r}-\omega^{\mathsf{T}}\sigma(W\hat{X})\|^2.$$

Extreme Learning Machines

Objectives: evaluate training and testing MSE performance as $n, p, T \to \infty$

- Training MSE:

$$E_\gamma(X,r) = \frac{1}{T}\|r - \omega^\mathsf{T}\Sigma\|^2$$

  with

$$\Sigma = [\sigma(Wx_1), \ldots, \sigma(Wx_T)]$$
$$\omega = \frac{1}{T}\Sigma\left(\frac{1}{T}\Sigma^\mathsf{T}\Sigma + \gamma I_T\right)^{-1} r.$$

- Testing MSE: upon new pair $(\hat{X}, \hat{r})$ of length $\hat{T}$,

$$\hat{E}_\gamma(X, r; \hat{X}, \hat{r}) = \frac{1}{\hat{T}}\|\hat{r} - \omega^\mathsf{T}\sigma(W\hat{X})\|^2.$$

- Optimize over γ.

Technical Aspects

Training MSE:

- Training MSE given by

$$E_\gamma(X,r) = \gamma^2 \frac{1}{T} r^\mathsf{T} Q_\gamma^2 r$$

$$Q_\gamma = \left(\frac{1}{T} \Sigma^\mathsf{T} \Sigma + \gamma I_T \right)^{-1}.$$

Technical Aspects

Training MSE:

- Training MSE given by

$$\begin{aligned} E_\gamma(X,r) &= \gamma^2 \frac{1}{T} r^\mathsf{T} Q_\gamma^2 r \\ Q_\gamma &= \left(\frac{1}{T} \Sigma^\mathsf{T} \Sigma + \gamma I_T \right)^{-1}. \end{aligned}$$

- Testing MSE given by

$$\hat{E}_\gamma(X,r;\hat{X},\hat{r}) = \frac{1}{\hat{T}} \left\| \hat{r} - \frac{1}{T} \sigma(W\hat{X})^\mathsf{T} \Sigma Q_\gamma r \right\|^2$$

Technical Aspects

Training MSE:

- Training MSE given by

$$E_\gamma(X,r) = \gamma^2 \frac{1}{T} r^\mathsf{T} Q_\gamma^2 r$$

$$Q_\gamma = \left(\frac{1}{T}\Sigma^\mathsf{T}\Sigma + \gamma I_T \right)^{-1}.$$

- Testing MSE given by

$$\hat{E}_\gamma(X,r;\hat{X},\hat{r}) = \frac{1}{\hat{T}} \left\| \hat{r} - \frac{1}{T}\sigma(W\hat{X})^\mathsf{T}\Sigma Q_\gamma r \right\|^2$$

- Requires first a deterministic equivalent $\bar{Q}_\gamma$ for Q_γ with non-linear $\sigma(\cdot)$.

Technical Aspects

Training MSE:

- Training MSE given by

$$E_\gamma(X,r) = \gamma^2 \frac{1}{T} r^\mathsf{T} Q_\gamma^2 r$$
$$Q_\gamma = \left(\frac{1}{T} \Sigma^\mathsf{T} \Sigma + \gamma I_T \right)^{-1}.$$

- Testing MSE given by

$$\hat{E}_\gamma(X,r;\hat{X},\hat{r}) = \frac{1}{\hat{T}} \left\| \hat{r} - \frac{1}{T} \sigma(W\hat{X})^\mathsf{T} \Sigma Q_\gamma r \right\|^2$$

- Requires first a deterministic equivalent $\bar{Q}_\gamma$ for Q_γ with non-linear $\sigma(\cdot)$.
- Then deterministic approximation of $\frac{1}{T}\sigma(Wa)^\mathsf{T} \Sigma Q_\gamma b$ for deterministic a, b.

Technical Aspects

Main technical difficulty: $\Sigma = \sigma(WX) \in \mathbb{R}^{n \times T}$ has

- independent rows
- a highly non trivial columns dependence!

Technical Aspects

Main technical difficulty: $\Sigma = \sigma(WX) \in \mathbb{R}^{n \times T}$ has

- independent rows
- a highly non trivial columns dependence!

Broken trace lemma!: for $w \sim \mathcal{N}(0, n^{-1} I_n)$, X, A deterministic of bounded norm,

$$w^\mathsf{T} X A X^\mathsf{T} w \simeq \frac{1}{n} \operatorname{tr} X A X^\mathsf{T}$$

Technical Aspects

Main technical difficulty: $\Sigma = \sigma(WX) \in \mathbb{R}^{n\times T}$ has

- independent rows
- a highly non trivial columns dependence!

Broken trace lemma!: for $w \sim \mathcal{N}(0, n^{-1}I_n)$, X, A deterministic of bounded norm,

$$w^{\mathsf{T}} X A X^{\mathsf{T}} w \simeq \frac{1}{n}\operatorname{tr} X A X^{\mathsf{T}}$$

BUT what about:

$$\sigma(w^{\mathsf{T}} X) A \sigma(X^{\mathsf{T}} w) \simeq ?$$

Technical Aspects

Updated trace lemma:

Lemma

For A deterministic and $\sigma(t)$ Lipschitz, $w \in \mathbb{R}^p$ with i.i.d. entries, $E[w_i] = 0$, $E[w_i^k] = \frac{m_k}{n^{k/2}}$,

$$\frac{1}{T}\sigma(w^\mathsf{T} X) A \sigma(X^\mathsf{T} w) - \frac{1}{T} \mathit{tr}\, \Phi_X A \xrightarrow{\text{a.s.}} 0$$

with

$$\Phi_X = E\left[\sigma(X^\mathsf{T} w)\sigma(w^\mathsf{T} X)\right].$$

Technical Aspects

Updated trace lemma:

Lemma

For A deterministic and $\sigma(t)$ Lipschitz, $w \in \mathbb{R}^p$ with i.i.d. entries, $E[w_i] = 0$, $E[w_i^k] = \frac{m_k}{n^{k/2}}$,

$$\frac{1}{T}\sigma(w^\mathsf{T} X)A\sigma(X^\mathsf{T} w) - \frac{1}{T}\mathit{tr}\,\Phi_X A \xrightarrow{\text{a.s.}} 0$$

with

$$\Phi_X = E\left[\sigma(X^\mathsf{T} w)\sigma(w^\mathsf{T} X)\right].$$

Technique of proof:

- Use concentration of vector w
- transfer concentration by Lipschitz property through mapping $w \mapsto \sigma(w^\mathsf{T} X)$, i.e.,

$$P\left(f\left(\sigma(w^\mathsf{T} X)\right) - E\left[f\left(\sigma(w^\mathsf{T} X)\right)\right] > t\right) \leq c_1 e^{-c_2 n t^2}$$

for all Lipschitz f (and beyond...), with $c_1, c_2 > 0$.

Results

Results:

- Deterministic equivalent: as $n, p, T \to \infty$ with $\sigma(t)$ smooth, W_{ij} i.i.d. $E[W_{ij}] = 0$, $E[W_{ij}^k] = \frac{m_k}{n^{k/2}}$,

  $$Q_\gamma \leftrightarrow \bar{Q}_\gamma$$

  where

  $$Q_\gamma = \left(\frac{1}{T}\Sigma\Sigma^\mathsf{T} + \gamma I_T\right)^{-1}$$
  $$\bar{Q}_\gamma = \left(\frac{n}{T}\frac{1}{1+\delta}\Phi_X + \gamma I_T\right)^{-1}$$

  with δ unique solution to

  $$\delta = \frac{1}{T}\mathrm{tr}\,\Phi_X\left(\frac{n}{T}\frac{1}{1+\delta}\Phi_X + \gamma I_T\right)^{-1}.$$

Results

Neural Network Performances:

- Training performance:

$$E_\gamma(X,r) \leftrightarrow \gamma^2 \frac{1}{T} r^\mathsf{T} \bar{Q}_\gamma \left[\frac{\frac{1}{n}\operatorname{tr}\left(\Psi_X \bar{Q}_\gamma^2\right)}{1-\frac{1}{n}\operatorname{tr}\left(\Psi_X \bar{Q}_\gamma\right)^2} \Psi_X + I_T \right] \bar{Q}_\gamma r.$$

Results

Neural Network Performances:

- Training performance:

$$E_\gamma(X,r) \leftrightarrow \gamma^2 \frac{1}{T} r^\mathsf{T} \bar{Q}_\gamma \left[\frac{\frac{1}{n}\operatorname{tr}(\Psi_X \bar{Q}_\gamma^2)}{1-\frac{1}{n}\operatorname{tr}(\Psi_X \bar{Q}_\gamma)^2} \Psi_X + I_T \right] \bar{Q}_\gamma r.$$

- Testing performance:

$$\hat{E}_\gamma(X,r;\hat{X},\hat{r}) \leftrightarrow \frac{1}{\hat{T}} \left\| \hat{r} - \Psi_{X,\hat{X}}^\mathsf{T} \bar{Q}_\gamma r \right\|^2 + \frac{\frac{1}{n} r^\mathsf{T} \bar{Q}_\gamma \Psi_X \bar{Q}_\gamma r}{1-\frac{1}{n}\operatorname{tr}(\Psi_X \bar{Q}_\gamma)^2} \times \left[\frac{1}{\hat{T}} \operatorname{tr} \Psi_{\hat{X}} - \frac{\gamma}{\hat{T}} \operatorname{tr}(\bar{Q}_\gamma \Psi_{X,\hat{X}} \Psi_{\hat{X},X} \bar{Q}_\gamma) - \frac{1}{\hat{T}} \operatorname{tr}(\Psi_{\hat{X},X} \bar{Q}_\gamma) \Psi_{X,\hat{X}}) \right].$$

where $\Psi_{A,B} = \frac{n}{T}\frac{1}{1+\delta}\Phi_{A,B}$, $\Psi_A = \Psi_{A,A}$, $\Phi_{A,B} = E[\frac{1}{n}\sigma(WA)^\mathsf{T}\sigma(WB)]$.

Results

Neural Network Performances:

- Training performance:

$$E_\gamma(X,r) \leftrightarrow \gamma^2 \frac{1}{T} r^\mathsf{T} \bar{Q}_\gamma \left[\frac{\frac{1}{n}\mathrm{tr}\,(\Psi_X \bar{Q}_\gamma^2)}{1-\frac{1}{n}\mathrm{tr}\,(\Psi_X \bar{Q}_\gamma)^2} \Psi_X + I_T \right] \bar{Q}_\gamma r.$$

- Testing performance:

$$\hat{E}_\gamma(X,r;\hat{X},\hat{r}) \leftrightarrow \frac{1}{\hat{T}} \left\| \hat{r} - \Psi_{X,\hat{X}}^\mathsf{T} \bar{Q}_\gamma r \right\|^2 + \frac{\frac{1}{n} r^\mathsf{T} \bar{Q}_\gamma \Psi_X \bar{Q}_\gamma r}{1-\frac{1}{n}\mathrm{tr}\,(\Psi_X \bar{Q}_\gamma)^2}$$
$$\times \left[\frac{1}{\hat{T}} \mathrm{tr}\, \Psi_{\hat{X}} - \frac{\gamma}{\hat{T}} \mathrm{tr}\,(\bar{Q}_\gamma \Psi_{X,\hat{X}} \Psi_{\hat{X},X} \bar{Q}_\gamma) - \frac{1}{\hat{T}} \mathrm{tr}\,(\Psi_{\hat{X},X} \bar{Q}_\gamma) \Psi_{X,\hat{X}} \right].$$

where $\Psi_{A,B} = \frac{n}{T}\frac{1}{1+\delta}\Phi_{A,B}$, $\Psi_A = \Psi_{A,A}$, $\Phi_{A,B} = E[\frac{1}{n}\sigma(WA)^\mathsf{T}\sigma(WB)]$.

In the limit where $n/p, n/T \to \infty$, taking $\gamma = \frac{n}{T}\Gamma$:

$$E_\gamma(X,r) \leftrightarrow \frac{1}{T}\Gamma^2 r^\mathsf{T} \left(\Phi_X + \Gamma I_T\right)^{-2} r$$
$$\hat{E}_\gamma(X,r) \leftrightarrow \frac{1}{\hat{T}} \left\| \hat{r} - \Phi_{\hat{X},X} \left(\Phi_X + \Gamma I_T\right)^{-1} r \right\|^2.$$

Results

Special Cases of $\Phi_{A,B}$:

$\sigma(t)$	W_{ij}	$[\Phi_{A,B}]_{ij}$
t	any	$\frac{m_2}{n} a_i^\intercal b_j$
$At^2 + Bt + C$	any	$A^2 \left[\frac{m_2^2}{n^2} \left(2(a_i^\intercal b_j)^2 + \lVert a_i \rVert^2 \lVert b_j \rVert^2 \right) + \frac{m_4 - 3m_2^2}{n^2} (a_i^2)^\intercal (b_j^2) \right]$ $+B^2 \frac{m_2}{n} a_i^\intercal b_j + AB \frac{m_3}{n^{3/2}} \left[(a_i^2)^\intercal b_j + a_i^\intercal (b_j^2) \right]$ $+AC \frac{m_2}{n} \left[\lVert a_i \rVert^2 + \lVert b_j \rVert^2 \right] + C^2$
$\max(t, 0)$	$\mathcal{N}(0, \frac{1}{n})$	$\frac{1}{2\pi n} \lVert a_i \rVert \lVert b_j \rVert \left(Z_{ij} \arccos(-Z_{ij}) + \sqrt{1 - Z_{ij}^2} \right)$
$\mathrm{erf}(t)$	$\mathcal{N}(0, \frac{1}{n})$	$\frac{2}{\pi} \arcsin \left(\frac{2 a_i^\intercal b_j}{\sqrt{(n + 2\lVert a_i \rVert^2)(n + 2\lVert b_j \rVert^2)}} \right)$
$1_{\{t>0\}}$	$\mathcal{N}(0, \frac{1}{n})$	$\frac{1}{2} - \frac{1}{2\pi} \arccos(Z_{ij})$
$\mathrm{sign}(t)$	$\mathcal{N}(0, \frac{1}{n})$	$1 - \frac{2}{\pi} \arccos(Z_{ij})$
$\cos(t)$	$\mathcal{N}(0, \frac{1}{n})$	$\exp \left(-\frac{1}{2} \left[\lVert a_i \rVert^2 + \lVert b_j \rVert^2 \right] \right) \cosh \left(a_i^\intercal b_j \right).$

Figure: $\Phi_{A,B}$ for W_{ij} i.i.d. zero mean, k-th order moments $m_k n^{-\frac{k}{2}}$, $Z_{ij} \equiv \frac{a_i^\intercal b_j}{\lVert a_i \rVert \lVert b_j \rVert}$, $(a^2) = [a_i^2]_{i=1}^n$.

Test on MNIST data

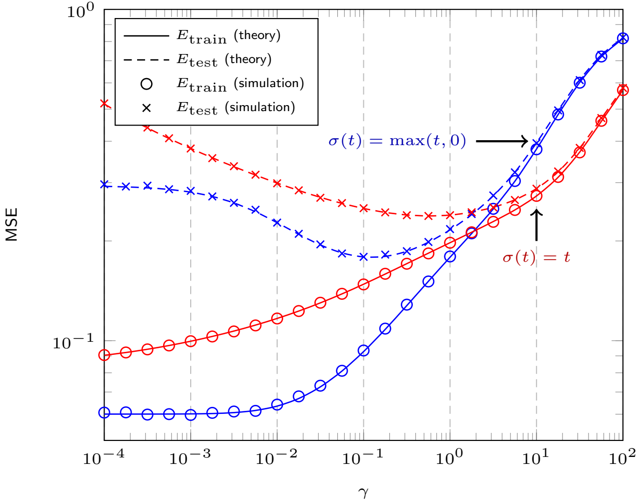


Figure: MSE performance for $\sigma(t) = t$ and $\sigma(t) = \max(t, 0)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = 1024$, $p = 784$.

Test on MNIST data

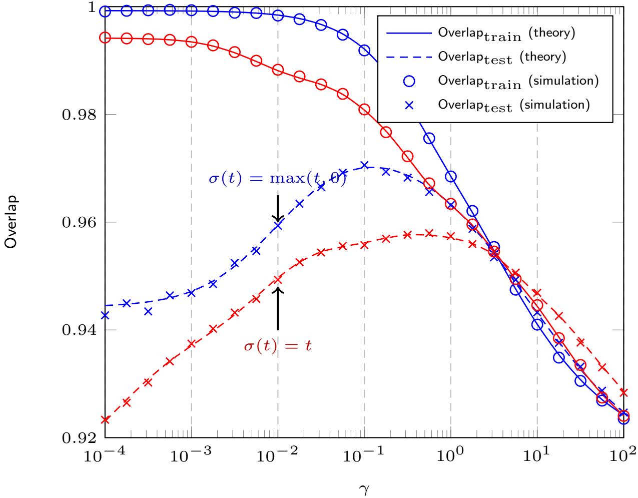


Figure: Overlap performance for $\sigma(t) = t$ and $\sigma(t) = \max(t, 0)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = 1024$, $p = 784$.

Next Investigations

Interpretations and Improvements:

- General formulas for Φ_X, $\Phi_{X,\hat{x}}$
- On-line optimization of γ, $\sigma(\cdot)$, n?

Next Investigations

Interpretations and Improvements:

- General formulas for Φ_X, $\Phi_{X,\hat{x}}$
- On-line optimization of γ, $\sigma(\cdot)$, n?

Generalizations:

- Multi-layer ELM?
- Optimize layers vs. number of neurons?
- Backpropagation error analysis?
- Connection to auto-encoders?
- Introduction of non-linearity to more involved structures (ESN, deep nets?).

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models

Applications

- Random Matrices and Robust Estimation
- Spectral Clustering Methods and Random Matrices
- Community Detection on Graphs
- Kernel Spectral Clustering
- Kernel Spectral Clustering: Subspace Clustering
- Semi-supervised Learning
- Support Vector Machines
- Neural Networks: Extreme Learning Machines

Perspectives

Summary of Results and Perspectives I

Robust statistics.

- ✔ Tyler, Maronna (and regularized) estimators
- ✔ Elliptical data setting, deterministic outlier setting
- ✔ Central limit theorem extensions
- 💡 Joint mean and covariance robust estimation
- 💡 Study of robust regression (preliminary works exist already using strikingly different approaches)

Applications.

- ✔ Statistical finance (portfolio estimation)
- ✔ Localisation in array processing (robust GMUSIC)
- ✔ Detectors in space time array processing

References.

R. Couillet, F. Pascal, J. W. Silverstein, "Robust Estimates of Covariance Matrices in the Large Dimensional Regime", IEEE Transactions on Information Theory, vol. 60, no. 11, pp. 7269-7278, 2014.

R. Couillet, F. Pascal, J. W. Silverstein, "The Random Matrix Regime of Maronna's M-estimator with elliptically distributed samples", Elsevier Journal of Multivariate Analysis, vol. 139, pp. 56-78, 2015.

Summary of Results and Perspectives II

T. Zhang, X. Cheng, A. Singer, "Marchenko-Pastur Law for Tyler's and Maronna's M-estimators", arXiv:1401.3424, 2014.

R. Couillet, M. McKay, "Large Dimensional Analysis and Optimization of Robust Shrinkage Covariance Matrix Estimators", Elsevier Journal of Multivariate Analysis, vol. 131, pp. 99-120, 2014.

D. Morales-Jimenez, R. Couillet, M. McKay, "Large Dimensional Analysis of Robust M-Estimators of Covariance with Outliers", IEEE Transactions on Signal Processing, vol. 63, no. 21, pp. 5784-5797, 2015.

L. Yang, R. Couillet, M. McKay, "A Robust Statistics Approach to Minimum Variance Portfolio Optimization", IEEE Transactions on Signal Processing, vol. 63, no. 24, pp. 6684–6697, 2015.

R. Couillet, "Robust spiked random matrices and a robust G-MUSIC estimator", Elsevier Journal of Multivariate Analysis, vol. 140, pp. 139-161, 2015.

A. Kammoun, R. Couillet, F. Pascal, M.-S. Alouini, "Optimal Design of the Adaptive Normalized Matched Filter Detector", (submitted to) IEEE Transactions on Information Theory, 2016, arXiv Preprint 1504.01252.

R. Couillet, A. Kammoun, F. Pascal, "Second order statistics of robust estimators of scatter. Application to GLRT detection for elliptical signals", Elsevier Journal of Multivariate Analysis, vol. 143, pp. 249-274, 2016.

Summary of Results and Perspectives III

D. Donoho, A. Montanari, "High dimensional robust m-estimation: Asymptotic variance via approximate message passing", Probability Theory and Related Fields, 1-35, 2013.

N. El Karoui, "Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results." arXiv preprint arXiv:1311.2445, 2013.

Summary of Results and Perspectives I

Kernel methods.

- ✔ Subspace spectral clustering
- ✔ Subspace spectral clustering for $f'(\tau) = 0$
- ✎ Spectral clustering with outer product kernel $f(x^\mathsf{T} y)$
- ✔ Semi-supervised learning, kernel approaches.
- ✔ Least square support vector machines (LS-SVM).
- ✎ Support vector machines (SVM).

Applications.

- ✔ Massive MIMO user clustering

References.

N. El Karoui, "The spectrum of kernel random matrices", The Annals of Statistics, 38(1), 1-50, 2010.

R. Couillet, F. Benaych-Georges, "Kernel Spectral Clustering of Large Dimensional Data", Electronic Journal of Statistics, vol. 10, no. 1, pp. 1393-1454, 2016.

R. Couillet, A. Kammoun, "Random Matrix Improved Subspace Clustering", Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2016.

Summary of Results and Perspectives II

Z. Liao, R. Couillet, “Random matrices meet machine learning: a large dimensional analysis of LS-SVM”, (submitted to) IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’17), New Orleans, USA, 2017.

X. Mai, R. Couillet, “The counterintuitive mechanism of graph-based semi-supervised learning in the big data regime”, (submitted to) IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’17), New Orleans, USA, 2017.

Summary of Results and Perspectives I

Community detection.

- ✔ Complete study of eigenvector contents in adjacency/modularity methods.
- 💡 Study of Bethe Hessian approach for the DCSBM model.
- 💡 Analysis of non-necessarily spectral approaches (wavelet approaches).

References.

H. Tiomoko Ali, R. Couillet, "Spectral community detection in heterogeneous large networks", (submitted to) Journal of Multivariate Analysis, 2016.

F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, P. Zhang, "Spectral redemption in clustering sparse networks. Proceedings of the National Academy of Sciences", 110(52), 20935-20940, 2013.

C. Bordenave, M. Lelarge, L. Massoulié, "Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs", Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on, pp. 1347-1357, 2015

A. Saade, F. Krzakala, L. Zdeborová, "Spectral clustering of graphs with the Bethe Hessian", In Advances in Neural Information Processing Systems, pp. 406-414, 2014.

Summary of Results and Perspectives I

Neural Networks.

- ✔ Non-linear extreme learning machines (ELM)
- ✎ Multi-layer ELM
- 💡 Backpropagation in ELM
- ✎ Random convolutional networks for image processing
- ✔ Linear echo-state networks (ESN)
- 💡 Non-linear ESN

References.

C. Williams, "Computation with infinite neural networks", Neural Computation, 10(5), 1203-1216, 1998.

N. El Karoui, "Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond", The Annals of Applied Probability, 19(6), 2362-2405, 2009.

C. Louart, R. Couillet, "Harnessing neural networks: a random matrix approach", (submitted to) IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), New Orleans, USA, 2017.

R. Couillet, G. Wainrib, H. Sevi, H. Tiomoko Ali, "The asymptotic performance of linear echo state neural networks", Journal of Machine Learning Research, vol. 17, no. 178, pp. 1-35, 2016.

Summary of Results and Perspectives I

Sparse PCA

- ✔ Spike random matrix sparse PCA
- ✎ Sparse kernel PCA

References.

R. Couillet, M. McKay, "Optimal block-sparse PCA for high dimensional correlated samples", (submitted to) Journal of Multivariate Analysis, 2016.

Signal processing on graphs, distributed optimization, etc.

- 💡 Turning signal processing on graph methods random.
- 💡 Random matrix analysis of diffusion networks performance.

The End

Thank you.