Random Matrices and Machine Learning
(Summer School on "Large Random Matrices and High Dimensional Statistical Signal Processing")

Romain COULLIET

CentraleSupélec (Paris, France)

June 8, 2016

CentraleSupélec

Outline

Outline

Context and Taskforce

General theme:

Understand and improve machine learning methods in the large dimensional regime

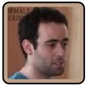
Context and Taskforce

General theme:

Understand and improve machine learning methods in the large dimensional regime

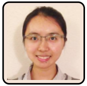
Collaborators:

Florent BENAYCH-GEORGES (Professor)
Kernel Spectral Clustering



Gilles WAINRIB (Assistant Professor)
Cosme LOUART (Intern)
Neural Networks

Hafiz TIOMOKO ALI (PhD student)
Community detection on graphs



Xiaoyi MAI (Intern)
Semi-supervised learning

Zhenyu LIAO (Intern)
Support vector machines

Outline

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: $x_1, \dots, x_n \in \mathbb{C}^N$ (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: $x_1, \ldots, x_n \in \mathbb{C}^N$ (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

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 $n = 100 \times N$.

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

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

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

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

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

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

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

$\Rightarrow$ no convergence in spectral norm.

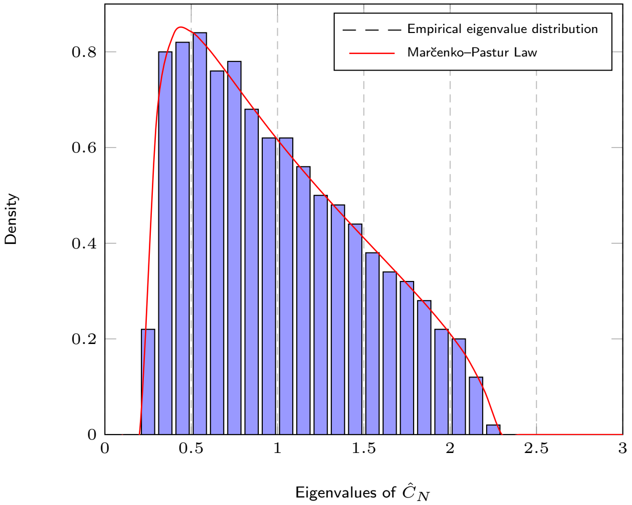


Figure: Histogram of the eigenvalues of $\hat{C}_N$ for $N = 500$, $n = 2000$, $C_N = I_N$.

Definition (Empirical Spectral Density)

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

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

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

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

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

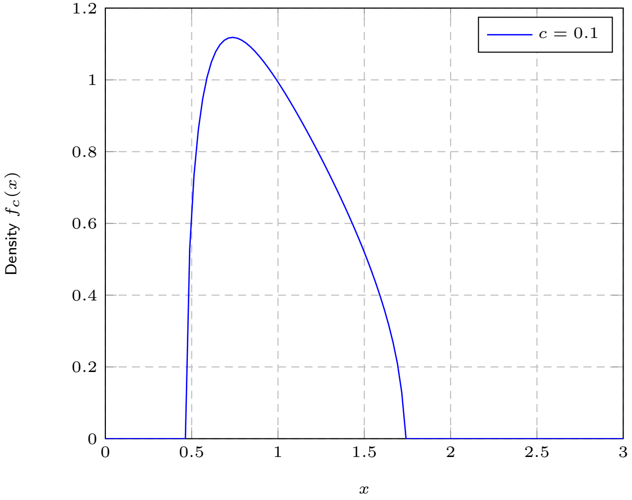


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{N\to\infty} N/n$.

The Marčenko–Pastur law

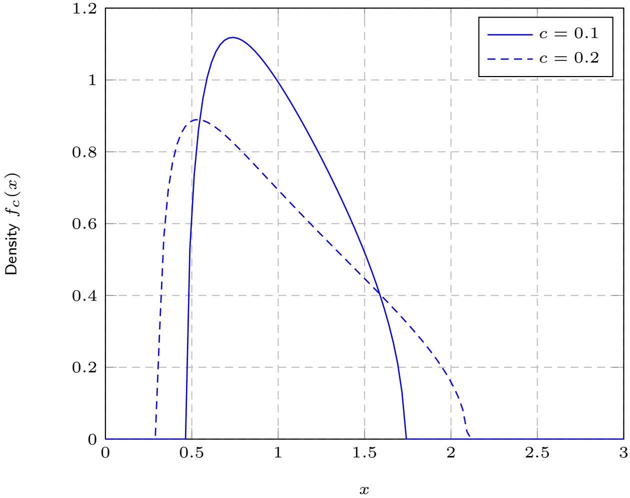


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{N \to \infty} N/n$.

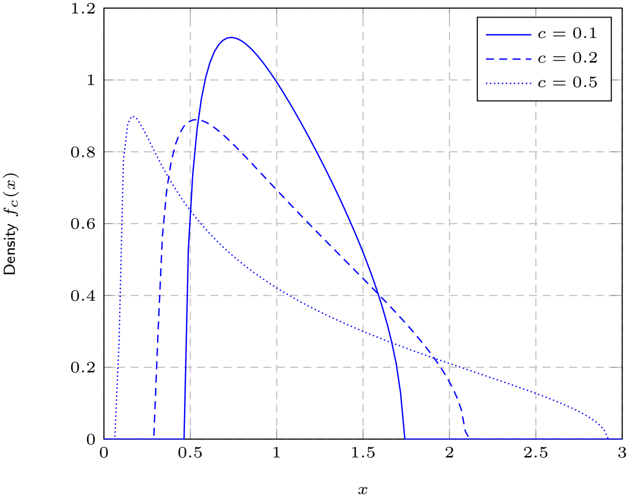


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{N\to\infty} N/n$.

Let X_N with i.i.d. (0,1) entries, P a rank-K matrix with K finite as $N, n \to \infty$. In either of these scenarios:

$$\hat{C}_N = (I_N + P)^{\frac{1}{2}} \frac{1}{n} X_N X_N^* (I_N + P)^{\frac{1}{2}}$$
$$\hat{C}_N = \frac{1}{n}(X_N + P)(X_N + P)^*$$
$$\hat{C}_N = \frac{1}{n} X_N X_N^* + P$$

we have $\mu_N \xrightarrow{\text{a.s.}} \mu_c$ but some eigenvalues can escape the support!

Spiked models

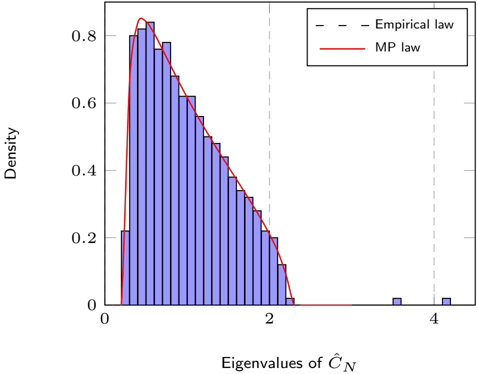
Let X_N with i.i.d. (0,1) entries, P a rank-K matrix with K finite as $N, n \to \infty$. In either of these scenarios:

$$\hat{C}_N = (I_N + P)^{\frac{1}{2}} \frac{1}{n} X_N X_N^* (I_N + P)^{\frac{1}{2}}$$

$$\hat{C}_N = \frac{1}{n}(X_N + P)(X_N + P)^*$$

$$\hat{C}_N = \frac{1}{n} X_N X_N^* + P$$

we have $\mu_N \xrightarrow{\text{a.s.}} \mu_c$ but some eigenvalues can escape the support!



Spiked models

Two fundamental properties (assume here $\hat{C}_N = (I_N + P)^{\frac{1}{2}} \frac{1}{n} X_N X_N^* (I_N + P)^{\frac{1}{2}}$):

- **Phase transition phenomenon:** for $\omega_1 > \ldots > \omega_K \geq 0$ eigenvalues of P,

$$\lambda_i(\hat{C}_N) \xrightarrow{\text{a.s.}} \begin{cases} (1+\sqrt{c})^2, & \omega_i < \sqrt{c} \\ 1 + \omega_i + c\frac{1+\omega_i}{\omega_i}, & \omega_i \geq \sqrt{c} \end{cases}$$

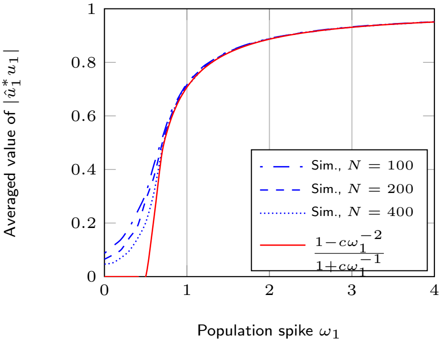
Two fundamental properties (assume here $\hat{C}_N = (I_N + P)^{\frac{1}{2}} \frac{1}{n} X_N X_N^* (I_N + P)^{\frac{1}{2}}$):

- **Phase transition phenomenon:** for $\omega_1 > \ldots > \omega_K \geq 0$ eigenvalues of P,

$$\lambda_i(\hat{C}_N) \xrightarrow{\text{a.s.}} \begin{cases} (1+\sqrt{c})^2, & \omega_i < \sqrt{c} \\ 1 + \omega_i + c\frac{1+\omega_i}{\omega_i}, & \omega_i \geq \sqrt{c} \end{cases}$$

- **Eigenvector angle:** for $u_1, \ldots, u_K$ eigenvectors of P and $\hat{u}_1, \ldots, \hat{u}_N$ of $\hat{C}_N$,

$$|\hat{u}_i^* u_i|^2 \xrightarrow{\text{a.s.}} \begin{cases} 0, & \omega_i < \sqrt{c} \\ \frac{1-cw_i^{-2}}{1+c\omega_i^{-1}}, & \omega_i \geq \sqrt{c} \end{cases}$$



Other classical examples.

- If $X_N \in \mathbb{C}^{N\times N}$ Hermitian with i.i.d. entries of mean 0, variance $1/N$, then (almost surely) $\mu_N \to \mu$ where μ has density f the semi-circle law

$$f(x) = \frac{1}{2\pi}\sqrt{(4-x^2)^+}.$$

- If $X_N \in \mathbb{C}^{N\times N}$ has with i.i.d. 0 mean, variance $1/N$ entries, then asymptotically its complex eigenvalues distribute uniformly on the complex unit circle, i.e. $\mu_N \to \mu$ with density

$$f(z) = \frac{1}{\pi}\delta_{|z|\leq 1}.$$

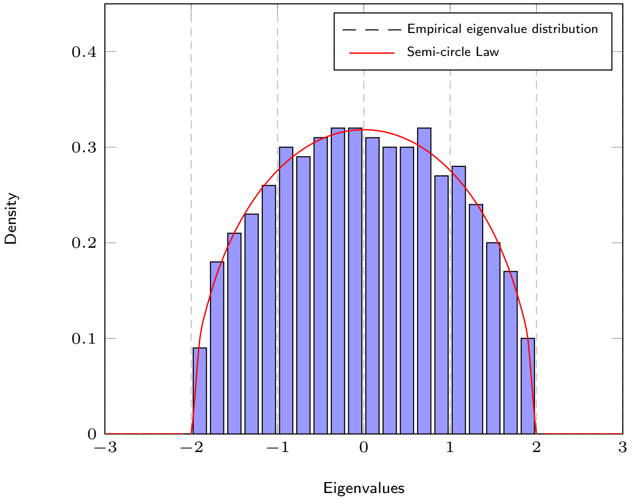


Figure: Histogram of the eigenvalues of Wigner matrices and the semi-circle law, for $N = 500$

Circular law

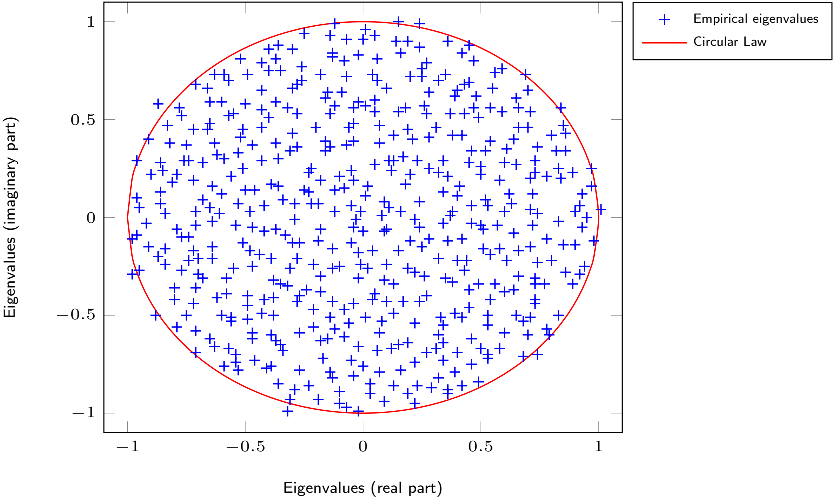


Figure: Eigenvalues of X_N with i.i.d. standard Gaussian entries, for $N = 500$.

Outline

System Setting

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

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

System Setting

Assume n-node **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

Assume n-node **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 $i \in \mathcal{C}_a$, $j \in \mathcal{C}_b$,

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

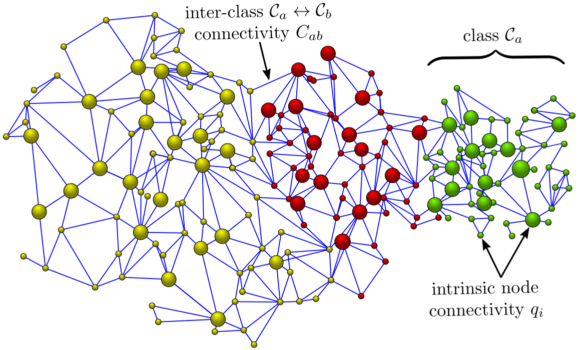
System Setting

Assume n-node **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 $A_{ij} \sim \text{Bernoulli}(q_i q_j C_{ab})$.



System Setting

Objective:

Understand and improve performance of spectral community detection methods:

- based on adjacency A or modularity $A - \frac{dd^\mathsf{T}}{d^\mathsf{T} 1_n}$ matrices (adapted to **dense nets**)

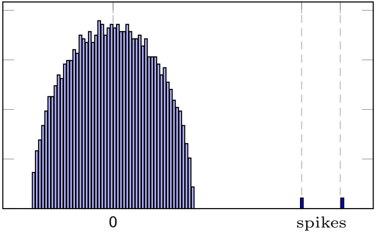
System Setting

Objective:
Understand and improve performance of spectral community detection methods:

- based on adjacency A or modularity $A - \frac{dd^{\mathsf{T}}}{d^{\mathsf{T}}1_n}$ matrices (adapted to **dense nets**)
- based on Bethe Hessian $(r^2 - 1)I_n - rA + D$ (adapted to **sparse nets!**).

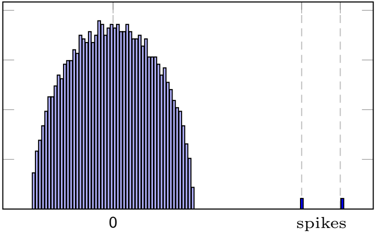
Objective:
Understand and improve performance of spectral community detection methods:

- based on adjacency A or modularity $A - \frac{dd^\mathsf{T}}{d^\mathsf{T} 1_n}$ matrices (adapted to **dense nets**)
- based on Bethe Hessian $(r^2 - 1)I_n - rA + D$ (adapted to **sparse nets!**).

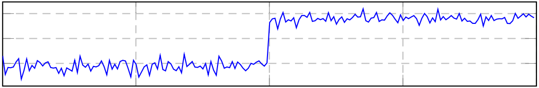


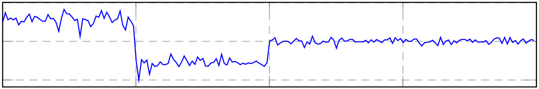
Objective:
Understand and improve performance of spectral community detection methods:

- based on adjacency A or modularity $A - \frac{dd^{\mathsf{T}}}{d^{\mathsf{T}} 1_n}$ matrices (adapted to **dense nets**)
- based on Bethe Hessian $(r^2 - 1)I_n - rA + D$ (adapted to **sparse nets!**).



$\Downarrow$ **Eigenvectors** $\Downarrow$

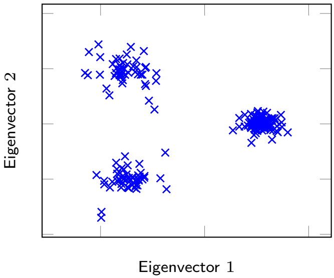




System Setting

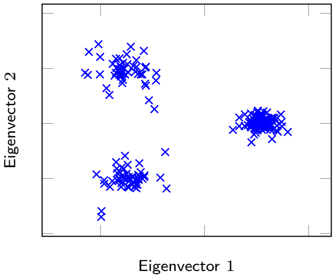
System Setting

$\Downarrow$ p-**dimensional representation** $\Downarrow$



System Setting

⇓ p-**dimensional representation** ⇓



⇓

EM or k-means clustering.

Limitations of Adjacency/Modularity Approach

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

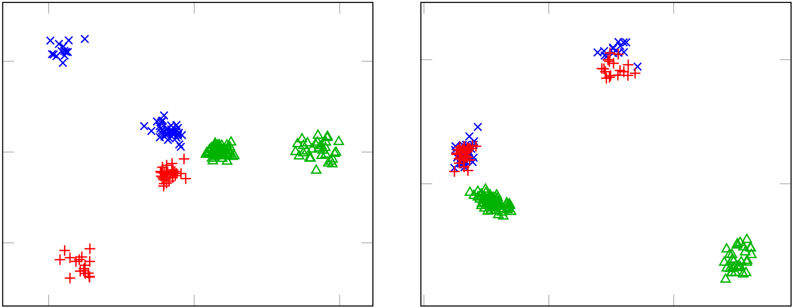
$\rightarrow$ Leading eigenvectors of A (or modularity $A - \frac{dd^\mathsf{T}}{d^\mathsf{T} 1_n}$) biased by q_i distribution.

$\rightarrow$ Similar behavior for Bethe Hessian.

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

$\rightarrow$ Leading eigenvectors of A (or modularity $A - \frac{dd^{\mathsf{T}}}{d^{\mathsf{T}}1_n}$) biased by q_i distribution.

$\rightarrow$ Similar behavior for Bethe Hessian.



(Modularity) (Bethe Hessian)

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 $D = \mathrm{diag}(A1_n) = \mathrm{diag}(d)$ the degree matrix)

$$L_\alpha = n^{2\alpha - \frac{1}{2}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{d^\mathsf{T} 1_n} \right] D^{-\alpha}.$$

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 $D = \operatorname{diag}(A 1_n) = \operatorname{diag}(d)$ the degree matrix)

$$L_\alpha = n^{2\alpha - \frac{1}{2}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{d^\mathsf{T} 1_n} \right] D^{-\alpha}.$$

Our results in a nutshell:

- we find optimal α_{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 $D = \mathrm{diag}(A1_n) = \mathrm{diag}(d)$ the degree matrix)

$$L_\alpha = n^{2\alpha - \frac{1}{2}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{d^\mathsf{T} 1_n} \right] D^{-\alpha}.$$

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)

$$L_\alpha = n^{2\alpha - \frac{1}{2}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{d^\mathsf{T} 1_n} \right] D^{-\alpha}.$$

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.
  $\Rightarrow$ **Never proposed before!**

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*

$$L_\alpha = n^{2\alpha-1} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{d^\mathsf{T} 1_n} \right] D^{-\alpha}$$

$$\tilde{L}_\alpha = \frac{1}{m_\mu^{2\alpha}} \left[\frac{1}{\sqrt{n}} D_q^{-\alpha} X D_q^{-\alpha} + U \Lambda U^\mathsf{T} \right]$$

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

$$U = \left[D_q^{1-\alpha} \frac{J}{\sqrt{n}} \quad \frac{1}{nm_\mu} D_q^{-\alpha} X 1_n \right], \quad \textit{rank } k+1$$

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

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* $\mathcal{C}_a$.

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

$$L_\alpha = n^{2\alpha-1} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{d^\mathsf{T} 1_n} \right] D^{-\alpha}$$

$$\tilde{L}_\alpha = \frac{1}{m_\mu^{2\alpha}} \left[\frac{1}{\sqrt{n}} D_q^{-\alpha} X D_q^{-\alpha} + U \Lambda U^\mathsf{T} \right]$$

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

$$U = \left[D_q^{1-\alpha} \frac{J}{\sqrt{n}} \quad \frac{1}{nm_\mu} D_q^{-\alpha} X 1_n \right], \quad \textit{rank } k+1$$

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

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 $\leftrightarrow$ $\lambda(M) >$ "spectrum edge"
  $\Rightarrow$ **optimal choice** α_{opt} **of** α from study of noise spectrum.

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

$$L_\alpha = n^{2\alpha-1} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{d^\mathsf{T} 1_n} \right] D^{-\alpha}$$

$$\tilde{L}_\alpha = \frac{1}{m_\mu^{2\alpha}} \left[\frac{1}{\sqrt{n}} D_q^{-\alpha} X D_q^{-\alpha} + U \Lambda U^\mathsf{T} \right]$$

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

$$U = \left[D_q^{1-\alpha} \frac{J}{\sqrt{n}} \quad \frac{1}{nm_\mu} D_q^{-\alpha} X 1_n \right], \quad \textit{rank } k+1$$

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

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 $\leftrightarrow$ $\lambda(M) >$ "spectrum edge"
  $\Rightarrow$ **optimal choice** α_{opt} **of** α from study of noise spectrum.
- eigenvectors correlated to $D_q^{1-\alpha} J$
  $\Rightarrow$ **Natural regularization by** $D^{\alpha-1} J$**!**

Eigenvalue Spectrum

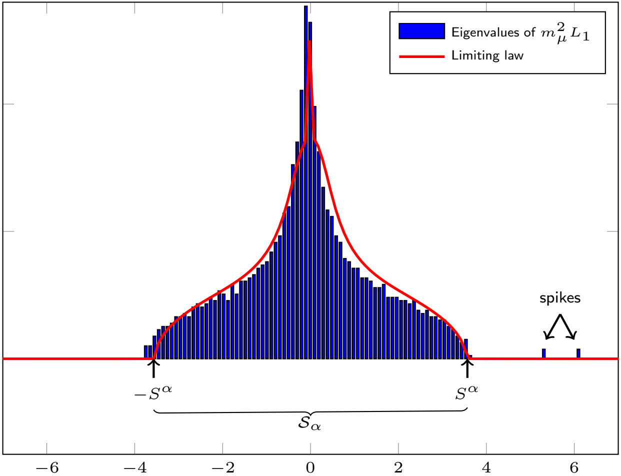


Figure: Eigenvalues of $m_\mu^2 L_1$, $K = 3$, $n = 2000$, $c_1 = 0.3, c_2 = 0.3, c_3 = 0.4$, $\mu = \frac{1}{2}\delta_{q_1} + \frac{1}{2}\delta_{q_2}$, $q_1 = 0.4$, $q_2 = 0.9$, M defined by $M_{ii} = 12$, $M_{ij} = -4, i \neq j$.

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* $m_\mu^{2\alpha} L_\alpha$ *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, $-\frac{1}{e_2^\alpha(\lambda_i(m_\mu^{2\alpha} L_\alpha))} = \lambda_i(\bar{M})$.

Theorem (Phase Transition)

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

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

with $[S_-^\alpha, S_+^\alpha]$ *limiting eigenvalue support of* $m_\mu^{2\alpha} L_\alpha$ *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, $-\frac{1}{e_2^\alpha(\lambda_i(m_\mu^{2\alpha} L_\alpha))} = \lambda_i(\bar{M})$.

Worst-case clustering for $\lambda_i(\bar{M}) = \min_\alpha \tau_\alpha$.

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

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

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 $m_\mu^{2\alpha} L_\alpha$ 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, $-\frac{1}{e_2^\alpha(\lambda_i(m_\mu^{2\alpha} L_\alpha))} = \lambda_i(\bar{M})$.

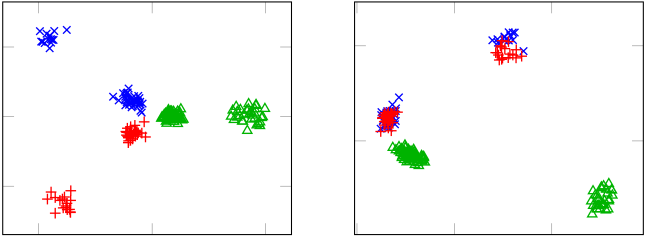
Worst-case clustering for $\lambda_i(\bar{M}) = \min_\alpha \tau_\alpha$.

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

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

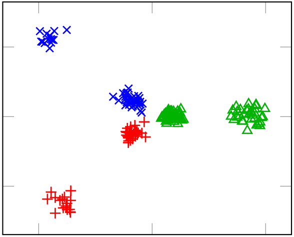
- From $\max_i \left| \frac{d_i}{\sqrt{d^\mathsf{T} 1_n}} - q_i \right| \xrightarrow{\rm a.s.} 0$, we obtain consistent estimator $\hat{\alpha}_{\rm opt}$ of $\alpha_{\rm opt}$.

Simulated Performance Results (2 masses of q_i)

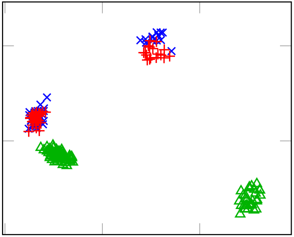


(Modularity) (Bethe Hessian)

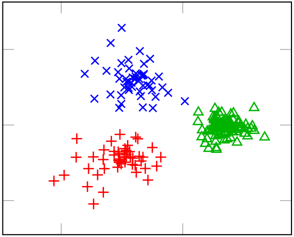
Simulated Performance Results (2 masses of q_i)



(Modularity)



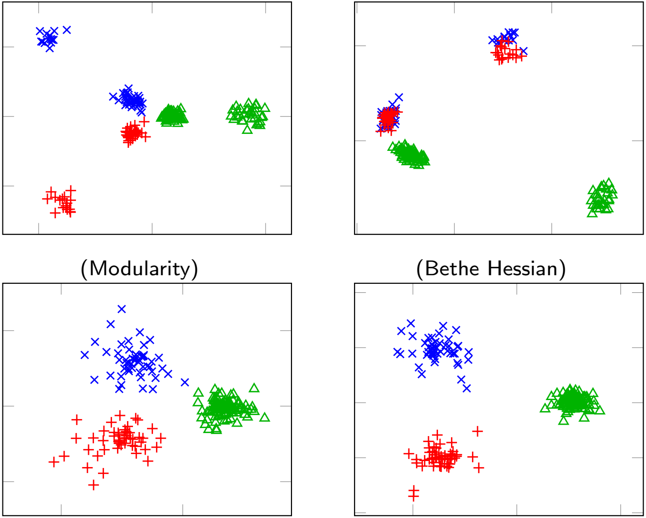
(Bethe Hessian)



(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)

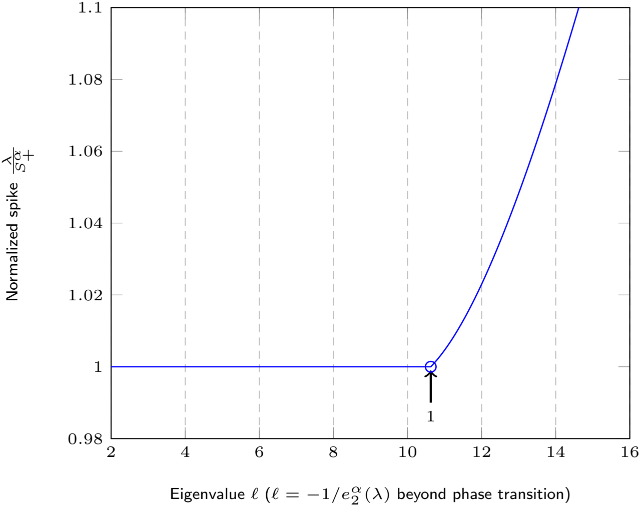


Figure: Largest eigenvalue λ of $m_\mu^2 L_\alpha$ as a function of the largest eigenvalue ℓ of $(\mathcal{D}(c) - cc^\mathsf{T})M$, for $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$ with $q_1 = 0.1$ and $q_2 = 0.5$, for $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \alpha_{\rm opt}\}$ (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)

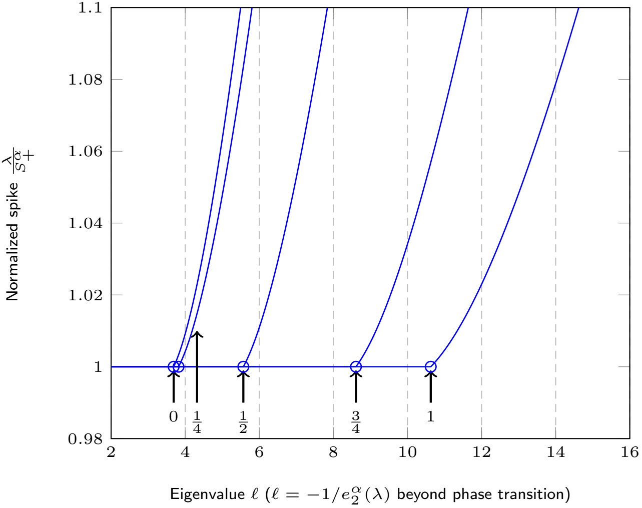


Figure: Largest eigenvalue λ of $m_\mu^2 L_\alpha$ as a function of the largest eigenvalue ℓ of $(\mathcal{D}(c) - cc^\mathsf{T})M$, for $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$ with $q_1 = 0.1$ and $q_2 = 0.5$, for $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \alpha_{\rm opt}\}$ (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)

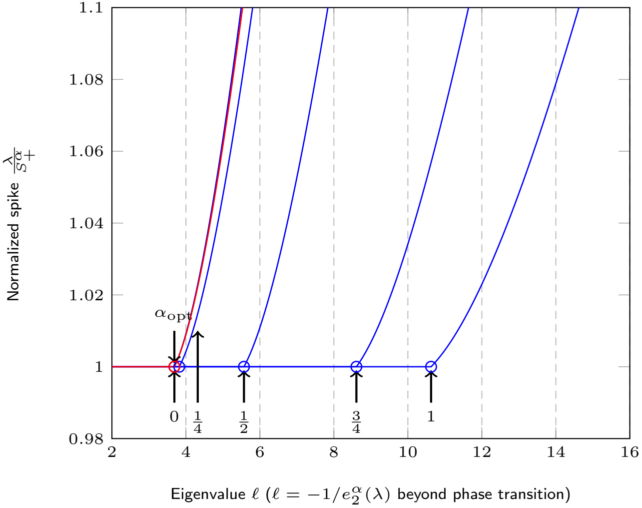


Figure: Largest eigenvalue λ of $m_\mu^2 L_\alpha$ as a function of the largest eigenvalue ℓ of $(\mathcal{D}(c) - cc^{\mathsf{T}})M$, for $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$ with $q_1 = 0.1$ and $q_2 = 0.5$, for $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \alpha_{\mathrm{opt}}\}$ (indicated below the graph). Here, $\alpha_{\mathrm{opt}} = 0.07$. Circles indicate phase transition. Beyond phase transition, $\ell = -1/e_2^\alpha(\lambda)$.

Simulated Performance Results (2 masses for q_i)

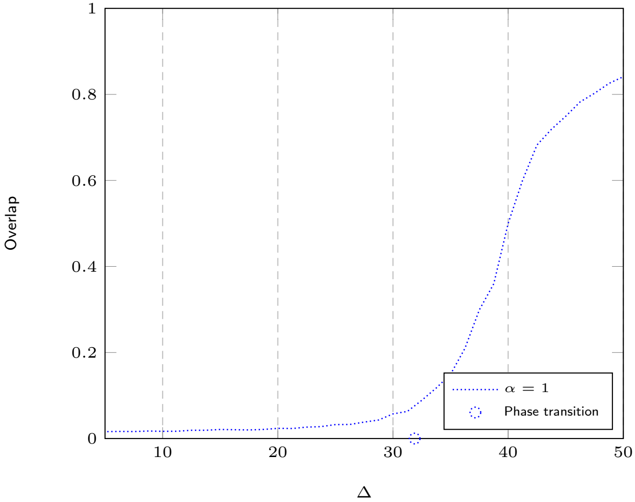


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$ with $q_1 = 0.1$ and $q_2 = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\rm opt} = 0.07$.

Simulated Performance Results (2 masses for q_i)

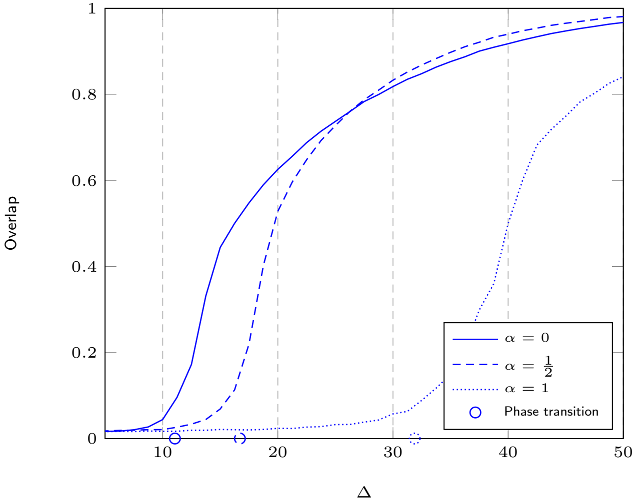


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$ with $q_1 = 0.1$ and $q_2 = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\rm opt} = 0.07$.

Simulated Performance Results (2 masses for q_i)

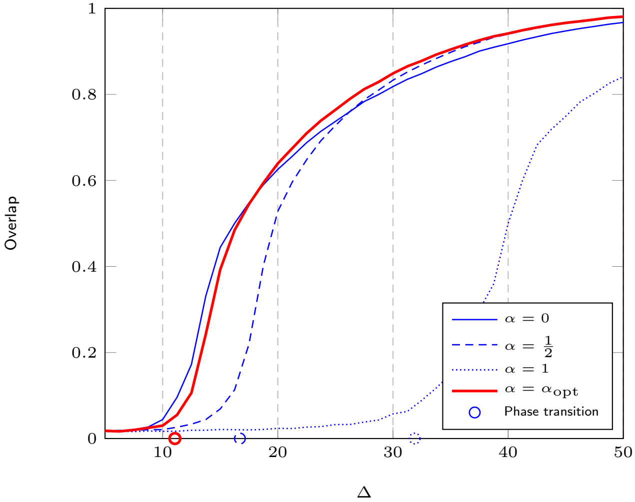


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$ with $q_1 = 0.1$ and $q_2 = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\text{opt}} = 0.07$.

Simulated Performance Results (2 masses for q_i)

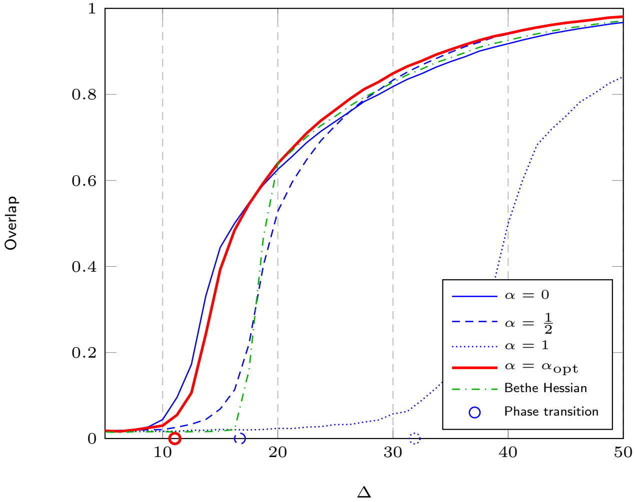


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$ with $q_1 = 0.1$ and $q_2 = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\rm opt} = 0.07$.

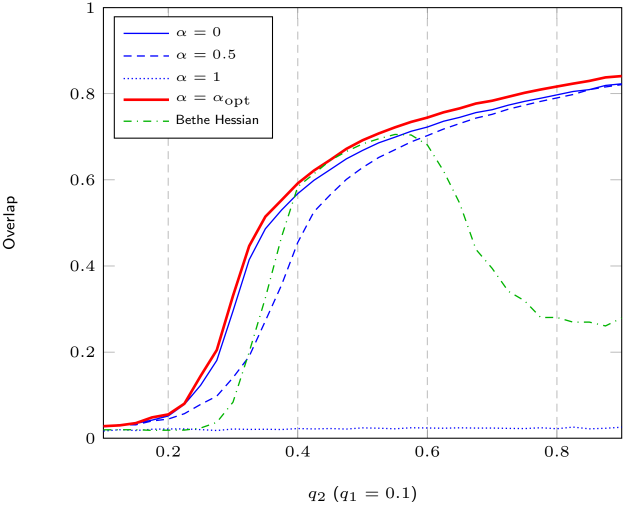


Figure: Overlap performance for $n = 3000$, $K = 3$, $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$ with $q_1 = 0.1$ and $q_2 \in [0.1, 0.9]$, $M = 10(2I_3 - 1_3 1_3^{\mathsf{T}})$, $c_i = \frac{1}{3}$.

Simulated Performance Results (2 masses for q_i)

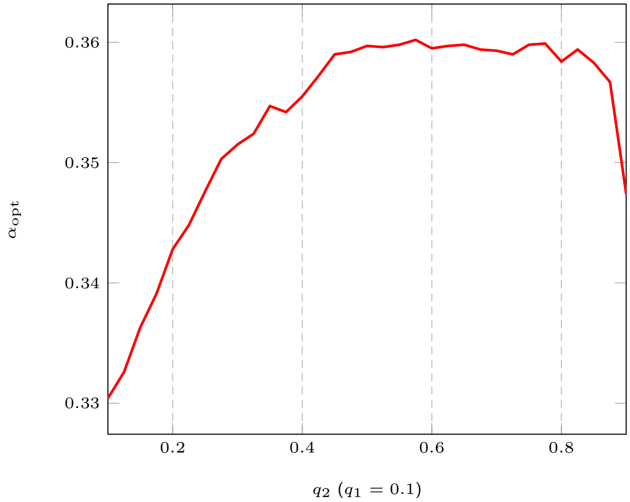


Figure: Evolution of $\alpha_{\rm opt}$ for $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$ with $q_1 = 0.1$, $q_2 \in [0.1, 0.9]$, $M = 10(2I_3 - 1_3 1_3^{\sf T})$, $c_i = \frac{1}{3}$.

Simulated Performance Results ("sparse" power law for q_i)

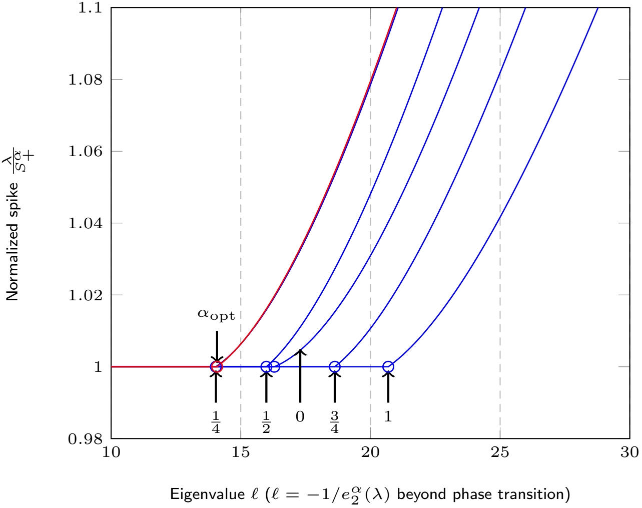


Figure: Largest eigenvalue λ of $m_\mu^2 L_\alpha$ as a function of the largest eigenvalue ℓ of $(\mathcal{D}(c) - cc^{\mathsf{T}})M$, for μ a power law with exponent 3 and support $[0.05, 0.3]$, for $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \alpha_{\rm opt}\}$ (indicated below the graph). Here, $\alpha_{\rm opt} = 0.28$. Circles indicate phase transition. Beyond phase transition, $\ell = -1/e_2^\alpha(\lambda)$.

Simulated Performance Results (“sparse” power law for q_i)

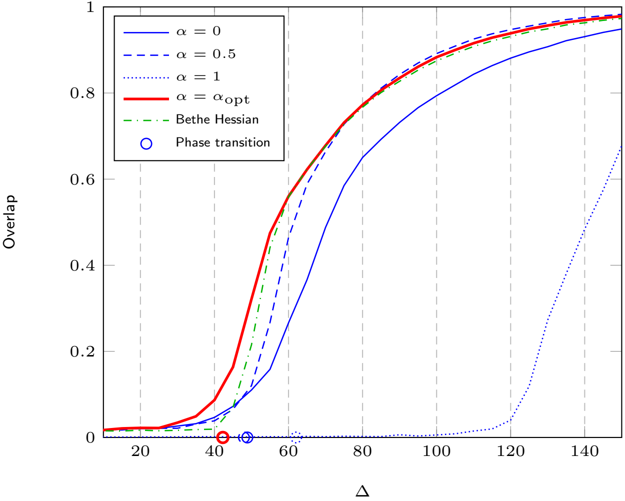


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, μ a power law with exponent 3 and support $[0.05, 0.3]$, $M = \Delta I_3$, for $\Delta \in [10, 150]$. Here $\alpha_{\rm opt} = 0.28$.

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 same class
- 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)

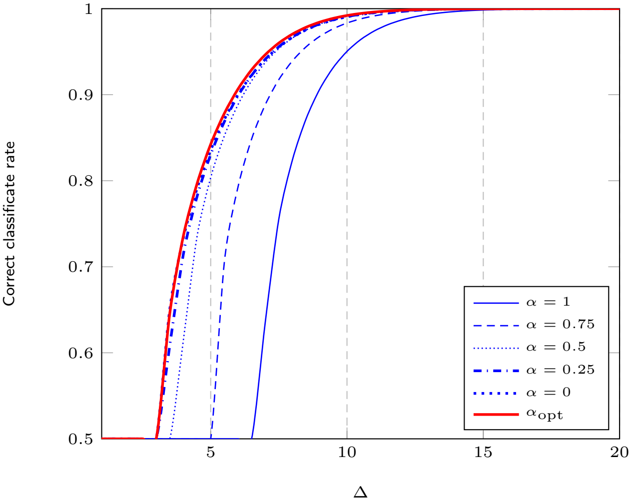


Figure: Theoretical probability of correct recovery for $n = 2000$, $K = 2$, $c_1 = 0.6$, $c_2 = 0.4$, μ uniformly distributed in $[0.2, 0.8]$, $M = \Delta I_2$, for $\Delta \in [0, 20]$.

Theoretical Performance Results (uniform distribution for q_i)

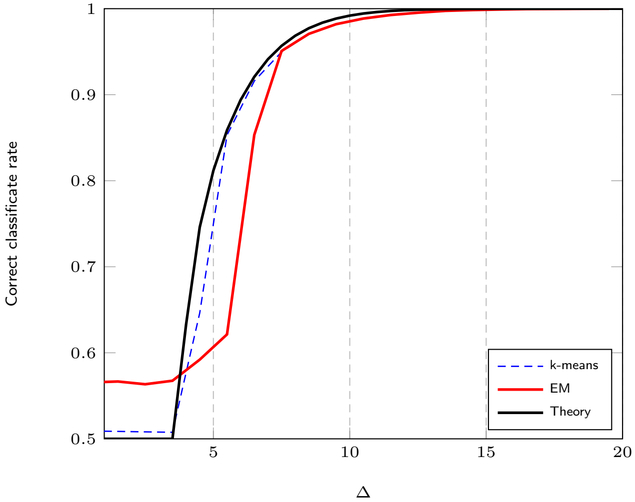


Figure: Probability of correct recovery for $n = 2000$, $K = 2$, $c_1 = 0.6$, $c_2 = 0.4$, μ uniformly distributed in $[0.2, 0.8]$, $M = \Delta I_2$, for $\Delta \in [0, 20]$.

Results on Benchmark Graphs

Graph (n, K)	$\alpha = 0$	$\alpha = \frac{1}{2}$	$\alpha = 1$	$\alpha = \alpha_{\text{opt}}$	(value)	BH
Polbooks $(105, 3)$	*0.743*	**0.757**	0.214	*0.743*	(0)	**0.757**
Adjnoun $(112, 2)$	0.571	**0.714**	0.000	0.571	(0)	0.661
Karate $(34, 2)$	0.176	*0.941*	0.353	0.176	(0)	**1.000**
Dolphins $(62, 2)$	**0.968**	**0.968**	0.387	**0.968**	(0.07)	*0.935*
Polblogs $(1221, 2)$	**0.897**	0.035	0.040	**0.897**	(0)	0.304
Football $(115, 12)$	0.858	*0.905*	*0.905*	*0.905*	(0.16)	**0.924**

Table: Overlap performance on benchmark graphs.

Some Takeaway messages

Main findings:

Some Takeaway messages

Main findings:

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

Some Takeaway messages

Main findings:

- Degree heterogeneity breaks community structures in eigenvectors.
  $\Rightarrow$ Compensation by D^{-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^{-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^{-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^{-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^{-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.

Some Takeaway messages

Main findings:

- Degree heterogeneity breaks community structures in eigenvectors.
  $\Rightarrow$ Compensation by D^{-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.
- Simulations on small networks in fact give ridiculous arbitrary results.
- When is sparse sparse and dense dense?
  - in theory, $d_i = O(\log(n))$ is dense...
  - in practice, assuming dense regime, eigenvalues smear beyond support edges in critical scenarios.

Outline

Problem Statement

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

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

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

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.

Towards kernel spectral clustering

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

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

  i.e., eigenvector problem:

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

Towards kernel spectral clustering

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

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

  i.e., eigenvector problem:
  1. find eigenvectors of smallest eigenvalues
  2. retrieve classes from eigenvector components

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

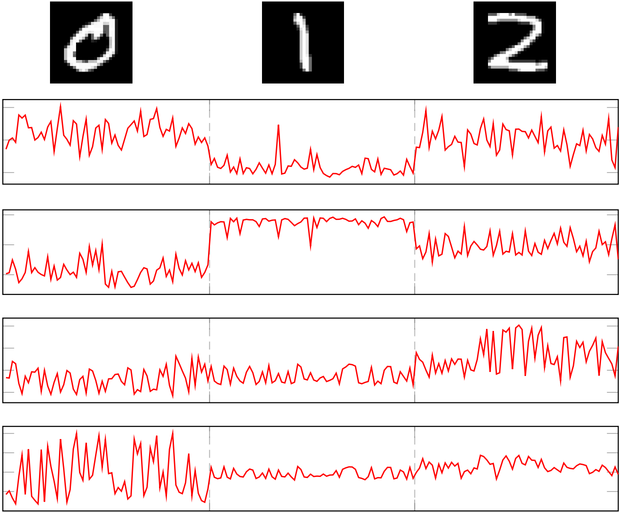


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ for MNIST data.

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)

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*
$$\|C_a\| = O(1), \quad \frac{1}{\sqrt{p}} \operatorname{tr} C_a^\circ = O(1).$$

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.

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 $D = \mathrm{diag}(K1_n)$.

Difficulty: L is a very intractable random matrix

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

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

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

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.

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:

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

- difficult to handle (3 orders to manipulate!)

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:

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

- difficult to handle (3 orders to manipulate!)

Observation: Spectrum of L:

- 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:

- 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:

$$L' = nD^{-\frac{1}{2}}KD^{-\frac{1}{2}} - n\frac{D^{\frac{1}{2}}1_n1_n^\mathsf{T}D^{\frac{1}{2}}}{1_n^\mathsf{T}D1_n}.$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Important Notations:

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

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

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

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

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

Important Notations:

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

Some consequences:

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

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

Isolated eigenvalues: Gaussian inputs

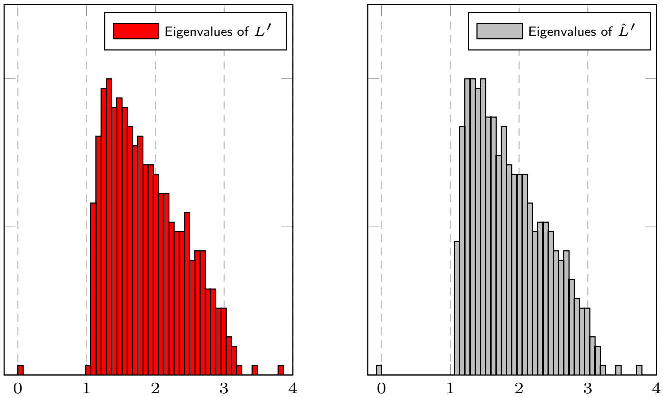


Figure: Eigenvalues of L' and $\hat{L}'$, $k = 3$, $p = 2048$, $n = 512$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $[\mu_a]_j = 4\boldsymbol{\delta}_{aj}$, $C_a = (1 + 2(a-1)/\sqrt{p})I_p$, $f(x) = \exp(-x/2)$.

Two-step Strategy:

1. Study limiting eigenvalue distribution (and its support $\mathcal{S}$) of $\frac{1}{p}PW^{\mathsf{T}}WP$
2. Solve, for $\lambda \notin \mathcal{S}$,

$$\det\left(\frac{1}{p}PW^{\mathsf{T}}WP + UBU^{\mathsf{T}} - \lambda I_n\right) = 0.$$

Equivalent to solving smaller dimensional:

$$\det\left(BU^{\mathsf{T}}Q_\lambda U\right) = 0$$

with $Q_\lambda = (\frac{1}{p}PW^{\mathsf{T}}WP - \lambda I_n)^{-1}$.

Lemma (Deterministic Equivalent)

For $z \in \mathbb{C}$ away from eigenvalues of $\frac{1}{p} P W^\mathsf{T} W P$ and

$$Q_z = \left(\frac{1}{p} P W^\mathsf{T} W P - z I_n \right)^{-1}, \quad \tilde{Q}_z = \left(\frac{1}{p} W P W^\mathsf{T} - z I_p \right)^{-1}.$$

Then, as $n \to \infty$,

$$Q_z \leftrightarrow \bar{Q}_z \triangleq c_0 \operatorname{diag} \{ g_a(z) 1_{n_a} \}_{a=1}^k - \left\{ \left(\frac{1}{z} + c_0 \frac{g_a(z) g_b(z)}{\sum_{i=1}^k c_i g_i(z)} \right) \frac{1_{n_a} 1_{n_b}^\mathsf{T}}{n} \right\}_{a,b=1}^k$$

$$\tilde{Q}_z \leftrightarrow \bar{\tilde{Q}}_z \triangleq \left(-z \left[I_p + \sum_{a=1}^k c_a g_a(z) C_a \right] \right)^{-1}$$

where $(g_1, \ldots, g_k)$ are the unique (Stieltjes transforms) solutions to

$$g_a(z) = \left(-z c_0 \left[1 + \frac{1}{p} \operatorname{tr} C_a \bar{\tilde{Q}}_z \right] \right)^{-1}$$

and $A_n \leftrightarrow B_n$ means $\frac{1}{n} \operatorname{tr} D_n A_n - \frac{1}{n} \operatorname{tr} D_n B_n \xrightarrow{\text{a.s.}} 0$ and $d_{1,n}^\mathsf{T} (A_n - B_n) d_{2,n} \xrightarrow{\text{a.s.}} 0$ for deterministic bounded D_n, $d_{i,n}$.

Theorem ((Useful) isolated eigenvalues)

Define the $k \times k$ matrix

$$G_z = h(\tau, z) I_k + D_{\tau,z}\Gamma_z$$

where

$$D_{\tau,z} = -zh(\tau,z)M^{\mathsf{T}}\bar{\tilde{Q}}_z M - h(\tau,z)\frac{f''(\tau)}{f'(\tau)}T + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^{\mathsf{T}}$$

$$\Gamma_z = \operatorname{diag}\left\{c_a g_a(z)\right\}_{a=1}^{k} - \left\{\frac{c_a g_a(z) c_b g_b(z)}{\sum_{i=1}^{k} c_i g_i(z)}\right\}_{a,b=1}^{k}$$

$$h(\tau,z) = 1 + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)\sum_{a=1}^{k} c_a g_a(z)\frac{2}{p}\operatorname{tr} C_a^2.$$

Theorem ((Useful) isolated eigenvalues)

Define the $k \times k$ matrix

$$G_z = h(\tau, z) I_k + D_{\tau,z}\Gamma_z$$

where

$$D_{\tau,z} = -zh(\tau,z)M^\mathsf{T}\bar{\tilde{Q}}_z M - h(\tau,z)\frac{f''(\tau)}{f'(\tau)}T + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^\mathsf{T}$$

$$\Gamma_z = \operatorname{diag}\left\{c_a g_a(z)\right\}_{a=1}^k - \left\{\frac{c_a g_a(z) c_b g_b(z)}{\sum_{i=1}^k c_i g_i(z)}\right\}_{a,b=1}^k$$

$$h(\tau,z) = 1 + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)\sum_{a=1}^k c_a g_a(z)\frac{2}{p}\operatorname{tr} C_a^2.$$

If $\rho \notin \mathcal{S}$ is such that $h(\tau,\rho) \neq 0$ and G_ρ has a zero eigenvalue of multiplicity m_ρ, then

$-2\frac{f(\tau)}{f'(\tau)}(L - \alpha(\tau)I_n)$ *has m_ρ isolated eigenvalues converging to ρ.*

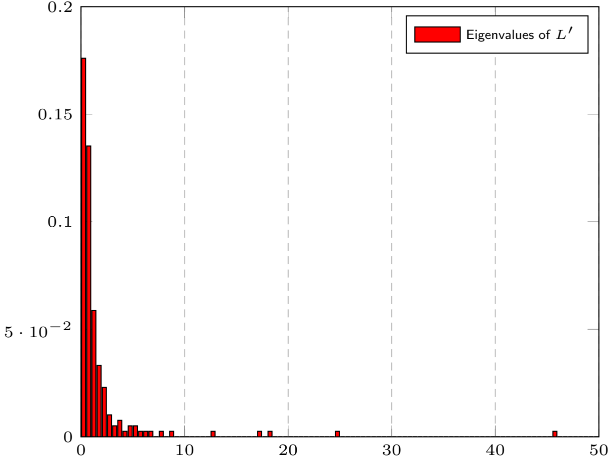


Figure: Eigenvalues of L' (red) and (equivalent Gaussian model) $\hat{L}'$ (white), MNIST data, $p = 784$, $n = 192$.

Isolated eigenvalues: MNIST

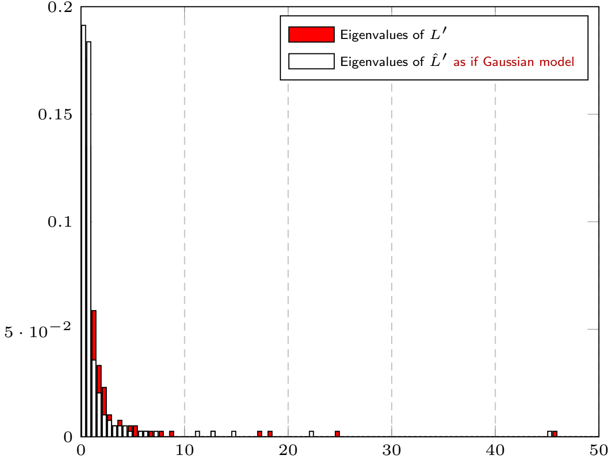


Figure: Eigenvalues of L' (red) and (equivalent Gaussian model) $\hat{L}'$ (white), MNIST data, $p = 784$, $n = 192$.

Eigenvectors

Strategy:

- Study "easy" eigenvector $D^{\frac{1}{2}} 1_n$

Eigenvectors

Strategy:

- Study "easy" eigenvector $D^{\frac{1}{2}}1_n$
- Independently, for each spike eigenvalue, study eigenvector projections on basis J

Strategy:

- Study "easy" eigenvector $D^{\frac{1}{2}}1_n$
- Independently, for each spike eigenvalue, study eigenvector projections on basis J

Dominant Eigenvector:

Proposition (Eigenvector $D^{\frac{1}{2}}1_n$)

We have

$$\frac{D^{\frac{1}{2}}1_n}{\sqrt{1_n^\mathsf{T} D 1_n}} = \frac{1_n}{\sqrt{n}} + \frac{1}{n\sqrt{c_0}} \frac{f'(\tau)}{2f(\tau)} \left[\{t_a 1_{n_a}\}_{a=1}^k + \operatorname{diag}\left\{ \sqrt{\frac{2}{p}\operatorname{tr}(C_a^2)} 1_{n_a} \right\}_{a=1}^k \varphi \right] + o(n^{-1})$$

with $\varphi \sim \mathcal{N}(0, I_n)$.

Strategy:

- Study "easy" eigenvector $D^{\frac{1}{2}} 1_n$
- Independently, for each spike eigenvalue, study eigenvector projections on basis J

Dominant Eigenvector:

Proposition (Eigenvector $D^{\frac{1}{2}} 1_n$)

We have

$$\frac{D^{\frac{1}{2}} 1_n}{\sqrt{1_n^\mathsf{T} D 1_n}} = \frac{1_n}{\sqrt{n}} + \frac{1}{n\sqrt{c_0}} \frac{f'(\tau)}{2f(\tau)} \left[\{t_a 1_{n_a}\}_{a=1}^k + \operatorname{diag}\left\{ \sqrt{\frac{2}{p} \operatorname{tr}(C_a^2)} 1_{n_a} \right\}_{a=1}^k \varphi \right] + o(n^{-1})$$

with $\varphi \sim \mathcal{N}(0, I_n)$.

Remark:

- $D^{\frac{1}{2}} 1_n$ block-wise constant + noise
- only information about $\operatorname{tr} C_a^\circ$!

Isolated eigenvectors

Theorem (Eigenvector projections)

Let ρ isolated eigenvalue and Π_ρ its associated subspace in L, then

$$\frac{1}{p} J^\mathsf{T} \hat{\Pi}_\rho J = -h(\tau, \rho) \Gamma_\rho \Xi_\rho + o(1)$$

where $J = [j_1, \ldots, j_k]$ canonical class-basis, and

$$\Xi_\rho = \sum_{i=1}^{m_\rho} \frac{(V_{r,\rho})_i (V_{l,\rho})_i^\mathsf{T}}{(V_{l,\rho})_i^\mathsf{T} G'_\rho (V_{r,\rho})_i}$$

with $V_{r,\rho}, V_{l,\rho} \in \mathbb{C}^{k \times m_\rho}$ right and left eigenvectors of G_ρ associated with eig. zero.

Isolated eigenvectors

Theorem (Eigenvector projections)

Let ρ isolated eigenvalue and Π_ρ its associated subspace in L, then

$$\frac{1}{p}J^\mathsf{T}\hat{\Pi}_\rho J = -h(\tau,\rho)\Gamma_\rho\Xi_\rho + o(1)$$

where $J = [j_1,\ldots,j_k]$ canonical class-basis, and

$$\Xi_\rho = \sum_{i=1}^{m_\rho} \frac{(V_{r,\rho})_i (V_{l,\rho})_i^\mathsf{T}}{(V_{l,\rho})_i^\mathsf{T} G'_\rho (V_{r,\rho})_i}$$

with $V_{r,\rho}, V_{l,\rho} \in \mathbb{C}^{k\times m_\rho}$ right and left eigenvectors of G_ρ associated with eig. zero.

Remark: $m_\rho = 1$ case

- $[J^\mathsf{T} uu^\mathsf{T} J]_{aa} = |j_a^\mathsf{T} u|^2$: eigenvector "level" in class $\mathcal{C}_a$
- $E = 1 - \frac{1}{n}\operatorname{tr}\left(\operatorname{diag}(\{1/c_i\}) J^\mathsf{T} uu^\mathsf{T} J\right)$: total noise energy
- Eigenvector levels given by eigenvectors of $G_\rho = h(\tau,\rho)I_k + D_{\tau,\rho}\Gamma_\rho$.

Eigenvectors

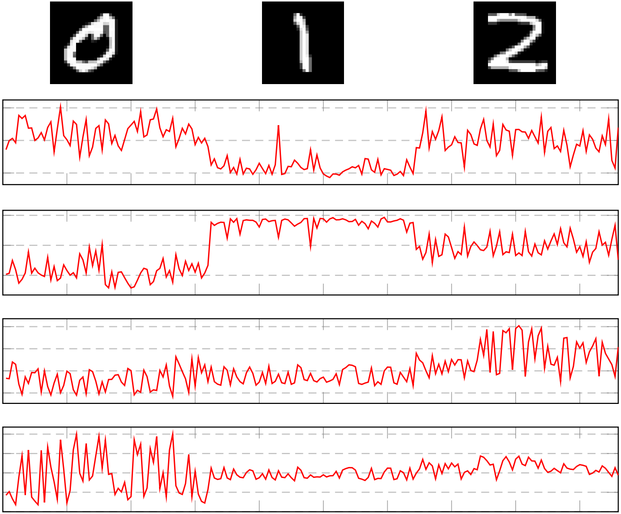


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data (**red**), versus Gaussian equivalent model (**black**), and theoretical findings (**blue**).

Eigenvectors

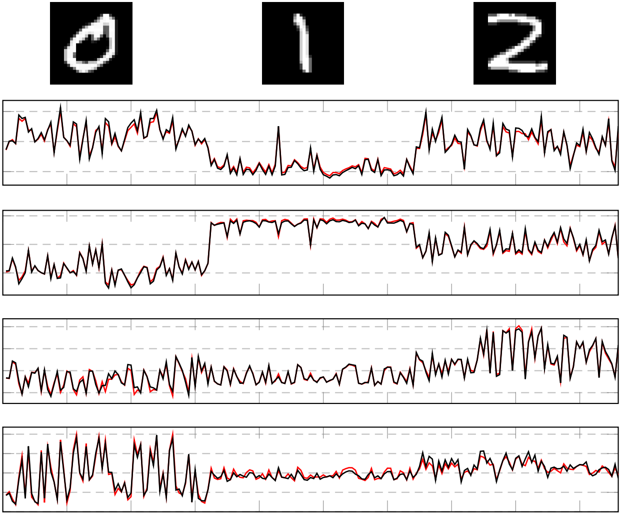


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data (**red**), versus Gaussian equivalent model (**black**), and theoretical findings (**blue**).

Eigenvectors

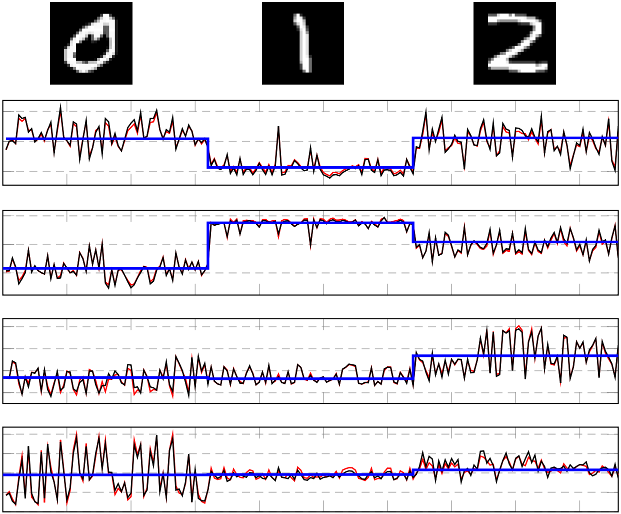


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data (**red**), versus Gaussian equivalent model (**black**), and theoretical findings (**blue**).

Case $C_1 = \ldots = C_k = I_k$

Corollary: let (ℓ, Υ) isolated eigenpair of $I_p + M \operatorname{diag}(\{c_i\}) M^\mathsf{T}$,

Case $C_1 = \ldots = C_k = I_k$

Corollary: let (ℓ, Υ) isolated eigenpair of $I_p + M\,\mathrm{diag}(\{c_i\})M^\mathsf{T}$,

- **Condition for Existence**: $|\ell - 1| > \sqrt{c_0}$ (classical spike random matrix result)

Case $C_1 = \ldots = C_k = I_k$

Corollary: let (ℓ, Υ) isolated eigenpair of $I_p + M \operatorname{diag}(\{c_i\}) M^\mathsf{T}$,

- **Condition for Existence**: $|\ell - 1| > \sqrt{c_0}$ (classical spike random matrix result)
- **Eigenvalues**: isolated eigenvalue ρ of $-\frac{f(\tau)}{2f'(\tau)}(L - \alpha(\tau) I_n)$

$$\rho = \frac{\ell}{c_0} + \frac{\ell}{\ell - 1}$$

Corollary: let (ℓ, Υ) isolated eigenpair of $I_p + M \operatorname{diag}(\{c_i\}) M^\mathsf{T}$,

- **Condition for Existence**: $|\ell - 1| > \sqrt{c_0}$ (classical spike random matrix result)
- **Eigenvalues**: isolated eigenvalue ρ of $-\frac{f(\tau)}{2f'(\tau)}(L - \alpha(\tau) I_n)$

$$\rho = \frac{\ell}{c_0} + \frac{\ell}{\ell - 1}$$

- **Eigenvectors**:

$$\frac{1}{n} J^\mathsf{T} \Pi_\rho J = \left(\frac{1}{\ell} - \frac{c_0}{\ell(\ell-1)^2} \right) \operatorname{diag}(\{c_i\}) M^\mathsf{T} \Upsilon_\rho \Upsilon_\rho^\mathsf{T} M \operatorname{diag}(\{c_i\}) + o(1).$$

Case $C_1 = \ldots = C_k = I_k$

Corollary: let (ℓ, Υ) isolated eigenpair of $I_p + M \operatorname{diag}(\{c_i\}) M^\mathsf{T}$,

- **Condition for Existence**: $|\ell - 1| > \sqrt{c_0}$ (classical spike random matrix result)
- **Eigenvalues**: isolated eigenvalue ρ of $-\frac{f(\tau)}{2f'(\tau)}(L - \alpha(\tau) I_n)$

$$\rho = \frac{\ell}{c_0} + \frac{\ell}{\ell - 1}$$

- **Eigenvectors**:

$$\frac{1}{n} J^\mathsf{T} \Pi_\rho J = \left(\frac{1}{\ell} - \frac{c_0}{\ell(\ell-1)^2} \right) \operatorname{diag}(\{c_i\}) M^\mathsf{T} \Upsilon_\rho \Upsilon_\rho^\mathsf{T} M \operatorname{diag}(\{c_i\}) + o(1).$$

Remark: **Does not depend on f!**

Case $M = 0$, $C_a = (1 + \gamma_a/\sqrt{p})I_p$

Corollary: let $\gamma = [\gamma_1, \ldots, \gamma_k]^\mathsf{T}$ and

$$\ell = \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)\left(2 + \sum_{a=1}^{k} c_a \gamma_a^2\right).$$

Then,

Case $M = 0$, $C_a = (1 + \gamma_a/\sqrt{p})I_p$

Corollary: let $\gamma = [\gamma_1, \ldots, \gamma_k]^\mathsf{T}$ and

$$\ell = \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) \left(2 + \sum_{a=1}^{k} c_a \gamma_a^2 \right).$$

Then,

- **Condition for Existence**: $|\ell - 1| > \sqrt{c_0}$ (classical spike random matrix result)

Corollary: let $\gamma=[\gamma_1,\ldots,\gamma_k]^\mathsf{T}$ and

$$\ell=\left(\frac{5f'(\tau)}{8f(\tau)}-\frac{f''(\tau)}{2f'(\tau)}\right)\left(2+\sum_{a=1}^{k}c_a\gamma_a^2\right).$$

Then,

- **Condition for Existence**: $|\ell-1|>\sqrt{c_0}$ (classical spike random matrix result)
- **Eigenvalues**: isolated eigenvalue ρ of $-\frac{f(\tau)}{2f'(\tau)}(L-\alpha(\tau)I_n)$

$$\rho=\frac{\ell}{c_0}+\frac{\ell}{\ell-1}$$

Case $M = 0$, $C_a = (1 + \gamma_a/\sqrt{p})I_p$

Corollary: let $\gamma = [\gamma_1, \ldots, \gamma_k]^\mathsf{T}$ and

$$\ell = \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) \left(2 + \sum_{a=1}^{k} c_a \gamma_a^2 \right).$$

Then,

- **Condition for Existence**: $|\ell - 1| > \sqrt{c_0}$ (classical spike random matrix result)
- **Eigenvalues**: isolated eigenvalue ρ of $-\frac{f(\tau)}{2f'(\tau)}(L - \alpha(\tau)I_n)$

$$\rho = \frac{\ell}{c_0} + \frac{\ell}{\ell - 1}$$

- **Eigenvectors**:

$$\frac{1}{n} J^\mathsf{T} \Pi_\rho J = \frac{1 - \frac{c_0}{(\ell-1)^2}}{2 + \sum_{a=1}^{k} c_a \gamma_a^2} \operatorname{diag}(\{c_i\}) \gamma \gamma^\mathsf{T} \operatorname{diag}(\{c_i\}) + o(1).$$

Case $M = 0$, $C_a = (1 + \gamma_a/\sqrt{p})I_p$

Corollary: let $\gamma = [\gamma_1, \ldots, \gamma_k]^\mathsf{T}$ and

$$\ell = \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)\left(2 + \sum_{a=1}^{k} c_a \gamma_a^2\right).$$

Then,

- **Condition for Existence**: $|\ell - 1| > \sqrt{c_0}$ (classical spike random matrix result)
- **Eigenvalues**: isolated eigenvalue ρ of $-\frac{f(\tau)}{2f'(\tau)}(L - \alpha(\tau)I_n)$

$$\rho = \frac{\ell}{c_0} + \frac{\ell}{\ell - 1}$$

- **Eigenvectors**:

$$\frac{1}{n}J^\mathsf{T}\Pi_\rho J = \frac{1 - \frac{c_0}{(\ell-1)^2}}{2 + \sum_{a=1}^{k} c_a \gamma_a^2} \operatorname{diag}(\{c_i\})\gamma\gamma^\mathsf{T}\operatorname{diag}(\{c_i\}) + o(1).$$

Remark:

- only ONE isolated eigenvalue
- eigenvector alignment directly linked to γ_a's.

Further Results

Beyond Class-wise means:

- per-class fluctuations
- per-class cross-eigenvector fluctuations

Further Results

Beyond Class-wise means:

- per-class fluctuations
- per-class cross-eigenvector fluctuations

Consequences:

- see M isolated eigenvectors as n points in $\mathbb{R}^M$
- clustering $x_1, \ldots, x_n \Leftrightarrow$ clustering n points in $\mathbb{R}^M$

Further Results

Beyond Class-wise means:

- per-class fluctuations
- per-class cross-eigenvector fluctuations

Consequences:

- see M isolated eigenvectors as n points in $\mathbb{R}^M$
- clustering $x_1, \ldots, x_n \Leftrightarrow$ clustering n points in $\mathbb{R}^M$

Method:

- per-class fluctuations: for each a, estimate

$$\operatorname{tr}\left(\operatorname{diag}(j_a)\hat{\Pi}_\rho\right)$$

$\Rightarrow$ for $\hat{\Pi}_\rho = u_\rho u_\rho^*$, gives access to $\operatorname{tr}\left(\operatorname{diag}(j_a)u_\rho u_\rho^*\right) = u_\rho^* \operatorname{diag}(j_a)u_\rho$

Beyond Class-wise means:

- per-class fluctuations
- per-class cross-eigenvector fluctuations

Consequences:

- see M isolated eigenvectors as n points in $\mathbb{R}^M$
- clustering $x_1, \ldots, x_n$ $\Leftrightarrow$ clustering n points in $\mathbb{R}^M$

Method:

- per-class fluctuations: for each a, estimate

$$\operatorname{tr}\left(\operatorname{diag}(j_a)\hat{\Pi}_\rho\right)$$

$\Rightarrow$ for $\hat{\Pi}_\rho = u_\rho u_\rho^*$, gives access to $\operatorname{tr}\left(\operatorname{diag}(j_a) u_\rho u_\rho^*\right) = u_\rho^* \operatorname{diag}(j_a) u_\rho$

- cross-eigenvector fluctuations: for each a and (ρ_1, ρ_2), estimate

$$\frac{1}{p} J^\mathsf{T} \hat{\Pi}_{\rho_1} \operatorname{diag}(j_a) \hat{\Pi}_{\rho_2} J$$

$\Rightarrow$ for $\hat{\Pi}_\rho = u_\rho u_\rho^*$, gives access to $(u_{\rho_1}^* \operatorname{diag}(j_a) u_{\rho_2}) \times (\frac{1}{\sqrt{p}} J^\mathsf{T} u_{\rho_1})(\frac{1}{\sqrt{p}} u_{\rho_2}^* J)$

Theoretical Findings versus MNIST

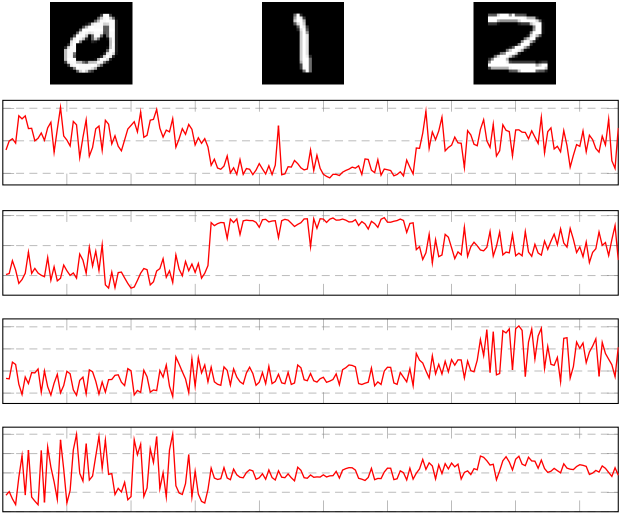


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data (**red**), versus Gaussian equivalent model (**black**), and theoretical findings (**blue**).

Theoretical Findings versus MNIST

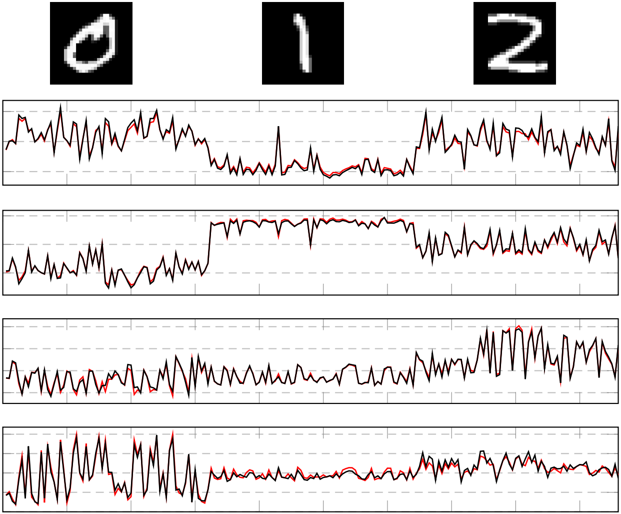


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data (**red**), versus Gaussian equivalent model (**black**), and theoretical findings (**blue**).

Theoretical Findings versus MNIST

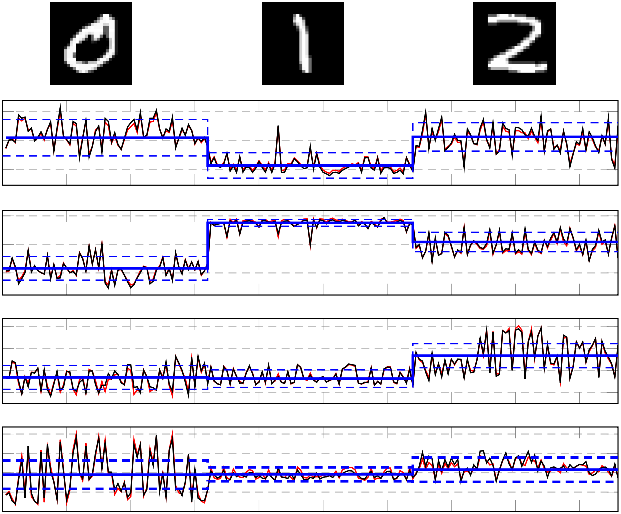


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ for MNIST data (**red**), versus Gaussian equivalent model (**black**), and theoretical findings (**blue**).

Theoretical Findings versus MNIST

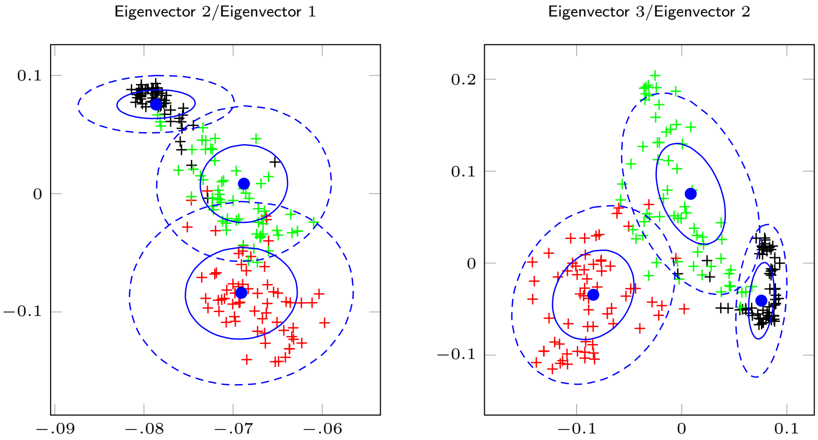


Figure: 2D representation of eigenvectors of L, for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in **blue**. Class 1 in **red**, Class 2 in **black**, Class 3 in **green**.

Some Takeaway messages

Surprising findings:

Some Takeaway messages

Surprising findings:

- "Good kernel functions" f need not be decreasing.

Some Takeaway messages

Surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.

Some Takeaway messages

Surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.
- More importantly, clustering possible despite $\|x_i - x_j\|^2 \to \tau$, i.e., no first order data difference
  $\Rightarrow$ **Breaks original intuitions and problem layout!**

Some Takeaway messages

Surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.
- More importantly, clustering possible despite $\|x_i - x_j\|^2 \to \tau$, i.e., no first order data difference
  $\Rightarrow$ **Breaks original intuitions and problem layout!**

Validity of the Results:

Surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.
- More importantly, clustering possible despite $\|x_i - x_j\|^2 \to \tau$, i.e., no first order data difference
  $\Rightarrow$ **Breaks original intuitions and problem layout!**

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).

Some Takeaway messages

Surprising findings:

- "Good kernel functions" f need not be decreasing.
- Dominant parameters in large dimensions are first three derivatives at τ.
- More importantly, clustering possible despite $\|x_i - x_j\|^2 \to \tau$, i.e., no first order data difference
  $\Rightarrow$ **Breaks original intuitions and problem layout!**

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 large images essentially Gaussian vectors?
  - kernels extract primarily first order properties (means, covariances)
  - with no fancy image processing (rotations, scale invariance), may be strong enough features.

Outline

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: $(d_i = [K 1_n]_i)$

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^{k} \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2$$

such that $F_{ia} = \delta_{\{x_i \in \mathcal{C}_a\}}$, for all labelled x_i.

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: $(d_i = [K 1_n]_i)$

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^{k} \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2$$

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

$$K = \begin{bmatrix} K_{(l,l)} & K_{(l,u)} \\ K_{(u,l)} & K_{(u,u)} \end{bmatrix}$$

$$D = \begin{bmatrix} D_{(l)} & 0 \\ 0 & D^{(u)} \end{bmatrix} = \operatorname{diag}\left\{ K 1_n \right\}.$$

Using $F^{(u)}$:

- From $F^{(u)}$, classification algorithm:

$$\text{Classify } x_i \text{ in } \mathcal{C}_a \quad \Leftrightarrow \quad F_{ia} = \max_{b \in \{1,\ldots,k\}} \{F_{ib}\}.$$

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?

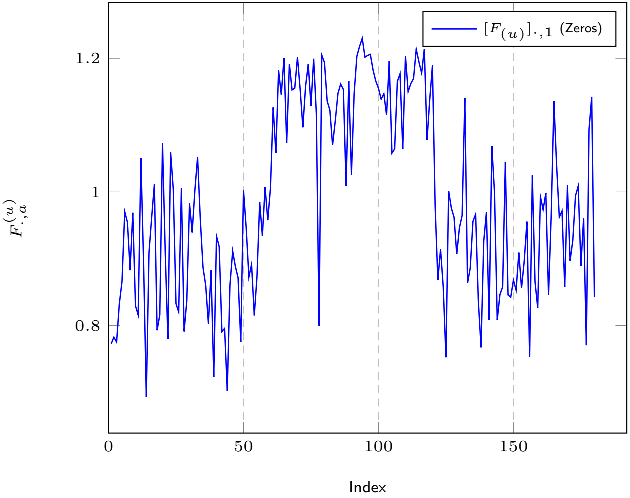


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

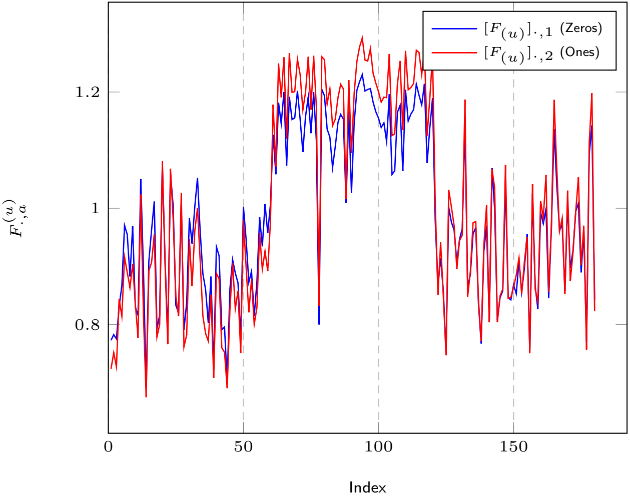


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

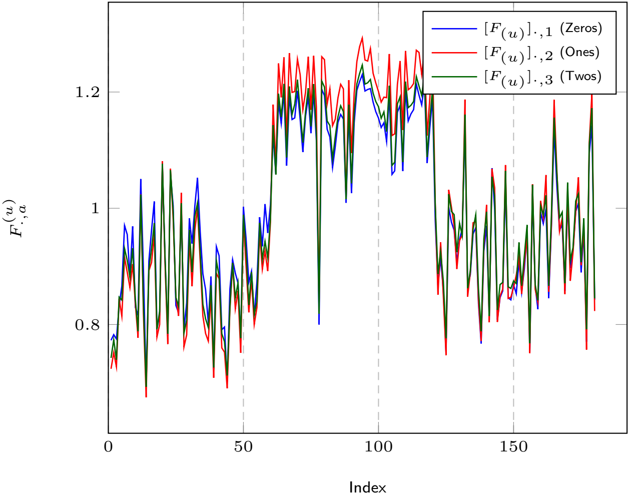


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

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

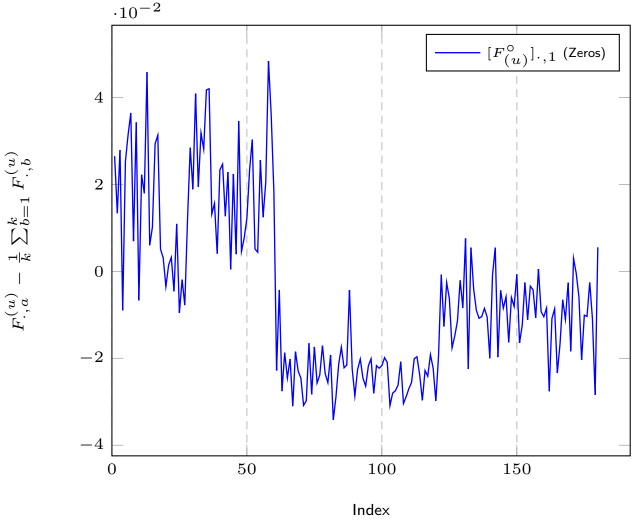


Figure: Centered Vectors $[F^\circ_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k} F_{(u)} 1_k 1_k^\mathsf{T}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

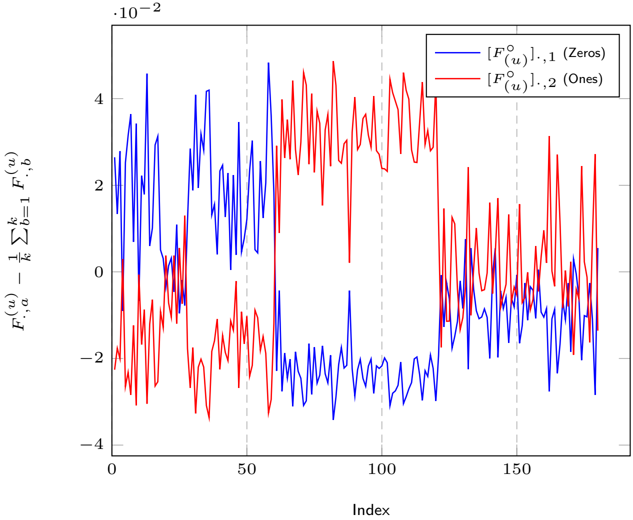


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k1_k^{\mathsf{T}}]_{\cdot,a}$, $a = 1,2,3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

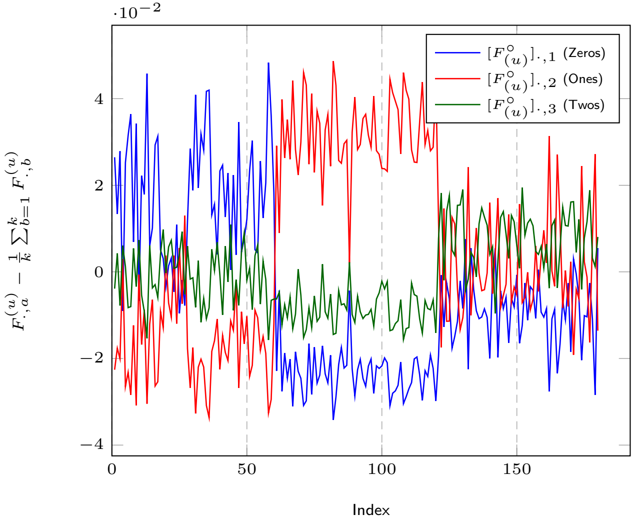


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k 1_k^{\mathsf{T}}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

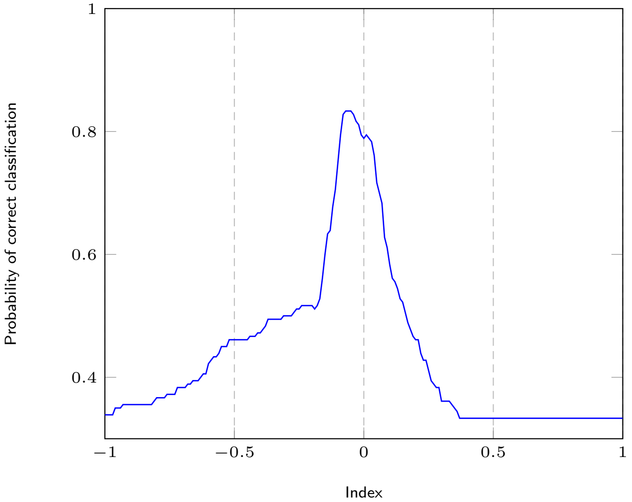


Figure: Performance as a function of α, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

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!

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:

Main Results (so far)

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

Main Results (so far)

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** $F^{(l)}$ need be scaled!

Main Results (so far)

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 $t_a = \frac{1}{\sqrt{p}}\operatorname{tr} C^\circ_a$.

- Many consequences:
  - Random non-informative bias linked to v
  - Strong Impact of $n_{l,a}$!

    $\Rightarrow$ All $n_{l,a}$ must be equal **OR** $F^{(l)}$ need be scaled!
  - Additional per-class bias $\alpha t_a 1_{n_u}$: no information here

    $\Rightarrow$ **Forces the choice**

$$\alpha = 0 + \frac{\beta}{\sqrt{n}}.$$

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** $F^{(l)}$ need be scaled!
  - Additional per-class bias $\alpha t_a 1_{n_u}$: no information here
    $\Rightarrow$ **Forces the choice**

$$\alpha = 0 + \frac{\beta}{\sqrt{n}}.$$

  - Relevant information hidden in smaller order terms!

MNIST Data Example

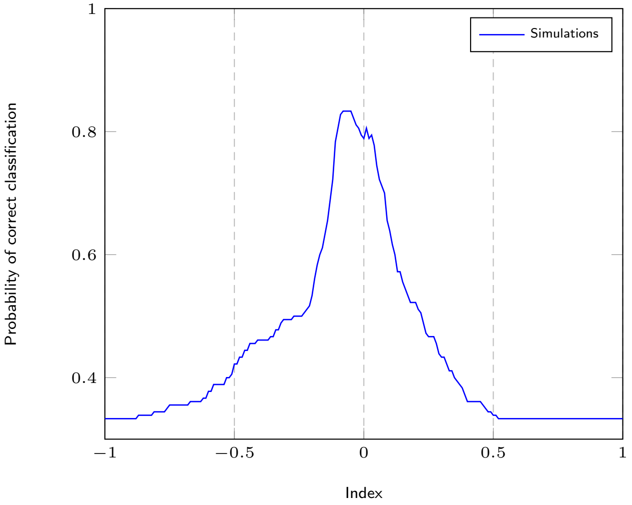


Figure: Performance as a function of α, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

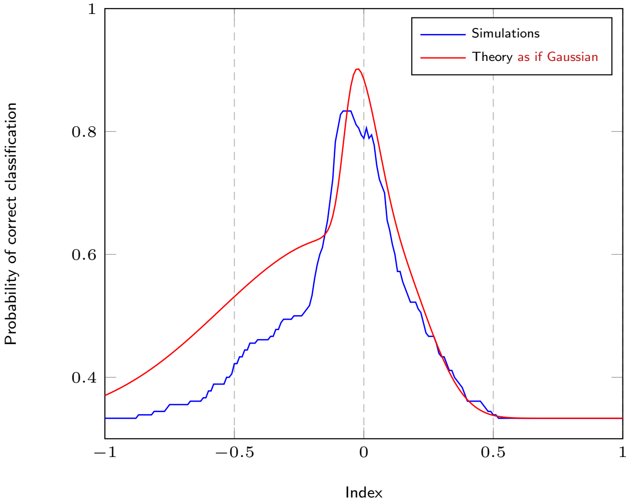


Figure: Performance as a function of α, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

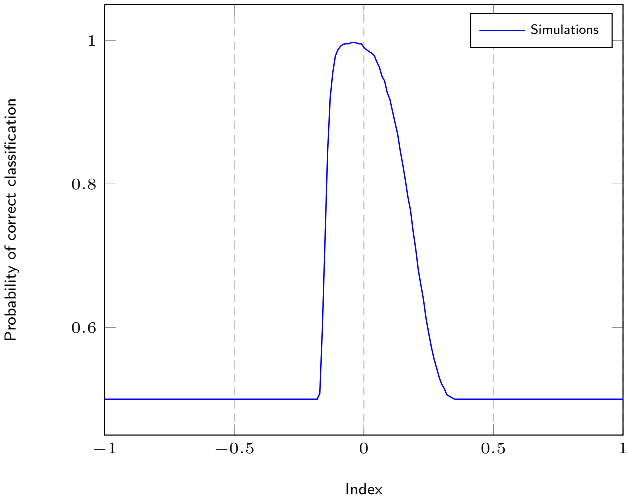


Figure: Performance as a function of α, for 2-class MNIST data (zeros, ones), $n = 1568$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

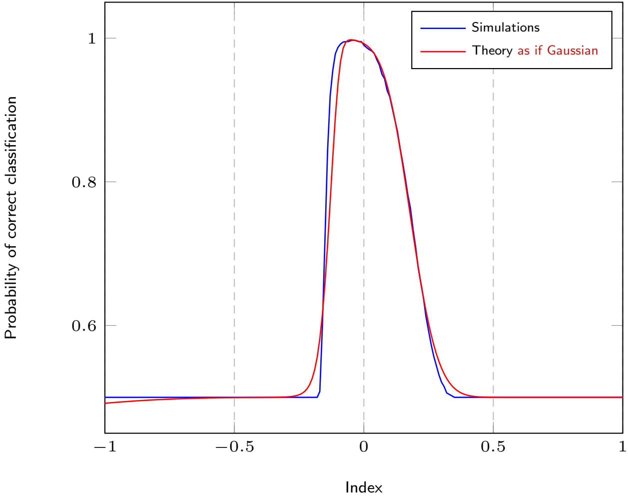


Figure: Performance as a function of α, for 2-class MNIST data (zeros, ones), $n = 1568$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

Outline

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.
$\Rightarrow$ Solved by quadratic programming methods.
$\Rightarrow$ 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**:

$$c(x_i; w, b) = \imath_{\{y_i(w^\mathsf{T}\phi(x_i)+b) \geq 1\}}$$

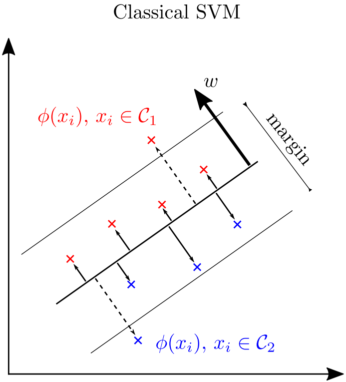
with $y_i = \pm 1$ depending on class.
$\Rightarrow$ Solved by quadratic programming methods.
$\Rightarrow$ Analysis requires joint RMT + convex optimization tools (very interesting but left for later...).

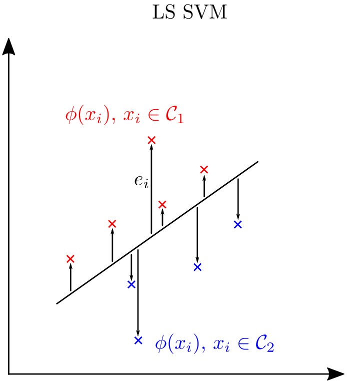
- **LS SVM**:

$$c(x_i; w, b) = \gamma e_i^2 \equiv (y_i - w^\mathsf{T}\phi(x_i) - b)^2.$$

$\Rightarrow$ Explict solution (but not sparse!).

Problem Statement





For new datum x, decision based on (sign of)

$$g(x) = \alpha^{\mathsf{T}} K(\cdot, x) + b$$

where $\alpha \in \mathbb{R}^n$ and b are solution to

$$\begin{bmatrix} 0 & 1_n^{\mathsf{T}} \\ 1_n & K + \frac{n}{\gamma} I_n \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}$$

with $y = [y_i]_{i=1}^n$, γ some parameter to set.

For new datum x, decision based on (sign of)

$$g(x) = \alpha^{\mathsf{T}} K(\cdot, x) + b$$

where $\alpha \in \mathbb{R}^n$ and b are solution to

$$\begin{bmatrix} 0 & 1_n^{\mathsf{T}} \\ 1_n & K + \frac{n}{\gamma} I_n \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}$$

with $y = [y_i]_{i=1}^n$, γ some parameter to set.

Objectives:

- Study behavior of $g(x)$

For new datum x, decision based on (sign of)

$$g(x) = \alpha^{\mathsf{T}} K(\cdot, x) + b$$

where $\alpha \in \mathbb{R}^n$ and b are solution to

$$\begin{bmatrix} 0 & 1_n^{\mathsf{T}} \\ 1_n & K + \frac{n}{\gamma} I_n \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}$$

with $y = [y_i]_{i=1}^n$, γ some parameter to set.

Objectives:

- Study behavior of $g(x)$
- For $x \in \mathcal{C}_a$, determine probability of success.

For new datum x, decision based on (sign of)

$$g(x) = \alpha^\mathsf{T} K(\cdot, x) + b$$

where $\alpha \in \mathbb{R}^n$ and b are solution to

$$\begin{bmatrix} 0 & 1_n^\mathsf{T} \\ 1_n & K + \frac{n}{\gamma} I_n \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}$$

with $y = [y_i]_{i=1}^n$, γ 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.

Early Results

As before, $x_i \sim \mathcal{N}(\mu_a, C_a)$, $a = 1, \ldots, k$, with identical growth conditions, here for $k = 2$.

Early 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

$$g(x) = \frac{n_2 - n_1}{n} + \frac{0}{\sqrt{p}} + \underbrace{\frac{G(x)}{p}}_{\text{Relevant terms here!}}$$

Early 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

$$g(x) = \frac{n_2 - n_1}{n} + \frac{0}{\sqrt{p}} + \underbrace{\frac{G(x)}{p}}_{\text{Relevant terms here!}}$$

- $G(x)$ proportional to γ

Early 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

$$g(x) = \frac{n_2 - n_1}{n} + \frac{0}{\sqrt{p}} + \underbrace{\frac{G(x)}{p}}_{\text{Relevant terms here!}}$$

- $G(x)$ proportional to γ
- $G(x)$ asymptotically Gaussian with in particular

$$E[G(x)] \to \begin{cases} -c_1 M & , x \in \mathcal{C}_1 \\ c_2 M & , x \in \mathcal{C}_2 \end{cases}$$

$$M = \frac{2c_1c_2}{\gamma}\left[-2f'(\tau)\|\mu_2 - \mu_1\|^2 + f''(\tau)(t_2 - t_1)^2 + \frac{4f''(\tau)}{p}\operatorname{tr}(C_1 - C_2)^2\right].$$

Early 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

$$g(x) = \frac{n_2 - n_1}{n} + \frac{0}{\sqrt{p}} + \underbrace{\frac{G(x)}{p}}_{\text{Relevant terms here!}}$$

- $G(x)$ proportional to γ
- $G(x)$ asymptotically Gaussian with in particular

$$E[G(x)] \to \begin{cases} -c_1 M & , x \in \mathcal{C}_1 \\ c_2 M & , x \in \mathcal{C}_2 \end{cases}$$

$$M = \frac{2c_1c_2}{\gamma}\left[-2f'(\tau)\|\mu_2 - \mu_1\|^2 + f''(\tau)(t_2 - t_1)^2 + \frac{4f''(\tau)}{p}\operatorname{tr}(C_1 - C_2)^2\right].$$

Consequences:

- Strong class-size bias

  $\Rightarrow$ Proper threshold must depend on $n_2 - n_1$.

Early 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

$$g(x) = \frac{n_2 - n_1}{n} + \frac{0}{\sqrt{p}} + \underbrace{\frac{G(x)}{p}}_{\text{Relevant terms here!}}$$

- $G(x)$ proportional to γ
- $G(x)$ asymptotically Gaussian with in particular

$$E[G(x)] \to \begin{cases} -c_1 M & , x \in \mathcal{C}_1 \\ c_2 M & , x \in \mathcal{C}_2 \end{cases}$$

$$M = \frac{2c_1c_2}{\gamma}\left[-2f'(\tau)\|\mu_2 - \mu_1\|^2 + f''(\tau)(t_2 - t_1)^2 + \frac{4f''(\tau)}{p}\operatorname{tr}(C_1 - C_2)^2\right].$$

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.

Early 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

$$g(x) = \frac{n_2 - n_1}{n} + \frac{0}{\sqrt{p}} + \underbrace{\frac{G(x)}{p}}_{\text{Relevant terms here!}}$$

- $G(x)$ proportional to γ
- $G(x)$ asymptotically Gaussian with in particular

$$E[G(x)] \to \begin{cases} -c_1 M & , \; x \in \mathcal{C}_1 \\ c_2 M & , \; x \in \mathcal{C}_2 \end{cases}$$

$$M = \frac{2c_1c_2}{\gamma}\left[-2f'(\tau)\|\mu_2 - \mu_1\|^2 + f''(\tau)(t_2 - t_1)^2 + \frac{4f''(\tau)}{p}\operatorname{tr}(C_1 - C_2)^2\right].$$

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 $g(x)$

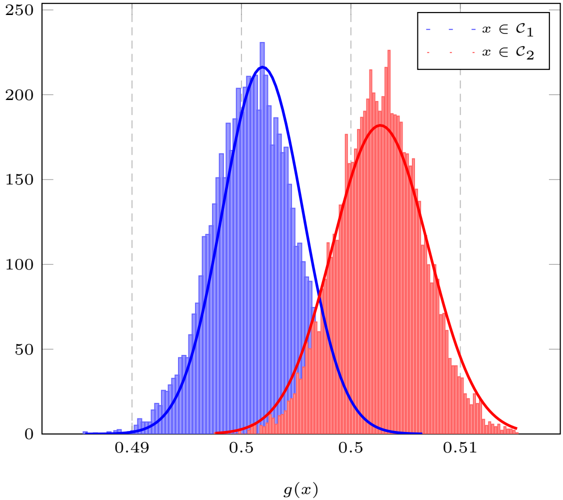


Figure: Values of $g(x)$ for Gaussian x_i's (different means and covariances) versus limiting theoretical distribution, $n = 512$, $p = 1024$.

Outline

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.

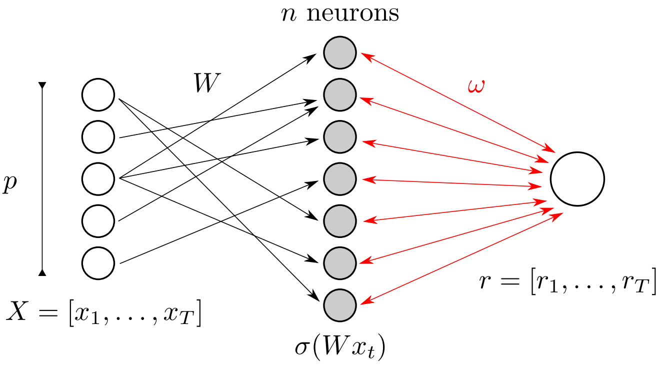
Problem Statement

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.

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.



Objectives: evaluate training and testing MSE performance as $n, p, T \to \infty$

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

$$\Sigma = [\sigma(Wx_1), \ldots, \sigma(Wx_T)]$$
$$\omega = \frac{1}{T}\Sigma\left(\frac{1}{T}\Sigma^\mathsf{T}\Sigma + \gamma I_T\right)^{-1} r.$$

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 $(\hat{x}, \hat{r})$,

$$\hat{E}_\gamma(X, r; \hat{x}, \hat{r}) = \|\hat{r} - \omega^\mathsf{T}\sigma(W\hat{x})\|^2.$$

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 $(\hat{x}, \hat{r})$,

$$\hat{E}_\gamma(X, r; \hat{x}, \hat{r}) = \|\hat{r} - \omega^\mathsf{T}\sigma(W\hat{x})\|^2.$$

- Optimize over γ.

Training MSE:

- Training MSE given by

$$E_\gamma(X,r) = \gamma^2 \frac{1}{T} r^\mathsf{T} \tilde{Q}_\gamma^2 r$$
$$\tilde{Q}_\gamma = \left(\frac{1}{T} \Sigma^\mathsf{T} \Sigma + \gamma I_T \right)^{-2}.$$

Training MSE:

- Training MSE given by

$$E_\gamma(X, r) = \gamma^2 \frac{1}{T} r^\mathsf{T} \tilde{Q}_\gamma^2 r$$
$$\tilde{Q}_\gamma = \left(\frac{1}{T} \Sigma^\mathsf{T} \Sigma + \gamma I_T \right)^{-2}.$$

- Testing MSE given by

$$\hat{E}_\gamma(X, r; \hat{x}, \hat{r}) = \left| \hat{r} - \frac{1}{T} \sigma(W\hat{x})^\mathsf{T} \Sigma \tilde{Q}_\gamma r \right|^2$$

Training MSE:

- Training MSE given by

$$E_\gamma(X,r) = \gamma^2 \frac{1}{T} r^\mathsf{T} \tilde{Q}_\gamma^2 r$$
$$\tilde{Q}_\gamma = \left(\frac{1}{T} \Sigma^\mathsf{T} \Sigma + \gamma I_T \right)^{-2}.$$

- Testing MSE given by

$$\hat{E}_\gamma(X,r;\hat{x},\hat{r}) = \left| \hat{r} - \frac{1}{T} \sigma(W\hat{x})^\mathsf{T} \Sigma \tilde{Q}_\gamma r \right|^2$$

- Requires first a deterministic equivalent $\bar{\tilde{Q}}_\gamma$ for $\tilde{Q}_\gamma$ with non-linear $\sigma(\cdot)$.

Training MSE:

- Training MSE given by

$$E_\gamma(X, r) = \gamma^2 \frac{1}{T} r^\mathsf{T} \tilde{Q}_\gamma^2 r$$
$$\tilde{Q}_\gamma = \left(\frac{1}{T} \Sigma^\mathsf{T} \Sigma + \gamma I_T \right)^{-2}.$$

- Testing MSE given by

$$\hat{E}_\gamma(X, r; \hat{x}, \hat{r}) = \left| \hat{r} - \frac{1}{T} \sigma(W\hat{x})^\mathsf{T} \Sigma \tilde{Q}_\gamma r \right|^2$$

- Requires first a deterministic equivalent $\bar{\tilde{Q}}_\gamma$ for $\tilde{Q}_\gamma$ with non-linear $\sigma(\cdot)$.
- Then deterministic approximation of $\frac{1}{T}\sigma(Wa)^\mathsf{T} \Sigma \tilde{Q}_\gamma b$ for deterministic vectors a, b.

Bai–Silverstein approach:

- Assume $\bar{\tilde{Q}}_\gamma = (F + \gamma I_T)^{-1}$ for some deterministic F.

Bai–Silverstein approach:

- Assume $\bar{\tilde{Q}}_\gamma = (F + \gamma I_T)^{-1}$ for some deterministic F.
- For A deterministic, we manipulate $\frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma - \frac{1}{T}\operatorname{tr} A\bar{\tilde{Q}}_\gamma$, to obtain

$$
\begin{aligned}
\frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma - \frac{1}{T}\operatorname{tr} A\bar{\tilde{Q}}_\gamma &= \frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma \left(F - \frac{1}{T}\Sigma^{\mathsf{T}}\Sigma\right)\bar{\tilde{Q}}_\gamma \\
&= \frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma F\bar{\tilde{Q}}_\gamma - \frac{1}{T}\sum_{i=1}^{n} \frac{1}{T}\Sigma_{i,\cdot}\bar{\tilde{Q}}_\gamma A\tilde{Q}_\gamma\Sigma_{i,\cdot}^{\mathsf{T}} \\
&= \frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma F\bar{\tilde{Q}}_\gamma - \frac{1}{T}\sum_{i=1}^{n} \frac{\frac{1}{T}\Sigma_{i,\cdot}\bar{\tilde{Q}}_\gamma A\tilde{Q}_{\gamma,-i}\Sigma_{i,\cdot}^{\mathsf{T}}}{1 + \frac{1}{T}\Sigma_{i,\cdot}\tilde{Q}_{\gamma,-i}\Sigma_{i,\cdot}^{\mathsf{T}}}
\end{aligned}
$$

where $\tilde{Q}_{\gamma,-i} = (\frac{1}{T}\Sigma^T\Sigma - \frac{1}{T}\Sigma_{i,\cdot}^T\Sigma_{i,\cdot} + \gamma I_T)^{-1}$.

Bai–Silverstein approach:

- Assume $\bar{\tilde{Q}}_\gamma = (F + \gamma I_T)^{-1}$ for some deterministic F.
- For A deterministic, we manipulate $\frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma - \frac{1}{T}\operatorname{tr} A\bar{\tilde{Q}}_\gamma$, to obtain

$$\begin{aligned}
\frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma - \frac{1}{T}\operatorname{tr} A\bar{\tilde{Q}}_\gamma &= \frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma \left(F - \frac{1}{T}\Sigma^\mathsf{T}\Sigma \right) \bar{\tilde{Q}}_\gamma \\
&= \frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma F\bar{\tilde{Q}}_\gamma - \frac{1}{T}\sum_{i=1}^{n} \frac{1}{T}\Sigma_{i,\cdot}\bar{\tilde{Q}}_\gamma A\tilde{Q}_\gamma \Sigma_{i,\cdot}^\mathsf{T} \\
&= \frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma F\bar{\tilde{Q}}_\gamma - \frac{1}{T}\sum_{i=1}^{n} \frac{\frac{1}{T}\Sigma_{i,\cdot}\bar{\tilde{Q}}_\gamma A\tilde{Q}_{\gamma,-i}\Sigma_{i,\cdot}^\mathsf{T}}{1 + \frac{1}{T}\Sigma_{i,\cdot}\tilde{Q}_{\gamma,-i}\Sigma_{i,\cdot}^\mathsf{T}}
\end{aligned}$$

  where $\tilde{Q}_{\gamma,-i} = (\frac{1}{T}\Sigma^T\Sigma - \frac{1}{T}\Sigma_{i,\cdot}^T\Sigma_{i,\cdot} + \gamma I_T)^{-1}$.
- Here $\Sigma_{i,\cdot} = \sigma(W_{i,\cdot}X)$ independent of $\tilde{Q}_{\gamma,-i}$

Bai–Silverstein approach:

- Assume $\bar{\tilde{Q}}_\gamma = (F + \gamma I_T)^{-1}$ for some deterministic F.
- For A deterministic, we manipulate $\frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma - \frac{1}{T}\operatorname{tr} A\bar{\tilde{Q}}_\gamma$, to obtain

$$\begin{aligned}
\frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma - \frac{1}{T}\operatorname{tr} A\bar{\tilde{Q}}_\gamma &= \frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma \left(F - \frac{1}{T}\Sigma^\mathsf{T}\Sigma\right)\bar{\tilde{Q}}_\gamma \\
&= \frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma F\bar{\tilde{Q}}_\gamma - \frac{1}{T}\sum_{i=1}^{n} \frac{1}{T}\Sigma_{i,\cdot}\bar{\tilde{Q}}_\gamma A\tilde{Q}_\gamma \Sigma_{i,\cdot}^\mathsf{T} \\
&= \frac{1}{T}\operatorname{tr} A\tilde{Q}_\gamma F\bar{\tilde{Q}}_\gamma - \frac{1}{T}\sum_{i=1}^{n} \frac{\frac{1}{T}\Sigma_{i,\cdot}\bar{\tilde{Q}}_\gamma A\tilde{Q}_{\gamma,-i}\Sigma_{i,\cdot}^\mathsf{T}}{1 + \frac{1}{T}\Sigma_{i,\cdot}\tilde{Q}_{\gamma,-i}\Sigma_{i,\cdot}^\mathsf{T}}
\end{aligned}$$

  where $\tilde{Q}_{\gamma,-i} = (\frac{1}{T}\Sigma^T\Sigma - \frac{1}{T}\Sigma_{i,\cdot}^T\Sigma_{i,\cdot} + \gamma I_T)^{-1}$.

- Here $\Sigma_{i,\cdot} = \sigma(W_{i,\cdot}X)$ independent of $\tilde{Q}_{\gamma,-i}$
  $\longrightarrow$ **reasoning broken on co-resolvent!** (lucky that we need $\tilde{Q}_\gamma$ and not Q_γ)

(Conjectured) updated trace lemma:

Lemma

For A deterministic and $\sigma(t)$ polynomial, W_{ij} i.i.d. $E[W_{ij}] = 0$, $E[W_{ij}^k] = \frac{m_k}{n^{k/2}}$,

$$\frac{1}{T}\Sigma_{i,\cdot} A \Sigma_{i,\cdot}^{\mathsf{T}} - \frac{1}{T}\operatorname{tr}\Phi_X A \xrightarrow{\text{a.s.}} 0$$

with

$$\Phi_X = E\left[\frac{1}{n}\sigma(WX)^{\mathsf{T}}\sigma(WX)\right].$$

(Conjectured) updated trace lemma:

Lemma

For A deterministic and $\sigma(t)$ polynomial, W_{ij} i.i.d. $E[W_{ij}] = 0$, $E[W_{ij}^k] = \frac{m_k}{n^{k/2}}$,

$$\frac{1}{T}\Sigma_{i,\cdot}A\Sigma_{i,\cdot}^{\mathsf{T}} - \frac{1}{T}\operatorname{tr}\Phi_X A \xrightarrow{\text{a.s.}} 0$$

with

$$\Phi_X = E\left[\frac{1}{n}\sigma(WX)^{\mathsf{T}}\sigma(WX)\right].$$

For instance,

- for $\sigma(t) = t$,

$$\Phi_X = \frac{m_2}{n}X^{\mathsf{T}}X.$$

- for $\sigma(t) = t^2$,

$$\Phi_X = \frac{m_2^2}{n^2}\left(\sigma(X^{\mathsf{T}}X) + 2\sigma(X)^{\mathsf{T}}1_p1_p^{\mathsf{T}}\sigma(X)\right) + \frac{m_4 - 3m_2^2}{n^2}\sigma(X)^{\mathsf{T}}\sigma(X).$$

Results

Early Results:

- (Conjectured) deterministic equivalent: as $n, p, T \to \infty$ with $\sigma(t)$ polynomial, W_{ij} i.i.d. $E[W_{ij}] = 0$, $E[W_{ij}^k] = \frac{m_k}{n^{k/2}}$,

$$\tilde{Q}_\gamma \leftrightarrow \bar{\tilde{Q}}_\gamma$$

where

$$\bar{\tilde{Q}}_\gamma = \left(\frac{n}{T} \frac{1}{1+\delta} \Phi_X + \gamma I_T \right)^{-1}$$

$$\delta = \frac{1}{T} \operatorname{tr} \Phi_X \left(\frac{n}{T} \frac{1}{1+\delta} \Phi_X + \gamma I_T \right)^{-1}$$

Results

Early Results:

- (Conjectured) deterministic equivalent: as $n, p, T \to \infty$ with $\sigma(t)$ polynomial, W_{ij} i.i.d. $E[W_{ij}] = 0$, $E[W_{ij}^k] = \frac{m_k}{n^{k/2}}$,

$$\tilde{Q}_\gamma \leftrightarrow \bar{\tilde{Q}}_\gamma$$

where

$$\bar{\tilde{Q}}_\gamma = \left(\frac{n}{T} \frac{1}{1+\delta} \Phi_X + \gamma I_T \right)^{-1}$$

$$\delta = \frac{1}{T} \operatorname{tr} \Phi_X \left(\frac{n}{T} \frac{1}{1+\delta} \Phi_X + \gamma I_T \right)^{-1}$$

We also denote

$$\delta' = (1+\delta) \frac{\frac{1}{T} \operatorname{tr} \Phi_X \bar{\tilde{Q}}_\gamma^2}{1 + \gamma \frac{1}{T} \operatorname{tr} \Phi_X \bar{\tilde{Q}}_\gamma^2}.$$

Results

Early Results:

- Training performance:

$$E_\alpha(X,r) \leftrightarrow \gamma^2 \frac{1}{T} r^\mathsf{T} \bar{\tilde{Q}}_\gamma \left[\frac{n}{T} \frac{\delta'}{(1+\delta)^2} \Phi_X + I_T \right] \bar{\tilde{Q}}_\gamma r.$$

Results

Early Results:

- Training performance:

$$E_\alpha(X,r) \leftrightarrow \gamma^2 \frac{1}{T} r^\mathsf{T} \bar{\bar{Q}}_\gamma \left[\frac{n}{T} \frac{\delta'}{(1+\delta)^2} \Phi_X + I_T \right] \bar{\bar{Q}}_\gamma r.$$

- Testing performance:

$$\hat{E}_\alpha(X,r;\hat{x},\hat{r}) \leftrightarrow \left| \hat{r} - \frac{n}{T} \frac{1}{1+\delta} \Phi_{X,\hat{x}}^\mathsf{T} \bar{\bar{Q}}_\gamma r \right|^2$$

with

$$\Phi_{X,\hat{x}} = E\left[\frac{1}{n} \sigma(WX)^\mathsf{T} \sigma(W\hat{x}) \right].$$

In particular, for $\sigma(t) = t$, $\Phi_{X,\hat{x}} = \frac{m_2}{n} X^\mathsf{T} \hat{x}$, and, for $\sigma(t) = t^2$, $\Phi_{X,\hat{x}} = \frac{m_2^2}{n^2} \left(\sigma(X^\mathsf{T}\hat{x}) + 2\sigma(X)^\mathsf{T} 1_p 1_p^\mathsf{T} \sigma(\hat{x}) \right) + \frac{m_4 - 3m_2^2}{n^2} \sigma(X)^\mathsf{T} \sigma(\hat{x})$.

Test on MNIST data

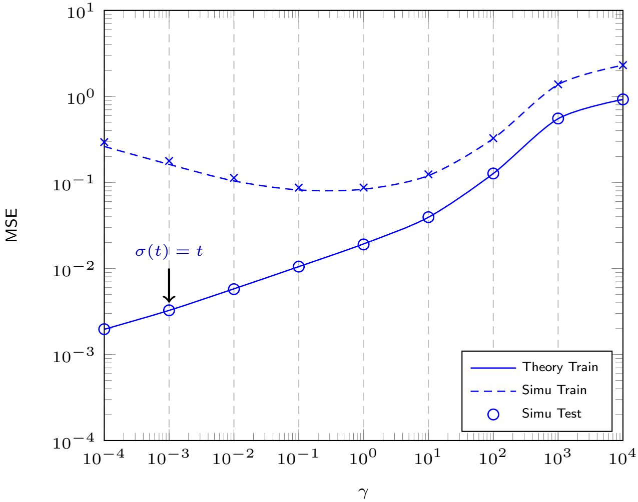


Figure: MSE Train and Test Performance for $\sigma(t) = t$ and $\sigma(t) = t^2$, as a function of γ, for 2-class MNIST data (zeros, ones), $n = 512$, $T = 512$, $p = 784$.

Test on MNIST data

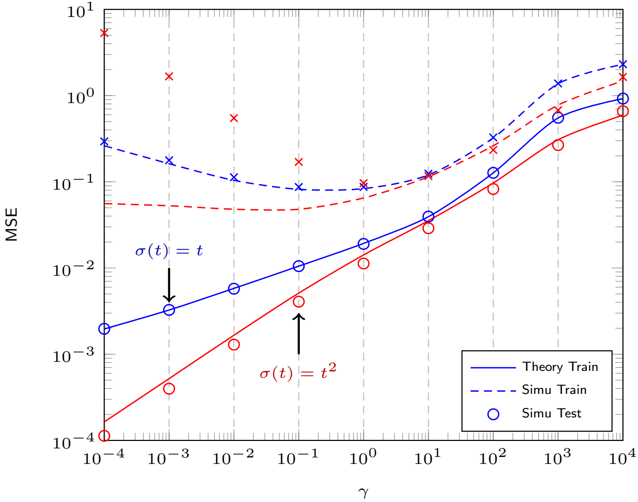


Figure: MSE Train and Test Performance for $\sigma(t) = t$ and $\sigma(t) = t^2$, as a function of γ, for 2-class MNIST data (zeros, ones), $n = 512$, $T = 512$, $p = 784$.

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?
- Connection to auto-encoders?
- Introduction of non-linearity to more involved structures (ESN, deep nets?).

Outline

Problem Statement

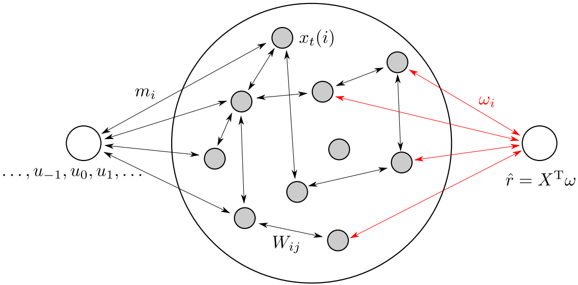
Echo-state Neural Networks (ESN)
Neural Net with n nodes, states $x_t \in \mathbb{R}^n$, defined recursively through

$$x_{t+1} = \sigma\left(Wx_t + mu_{t+1} + \eta\varepsilon_{t+1}\right)$$

where

- W **fixed (often random) connectivity matrix**
- m input to network connectivity (also fixed)
- ε_t in-network noise (ensures stability)

$\Rightarrow$ We take here $\sigma(x) = x$.



ESN Performance

Training and Testing tasks
From input $u \in \mathbb{R}^T$ and expected output $r \in \mathbb{R}^T$,

- Given r, train the ESN by setting **network to sink link**

$$\omega = \begin{cases} (XX^\mathsf{T})^{-1}Xr & , T > n \\ X(X^\mathsf{T}X)^{-1}r & , T \leq n \end{cases}$$

with $X = [x_1, \ldots, x_T] \in \mathbb{R}^{n \times T}$ (so that $\|r - X^\mathsf{T}\omega\|$ minimized).

ESN Performance

Training and Testing tasks

From input $u \in \mathbb{R}^T$ and expected output $r \in \mathbb{R}^T$,

- Given r, train the ESN by setting **network to sink link**

$$\omega = \begin{cases} (XX^\mathsf{T})^{-1}Xr & , T > n \\ X(X^\mathsf{T}X)^{-1}r & , T \leq n \end{cases}$$

  with $X = [x_1, \ldots, x_T] \in \mathbb{R}^{n \times T}$ (so that $\|r - X^\mathsf{T}\omega\|$ minimized).
- For unknown $\hat{r} \in \mathbb{R}^{\hat{T}}$ and input $\hat{u} \in \mathbb{R}^{\hat{T}}$, test the ESN by setting

$$\hat{y} = \hat{X}^\mathsf{T}\omega.$$

ESN Performance

Training and Testing tasks
From input $u \in \mathbb{R}^T$ and expected output $r \in \mathbb{R}^T$,

- Given r, train the ESN by setting **network to sink link**

$$\omega = \begin{cases} (XX^\mathsf{T})^{-1}Xr & , T > n \\ X(X^\mathsf{T}X)^{-1}r & , T \leq n \end{cases}$$

  with $X = [x_1, \ldots, x_T] \in \mathbb{R}^{n \times T}$ (so that $\|r - X^\mathsf{T}\omega\|$ minimized).

- For unknown $\hat{r} \in \mathbb{R}^{\hat{T}}$ and input $\hat{u} \in \mathbb{R}^{\hat{T}}$, test the ESN by setting

$$\hat{y} = \hat{X}^\mathsf{T}\omega.$$

Training Performance

$$E_\eta(u, r) \equiv \frac{1}{T}\left\| r - X^\mathsf{T}\omega \right\|^2 = \lim_{\gamma \downarrow 0} \gamma \frac{1}{T} r^\mathsf{T} \tilde{Q}_\gamma r.$$

with $\tilde{Q}_\gamma \equiv (\frac{1}{T}X^\mathsf{T}X + \gamma I_T)^{-1}$, random matrix resolvent.

Training and Testing tasks
From input $u \in \mathbb{R}^T$ and expected output $r \in \mathbb{R}^T$,

- Given r, train the ESN by setting **network to sink link**

$$\omega = \begin{cases} (XX^\mathsf{T})^{-1}Xr & , T > n \\ X(X^\mathsf{T}X)^{-1}r & , T \leq n \end{cases}$$

  with $X = [x_1, \ldots, x_T] \in \mathbb{R}^{n \times T}$ (so that $\|r - X^\mathsf{T}\omega\|$ minimized).
- For unknown $\hat{r} \in \mathbb{R}^{\hat{T}}$ and input $\hat{u} \in \mathbb{R}^{\hat{T}}$, test the ESN by setting

$$\hat{y} = \hat{X}^\mathsf{T}\omega.$$

Training Performance

$$E_\eta(u,r) \equiv \frac{1}{T}\left\|r - X^\mathsf{T}\omega\right\|^2 = \lim_{\gamma \downarrow 0} \gamma \frac{1}{T} r^\mathsf{T} \tilde{Q}_\gamma r.$$

with $\tilde{Q}_\gamma \equiv (\frac{1}{T}X^\mathsf{T}X + \gamma I_T)^{-1}$, random matrix resolvent.

Testing Performance

$$\begin{aligned} \hat{E}_\eta(u,r;\hat{u},\hat{r}) &= \frac{1}{\hat{T}}\left\|\hat{r} - \hat{X}^\mathsf{T}\omega\right\|^2 \\ &= \lim_{\gamma \downarrow 0} \frac{1}{\hat{T}}\|\hat{r}\|^2 + \frac{1}{T^2\hat{T}} r^\mathsf{T}\tilde{Q}_\gamma X^\mathsf{T}\hat{X}\hat{X}^\mathsf{T}X\tilde{Q}_\gamma r - \frac{2}{T\hat{T}}\hat{r}^\mathsf{T}\hat{X}^\mathsf{T}X\tilde{Q}_\gamma r \end{aligned}$$

Training Performance

Theorem (Training MSE for fixed W)

As $n, T \to \infty$, $n/T \to c < 1$,

$$E_\eta(u,r) \leftrightarrow \frac{1}{T} r^\mathsf{T} \left(I_T + \mathcal{R} + \frac{1}{\eta^2} U^\mathsf{T} \left\{ m^\mathsf{T} (W^i)^\mathsf{T} \tilde{\mathcal{R}}^{-1} W^j m \right\}_{i,j=0}^{T-1} U \right)^{-1} r.$$

where $U_{ij} = u_{i-j}$ and $\mathcal{R}$, $\tilde{\mathcal{R}}$, solution to

$$\mathcal{R} = c \left\{ \frac{1}{n} \mathrm{tr} \left(S_{i-j} \tilde{\mathcal{R}}^{-1} \right) \right\}_{i,j=1}^{T}$$

$$\tilde{\mathcal{R}} = \sum_{q=-\infty}^{\infty} \frac{1}{T} \mathrm{tr} \left(J^q (I_T + \mathcal{R})^{-1} \right) S_q.$$

with $[J^q]_{ij} \equiv \boldsymbol{\delta}_{i+q,j}$ and $S_q \equiv \sum_{k \geq 0} W^{k+(-q)^+} (W^{k+q^+})^\mathsf{T}$.

Theorem (Training MSE for fixed W)

As $n, T \to \infty$, $n/T \to c < 1$,

$$E_\eta(u,r) \leftrightarrow \frac{1}{T} r^\mathsf{T} \left(I_T + \mathcal{R} + \frac{1}{\eta^2} U^\mathsf{T} \left\{ m^\mathsf{T} (W^i)^\mathsf{T} \tilde{\mathcal{R}}^{-1} W^j m \right\}_{i,j=0}^{T-1} U \right)^{-1} r.$$

where $U_{ij} = u_{i-j}$ and $\mathcal{R}$, $\tilde{\mathcal{R}}$, solution to

$$\mathcal{R} = c \left\{ \frac{1}{n} \operatorname{tr} \left(S_{i-j} \tilde{\mathcal{R}}^{-1} \right) \right\}_{i,j=1}^{T}$$

$$\tilde{\mathcal{R}} = \sum_{q=-\infty}^{\infty} \frac{1}{T} \operatorname{tr} \left(J^q (I_T + \mathcal{R})^{-1} \right) S_q.$$

with $[J^q]_{ij} \equiv \boldsymbol{\delta}_{i+q,j}$ and $S_q \equiv \sum_{k \geq 0} W^{k+(-q)^+} (W^{k+q^+})^\mathsf{T}$.

$\longrightarrow$ **When $c = 0$,**

$$E_\eta(u,r) \leftrightarrow \frac{1}{T} r^\mathsf{T} \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} \left\{ m^\mathsf{T} (W^i)^\mathsf{T} S_0^{-1} W^j m \right\}_{i,j=0}^{T-1} U \right)^{-1} r.$$

- Note that columns of U are delayed versions of u_t.

Theorem (Testing MSE for fixed W)

As $n, T \to \infty$, $n/T \to c < 1$,

$$\hat{E}_\eta(u, r; \hat{u}, \hat{r}) \leftrightarrow \left\| \frac{1}{\eta^2 \sqrt{T}} \hat{A}^\mathsf{T} \mathcal{Q} A (\boldsymbol{\delta}_{c<1} I_T + \mathcal{R})^{-1} r - \frac{1}{\sqrt{\hat{T}}} \hat{r} \right\|^2 + \frac{1}{T} r^\mathsf{T} \tilde{\mathcal{Q}} \mathcal{G} \tilde{\mathcal{Q}} r$$
$$+ \frac{1}{\eta^2 T} r^\mathsf{T} (\boldsymbol{\delta}_{c<1} I_T + \mathcal{R})^{-1} A^\mathsf{T} \mathcal{Q} \left[S_0 + \tilde{\mathcal{G}} \right] \mathcal{Q} A (\boldsymbol{\delta}_{c<1} I_T + \mathcal{R})^{-1} r$$

where $A = MU$, $\hat{A} = \hat{M}\hat{U}$, $M = [m, Wm, \ldots, W^{T-1} m]$, *and* $\mathcal{G}$, $\tilde{\mathcal{G}}$, *solution to*

$$\mathcal{G} = c \left\{ \frac{1}{n} \operatorname{tr} \left(S_{i-j} \tilde{\mathcal{R}}^{-1} \left[S_0 + \tilde{\mathcal{G}} \right] \tilde{\mathcal{R}}^{-1} \right) \right\}_{i,j=1}^{T}$$
$$\tilde{\mathcal{G}} = \sum_{q=-\infty}^{\infty} \frac{1}{T} \operatorname{tr} \left(J^q (I_T + \mathcal{R})^{-1} \mathcal{G} (I_T + \mathcal{R})^{-1} \right) S_q.$$

Theorem (Testing MSE for fixed W)

As $n, T \to \infty$, $n/T \to c < 1$,

$$\hat{E}_\eta(u, r; \hat{u}, \hat{r}) \leftrightarrow \left\| \frac{1}{\eta^2 \sqrt{T}} \hat{A}^\mathsf{T} \mathcal{Q} A (\boldsymbol{\delta}_{c<1} I_T + \mathcal{R})^{-1} r - \frac{1}{\sqrt{\hat{T}}} \hat{r} \right\|^2 + \frac{1}{T} r^\mathsf{T} \tilde{\mathcal{Q}} \mathcal{G} \tilde{\mathcal{Q}} r$$
$$+ \frac{1}{\eta^2 T} r^\mathsf{T} (\boldsymbol{\delta}_{c<1} I_T + \mathcal{R})^{-1} A^\mathsf{T} \mathcal{Q} \left[S_0 + \tilde{\mathcal{G}} \right] \mathcal{Q} A (\boldsymbol{\delta}_{c<1} I_T + \mathcal{R})^{-1} r$$

where $A = MU$, $\hat{A} = \hat{M}\hat{U}$, $M = [m, Wm, \ldots, W^{T-1}m]$, and $\mathcal{G}$, $\tilde{\mathcal{G}}$, solution to

$$\mathcal{G} = c \left\{ \frac{1}{n} \operatorname{tr} \left(S_{i-j} \tilde{\mathcal{R}}^{-1} \left[S_0 + \tilde{\mathcal{G}} \right] \tilde{\mathcal{R}}^{-1} \right) \right\}_{i,j=1}^T$$
$$\tilde{\mathcal{G}} = \sum_{q=-\infty}^{\infty} \frac{1}{T} \operatorname{tr} \left(J^q (I_T + \mathcal{R})^{-1} \mathcal{G} (I_T + \mathcal{R})^{-1} \right) S_q.$$

$\longrightarrow$ **When $c = 0$,**

$$\hat{E}_\eta(u, r; \hat{u}, \hat{r}) \leftrightarrow \left\| \frac{1}{\sqrt{T}} \hat{A}^\mathsf{T} \left(\eta^2 S_0 + AA^\mathsf{T} \right)^{-1} Ar - \frac{1}{\sqrt{\hat{T}}} \hat{r} \right\|^2 + \frac{1}{T} r^\mathsf{T} A^\mathsf{T} \left(\eta^2 S_0 + AA^\mathsf{T} \right)^{-2} Ar.$$

$(S_0 = \sum_{k \geq 0} W^k (W^k)^\mathsf{T})$.

ESN Performance for Random Haar W

- Letting $W = \sigma Z$ with Z orthogonal and orthogonally invariant,

$$E_\eta(u,r) \leftrightarrow (1-c)\frac{1}{T} r^\mathsf{T} \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} D U \right)^{-1} r$$

$$\hat{E}_\eta(u,r;\hat{u},\hat{r}) \leftrightarrow \left\| \frac{1}{\eta^2\sqrt{T}} \hat{U}^\mathsf{T} \hat{D} U \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} D U \right)^{-1} r - \frac{1}{\sqrt{\hat{T}}} \hat{r} \right\|^2$$
$$+ \frac{1}{1-c} \frac{1}{T} r^\mathsf{T} \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} D U \right)^{-1} r - \frac{1}{T} r^\mathsf{T} \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} D U \right)^{-2} r$$

where

$$D \equiv \left\{ m^\mathsf{T} (W^i)^\mathsf{T} S_0^{-1} W^j m \right\}_{i,j=0}^{T-1}$$
$$\hat{D} \equiv \left\{ m^\mathsf{T} (W^i)^\mathsf{T} S_0^{-1} W^j m \right\}_{i,j=0}^{\hat{T}-1,T-1}.$$

ESN Performance for Random Haar W

- Letting $W = \sigma Z$ with Z orthogonal and orthogonally invariant,

$$E_\eta(u,r) \leftrightarrow (1-c)\frac{1}{T}r^\mathsf{T}\left(I_T + \frac{1}{\eta^2}U^\mathsf{T}DU\right)^{-1} r$$

$$\hat{E}_\eta(u,r;\hat{u},\hat{r}) \leftrightarrow \left\|\frac{1}{\eta^2\sqrt{T}}\hat{U}^\mathsf{T}\hat{D}U\left(I_T + \frac{1}{\eta^2}U^\mathsf{T}DU\right)^{-1} r - \frac{1}{\sqrt{\hat{T}}}\hat{r}\right\|^2$$
$$+ \frac{1}{1-c}\frac{1}{T}r^\mathsf{T}\left(I_T + \frac{1}{\eta^2}U^\mathsf{T}DU\right)^{-1} r - \frac{1}{T}r^\mathsf{T}\left(I_T + \frac{1}{\eta^2}U^\mathsf{T}DU\right)^{-2} r$$

where

$$D \equiv \left\{m^\mathsf{T}(W^i)^\mathsf{T}S_0^{-1}W^jm\right\}_{i,j=0}^{T-1}$$
$$\hat{D} \equiv \left\{m^\mathsf{T}(W^i)^\mathsf{T}S_0^{-1}W^jm\right\}_{i,j=0}^{\hat{T}-1,T-1}.$$

- If m independent of W, D diagonal,

$$D_{ii} \leftrightarrow (1-\sigma^2)\sigma^{2(i-1)}.$$

Multimemory Connectivity

Analysis suggests taking $W = \operatorname{diag}(W_1, \ldots, W_k)$, $W_j = \sigma_j Z_j$, $Z_j \in \mathbb{R}^{n_j \times n_j}$ Haar, so that

$$D_{ii} \leftrightarrow \frac{\sum_{j=1}^{k} c_j \sigma_j^{2(i-1)}}{\sum_{j=1}^{k} c_j (1-\sigma_j^2)^{-1}}.$$

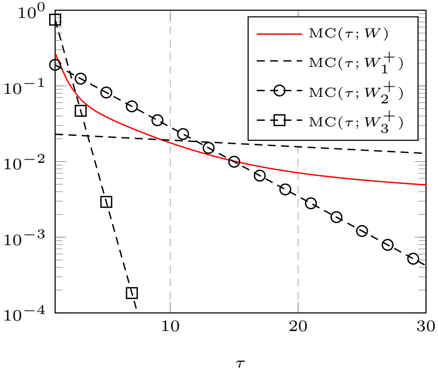


Figure: Memory curve (MC) for $W = \operatorname{diag}(W_1, W_2, W_3)$, $W_j = \sigma_j Z_j$, $Z_j \in \mathbb{R}^{n_j \times n_j}$ Haar distributed, $\sigma_1 = .99$, $n_1/n = .01$, $\sigma_2 = .9$, $n_2/n = .1$, and $\sigma_3 = .5$, $n_3/n = .89$. The matrices W_i^+ are defined by $W_i^+ = \sigma_i Z_i^+$, with $Z_i^+ \in \mathbb{R}^{n \times n}$ Haar distributed.

Multimemory Connectivity

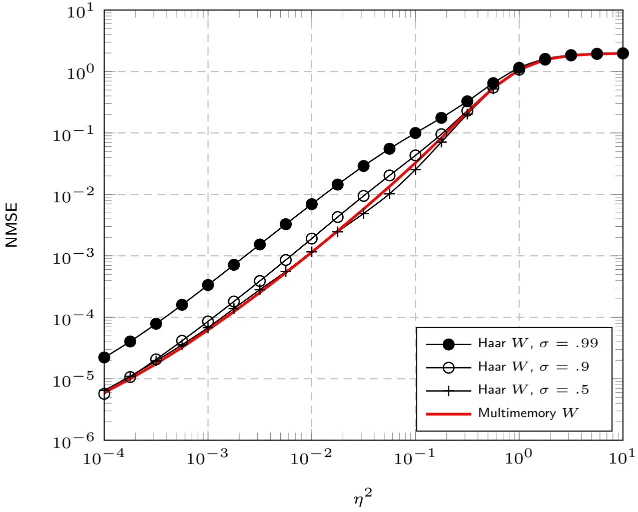


Figure: Mackey Glass one-step ahead task, W (multimemory) versus $W_1^+ = .99Z_1^+$, $W_2^+ = .9Z_2^+$, $W_3^+ = .5Z_3^+$, $n = 400$, $T = \hat{T} = 800$.

Example: Mackey-Glass Model, random matrix convergence

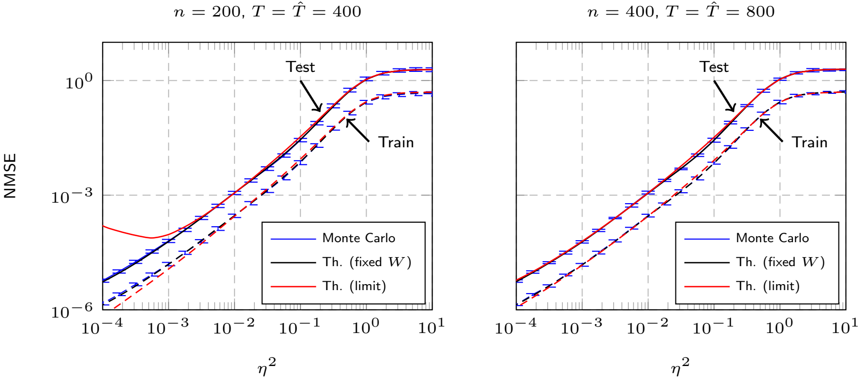


Figure: Mackey Glass one-step ahead task, W multimemory, $n = 200$, $T = \hat{T} = 400$ (left) and $n = 400$, $T = \hat{T} = 800$ (right).

Robustness to outliers

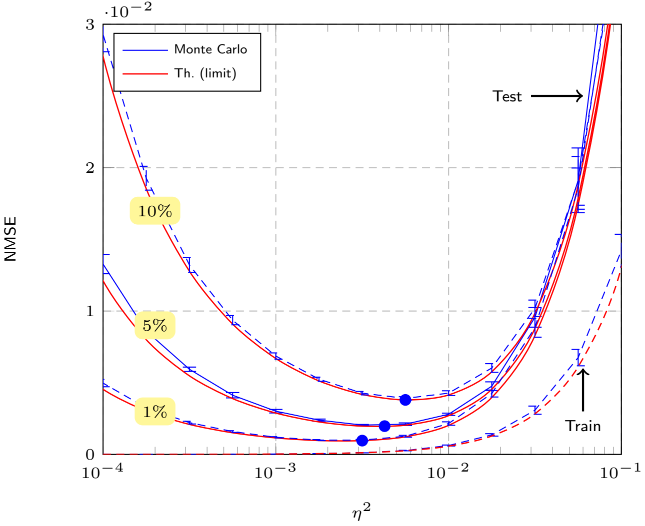


Figure: Mackey-Glass one-step ahead task with 1% or 10% impulsive $\mathcal{N}(0, .01)$ noise pollution in test data inputs, W Haar with $\sigma = .9$, $n = 400$, $T = \hat{T} = 1000$.

Robustness to outliers

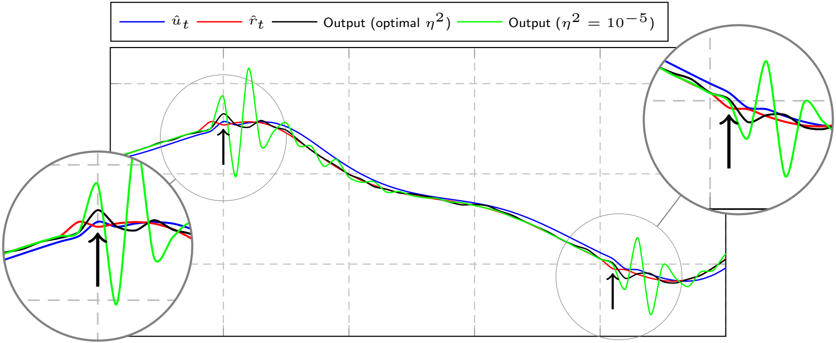


Figure: Realization of a 1% $\mathcal{N}(0, .01)$-noisy Mackey-Glass sequence versus network output, W Haar with $\sigma = .9$, $n = 400$, $T = \hat{T} = 1000$.

Non-symmetric versus symmetric W

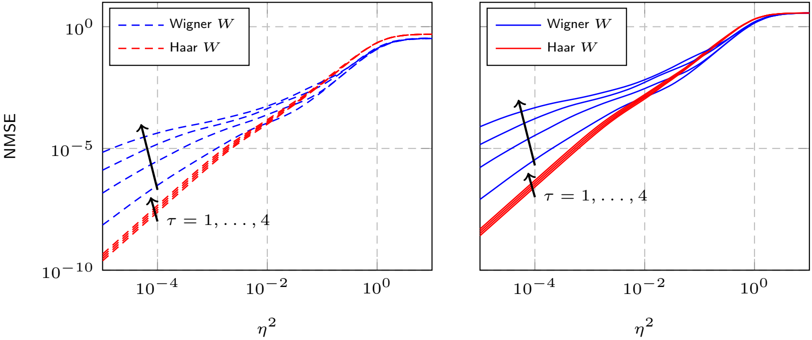


Figure: Training (left) and testing (right) performance of a τ-delay task for $\tau \in \{1, \ldots, 4\}$ for Haar versus Wigner W, $\sigma = .9$ and $n = 200$, $T = \hat{T} = 400$.

Outline

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 ($n > N$)

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

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 ($n > N$)

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

- **[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 $E[x_1] = 0$, $E[x_1 x_1^*] = C_N$:

- 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 ($n > N$)

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

- **[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

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

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

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)

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

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

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 $\phi(x) \triangleq xu(x)$ increasing of supremum ϕ_∞ with

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

Recent Theoretical Results

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 $\hat{S}_N$.

Recent Theoretical Results

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 $\hat{S}_N$.

- 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.

Recent Theoretical Results

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 $\hat{S}_N$.

- 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.

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,

$$\hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} v(\tau_i \gamma_N) x_i x_i^*$$

and γ_N *unique solution of*

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

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,

$$\hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} v(\tau_i \gamma_N) x_i x_i^*$$

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. ($\mu_N^X \triangleq \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\delta}_{\lambda_i(X)}$)

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,

$$\hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} v(\tau_i \gamma_N) x_i x_i^*$$

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**: $\max_{1 \le i \le N} |\lambda_i(\hat{C}_N) - \lambda_i(\hat{S}_N)| \xrightarrow{\text{a.s.}} 0.$

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,

$$\hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^{n} v(\tau_i \gamma_N) x_i x_i^*$$

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!)

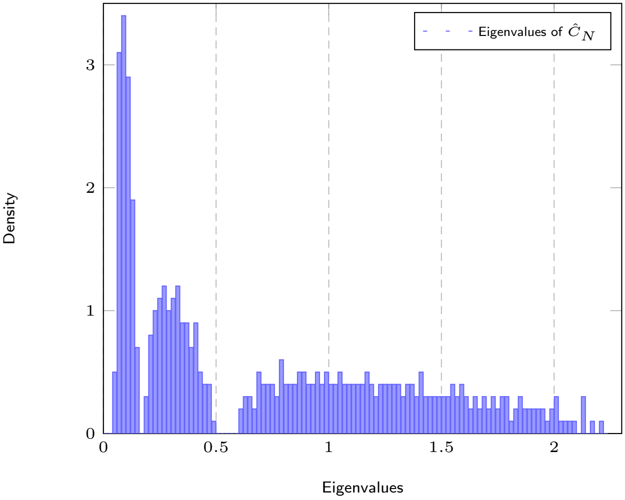


Figure: $n = 2500$, $N = 500$, $C_N = \text{diag}(I_{125}, 3I_{125}, 10I_{250})$, $\tau_i \sim \Gamma(.5, 2)$ i.i.d.

Large dimensional behavior

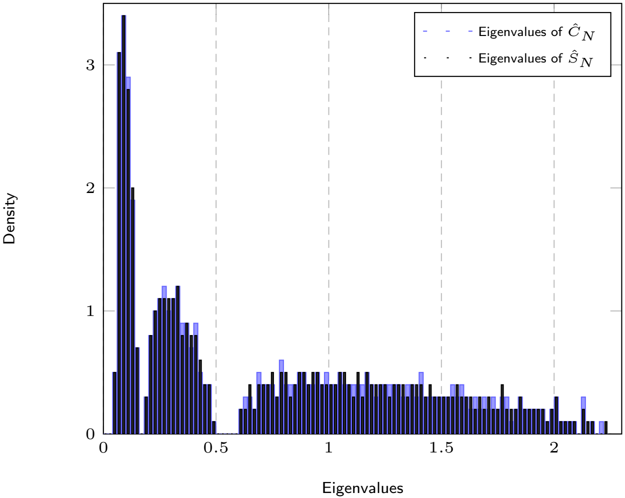


Figure: $n = 2500$, $N = 500$, $C_N = \mathrm{diag}(I_{125}, 3I_{125}, 10I_{250})$, $\tau_i \sim \Gamma(.5, 2)$ i.i.d.

Large dimensional behavior

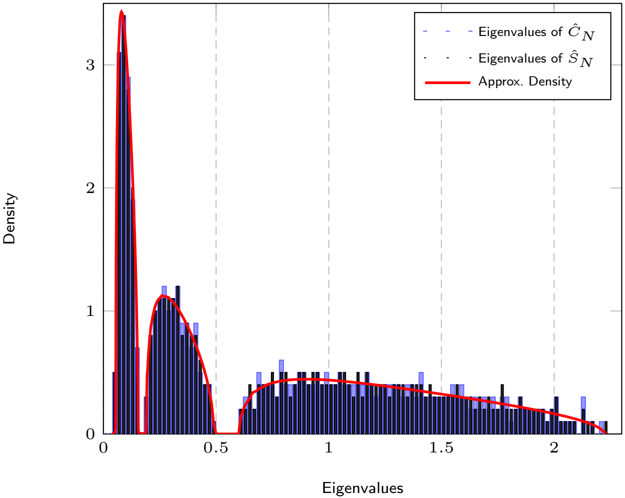


Figure: $n = 2500$, $N = 500$, $C_N = \operatorname{diag}(I_{125}, 3I_{125}, 10I_{250})$, $\tau_i \sim \Gamma(.5, 2)$ i.i.d.

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*

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

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

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

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

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

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

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

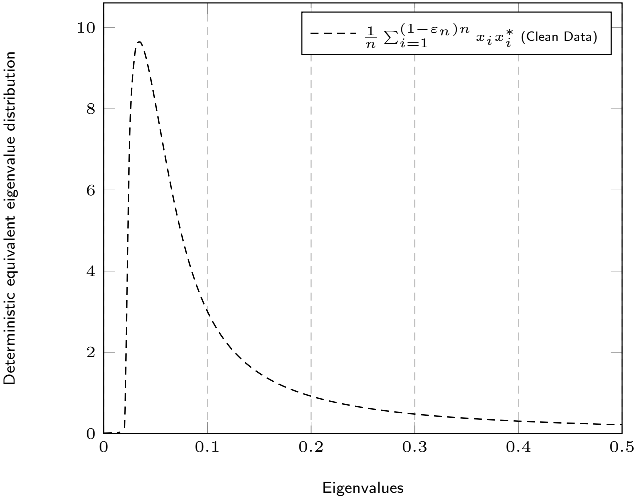


Figure: Limiting eigenvalue distributions. $[C_N]_{ij} = .9^{|i-j|}$, $D_N = I_N$, $\varepsilon = .05$.

Outlier Data

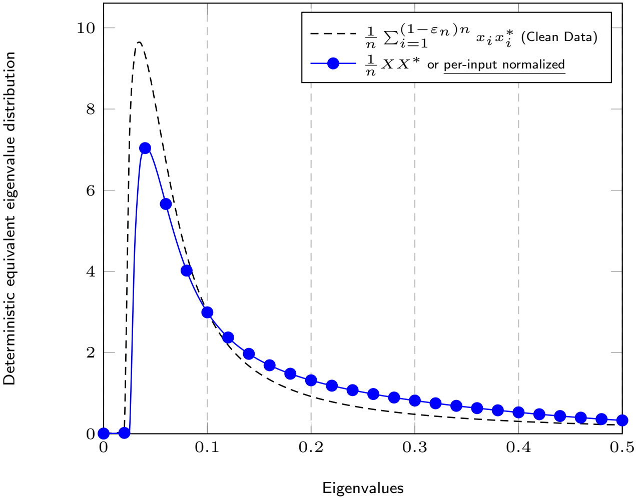


Figure: Limiting eigenvalue distributions. $[C_N]_{ij} = .9^{|i-j|}$, $D_N = I_N$, $\varepsilon = .05$.

Outlier Data

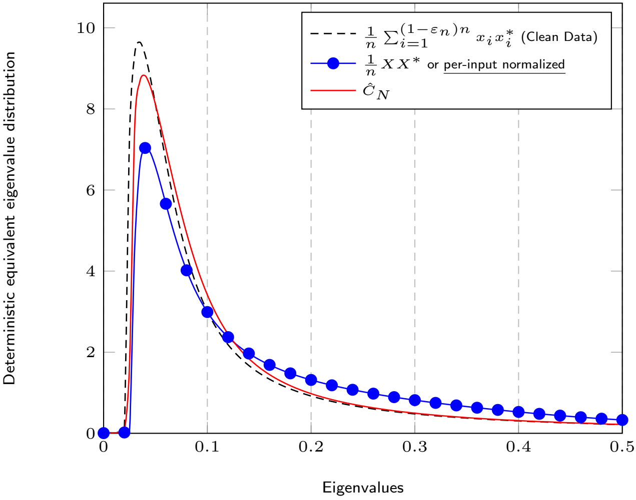


Figure: Limiting eigenvalue distributions. $[C_N]_{ij} = .9^{|i-j|}$, $D_N = I_N$, $\varepsilon = .05$.

Short Term Objectives:

- **Robust statistics**.
  - ✎ Joint mean and covariance robust estimation
  - Study of robust regression (preliminary works exist already using strikingly different approaches)

Short Term Objectives:

- **Robust statistics.**
  - ✎ Joint mean and covariance robust estimation
  - Study of robust regression (preliminary works exist already using strikingly different approaches)
- **Kernel methods.**
  - ✔ Subspace spectral clustering (dramatically different case of $f'(\tau) = 0$)
  - ✎ Spectral clustering with outer product kernel $f(x^{\mathsf{T}} y)$
  - ✎ Semi-supervised learning, kernel approaches.
  - ✎ Support vector machines (SVM).

Short Term Objectives:

- **Robust statistics.**
  - ✎ Joint mean and covariance robust estimation
  - Study of robust regression (preliminary works exist already using strikingly different approaches)
- **Kernel methods.**
  - ✔ Subspace spectral clustering (dramatically different case of $f'(\tau) = 0$)
  - ✎ Spectral clustering with outer product kernel $f(x^{\mathsf{T}}y)$
  - ✎ Semi-supervised learning, kernel approaches.
  - ✎ Support vector machines (SVM).
- **Community detection.**
  - ✔ Complete study of eigenvector contents in adjacency/mpdularity methods.
  - Study of Bethe Hessian approach.
  - Analysis of non-necessarily spectral approaches (wavelet approaches).

Short Term Objectives:

- **Robust statistics**.
  - ✎ Joint mean and covariance robust estimation
  - Study of robust regression (preliminary works exist already using strikingly different approaches)
- **Kernel methods**.
  - ✔ Subspace spectral clustering (dramatically different case of $f'(\tau) = 0$)
  - ✎ Spectral clustering with outer product kernel $f(x^{\mathsf{T}} y)$
  - ✎ Semi-supervised learning, kernel approaches.
  - ✎ Support vector machines (SVM).
- **Community detection**.
  - ✔ Complete study of eigenvector contents in adjacency/mpdularity methods.
  - Study of Bethe Hessian approach.
  - Analysis of non-necessarily spectral approaches (wavelet approaches).
- **Neural Networks**.
  - ✎ Analysis of non-linear extreme learning machines
  - non-linear echo-state

Short Term Objectives:

- **Robust statistics**.
  - ✎ Joint mean and covariance robust estimation
  - Study of robust regression (preliminary works exist already using strikingly different approaches)
- **Kernel methods**.
  - ✔ Subspace spectral clustering (dramatically different case of $f'(\tau) = 0$)
  - ✎ Spectral clustering with outer product kernel $f(x^\mathsf{T} y)$
  - ✎ Semi-supervised learning, kernel approaches.
  - ✎ Support vector machines (SVM).
- **Community detection**.
  - ✔ Complete study of eigenvector contents in adjacency/mpdularity methods.
  - Study of Bethe Hessian approach.
  - Analysis of non-necessarily spectral approaches (wavelet approaches).
- **Neural Networks**.
  - ✎ Analysis of non-linear extreme learning machines
  - non-linear echo-state
- **Signal processing on graphs, further graph inference, etc.**
  - Making graph methods random.

Thank you.