New methods for large scale unsupervised learning

Hafiz TIOMOKO ALI

CentraleSupélec, University of ParisSaclay, France.

September 24, 2018

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Outline

Introduction and motivation

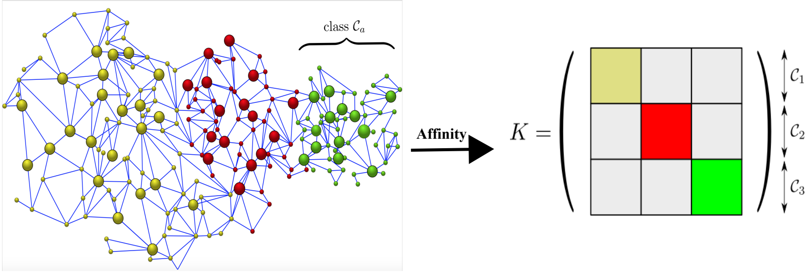
Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

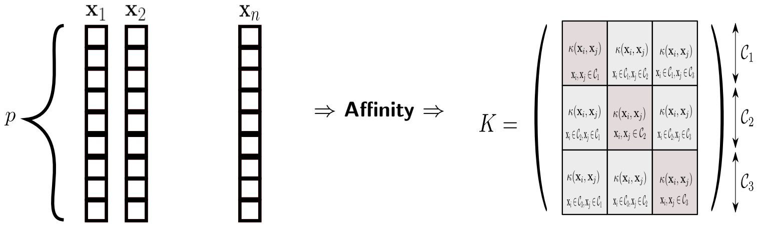
Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Motivation: Graph community detection



- Dense graph clustering on realistic block models.
- Asymptotic regime: number of nodes $n \to \infty$.
- Understanding of spectral clustering: Non-trivial behavior.

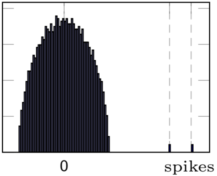
Motivation: Data clustering



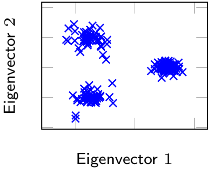
- Clustering of n data vectors of dimension p ($n, p \to \infty$).
- Affinity between vectors $\Rightarrow$ Graph clustering
- Understanding of kernel spectral methods in the big-data regime.

$$K = \begin{array}{c} \left(\begin{array}{ccc} \underset{x_i, x_j \in \mathcal{C}_1}{\kappa(\mathbf{x}_i, \mathbf{x}_j)} & \underset{x_i \in \mathcal{C}_1, x_j \in \mathcal{C}_2}{\kappa(\mathbf{x}_i, \mathbf{x}_j)} & \underset{x_i \in \mathcal{C}_1, x_j \in \mathcal{C}_3}{\kappa(\mathbf{x}_i, \mathbf{x}_j)} \\ \underset{x_i \in \mathcal{C}_2, x_j \in \mathcal{C}_1}{\kappa(\mathbf{x}_i, \mathbf{x}_j)} & \underset{x_i, x_j \in \mathcal{C}_2}{\kappa(\mathbf{x}_i, \mathbf{x}_j)} & \underset{x_i \in \mathcal{C}_2, x_j \in \mathcal{C}_3}{\kappa(\mathbf{x}_i, \mathbf{x}_j)} \\ \underset{x_i \in \mathcal{C}_3, x_j \in \mathcal{C}_1}{\kappa(\mathbf{x}_i, \mathbf{x}_j)} & \underset{x_i \in \mathcal{C}_3, x_j \in \mathcal{C}_2}{\kappa(\mathbf{x}_i, \mathbf{x}_j)} & \underset{x_i, x_j \in \mathcal{C}_3}{\kappa(\mathbf{x}_i, \mathbf{x}_j)} \end{array}\right) \end{array} \begin{array}{c} \updownarrow \mathcal{C}_1 \\ \updownarrow \mathcal{C}_2 \\ \updownarrow \mathcal{C}_3 \end{array}$$

$\Rightarrow$ **Eigenvalues** $\Rightarrow$



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



$\Downarrow$

EM or k-means clustering.

Spectral clustering: objective

State-of-the-art

- *Spectral community detection in graphs*
  - Detectability phase transition threshold in dense and sparse graph models
  - Regime of study where clustering is asymptotically perfect.
  - Studies performed mostly on (too simple) SBM models.
  - Lack of eigenvectors characterization in involved models.
- *Kernel spectral clustering*
  - Algorithms derived from ad-hoc procedures (e.g., relaxation).
  - Little understanding of performance, even for Gaussian mixtures.
  - Study of n large with p fixed.

Objectives

- Study advanced statistical models (e.g., DC-SBM)
- Study in big-data (both p and n large) and non-trivial regime of clustering.
- Benefit from concentration effect to study affinity matrices in both applications.
- Characterization of phase transition, eigenvectors content.

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Context

Baseline scenario: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ i.i.d. with $E[\mathbf{x}_1] = \mathbf{0}$, $E[\mathbf{x}_1 \mathbf{x}_1^*] = \mathbf{C}_p$:

- If $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{C}_p)$, ML estimator for $\mathbf{C}_p$ is the sample covariance matrix (SCM)

$$\hat{\mathbf{C}}_p = \frac{1}{n} \mathbf{X}^{(p)} (\mathbf{X}^{(p)})^* = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^*$$

  ($\mathbf{X}^{(p)} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$).

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

$$\hat{\mathbf{C}}_p \xrightarrow{\text{a.s.}} \mathbf{C}_p.$$

  or equivalently, in spectral norm

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

Random Matrix Regime

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

$$\left\| \hat{\mathbf{C}}_p - \mathbf{C}_p \right\| \not\to 0.$$

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

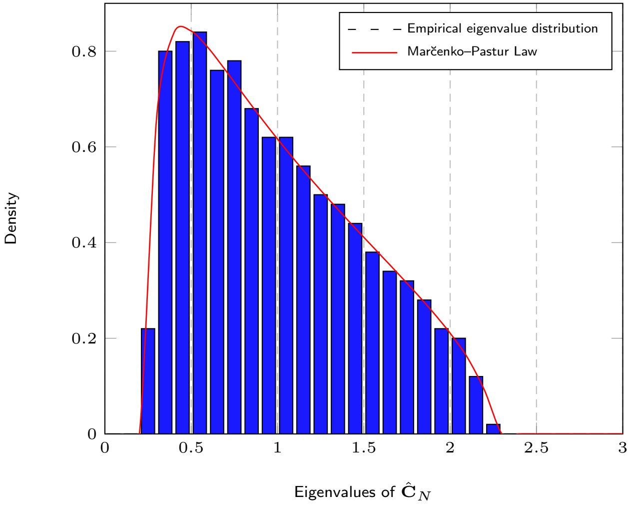


Figure: Histogram of the eigenvalues of $\hat{\mathbf{C}}_p$ for $p = 500$, $n = 2000$, $\mathbf{C}_p = \mathbf{I}_p$.

Definition (Empirical Spectral Density)

Empirical spectral density (e.s.d.) μ_n of Hermitian matrix $\mathbf{X}^{(n)} \in \mathbb{R}^{n \times n}$ is

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

Definition (Stieltjes transform)

Stieltjes transform $m_\mu(z)$ of real measurable function μ

$$m_\mu(z) = \int_{-\infty}^{\infty} \frac{1}{t - z} \mathrm{d}\mu(t).$$

for $z \in Supp(\mu)^c$.

Theorem (Inverse transformation)

If μ has a density $f_\mu(x)$ at x

$$f_\mu(x) = \frac{1}{\pi} \lim_{y \to 0^+} \mathcal{I}[m_\mu(x + iy)].$$

Theorem ([Silverstein, Bai'95])

- $\mathbf{X}^{(p)} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$, $\mathbf{x}_1$ *i.i.d. with* $E[\mathbf{x}_1] = 0$, $E[\mathbf{x}_1 \mathbf{x}_1^*] = \mathbf{C}_p$
- $\mathbf{C}_p \in \mathbb{C}^{p \times p}$ *and* $\mu_{\mathbf{C}_p} \xrightarrow{\text{a.s.}} \nu$.

As $p, n \to \infty$ *with* $p/n \to c \in (0, \infty)$, *e.s.d.* μ_p *of* $\frac{1}{n}\mathbf{X}^{(p)}(\mathbf{X}^{(p)})^*$ *satisfies*

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

weakly, where m_{μ_c} *Stieltjes transform of* μ_c *the unique solution for* $z \in \mathbb{R}$ *of*

$$m_{\mu_c}(z) = \frac{1}{c}\left(-z + c\int \frac{t}{1 + ct\left[m_{\mu_c}(z) + \frac{c-1}{zc}\right]}\mathrm{d}\nu(t)\right)^{-1} - \frac{c-1}{zc}.$$

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

$\mathbf{X}^{(p)} \in \mathbb{R}^{p \times n}$ *with i.i.d. zero mean, unit variance entries.*
As $p, n \to \infty$ *with* $p/n \to c \in (0, \infty)$*, e.s.d.* μ_p *of* $\frac{1}{n}\mathbf{X}^{(p)}(\mathbf{X}^{(p)})^*$ *satisfies*

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

weakly, where

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

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

And m_{μ_c} *Stieltjes transform of* μ_c *is given by*

$$m_{\mu_c}(z) = \frac{1}{1 - c - z - czm_{\mu_c}(z)}.$$

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Theorem (Deformed semi-circle Law [Th1.1, Girko, V.L'2001 , P.2])

$\mathbf{X}^{(n)} \in \mathbb{R}^{n\times n}$ *symmetric, with independent entries* $X_{ij}^{(n)}$, $\mathbb{E}[X_{ij}^{(n)}] = 0$,
$Var[X_{ij}^{(n)}] = \sigma_{ij}^{(n)}$.
As $n \to \infty$, *e.s.d.* μ_n *of* $n^{-\frac{1}{2}} X^{(n)}$ *satisfies*

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

weakly where

$$m_\mu(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{-z - m_i(z)}$$

$$m_i(z) = \frac{1}{n}\sum_{j=1}^{n} \frac{\sigma_{ji}^{(n)}}{-z - m_j(z)}$$

with $m_\mu(z)$ *Stieltjes transform of* μ.

Particular case: the semi-circle law

Theorem (Semi-circle Law)

$\mathbf{X}^{(n)} \in \mathbb{R}^{n\times n}$ *symmetric, with i.i.d entries* $X_{ij}^{(n)}$, $\mathbb{E}[X_{ij}^{(n)}] = 0$, $Var[X_{ij}^{(n)}] = 1$.
As $n \to \infty$, *e.s.d.* μ_n *of* $n^{-\frac{1}{2}}\mathbf{X}^{(n)}$ *satisfy*

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

weakly where μ *has a density* f *supported on* $[-2, 2]$ *and defined as*

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

And m_μ *Stieltjes transform of* μ *is given by*

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

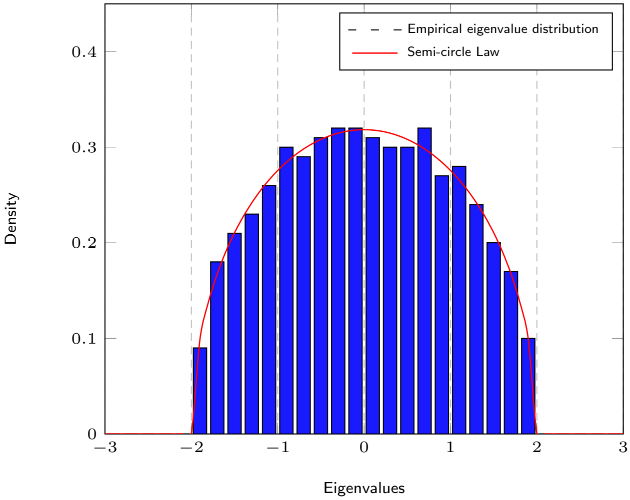


Figure: Histogram of the eigenvalues of Wigner matrices and the semi-circle law, for $n = 500$.

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Theorem (Eigenvalues)

- *Let* $\mathbf{Y}^{(n)} = \frac{\mathbf{X}^{(n)}}{\sqrt{n}} + \sum_{i=1}^{k} w_i \mathbf{v}_i \mathbf{v}_i^{\mathsf{T}}$ *with ordered eigenvalues* $\lambda_1(\mathbf{Y}^{(n)}) \geq \cdots \geq \lambda_n(\mathbf{Y}^{(n)})$ *and* $w_1 \geq \cdots \geq w_k$.
- $\mathbf{X}^{(n)} \in \mathbb{R}^{n \times n}$ *symmetric, with i.i.d entries* $X_{ij}^{(n)}$, $\mathbb{E}[X_{ij}^{(n)}] = 0$, $Var[X_{ij}^{(n)}] = 1$.
- $\mu_n \xrightarrow{\text{a.s.}} \mu$ *with support* $[-2, 2]$, *and* $m(z)$ *Stieltjes transform of* μ.

As $n \to \infty$, *for* $i = 1, \cdots, k$

- *If* $|\omega_i| > \lim_{z \downarrow 2} -\frac{1}{m(z)} = 1$

$$\lambda_i(\mathbf{Y}^{(n)}) \xrightarrow{\text{a.s.}} m^{-1}\left(-\frac{1}{\omega_i}\right) = \frac{1 + \omega_i^2}{\omega_i} > 2.$$

- *Otherwise*

$$\lambda_i(\mathbf{Y}^{(n)}) \xrightarrow{\text{a.s.}} 2.$$

Theorem (Eigenvectors)

- *Let* $\mathbf{Y}^{(n)} = \frac{\mathbf{X}^{(n)}}{\sqrt{n}} + \sum_{i=1}^{k} w_i \mathbf{v}_i \mathbf{v}_i^\mathsf{T}$ *with ordered eigenvalues* $\lambda_1(\mathbf{Y}^{(n)}) \geq \cdots \geq \lambda_n(\mathbf{Y}^{(n)})$ *and* $w_1 \geq \cdots \geq w_k$*.*
- $\mathbf{X}^{(n)} \in \mathbb{R}^{n\times n}$ *symmetric, with i.i.d entries* $X_{ij}^{(n)}$, $\mathbb{E}[X_{ij}^{(n)}] = 0$, $Var[X_{ij}^{(n)}] = 1$.
- $\mu_n \xrightarrow{\text{a.s.}} \mu$ *with support* $[-2, 2]$*, and* $m(z)$ *Stieltjes transform of* μ*.*

As $n \to \infty$*, for* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ *deterministic vectors and* $\mathbf{u}_i$ *eigenvector of* $\mathbf{Y}^{(n)}$ *associated with eigenvalue* $\lambda_i(\mathbf{Y}^{(n)})$*,*

$$\mathbf{a}^\mathsf{T}\mathbf{u}_i\mathbf{u}_i^\mathsf{T}\mathbf{b} - \frac{\omega_i^2 - 1}{\omega_i^2}\mathbf{a}^\mathsf{T}\mathbf{v}_i\mathbf{v}_i^\mathsf{T}\mathbf{b} \cdot 1_{|\omega_i|>1} \xrightarrow{\text{a.s.}} 0.$$

In particular

$$\left|\mathbf{v}_i^\mathsf{T}\mathbf{u}_i\right|^2 \xrightarrow{\text{a.s.}} \frac{\omega_i^2 - 1}{\omega_i^2} \cdot 1_{|\omega_i|>1}.$$

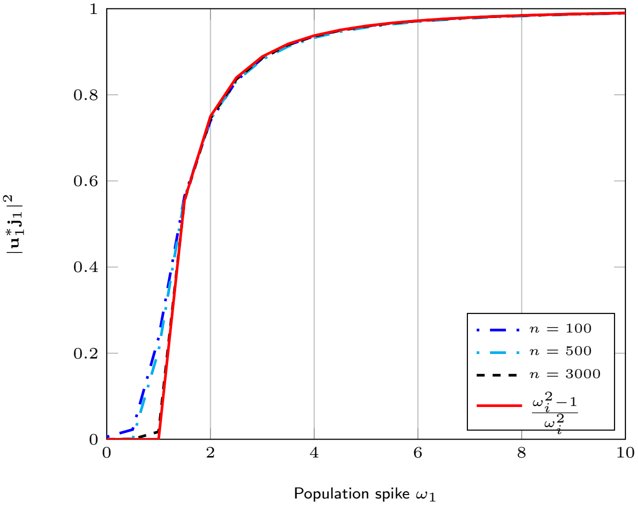


Figure: Simulated versus limiting $|\mathbf{u}_1^\mathsf{T} \mathbf{j}_1|^2$ for $\mathbf{Y} = \mathbf{X} + \omega_1 \mathbf{j}_1 \mathbf{j}_1^\mathsf{T}$, $\mathbf{j}_1 = \frac{2}{\sqrt{n}}[\mathbf{1}_{n/2}, \mathbf{0}_{n/2}]$, $X_{ij} \sim \mathcal{N}(0, 1/n)$, varying ω_1.

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Outline

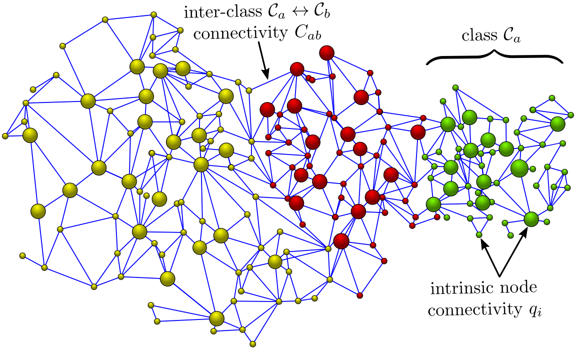
Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions



Undirected graph with n nodes, m edges:

- "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 $\mathcal{C}_b$
- edge probability for nodes $i \in \mathcal{C}_{g_i}, j \in \mathcal{C}_{g_j}$:

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

- adjacency matrix $\mathbf{A}$ with

$$\boxed{A_{ij} \sim \text{Bernoulli}(q_i q_j C_{g_i g_j})}$$

Affinity matrices used for spectral methods

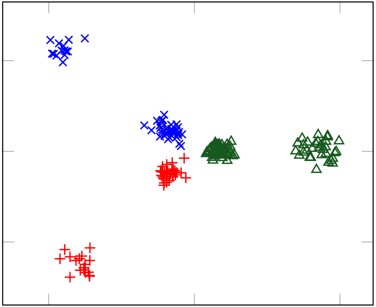
Dense graphs:

- Adjacency: $\mathbf{A}$.
- Modularity: $\mathbf{A} - \mathbf{d}\mathbf{d}^\mathsf{T}$, $\mathbf{d} = \mathbf{A}\mathbf{1}$.
- Laplacian: $\mathcal{D}(\mathbf{d})^{-\frac{1}{2}}\mathbf{A}\mathcal{D}(\mathbf{d})^{-\frac{1}{2}}$.

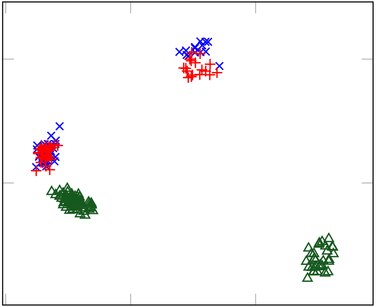
Sparse graphs:

- Non-backtracking matrix (affinity between edges).
- Bethe-Hessian matrix: $\mathcal{D}(\mathbf{d}) - \mathbf{A}$.

- 3 classes with μ bi-modal ($\mu = \frac{3}{4}\boldsymbol{\delta}_{0.1} + \frac{1}{4}\boldsymbol{\delta}_{0.5}$)



(Modularity $\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{2m}$)



(Bethe Hessian $\mathbf{D} - r\mathbf{A}$)

Proposed Regularized Modularity Approach

Recall: $P(i \sim j) = q_i q_j C_{g_i g_j}$.

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

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

Community information is **weak but highly redundant**

Considered Matrix:

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

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Theorem (Limiting Random Matrix Equivalent)

As $n \to \infty$, $\|\mathbf{L}_\alpha - \tilde{\mathbf{L}}_\alpha\| \xrightarrow{\text{a.s.}} 0$, *where*

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

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

with $\mathbf{D}_q = \operatorname{diag}(\{q_i\})$, $\mathbf{X}$ *zero-mean random matrix with variance profile,*

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

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

and $\mathbf{J} = [\mathbf{j}_1, \ldots, \mathbf{j}_k]$, $\mathbf{j}_a = [0, \ldots, 0, \mathbf{1}_{n_a}^\mathsf{T}, 0, \ldots, 0]^\mathsf{T} \in \mathbb{R}^n$, $\mathbf{c} = \{c_a\}_{a=1}^k$, $c_a = n_a / n$.

Consequences:

- isolated eigenvalues beyond phase transition $\Leftrightarrow$ $\lambda(\mathbf{M}) >$ "spectrum edge"

  Optimal choice α_{opt} of α from study of limiting spectrum.

- **eigenvectors correlated to** $\mathbf{D}_q^{1-\alpha} \mathbf{J}$

  Necessary regularization by $\mathbf{D}^{\alpha-1}$.

Eigenvalue Spectrum

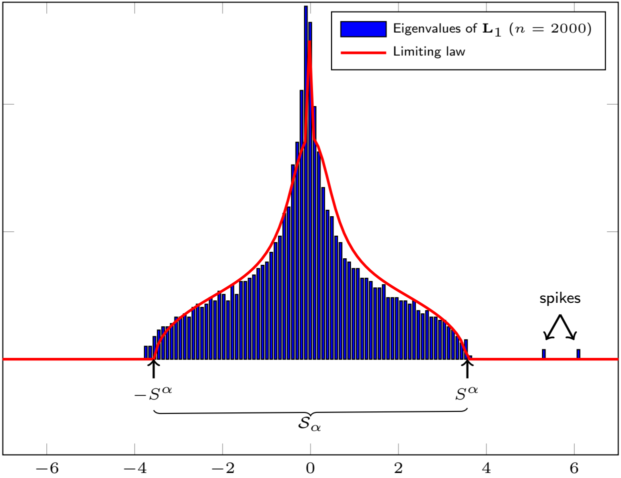


Figure: 3 classes, $c_1 = c_2 = 0.3, c_3 = 0.4$, $\mu = \frac{1}{2}\boldsymbol{\delta}_{0.4} + \frac{1}{2}\boldsymbol{\delta}_{0.9}$, $\mathbf{M} = 4\begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}$.

Theorem (Phase Transition)

Isolated eigenvalue $\lambda_i(\mathbf{L}_\alpha)$ *if* $|\lambda_i(\bar{\mathbf{M}})| > \tau^\alpha$, $\bar{\mathbf{M}} = (\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}$, *where*

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

with $[S_-^\alpha, S_+^\alpha]$ *limiting eigenvalue support of* $\mathbf{L}_\alpha$ *and* $g^\alpha(x)$ *(*$|x| > S_+^\alpha$*) solution of*

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

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

In this case, $\lambda_i(\mathbf{L}_\alpha) \xrightarrow{\text{a.s.}} (g^\alpha)^{-1}\left(-1/\lambda_i(\bar{\mathbf{M}})\right)$.

Clustering possible when $\lambda_i(\bar{\mathbf{M}}) > (\min_\alpha \tau_\alpha)$:

- **"Optimal"** $\alpha_{\text{opt}} \equiv \operatorname{argmin}_\alpha \{\tau_\alpha\}$.
- From $\hat{q}_i \equiv \frac{\mathbf{d}_i}{\sqrt{\mathbf{d}^\mathsf{T} \mathbf{1}_n}} \xrightarrow{\text{a.s.}} q_i$, $\mu \simeq \hat{\mu} \equiv \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{\hat{q}_i}$ and thus:

$$\boxed{\text{Consistent estimator } \hat{\alpha}_{\text{opt}} \text{ of } \alpha_{\text{opt}}.}$$

Outline

Introduction and motivation

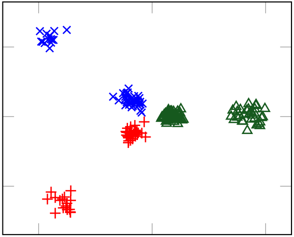
Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

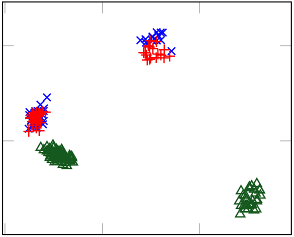
Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

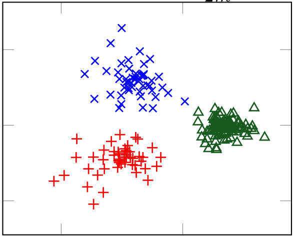
Simulated Performance Results (2 masses of q_i)



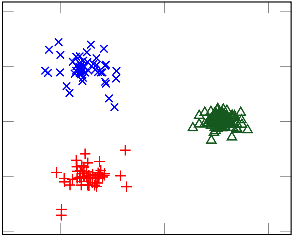
(Modularity $\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{2m}$)



(Bethe Hessian $\mathbf{D} - r\mathbf{A}$)



(Proposed, $\alpha = 1$)



(Proposed, $\hat{\alpha}_{\rm opt}$)

Figure: 3 classes, $\mu = \frac{3}{4}\boldsymbol{\delta}_{0.1} + \frac{1}{4}\boldsymbol{\delta}_{0.5}$, $c_1 = c_2 = \frac{1}{4}$, $c_3 = \frac{1}{2}$, $\mathbf{M} = 100\mathbf{I}_3$.

Simulated Performance Results (2 masses for q_i)

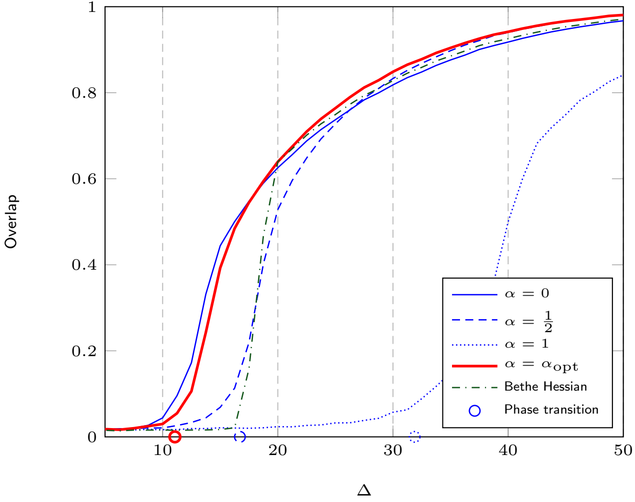


Figure: Overlap performance for $n = 3000$, $k = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, $\mathbf{M} = \Delta \mathbf{I}_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\text{opt}} = 0.07$.

Simulated Performance Results (2 masses for q_i)

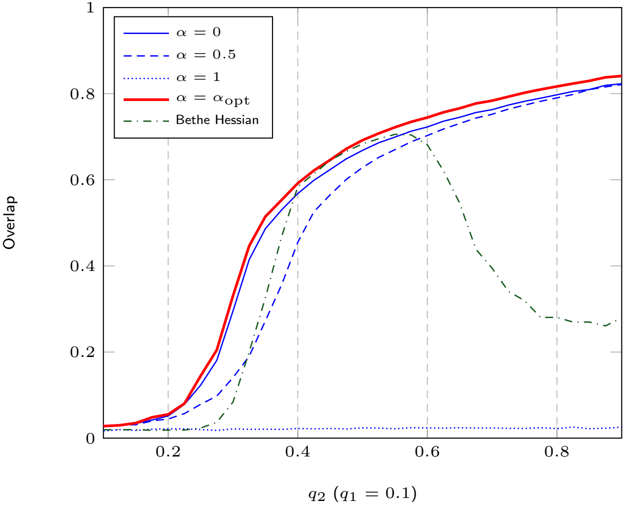
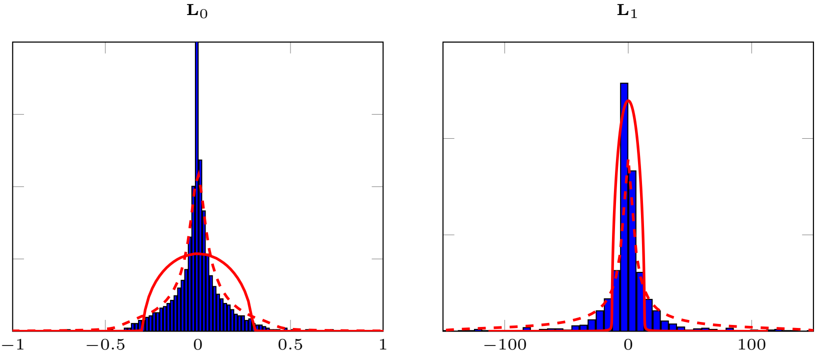


Figure: Overlap performance for $n = 3000$, $k = 3$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} \in [0.1, 0.9]$, $\mathbf{M} = 10(2\mathbf{I}_3 - \mathbf{1}_3\mathbf{1}_3^{\mathsf{T}})$, $c_i = \frac{1}{3}$.

Real Graph Example: PolBlogs ($n = 1490$, two classes)



($\lambda_{\max} \simeq 1.75$) ($\lambda_{\max} \simeq 483$)

Algorithms	Overlap	Modularity
α_{opt} ($\simeq 0$)	$\mathbf{0.897}$	$\mathbf{0.4246}$
$\alpha = 0.5$	0.035	$\simeq 0$
$\alpha = 1$	0.040	$\simeq 0$
BH	0.304	0.2723

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Setting and Basic Assumptions

Data: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, with $\begin{cases} \mathbf{x}_1, \ldots, \mathbf{x}_{n_1} & \in \mathcal{C}_1 \\ \ldots & \ldots \\ \mathbf{x}_{n-n_k+1}, \ldots, \mathbf{x}_n & \in \mathcal{C}_k. \end{cases}$

Class definition:

$$\mathbf{x}_i \in \mathcal{C}_a \Leftrightarrow \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a).$$

Growth rates: As $n \to \infty$, k remains fixed, and

$$p/n \to c_0 > 0, \quad n_a/n \to c_a > 0.$$

Growing $p \Rightarrow$ Control of $\boldsymbol{\mu}_a$, $\mathbf{C}_a$ to avoid trivial solutions.

Neyman-Pearson optimal rates:
Knowing $\boldsymbol{\mu}$ and ϵ,

- For $\mathbf{x} \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I})$, can decide on classes when $\|\boldsymbol{\mu}\| \geq \mathcal{O}(1)$.
- For $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I} \pm \mathbf{E})$, can decide on classes when $\|\mathbf{E}\| \geq \mathcal{O}(p^{-\frac{1}{2}})$.

Neyman-Pearson separability rates

As $p \to \infty$, for all $a, b \in \{1, \ldots, k\}$, when $\boldsymbol{\mu}_a$, $\mathbf{C}_a$ known,

- $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = O(1)$
- $\|\mathbf{C}_a\|$ bounded, $|\operatorname{tr}(\mathbf{C}_a - \mathbf{C}_b)| = O(\sqrt{p})$, $\boxed{\operatorname{tr}((\mathbf{C}_a - \mathbf{C}_b)^2) = O(1)}$.

(Inner Product) Kernel Matrix

Object of interest: With $\mathbf{x}_i^\circ = \mathbf{x}_i - \frac{1}{n}\sum_{j=1}^n \mathbf{x}_j$, define

$$\mathbf{K} = \mathbf{P}\left\{ f\left(\frac{1}{p}(\mathbf{x}_i^\circ)^\mathsf{T}\mathbf{x}_j^\circ\right) 1_{i\neq j} \right\}\mathbf{P}$$

where $\mathbf{P} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ and f three-times differentiable around 0.

($f(\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ could be treated similarly)

Objective: Study:

- limiting spectrum of $\mathbf{K}$ (eigenvalues + eigenvectors)
- clustering performances.

Previous work: Kernel spectral clustering ($f'(0) \neq 0$)

Assumption (Separability Rate)

As $p \to \infty$, for all $a, b \in \{1, \ldots, k\}$,

- $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = O(1)$
- $\|\mathbf{C}_a\|$ *bounded,* $|tr(\mathbf{C}_a - \mathbf{C}_b)| = O(\sqrt{p})$, $\boxed{tr((\mathbf{C}_a - \mathbf{C}_b)^2) = O(p)}$.
- f *such that* $f(0) = \mathcal{O}(1)$, $f'(0) = \mathcal{O}(1)$, $f''(0) = \mathcal{O}(1)$.

Key Result: As $n, p \to \infty$, and assumption above,

- for all $i \neq j$, irrespective of the class,

$$\boxed{\frac{1}{p}(\mathbf{x}_i^\circ)^\mathsf{T}\mathbf{x}_j^\circ \to 0, \quad \frac{1}{2p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 \to \tau.}$$

- counter-intuitive curse of dimensionality: **all vectors are far**.
- but allows for Taylor-expansion of K_{ij} around $f(0)$:

$$K_{ij} = f(0) + f'(0)\left[\frac{1}{p}\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b + \frac{1}{p}\mathbf{w}_i^\mathsf{T}\mathbf{w}_j + \ldots\right] + \frac{1}{2}f''(0)\left[\frac{1}{p^2}\operatorname{tr}(\mathbf{C}_a - \mathbf{C}_b)^2\right] + o(1)$$

(for $x_i = \boldsymbol{\mu}_a + \mathbf{w}_i$, $x_j = \boldsymbol{\mu}_b + \mathbf{w}_j$)

- **Model type**: Marčenko-pastur+ Spikes.

Previous work: Kernel spectral clustering $f'(0) \neq 0$

Assumption (Separability Rate)

As $p \to \infty$, *for all* $a, b \in \{1, \ldots, k\}$,

- $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = O(1)$
- $\|\mathbf{C}_a\|$ *bounded*, $|\mathit{tr}(\mathbf{C}_a - \mathbf{C}_b)| = O(\sqrt{p})$, $\boxed{\mathit{tr}((\mathbf{C}_a - \mathbf{C}_b)^2) = O(p)}$.
- f *such that* $f(0) = \mathcal{O}(1)$, $f'(0) = \mathcal{O}(1)$, $f''(0) = \mathcal{O}(1)$.

Conclusions:

- limiting e.s.d of kernel: Marčenko-pastur law.
- Can do better on *class covariance* rates.
- $f'(0) = 0$: **asymptotic trivial clustering** $\Rightarrow$ Can improve growth rates for covariances.

Assumption (Separability Rate)

As $p \to \infty$, for all $a, b \in \{1, \dots, k\}$,

- $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = O(1)$
- $\|\mathbf{C}_a\|$ *bounded,* $|\mathrm{tr}(\mathbf{C}_a - \mathbf{C}_b)| = O(\sqrt{p})$, $\boxed{\mathrm{tr}((\mathbf{C}_a - \mathbf{C}_b)^2) = O(\sqrt{p})}$.
- f *such that* $f(0) = \mathcal{O}(1)$, $f'(0) = 0$, $f''(0) = \mathcal{O}(1)$.

$$\sqrt{p} K_{ij} = f(0) + \frac{f''(0)}{2} \left[\frac{1}{p\sqrt{p}} (\mathbf{w}_i^{\mathsf{T}} \mathbf{w}_j)^2 + \frac{1}{p\sqrt{p}} \mathrm{tr}\, (\mathbf{C}_a - \mathbf{C}_b)^2 \right] + o(1)$$

Conclusions:

- **Model type**: Semi-circle + spikes.
- Better rates in class-covariance.
- Sub-optimal in class means discrimination (means discarded).

Assumption (Separability Rate)

As $p \to \infty$, for all $a, b \in \{1, \ldots, k\}$,

- $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = O(1)$
- $\|\mathbf{C}_a\|$ *bounded,* $|tr(\mathbf{C}_a - \mathbf{C}_b)| = O(\sqrt{p})$, $\boxed{tr((\mathbf{C}_a - \mathbf{C}_b)^2) = O(\sqrt{p})}$.
- $f'(0) = \frac{\alpha}{\sqrt{p}}$, $\frac{1}{2}f''(0) = \beta$.

Taylor approximation:

$$\sqrt{p}K_{ij} = f(0) + \alpha\left[\frac{1}{p}\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b + \frac{1}{p}\mathbf{w}_i^\mathsf{T}\mathbf{w}_j + \ldots\right] + \beta\left[\frac{1}{p\sqrt{p}}\mathrm{tr}\,(\mathbf{C}_a - \mathbf{C}_b)^2 + \frac{1}{p\sqrt{p}}(\mathbf{w}_i^\mathsf{T}\mathbf{w}_j)^2\right] + o(1)$$

(for $x_i = \boldsymbol{\mu}_a + \mathbf{w}_i$, $x_j = \boldsymbol{\mu}_b + \mathbf{w}_j$)

Findings:

- **Model type**: Marcenko-pastur+Semi-circle+spikes.
- Balance between class means and covariances.

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Theorem (Asymptotic Equivalent for $\mathbf{K}$)

For f such that $f'(0) = \frac{\alpha}{\sqrt{p}}$, $\frac{1}{2}f''(0) = \beta$, as $n, p \to \infty$,

$$\|\mathbf{K} - \hat{\mathbf{K}}\| \xrightarrow{\text{a.s.}} 0$$

where

$$\boxed{\sqrt{p}\hat{\mathbf{K}} \equiv \alpha \mathbf{P}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{P} + \beta \mathbf{P}\Phi\mathbf{P} + \mathbf{U}\mathbf{A}\mathbf{U}^{\mathsf{T}} - (f(0) + \tau f'(0))\mathbf{P}}$$

with

$$\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n], \quad \Phi_{ij} = \sqrt{p}\left[((\mathbf{w}_i)^{\mathsf{T}}\mathbf{w}_j)^2 - \frac{1}{p^2} tr\, \mathbf{C}_a \mathbf{C}_b\right] 1_{i \neq j}$$

$$\mathbf{U} = \left[\frac{\mathbf{J}}{\sqrt{p}}, \mathbf{P}\mathbf{W}^{\mathsf{T}}\mathbf{M}\right], \quad \mathbf{J} = [\mathbf{j}_1, \ldots, \mathbf{j}_k], \quad \mathbf{j}_a = (0, \ldots, \mathbf{1}_{n_a}, \ldots, 0)^{\mathsf{T}}$$

$$\mathbf{A} = \begin{bmatrix} \alpha \mathbf{M}^{\mathsf{T}}\mathbf{M} + \beta \mathbf{T} & \alpha \mathbf{I}_k \\ \alpha \mathbf{I}_k & 0 \end{bmatrix}, \quad \mathbf{M} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k], \quad \mathbf{T} = \frac{1}{p\sqrt{p}} \left\{ tr (\mathbf{C}_a - \mathbf{C}_b)^2 \right\}.$$

Role of α, β:

- Weighs Marčenko–Pastur versus semi-circle parts.
- Trade-off between means and covariance discrimination.

Theorem (Limiting Eigenvalue Distribution)

As $n, p \to \infty$,

$$\mu_n \equiv \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{K})} \xrightarrow{\mathcal{L}} \mu$$

with μ (having compact support $\mathcal{S}$) given by its Stieltjes transform $m(z) = \int \frac{\mu(d\lambda)}{\lambda - z}$, unique solution of

$$\frac{1}{m(z)} = -z + \frac{\alpha}{p} \operatorname{tr} \mathbf{C}^{\circ} \left(\mathbf{I}_p + \frac{\alpha m(z)}{c_0} \mathbf{C}^{\circ} \right)^{-1} - \frac{2\beta^2}{c_0} m(z) \left(\frac{1}{p} \operatorname{tr} (\mathbf{C}^{\circ})^2 \right)^2 .$$

where

$$\mathbf{C}^{\circ} \triangleq \sum_{a=1}^{k} c_a \mathbf{C}_a .$$

Mixed Marcenko–Pastur & Wigner Spectrum

- $\mathbf{PW}^{\mathsf{T}}\mathbf{WP}$: Marcenko–Pastur like spectrum
- $\mathbf{P\Phi P}$: semi-circle (Wigner) like spectrum
- $\mathbf{UAU}^{\mathsf{T}}$: produces spikes under phase transition!

$$\text{Here for } f(x) = \frac{1}{2}\beta\left(x + \frac{1}{\sqrt{p}}\frac{\alpha}{\beta}\right)^2,$$

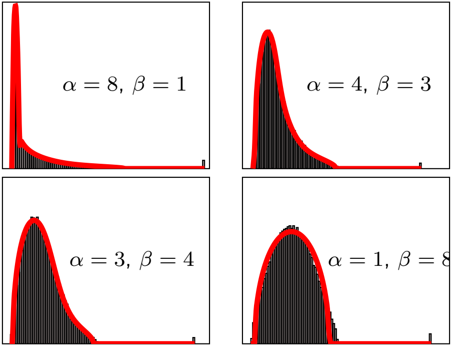


Figure: Eigenvalues of $\mathbf{K}$ versus limiting law, $p = 2048$, $n = 4096$, $k = 2$, $n_1 = n_2$, $\boldsymbol{\mu}_i = 3\boldsymbol{\delta}_i$.

Theorem

Let $\rho \in \mathbb{R} \setminus \mathcal{S}$ *be such that*

$$\frac{m(\rho)}{4c_0}(\alpha\delta g(\rho) + \beta\theta) + 1 = 0$$

with

$$g(\rho) = \frac{1}{p} tr(\mathbf{I}_p + \frac{\alpha m(\rho)}{c_0}\mathbf{C}^\circ)^{-1}$$
$$\delta = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$$
$$\theta = \frac{1}{\sqrt{p}} tr(\mathbf{C}_1 - \mathbf{C}_2)^2.$$

Then, there exists λ_j *eigenvalue of* $\hat{\mathbf{K}}$ *such that*

$$|\lambda_j - \rho| \xrightarrow{\text{a.s.}} 0.$$

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Datasets	$\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$	$\frac{1}{\sqrt{p}} \operatorname{tr} (\mathbf{C}_1 - \mathbf{C}_2)^2$	Ratio
MNIST (digits $1, 7$)	612	1990	3.3
MNIST (digits $3, 6$)	441	1119	2.5
MNIST (digits $3, 8$)	212	652	3.0
EEG (sets A, E)	2.4	109	45.4

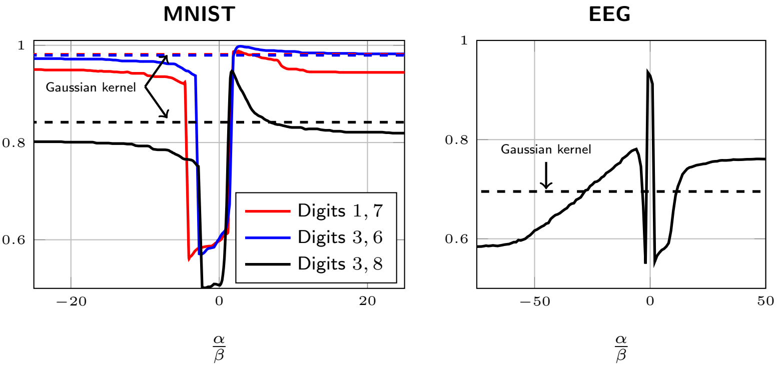


Figure: Spectral clustering accuracy for MNIST and EEG, versus Gaussian kernel ($K_{ij} = e^{-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2}$).

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Conclusions

Spectral community detection in dense and heterogeneous graphs:

- In heterogeneous dense graph models, adjacency, Laplacian not always optimal even with proper eigenvector normalization.
- We found a matrix with better phase transition in challenging cases.
- We characterize content of the eigenvectors $\Rightarrow$ Improved version of E.M algorithm.

Kernel spectral clustering:

- Original intuitions of ML algorithms in small dimensions most often no longer valid in high dimensions.
- Under non-trivial regime, concentration of key quantities allows for understanding of kernel matrices.
- We conciliate previous findings to propose new kernel design with better means and covariances discriminative rates.

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Perspectives

Spectral Community detection.

- ✎ Study of Fisher-score matrix for spectral clustering.
- Sparse DCSBM
- Complexity/Performance tradeoff of spectral methods.

Kernel spectral clustering.

- ✎ Off-line estimation of kernel (α, β).
- Study of kernel spectral clustering under other statistical models (heavy tails,. . .).

Use of the technical developed tools in the understanding of performances of other machine learning algorithms beyond spectral ones?

Outline

Introduction and motivation

Basics of Random Matrix Theory
- Large Sample Covariance Matrices
- Semi-circle law
- Spiked models

Community detection in graphs
- Motivations and model
- Main results
- Simulations

Kernel spectral clustering
- Model and assumptions
- Main results
- Applications

Conclusions and perspectives
- Conclusions
- Perspectives
- Contributions

Contributions I

Journals (2 published)

Tiomoko Ali, H. and Couillet, R. Improved spectral community detection in large heterogeneous networks. Journal of Machine Learning Research, 18:149.

Couillet, R., Wainrib, G., Sevi, H., and Tiomoko Ali, H. The asymptotic performance of linear echo state neural networks. Journal of Machine Learning Research, 17(178):135.

Conferences (6 published, 1 submitted)

Tiomoko Ali, H. and Couillet, R. Performance analysis of spectral community detection in realistic graph models. In ICASSP16.

Tiomoko Ali, H. and Couillet, R. community detection in heterogeneous networks. In Signals, Systems and Computers, 2016 50th Asilomar Conference.

Tiomoko Ali, H., Kammoun, A., and Couillet, R. Random matrix asymptotic of inner product kernel spectral clustering. In ICASSP18.

Tiomoko Ali, H., Kammoun, A., and Couillet, R. Random matrix-improved kernels for large dimensional spectral clustering. In Statistical Signal Processing Workshop (SSP), 2018.

Couillet, R., Wainrib, G., Sevi, H., and Tiomoko Ali, H. Training performance of echo state neural networks. In Statistical Signal Processing Workshop (SSP), 2016.

Contributions II

Couillet, R., Wainrib, G., Tiomoko Ali, H., and Sevi, H. A random matrix approach to echo-state neural networks. In International Conference on Machine Learning (ICML 2016).

Ali, H. T., Liu, S., Yilmaz, Y., Hero, A., Couillet, R., and Rajapakse, I. Latent heterogeneous multilayer community detection.

Thank you.