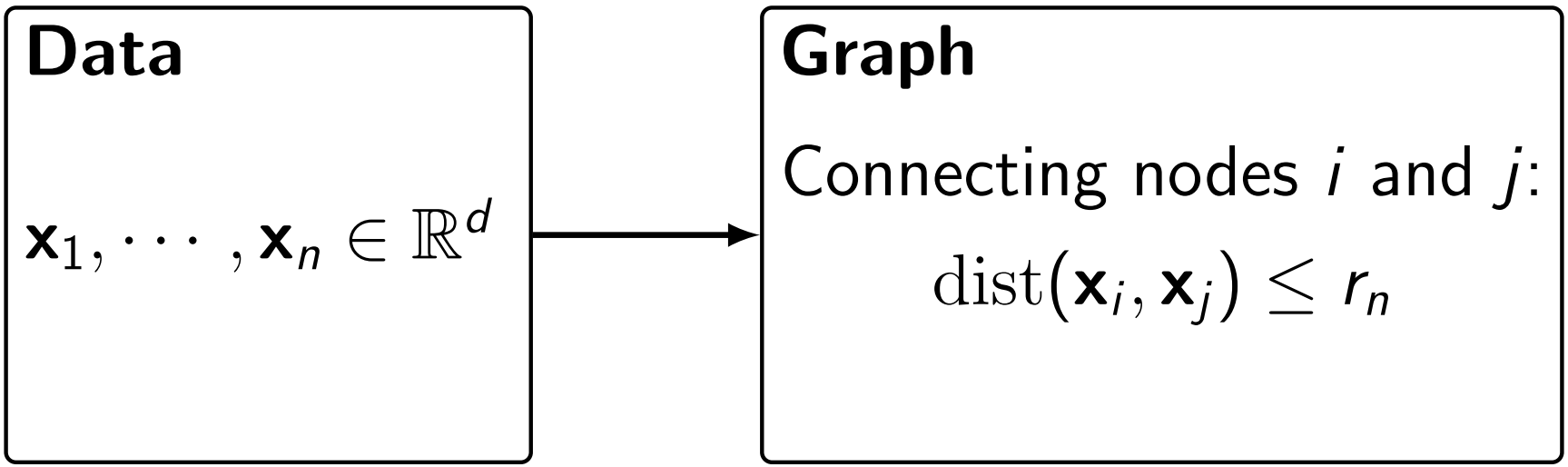
Recent Developments in Random Geometric Graphs and Their Applications

Xiucai Ding

Department of Statistics
University of California, Davis

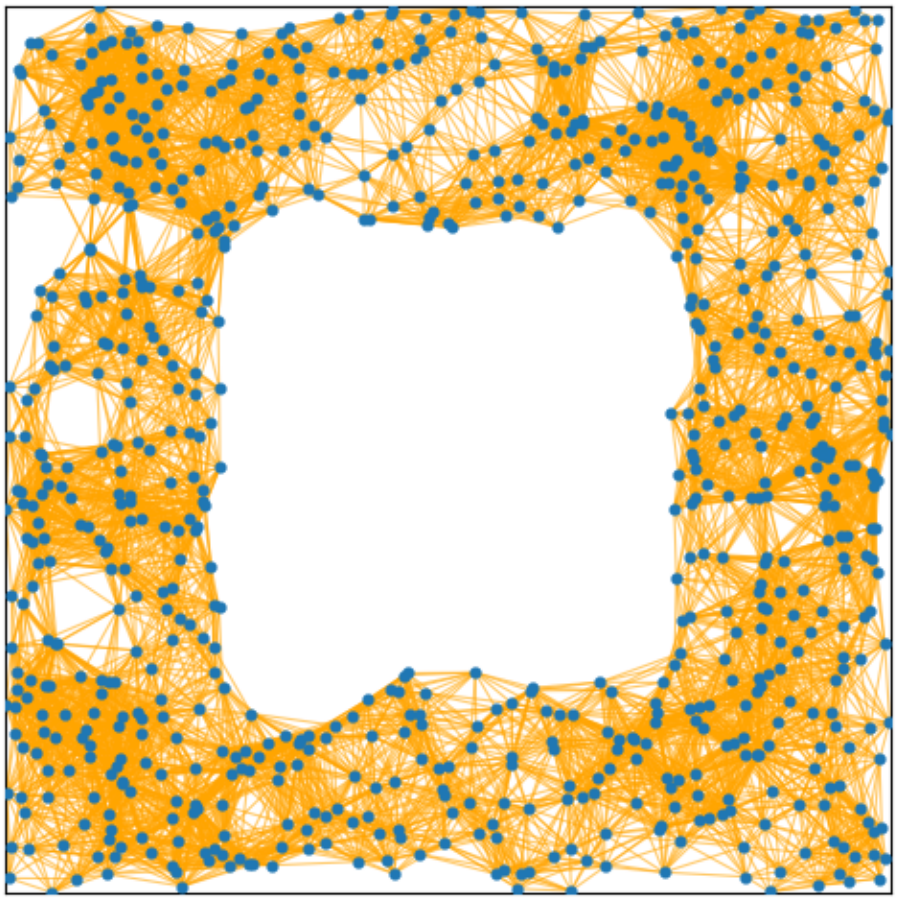
Joint work with Yichen Hu (UC Davis)/Haixiao Wang (UW–Madison)

From Data to RGG

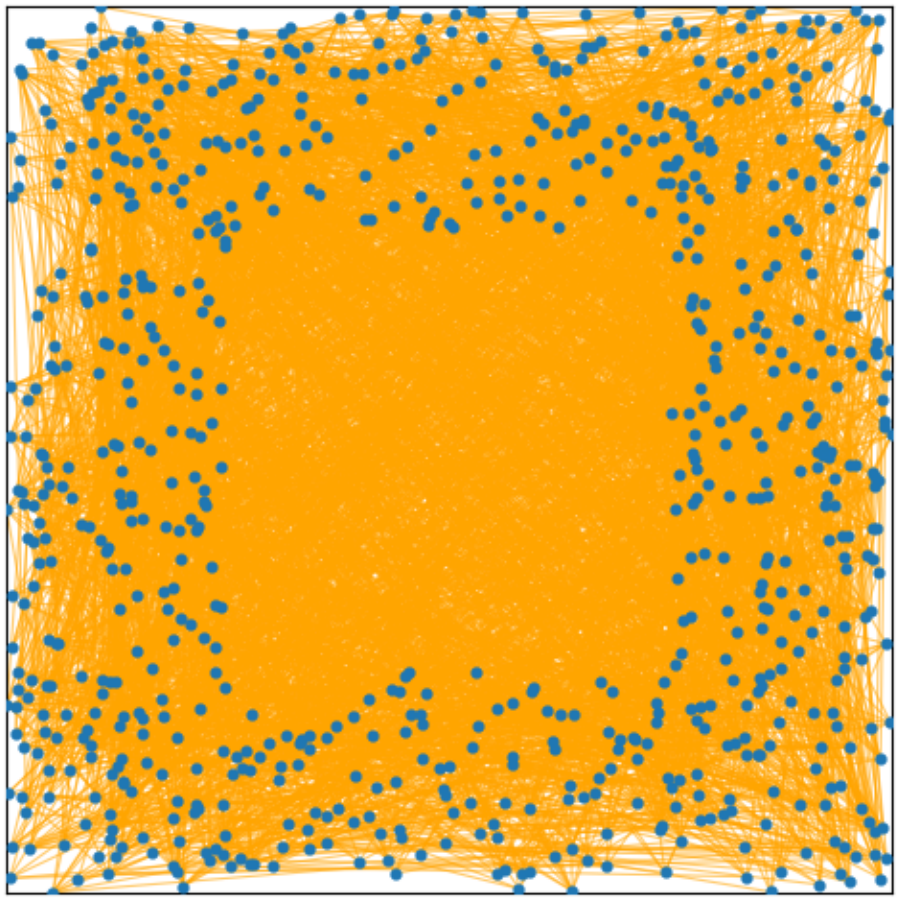


Example

$\mathbf{x}_1, \cdots, \mathbf{x}_n$ be i.i.d. uniform on $[0,1]^2 \setminus [1/4, 3/4]^2$



(a) RGG



(b) ERG

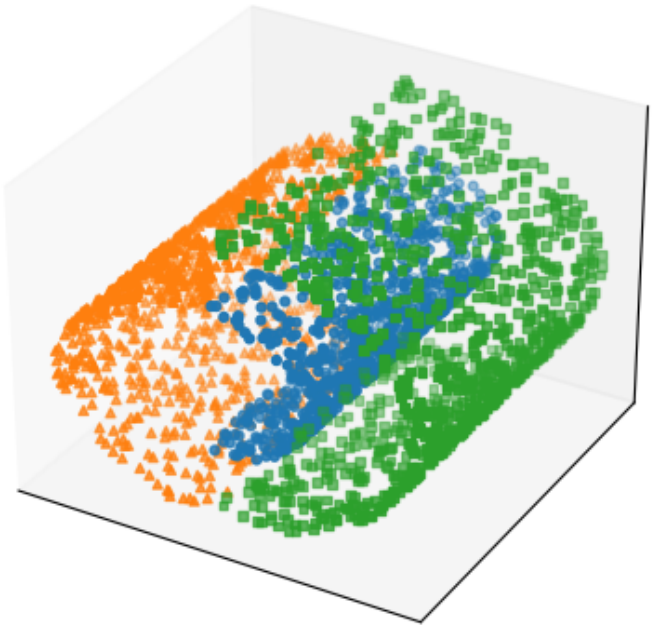
Application: manifold learning

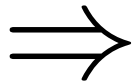
Most manifold learning algorithms are based on graph structures, e.g., Laplacian eigenmap. It can be summarized in the following framework:

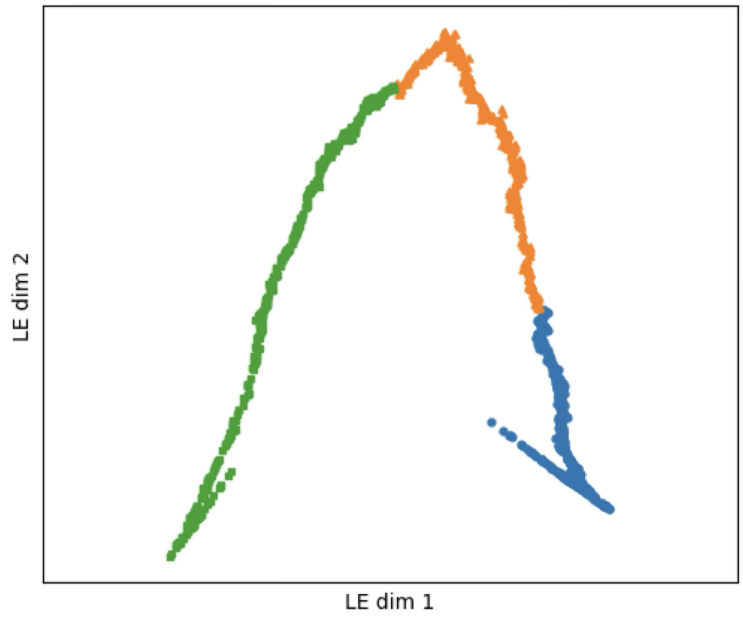
Algorithm

1: Input: nodes $1, \cdots, n$ associated with features $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$
2: Connect node i and node j if $\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) \leq r_n$.
3: Assign a weight $g(\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j)/r_n)$ if node i and node j are connected.
4: Construct affinity matrix $\mathbf{K}$ and degree matrix $\mathbf{D}$.
5: Compute random-walk Laplacian matrix $\mathbf{L}_{rw} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{K}$ and its eigenvalues $0 \leq \lambda_1 \leq \cdots \leq \lambda_n$ with eigenvectors $\nu_1, \cdots, \nu_n$.
6: Embed $\mathbf{x}_j \longrightarrow (\nu_{2,j}, \cdots, \nu_{M+1,j})^\top$.

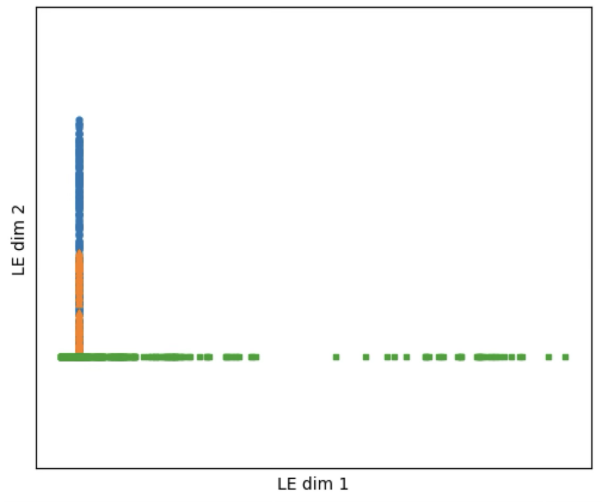
This uses the spectrum of an RGG via $\mathbf{L}_{rw}$.



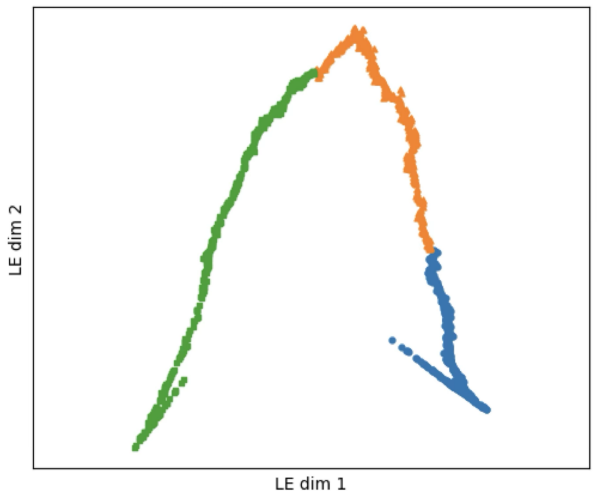



LE dim 2
LE dim 1

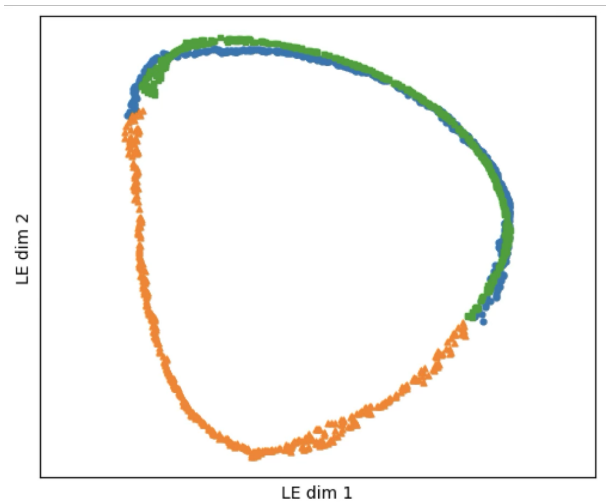
Impact of r_n



(a) $r_n = 0.05$

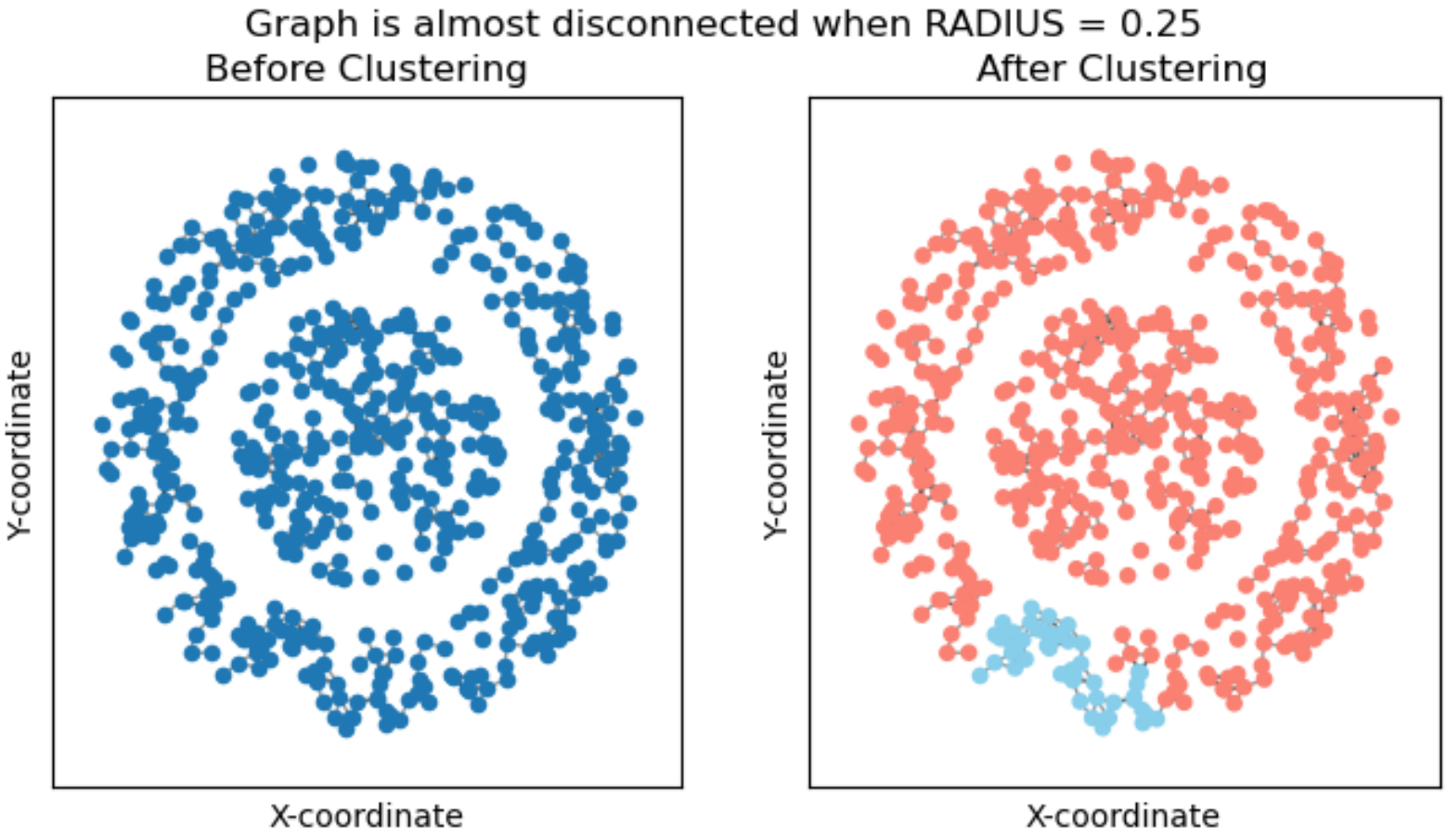


(b) $r_n = 0.2$

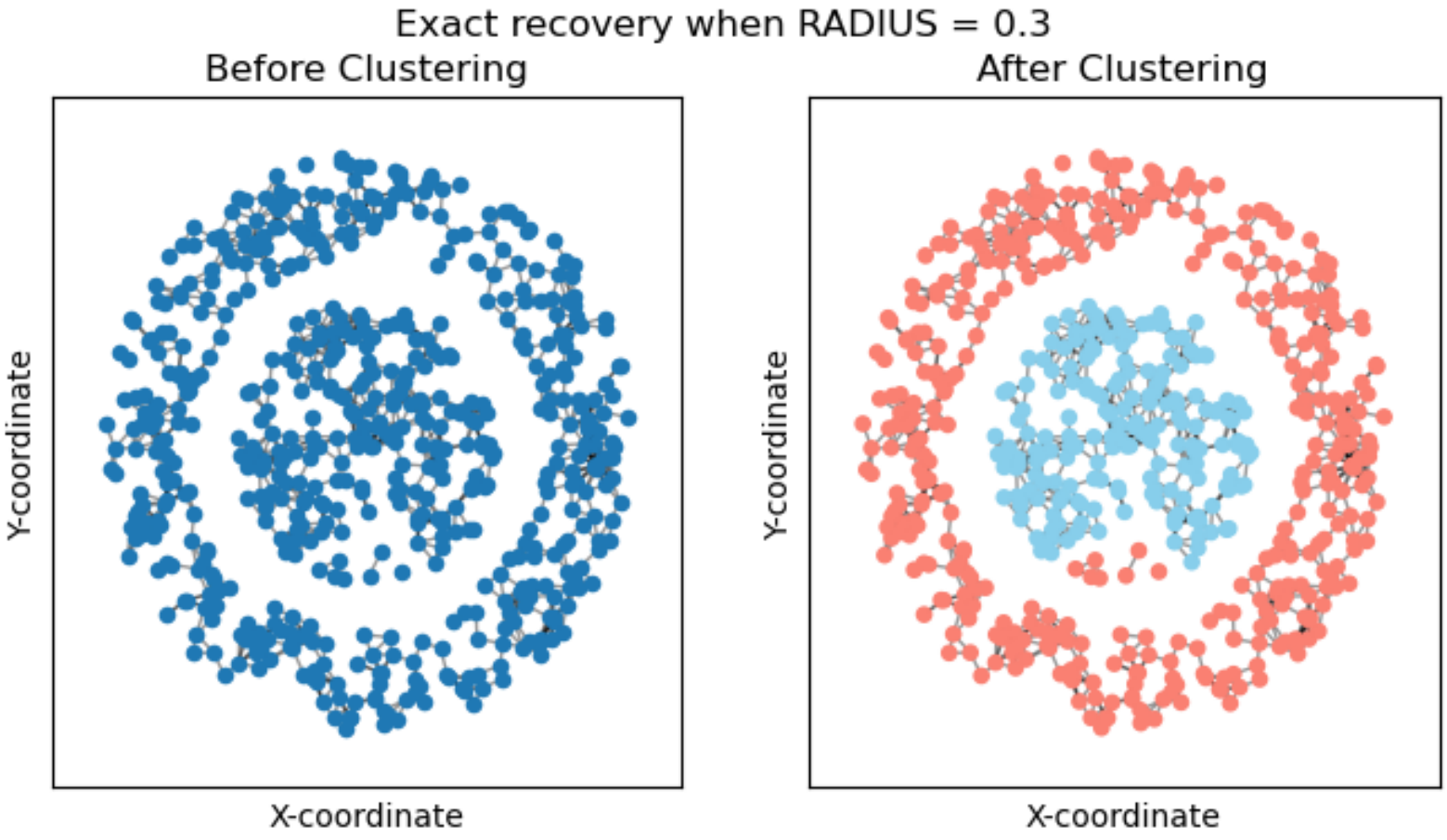


(c) $r_n = 0.5$

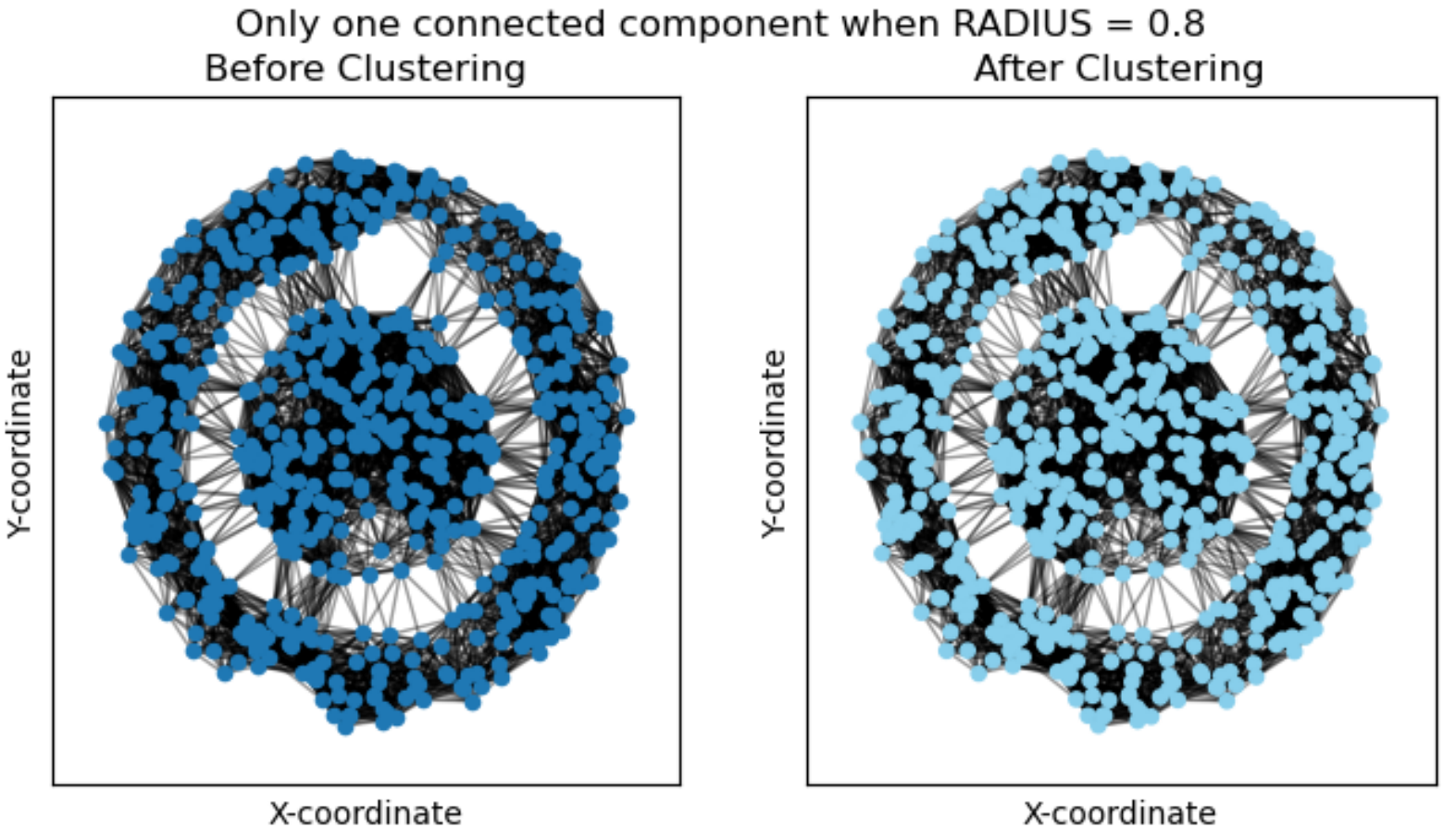
Clustering


Graph is almost disconnected when RADIUS = 0.25
Before Clustering
After Clustering
Y-coordinate
X-coordinate
Y-coordinate
X-coordinate

Clustering


Exact recovery when RADIUS = 0.3
Before Clustering
After Clustering
Y-coordinate
X-coordinate
Y-coordinate
X-coordinate

Clustering


Only one connected component when RADIUS = 0.8
Before Clustering
After Clustering
Y-coordinate
X-coordinate
Y-coordinate
X-coordinate

Related literature

- When $r_n \equiv r$, $g \equiv 1$, ℓ_∞ metric is used and the sampling distribution is uniform on $[-1,1]^d$, the edge eigenvalues of $\mathbf{L}_{rw}$ are studied by Adhikari et al. (2022)
- $d = 2, r = 1$:

$$\lambda_1(\mathbf{L}_{rw}) = 1, \lambda_2(\mathbf{L}_{rw}) \approx 1/2, \lambda_3(\mathbf{L}_{rw}) \approx 1/2, \lambda_4(\mathbf{L}_{rw}) \approx 3/4.$$

- When $r_n = o(1)$, ℓ_2 metric is used and the sampling distribution is compactly supported with density bounded from above and below, García Trillos et al. (2020) and García Trillos et al. (2021) show the (scaled) edges eigenvalues of $\mathbf{L}_{rw}$ converge to those of a weight Lapace-Beltrami operator.

Our results

Setup

- Gaussian samples $\mathbf{x}_j \sim \mathcal{N}_d(0, \Sigma)$ with $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_d^2)$;
- ℓ_p distance, i.e., $\|\cdot\|_p$, with $1 \le p \le \infty$;
- radius r_n,

$$n^{-\frac{1}{d+4}+\varepsilon} \ll r_n \ll n^{-\varepsilon},$$

  for some small $0 < \varepsilon < \frac{1}{2(d+4)}$;
- weight function $g \in C^2([0,\infty))$ and bounded from above and below by positive constants on $[0,1]$

$$\mathbf{K}(i,j) = \mathbf{1}(\|\mathbf{x}_i - \mathbf{x}_j\|_p \le r_n)\; g\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_p}{r_n}\right).$$

Setup

- scaling constants

$$m_\ell = \int_{\|\mathbf{u}\|_p \le 1} u_i^\ell \, g(\|\mathbf{u}\|_p) \, d\mathbf{u}$$

- our objective matrix is

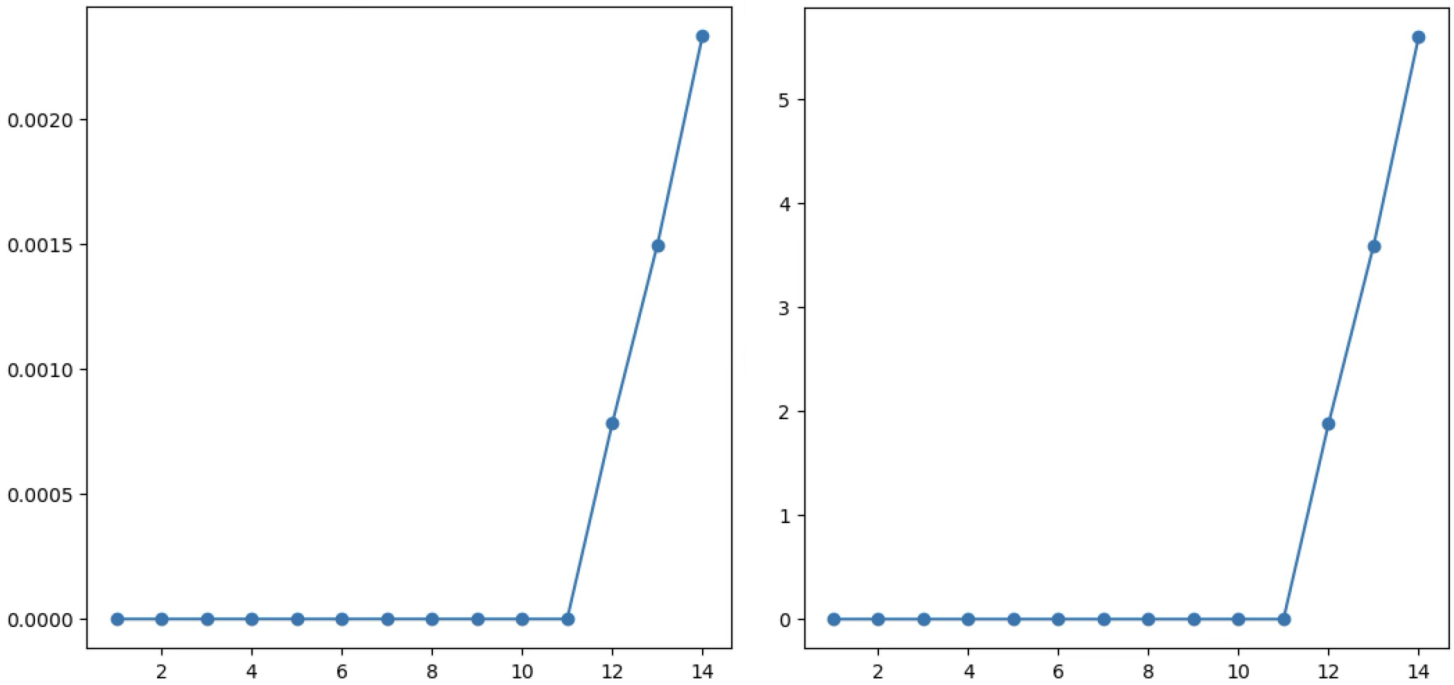
$$\mathbf{L} = \frac{2m_0}{m_2 r_n^2} \left(\mathbf{I} - \mathbf{D}^{-1}\mathbf{K}\right) = \frac{2m_0}{m_2 r_n^2} \mathbf{L}_{rw},$$

  where $\mathbf{K}(i,j) = \mathbf{1}(\|\mathbf{x}_i - \mathbf{x}_j\|_p \le r_n) \; g\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_p}{r_n}\right)$.

- eigenvalues of $\mathbf{L}$

$$\lambda_1(\mathbf{L}) \le \lambda_2(\mathbf{L}) \le \cdots \le \lambda_n(\mathbf{L})$$

Setup



- exclude small eigenvalues

$$K_0 = \max_j \{ j : \lambda_j(\mathbf{L}) \leq \delta \},$$

for some small $0 < \delta < \min\{\sigma_1^{-2}, \cdots, \sigma_d^{-2}\}$.

Continuum limits

- Hilbert space

$$\mathcal{F} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \int_{\mathbb{R}^d} f^2(\mathbf{x})\, \varrho^2(\mathbf{x})\, d\mathbf{x} < \infty \right\}$$

with weight

$$\varrho(\mathbf{x}) = \exp\left(-\sum_{i=1}^{d} \frac{x_i^2}{2\sigma_i^2} \right)$$

and inner product

$$\langle f_1, f_2 \rangle = \int_{\mathbb{R}^d} f_1(\mathbf{x})\, f_2(\mathbf{x})\, \varrho^2(\mathbf{x})\, d\mathbf{x}$$

Continuum limits

- weighted Laplace–Beltrami operator

$$\Delta_\varrho f = -\frac{1}{\varrho^2}\operatorname{div}\bigl(\varrho^2 \nabla f\bigr) = \sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2} - \frac{2x_i}{\sigma_i^2}\frac{\partial f}{\partial x_i}$$

- eigenvalues of Δ_ϱ on $(\mathcal{F}, \langle \cdot, \cdot \rangle)$

$$\mu_1(\Delta_\varrho) \leq \mu_2(\Delta_\varrho) \leq \cdots$$

Main results (I)

Theorem

For the weighted Laplace-Beltrami operator Δ_ϱ on $(\mathcal{F}, \langle\cdot,\cdot\rangle)$, its spectrum is discrete and consists of eigenvalues $\sum_{i=1}^{d} 2(k_i-1)/\sigma_i^2$ with corresponding orthonormal eigenfunctions

$$\prod_{i=1}^{d} \psi_{i,k_i-1}(x_i), \quad k_1, \cdots, k_d \in \mathbb{N}^+,$$

where

$$\psi_{i,k_i-1}(x_i) = \left(\frac{1}{\sqrt{\pi}(k_i-1)!2^{k_i-1}\sigma_i}\right)^{1/2} H_{k_i-1}(x_i/\sigma_i),$$

and H_{k_i-1} is the (k_i-1)-th physicist's Hermite polynomial.

Main results (II)

Theorem

Let M be any fixed integer. For sufficiently large n, with $1 - \mathrm{o}(1)$ probability

$$K_0 \leq n^{1-\frac{2}{(d+2)^2+\eta}} \quad \text{for some small } \eta > 0.$$

Moreover,

$$\lambda_{K_0+k}(\mathbf{L}) = \mu_{k+1}(\Delta_\varrho) + \mathrm{o}(1), \quad \text{for } k = 1, \cdots, M.$$

- scaling $2m_0/(m_2 r_n^2)$
- range of r_n as $n^{-\frac{1}{d+4}+\varepsilon} \ll r_n \ll n^{-\varepsilon}$
- bound of K_0

Main results (III)

For $k_1, k_2, \cdots, k_d \in \mathbb{N}^+$, denote a sequence

$$\left\{ \sum_{i=1}^{d} \frac{2(k_i - 1)}{\sigma_i^2} \right\},$$

ordered by $a_1 \leq a_2 \leq \cdots$.

Proposition

Under mild assumptions, $\mu_j(\Delta_\varrho) = a_j$. Consequently, for sufficiently large n and any fixed M, with probability $1 - \mathrm{o}(1)$,

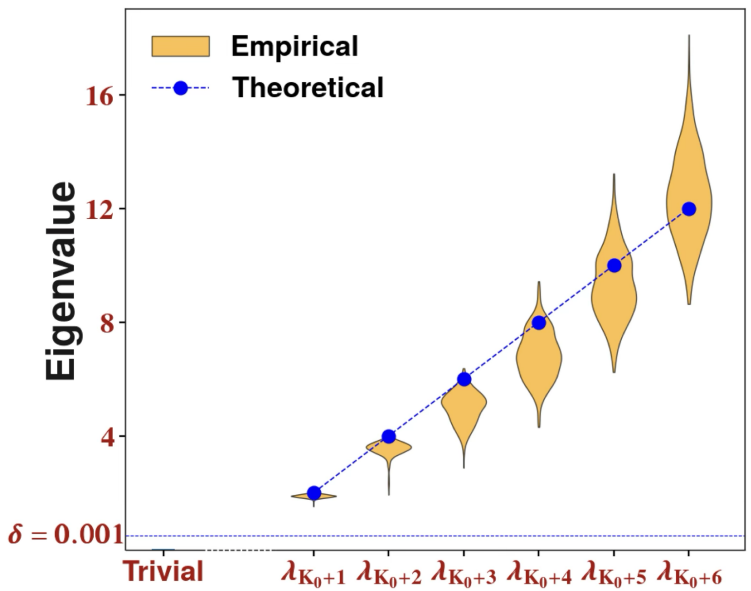
$$\lambda_{K_0+k}(\mathbf{L}) = a_{k+1} + \mathrm{o}(1), \quad \text{for } k = 1, \cdots, M.$$

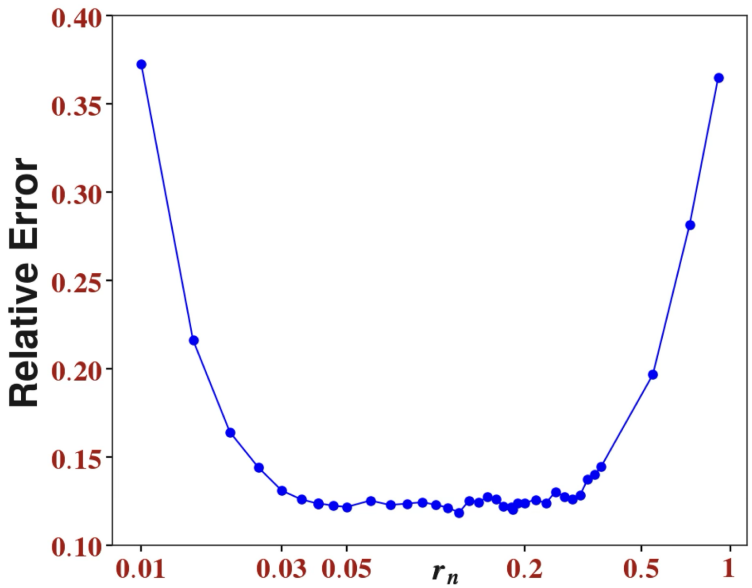
Examples

- $d = 1$, $\sigma_1 \equiv \sigma$. For $1 \leq k \leq M$,

$$\lambda_{K_0+k}(\mathbf{L}) = \frac{2k}{\sigma^2} + o_{\mathbb{P}}(1).$$

- $\sigma = 1, g \equiv 1, n = 5000$





$$\mathrm{RE} = \frac{1}{M}\sum_{k=1}^{M} |\lambda_{K_0+k}(\mathbf{L}) - \mu_{k+1}(\Delta_\varrho)|/\mu_{k+1}(\Delta_\varrho)$$

Examples

- $d = 2$, $\sigma_1 = \sigma_2 = \sigma$.

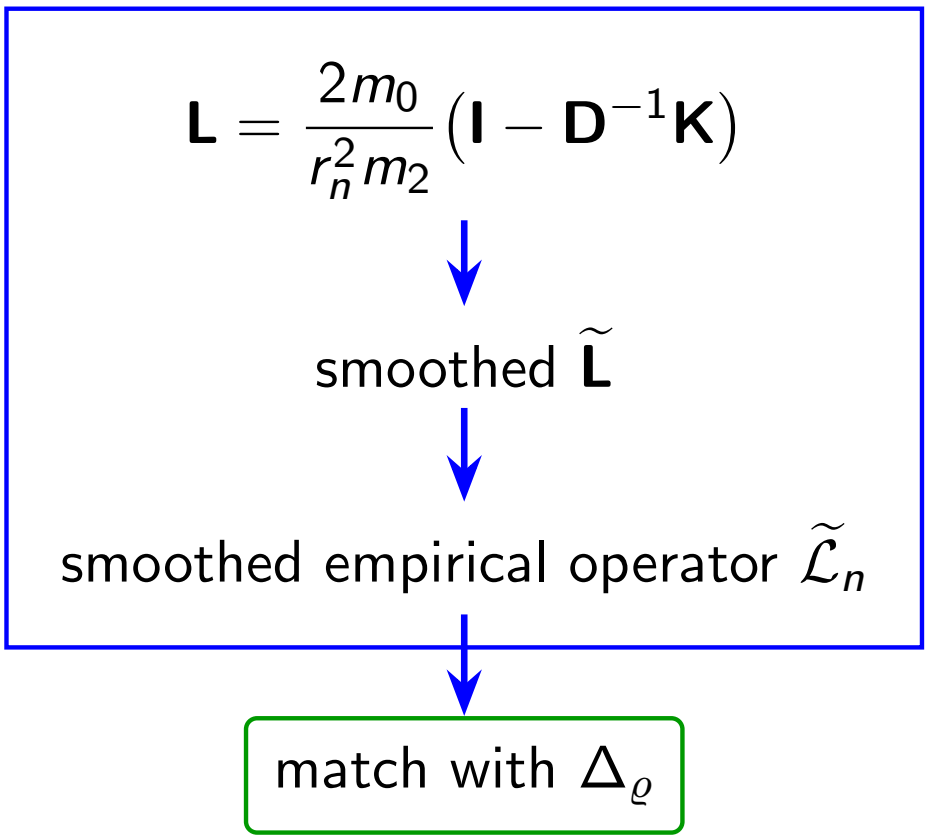
$$\lambda_{K_0+k}(\mathbf{L}) = \frac{2}{\sigma^2} \min \left\{ l : \sum_{s=1}^{l} s \geq k+1,\ l \in \mathbb{N} \right\} - \frac{2}{\sigma^2} + o_{\mathbb{P}}(1).$$

- $d = 2$, $\sigma_1^2 = \sigma^2$ and $\sigma_2^2 = 2\sigma^2$.

$$\lambda_{K_0+k}(\mathbf{L}) = \frac{1}{\sigma^2} \min \left\{ l : \sum_{s=0}^{l} \left(\lfloor s/2 \rfloor + 1 \right) \geq k+1,\ l \in \mathbb{N} \right\} + o_{\mathbb{P}}(1).$$

Proof strategy

- smoothing-matching



smoothing

matching

Smoothing

- $\mathbf{L} \to \widetilde{\mathbf{L}}$

$$\mathbf{K}(i,j) = \mathbf{1}(\|\mathbf{x}_i - \mathbf{x}_j\|_p \le r_n)\, g(\|\mathbf{x}_i - \mathbf{x}_j\|_p / r_n)$$

$$\widetilde{\mathbf{K}}(i,j) = s\big(\alpha_n\big(r_n^2 - \|\mathbf{x}_i - \mathbf{x}_j\|_p^2\big)\big)\, g^*(\|\mathbf{x}_i - \mathbf{x}_j\|_p / r_n)$$

  where $s(t) = 1/(1+\exp(-t))$, $\alpha_n \gg n^{7/2}$ and $g^*(t)$ is some extension of $g(t)$ from $[0,1]$ to $[0,\infty)$.

- $\|\mathbf{L} - \widetilde{\mathbf{L}}\| = o_{\mathbb{P}}(1)$ under $\alpha_n \gg n^{7/2}$

$$\lambda_j(\mathbf{L}) = \lambda_j(\widetilde{\mathbf{L}}) + o_{\mathbb{P}}(1), \quad 1 \le j \le n.$$

Smoothing

- $\widetilde{\mathbf{L}} \to \widetilde{\mathcal{L}}_n$

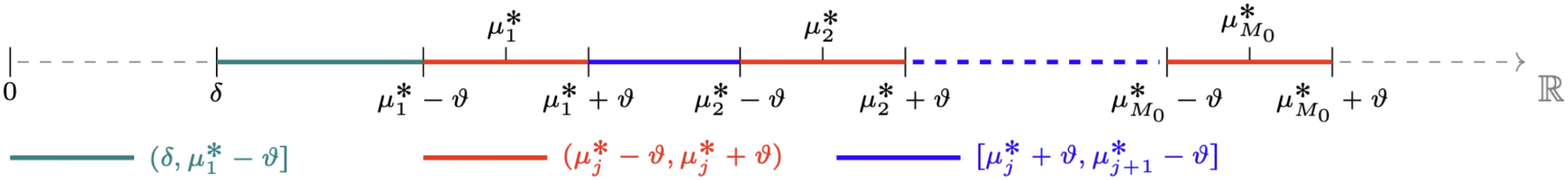
$$\widetilde{\mathcal{L}}_n f(\mathbf{x}) = \frac{2m_0}{r_n^2 m_2} \frac{\frac{1}{n}\sum_{j=1}^n s\big(\alpha_n\big(r_n^2 - \|\mathbf{x}-\mathbf{x}_j\|_p^2\big)\big)\; g^*\big(\|\mathbf{x}-\mathbf{x}_j\|_p / r_n\big)\,\big(f(\mathbf{x}) - f(\mathbf{x}_j)\big)}{\frac{1}{n}\sum_{j=1}^n s\big(\alpha_n\big(r_n^2 - \|\mathbf{x}-\mathbf{x}_j\|_p^2\big)\big)\; g^*\big(\|\mathbf{x}-\mathbf{x}_j\|_p / r_n\big)}$$

- well-defined
- Eigenvalues of $\widetilde{\mathbf{L}}$ and $\widetilde{\mathcal{L}}_n$ coincide for all $\lambda_j(\widetilde{\mathcal{L}}_n) < 2m_0/(r_n^2 m_2)$.

$$\lambda_j(\mathbf{L}) = \lambda_j(\widetilde{\mathcal{L}}_n) + o_{\mathbb{P}}(1), \quad \text{for } j \text{ such that } \lambda_j(\widetilde{\mathcal{L}}_n) < \frac{2m_0}{r_n^2 m_2}.$$

Matching

- $0 < \mu_1^* < \mu_2^* < \cdots < \mu_{M_0}^*$ with multiplicities $n_1, n_2, \cdots, n_{M_0}$, $\sum_{j=1}^{M_0} n_j \geq M$ and $\mu_{j+1}^* - \mu_j^*$ lower bounded
- partition of real line



- counting regions
- empty regions

Matching

- locating

  $\widetilde{\mathcal{L}}_n$ has at least one eigenvalue in each counting region

- counting

  comparing $-\frac{1}{2\pi i}\oint_\Gamma \mathrm{Tr}\Big(R(z,\widetilde{\mathcal{L}}_n)\Big)\,dz$ and $-\frac{1}{2\pi i}\oint_\Gamma \mathrm{Tr}(R(z,\Delta_\varrho))\,dz$

- empty regions

$$\lambda_j(\widetilde{\mathcal{L}}_n)=\mu_\ell^*+\mathrm{o}_{\mathbb{P}}(1),\ \text{for}\ K_0+1+\sum_{s=1}^{j-1}n_s\le j\le K_0+\sum_{s=1}^{j}n_s, 1\le \ell\le M_0.$$

Key ingredient

- closeness by eigenfunctions ψ_j of Δ_ϱ

$$\left\|(\Delta_\varrho - \widetilde{\mathcal{L}}_n)\psi_j\right\|_{\mathcal{F}} = o_{\mathbb{P}}(1)$$

- intermediate operator

$$\widetilde{\mathcal{T}}_n f(\mathbf{x}) = \frac{2}{r_n^{d+2} m_2\, \varrho(\mathbf{x})} \int_{\|\mathbf{x}-\mathbf{y}\|_p \le r_n} g(\|\mathbf{x}-\mathbf{y}\|_p / r_n)\, \big(f(\mathbf{x}) - f(\mathbf{y})\big)\, \varrho(\mathbf{y})\, d\mathbf{y}$$

- decomposition of $(\Delta_\varrho - \widetilde{\mathcal{L}}_n)\psi_j$

$$\left\|(\Delta_\varrho - \widetilde{\mathcal{L}}_n)\psi_j\right\|_{\mathcal{F}} \le \left\|(\Delta_\varrho - \widetilde{\mathcal{T}}_n)\psi_j\right\|_{\mathcal{F}} + \left\|(\widetilde{\mathcal{T}}_n - \widetilde{\mathcal{L}}_n)\psi_j\right\|_{\mathcal{F}}$$

Applications to manifold data clustering

- $\mathbf{x}_j = \mathbf{s}_j + \mathbf{z}_j \in \mathbb{R}^p$
- $\mathbf{s}_j$ are sampled from manifolds
- $\varrho(\mathbf{s}_j) = \sum_{k=1}^{K} \mathbb{P}(Y_j = k)\varrho_k(\mathbf{s}_j|k)$
- Distinct vertices $i, j \in \mathcal{V}$ are connected if and only if $\|\mathbf{x}_i - \mathbf{x}_j\| \leq r_n$, where the threshold radius $r_n > 0$ is such that

$$\mathbb{P}(\|\mathbf{s}_i - \mathbf{s}_j\| \leq r_n) = \mathrm{Q}_{y(i)y(j)},$$

  where $\mathrm{Q} \in [0,1]^{K \times K}$ is some symmetric probability matrix
- Data $\rightarrow$ Random Geometric Graph (RGG) construction with parameter selection based on spectral analysis $\rightarrow$ Eigenvector-based embedding $\rightarrow$ Clustering