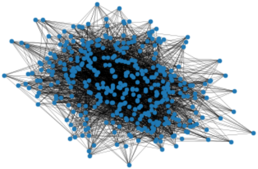
Higher-order spectral clustering for geometric graphs

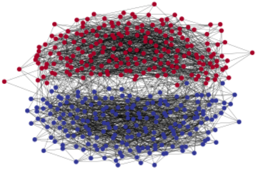
K. Avrachenkov[2] A. Bobu[2] **M. Dreveton**[1,2]

[1] Inria Sophia-Antipolis, France
[2] EPFL, Switzerland

INFORMS APS – 28 June 2023

Introduction: spectral methods for graph clustering





Spectral methods:

Input: Matrix M (*e.g.*, L, $\mathcal{L}$, A);

Output: Node labeling $\widehat{\sigma} \in \{-1, 1\}^n$.

Algorithm: Compute $v^{(2)}$, the eigenvector associated with the second smallest (or largest) eigenvalue of M;

- Let $\widehat{\sigma}_i = \mathrm{sign}\left(v_i^{(2)}\right)$ for $i = 1, \ldots, n$.

Aim of this talk

Explain why spectral methods can fail when nodes have geometric attributes, and propose a solution.

[1] Fiedler, M.: A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. Czechoslov. Math. J. 25(4), 619–633 (1975)

Soft Geometric Block Model (SGBM)

Model parameters

number of nodes n, geometric dimension d and two measurables functions $F_{\mathrm{in}}, F_{\mathrm{out}} : \mathbb{T}^d \to [0,1]$.

Model definition

- Set of nodes $V = \{1, \dots, n\}$;
- Each node i has a random position X_i on the torus $\mathbb{T}^d$;
- Each node i is randomly assigned a community label $\sigma_i \in \{-1, 1\}$;
- Each pair of nodes (i,j) is connected with probability

$$p_{ij} = \begin{cases} F_{\mathrm{in}}\left(X_i - X_j\right) & \text{if } \sigma_i = \sigma_j, \\ F_{\mathrm{out}}\left(X_i - X_j\right) & \text{if } \sigma_i \neq \sigma_j. \end{cases}$$

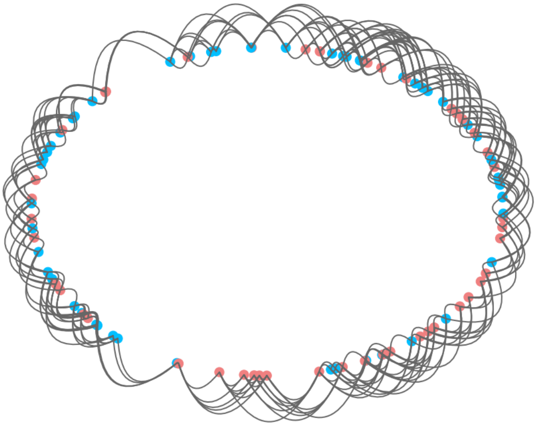
Inference problem

Estimate the latent node labelling σ given the observation of A, and possibly the knowledge of $F_{\mathrm{in}}, F_{\mathrm{out}}$.

Example: GBM

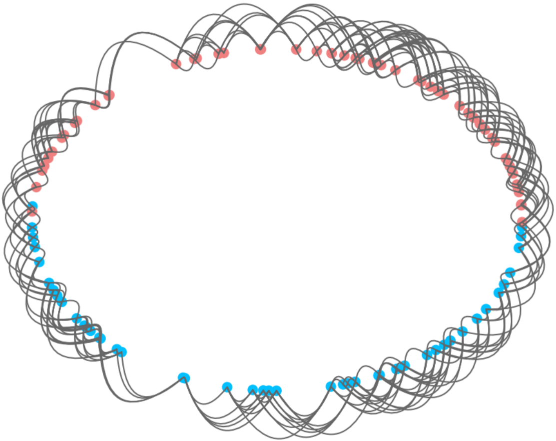
Geometric Block Model (GBM) [1]

Consider $d = 1$ and $F_{\mathrm{in}}(x) = 1(|x| \leq r_{\mathrm{in}})$, $F_{\mathrm{out}}(x) = 1(|x| \leq r_{\mathrm{out}})$ with fixed $r_{\mathrm{in}} > r_{\mathrm{out}}$.



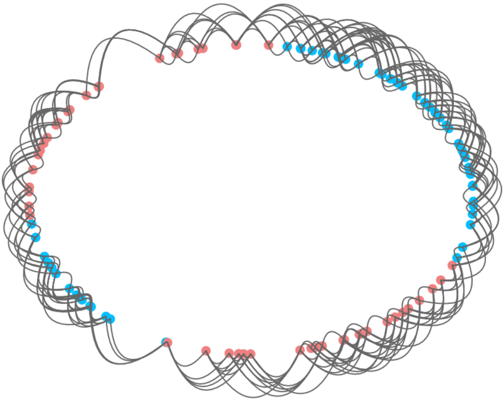
[1] Galhotra, Mazumdar, Pal, Saha: The geometric block model. In Proceedings of the AAAI Conference on Artificial Intelligence.

Spectral clustering on the GBM (1)



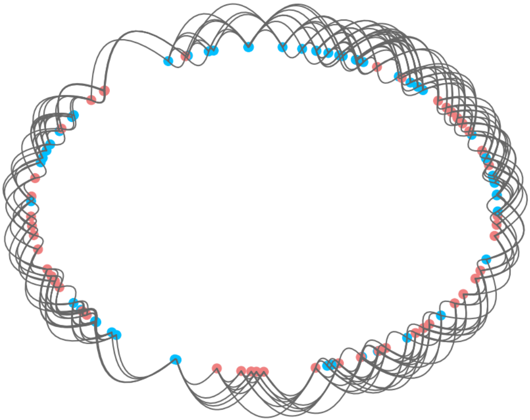
Geometric partitioning!

Spectral clustering on the GBM (2)



The eigenvector v_4 associated with λ_4 (the fourth smallest eigenvalue) gives the partition into 4 regions.
The eigenvector v_6 divides the circle into 6 regions, and so on... Nothing useful?

Spectral clustering on the GBM (3)



The eigenvector v_{10} gives accuracy 87%!
It contains useful information about the true community structure.

How to choose the best eigenvector?

Suppose the two clusters are $V_1 = \{1, \ldots, n/2\}$, $V_2 = \{n/2+1, \ldots, n\}$. The ideal vector for recovery is then

$$v_* = (\underbrace{1, \ldots, 1}_{n/2}, \underbrace{-1, \ldots, -1}_{n/2})^T.$$

Denote

$$\begin{aligned} \mu_{\rm in} &= \int_{\mathbf{T}^d} F_{in}(x)dx \quad \text{average intra-cluster edge density,} \\ \mu_{\rm out} &= \int_{\mathbf{T}^d} F_{out}(x)dx \quad \text{average inter-cluster edge density.} \end{aligned}$$

Hence v_* is an eigenvector of $\mathbb{E}A$, associated to λ_* such that

$$\boxed{\lambda_* = \mathbb{E}\sum_{j=1}^{n/2} A_{ij} - \mathbb{E}\sum_{j=n/2+1}^{n} A_{ij} = \frac{(\mu_{in} - \mu_{out})n}{2}}$$

Idea: take the eigenvector $\tilde{v}$ of A associated with $\tilde{\lambda}$ the closest to $\lambda_* = (\mu_{in} - \mu_{out})n/2$

Higher-order spectral clustering algorithm

Higher-order spectral clustering algorithm (HOSC):

1. Compute the eigenvalues of the adjacency matrix A;
2. Take the eigenvector $\tilde{v}$ associated with the eigenvalue $\tilde{\lambda}$ closest to $\lambda_* = (\mu_{\text{in}} - \mu_{\text{out}})n/2$;
3. Let $\widehat{\sigma}_i = \text{sign}(\tilde{v}_i)$ for $i = 1, \ldots, n$.

Higher-order spectral clustering algorithm

Higher-order spectral clustering algorithm (HOSC):

1. Compute the eigenvalues of the adjacency matrix A;
2. Take the eigenvector $\tilde{v}$ associated with the eigenvalue $\tilde{\lambda}$ closest to $\lambda_* = (\mu_{\rm in} - \mu_{\rm out})n/2$;
3. Let $\widehat{\sigma}_i = \text{sign}(\tilde{v}_i)$ for $i = 1, \ldots, n$.

Clustering error: $\ell(\sigma, \widehat{\sigma}) = \min\{\text{Ham}(\sigma, \widehat{\sigma}), \text{Ham}(\sigma, -\widehat{\sigma})\}$ where Ham is the Hamming distance.

Higher-order spectral clustering algorithm

Higher-order spectral clustering algorithm (HOSC):

1. Compute the eigenvalues of the adjacency matrix A;
2. Take the eigenvector $\tilde{v}$ associated with the eigenvalue $\tilde{\lambda}$ closest to $\lambda_* = (\mu_{\rm in} - \mu_{\rm out})n/2$;
3. Let $\widehat{\sigma}_i = \mathrm{sign}(\tilde{v}_i)$ for $i = 1, \ldots, n$.

Clustering error: $\ell(\sigma, \widehat{\sigma}) = \min\{\mathrm{Ham}(\sigma, \widehat{\sigma}), \mathrm{Ham}(\sigma, -\widehat{\sigma})\}$ where Ham is the Hamming distance.

Theorem (Avrachenkov, Bobu, Dreveton 2021)

In the GBM, for almost all choices of parameters $(r_{\rm in}, r_{\rm out})$, we have with high probability $\ell(\sigma, \widehat{\sigma}) = o(n)$.

Remark. We can have $\ell(\sigma, \widehat{\sigma}) = o(1)$ with an additional step (not shown here).

Spectrum of the SGBM

For $k \in \mathbb{Z}^d$ and $F : \mathbf{T}^d \to \mathbb{R}$ we define the Fourier transform as

$$\widehat{F}(k) = \int_{\mathbf{T}^d} F(x) e^{-2i\pi \langle k,x \rangle} \, dx.$$

Theorem (Informal statement)

Assume that $F_{\mathrm{in}}(0), F_{\mathrm{out}}(0)$ are equal to the Fourier series of $F_{\mathrm{in}}, F_{\mathrm{out}}$ evaluated at 0. Then, the limiting spectrum of the adjacency matrix of the SGBM is

$$S = \left\{ \frac{\widehat{F}_{\mathrm{in}}(k) + \widehat{F}_{\mathrm{out}}(k)}{2} n \text{ for } k \in \mathbb{Z}^d \right\} \bigcup \left\{ \frac{\widehat{F}_{\mathrm{in}}(k) - \widehat{F}_{\mathrm{out}}(k)}{2} n \text{ for } k \in \mathbb{Z}^d \right\}.$$

This extends [1] to clustered geometric graphs.

Good news: $\lambda_* = \frac{\mu_{\mathrm{in}} - \mu_{\mathrm{out}}}{2} n = \frac{\widehat{F}_{\mathrm{in}}(0) - \widehat{F}_{\mathrm{out}}(0)}{2} n \in S$.

Now: Need to establish that λ_* is of multiplicity one and is separated from other eigenvalues.

[1] Bordenave, C. (2008). Eigenvalues of Euclidean random matrices. Random Structures Algorithms, 33(4), 515-532.

Separation of λ_*

Proposition

Consider the adjacency matrix A of an SGBM and assume that:

$$\mu_{\text{in}} - \mu_{\text{out}} \neq \widehat{F}_{\text{in}}(k) + \widehat{F}_{\text{out}}(k), \qquad \forall k \in \mathbb{Z}^d, \tag{1}$$

$$\mu_{\text{in}} - \mu_{\text{out}} \neq \widehat{F}_{\text{in}}(k) - \widehat{F}_{\text{out}}(k), \qquad \forall k \in \mathbb{Z}^d \backslash \{0\}, \tag{2}$$

with $\mu_{\text{in}} \neq \mu_{\text{out}}$. Then:

- *the eigenvalue of A the closest to $\lambda_* = \frac{\mu_{\text{in}} - \mu_{\text{out}}}{2} n$ is of multiplicity one;*
- *there exists $\epsilon > 0$ such that for large enough n every other eigenvalue is at a distance at least ϵn.*

Remark 1. This implies that μ_{in} and μ_{out} are constant, so the average degrees are $\Theta(n)$.

Remark 2. In the case of the GBM ($F_{\text{in}}(x) = 1(|x| \leq r_{\text{in}})$ and $F_{\text{out}}(x) = 1(|x| \leq r_{\text{out}})$), we showed that conditions (1)-(2) hold true for all but a zero Lebesgue measure set of parameters $r_{\text{in}}, r_{\text{out}}$.

Numerical experiments on GBM (1)

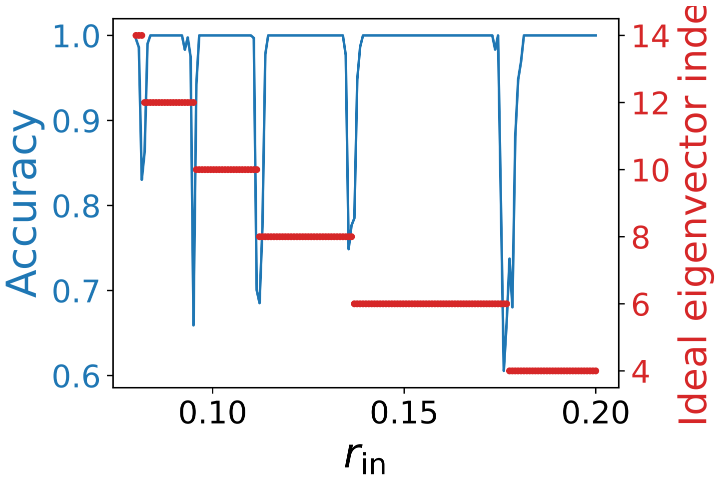


Figure: Blue curve: evolution of the accuracy with r_{in}, for a GBM with $n = 3000$ and $r_{\text{out}} = 0.06$. Red curve: index of the ideal eigenvector.

Numerical experiments on GBM (2)

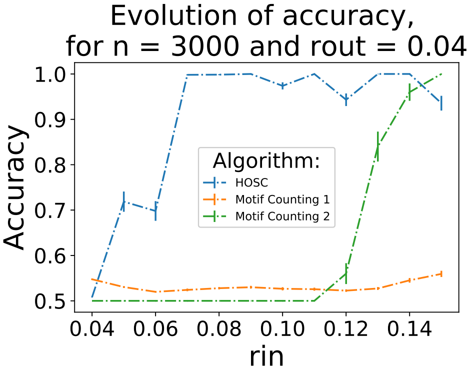


Figure: Accuracy obtained on 1-dimensional GBM for different clustering methods. Results are averaged over 50 realizations, and error bars show the standard error.

[1] Galhotra, Mazumdar, Pal, Saha: The geometric block model. In Proceedings of the AAAI Conference on Artificial Intelligence.
[2] Galhotra, Mazumdar, Pal, Saha: Connectivity of random annulus graphs and the geometric block model. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* (2019)

Real data set (1)

k-nearest neighbour graph ($k = 10$) of 1000 images of digits (7,9) selected from MNIST.

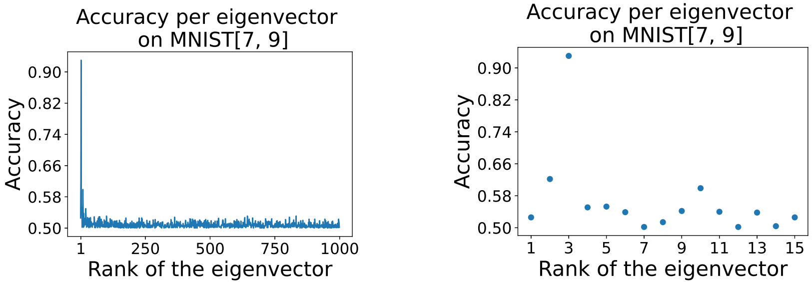
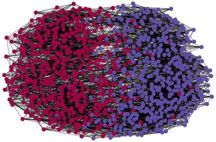


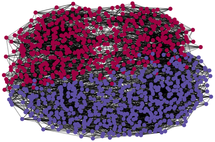
Figure: Clustering accuracy per eigenvector. Right: all eigenvectors. Left: zoom on the first 15 eigenvectors.

Real data set (2)

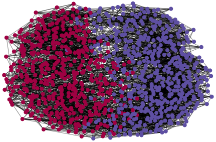
Same subsample of the MNIST (7,9) data set, representation using the Kamada-Kawai layout.



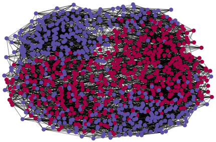
(a) True labels



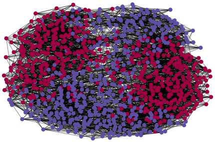
(b) 2nd eigenvector



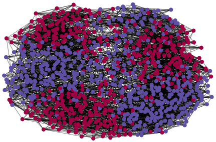
(c) 3rd eigenvector



(d) 4th eigenvector



(e) 6th eigenvector



(f) 9th eigenvector

Directions of further research

Takeaway message

If you use spectral clustering methods, check higher-order eigenvectors, they can be more effective! Especially if you deal with geometry.

Future work

- **Model parameters**
  Is it possible to determine μ_{in} and μ_{out} from the observed graph?
- **More clusters**
  How to choose the eigenvector(s) if we have $K > 2$ clusters?
- **Sparse regime**
  The current technique does not work if the average degree is $o(n)$. What to do?
- **Weighted graphs**
  Can the results be easily transferred to models with weighted edges instead of the probability of edge appearance?

Reference: Avrachenkov, Bobu, Dreveton (2021). Higher-order spectral clustering for geometric graphs. Journal of Fourier Analysis and Applications, 27(2), 22.