Spectral clustering in sparse heterogeneous networks

Lorenzo Dall’Amico, lorenzo.dall-amico@gipsa-lab.fr
Romain Couillet, romain.couillet@gipsa-lab.fr
Nicolas Tremblay, nicolas.tremblay@gipsa-lab.fr

Abstract

- Study of the problem of unsupervised clustering in sparse heterogeneous networks. In this regime classical Laplacians fail.
- Proposition of an original analysis of the deformed Laplacian matrix $D - rA$, showing that for a particular value of $r = \zeta$ it can perform clustering on sparse networks down to the detectability threshold, reaching Bayes optimality and showing resilience to degree heterogeneity.
- We then provide unsupervised tool to estimate $r_{opt} = \zeta$.

Problem statement (I)

Community detection

Observation: nodes with similar properties connect more easily.
Problem: reconstruct classes given the network Constraint: sparse heterogeneous networks
Method: spectral techniques (matrix eigenvectors)

Notation – DC-SBM

$P(A_{ij} = 1 | \sigma_i, \sigma_j, q, q_j) = q_q C(\sigma_i, \sigma_j) / n$
- $A$: adjacency matrix
- $\sigma$: label vector
- $q$: intrinsic connection probability, i.i.d with $(q) = 1$
- $C(\sigma_i, \sigma_j)$: affinity matrix
- $P(\cdot)$: sparse regime
- $n = \sqrt{\frac{e}{\lambda_{\max}}}$: control parameter

Bayes point of view (II)

A property of the DC-SBM

$P(\sigma_i, \sigma_j | A_{ij} = 1) = \frac{1}{Z} \int \int dqdq_{ij} \tilde{P}(A_{ij} | \sigma_i, \sigma_j, q, q_j) = \frac{C(\sigma_i, \sigma_j)}{2c_{in} + c_{out}}$

- Does not depend on the degree distribution
- Is relevant in the tree-like regime
- $|d_i^{(\text{in})}|$: size of the neighborhood of $i$ in the same/opposite class

Sum of i.i.d Bernoulli RV with parameter $p = c_{in}/c_{in} + c_{out}$

The deformed laplacian

- $D - rA$: Deformed Laplacian
- Average behaviour and Bayes optimal choice of $r$

$E[(D - rA\sigma)_{ij}^2] = d_{ij} \left( 1 - r \left( \frac{\partial^2 \mathcal{L}}{\partial \sigma_i \partial \sigma_j} - \frac{\partial^2 \mathcal{L}}{\partial \sigma_i \partial \sigma_{ij}} \right) \right) = d_{ij} \left( 1 - \frac{c_{in} - c_{out}}{c_{in} + c_{out}} \right)$

Optimal choice of $r$

$r_{opt} = \zeta \frac{c_{in} + c_{out}}{c_{in} - c_{out}}$

Estimate $\zeta$ (IV)

Property

- The non-backtracking matrix $B_{\text{non-back}} = d_{ij} (1 - d_{ij})$
- The Bethe-Hessian matrix $H_{\text{B}} = (r^2 - 1) I + D - rA$
- Property

$Br \rightarrow \text{det}(H_{\text{B}}) = 0$

Spectrum of $B$

More classes (V)

Straightforward generalization using $\{\zeta_p\}$ with

$C(\zeta_p) = \lambda_p \zeta_p$

$\zeta_p = \frac{c_{in}}{\lambda_p}$

The algorithm

- Estimate the number of classes ($r = \sqrt{c_{in}}$)
- Estimate the values of $r$ (linesearch using $H_{\text{B}}$)
- Look into the eigenvectors one by one
- Perform k-means

The result

- Resilience to degree heterogeneity
- Bayes optimality
- Recovery down to the detectability threshold

REFERENCES

2. Saade et al.: Spectral clustering of graphs with the bethe hessian (2014)