An l_1-version of the spectral clustering to promote sparse eigenvectors basis

Camille Champion [1] & Mélanie Blazère [1] & Fabrice Gamboa [1] & Jean-Michel Loubes [1]

[1] *Toulouse Mathematics Institute*
camille.champion@math.univ-toulouse.fr
fabrice.gamboa@math.univ-toulouse.fr
loubes@math.univ-toulouse.fr

Résumé. Les graphes jouent un rôle central dans la modélisation des systèmes complexes. Leur analyse est une problématique importante qui couvre une grande variété de domaines et d'applications. Dans ce contexte, nous proposons une variante d'une des méthodes les plus connues d'analyse de graphe, le spectral clustering. Cette nouvelle méthode, appelée l_1-spectral clustering, ne requiert pas l'utilisation du k-means pour regrouper les nœuds du graphe, mais estime directement les indicateurs des communautés en déterminant une base propre spécifique à partir d'une pénalité l_1.

Mots-clés. Spectral clustering, détection de communautés, base de vecteurs propres, pénalité l_1.

Abstract. Graphs play a central role in complex systems. Analyzing a graph is a major issue that covers a wide range of fields and applications. In this context, we propose a variant of one of the most well-known graph clustering method, the spectral clustering. This procedure, called l_1-spectral clustering, does not require the use of k-means to cluster the nodes of the graph, but directly estimates the indicators of the communities by computing a specific eigenbasis using l_1 penalty.

Keywords. Spectral clustering, community detection, eigenvectors basis, l_1-penalty.

Introduction

Graphs play a central role in complex systems as they are involved in the extraction of useful information from the related problem. They cover a wide field of applications, ranging from mathematics to physics, sociology, marketing, informatics or biology. Graphs are a convenient way to model and study interactions between individuals represented by nodes. One of the challenges when analyzing graphs is the estimation of these interactions when the structure is not known or not fixed. Another challenging task is the understanding of the graph structure by clustering highly connected subsets of nodes. For instance, in genetics, groups of genes with high interactions are likely to be involved in a same function that drives a specific biological process. More generally, such community structures (groups of nodes that are densely connected with sparse connections in between) are believed to play an important role in the functioning of complex systems modelled

by graphs, so that detecting theses structures is of the highest importance (Girvan and Newman(2002), Newman and Girvan(2004)) .

Among different graph community detection methods, spectral clustering is currently one of the most popular (Von Luxburg (2007)). This method uses the eigenvectors of adjacency type matrices to cluster the nodes of a graph into a given number of communities. The nodes are not directly clustered but k-means is applied to the eigenvectors to detect the communities. If this method is so popular it is mainly because spectral clustering is very easy to implement and easy to use: computations are thus very fast and efficient, even for very large graphs. However, there is no guarantee to reach the best or most natural partitioning for general models.

In reply to this issue, we developed a first regularization technique, which is cast into a specific graph clustering strategy (Champion and Brunet and Loubes and Risser(2018)). this edge aims at detecting highly connected subnetworks containing representative variables selected as the centers of specific variables clusters, called Core-clusters. Core-clusters have two properties: they contain more than a predefined amount of variables and each of their possible variable pairs have a coherent observed behavior. In addition, the representative variables selected as the centers of Core-clusters are clearly more interpretable and pertinent than those estimated using spectral clustering method with an intuitive parameter tuning.

Then, we developed an alternative method to Spectral clustering, called l_1-spectral clustering. This procedure does not require the use of k-means to cluster the nodes of the graph, but directly estimates the indicators of the communities by computing a specific eigenbasis, better suited for clustering, using l_1 penalty in a specific random graph model.

1 Notation and model

1.1 Graph notations

We consider a graph $G(V, E)$ with n fixed nodes which refers to a set of vertices (nodes) and a set of edges (links). The nodes, labeled from 1 to n, represent the individuals or objects and the edges the interactions and relationships between them so that $V = \{1, \ldots, n\}$. From a mathematical point of view, a graph G is a pair (V, E) where V is the set of vertices and E refers to the set of edges that pairwise connect the vertices. An edge $e \in E$ that connects a node i and a node j is denoted by $e = (i, j)$. In our setting, we consider only unweighted and undirected graphs with fixed vertices and no loops. An important object associated to the graph is the adjacency matrix $A = (A_{ij})_{(i,j)\in V^2}$ defined by

$$A_{i,j} = \begin{cases} 1 & \text{if there is an edge between } i \text{ and } j, \\ 0 & \text{otherwise.} \end{cases}$$

Since the graph is undirected, $A \in \mathbb{M}_n(\mathbb{R})$ is a symmetric matrix that is $A_{ij} = A_{ji}$. Moreover, $A_{ii} = 0$ because there are no loops.

The degree d_i of a node i is equal to the number of edges incident to i (whose root is node i), so that

$$d_i = \sum_{j=1}^{n} A_{ij}.$$

The matrix D, called the degree matrix, contains $(d_1, \ldots, d_n)$ on the diagonal and zero anywhere else

$$D = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{pmatrix}.$$

Given a subset of vertices $C \in V$, we define the indicator $\mathbf{1}_C \in \mathbb{R}^n$ as the vector whose entries are defined by

$$(\mathbf{1}_C)_i = \begin{cases} 1 & \text{if vertex } i \text{ belong to } C, \\ 0 & \text{otherwise.} \end{cases}$$

where a subset $C \in V$ of a graph is said to be connected if any two vertices in C are connected by a path in C (sequence of vertices in C connected by edges that joined the two initial vertices). In addition, C is called a connected component if there are no connections between vertices in C and $\overline{C}$. Non empty sets $C_1, \ldots, C_k$ form a partition of the graph if $C_i \cap C_j = \emptyset$ and $C_1 \cup \cdots \cup C_k = V$.

1.2 Presentation of the model

The ideal graph G is assumed to be the union of k complete graphs that are disconnected from each other. We allow the number of vertices in each subgraph to be different. We denote by $s_1, \ldots, s_k$ (≥ 2) their respective size $(\sum_{i=1}^{k} s_i = n)$. To simplify, we assume that the nodes $\{1, \ldots, n\}$ are ordered with respect to their block membership and in increasing order with respect to the size of the blocks.

$$A = \begin{bmatrix} \underbrace{\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 \end{bmatrix}}_{s_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \underbrace{\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 \end{bmatrix}}_{s_k} \end{bmatrix}$$

The reality is that we do not have access to the graph G with such an adjacency matrix but we observe a perturbed version of G denoted by $\hat{G}$. We actually assume that the perturbed graph results form a deterministic graph with an exact community structure, whose edges have been dusturbed by Bernoulli variables.

The associated adjacency matrix $\hat{A}$ satisfies

$$\hat{A} = A \overset{2}{\bigoplus} B$$

where

- $\hat{A}_{ij} = \{A \overset{2}{\bigoplus} B\}_{ij} = A + B \ (mod 2)$.
- B is a symmetric matrix of size n, whose upper entries are realizations of independent Bernoulli variables i.e. $B_{ij} \sim \mathcal{B}(p) \ \ i.i.d, \ i < j$ with $B_{ii} = 0$ and $B_{ij} = Bji$.

B is therefore the adjacency matrix of an Erdos-Renyi graph $G(n, p)$.

In this framework, we suggest the use of spectral clustering algorithm to find the underlying communities of G from the observation of the noisy graph $\hat{G}$.

2 ℓ_1-spectral clustering, a new graph community detection method

2.1 Model of the traditional spectral clustering

The original spectral clustering has been proposed by Von Luxburg (2007) to cluster the nodes of the graph into communities using the first k eigenvectors (corresponding the k smallest eigenvalues) of a normalized or unormalized version of the Laplacian matrix (derived from the adjacency one). Let $G = (V, E)$ be a graph made of k connected components $C_1, \ldots, C_k$. Let A be the associated adjacency matrix and D its degree matrix. Let $L = D - A$, $L_{sym} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ and $L_{rw} = I - D^{-1}A$ be respectively the Laplacian, Symmetric Laplacian and random walk Laplacian matrix. Then, the following spectral proposition holds:

Proposition 1 *(Number of connected components and spectra of L_{sym} and L_{rw}) :*

Let G be an undirected graph. Then the multiplicity k of the eigenvalue 0 of both L and L_{sym} equals the number of connected components $C_1, \ldots, C_k$ in the graph. For L_{rw} , the eigenspace of 0 is spanned by the indicator vectors $\{1_{C_i}\}_{1 \leq i \leq n}$ of those components. For L_{sym}, the eigenspace of 0 is spanned by the vectors $\{D^{1/2} 1_{C_i}\}_{1 \leq i \leq n}$.

We deduce from **Proposition 1** (Von Luxburg (2007)) that the multiplicity of the null eigenvalue (that corresponds to the smallest eigenvalue) is equal to the number of connected components. Thus, a particular basis of the associated eigenspace is spanned by the community indicators. In addition, the rows of the matrix resulting from the

concatenation of the k first eigenvectors, corresponding to indices of nodes, in the same class are equal. Therefore, it is natural to apply k-means to these rows to provide, by the same way, the knowledge of the blocks which make them very attractive.

However, the graph is not made of connected components, but of densely connected subgraphs that are sparsely connected to each other. These densely connected subgraphs represent somehow a perturbed version of the initial connected components that form the communities. If the perturbation is not too high, we can still hope that the eigenvectors of the perturbed Laplacian matrix (associated to this perturbed graph) still contain enough information on the graph structure to detect these communities. Therefore, there is no theoretical guarantee that we recover the true communities.

In the next section, we present the heart of our contribution to spectral clustering methods.

2.2 ℓ_1-spectral clustering

The way the eigenvectors basis of the matrix is built is of the highest importance to ensure a good recovery of the communities. The key is to select relevant eigenvectors that provide useful information about the natural grouping of the data.

Unlike the spectral clustering method, the alternative one aims at finding reliably the k underlying communities of a graph G from the observation of the noisy graph $\hat{G}$. We still focus on the space spanned by the k first eigenvectors but directly computed from the adjacency matrix. The idea remains the same, the only difference is that the eigenvalues associated to the eigenvectors $\{\mathbf{1}_{C_i}\}_{1\leq i\leq k}$ have a different value: $d_1, \ldots, d_k$. The other eigenvalues are equal to -1. The community indicators are the eigenvectors associated this time to the largest eigenvalues. We assume that the eigenvalues of A denoted are sorted in decreasing order.

Let $u_1, \ldots, u_n$ be the associated normalized eigenvectors given by any eigensolvers, so that the k first eigenvectors of A (associated to the k largest eigenvalues) are denoted by $u_1, \ldots, u_k$. We denote by U_k the matrix that contains $u_1, \ldots, u_k$ in columns and by V_k the one that contains $u_{k+1}, \ldots, u_n$. We define $\mathcal{U}_k = \text{Span}\{u_1, \ldots, u_k\}$. The first community indicator is solution of some specific minimization problem:

Proposition 2 *The minimization problem*

$$\arg\min_{v\in\mathcal{U}_k\setminus\{0\}} \|v\|_0$$

has a unique solution (up to a constant) given by $\mathbf{1}_{C_1}$.

In other words, $\mathbf{1}_{C_1}$ is the sparsest non-zero eigenvector in the space spanned by the first k eigenvectors.

Notice that the constraints are linear. However, because of the l_0-norm this minimization problem is NP-hard. But assuming the knowledge of one representative for each group, we can replace the l_0-norm by its convex relaxation given by the l_1-norm. Thus, in addition to the number of communities, we assume that we know one representative

of each community. By a representative, we mean a node belonging to this community. This assumption is not so restrictive compared to traditional spectral clustering where the number of communities is assumed to be known.

Let $i_1, \ldots, i_k$ be the indices of these representative elements and let $\tilde{\mathcal{U}}_k = \{v \in \mathcal{U}_k : v_{i_1} = 1\}$.

This is straightfoward to see that the community indicator of the smallest community is solution of the following optimization problem.

Proposition 3 *The minimization problem ($\mathcal{P}_1$)*

$$\arg\min_{v \in \tilde{\mathcal{U}}_k} \|v\|_1$$

has a unique solution given by $\mathbf{1}_{C_1}$.

To simplify and without loss of generality, we assume that i_1 corresponds to the first index (up to a permutation).

Proposition 4 *Problem ($\mathcal{P}_1$) is equivalent to*

$$\arg\min_{\tilde{v} \in \mathcal{S}} \|\tilde{v}\|_1$$

where S is a subset of $\mathbb{R}^{n-1}$ based on some transformations of the initial eigenvectors of the adjacency matrix. These eigenvectors depend on the perturbation of the adjacency matrix and more specifically on the Frobenius norm of the noise matrix (Stewart and Sun and Jovanovich (1990)).

These propositions are then generalized to find the other indicators of the communities. Hence, from the adjacency matrix, the ℓ_1-spectral clustering does not directly use the subspace spanned by the first eigenvectors to find the communities but computes another eigenbasis that promotes sparse solutions for the eigenvectors. The indicators of the communities are characterized as the ones that have the minimal ℓ_1-norm with respect to a specific restricted space.

Bibliographie

Von Luxburg, U. (2007). A tutorial on spectral clustering, *Statistics and computing*, 17(4), 395–416.

Girvan, M. and Newman, E.J.M. (2002). Community structure in social and biology networks, *Proceedings of the national academy of sciences*, 99(12), 7821-7826.

Newman, E.J.M. and Girvan, M. (2004). Finding and evaluating community structure in networks, *Physical review E*, 69(2), 026–113.

Stewart, W.G. and Sun, J. and Jovanovich, B.H. (1990). Matrix pertubation theory, *Academic press New York*, 175.

Champion, C. and Brunet, C.A. and Loubes, J.M. and Risser, L. (2018). COREclust: a new package for a robust and scalable analysis of complex data, *HAL*.