Le spectral clustering

Le spectral clustering est un algorithme moderne de clustering se basant sur l'étude des valeurs et vecteur propres des graphes de Laplace et leur matrices associées.

L'algorithme utilisé pour appliquer le spectral clustering aux données mono/oligo/polyclonales est le suivant :

Normalized spectral clustering according to Ng, Jordan, and Weiss (2002)

Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct.

- Construct a similarity graph by one of the ways described in Section 2. Let W be its weighted adjacency matrix.
- Compute the normalized Laplacian L_{sym}.
- **Compute the first k eigenvectors $u_1, \ldots, u_k$ of L_{sym}.**
- Let $U \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $u_1, \ldots, u_k$ as columns.
- **Form the matrix $T \in \mathbb{R}^{n \times k}$ from U by normalizing the rows to norm 1,** that is set $t_{ij} = u_{ij} / (\sum_k u_{ik}^2)^{1/2}$.
- For $i = 1, \ldots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of T.
- Cluster the points $(y_i)_{i=1,\ldots,n}$ with the k-means algorithm into clusters $C_1, \ldots, C_k$.

Output: Clusters $A_1, \ldots, A_k$ with $A_i = \{j | \, y_j \in C_i\}$.

Source : Ulrike von Luxburg, A tutorial on Spectral Clustering, Chapter 17 Volume 4, Springer 2007

Cette méthode est par ailleurs reprise dans l'article de Alberto Paccanaro, Spectral clustering of protein sequences, Nucleic Acids Res., 2006.

Spectral methods use the leading eigenvectors of a matrix derived from the similarity information. There are various ways in which this can be done. We used a method which has been proposed recently (14) (http://www-2.cs.cmu.edu/Groups/NIPS/NIPS2001/papers/psgz/AA35.ps.gz), and was shown to give good results in a variety of difficult problems. The algorithm, depicted in Figure 1 in the Supplementary Data (Appendix file), is the following:

(i) From the affinity matrix S construct a symmetric normalized matrix $L = D^{-1/2} S D^{-1/2}$.
(ii) Find a matrix U of eigenvectors: $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K]$, corresponding to the K largest eigenvalues of L.
(iii) Build a matrix Y by renormalizing each of U's rows to have unit length: $Y_{i,j} = \frac{U_{i,j}}{(\sum_j U_{i,j}^2)^{1/2}}$.
(iv) Treating the rows of Y as points in $\mathbb{R}^K$, cluster them into K clusters using K-Means.
(v) Assign node i to cluster k if and only if row i of the matrix Y was assigned to cluster k.

Protocole :

On construit la matrice de similarité entre toute les séquences = matrice d'adjacence = dérivé de matrice de distance de Hamming ainsi que la matrice diagonale
On construit la matrice de Laplace correspondant à notre matrice symétrique et notre matrice diagonale
On détermine la matrice de vecteurs propres de la matrice de Laplace. On obtient une matrice

contenant des vecteurs colonnes propres, dont chaque ligne de cette matrice correspond à un point à clusteriser.
On utilise l'algorithme K-moyenne pour déterminer les cluster contenant chacun des points envisagé ci-dessus.

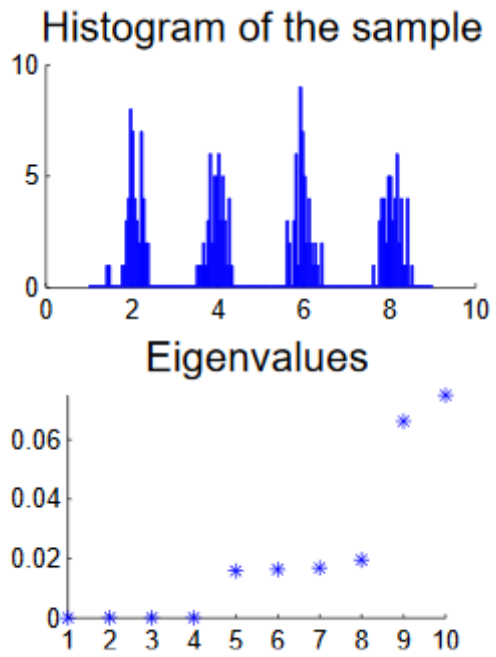
Heuristique pour choisir le nombre de clusters K :

Pour l'ensemble des algorithme de clustering, le choix du nombre de cluster à construire est toujours une étape difficile, car il n'existe pas à ce jour de méthode universelle parfaite qui fonctionne pour chaque groupe de données.

Certaines approches se basent sur l'étude de la log-likelihood des données, qui peut être traitée de manière fréquentiste ou Bayésienne. D'autre se basent sur les mesures ad-hoc telles que le ratio de similarités intra-cluster et inter-cluster, les criètes de théorie de l'information, les statistiques de gap ou encore les approches par stabilité.

Une de méthodes les plus utilisés et que nous retiendrons pour la suite est la méthode du **spectral gap**. Une fois avoir déterminé les vecteurs propres de la matrice de Laplace et les valeurs propres associées, les valeurs propres sont triées par ordre croissant et affichées sur un graphe. On admet alors l'existence d'un « gap » supérieur à la moyenne dans les valeurs propres indique le nombre de clusters supposés pour les données étudiées.

Exemple :



On observe ici l'affichage des valeurs propres du graphe de Laplace provenant de l'échantillon. On remarque que le premier gros gap en termes de valeur propre s'opère entre la 4ème et la 5ème valeur propre. Cela suggère que les données contiennent 4 clusters. On s'arrête au premier gap rencontré.