**EE378B Inference, Estimation, and Information Processing**

# Spectral clustering and stochastic block model

*Andrea Montanari* *Lecture 3 - Due on 1/25/2021*

Homework should be submitted via Gradescope, by Monday afternoon: the code will be communicated by an announcement on Canvas. This homework requires some

For getting credit for the class, you are required to present solutions of some of these homeworks during the first 15 minutes of class starting on 1/20. Please, sign up for (at least) one slot, and be sure that your explanation lasts 15 minutes (or less). For these presentations, you are free to choose whatever format you prefer (slides, typed notes, handwriting, ...).

This week, the presentations will be:

- Monday 1/25: Questions $(a)$, $(b)$, $(c)$
- Wednesday 1/27: Problem 2, points $(d)$, $(e)$, $(f)$.

**Spectral clustering**

Given a graph $G = (V = [n], E)$, let $\boldsymbol{A}_G$ be its adjacency matrix, $\boldsymbol{d}_G = \boldsymbol{A}_G\mathbf{1}$ the vector of degrees and $\hat{d} = \langle \boldsymbol{d}_G, \mathbf{1}\rangle/n$ denote its average degree. Also, denote by $\boldsymbol{D}_G = \mathrm{diag}(\boldsymbol{d}_G)$ be the diagonal matrix containing the degrees.

Spectral clustering is a general approach to generate a partition of $V$ in $k$ clusters. It has many variants and possible implementations, but at high level it proceeds as follows.

SPECTRAL CLUSTERING

**Input :** Data $G$, number of clusters $k$
**Output :** Vertex labels $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}_1, \dots, \hat{\sigma}_n)$, $\hat{\sigma}_i \in \{1, \dots, k\}$
1: Construct matrix $\boldsymbol{M}_G \in \mathbb{R}^{n\times n}$ from $\boldsymbol{A}_G$;
2: Compute the top $(k-1)$ eigenvectors $\boldsymbol{v}_1(\boldsymbol{M}_G)$, ..., $\boldsymbol{v}_{k-1}(\boldsymbol{M}_G)$;
3: Let $\boldsymbol{x}_i \in \mathbb{R}^{k-1}$ the $i$-th row of the matrix $\boldsymbol{V} \in \mathbb{R}^{n\times(k-1)}$ whose $i$-th column is $\boldsymbol{v}_i(\boldsymbol{M}_G)$;
4: Cluster points $\{\boldsymbol{x}_i\}_{i\in[n]}$ into $k$ groups, assigning labels $\hat{\sigma}_i \in \{1, \dots, k\}$;
5: Return vector $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}_1, \dots, \hat{\sigma}_n)$

Step 1 (construction of the matrix $\boldsymbol{M}_G$) and step 4 (clustering of points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^{k-1}$) admit several concrete implementation. We will consider a few of them.

- We will consider two specific choices for the matrix $\boldsymbol{M}_G$ at step 1:
  1. The centered adjacency matrix $\boldsymbol{M}_G = \boldsymbol{A}_G^{\mathrm{cen}}$, where
  $$\boldsymbol{A}_G^{\mathrm{cen}} = \boldsymbol{A}_G - \frac{\hat{d}}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}\,. \tag{1}$$
  2. The normalized (subtracted) Laplacian. Letting $\boldsymbol{z}_G = \boldsymbol{d}_G^{1/2}/\sqrt{n\hat{d}}$ (where $\boldsymbol{d}_G^{1/2}$ is the vector of square roots of degrees), we let $\boldsymbol{M}_G = \boldsymbol{L}_G$, where
  $$\boldsymbol{L}_G = \boldsymbol{D}_G^{-1/2}\boldsymbol{A}_G\boldsymbol{D}_G^{-1/2} - \boldsymbol{z}_G\boldsymbol{z}_G^{\mathsf{T}}\,. \tag{2}$$
  (We will study some interesting properties of the graph Laplacian later in the class.)
- We will consider two possibilities for the clustering step 4:

1. For $k = 2$, the points $x_1, \ldots, x_n$ generate at step 3 are real numbers. We partition them by thresholding at 0:

$$\hat{\sigma}_i = \begin{cases} 1 & \text{if } x_i \geq 0, \\ 2 & \text{if } x_i < 0, \end{cases} \tag{3}$$

2. For general $k$, we can cluster points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ using the $k$-means algorithm. (You can read about it and implement it, or read about it and use an available implementation.)

We will test spectral clustering in two different settings: (*i*) Synthetic data generated according to the stochastic block model; (*ii*) A real data set concerning political blogs.

In both cases, if $\boldsymbol{\sigma}$ denotes the true vertex labels, we will evaluate a proposed clustering $\hat{\boldsymbol{\sigma}}$ using the overlap

$$Q(\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}) \equiv \frac{k}{k-1} \max_{\pi \in S_k} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\pi(\hat{\sigma}_i) = \sigma_i\}} - \frac{1}{k} \right\} \tag{4}$$

Here $\pi : [k] \to [k]$ is a permutation of the vertex labels $1, \ldots, k$ and accounts for the fact that labels are arbitrary. $S_k$ is the set for all such permutations. Note that $Q \in [0, 1]$ with $Q \approx 1$ corresponding to nearly perfect clustering, and $Q \approx 0$ to random clusters.

**Stochastic block model**

Recall that a graph from the two-groups symmetric block model $G = (V, E) \sim \mathsf{G}(n, a/n, b/n)$ has vertices $V = [n]$ and is generated as follows. Attribute to each vertex $i \in [n]$ a true label $\sigma_i \in \{+1, -1\}$ uniformly at random. Conditional on these labels, add edges independently with probabilities

$$\mathbb{P}\big\{(i,j) \in E \big| \boldsymbol{\sigma}\big\} = \begin{cases} a/n & \text{if } \sigma_i = \sigma_j, \\ b/n & \text{if } \sigma_i \neq \sigma_j. \end{cases} \tag{5}$$

(*a*) Generate random graphs from the stochastic block model with $n = 10,000$, $b = 10$ and each of $a \in \{10, 15, 20, 25, 30\}$. In each case, plot the histogram of eigenvalues of the centered adjacency matrix $\boldsymbol{A}_G^{\text{cen}}$, along with the largest eigenvalue.

(*b*) Generate 20 random graphs for each of the choices of parameters at the previous point (hence a total of 100 instances). Apply the spectral clustering algorithm described in the previous section, with $k = 2$, $\boldsymbol{M}_G = \boldsymbol{A}_G^{\text{cen}}$ the centered adjacency matrix, and using Eq. (3) to produce the vertex labels.

Report the empirical overlap for each choice of the parameters $(n, a, b)$, averaged over the 20 samples.

(*c*) Compare your results with the theory we studied in class.

**A real dataset**

We will study a political blogs dataset first compiled for the paper

> Lada A. Adamic and Natalie Glance, *The political blogosphere and the 2004 US Election*, in Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem (2005)

The dataset polblogs contains a graph with $n = 1490$ vertices ('nodes') corresponding to political blogs. Each vertex has a $0 - -1$ label (indicated as 'value') corresponding to the political orientation of that blog. We will consider this as the true label $\sigma_i$ and try to reconstruct the vector $\boldsymbol{\sigma}$ from the graph using the spectral clustering approach introduced above.

For each of the implementations given below, report the resulting overlap $Q(\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}})$:

(*d*) Use SPECTRAL CLUSTERING with $\boldsymbol{M}_G = \boldsymbol{A}_G^{\text{cen}}$ the centered adjacency matrix and using simple thresholding (as in Eq. (3)) to produce the clusters.

(*e*) Repeat the experiment at the previous point, but replacing the centered adjacency matrix with the normalized Laplacian $\boldsymbol{L}_G$.

(*f*) As in the previous point, we use the Laplacian matrix $\boldsymbol{L}_G$. However, here we assume that the number of clusters in $G$ is estimated to be $k = 3$, and hence we use the top 2 principal components $\boldsymbol{v}_1(\boldsymbol{L}_G)$, $\boldsymbol{v}_2(\boldsymbol{L}_G)$. Produce a scatterplot of the datapoints $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^2$ produced at step 3 of the algorithm (it is useful to color-code them, using the true labels).

Use $k$-Means (with $k = 3$) to cluster the datapoints $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. How would you evaluate the resulting clustering?