Notes on Graph Spectral Clustering

Max Daniels
daniels.g@northeastern.edu

Northeastern University — June 22, 2020

1 Spectral Theorem

Theorem 1.1. *Let $A \in \mathbb{R}^{n\times n}$ be a real symmetric matrix. Then A is orthogonally diagonalizable:*

$$A = U\Sigma U^T \qquad U \text{ orthogonal and } \Sigma \text{ diagonal} \tag{1}$$

The columns of U are eigenvectors of A, and moreover, the associated eigenvalues are real.

Proof. The eigenvectors of A diagonalize A. It remains to show these eigenvectors are orthogonal and that the corresponding eigenvalues are real. □

Theorem 1.2. *The eigenvalues of a real symmetric matrix $A \in \mathbb{R}^{n\times n}$ are real.*

Proof. The characteristic polynomial of A has real coefficients by construction. Hence by (1.4), its eigenvalues must come in conjugate pairs: $P(z) = 0 \implies P(z^*) = 0$.

Let v be a nonzero eigenvector of A. By definition, $A = A^H$, so

$$\lambda \|v\|^2 = v^H A v = (v^H A v)^H = \lambda^* \|v\|^2 \tag{2}$$

It follows that $\lambda = \lambda^*$ and so $\lambda \in \mathbb{R}$. □

Theorem 1.3. *The eigenvectors of a real symmetric matrix $A \in \mathbb{R}^{n\times n}$ are orthogonal.*

Proof. There are two cases:

1. u, v are independent eigenvectors sharing an eigenvalue. Then any linear combination of u, v is another eigenvector. They may be orthogonalized with Graham Schmidt.
2. u, v are independent eigenvectors with eigenvalues λ, μ respectively. Suppose $(u^T v) \neq 0$. Then,

   $$u^T A^T A v = \lambda^2 (u^T v) = \mu^2 (u^T v) = \lambda\mu (u^T v) \tag{3}$$

   This would imply

   $$\mu(\mu - \lambda)(u^T v) = 0 \tag{4}$$
   $$\lambda(\mu - \lambda)(u^T v) = 0 \tag{5}$$

   By assumption, $(\mu - \lambda)(u^T v) \neq 0$, while $\mu = \lambda = 0$ is contradictory. It follows that $u^T v = 0$.

□

Lemma 1.4. *Let $P \in \mathbb{R}[x]$ be a polynomial with real coefficients. Then $(P(z))^* = P(z^*)$*

Proof. Key idea:

$$(-i)^k = \begin{cases} i^k & k \text{ even} \\ -i^k & k \text{ odd} \end{cases} \tag{6}$$

Take a monomial z^n where $z = a + ib \in \mathbb{C}$. Expand using the binomial theorem and collect separately the real and imaginary parts. Each term of the real part has even powers of $(ib)^k$ and is therefore unaffected. Each term of the imaginary part is negated. The property "distributes" for sums of many of these atoms, as long as they have strictly real coefficients. □

2 Notes on Spectral Clustering

Graph Spectral Clustering leverages the idea that the eigenvectors of $\mathcal{L}$, the so-called "Graph Laplacian," approximate indicator functions of clusters in the graph. In particular, $f^T \mathcal{L} f$ can (mainly by coincidence) be written like a min-cut problem on the graph.

2.1 Graph Laplacian

Definition 2.1. Let $G = (V, E)$ be a weighted graph with degree matrix D and weight matrix W. Its Laplacian is defined $\mathcal{L} = D - W$.

By coincidence, $\mathcal{L}$ yields a nice quadratic form:

$$\langle f, \mathcal{L} f \rangle = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^2 \tag{7}$$

It turns out that (7) can be manipulated into the form of an optimization objective for variations of the GraphCut problem. So, minimizing (7) is either finding the smallest eigenvectors of $\mathcal{L}$, or finding minimum cuts of a graph, depending on how you look at it.

Theorem 2.2. *Corollaries of* (7)
The Laplacian $\mathcal{L}$ satisfies:

1. *By construction $\mathcal{L}$ is real and symmetric.*
2. *Since $\langle f, \mathcal{L} f \rangle \geq 0$ for all $f \in \mathbb{R}^n$, $\mathcal{L}$ is positive semi-definite.*
3. *$\mathbb{1}_V$ is an eigenvector with eigenvalue 0.*
4. *If $A \subset V$ is disconnected from $\bar{A}$, then $\mathbb{1}_A$ and $\mathbb{1}_{\bar{A}}$ are eigenvectors with eigenvalues 0.*

Remark. The last two properties hint at the connection between 0 eigenvalues and minimal cuts between disconnected components of G. In both cases, the "objective value" is zero. We will see this relationship extends to min-cuts of non-zero cost.

Definition 2.3. Normalized Laplacian
There are two definitions of the Normalized Laplacian:

$$\mathcal{L}_{\text{rw}} = D^{-1} \mathcal{L} = I - D^{-1} W \tag{8}$$

$$\mathcal{L}_{\text{sym}} = D^{-1/2} \mathcal{L} D^{-1/2} = I - D^{-1/2} W D^{-1/2} \tag{9}$$

Remark. Note that $D^{-1}W$ has rows matching W but scaled to be probability vectors. So $\mathcal{L}_{\text{rw}}$ is associated with the *random walk* induced on G by the edge weights. $\mathcal{L}_{\text{sym}}$ is normalized in a way that maintains an interpretation like (7), and also symmetry.

Theorem 2.4. *Analogous Properties of Normalized Laplacians*
The normalized Laplacians satisfy the following:

1. *$\mathcal{L}_{sym}$ adds normalization to* (7)*:*

$$f^T \mathcal{L}_{sym} f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2 \tag{10}$$

2. *By its definition, $\mathcal{L}_{rw}$ has 0 eigenvalues for the same vectors that $\mathcal{L}$ does, so it shares the cluster intuition.*

Remark. Together, the two normalized Laplacians have all the same properties as $\mathcal{L}$, but shared in a weird way. From the Min-cut perspective they are equivalent.

2.2 Algorithms

Normalized and Unnormalized Graph Spectral Clustering

Unnormalized Graph Spectral Clustering:

1. Fix k the number of clusters
2. Compute a rank-k approximation of $\mathcal{L}$
3. These columns approximate the indicator functions for k different (possibly overlapping) clusters. Apply k-means to the rows to group the points indicated to be in the same clusters.

Normalized ($\mathcal{L}_{\text{rw}}$) Graph Spectral Clustering:

1. Fix k the number of clusters
2. Compute a rank-k approximation of $\mathcal{L}_{\text{rw}}$
3. These columns approximate the indicator functions for k different (possibly overlapping) clusters. Apply k-means to the rows to group the points indicated to be in the same clusters.

Normalized ($\mathcal{L}_{\text{sym}}$) Graph Spectral Clustering:

1. Fix k the number of clusters
2. Compute a rank-k approximation of $\mathcal{L}_{\text{sym}}$
3. Normalize each row of $\mathcal{L}_{\text{sym}}$
4. These columns approximate the indicator functions for k different (possibly overlapping) clusters. Apply k-means to the rows to group the points indicated to be in the same clusters.

RatioCut

RatioCut and NCut are *graph cut problems*. A graph cut is a partition of vertices V into disjoint sets $\{A_i\}_{i=1}^N$. The minimum cut problem seeks partitions which minimize the weight of the cut edges, or some normalized version of this. Here are some common objectives:

$$\text{cut}(A_1, ..., A_k) := \frac{1}{2}\sum_{i=1}^{k} W(A_i, \bar{A}_i) \tag{11}$$

$$\text{NCut}(A_1, ..., A_k) := \sum_{i=1}^{k} \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|} \tag{12}$$

$$\text{RatioCut}(A_1, ..., A_k) := \frac{1}{2}\sum_{i=1}^{k} \frac{\text{cut}(A_i, \bar{A}_i)}{\text{vol}(A_i)} \tag{13}$$

In the simplest case ($k = 2$ for cut), we can demonstrate the connection to $\mathcal{L}$.
The trick:

$$\text{Let } f \in \mathbb{R}^n \qquad f_i = \begin{cases} \sqrt{|\bar{A}|/|A|} & v_i \in A \\ \sqrt{|A|/|\bar{A}|} & v_i \in \bar{A} \end{cases} \tag{14}$$

The derivation:

$$f^T \mathcal{L} f = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 \tag{15}$$

$$= \frac{1}{2}\left(\sqrt{\frac{|\bar{A}|}{|A|}} + \sqrt{\frac{|A|}{|\bar{A}|}}\right)^2 \left[\sum_{i\in A, j\in\bar{A}} w_{ij} + \sum_{i\in A, j\in\bar{A}} w_{ij}\right] \tag{16}$$

$$= \text{cut}(A, \bar{A})\left(\frac{|A| + |\bar{A}|}{|A|} + \frac{|A| + |\bar{A}|}{|\bar{A}|}\right) \tag{17}$$

$$= |V| \cdot \text{RatioCut}(A, \bar{A}) \tag{18}$$

The $(f_i - f_j)$ makes this work. Any indicator function passed into $\mathcal{L}$ will cancel all terms from the same cluster, setting up the connection to cut. Symmetry of entries of f gives each coordinate the same coefficient, which factors out, so we just need to pick coefficients working out to the appropriate scale.

Hence we can minimize RatioCut by minimizing $f^T \mathcal{L} f$ under appropriate constraints. What are the constraints? Can we use them to find a recipe for this minimization problem?

1. $f^T \perp \mathbb{1}$ - check by inspection (something one might expect from symmetry)
2. $\|f\|^2 = n$ - the value n in less important than the fact that $\|f\|$ is fixed. It falls out of the scale of f to normalize cut.

Now, using this structure:

$$\min_{A \subset V} f^T \mathcal{L} f \text{ subject to } f \perp \mathbb{1},\ f_i \text{ as (14)},\ \|f\| = \sqrt{n} \qquad \text{[this is NP-Hard]} \tag{19}$$

$$\min_{A \subset V} f^T \mathcal{L} f \text{ subject to } f \perp \mathbb{1},\ \|f\| = \sqrt{n} \qquad \text{[pay discreteness for feasibility]} \tag{20}$$

Interestingly, $\mathbb{1}$ is an eigenvector of $\mathcal{L}$, so by Rayleigh-Ritz this minimization asks for the second eigenvector.

Extension of RatioCut for $k > 2$

1. Rather than $\pm c$ entries of f, we can use entries c or 0.
2. This computes the cut cost for a single A_i, rather than the cut cost for A_i and $\bar{A}_i$ at the same time.
3. So we simply compute the cost for each of the $h_i^T A_i h_i$, which can be done in a matrix.

Specifically:

$$h_i = \begin{cases} 1/\sqrt{|A_i|} & v_i \in A_i \\ 0 & else \end{cases} \tag{21}$$

$$h_i^T \mathcal{L} h_i = \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|} \tag{22}$$

$$\text{RatioCut}(A_1, ..., A_k) = \sum_{i=1}^{k} h_i^T \mathcal{L} h_i = \text{Tr}(H^T \mathcal{L} H) \tag{23}$$

Now, $f \perp \mathbb{1}$ may have seemed strange, but with the new h_i the requirement is orthogonality. There is a similar optimization setup:

$$\min_{A \subset V} \text{Tr}(H^T \mathcal{L} H) \text{ subject to } H^T H = I,\ h_i \text{ as (21)} \qquad \text{[this is NP-Hard]} \tag{24}$$

$$\min_{A \subset V} \text{Tr}(H^T \mathcal{L} H) \text{ subject to } H^T H = I \qquad \text{[pay discreteness for feasibility]} \tag{25}$$

By Rayleigh-Ritz: H has to be the k smallest eigenvectors.

Extension to NCut for $k > 2$

NCut follows the same template: construct indicator vectors h_i whose elements are normalized. Bundle these vectors into a matrix and do trace minimization against $\mathcal{L}_{\text{rw}}$ or $\mathcal{L}_{\text{sym}}$, relaxing a discreteness constraint on the indicator vectors.

A notable point: the naive setup is to weight points by cluster volume:

$$h_{i,j} = \begin{cases} 1/\sqrt{\text{vol}(A_j)} & v_i \in A_i \\ 0 & \text{else} \end{cases} \tag{26}$$

But this leads to

$$\langle h_i, h_j \rangle = \delta_{ij} \cdot \frac{|A_i|}{\text{vol}(A_i)} \tag{27}$$

Each coordinate of h_i has the same "weight" in this sum, and they need to be reweighted for the h_i to be orthonormal. That is, we need $\langle h_i, h_j \rangle_D = h_i^T D h_j = \delta_{i,j}$. *RatioCut normalization is implicitly orthormalized, while NCut needs to be weighted by node degrees, introducing the D factor that brings normalized Laplacians into the minimization problem.* The corresponding optimization has two equivalent forms:

$$\min_{K \in \mathbb{R}^{n \times k}} \mathrm{Tr}(K^T \mathcal{L} K) \text{ subject to } K^T D K = I \tag{28}$$

$$\min_{R \in \mathbb{R}^{n \times k}} \mathrm{Tr}(R^T D^{-1/2} \mathcal{L} D^{-1/2} R) \text{ subject to } R^T R = I \tag{29}$$

By (29) and Rayleigh-Ritz, R has the smallest eigenvectors of $\mathcal{L}_{\mathrm{sym}}$. Identifying $D^{-1/2} R = K$, it is clear that K has eigenvectors of $\mathcal{L}_{\mathrm{rw}}$:

$$L D^{-1/2} u_R = \lambda u_R \implies L D^{-1} u_K = \lambda u_K \implies D^{-1} L u_K = \lambda u_K \tag{30}$$

3 Probabalistic Interpretation

Consider a random walk on the graph with probabilities according to edge weights. The transition matrix is $D^{-1}\mathcal{L}$. Using Bayes' rule,

$$\Pr(\bar{A}_i | A_i) = \mathrm{NCut}(A_i, \bar{A}_i) \tag{31}$$

And so we are minimizing the probability that a random walker jumps between clusters.

4 Perturbation Approach

Matrix perturbation theory studies the impact that a small matrix perturbation can have on eigenvalues and eigenvectors of a matrix.

For graphs which are perfectly separable into k disconnected clusters, the Laplacian eigenvectors are exactly the indicator vectors for these clusters. We may view a generic Laplacian as a perturbed version of the idealized case to prove that generic Laplacian eigenvectors approximate proper indicator functions.

Theorem 4.1. *Paraphrased Davis-Kahan*

Let $A' = A + \mathcal{E} \in \mathbb{R}^{n \times n}$ be a perturbation of the symmetric matrix A by a symmetric matrix $\mathcal{E}$. Let $S \subset \mathbb{R}$ be an interval. Denote by σ_S the set of eigenvalues of A contained in S, and by V the span of the corresponding eigenvectors. Define S', V' analogously for A'.

Define the distance between S and the spectrum of A:

$$\delta = \min\{|\lambda - s| \ : \ \lambda \text{ eigenvalue of } A,\ \lambda \notin S,\ s \in S\} \tag{32}$$

Then the distance $d(V, V') = \|A - A'\|_F = \|\vec{\sigma}_A - \vec{\sigma}_{A'}\|_2$ satisfies:

$$d(V, V') \leq \frac{\|\mathcal{E}\|_F}{\delta} \tag{33}$$

Identifying δ with the spectral gap, Davis-Kahan indicates that two cases when $\mathcal{L}$ clusters well are when the spectral gap is large and when $\mathcal{L}$ (in turn, D and W) are close to that of an idealized cluster graph.

Davis-Kahan guarantees *only* that the eigenvectors of A' stay near A, under conditions. It is not sufficient to guarantee clustering unless the idealized version A is good for clustering:

1. Potential pitfall: any block diagonal matrix has "indicator-like" eigenvectors localized on the blocks
2. Also needed: eigenvalues induce an order of eigenvectors by relevance. (The first k eigenvectors indicate the first k clusters)
3. Also needed: in the idealized case, indicator components are bounded away from zero (they are either on or off). Otherwise, even a small amount of noise yields an ambiguous approximation.

The second special property does not hold for $\mathcal{L}_{\mathrm{sym}}$, whose eigenvectors are $D^{1/2}\mathbb{1}_{A_i}$. Vertices with very low degree are nearly zeroed out. This is the reason for the row normalization step in $\mathcal{L}_{\mathrm{sym}}$.

References

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec 2007. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-007-9033-z.