Graph Spectral Clustering

Shubham Makharia
shubham@brown.edu

June 24, 2020

The Spectral Theorem is a result in linear algebra that unites the characteristic eigenvectors of a linear map with its matrix representation. Recognizing linear maps with eigenvectors formalizes the intuition of a linear map acting as "scalings and rotations of standard basis vectors," and the Spectral Theorem gives a condition on the matrix representation of linear maps that recovers this idea beautifully. Any linear operator with a symmetric matrix over an orthonormal basis can be constructed precisely as a rotation of standard basis vectors, and then an individual scaling of the vectors (or vice versa).

Preliminaries

Adjoint
Let $v \in V, w \in W$. $T \in \mathcal{L}(V, W)$ can be identified with its **adjoint**, $T^* \in \mathcal{L}(W, V)$ by solving the system

$$\langle Tv, w \rangle = \langle v, T^* w \rangle \tag{1}$$

Self-Adjoint
Let $v \in V, w \in W$. $T \in \mathcal{L}(V, W)$ is self-adjoint if $T = T^*$
The definition of adjoint extends to operators $T \in \mathcal{L}(V)$ and can be found with respect to an orthonormal basis by taking the conjugate transpose of the matrix representation of T. This often leads to an abuse of notation, where * is defined to be the conjugate transpose of a matrix. This way, we recognize that operators over real vector spaces are self-adjoint iff they are symmetric matrices. This motivates building undirected graphs as objects of study in data science. The following proofs and lemmas are geared to prove statements only about real vector spaces. To extend the spectral theorem to complex vector spaces, relax the condition on T to being normal, and the spectral theorem results from the fact that every operator over a complex vector space has an eigenvalue.

1 Real Spectral Theorem

Theorem 1.1. *Suppose V is a finite-dimensional vector space taking scalars over the field $\mathbb{R}$. Let $T \in \mathcal{L}(V)$ be a linear operator. T is a self-adjoint operator $\iff$ $\exists$ an orthonormal basis of V consisting of eigenvectors of T.*

Proof. Begin by supposing that $\exists$ an orthonormal basis, $\mathcal{B}$, consisting of eigenvectors of T. Then we can diagonalize T with respect to $\mathcal{B}$ by setting diagonal entries to be the eigenvalues of each basis vector v_i, where $i = 1, ..., n$ and $\dim(V) = n$, giving us the following matrix:

$$T_{\mathcal{B}} \doteq \Sigma = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix}$$

This matrix representation of $T_{\mathcal{B}}$ is with respect to an orthonormal basis, so we can identify the adjoint $T^*_{\mathcal{B}}$ of $T_{\mathcal{B}}$ with the conjugate transpose of this matrix. Every eigenvalue is real and $T_{\mathcal{B}}$ is a diagonal matrix, thus $T_{\mathcal{B}} = T^*_{\mathcal{B}}$ and $T_{\mathcal{B}}$ is a self-adjoint operator.

To prove the other direction of this equivalence is slightly trickier, and involves proving general properties of self-adjoint operators acting on real inner product spaces. We state these properties without proof here as a claim, and defer the proof to a subsequent lemma.

- **Claim:** Self-adjoint operators have at least one eigenvalue.

Now suppose that T is a self-adjoint operator over real vector space V. Then there exists at least one eigen vector-value pair (u, λ) associated to T. Note that λ must be real, as we are working over a real vector space. Let $U = \text{span}(u)$ be a 1-dimensional subspace of V with a basis $\{u\}$.

We complete the proof via induction on the dimension of V, n. Observe that $T|_U$ is a self-adjoint operator on U with an eigenbasis $\{u\}$, proving the base case for induction. To establish the inductive hypothesis, suppose that W, a vector space of dimension less than n, has an orthonormal basis consisting of eigenvectors of T. Consider the perpendicular complement of U, $U^\perp$, a subspace of V with dimension $n-1$. By our inductive hypothesis, there is an orthonormal basis $\mathcal{B}_{U^\perp}$ of $U^\perp$ consisting of eigenvectors of $T|_{U^\perp}$. Observing that $\mathcal{B} = \mathcal{B}_{U^\perp} \cup \{u\}$ is an orthonormal basis of V consisting of eigenvectors of T completes the proof of the Real Spectral Theorem. □

Lemma 1.2. *Self-adjoint operators over real inner product spaces have at least one eigenvalue.*

Proof. Recall that $\mathcal{L}(V)$ is a vector space itself, so we can preserve the notion of exponentiation through repeated composition of an operator. Let $T \in \mathcal{L}(V)$ be self-adjoint. Consider the set $\{v, Tv, \dots, T^n v\}$, a collection of $n+1$ linearly dependent vectors in V. Linear dependence permits the following equation of 0:

$$\begin{aligned} 0 &= c_0 v + c_1 T v + \cdots + c_n T^n v && (2) \\ &= (c_0 + \cdots + c_n T^n) v && (3) \\ &\doteq P v && (4) \end{aligned}$$

P looks like a polynomial. We're also working over a real vector space, so we can recognize P with a real polynomial $P \in \mathbb{R}[x]$. P can be solved as follows:

$$\begin{aligned} 0 = P &= c_n x^n + \cdots + c_0 && (5) \\ &= c \prod_{i=1,\dots,k} (x^2 + b_i x + c_i) \prod_{j=1,\dots,l} (x - \lambda_j) && (6) \end{aligned}$$

The first product are the irreducible factors of P over $\mathbb{R}$. The irreducibility condition is enforced as $b_i^2 < 4c_i$ for every $i = 1, \ldots, k$. Applying an evaluation map of T to P results in the first product being an invertible operator constructed of invertible operators of the form $T^2 + b_i T + c_i I$. Thus, all roots of P must reside in the rightmost product, constructed of linear terms, exactly defining at least one eigenvalue of T. □

The previous lemma sufficiently proves the claim in the proof of Real Spectral Theorem. However the proof of the lemma itself made implicit two assumptions: we can freely associate a matrix P with a polynomial over the same scalar field, and P cannot have complex eigenvalues. To resolve the former, it suffices to recognize each row of Pv as a polynomial of n variables, where each variable is an element of an orthonormal basis of V, so P actually stores n^2 polynomials over the same scalar field. We resolve the latter and end this section with a neat proof about eigenvalues of self-adjoint operators.

Lemma 1.3. *Eigenvalues of self-adjoint operators are real.*

Proof. Let $T \in \mathcal{L}(V)$ be a self-adjoint operator with eigen vector-value pair (v, λ). Consider:

$$\lambda\langle v, v\rangle = \langle \lambda v, v\rangle = \langle Tv, v\rangle = \langle v, Tv\rangle = \bar{\lambda}\langle v, v\rangle \tag{7}$$

□

So What?

Consider a symmetric matrix $T \in \mathcal{L}(V)$ with respect to the standard basis. Spectral Theorem says we can find a matrix $U \in \mathcal{L}(V)$ such that each column is a unit eigenvector of T, and columns(U) is an orthonormal basis of V. We can treat U as a change of basis, rotating standard basis vectors to form an eigenbasis $\mathcal{B}_T$ of V. Now, the identity matrix

$$I_U = U = \begin{bmatrix} e_1 & \ldots & e_n \end{bmatrix} \text{ where } e_i = \text{the } i\text{'th eigenvector of } T$$

With Σ as defined in 2.1, it's clear that matrix $I_U\Sigma$ is exactly the same as the multiplication TU, because U is the eigenvector basis of V. Now we can compute how T might transform $v \in V$ recomputing v as a linear combination of eigenvectors, and then scaling each term of the combination by referencing Σ. To recompute Tv in terms of standard basis vectors apply an inverse change of basis, resulting in the well known basis decomposition of linear operators over real vector spaces:

$$T = U\Sigma U^* \tag{8}$$

2 Spectral Clustering

With the Spectral Theorem in hand, we can explore clustering methods in data analysis by creating a similarity graph of some data set, and investigate the effect of different quantifications of graph similarities on by constructing a matrix denoted the Graph Laplacian, L. The Laplacian matrix L takes on many entries, as there exist many valid definitions of it, though the commonly used matrices oft make use of the adjacency matrix W and degree matrix D consequent of constructing the similarity graph.

2.1 Graph Laplacian Matrix

Definition *L is a Laplacian matrix of a graph G iff:*

- L is symmetric and positive semi-definite
- The smallest eigenvalue of L is 0 (with eigenvector $\mathbb{1}_v$ being the constant one vector with respect to an orthonormal basis).
- If G contains disconnected components A_i, then the eigenvalue 0 corresponds to the eigenvectors $\mathbb{1}_{A_i}$
- $\forall f \in \mathbf{R}^n$,

$$f^* L f = \frac{1}{2} \sum_{(i,j)=(1,1)}^{(n,n)} w_{ij}(f_i - f_j)^2 \tag{9}$$

The relation between eigenvectors of 0 and disconnected components of the graph is key in extending the clustering problem to minimal cuts of G. When working with less than ideal graphs G, we can extend this notion to weakly connected clusters.
Three formulations of a graph laplacian are offered below.

$$L \coloneqq D - W \tag{10}$$
$$L_{sym} \coloneqq I - D^{-1/2} W D^{1/2} \tag{11}$$
$$L_{rw} \coloneqq I - D^{-1} W \tag{12}$$

The first Laplacian is known as the unnormalized laplacian, and is the most conventional definition of the laplacian. The latter two definitions are known as normalized Laplacians, and are closely related by the way $D^{1/2}$ is defined. The laplacian L_{rw} is a laplacian associated with an interpretation of clusters that takes perspective from looking at various *random walks* induced by the adjacency matrix W. It is worth noting that the left action of D^{-1} on W is strictly to normalize W, as we can calculate the transition matrix with $P = D^{-1}W$.

2.2 Algorithms for Clustering

We can describe the graph spectral clustering algorithms for all of three aforementioned laplacians with the same approach, although each implementation varies slightly. A brief discussion of these differences will follow the statement of the algorithm, for details reference "INSERT CITATION".
Graph Spectral Clustering:

1. Fix k to be the number of clusters to apply.
2. Solve an eigenvalue problem to compute a rank-k approximation of the laplacian matrix U.

3. Cluster along rows of U through an algorithm like k-means.

Why clusters generated by k-means works in step 3 is because the columns of U are approximative of the indicator functions $\mathbb{1}_{A_k}$, where each cluster is constructed as the span of these vectors, so clustering the rows of U identify which data points are of a similar coordinate with respect to the eigenvectors. When using L_{rw} it is important to note that L_{rw} is no longer symmetric (because the normalization on W), so spectral theorem no longer applies, and we have to solve a generalized eigenvalue problem to recover the conditions to apply spectral theorem. The upshot is that the conditions are recovered by applying a change of basis, essentially mapping V to $W \coloneqq D^{1/2}V$. The resulting laplacian from this change is exactly L_{sym}. So mathematically, there is no difference in the latter two uses of the Laplacian, but there are broader computational considerations to be made when deciding which route to proceed with.

Graph Cuts

A *graph cut* is a partition of a the vertices set V into disjoint sets $\{A_i\}$. Given some set of edge weights, there is a natural minimization problem that appears: Which graph cut of G minimizes the sum weight of cut edges? Three objective functions are formulated below.

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

$$\text{RatioCut}(A_1, ..., A_k) := \sum_{i=1}^{k} \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|} \tag{14}$$

$$\text{NCut}(A_1, ..., A_k) := \frac{1}{2}\sum_{i=1}^{k} \frac{\text{cut}(A_i, \bar{A}_i)}{\text{vol}(A_i)} \tag{15}$$

After some involved algebra and induction on the number of clusters k, the discrete minimization problem can be represented as the following optimization setup:

$$\min_{A \subset V} \text{Tr}(H^T \mathcal{L} H) \text{ subject to } H^T H = I \tag{16}$$

This is not exactly the optimization of the posed problem, as the exact solution requires entries of orthogonal matrix H to meet stricter conditions about entry values relative to cluster size. This restriction makes an otherwise tractable problem NP-hard.
The optimization for NCut, when edges are weighted, follow a similar recipe to produce a trace minimization problem on L_{rw} or L_{sym}. It's important to note that when working with edge weights, most setups initialize weights by cluster volume, and these result in not necessarily orthonormal matrices. To resolve this, a factor of D is introduced to allow for normalized Laplacians as the following minimization:

$$\min_{A \subset V} \text{Tr}(K^T \mathcal{L} K) \text{ subject to } K^T D K = I \tag{17}$$

The subject of this minimization can be modified to an orthogonal matrix R, by identifying $K \coloneqq D^{-1/2}R$, and defining the square root of D as $D^{1/2} = U\Sigma^{1/2}U^T$, where U is the unitary horizontal stack of k eigenvectors and Σ is the diagonal of eigenvalues.

We also recognize a probabilstic approach to the problem that is embedded in the construction of L_{rw} as minimizing the probability of a random walker leaving the cluster it began in, or walking to another cluster.

Perturbation

Matrix perturbation theory offers an analytical framework with which to understand how eigen vector-value pairs (v, λ) of a matrix A varies upon a certain degree of *perturbation*, defined as adding a bounded matrix H to A.

The Davis-Kahan theorem defines a sufficient condition for a peturbed matrix A' has similar eigenvectors. Informally, it states that the distance between the subspaces V, the span of the eigenvectors associated with the first k eigenvalues of A, and V', analagously defined w.r.t. A', is bounded by ratio of these two quantities: $\epsilon = A - A'$ and δ, given in the following equation:

$$\delta \coloneqq \min\{|\lambda - s| \ : \ \lambda \text{ eigenvalue of } A,\ \lambda \notin S,\ s \in S\} \tag{18}$$

With this bound in terms of degree of perturbation, we can relate the notion of a spectral gap to the value of δ. The sufficient conditions of Davis-Kahan do inform us that for certain matrices A, even small perturbations have meaningful impact on the eigen pairs of A'. These pathological counterexamples arise as linear operators that exhibit a small spectral gap, meaning that for a fixed k, $\lambda_{k+1} - \lambda_k$ is small. We can use the spectral gap to contravariantly identify more constraints to cluster our data set. For starters, we can seek to order the eigenvectors in orthonormal basis respective to the increasing eigenvalues, making explicit an ordering of the first k clusters. Also a word of caution: L_{sym} is not ordered with respect to eigenvalues, which naturally penalizes vertices with low degree, supporting the need for row normalization when clustering with respect to L_{sym}. But generally, finding the spectral gap from a list of eigenvalues will be very informative in telling how many reasonable clusters exist for some data.