4 Multi-layer Graphs and Grassmannian Manifolds

A multi-layer graph is a set of graphs $G_i = (E_i, \Gamma)$, $i = 1, \dots, \ell$, on a common set Γ of vertices, with each graph G_i capturing a different kind of relations between the nodes via the weighted edge set E_i. In general, clustering nodes on a single graph G_i may differ from that on another graph G_j. The problem that we would like to address in this Chapter is how to find a consensus clustering that summarizes the cluster patterns encoded in individual graphs G_i. To accomplish this task, let us reformulate spectral embedding using a single graph Laplacian L:

Find k vectors $v_i \in \mathbb{R}^m$, $i = 1, \dots, k$, such that $V = (v_1 \cdots v_k)$ satisfies $V^t V = I_{k\times k}$ and

$$\operatorname{tr}(V^t L V)$$

is minimized; the rows of V give the desired spectral embedding, unique up to an orthogonal transformation $V \mapsto VU, U \in O(k)$.

EXERCISE 4.1. *Check this claim.*

REMARK 4.1. *The condition $V^t V = I_{k\times k}$ implies that the columns of V are orthonormal.*

REMARK 4.2. *In spectral embedding, one typically chooses the first k lowest-eigenvalue eigenvectors of L as the columns of V.*

When we have ℓ different graphs on a common vertex set, each graph G_i will give rise to a graph Laplacian L_i. One possible consensus clustering would be to study:

Find k vectors $v_i \in \mathbb{R}^m$, $i = 1, \dots, k$, such that $V = (v_1 \cdots v_k)$ satisfies $V^t V = I_{k\times k}$ and

$$\operatorname{tr}\left(\sum_{i=1}^{\ell} V^t L_i V\right) = \operatorname{tr}\left(V^t \left(\sum_{i=1}^{\ell} L_i\right) V\right) \tag{4.1}$$

is minimized;

This approach, however, depends on the relative scale of the eigenvalues of L_i and may get dominated by a single graph Laplacian that has the largest set of eigenvalues. We thus add a penalty term to (4.1) based on how close the consensus space V is to the individual eigen-subspaces V_i of L_i. To accomplish this task, we need to define a set of k-dimensional subspaces in $\mathbb{R}^m$ and measure the distance between two subspaces in this set.

4.1 Stiefel and Grassmannian Manifolds

Definition 4.1 (Frame)**.** *A k-frame in $\mathbb{R}^m$ is an ordered set of k orthonormal vectors in $\mathbb{R}^m$. It is thus an ordered orthonormal basis of a k-dimensional subspace in $\mathbb{R}^m$.*

Definition 4.2 (Stiefel Manifold)**.** *The set $V(k, m)$ of all k-frames in $\mathbb{R}^m$ is called a Stiefel manifold.*

Let W be a fixed $m \times k$ matrix with orthonormal vectors along the columns; i.e. the columns of W form a k-frame. Then, all k-frames in $\mathbb{R}^m$ are related to W via an orthogonal transform:

$$\forall\, W' \in V(k,m), \exists\, U \in O(m), \text{ such that } W' = UW.$$

The choice of U is not unique, because any rotation in the subspace orthogonal to W will not affect W. In other words, we have the following correspondence:

$$V(k,m) \cong O(m)/O(m-k).$$

Definition 4.3 (Grassmannian Manifold)**.** *The set $Gr(k,m)$ of all k-dimensional subspaces in $\mathbb{R}^m$ is called a Grassmannian manifold.*

An element in $Gr(k,m)$ is obtained by identifying elements in $V(k,m)$ that are related by orthogonal transformations within the k-dimensional subspace. That is, we identify $W, W' \in V(k,m)$ if $W' = WR$ for some $R \in O(k)$, since column span(W) = column span(W') in that case. In terms of Lie groups, this identification amounts to

$$Gr(k,m) = V(k,m)/O(m) = O(m)/(O(m-k) \times O(k)).$$

Example 4.1. *Since* $\dim(O(m)) = m(m-1)/2$*, we have*

$$\dim(V(k,m)) = \dim(O(m)) - \dim(O(m-k)) = mk - \frac{k(k+1)}{2},$$

and

$$\dim(Gr(k,m)) = \dim(O(m)) - \dim(O(m-k)) - \dim(O(k)) = k(m-k).$$

4.1.1 Principal Angles via Optimization

A distance between k-dimensional subspaces in $\mathbb{R}^m$, which are now viewed as points on $Gr(k,m)$, can be obtained by defining the notion of principal angles between them.

Definition 4.4 (Principal Vectors and Principal Angles)**.** *Let $A, B \in Gr(k,m)$. The principal vectors (p_i, q_i), $i = 1, \dots, k$ are defined recursively as*

$$(p_i, q_i) = \arg \max_{\substack{p \in A, q \in B \\ p \perp p_j, q \perp q_j, j = 1, \dots, i-1 \\ \|p\|_2 = 1, \|q\|_2 = 1}} p^t q.$$

The principal angles are defined as $\theta_i = \cos^{-1}(p_i^t q_i)$, $i = 1, \dots, k$.

For subspaces $A \in Gr(k,m)$ and $B \in Gr(l,m)$ of different dimensions, we can similarly define $r = \min(k,l)$ principal vectors and angles.

EXERCISE 4.2. *Are the principal angles uniquely defined?*

4.1.2 Principal Angles via SVD

Let $A \in Gr(k,m)$ and $B \in Gr(l,m)$. Choose a k-frame and l-frame in these subspaces to represent them as $A \in \mathbb{R}^{m\times k}$ and $B \in \mathbb{R}^{m\times l}$. Then, perform the full SVD of A^tB:

$$A^tB = U\Sigma V^t$$

where $U \in O(k), V \in O(l)$ and Σ is a $k \times l$ diagonal matrix of singular values σ_i along the diagonal. As before, we assume that the singular values are ordered: $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_r \geq 0$, where $r = \min(k,l)$.

Theorem 4.1. *The principal angles are given by*

$$\theta_i = \cos^{-1}\sigma_i,\ \ i = 1,\ldots,r.$$

The corresponding principal vectors are the first r columns of $AU = (p_1 \cdots p_k)$ and $BV = (q_1 \cdots q_l)$.

Proof. Let $M_1 = A^tB$. The first pair (p_1, q_1) of principal vectors can be written as linear combinations of the columns of A and B; i.e. $p_1 = Aa_1$ and $q_1 = Bb_1$ for some $a_1 \in \mathbb{R}^k$, $b_1 \in \mathbb{R}^l$, and

$$p_1^tq_1 = \max_{a,b,\|a\|_2=1,\|b\|_2=1} a^tM_1b.$$

Using the Lagrange multiplier, we need to find the critical point of

$$\mathcal{L} = a^tM_1b - \frac{\lambda}{2}(\|a\|_2^2 - 1) - \frac{\mu}{2}(\|b\|_2^2 - 1).$$

Differentiating $\mathcal{L}$ with respect to a or b and setting the result to zero, we get

$$M_1b_1 = \lambda a_1 \ \text{ and } \ M_1^ta_1 = \mu b_1.$$

Combining the two equations, we get

$$M_1^tM_1b_1 = \lambda\mu b_1 \ \text{ and } \ M_1M_1^ta_1 = \lambda\mu a_1.$$

That is, a_1 and b_1 are left and right singular vector pairs of M_1. Because we want to maximize $a_1^tM_1b_1$, we need to choose them to be the first left and right singular vectors with the largest singular value σ_1, which is equal to $a_1^tM_1b_1 \equiv p_1^tq_1$. To find the next pair of principal vectors, we deflate M_1 by subtracting off the first SVD term and define

$$M_2 = M_1 - \sigma_1a_1b_1^t.$$

By construction M_2 and M_1 act identically on the subspace orthogonal to b_1 in B. Similarly, M_2^t and M_1^t act identically on the subspace orthogonal to a_1 in A. The only difference between M_2 and M_1 is that a_1 and b_1 are in the cokernel and kernel of M_2, respectively. Hence, maximizing a^tM_1b in the subspaces orthogonal to span$\{a_1\}$ and span$\{b_1\}$ in A and B, respectively, is equivalent to maximizing a^tM_2b in the entire A and B. Proceeding as before, we can easily see that the second pair of principal vectors should correspond to the second pair of left and right singular vectors of M_2 (and thus M_1) with singular value σ_2. Repeat this process of deflating and optimizing to find all the principal vectors as claimed. □

4.1.3 Distance between Subspaces

Using the principal angles, we can define several distance measures.

Theorem 4.2. *Let $A \in Gr(k,m)$ and $B \in Gr(l,m)$, and let $r = \min(k,l)$. Then, the following are well-defined distance measures:*

$$\begin{aligned} d_g(A,B) &= \sqrt{\sum_{i=1}^{r} \theta_i^2} && \textit{(Geodesic Distance)}, \\ d_c(A,B) &= \sqrt{\sum_{i=1}^{r} \sin^2 \theta_i} && \textit{(Chordal or Projection Distance)}. \end{aligned}$$

EXERCISE 4.3. *When $k = l$, show that the chordal distance can be written as*

$$d_c(A,B) = \frac{\|AA^t - BB^t\|_F}{\sqrt{2}} = \sqrt{k - \operatorname{tr}(AA^tBB^t)}. \tag{4.2}$$

REMARK 4.3. *The expression in (4.2) is independent of the choice of a basis.*

4.2 Multi-layer Spectral Clustering

We can now define an easy penalty term to be added to (4.1) to impose that the consensus subspace V be not too far from the individual eigen-subspaces V_i of each graph Laplacian L_i.

Let $\alpha > 0$ be a fixed number. Find k vectors $v_i \in \mathbb{R}^m$, $i = 1, \ldots, k$, such that $V = (v_1 \cdots v_k)$ satisfies $V^tV = I_{k\times k}$ and

$$\mathcal{L} = \operatorname{tr}\left(V^t \left(\sum_{i=1}^{\ell} L_i \right) V \right) + \alpha \sum_{i=1}^{\ell} d_c^2(V, V_i) \tag{4.3}$$

is minimized.

Using (4.2), we can rewrite (4.3) as

$$\begin{aligned} \mathcal{L} &= \operatorname{tr}\left(V^t \left(\sum_{i=1}^{\ell} L_i \right) V \right) + \alpha \sum_{i=1}^{\ell} \left(k - \operatorname{tr}(V^tV_iV_i^tV) \right) && (4.4) \\ &= \operatorname{tr}\left(V^t \left[\sum_{i=1}^{\ell} (L_i - \alpha\, V_iV_i^t) \right] V \right) + \alpha k \ell\,. && (4.5) \end{aligned}$$

Hence, the overall effect of the penalty term is to shift each graph Laplacian L_i by a multiple of the operator $V_iV_i^t$ that projects onto the eigen-subspace of L_i.

4.2.1 Trace Optimization

Theorem 4.3. *Let M be a symmetric $m \times m$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$. Then,*

$$\min_{V \in \mathbb{R}^{m \times k},\ V^t V = I_{k \times k}} \operatorname{tr}(V^t M V) = \sum_{i=1}^{k} \lambda_i.$$

A particular solution to the minimization is $V = (v_1 \cdots v_k)$, where v_i are the eigenvectors of M corresponding to λ_i.

Proof. Let V be any $m \times k$ matrix, and denote the subspace spanned by its columns as S. Because the trace is invariant under conjugating $V^t M V$ with an orthogonal matrix $U \in O(k)$, we can choose any frame in S in place of V to evaluate the trace. Choose v_k such that

$$v_k = \arg \max_{v \in S, \|v\|_2 = 1} v^t M v$$

Iterating this process, choose v_i such that

$$v_i = \arg \max_{v \in S,\ \|v\|_2 = 1,\ v \perp v_{i+1}, \ldots, v_k} v^t M v.$$

Then, $v_1, \cdots, v_k$ form a k-frame in S and

$$\operatorname{tr}(V^t M V) = \sum_{i=1}^{k} v_i^t M v_i.$$

But, by the Courant-Fisher Theorem (Theorem A.8), we have

$$v_i^t M v_i \geq \lambda_i.$$

Hence, for any $m \times k$ matrix V satisfying $V^t V = I_{k \times k}$, we have $\operatorname{tr}(V^t M V) = \sum_{i=1}^{k} v_i^t M v_i \geq \sum_{i=1}^{k} \lambda_i$. The inequality is clearly saturated if we choose v_i to be a unit eigenvector of M corresponding to λ_i. □

4.3 Pareto Multi-objective Optimization

An alternative way of performing multi-layer graph clustering is to construct not just one, but a set of several comparable clustering results. Given a specific partition C of nodes into k clusters, let $\mathcal{L}_1(C), \mathcal{L}_2(C), \ldots, \mathcal{L}_\ell(C)$ denote the loss functions associated with the graphs $G_1, G_2, \ldots, G_\ell$, respectively. We would like to simultaneously minimize these loss functions, but there will be a trade-off between different loss functions. Let $S \subset \mathbb{R}^\ell$ denote the space of all values that these ℓ loss functions can take. Then, we introduce a partial ordering on this set as follows: for any $a = (a_1, \ldots, a_\ell) \in S$ and $b = (b_1, \ldots, b_\ell) \in S$, we say that b dominates a, written as $a \prec b$, if $\forall\, i \in \{1, \ldots, \ell\}$, $a_i \geq b_i$, and for at least one $i \in \{1, \ldots, \ell\}$, $a_i > b_i$. The set of all points in S that are not dominated by any other point in S is called the Pareto front. The partitions of nodes that give rise to this Pareto front constitute our candidate consensus clustering results. Sampling partitions can be done via discrete optimization algorithms, some of which we will study in the second half of this course.