COMS 4774 Unsupervised Learning Fall 2024
Problem Set #1

Brian Chen - bc2924@columbia.edu

September 19, 2024

Problem 1

Comparing Clusterings – An Axiomatic View [Meila, 2005]

The paper provides an overview of various ways of determining "clustering quality" and other ways of quantifying the "distance" between different clusterings. It then derives an impossibility theorem on clusterings that proves that no clustering can satisfy a set of separate but desirable properties at the same time. This result is important because it 1) characterizes properties of distance functions between clusters, 2) speaks to the ways that we can (and cannot) be re-embedded into some other spaces without breaking the lattice structure of the original partition graph. By reinterpreting clustering as a question of graph properties off of the lattice built from all possible partitions of a dataset, we can apply graph theory techniques to investigate the relationships between clusterings and functions that interpret the "distance" between different clusterings.

We define a *clustering* of a dataset D as a set of K nonempty partitions $C_1, \ldots, C_K$ such that

$$C_k \cap C_l = \emptyset, \qquad \bigcup_{k=1}^{K} C_k = D$$

We want to define a "distance" $d(\cdot, \cdot)$ between two clusterings C, C' that obeys some intuitive properties, e.g. that $d(C, C') = 0$ if $C = C'$.

Definition 1. A clustering distance function is **additive with respect to refinement** if for clusterings C, C', C'', where C' is a refinement of C and C'' is a refinement of C', that

$$d(C, C'') = d(C, C') + d(C', C'')$$

In the Hasse diagram representation of a set, this corresponds to a downwards traversal.

Definition 2. A clustering distance function is **additive with respect to the join** for a join defined as

$$C \times C' = \{C_k \cap C'_{k'} \mid C_k \in C, C'_{k'} \in C', C_k \cap C'_{k'} \neq \emptyset\}$$

if for any clusterings C, C',

$$d(C, C') = d(C, C \times C') + d(C', C \times C')$$

Definition 3. Let C be a clustering and C', C'' be refinements of C. Let $C'_k(C''_k)$ be the partitioning induced by C' on C_k, and let $P(k)$ represent the proportion of points in C_k. For a distance defined as

$$d(C', C'') = \sum_{k=1}^{K} P(k) d(C'_k, C''_k)$$

d has **convex additivity** if

$$d(C, C') = d(\tilde{C}, \tilde{C}').$$

The question now becomes: how can we define a "distance" function of clusterings that has all three of these properties? The properties of AV, CA, and AJ represent *iterative* steps that one could take to transform one clustering into another. In a very intuitive sense, then, d represents the "distance" between clusterings as a function of how many steps it takes to transform one into another.

Define the *variation of information* distance (d_{VI}) as

$$d_{VI}(C, C') = H(C) + H(C') - 2I(C, C')$$

where $H(C) = -\sum_{k=1}^{K} \frac{n_k}{n} \log \frac{n_k}{n}$, $I(C, C') = \sum_{k=1}^{K} \sum_{k'}^{K'} \frac{n_{k,k'}}{n} \log \frac{n_{k,k'}}{n} \frac{n_k}{n} \frac{n'_{k'}}{n}$, in other words the entropy and mutual information between the clusterings, respectively.

Lemma 1. *The variation of information distance satisfies additivity with respect to refinement.*

Proof. By [Meila, 2003] we can represent the distance as the sum of conditional entropies,

$$d_{VI}(C, C') = H(C \mid C') + H(C' \mid C)$$

Since C' is a refinement of C, $H(C \mid C') = 0$. Then by the chain rule for conditional entropy, we have

$$H(X, Y) = H(X) + H(Y \mid X)$$

where $H(X, Y) = -E \log p(X, Y)$ is the joint entropy of X and Y, and $H(Y \mid X) = -E \log p(Y \mid X)$ is the conditional entropy of X and Y. Then we have

$$\begin{aligned} d(C, C'') &= d(C, C') + d(C', C'') \\ &= H(C \mid C') + H(C' \mid C) + H(C' \mid C'') + H(C'' \mid C') \\ &= H(C' \mid C) + H(C'' \mid C') \\ &= H(C', C) - H(C) + H(C'', C) - H(C') \\ &= H(C \mid C'') + H(C'' \mid C) \end{aligned}$$

□

The paper then states that d_{VI} is uniquely defined by satisfying certain axioms.

Theorem 1. *d_{VI} uniquely satisfies (1)* ***symmetry****, (2)* ***additivity w.r.t. refinement****, (3)* ***additivity w.r.t. join****, (4)* ***convex additivity****, and (5)* ***scale****.*

Proof. (2) was proven in Lemma 1 as a straightforward application of the definition of refinement additivity. Then we have

$$\begin{aligned} d(\hat{0}, C) &= \sum_k \frac{n_k}{n} \log n_k \\ &= \log n - H(C) \end{aligned}$$

and similarly

$$d(\hat{1}, C) = H(C)$$

where $\hat{0}$ is the clustering with n clusters (every point in its own cluster), and $\hat{1}$ the clustering with 1 center (every point in the same cluster). Then

$$\begin{aligned} d(C, C \times C') &= \sum_{k=1}^{K} \frac{n_k}{n} d(\hat{1}, C_k) \\ &= H(C \mid C') \end{aligned}$$

and therefore

$$d(C, C') = H(C \mid C') + H(C' \mid C) = d_{VI}(C, C')$$

by the AJ property of (3). □

The author notes that defining clustering comparisons via their axioms is an interesting approach, and proceeds to do a similar analysis on several other metrics and distance functions for cluster comparison.

Theorem 2. *No cluster comparison function satisfies (1)-(4) and $d(\hat{1}, C_K^U) = 1 - 1/K$.*

Proof.

$$\begin{aligned} d(\hat{0}, C) &= \sum_k \frac{n_k}{n} d(\hat{1}, C_{n_k}^U) \\ &= \sum_k \frac{n_k}{n} \left(1 - \frac{1}{n_k}\right) \\ &= 1 - \frac{K}{n} \end{aligned}$$

which implies

$$\begin{aligned} d(\hat{1}, C) &= (1 - \frac{1}{n}) - (1 - \frac{K}{n}) \\ &= \frac{K-1}{n} \end{aligned}$$

for $|C| = K$. However, this is a contradiction with the theorem statement that $d(\hat{1}, C_K^U) = 1 - 1/K$. □

Note that the condition in (5) that we are changing effectively controls the *scale* of d, e.g. the logarithmic growth rate of d. The main theorem of the paper thus investigates under what conditions can we generate a distance function under a particular choice of the scaling function $d(\hat{1}, C_K^U) = h(K)$.

Theorem 3. *(Impossibility) Any clustering comparison function satisfying (1)-(4) and* $d(\hat{1}, C_K^U) = h(K)$ *is identical to* d_{VI} *up to a constant coefficient.*

Proof.

$$\begin{aligned} d(\hat{1}, \hat{0}) &= h(K) \text{ by definition} \\ d(\hat{1}, C) &= h(n) - d(\hat{0}, C) \text{ by (2)} \\ d(\hat{1}, C_K^U) &= h(n) - d(\hat{0}, C_K^U) \\ &= h(n) - h\left(\frac{n}{K}\right) \end{aligned}$$

Since n/K is an integer, we effectively have a distance that is the sum of integer inputs to the same scaled distance function h,

$$d(\hat{1}, C_K^U) = h(K) + h(M)$$

which implies that $h(n) = C \log n$, e.g. that h is just the original scaling from (5) with a constant coefficient. □

Because of this, d_{VI} is the only sensible measure of cluster dissimilarity under the conditions discussed at the beginning of the paper.

Hartigan's Method: k-means Clustering without Voronoi [Telgarsky, Vattani 2010] Hartigan's method gives a point-by-point method for k-means

Note the bias-variance tradeoff for k-means cost function

$$\phi(C,z) = \phi(C) + C\|\mu(C) - z\|^2 \tag{1}$$

Lemma 2. *The cost of merging two clusters A, B is*

$$\Delta(A,B) = \frac{AB}{A+B}\|\mu(A) - \mu(B)\|^2 \tag{2}$$

for $\Delta(A,B) = \phi(A \cup B) - \phi(A) - \phi(B)$

Proof.

$$\Delta(A,B) = \phi(A \cup B) - \phi(A) - \phi(B)$$

Note that if either A or B are empty, then $AB = 0$ and the merging cost is thus 0. □

We now consider the direct formulation of Hartigan's method in terms of choosing points to move between clusters according to how it improves the overall cost. Note that we can write this in terms of the merging cost function (2).

Lemma 3. *For clusers S, T, and a point $x \in S$, the cost of moving x to T is*

$$\Phi(x;S,T) = \frac{T}{T+1}\|\mu(T) - x\|^2 - \frac{S}{S-1}\|\mu(S) - x\|^2 \tag{3}$$

Proof. Using (2) we get

$$\begin{aligned}\Phi(x;S,T) &= \phi(S) + \phi(T) - \phi(S\backslash\{x\} - \phi(T \cup \{x\}) \\ &= -\frac{T}{T+1}\|\mu(T) - x\|^2 + \frac{S}{S-1}\|\mu(S) - x\|^2\end{aligned}$$

□

The standard Hartigan's method formulation thus takes the form

```
Algorithm 1: Hartigan's method (1)
    Data: points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, number of clusters $k$
    Result: cluster assignments of $X$
    while $\exists x, S, T \; s.t. \Delta(x, S, T) > 0$ do
        choose an $x_i$ with cluster label $y_i$ according to some SELECT function
        choose a new cluster to assign $x_i$ to according to $\arg\max_j \Phi(x_i; C_{y_i}, C_j)$
        update labels and means for the affected clusters
    end
    return cluster assignments
```

We note immediately that the actual choice of point to update is parameterized by the SELECT function. The paper notes that this can either be an in-order traversal of the original datapoints or a greedy selection of the point that would lower the cost the most.

There also exists a reformulation of the algorithm that rewrites the loop termination condition as

$$\exists i, j \quad s.t. \quad \|\mu(C_j) - x_i\| < \alpha(C_y, C_j)\|\mu(C_{y_i} - x_i\|$$

Based on these, the paper proposes a theorem on the properties of Hartigan's method.

Theorem 4. *For Hartigan's method,*
(1) The cost sequence of the method is strictly decreasing.
For $n \geq k$ distinct points,
(2) The resulting partition has no empty clusters.
(3) The resulting partition has distinct means.

Proof. (1) holds trivially as the update condition only happens if we find another cluster C_j such that the cost is lowered, e.g. the "improvement condition" from eq. (3) is positive. We terminate if this doesn't exist, so clearly the cost goes down.
(2) Assume that we finished partitioning the points, and we ended with an empty cluster. We could assign any two points from any other cluster to the empty cluster for free (since the cost eq.(3) would be zero due to the size of the empty cluster). But this clearly would decrease the entire cost, since the cost of the cluster the points were removed from would decrease. This contradicts the assumption that we had finished partitioning the points.
(3) Assume that the algorithm terminated with $\mu(S) = \mu(T)$ and all the points are distinct. Let $\alpha(S,T) = \sqrt{S(T+1)/(T(S-1))}$. Since $\alpha(S,T) > 1$, then we have $\|\mu(T) - x\| = \|\mu(S) - x\| < \alpha(S,T)\|\mu(S) - x\|$ which means that there is another swap that could be made to reduce the cost. Therefore we have a contradiction. □

Theorem 5. *The local optima of Hartigan's method is a subset of the local optima of Lloyd's method.*

Proof. Hartigan's method only continues if there is some swap that lowers the cost, e.g. if

$$\|\mu(C_j) - x_i\| < \|\mu(C_{y_i} - x_i)\|\sqrt{\frac{C_{y_i}(C_j + 1)}{C_j(C_{y_i} - 1)}}$$

The square root is always > 1 for $C_{y_i}, C_j \geq 1$, so equivalently it terminates when

$$\|\mu(C_{y_i} - x_i)\| < \|\mu(C_j) - x_i\|$$

which means that all points were assigned to its nearest mean. Therefore, k-means does not iterate either. □

The main theorem of section 3 is the following.

Theorem 6. *Given a k-means instance in* $[0,1]^d$*, consider a* (m, ϵ)*-Voronoi and circlonoi partition. Let* Γ *be the volume of space in* $[0,1]^d$ *that is not inside any circlonoi cell. Then*

$$Vol(\Gamma) \geq \left(\frac{1}{e}\right)^{\Theta(\epsilon d)} - \left(\frac{1}{e}\right)^{\Theta(d/m)}$$

Proof. Define

$$\epsilon_S = \frac{1}{2} \min_{T \in adj(S)} \|\mu(S) - \mu(T)\|$$

as the minimum distance from the center of S to the boundary of an adjacent cell. If we scaled down S, we would retain the same center S', so we have

$$\epsilon_{S'} = \epsilon_S(1 - \frac{1}{m}) = \epsilon_S - \frac{1}{2} \min_{T \in adj(S)} \frac{\|\mu(S) - \mu(T)\|}{m}$$

Therefore, the circlonoi cell of S must be within S' if S is not adjacent to the hypercube's face. Otherwise, if it is adjacent, we have

$$\sum_S Vol(C_S) \leq \sum_S (Vol(S') + Vol(B_S))$$

$$= \sum_S Vol(S)(1 - \frac{1}{m})^d + \sum_S Vol(B_S)$$

$$= (1 - \frac{1}{m})^d + (1 - (1 - 2\gamma)^d)$$

from which we can conclude that

$$Vol(\Gamma) \geq \left(1 - 2\epsilon\left(1 - \frac{1}{m}\right)\right)^d - \left(1 - \frac{1}{m}\right)^d$$

since $Vol(\Gamma) = 1 - \sum_S Vol(C_S)$. Then we notice that for any $m \geq 2$, $\gamma = \Theta(\epsilon)$, so using the fact that $(1 - \frac{1}{x}) = \frac{1}{e}^{\Theta(1/x)}$ we arrive at

$$\text{Vol}(\Gamma) \geq \left(\frac{1}{e}\right)^{\Theta(\epsilon d)} - \left(\frac{1}{e}\right)^{\Theta(d/m)}$$

□

Problem 2

Given a graph $G = (V, E)$ construct a metric as such over X:

$$d(u, v) = \begin{cases} 0, & u = v \\ 1, & uv \in E \\ 2, & uv \notin E \end{cases}$$

Then, if G has a dominating set of size K, it corresponds to a solution of the k-centers problem of 1, since all points are at most 1 away from its nearest center. Alternatively, if G does not have a dominating set of size K, then the corresponding vertices have a k-centers solution of > 1. Therefore there is a one to one correspondence between solutions of the k-center problem and a decision on the dominating set problem, so the reduction is complete.

Problem 3

Part (i)

Intuitively, an ϵ-cover corresponds to a set $C \subset X$ such that if we centered an ϵ-ball (with respect to the distance function d), all points are covered by the union of all of the balls. Likewise, an ϵ-packing can be defined intuitively as a sphere packing problem, where $P \subset X$ is an ϵ-packing if each ϵ-ball does not overlap.

We can construct the subset Y greedily to make it both an ϵ-cover and an ϵ-packing. Without loss of generality add the first point in X to Y. Then for every subsequent point in X, add the point if and only if it $d(x_i, y_i) > \epsilon \;\forall\; y_i \in Y$. Repeat this until all points in X have been considered.

The greedy construction is guaranteed to be an ϵ-cover because we ever encountered a point x in X that had $d(x, y) > \epsilon \;\; \forall y \in Y$, we would add it to the set Y.

The greedy construction is also guaranteed to be an ϵ-packing because we add a new point to Y only if it is at least $d(x_i, y_i) > \epsilon \;\forall y_i \in Y$, which is the definition of ϵ-packing.

Part (ii)

First we show the second half of the inequality.

Lemma 4.

$$N_\epsilon(X) \leq P_\epsilon(X)$$

where $N_\epsilon(X)$ and $P_\epsilon(X)$ are the ϵ-covering number and ϵ-packing number of X, respectively.

Proof. Consider the set $X \backslash P$, where P is the maximal ϵ-packing. Then by definition

$$\exists p_i \in P \text{ s.t. } d(x, p_i) \leq \epsilon \;\forall x \in X$$

because if this was not the case, then we could construct a larger ϵ-packing by adding the point x to P. Therefore, P is also an ϵ-cover. We defined $N_\epsilon(X)$ as the *minimal* size covering number, so clearly

$$N_\epsilon(X) \leq P_\epsilon(X)$$

□

Now we show the first half of the inequality, where I use $\epsilon \to 2\epsilon$ and $\epsilon/2 \to \epsilon$.

Lemma 5.

$$P_{2\epsilon}(X) \leq N_\epsilon(X)$$

for ϵ-covering number $N_\epsilon(X)$ and 2ϵ-packing number $P_{2\epsilon}(X)$.

Proof. By contradiction. Let $P = \{p_1, \dots, p_N\}$ be the 2ϵ-packing and $C = \{c_1 \dots, c_M\}$ be the ϵ-packing, and assume for same of contradiction $M < N \to M + 1 \leq N$. By the pigeonhole principle there must be some $x \in X, p_i, p_j \in P$ such that $d(x, p_i) \leq \epsilon$ and $d(x, p_j) \leq \epsilon$. However this implies that $d(p_i, p_j) \leq 2\epsilon$ by the triangle inequality, which contradicts the assumption that P is a 2ϵ-packing. □

Part (iii)

For $B^d(x, r)$ the volume of the closed Euclidean ball of radius r centered at x in $\mathbb{R}^d$, we have by definition of an ϵ-covering that

$$\text{vol}(B^d(0,1)) \leq \text{vol}\left(\bigcup_{i=1}^{N} B^d(x_i, \epsilon)\right)$$

$$\leq \sum_{i=1}^{N} \text{vol}(B^d(x_i, \epsilon))$$

$$\leq \sum_{i=1}^{N} \epsilon^d \text{vol}(B^d(x_i, 1))$$

$$1 \leq N\epsilon^d$$

$$\Rightarrow N \geq \frac{1}{\epsilon^d}$$

Similarly we can upper bound the covering number by using Lemma 5, taking the $\epsilon/2$ packing of the unit ball:

$$P_\epsilon(X) \leq N_{\epsilon/2}(X)$$

$$\text{vol}(B^d(0, \epsilon/2))N_\epsilon(X) \leq \text{vol}(B^d(0, 1 + \epsilon/2))$$

$$(\epsilon/2)^d \text{vol}(B^d(0,1))N_\epsilon(X) \leq (1 + \epsilon/2)^d \text{vol}(B^d(0,1))$$

$$\Rightarrow N_\epsilon(X) \leq \frac{(1 + \epsilon/2)^d}{(\epsilon/2)^d} = \left(1 + \frac{2}{\epsilon}\right)^d$$

which gives us the full inequality

$$\boxed{\left(\frac{1}{\epsilon}\right)^d \leq N_\epsilon(X) \leq \left(1 + \frac{2}{\epsilon}\right)^d}$$

Part (iv)

We can construct an approximation of the maximum singular value by 1) finding a cover of S^{n-1} and then 2) testing a (finite number of) points corresponding to each element of the cover to find the maximum singular value.

First, notice that we trivially lower bound the maximum singular value by checking points only in the ϵ-covering since this is a strict subset of the ball. Therefore,

$$\max_{x \in N_\epsilon} \|Ax\| \leq \max_{x \in S^{n-1}} \|Ax\|$$

We know that we need to evaluate at $x \in S^{n-1}$ with an approximation $y \in N_\epsilon$. Let the approximation of x be on order ϵ, e.g. $\|x - y\| \leq \epsilon$. Then we have

$$\|Ax - Ay\| = \|A\|\|x - y\|$$

$$= \|A\|\epsilon$$

In addition, by the triangle inequality

$$\|Ay\| \geq \|Ax\| - \|Ax - Ay\|$$

$$\geq \|A\| - \epsilon\|A\|$$

$$\geq (1 - \epsilon)\|A\|$$

$$\Rightarrow \|A\| \leq \frac{1}{1-\epsilon} \max_{x \in N_\epsilon} \|Ax\|$$

which gives us the full bounds:

$$\boxed{\max_{x \in N_\epsilon} \|Ax\| \leq \|A\| \leq \frac{1}{1-\epsilon} \max_{x \in N_\epsilon} \|Ax\|}$$

Problem 4

Part (i)

Assuming that the data is Euclidean in $\mathbb{R}^n$, we have

$$D_{ij} = \|\alpha_i - \alpha_j\|^2$$

$$= \langle \alpha_i, \alpha_i \rangle + \langle \alpha_j, \alpha_j \rangle - 2\langle \alpha_i, \alpha_j \rangle$$

which implies

$$\langle \alpha_i, \alpha_j \rangle = -\frac{1}{2}\left[\|\alpha_i - \alpha_j\|^2 - \langle \alpha_i, \alpha_i \rangle - \langle \alpha_j, \alpha_j \rangle\right]$$

Since we are working with centered points, we have

$$\langle \alpha_i - \overline{\alpha}, \alpha_j - \overline{\alpha} \rangle = -\frac{1}{2}\left[\|\alpha_i - \alpha_j\|^2 - \langle \alpha_i - \overline{\alpha}, \alpha_i - \overline{\alpha} \rangle - \langle \alpha_j - \overline{\alpha}, \alpha_j - \overline{\alpha} \rangle\right]$$

We immediately notice similarities with the Gram matrix $-\frac{1}{2}H^\intercal DH$. Namely, since the coefficient in the front is the same, we just need to show that the i,jth entry of $H^\intercal DH$ equal the i,jth entry of the expression in the bracket.

Defining $H = I - \frac{1}{n}\mathbb{1}\mathbb{1}^\intercal$, the i,jth entry of $H^\intercal DH$ is

$$H^\intercal DH = (I - \frac{1}{n}\mathbb{1}\mathbb{1}^\intercal)D(I - \frac{1}{n}\mathbb{1}\mathbb{1}^\intercal)$$

$$= D - \frac{1}{n}\mathbb{1}\mathbb{1}^\intercal D - \frac{1}{n}D\mathbb{1}\mathbb{1}^\intercal + \frac{1}{n^2}\mathbb{1}\mathbb{1}^\intercal D\mathbb{1}\mathbb{1}^\intercal$$

Note that the rows of $-\frac{1}{n}D\mathbb{1}\mathbb{1}^T$ are comprised of the average of the i-th *rows* of D, and the columns of $\frac{1}{n}\mathbb{1}\mathbb{1}^T D$ are comprised of the averages of the i-th *columns* of D. The matrix $\frac{1}{n^2}\mathbb{1}\mathbb{1}^\intercal D\mathbb{1}\mathbb{1}^\intercal$ is trivially the $n \times n$ matrix where every element is the *average* of all of the elements in the matrix. Thus, the i,jth component of the resultant matrix is

$$\left[-\frac{1}{2}H^\intercal DH\right]_{ij} = -\frac{1}{2}\left[D_{ij} - \langle \alpha_i - \overline{\alpha}, \alpha_i - \overline{\alpha} \rangle - \langle \alpha_j - \overline{\alpha}, \alpha_j - \overline{\alpha} \rangle\right]$$

$$= -\frac{1}{2}\left[\|\alpha_i - \alpha_j\|^2 - \langle \alpha_i - \overline{\alpha}, \alpha_i - \overline{\alpha} \rangle - \langle \alpha_j - \overline{\alpha}, \alpha_j - \overline{\alpha} \rangle\right]$$

which is the form that we were looking for.

Part (ii)

Lemma 6. *B is positive semidefinite.*

Proof. For any vector x, we need $x^T Bx \geq 0$ for B to be positive semidefinite. Let b_i be the columns of B. Then, we have

$$x^T Bx = \sum_{i,j} x_i \langle b_i, b_j \rangle x_j$$

by the definition of matrix multiplication. Since the inner product $\langle b_i, b_j \rangle$ is just a number, it commutes with the vectors x_i and x_j.

$$= \sum_{i,j} \langle x_i b_i, x_j b_j \rangle$$

$$= \langle \sum_i x_i b_i, \sum_j x_j b_j \rangle$$

$$= \left\| \sum_i x_i b_i \right\|^2$$

$$\geq 0$$

by the nonnegativeness of the norm. Thus, B is positive semidefinite[1]. We are dealing with real numbers only (e.g. distances), so we don't need to take the Hermitian conjugate, just the regular transpose of x. □

Since B_{ij} is positive semidefinite, it can be decomposed into

$$B = X^T X$$

. In addition, since it is symmetric it thus has an eigendecomposition

$$B = Q\Lambda Q^T$$

where Q is the matrix whose columns are the eigenvectors of B. Then, since B is positive semidefinite, its eigenvalues are all nonnegative, so we can take a square root of all the eigenvalues. We see that

$$B = Q\Lambda^{1/2}\Lambda^{1/2}Q^T$$

$$X^T X = (Q\Lambda^{1/2})(Q\Lambda^{1/2})^T$$

$$\Rightarrow X = Q\Lambda^{1/2}$$

[1] I referenced the Wikipedia page for Gram matrices: `https://en.wikipedia.org/wiki/Gram_matrix`, which actually does this calculation explicitly, as well as the page for positive semidefinite matrices `https://en.wikipedia.org/wiki/Positive_semidefinite`, which I have been referencing all semester.

Note that the effect of multiplying a matrix by a diagonal matrix is just to multiply the i-th column of Q with the i-th eigenvalue, e.g.

$$Q\Lambda^{1/2} = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n1} & q_{n2} & \cdots & q_{nn} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix} = \begin{bmatrix} q_{11}\sqrt{\lambda_1} & q_{12}\sqrt{\lambda_2} & \cdots & q_{1n}\sqrt{\lambda_n} \\ q_{21}\sqrt{\lambda_1} & q_{22}\sqrt{\lambda_2} & \cdots & q_{2n}\sqrt{\lambda_n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n1}\sqrt{\lambda_1} & q_{n2\sqrt{\lambda_2}} & \cdots & q_{nn}\sqrt{\lambda_n} \end{bmatrix}$$

The i-th row of this matrix is

$$\begin{bmatrix} q_{i1}\sqrt{\lambda_1} & q_{i2}\sqrt{\lambda_2} & \cdots & q_{in}\sqrt{\lambda_n} \end{bmatrix}$$

and thus the distance between any two points (without loss of generality, indices 1 and 2) here is

$$= \left\| q_{i1}\sqrt{\lambda_1} - q_{i2}\sqrt{\lambda_2} \right\|^2$$

$$= \lambda_1 q_{i1}^2 - \lambda_2 q_{i2}^2$$

which by definition is the original distance embedding

$$= \langle \alpha_i - \overline{\alpha}, \alpha_j - \overline{\alpha} \rangle$$

$$= \|\alpha_i - \alpha_j\|^2$$

$$= D_{ij}$$

Part (iii)

We aim to

$$\text{minimize } \sum_{i \neq j}^{n} (D_{ij} - \|x_i - x_j\|^2)^2$$

Via spectral decomposition, we can decompose D_{ij} and our learned distances x_{ij} into $D = Y\Lambda_D Y^T$, $x_{\text{matrix}} = X\Lambda_x X^T$. Going into $\mathbb{R}^{n\times n}$ space means we can transform the norms into a trace[2], namely that

$$\|A\|_F = \sqrt{\sum_i^m \sum_j^n |a_{ij}|^2} = \sqrt{\text{Tr}(A^* A)}$$

where A^* denotes the conjugate transpose (or just the transpose for real matrices). Our minimization problem becomes

$$\text{minimize}_{X,\Lambda_X} \text{Tr}(Y\Lambda_D Y^T - X\Lambda_X X^T)^2$$

$$= \text{minimize}_{X,\Lambda_X} \text{Tr}(\Lambda_D - Y^T X \Lambda_X X^T Y)^2$$

[2] `https://en.wikipedia.org/wiki/Matrix_norm`, specifically the section on the Frobenius norm

Letting $Q = Y^T X$ for notational convenience, we have

$$= \text{minimize}_{Q,\Lambda_X} \text{Tr}(\Lambda_D - Q\Lambda_X Q^T)^2$$

$$= \text{minimize}_{Q,\Lambda_X} \text{Tr}(\Lambda_D^2 - 2Q\Lambda_X Q^T + (Q\Lambda_X Q^T)^2)$$

$$= \text{minimize}_{Q,\Lambda_X} \text{Tr}(\Lambda_D^2 - \Lambda_X\Lambda_D + \Lambda_X^2)$$

$$= \text{minimize}_{Q,\Lambda_X} \text{Tr}(\Lambda_D - \Lambda_X)^2$$

since the minimizing values of Q do not change whether we apply H^T and H to the normed terms.

Since these Λ matrices are diagonal matrices with eigenvalues along the diagonal, the trace is clearly minimized by taking the k largest eigenvalues and eigenvectors in Λ_X.

Problem 5

Recall from PSET0 that if Y is a chi-squared distributed random variable with n degrees of freedom, $Y = X_1^2 + \cdots + X_n^2$ with X i.i.d. variables drawn from the standard normal distribution, then we have

$$\Pr\left[Y > (1+\epsilon)^2 n\right] < e^{-cn\epsilon^2}$$

Now note that

$$\|Gv\|^2 = v^\mathsf{T} G^\mathsf{T} G v$$

Since $G_{ij} = \mathcal{N}(0, 1/d)$,

$$(G^\mathsf{T} G)_{ij} \sim d\chi_d^2$$

since every term in $G^\mathsf{T} G$ is the sum of d squared normals. Now to complete the proof, we make use of the following result:

Lemma 7. *For unit vector $v \in \mathbb{R}^d$ and a $d \times d$ matrix M whose entries are all i.i.d random variables drawn from a chi-squared distribution, $v^\mathsf{T} M v \sim d\chi_d^2$ and is a scalar value.*

Proof.

$$v^\mathsf{T} M v = \sum_{i=1}^{d} \sum_{j=1}^{d} v_i M_{ij} v_j$$

$$= \sum_{i=1}^{d} v_i \left(\sum_{j=1}^{d} v_j M_{ij} \right)$$

The sum of chi-squared random variables is chi-squared distributed[3] for unit vector v, so the term in the parentheses is chi-squared distributed.

$$v^\mathsf{T} M v = \sum_{i=1}^{d} v_i M_i'$$

$$= M''$$

where the entries of M' are also chi-squared distributed, and similarly this weighted sum of the entries is also chi-squared distributed. □

Therefore, we have

$$\|Gv\|^2 / d \sim Y$$

[3] `https://online.stat.psu.edu/stat414/book/export/html/784`

so we can use the inequality from PSET 0:

$$\Pr\left[Y/d > (1+\epsilon)^2\right] = \Pr\left[Y > (1+\epsilon)^2 d\right]$$

$$< e^{-cd\epsilon^2}$$

Bibliography

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Wiley.