Andrew Li-Yang Liu
Nakul Verma
Unsupervised Machine Learning
Columbia University
November 19, 2022

# UML HW3

# Problem 1: Reading on "Horseshoes in Multidimensional Scaling and Local Kernel Methods" by Diaconis, Goel and Holmes

The authors apply multidimensional scaling (MDS) to the 2005 US House of Representatives roll call votes. They observe the horseshoe phenomenon (which is heuristically attributed to latent ordering of the data) and provide a theoretical justification as to why it occurs.

## Section 1: Review of MDS and preliminary results

Recall from HW1 that if we have an Euclidean distance matrix $(D_2)_{ij} = \sqrt{\sum_{k=1}^{p}\left(x_i^{(k)} - x_j^{(k)}\right)^2}$ defined for unknown points $x_1, \dots, x_n \in \mathbb{R}^p$, we can find an isometric embedding of $X$ with the following procedure: Define the centered similarity matrix $S = XX^T = -\frac{1}{2}HD_2H$, where $H = I - \frac{1}{n}11^T$. Since $S$ is PSD, we can take $S = U\Lambda U^T$ and subsequently $X = U\Lambda^{1/2}$. One can check that this embedding for $X$ is in fact isometric.

Say that instead we wanted to find a lower-dimensional embedding that is approximately isometric. This can be achieved simply by taking the dominant eigenvectors (similarity modes) of $S$. In other words, suppose we want a $k$-dimensional embedding. Then assume that the eigenvalues in the diagonal entries of $\Lambda$ are of descending order (if not, just reorder them) and reorder the eigenvector columns of $U$ accordingly. Then just take $Y_k$ to be the first $k$ columns of $U$ and rescaled such that the squared norm of each eigenvector is the corresponding eigenvalue. This rescaling ensures that, in the limit when $k = p$, we recover the non-compressed isometric embedding $Y_{k=p} = X$. Since we are taking the $k$ dominant eigenvectors, this compression scheme achieves the following objective:

$$\min_{y_i \in \mathbb{R}^k} \left( \|x_i - x_j\|_2^2 - \|y_i - y_j\|_2^2 \right)$$

(where $y_i$'s are columns of $Y$)

Relating the eigenvectors of $D_2$ to the eigenvectors of $S$: Suppose that $w$ is an eigenvector of $D_2$ with eigenvalue $\lambda$, and denote $\overline{w} = \left(\frac{1}{n}\sum_{i=1}^{n} w_i\right)\mathbf{1}$ (i.e. a constant vector with each entry being

the average of the entries of $w$). Using the fact that for any $v$ such that $1^T v = 0$ we have $Hv = v$, one can verify that:

$$S(w - \bar{w}) = -\frac{\lambda}{2}(w - \bar{w}) + \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right)\begin{bmatrix} r_1 - \bar{r} \\ \vdots \\ r_n - \bar{r} \end{bmatrix}$$

Where $r_i = \sum_{j=1}^{n}(D_2)_{ij}$ and $\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i$. This means that if $\bar{w} = 0$, $w$ is also an eigenfunction of $S$. In the case where $\bar{w} \cong 0$ or $\bar{r} \cong r_i \ \ \forall i$, we have that $w - \bar{w}$ is an approximate eigenfunction of $S$. In the case of the House of Representatives data, it turns out that both $\bar{w} \cong 0$ and $\bar{r} \cong r_i$ are satisfied, so $S$ and $D_2$ do indeed have approximately the same eigenvectors.

**Section 2: Modeling the data**

After some preprocessing, the data matrix $V$ is a 435x669 matrix representing the 669 roll calls of 435 legislators. In each roll call, $V_{ik} \in \left\{\frac{1}{2}, -\frac{1}{2}\right\}$, where ½ corresponds to a vote of "yea" and -1/2 corresponds to "nay". One defines the empirical distance between two legislators as $\hat{d}(l_i, l_j) = \frac{1}{669}\sum_{i=1}^{669} |V_{ik} - V_{jk}|$ and the empirical dissimilarity matrix as $P(i,j) = 1 - \exp\left(-\hat{d}(l_i, l_j)\right)$. The point of defining $P$ is so that for $\hat{d}(l_i, l_j) \ll 1$, we have $P(i,j) \cong \hat{d}(l_i, l_j)$ and that $P$ is not as sensitive to noise for large $\hat{d}(l_i, l_j)$. These aspects of $P$ reflect the fact that high similarities are more meaningful.

We make the assumption that the legislators $l_1, \ldots, l_n$ can be isometrically mapped to a unit interval $I = [0,1]$, with lower values being liberal and higher values being conservative. We can define the latent distance between two legislators simply as $d(l_i, l_j) = |l_i - l_j|$. The paper now employs a cut-point model for voting, where each bill $1 \le k \le m$ is represented as a tuple $(C_k, P_k) \in [0,1] \times \{0,1\}$. $C_k$ is the cutting point between legislators voting "yea" and "nay", and $P_k$ denotes whether the bill was proposed by a liberal (0) or conservative (1). Thus:

$$V_{ik} = \begin{cases} \frac{1}{2} - P_k, & l_i \le C_k \\ P_k - \frac{1}{2}, & l_i > C_k \end{cases}$$

We note that the system is underdetermined for determining $l_i$'s and $C_k$'s since slight perturbations to $l_i$ and $C_k$ keep the votes the same. One reduces the degrees of freedom by requiring $C_k \sim \text{Unif}([0,1])$ iid. Then we have $\mathbb{P}(V_{ik} \neq V_{jk}) = d(l_i, l_j)$ since two legislators are opposing if and only if $C_k \in [l_i, l_j]$.

With this distribution defined, one can show that the empirical distance $\hat{d}_m$ converges to the latent distance $d$, i.e. $\lim_{m\to\infty} \hat{d}_m(l_i, l_j) = d(l_i, l_j)$. More concretely, for any $\epsilon > 0$ if $m \ge \log\left(\frac{n}{\sqrt{\epsilon}}\right)/\epsilon^2$, we have that $\mathbb{P}\left(\left|\hat{d}_m(l_i, l_j) - d(l_i, l_j)\right| \le \epsilon \quad \forall 1 \le i, j \le n\right) \ge 1 - \epsilon$. This is proven trivially using Hoeffding's inequality and a union bound.

**Section 3: How horseshoes arise from this model.**

Consider a special form of the latent distance function $d(l_i, l_j) = \left|\frac{i}{n} - \frac{j}{n}\right|$, which corresponds to when legislators are evenly spaced in $I$. Consider the corresponding dissimilarity function $P(i,j) = 1 - \exp\left(-\left|\frac{i}{n} - \frac{j}{n}\right|\right)$. To perform an approximately isometric embedding, we want to find the top eigenvectors of $S = -\frac{1}{2}HPH = -\frac{1}{2}(P - JP - PJ + JPJ)$ where $J = \left(\frac{1}{n}\right)11^T$.

Solving this eigenproblem in the discrete case is hard, so consider the continuous relaxation by considering $S_n = \frac{1}{n}S$ and the eigenvalue problem $S_n v = \lambda v$. Taking the large $n$ limit, $S_n$ becomes an integral operator and the eigenproblem becomes:

$$\int_0^1 K(x,y)f(y)dy = \lambda f(x)$$

Where the kernel $K(x,y)$ is given as:

$$K(x,y) = \frac{1}{2}\left(e^{-|x-y|} - \int_0^1 e^{-|x-y|}dx - \int_0^1 e^{-|x-y|}dy + \int_0^1\int_0^1 e^{-|x-y|}dxdy\right)$$
$$= \frac{1}{2}\left(e^{-|x-y|} + e^{-y} + e^{-(1-y)} + e^{-x} + e^{-(1-x)}\right) + e^{-1} - 2$$

After solving the problem in the continuous domain, the following theorems provide the approximate eigenvector solution and error bounds in the discretized setting:

Theorem 3.1: The approximate eigenfunctions of $S_n$ are:

1. $f_{n,a}(x_i) = \cos\left(a\left(\frac{i}{n} - \frac{1}{2}\right)\right) - \frac{2}{a}\sin\left(\frac{a}{2}\right)$ where $a > 0$ is a solution to $\tan\left(\frac{a}{2}\right) = \frac{a}{2+3a^2}$. We have $S_n f_{n,a}(l_i) = \frac{1}{1+a^2}f_{n,a}(x_i) + R_{f,n}$, where the residual term is bounded: $\left|R_{f,n}\right| \leq \frac{a+4}{2n}$
2. $g_{n,a}(x_i) = \sin\left(a\left(\frac{i}{n} - \frac{1}{2}\right)\right)$, where $a > 0$ is a solution to $a\cot\left(\frac{a}{2}\right) = -1$. In this case, we have $S_n g_{n,a}(x_i) = \frac{1}{1+a^2}g_{n,a}(x_i) + R_{g,n}$, with $\left|R_{g,n}\right| \leq \frac{a+2}{2n}$

Theorem 3.2: We can similarly bound the error on the approximate eigenvalues.

1. For $a > 0$ solutions to $\tan\left(\frac{a}{2}\right) = \frac{a}{2+3a^2}$, we have $\min_{1\leq i\leq n}\left|\lambda_i - \frac{1}{1+a^2}\right| \leq \frac{a+4}{\sqrt{n}}$
2. For $a > 0$ solutions to $a\cot\left(\frac{a}{2}\right) = -1$, we have $\min_{1\leq i\leq n}\left|\lambda_i - \frac{1}{1+a^2}\right| \leq \frac{a+2}{\sqrt{n}}$

Proof sketch of theorems 3.1 and 3.2:

Form theorem 3.1, we want to show that in the continuous limit we have $f(x) = \cos\left(a\left(x-\frac{1}{2}\right)\right) - \frac{2}{a}\sin\left(\frac{a}{2}\right)$ where $\tan\left(\frac{a}{2}\right) = \frac{a}{2+3a^2}$ and $g(x) = \sin\left(a\left(x-\frac{1}{2}\right)\right)$ where $a\cot\left(\frac{a}{2}\right) = -1$ both satisfying the continuous eigenproblem $\int_0^1 K(x,y)f(y)dy = \lambda f(x)$, with $\lambda = \frac{1}{1+a^2}$.

One first notices that the eigenproblem is similar to the Fredholm equations of second type in Fredholm theory of integral equations, which has trigonometric solutions. One can thus guess a trigonometric form and check by plugging the solutions into the eigenproblem and performing straightforward integration.

The bound on the residual terms is simply the discretization error of the otherwise continuous kernel $K$. Approximating the integral as a Riemann sum gives rise to such bounds.

We also want to show that these particular eigenfunction solutions are in fact the only solutions with positive associated eigenvalues. This part is somewhat mathematically involved.

It can be shown that $\|K\|_\infty < 1$, so if $\lambda$ is an eigenvalue of $K$, then $|\lambda| < 1$. Let $f$ be an eigenfunction of $K$, then $\int_0^1 K(x,y)f(y)dy = \lambda f(x)$. Taking two derivatives wrt $x$ on both sides and relating this doubly-differentiated equation with the original equation, we then denote $g(x) = f'(x)$, which gives rise to the second order ODE:

$$g''(x) = \frac{\lambda - 1}{\lambda} g(x)$$

For $0 < \lambda < 1$ (in particular $\lambda = \frac{1}{1+a^2}$), the solution of this ODE is:

$$g(x) = A\sin\left(a\left(x-\frac{1}{2}\right)\right) + Bcos\left(a\left(x-\frac{1}{2}\right)\right)$$

Obtain $f(x)$ by integrating $g$:

$$f(x) = A\sin\left(a\left(x-\frac{1}{2}\right)\right) + B\cos\left(a\left(x-\frac{1}{2}\right)\right) + C$$

We note that $\int_0^1 K(x,y)dy = 0$, so the constant function $C$ is an eigenfunction of $K$ with eigenvalue 0. Furthermore, since $K$ is symmetric, eigenfunctions $f$ with nonzero eigenvalue are necessary orthogonal to $C$, i.e. $\int_0^1 f(x)dx = 0$. So we can write, WLOG:

$$f(x) = A\sin\left(a\left(x-\frac{1}{2}\right)\right) + B\left[\cos\left(a\left(x-\frac{1}{2}\right)\right) - \frac{2}{a}\sin\left(\frac{a}{2}\right)\right]$$

Assuming $B \neq 0$, then we can obtain $f\left(\frac{1}{2}\right) = 1 - \left(\frac{2}{a}\right)\sin\left(\frac{a}{2}\right)$. By symmetry of $K(x,.)$ with respect to $1/2$ and skew-symmetry of $\sin\left(a\left(x-\frac{1}{2}\right)\right)$, with respect to ½, one obtains another

complicated expression for $\lambda f\left(\frac{1}{2}\right)$. Equating the two forms and solving for $a$ gives rise to $\tan\left(\frac{a}{2}\right) = \frac{a}{2+3a^2}$. The solution for $a$ corresponds to approximately even multiples of $\pi$.

Assuming $A \neq 0$. One can obtain $f'\left(\frac{1}{2}\right) = a$ and $\lambda f'\left(\frac{1}{2}\right) = -\frac{e^{-\frac{1}{2}}}{1+a^2}\left(a\cos\left(\frac{a}{2}\right) + \sin\left(\frac{a}{2}\right)\right) + \frac{a}{1+a^2}$, giving rise to $a\cot\left(\frac{a}{2}\right) = -1$. The solution for $a$ corresponds to approximately odd multiples of $\pi$.

Since $\tan\left(\frac{a}{2}\right) = \frac{a}{2+3a^2}$ and $a\cot\left(\frac{a}{2}\right) = -1$ cannot be simultaneously satisfied, we have that $A = 0$ if and only if $B \neq 0$. Hence, the eigenfunctions given in Theorem 3.1 are in fact the only solutions with $0 < \lambda < 1$.

Hence, should we perform a 2-dimensional MDS using the first and second largest approximate eigenfunctions of $S$ as the embedding, the parametrized curve $\Lambda: x_i \mapsto \left(\sqrt{\lambda_1}f_1(x_i), \sqrt{\lambda_2}f_2(x_i)\right)$ will indeed look like a horseshoe.

In the case of voting data, we see 2 horseshoes. This is because when considering $\mathcal{X} = \{x_1, \dots, x_n, y_1, \dots, y_n\}$ with $d(x_i, x_j) = 1 - e^{-\left|\frac{i}{n} - \frac{j}{n}\right|}$, $d(y_i, y_j) = 1 - e^{-\left|\frac{i}{n} - \frac{j}{n}\right|}$ and $d(x_i, y_j) = 1$, one obtains a proximity matrix $\tilde{P}_{2n} = \begin{bmatrix} P_n & 1 \\ 1 & P_n \end{bmatrix}$, with $P_n(i,j) = 1 - e^{-\left|\frac{i}{n} - \frac{j}{n}\right|}$. Using similar analysis as above, this proximity matrix gives rise to three forms of approximate eigenfunctions $f_1, f_2, f_3$. When plotting the 3D embedding $\Lambda: z \mapsto \left(\sqrt{\lambda_1}f_1(z), \sqrt{\lambda_2}f_2(z), \sqrt{\lambda_3}f_3(z)\right)$, we see that we indeed get 2 horseshoes. Note that, as is also in the case of spectral clustering, the first eigenvector $f_1$ acts as the class separator.

**Section 4: Connecting the model to the data**

The authors then show that the eigenfunctions empirically obtained from the data roughly match the theoretically-derived eigenfunctions above.

# Problem 2: Further questions on embeddings

We take $d = O(\log^2 n)$. Based on a result from HW0, we have the following inequality relating L2 norms and L-infinity norms:

$$\|f(x) - f(x')\|_{L_2^d} \leq \|f(x) - f(x')\|_{L_\infty^d} d^{\frac{1}{2}} = \|f(x) - f(x')\|_{L_\infty^d} O(\log n)$$

By the given theorem for L-infinity Bourgain embeddings, we have that there exists $r > 0$ such that:

$$\|f(x) - f(x')\|_{L_\infty^d} \leq rA\rho(x, x')$$

Where $A$ is the distortion of embedding into L-infinity space. We note that the relation between $d$ and $A$ is $d = O\left(An^{\frac{2}{D}} \log n\right)$. We note that for fixed $n$, $n^{2/A}$ is bounded in $A$. So $d = O(A \log n)$. We were also given that $d = O(\log^2 n)$, so $A = O(\log n)$. Hence:

$$\|f(x) - f(x')\|_{L_2^d} \leq rO(\log^2 n)\rho(x, x')$$

We identify $D = O(\log^2 n)$, thereby completing the upper bound for $\|f(x) - f(x')\|_{L_2^d}$.

Now consider the lower bound. From another result in HW0, we found that:

$$\|f(x) - f(x')\|_{L_2^d} \geq \|f(x) - f(x')\|_{L_\infty^d}$$

But we know that $\|f(x) - f(x')\|_{L_\infty^d} \geq r\rho(x, x')$ by the lower bound requirement for $D$-embeddability. Hence we arrive at the conclusion that $\exists r > 0$ such that $\forall x, x' \in X$:

$$r\rho(x, x') \leq \|f(x) - f(x')\|_{L_2^d} \leq rD\rho(x, x')$$

Where $D = O(\log^2 n)$ and $d = O(\log^2 n)$, completing the proof.

ii.

Consider a finite tree $T$ with metric $\rho$ (we note that between two nodes there is a unique path and hence we do indeed have a metric space). Let $|T| = n$. In the base case, if $n = 1$ or $n = 2$, $T$ can be isometrically embedded in $L_1^n$.

Now, for the inductive hypothesis, we assume that for $|T| = k$ there is an isometry $f: T \to L_1^k$, i.e. for all $u, v \in T$, we have $\|f(u) - f(v)\|_1 = \rho(u, v)$. The embedding for each node is a $k$-dimensional vector $f(v) = (f_1(v), f_2(v), \dots, f_k(v))$. Then we append to the graph (make $|T| = k + 1$) by adding a leaf node $w$. Let $z$ be the unique parent of $w$. Define a new function $g: T \to L_1^{k+1}$ as follows:

$$g(v) = \begin{cases} (f_1(v), \dots, f_k(v), 0) \text{ if } v \neq w \\ (f_1(z), \dots, f_k(z), \rho(w, z)) \text{ if } v = w \end{cases}$$

We see that, for nodes $u, v \neq w$ (nodes that were part of the original tree) $\|g(u) - g(v)\|_1 = \|f(u) - f(v)\|_1 + 0 = \rho(u, v)$. For the new node $w$ and its parent $z$, we have $\|g(w) - g(z)\|_1 = \|(f_1(z), \dots, f_k(z), 0) - \big(f_1(z), \dots, f_k(z), \rho(w, z)\big)\|_1 = \rho(w, z)$. So $g$ is an isometry.

Hence, through induction we've proven that every finite metric tree $T$ with $|T| = n$ can be isometrically embedded into $L_1^n$.

# Problem 3: From finiteness to (structured) infinity (and beyond)

Define $w$ to be the difference vector between any two points $x, y \in S$. Consider the unit vector in the same direction $v = \frac{w}{|w|}$. From the hint, we saw that if $\|Gv\| \cong \|v\|$, then $\|Gw\| \cong \|w\|$ (where the approximate means up to $\epsilon$-distortion). Hence, we only need to prove that $G$ is approximately isometric for unit vectors, i.e. for $v \in S^{k-1}$.

We note that $S^{k-1}$ is compact, so it admits an $\alpha$-cover $C = (c_1, \dots, c_N) \subset S^{k-1}$. Note that in order for $C = (c_1, \dots, c_N)$ to cover $S^{k-1}$, we must have:

$$N\ vol(c_i) \geq vol(S^{k-1})$$
$$N \geq \frac{vol(S^{k-1})}{vol(c_i)} = \frac{1}{\alpha^k}$$

We can apply the regular JL-Lemma onto $C$. Recall that each $c \in C$ corresponds to the difference $x - y$ for some $x, y \in S$ such that $\|x - y\| \leq 1$. Hence, given $\epsilon > 0$, $G$ with entries iid $N\left(0, \frac{1}{d}\right)$, if $d \geq \Omega\left(\frac{\ln(N)}{\epsilon^2}\right)$, then with probability at least ¾, we get:

$$(1 - \epsilon)^2\|c\|^2 \leq \|Gc\|^2 \leq (1 + \epsilon)^2\|c\|^2$$

In other words, if $d \geq \Omega\left(\frac{k \ln\left(\frac{1}{\alpha}\right)}{\epsilon^2}\right)$, then we have that $G$ is approximately isometric when applied to $c_1, \dots, c_N$ (covering centers). Now consider the case for points that are not cover centers. Consider a point $p$ that's not a cover center. Since $C$ is an $\alpha$-cover, $p = c + r$ for some $c \in C$ and $\|r\| \leq \alpha$. We want to choose $\alpha$ sufficiently small such that:

$$(1 - \epsilon)\|p\| \leq \|Gp\| \leq (1 + \epsilon)\|p\|$$

We note that since we are considering unit vectors $p$, i.e. $\|p\| = 1$, this inequality becomes:

$$(1 - \epsilon) \leq \|Gp\| \leq (1 + \epsilon)$$

This is achieved by noting that:

$$\|Gp\| \leq (1 + \epsilon)\|c\| + \|Gr\|$$

Now we note that since $G$ is a linear transformation, it has bounded matrix norm, where matrix norm is defined as $\|G\| = \sup_{x,|x|=1} \|Gx\| \leq M$. Therefore, $\|Gr\| \leq M\alpha$. To satisfy $\|Gp\| \leq (1+\epsilon)$, we can simply choose $\alpha$ to satisfy $M\alpha + (1+\epsilon)\|c\| \leq 1+\epsilon$, so:

$$\alpha \leq \frac{(1+\epsilon)(1-\|c\|)}{M}$$

Note that this must hold true for all cluster centers $c$ (since it holds true for any point $p$). So, equivalently, denote $q_{min} = \inf_{c \in C}(1-\|c\|)$. Then just choose $\alpha$ so that:

$$\alpha \leq \frac{(1+\epsilon)q_{min}}{M}$$

Put more simply, $\alpha = O(\epsilon)$.

So, in conclusion: If $d \geq \Omega\left(\frac{k}{\epsilon^2}\ln\left(\frac{1}{\epsilon}\right)\right)$, then with probability at least $3/4$ we have that $G$ is an $(1 \pm \epsilon)$-isometry on $S$.

# Problem 4: Hardness of Sparse PCA

i.

If the vertices of $G$ form a clique, then the adjacency matrix is of the form $A_{ij} = \begin{cases} 1 & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$. So consider the eigenproblem:

$$Av = \lambda v$$

Clearly, $\lambda = n-1$ is an eigenvalue, with corresponding eigenvector $v = (1,1,1,1,\ldots,1)$.

Now prove the other direction. Suppose the adjacency matrix has eigenvalue $\lambda = n-1$.

Suppose $G$ is not a clique, then compare $A_G$ with $A_{clique}$. Suppose $v$ is the (unit) eigenvector corresponding to the largest eigenvalue of $A_G$. Then:

$$\lambda_{max}^{G} = v^T A_G v \leq v^T A_{clique} v \leq \max_{x,\|x\|=1} x^T A_{clique} x = n-1 = \lambda_{max}^{clique}$$

Where the first inequality arises because we can assume the elements in $v$ are of the same sign (otherwise we get cancellation of magnitudes, which leads to a lower value for $v^T A_G v$), so assume all of them are non-negative, which implies:

$$v^T A_G v = \sum_{ij}^{n} v_i (A_G)_{ij} v_j \leq \sum_{ij}^{n} v_i \left(A_{clique}\right)_{ij} v_j$$

(since $v_i v_j \geq 0$ and the matrix $A_{clique}$ has more 1 entries than $A_G$)

We note that $v_i v_j > 0$ (strictly positive) if all elements of $v$ are positive. This corresponds to the case when $G$ is a connected graph. Assuming $G$ is a connected graph, then we get the strict inequality $\lambda_{max}^{G} < n - 1$.

If $G$ is not a connected graph, we note that we can construct a connected graph $G'$ by adding edges connecting disconnected components. Moreover, we can add just enough edges so that $G'$ is both connected and not fully connected. By similar argument as above, we get that $\lambda_{max}^{G} \leq \lambda_{max}^{G'} < \lambda_{max}^{clique} = n - 1$.

So in either case, $\lambda_{max}^{G} < n - 1$, which contradicts the fact that $A_G$ has an eigenvalue $\lambda = n - 1$. So $G$ must be a clique.

So $A_G$ has eigenvalue $n - 1$ if and only if $G$ is a clique.

ii.
The clique problem is as follows: "Does G contain a clique of k vertices?" Consider an instance of the clique problem $\rho = (G, k)$. We note that a subgraph $S$ of $G$ is a clique of $k$ vertices if and only if $S$ has an eigenvalue of $k - 1$. So let $A_G$ be the adjacency matrix of G and $A_S$ be the adjacency matrix of the subgraph $S$.

**We want to prove that $S$ has eigenvalue of $k - 1$ if and only if there exists a unit vector $w \in \mathbb{R}^{|S|}$ such that $w^T A_S w \geq k - 1$.** In the forward direction, if $S$ has an eigenvalue of $k - 1$, then, trivially, there exists unit vector $w \in \mathbb{R}^{|S|}$ such that $w^T A_S w \geq k - 1$. Now prove the backward direction. We know the Rayleigh quotient $R = \frac{w^T A_S w}{w^T w} = w^T A_S w$, where $|w|^2 = 1$, is maximized by an the largest eigenvector of $A_S$ since constrained optimization with the Lagrangian $w^T A_S w - \lambda(|w|^2 - 1)$ gives rise to $A_S w = \lambda w$. So if there is a $w$ s.t. $w^T A_S w \geq k - 1$, there must also be an eigenvector satisfying the same inequality. We know that the eigenvalue of any adjacency matrix is at most $k - 1$. Hence, if there is a $w$ such that $w^T A_S w \geq k - 1$, this implies there is an eigenvector $x$ such that $x^T A_S x = k - 1$ (strict equality). This proves the statement.

Now let $v(w) \in \mathbb{R}^{|G|}$ be the vector whose non-zero elements are the entries of $w$ (i.e. $v$ is the extension of $w$ with other entries assigned to 0). Hence, G contains a clique of k vertices if and only if $v^T A_G v = w^T A_S w \geq k - 1$, where $v$ has at most $|S|$ non-zero elements.

So consider an instance of the sparse PCA problem $\phi = (A, m, M)$. Let $M = |S|$, $A = A_G$, and $m = |S|$. We see therefore that $\rho$ reduces to $\phi$, so the sparse PCA problem is also NP-hard. Clearly, the sparse PCA problem is in NP (since we can verify the solution in polynomial time). Hence, sparse PCA is NP-complete.

# Problem 5: SDP for Multidimensional Scaling

Objective:

$$\text{maximize } \frac{1}{2N}\sum_{i,j\in[N]}\|y_i-y_j\|_2^2$$

$$\text{subject to } d(x_i,x_j)=\|y_i-y_j\|_2 \text{ for all } (i,j)\in E$$

i.

$E$ defines a connected graph. Then, consider:

$$\frac{1}{2N}\sum_{i,j\in[N]}\left\|y_i-y_j\right\|_2^2 \le \frac{1}{2N}(2N)\max_{i,j}\left\|y_i-y_j\right\|_2^2 = \max_{i,j}\left\|y_i-y_j\right\|_2^2$$

Since the graph is connected, we can find a path of edges in E connecting the two farthest-distance nodes. Since the distance between two connected nodes is finite, by triangle inequality, $\max_{i,j}\left\|y_i-y_j\right\|_2^2 \le \text{length of path} < \infty$. So the objective function is bounded from above.

ii.

The objective function is:

$$\frac{1}{2N}\sum_{i,j=1}^{N}\left\|y_i-y_j\right\|_2^2 = \frac{1}{2N}\sum_{i,j=1}^{N}\left(\langle y_i,y_i\rangle+\langle y_j,y_j\rangle-2\langle y_i,y_j\rangle\right) = \frac{1}{2N}\sum_{i,j=1}^{N}\left(B_{ii}+B_{jj}-2B_{ij}\right)$$

$$= \frac{1}{2N}\left(2N\mathrm{Tr}(B)-2\sum_{i,j=1}^{N}B_{ij}\right) = \mathrm{Tr}(B)-\frac{1}{N}\mathrm{Tr}(11^TB) = \mathrm{Tr}\left(\left(I-\frac{1}{N}11^T\right)B\right)$$

$$= \mathrm{Tr}(HB)$$

Where $H = I-\frac{1}{N}11^T$

The constraints are:

$$d(x_i,x_j) = \left\|y_i-y_j\right\|_2$$

Note that $d(x_i,x_j)=d_{ij}$ are given. The constraint becomes:

$$B_{ii}+B_{jj}-2B_{ij}-d_{ij}=0 \text{ for } i\sim j$$

We also note that the Gram matrix must be PSD.

So the SDP is:

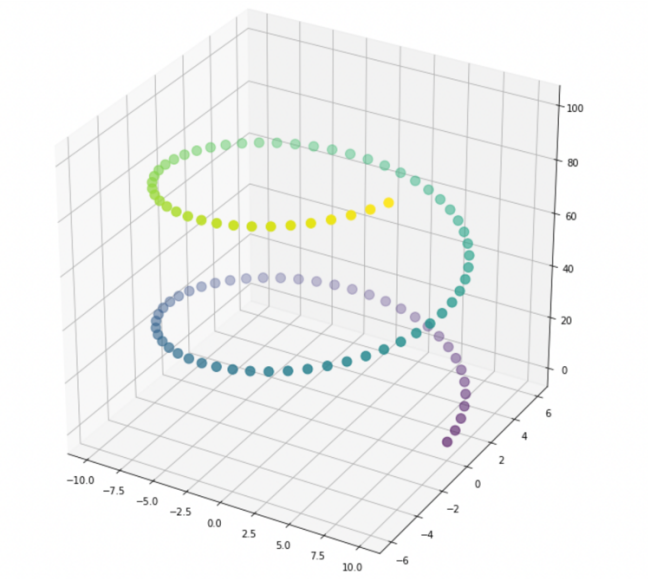
$$\max_B \mathrm{Tr}(HB)$$
$$\text{s.t. } B_{ii}+B_{jj}-2B_{ij}-d_{ij}=0 \text{ for } i\sim j$$
$$B \text{ is PSD}$$

Note that an equivalent formulation of the problem involves requiring $\sum_{i,j=1}^{N} B_{ij} = 0$. In this case, following the algebra above, we simply maximize $Tr(B)$. So the SDP can also be written as:

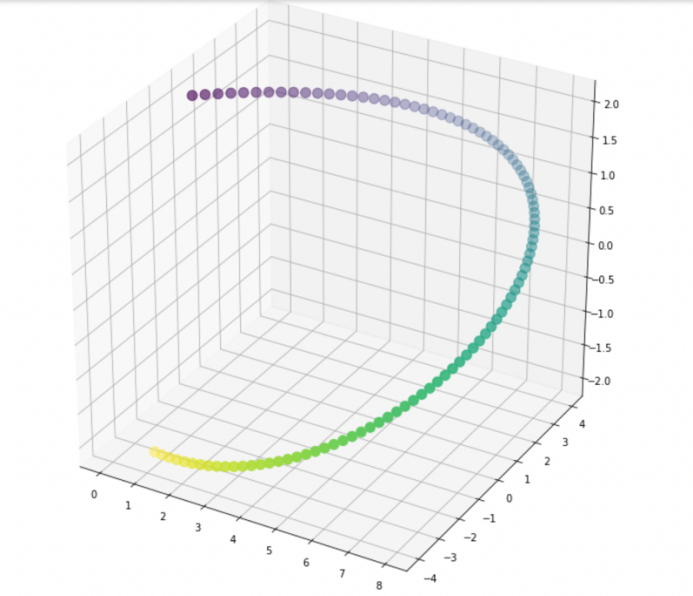
$$\max_{B} \operatorname{Tr}(B)$$
$$\text{s.t. } B_{ii} + B_{jj} - 2B_{ij} - d_{ij} = 0 \text{ for } i \sim j$$
$$\sum_{i,j=1}^{N} B_{ij} = 0$$
$$B \text{ is PSD}$$

iii.

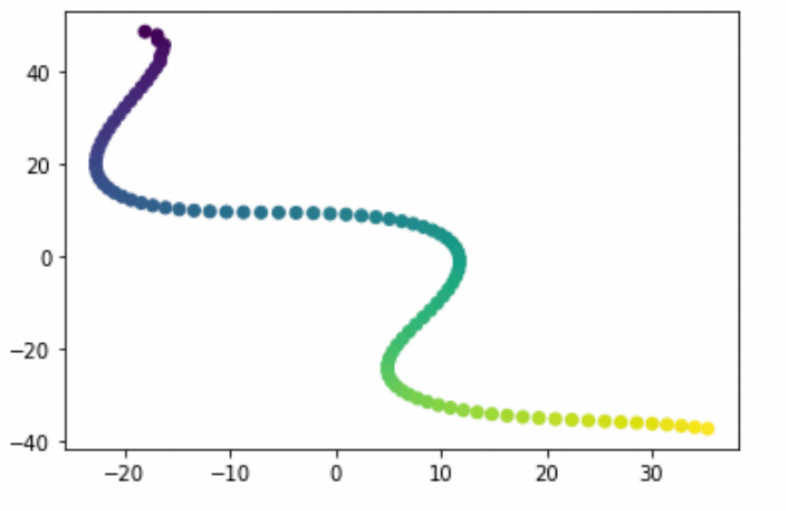
a. The first three dimensions form a spiral



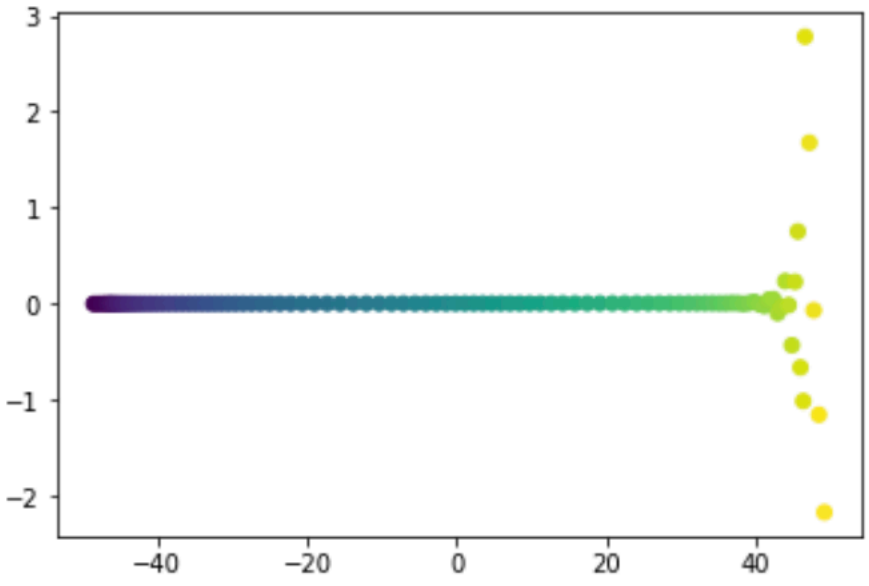

b. The last three dimensions form an incomplete loop:



c. Embedding with classical MDS:




d. Embedding using SDP Unfolding + MDS:




We see that the SDP unfolding procedure does a somewhat good job of turning geodesics on the manifold into straight line distances.