MATH-BIOINF-STATS 547: Mathematics of Data

Due Date: April 22, 2025

Problem Set 5: SVD, MDS, TDA (OMG!)

The first three problems are to refresh your memory about SVD and PCA and introduce multidimensional scaling (MDS).

Remark: Please submit a `.pdf` document with a write-up of your results and observations. We encourage using Overleaf, but MS Word or other similar word-processing software is OK. We have provided a LaTeX template to help get you started, which is available on Canvas and the course website.

Problem 1 - Singular Value Decomposition (SVD):

One of the best references for the SVD is Chapter 2 in the book Matrix Computations (Golub and Van Loan, 4th edition [1]).

(a) **Existence**: Prove the existence of the SVD. That is, show that if $\mathbf{A}$ is an $m \times n$ real valued matrix, then $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, where $\mathbf{U}$ is an $m \times m$ orthogonal matrix, $\mathbf{V}$ is an $n \times n$ orthogonal matrix, and $\mathbf{S} = \text{diag}(\sigma_1, \sigma_2, ..., \sigma_p)$ (where $p = min\{m, n\}$) is an $m \times n$ diagonal matrix. It is conventional to order the singular values in decreasing order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. Determine to what extent the SVD is unique. (See Theorem 2.4.1, page 76 in Golub and Van Loan).

(b) **Best rank$-k$ approximation - Frobenius norm**: Show that the SVD also provides the best rank$-k$ approximation for the Frobenius norm, that is, $\mathbf{A}_k$= $\mathbf{U}\mathbf{S}_k\mathbf{V}^T$ satisfies

$$\|\mathbf{A} - \mathbf{A}_k\|_F = \min_{rank(B)=k} \|\mathbf{A} - \mathbf{B}\|_F$$

Problem 2 - Multidimensional Scaling (MDS):

MDS is a popular technique for mapping a finite metric space into a low-dimensional Euclidean space in a way that best preserves pairwise distances. Given a distance matrix $\mathbf{D}^X$ from N points, find a set of N points $\mathbf{Y} = \{y_i \text{ for } i \in [1, N]\}$ in a k-dimensional space so that the distance matrix $\mathbf{D}^Y$ is as close as possible to $\mathbf{D}^X$.

Steps:

(i) **Input:** $N \times N$ distance matrix $\mathbf{D}^X$, where $\mathbf{D}^X = d_{ij}^2$ is a symmetric matrix of the squared distances between all points, and desired dimension k.

(ii) Let $\mathbf{B}^X = -\frac{1}{2}\mathbf{H}\mathbf{D}^X\mathbf{H}$. $\mathbf{H} = \mathbf{I}_N - \frac{1}{N}\mathbf{e}\mathbf{e}^\mathsf{T}$, where $\mathbf{I}_N$ is the identity matrix and $\mathbf{e}$ is an $N \times 1$ column vector of ones ($\mathbf{H}\mathbf{D}^X$ centers the columns and $\mathbf{D}^X\mathbf{H}$ centers the rows).

(iii) SVD (or the eigenvalue decomposition due to symmetry) of the centered matrix gives $\mathbf{B}^\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_N)$

$$\mathbf{B}^\mathbf{X} = \mathbf{Y}^\mathsf{T}\mathbf{Y}, \text{ so } \mathbf{Y}^\mathsf{T}\mathbf{Y} = (\mathbf{U}\mathbf{\Lambda}^{1/2})(\mathbf{\Lambda}^{1/2}\mathbf{U}^\mathsf{T})$$

(iv) Choose the top k non-zero eigenvalues and corresponding eigenvectors, $\tilde{\mathbf{X}}_k = \mathbf{U}_k\mathbf{\Lambda}_k^{1/2}$, where $\mathbf{U}_k = (u_1, \ldots, u_k)$, $u_k \in \mathbb{R}^n$, $\Lambda_k = \text{diag}(\lambda_1, \ldots, \lambda_k)$. $\tilde{\mathbf{X}}_k$ are the first k columns of $\mathbf{Y}^\mathsf{T}$, which are the new k-dimensional coordinates of the data.

Problem:

MDS of cities: Visit the following website to perform the following exercise. `http://geobytes.com/citydistancetool/`

(a) Input a few cities (no less than 7), and collect the pairwise air traveling distances shown on the website into a matrix $\mathbf{D}$.

(b) Make your own code for the MDS algorithm for $\mathbf{D}$;

(c) Plot the normalized eigenvalues $\frac{\lambda_i}{\sum \lambda_i}$ in a descending order of magnitudes, analyze your observations (did you see any negative eigenvalues? if yes, why?).

(d) Make a scatter plot of those cities using top 2 or 3 eigenvectors, and analyze your observations.

Problem 3 - Topological Data Analysis (TDA):

Spatial transcriptomics data represents gene expression levels mapped to specific locations within a tissue. Use the starter code and spatial transcriptomics data to solve the following questions.

(a) First, we will focus on the gene Trem2. Reformulate the data so we only consider cells with nonzero gene expression. Create a simplicial complex from the resultant point cloud. Provide plots for multiple simplicial complices as we increase the radius. If you are stuck, this is a good place to start.

(b) Instead of nonzero gene expression, threshold the gene expression and plot simplicial complices. Compare these results with your results from the previous question.

(c) For a given threshold and radius, compare the simplicial complices for the genes Trem2, Lpl, Lep, Cd36.

References

[1] GH Golub and CF Van Loan. *Matrix computations*, volume 4. JHU press, 2013.