Visualization
Prof. Bernhard Schmitzer, Uni Göttingen, summer term 2024

Problem sheet 7

- *Submission by **2024-06-10** 18:00 via StudIP **as a single PDF/ZIP**. Please combine all results into one PDF or archive. If you work in another format (markdown, jupyter notebooks), add a PDF converted version to your submission.*

- *Use Python 3 for the programming tasks as shown in the lecture. If you cannot install Python on your system, the GWDG jupyter server at https://jupyter-cloud.gwdg.de/ might help. Your submission should contain the final images as well as the code that was used to generate them.*

- *Work in groups of up to three. Clearly indicate names and enrollment numbers of all group members at the beginning of the submission.*

Exercise 7.1: dimensionality reduction on histograms, v2.
In this problem we study in more detail the ability of PCA to meaningfully reduce a dataset of histograms when the standard Euclidean distance is used between the histograms.

1. First we generate an example dataset. Based on example code from the lecture, write a function that generates a histogram of samples from a normal distribution for given standard deviation σ and mean μ. Set the number of samples to $n = 10000$ and fix the bins to be 100 uniform bins on the interval $[-4, 4]$. Then write a function which, for a single fixed value of σ, returns a list/array of histograms for all $\mu \in$ `np.linspace(-1,1,num=21)`. The output of this function should be an array of shape 21×100. Finally, generate these lists of histograms for $\sigma \in S := [0.8, 0.4, 0.2, 0.1]$. Subsequently, we will analyze theses lists of histograms separately with PCA, similar as in the lecture.

2. For each $\sigma \in S$, apply PCA to the corresponding list of histograms. Show the eigenvalues in a combined plot, with a focus on comparing the largest eigenvalues of each dataset. How do the dominant eigenvalues differ? How does the approximate dimensionality of the dataset depend on σ?

3. For each $\sigma \in S$, plot the 2-dimensional PCA embedding, i.e. the projection of all datapoints to the two dominant PCA eigenvectors. Add the original mean μ of each histogram as color. How well do the embeddings (for different σ) capture the underlying one-dimensional structure of the dataset?

4. For each $\sigma \in S$, visualize the change of histograms corresponding to moving along the first PCA direction by one standard deviation. How well does this capture the original dataset?

Exercise 7.2: embeddings with multidimensional scaling.
We test multidimensional scaling for embedding metric graphs.

1. The files `graphX.npz` for `X` $\in \{1, 2\}$ contain data for two metric graphs in `scipy.sparse.coo_array` format, i.e. each file contains three arrays labeled `data`, `i`, and `j`, corresponding to a sparse matrix in this format, which encode a weighted, symmetric graph. Each given

entry encodes an edge and its length. The matrix can be constructed with a standard constructor for `scipy.sparse.coo_array` (the shape should be set to $n \times n$, where $n = 512$ is the number of vertices in each graph). As in the lecture,

```
scipy.sparse.csgraph.shortest_path(csgraph=graph)**2
```

can then be used on these matrices to construct the matrix of squared pairwise distances. Compute these matrices for both graphs.

2. Then run MDS as in the lecture on both squared distance matrices. Plot the spectra of the gram matrices and briefly discuss them, in particular concerning how well the data can be embedded into Euclidean space, and how many dimensions are needed for a reasonable approximation.

3. Plot both two-dimensional embeddings. Do they look qualitatively different?