Topological and Geometric Data Reduction **19/02/2021**

Homework 1. PCA and MDS

YuiHong NG (20434594)

1 PCA experiments

I took digit data ' 3' to perform experiments in this report.

(a)(b)(c)(d) After the construction of centered data matrix $\tilde{X}$, the sample data covariance matrix Σ_n, the eigenvalue of both matrices are computed via singular value decomposition and eigendecomposition, the normalized eigenvalue of the covariance matrix $\tilde{\Sigma}_n$ is plotted in Fig. 1.

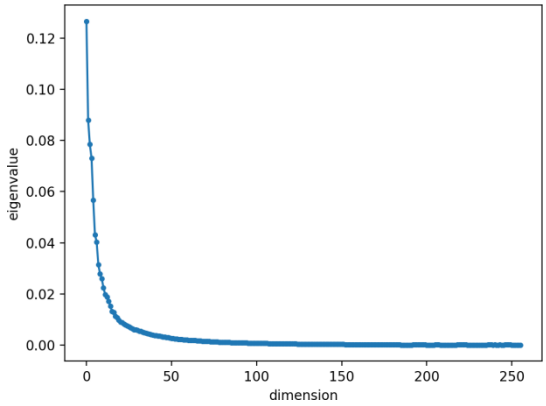

Figure 1: Percentage of singular values over total sum.

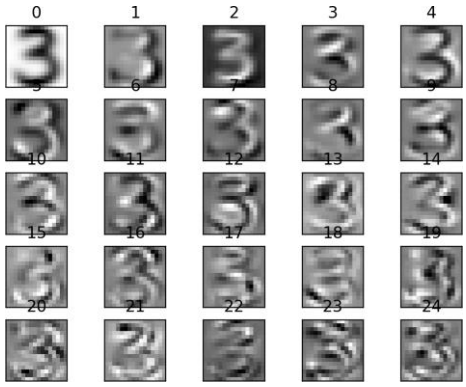

Figure 2: Images of the sample mean and top 24 principle components.

(e) The eigenvectors of data covariance matrix $\hat{\Sigma}_n$, and *left* singular vectors are the same, where top 24 principle components are visualized in Fig. 2 together with the sample mean $\hat{\mu}_n$.

(f) For $k = 1$, the image data x_i $(i = 1, \ldots, n)$ are sorted according the the embedding value to the top *right* singular vectors. In Fig. 3, the images are sorted in a ascending order, and the 25 images are uniformly sampled from 658 samples according to the sorted embedding.

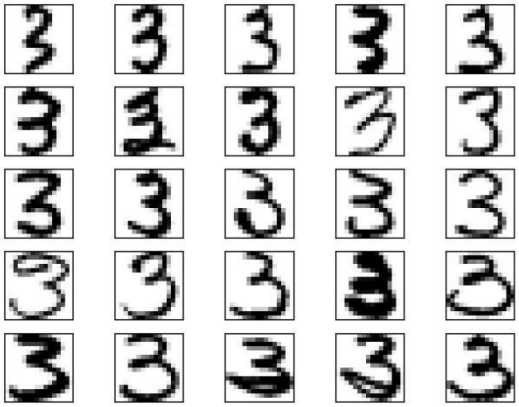

Figure 3: Sort the image data samples according to top 1 eigenvector embedding. Listed images are uniformly sampled according to the embedding.

(g) Similarly, for $k = 2$, the embedding on top 2 *right* singular vectors are displayed on the following scatter plot in Fig. 4, the sample images are also visualized located on their 2D embedding location.

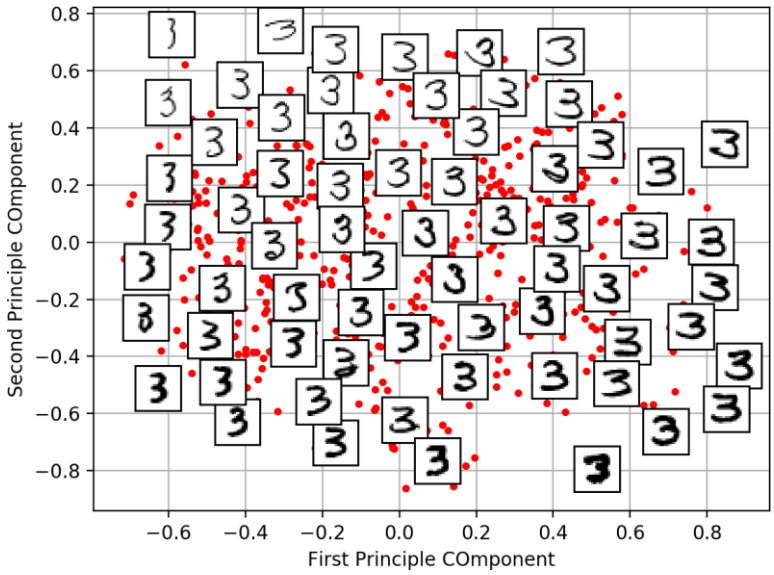

Figure 4: The first two principal components of the hand-written threes. The images show the nature of the first two principal components.

(h)* Finally, parallel analysis is tested. The elements within rows in centered data matrix $\tilde{X}$ is randomly permuted $\tilde{X}^k$. Noted that permutation within rows keep the sample mean, thus direct permutation on centered data matrix will not affect the computation. Then the singular values of permuted data matrix $\lambda^k{}_i$ are computed via SVD, as shown in Fig. 5 compared with original singular values. Then, the same procedure is repeated by $k = 100$ times and the $pval$ curve is calculated via

$$pval_i = \frac{1}{R}\{\hat{\lambda}_i^k > \hat{\lambda}_i\}$$

From the $pval$ curve, the top 19 principle components are suggested by parallel analysis.

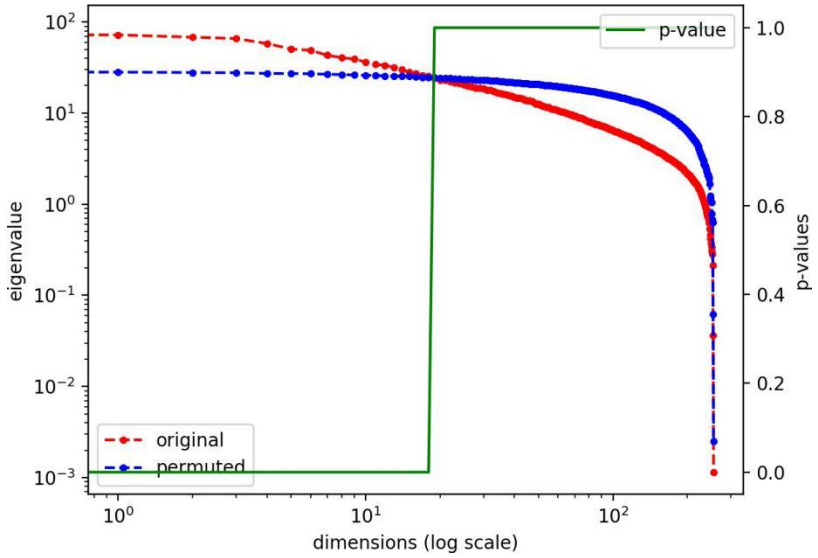

Figure 5: The first two principal components of the hand-written threes. The images show the nature of the first two principal components.

2 MDS of cities

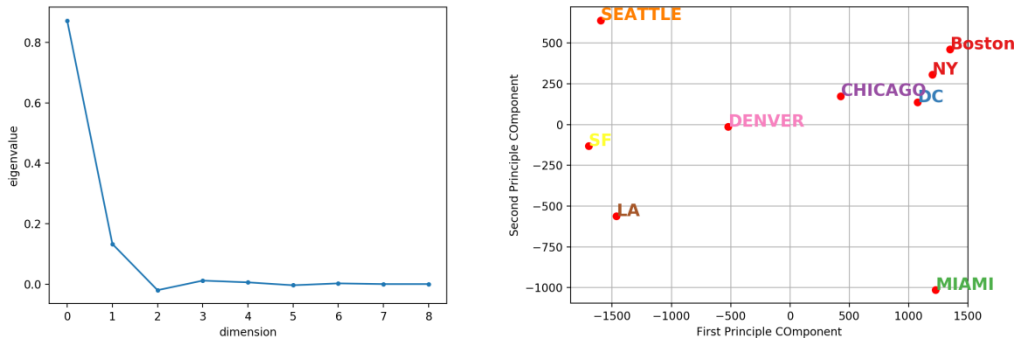

Figure 6: Eigenvalue curve and 2D embedding plot using the data in lecture notes.

(a) To prove the correctness of implementation of MDS algorithm, firstly the procedures (a)-(d) are tested on the distance data in lecture notes where 9 pairwise cities distance in US are considered. The results are shown in Fig. 6, which are identical to the results in lecture notes.

Then, pairwise cities distance are calculated via the suggested websites, 10 famous cities from all of the world are involved, shown in Fig. 8.

(b)(c) The MDS algorithm is performed by calculating the $\boldsymbol{B} = -\frac{1}{2}\boldsymbol{H}\boldsymbol{D}\boldsymbol{H}^T$, and then compute the eigen decomposition on inner product matrix $\boldsymbol{B}$. The eigenvalues of $\boldsymbol{B}$ is shown in Fig. 7. This example shows that MDS may obtain negative eigenvalues, when the distance is not Euclidean.

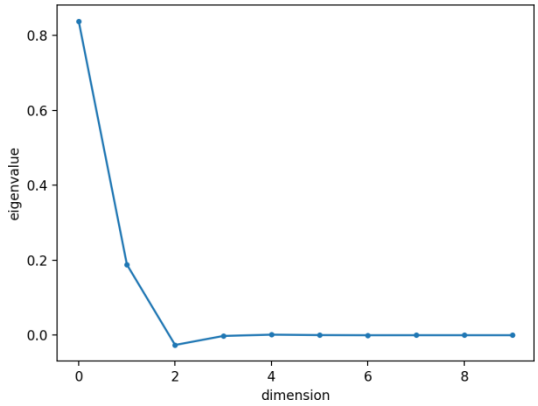

Figure 7: Normalized eigenvalue of MDS, where the third eigenvalue is negative.

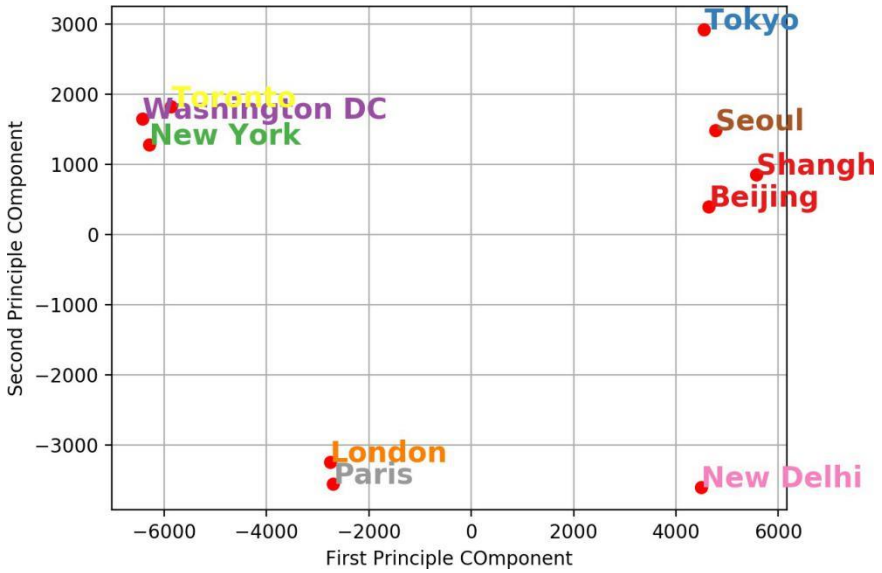

Figure 8: Sort the image data samples according to top 1 eigenvector embedding. Listed images are uniformly sampled according to the embedding.

(d) Then the scatter plot is shown in Fig. 8 using the top 2 eigenvectors, from the results, the relative location in first two principle components partially reflects the geographical relationship among the cities like the results in Fig. 6. The relative location in Asian cities or American cities are very close to the true geography, while the relative position of European cities may be different, maybe due to the distance indicates the shortest path on the sphere by air.

Q3 - Q5 are in CSIC5011_Q3_Q4_Q5.ipynb