Ryan_Plain_Homework_2_report

February 13, 2022

1 Homework 2

1.1 1. Conceptual Questions

1. Please prove the first principle component direction v corresponds to the largest eigenvector of the sample covariance matrix.

$$v = \arg \max_{w:||w||\leq 1} \frac{1}{m} \sum_{i=1}^{m} (w^T x^i - w^T \mu)^2$$

The covariance matrix for matrix X of m features is $M = E[(X-\mu)(X-\mu)^T]$.

PCA finds the vector (direction) where the variance is maximized. The first principal component will, due to the maximization objective, will correspond to the largest eigenvector.

2. What is the relationship between SVD and eigendecomposition? Please explain this mathematically, and touch on why this is relevant for PCA. A factorization of a rel matrix. $C = U$

Eigendecomposition relies on square matricies (typically symetrical), where as SVD will exist for any rectangular or square matrix. SVD is more general in that since.

SVD is a product of 3 matricies:

$$M = U\Sigma V^T$$

- $U \in \mathbb{R}^{nxm}$ – left singular vectors (orthonormal)
- $\Sigma \in \mathbb{R}^{nxm}$ – singular values
- $V \in \mathbb{R}^{mxm}$ – Right singular vectors (orthonormal)

The eigenvectors of $C := MM^T$ is the U left singular vectors

The eigenvalues of C is σ_i^2 (squared singular values of M)

To use it in PCA, the eigenvectors of U and V can be used with the number of columns as the number of principal components from the eigenvalues in the Σ matrix. To understand the amount of variance explained, you can take the sum of the principal components used, divided by the sum of all principal components.

3. Explain the three key ideas of ISOMAP (for manifold learning and non-linear dimensionality reduction.) Geodesic distance to capture distance between the points.

- A weighted nearest neighbors method is applied

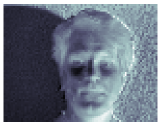
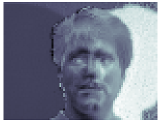
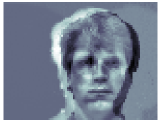
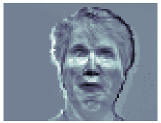
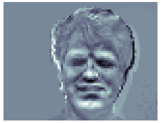
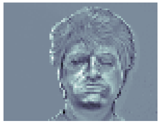
- Find the shortest path distance matrix D between each pairs of points
- Find low dimensional representation that preservce Produce low dimensional representation which preserves “walking distance” over the manifold.

Shortest distance

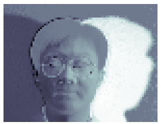
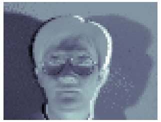
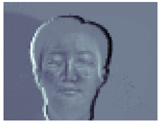
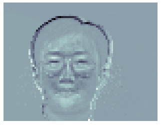
1.2 2. Eigenfaces and simple face recognition

1.2.1 a.)

Subject 01 Eigenfaces

1. 
2. 
3. 
4. 
5. 
6. 

Subject 02 Eigenfaces

1. 
2. 
3. 
4. 

5.
6.

Each image was vectorized, reduced by a factor of 4, and added to a dataset that was $M = (10, 4880)$ The first 6 eigenfaces are represented by individual eigenvectors extracted from SVD on the covariance matrix.

The first pictures are the most recognizable, because the first eigenvalue and eigenvector used captures the most variance in the dataset. As we progress through the eigenvectors, different shapes and features from unqique poses are captured. Eigenvectors such as those ≥ 4 contain more characteristicss that are represented of the individual.

1.2.2 b.)

To do face recognition, we need to use the top eigenvector from PCA to test against the new image. I used the SVD approach again, in which the eigenvectors are already sorted by the largest eigenvalues. The new test image was centered on the mean of the original training images.

Taking the squared $L2$ norm of the projection residual, we get the following values for the eigenvector and test subject:

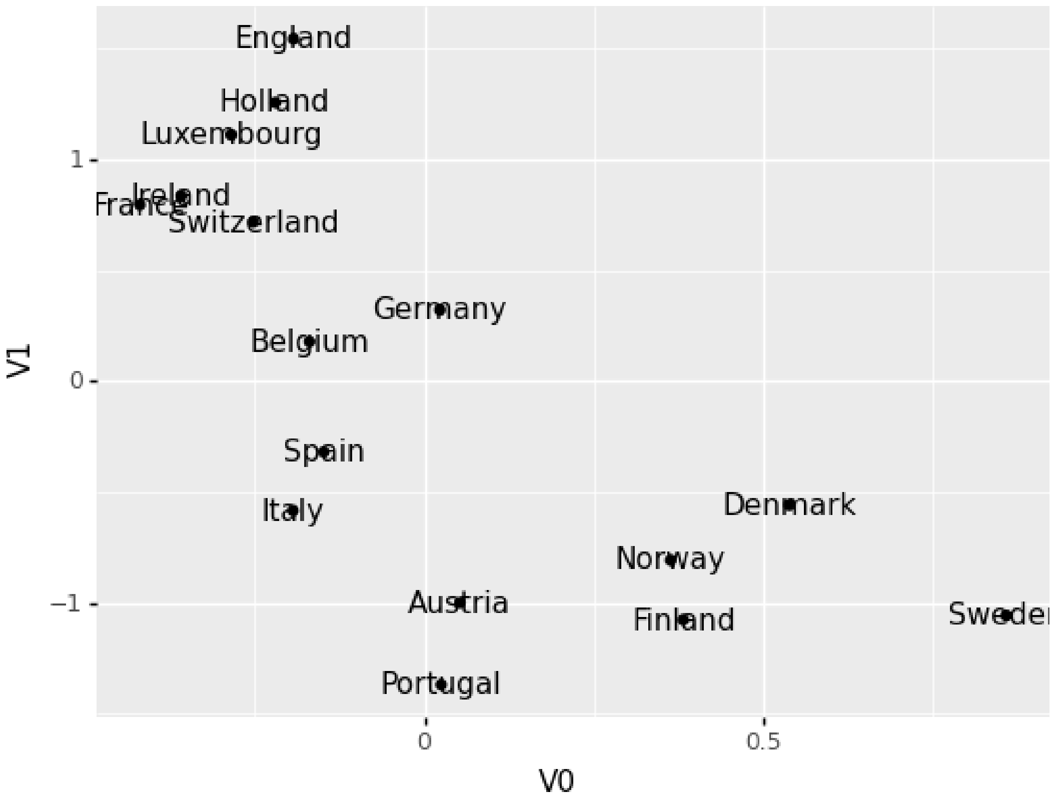
Subject 01
- s_{11}: 49,933,494
- s_{12}: 113,948,550

Subject 02
- s_{21}: 199,761,842
- s_{22}: 181,971,406

In both test cases, the residuals of the incorrect test subject are larger than the correct test subject. This can be used to classify which subject the test images belong to.

1.3 4. PCA: Food consumption in European Countries

a.) Using SVD, I was able to recreate the matrix using the first 2 principle components and respective eigenvectors. The scatter plot below shows the relationship of the first 2 principle components.



The countries that are clustered together show similar food characteristics. Doing this analysis makes it easier to view high dimensional data in a low dimensional setting.