EECS 553 HW 6

Due Thursday, Oct. 31, by 11:59 PM Eastern Time

0. Project Proposal

As a reminder, the project proposal is due at the same time as this homework assignment. To give you time to work on the proposal, this assignment is shorter than usual. Please see the project guidelines on Canvas for details regarding the proposal.

1. Eigenfaces (5 points each)

In this exercise you will apply PCA to a modified version of the `FairFace` dataset, which has been recently introduced as a demographically balanced facial imaging dataset. The modified dataset is available in the file `fairface.npy` on Canvas. The modification subsamples the original data and each image has been downsized by a factor of $4 \times 4 = 16$ to avoid computational and memory bottlenecks.

For a whirlwind tour of the data, issue the following commands. In Python:

```
import numpy as np
import matplotlib.pyplot as plt

X = np.load("fairface.npy")

fig = plt.figure(figsize=(10, 5))
for i in range(15):
    ax = fig.add_subplot(3, 5, i + 1, xticks=[], yticks=[])
    idx = np.random.randint(0, len(X) - 1)
    ax.imshow(X[idx], cmap=plt.get_cmap("gray"))
```

The data X here will be of size `(4500, 56, 56)`, i.e., 4500 facial images of size 56×56. What you will see running the code above are several facial images (in gray-scale) of different subjects, potentially under a variety of lighting conditions.

a. By viewing each image as a vector in a high dimensional space, perform PCA on the full dataset. Hand in a plot the sorted eigenvalues (use the `plt.semilogy` in Python) of the sample covariance matrix. How many principal components are needed to represent 95% of the total variation? 99%? What is the percentage reduction in dimension in each case? Useful commands in Python: `np.reshape, np.linalg.eig, np.linalg.svd, np.mean, np.diag`.

b. Hand in a 5×5 array of subplots (in gray-scale, see the sample code above) showing principal eigenvectors ('eigenfaces') 0 through 24 as images, treating the sample mean as the zeroth order principal eigenvector. Comment on what facial or lighting variations some of the different principal components are capturing.

c. Here you will compute the scores associated with each sample for two principal components in order interpret which particular images contributed to these principle components. Recall that the score is the inner product between a principal eigenvector and a sample instance, and the magnitude of the score determines how well the sample is explained by the PC. Compute the scores between the second principal component and all the original images in the database (remember to subtract the mean from the original images). Plot the first five facial images that result in the five scores of highest magnitude. Do the same for the eleventh principle component. Comment on the results for each case.

2. Uniqueness of PCA Subspace (5 points)

Give a simple condition involving the spectral decomposition of the sample covariance matrix that is both necessary and sufficient for the subspace $\langle \boldsymbol{A} \rangle$ in PCA to be unique. Justify your answer.

Hint: Consult the proof of Lemma 1 in the PCA notes.

3. PCA Optimal Object Function Value (5 points)

Let $k \in \{0, 1, \dots, d\}$ be arbitrary. Show that

$$\min_{\boldsymbol{\mu}, \boldsymbol{A}, \{\boldsymbol{\theta}_i\}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\theta}_i\|^2 = n \sum_{j=k+1}^{d} \lambda_j,$$

where $\boldsymbol{A}$ ranges over all $d \times k$ matrices with orthonormal columns.

Hint: Use properties of the trace operator.