CIS 520, Machine Learning, Fall 2021
Homework 6
Due: Tuesday, October 26th, 11:59pm
Submit to Gradescope

Yixuan Meng, Zhouyang Fang

1 Singular Value Decomposition

1. To solve

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}),$$

we have

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}}(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) &= \mathbf{0}\\ -\mathbf{2}\mathbf{X}^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\widehat{\mathbf{w}}) &= \mathbf{0}\\ \mathbf{X}^{\mathbf{T}}\mathbf{y} &= \mathbf{X}^{\mathbf{T}}\mathbf{X}\widehat{\mathbf{w}}.\end{aligned}$$

For $\mathbf{n} \times \mathbf{p}$ matrix $\mathbf{X}$ ($\mathbf{n} > \mathbf{p}$), for $\mathbf{v} \in \mathbb{R}^{\mathbf{p}}$, $\mathbf{X}\mathbf{v} = \mathbf{0}$ if and only if $\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{v} = \mathbf{0}$, which leads to

$$\begin{aligned}&\mathbf{v}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{v} = \mathbf{0}\\ \iff\ &(\mathbf{X}\mathbf{v})^{\mathbf{T}}\mathbf{X}\mathbf{v} = \mathbf{0}.\end{aligned}$$

Since the rank of $\mathbf{X}$ is p, meaning $\mathbf{X}$ is one-to-one when acting on $\mathbb{R}^{\mathbf{p}}$. So by observation, $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ is one-to-one, making it invertible.

Since $\mathbf{X}$ rank is $\mathbf{p}$, assume $\mathbf{X} = \mathbf{U_p}\boldsymbol{\Lambda}_{\mathbf{p}}\mathbf{V}_{\mathbf{p}}^{\mathbf{T}}$ using singular value decomposition. After substituting $\mathbf{X}$ in the equation, we get

$$\begin{aligned}\mathbf{w} &= (\mathbf{V_p}\boldsymbol{\Lambda}_{\mathbf{p}}\mathbf{U}_{\mathbf{p}}^{\mathbf{T}}\mathbf{U_p}\boldsymbol{\Lambda}_{\mathbf{p}}\mathbf{V}_{\mathbf{p}}^{\mathbf{T}})^{-\mathbf{1}}\mathbf{V_p}\boldsymbol{\Lambda}_{\mathbf{p}}\mathbf{U}_{\mathbf{p}}^{\mathbf{T}}\mathbf{y}\\ &= \left(\mathbf{V_p}\boldsymbol{\Lambda}_{\mathbf{p}}^{\mathbf{2}}\mathbf{V}_{\mathbf{p}}^{\mathbf{T}}\right)^{-\mathbf{1}}\mathbf{V_p}\boldsymbol{\Lambda}_{\mathbf{p}}\mathbf{U}_{\mathbf{p}}^{\mathbf{T}}\mathbf{y}\\ &= \mathbf{V_p}\boldsymbol{\Lambda}_{\mathbf{p}}^{-\mathbf{2}}\mathbf{V}_{\mathbf{p}}^{\mathbf{T}}\mathbf{V_p}\boldsymbol{\Lambda}_{\mathbf{p}}\mathbf{U}_{\mathbf{p}}^{\mathbf{T}}\mathbf{y}\\ &= \mathbf{V_p}\boldsymbol{\Lambda}_{\mathbf{p}}^{-\mathbf{1}}\mathbf{U}_{\mathbf{p}}^{\mathbf{T}}\mathbf{y}.\end{aligned}$$

Comparing with $\widehat{\mathbf{w}} = \mathbf{X}^{+}\mathbf{y}$, we get the pseudo-inverse $\mathbf{X}^{+} = \mathbf{V_k}\boldsymbol{\Lambda}_{\mathbf{k}}^{-\mathbf{1}}\mathbf{U}_{\mathbf{k}}^{\mathbf{T}}$, where $\mathbf{k} = \mathbf{p}$.

2. Using singular value decomposition, $\mathbf{X}$ can be represented as

$$\mathbf{X} = \mathbf{U_k \Lambda_k V_k^T},$$

where $\mathbf{k}$ is the rank of $\mathbf{X}$. As a result,

$$\begin{aligned}\mathbf{XX^T} &= \mathbf{U_k \Lambda_k V_k^T (U_k \Lambda_k V_k^T)^T} \\ &= \mathbf{U_k \Lambda_k V_k^T V_k \Lambda_k U_k^T} \\ &= \mathbf{U_k \Lambda_k^2 U_k^T}\end{aligned}$$

$$\begin{aligned}\mathbf{X^T X} &= \mathbf{(U_k \Lambda_k V_k^T)^T U_k \Lambda_k V_k^T} \\ &= \mathbf{V_k \Lambda_k U_k^T U_k \Lambda_k V_k^T} \\ &= \mathbf{V_k \Lambda_k^2 V_k^T}\end{aligned}$$

As we can see, both $\mathbf{XX^T}$ and $\mathbf{X^T X}$ have the same eigenvalue matrix $\mathbf{\Lambda_k^2}$, so they have the same eigenvalues. Also, the singular values for $\mathbf{X}$ is $\sqrt{\lambda_\mathbf{i}}$.

Multiply $\mathbf{\Lambda_k^{-1} U_k^T}$ on the left side of $\mathbf{X}$, we get

$$\begin{aligned}\mathbf{\Lambda_k^{-1} U_k^T X} &= \mathbf{\Lambda_k^{-1} U_k^T U_k \Lambda_k V_k^T} \\ \mathbf{\Lambda_k^{-1} U_k^T X} &= \mathbf{V_k^T} \\ \mathbf{V_k} &= \mathbf{X^T U_k \Lambda_k^{-1}}.\end{aligned}$$

So for each unit eigenvector $\mathbf{v_i}$, we get

$$\mathbf{v_i} = \frac{\mathbf{X^T u_i}}{\sqrt{\lambda_\mathbf{i}}}.$$

3. Since the size of the eigenvectors are the same for both $\mathbf{X^T X}$ and $\mathbf{XX^T}$, we just compare the size of $\mathbf{X^T X}$ and $\mathbf{XX^T}$ to compute a smaller one. From 1.2, it's efficient to compute this way iff $\mathbf{XX^T}$ has a smaller size, which is $\mathbf{n} \times \mathbf{n}$. So the condition is at least $\mathbf{n} < \mathbf{p}$ and would be better if $\mathbf{n} << \mathbf{p}$.

2 Simple Principal Component Analysis

2.1 Part 1: Comparing Principal Components

1. Report the eigenvectors and eigenvalues here.

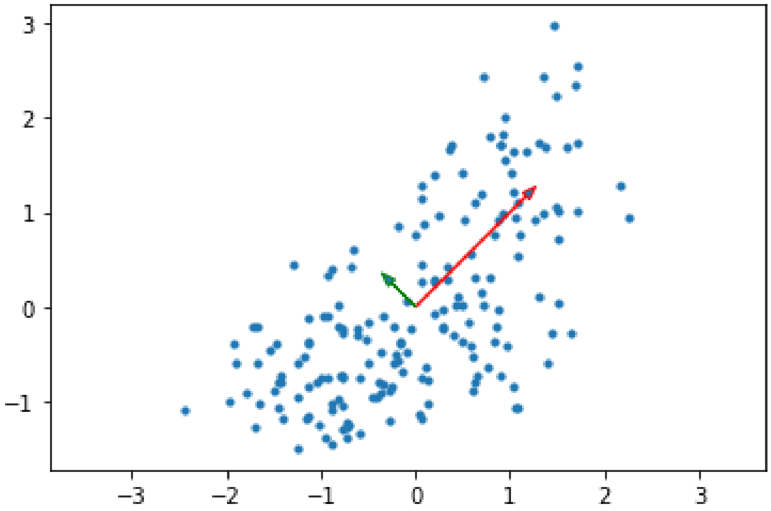
   Eigenvalues: 1.653, 0.3583
   Eigenvector 1: [0.7071,0.7071]
   Eigenvector 2: [-0.7071,0.7071]

2. What can you say about the relationship between the first principal component and the second?

   They are perpendicular.

2.2 Part 2: Plotting Principal Components in Original Space

1. Paste the plot here.

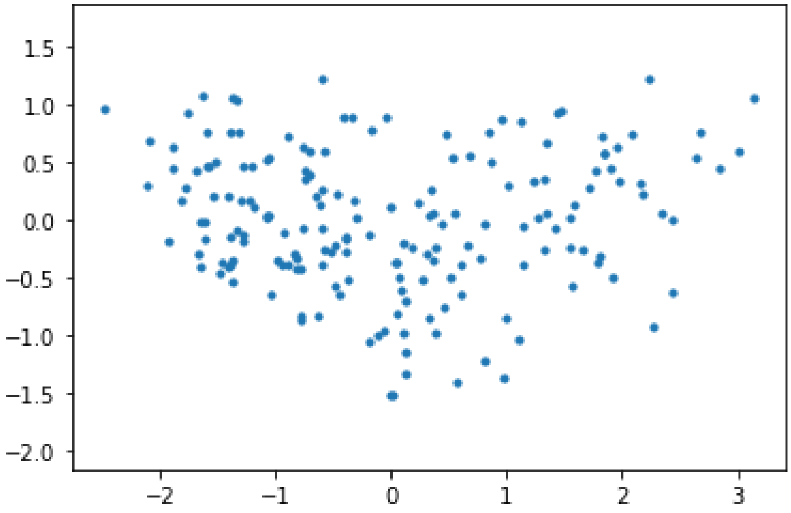


2. Describe how the principal components relate to the points in terms of variance.

   The direction of principle components maximize the variance of points.

2.3 Part 3: Plotting Data Projected onto Component Space

1. Paste the plot here.



2. Explain how the graph of points on principal component space relates to the graph of points on original space above.

   Applying a rotation operation to the graph of points on original space results in the graph of points on principal component space.

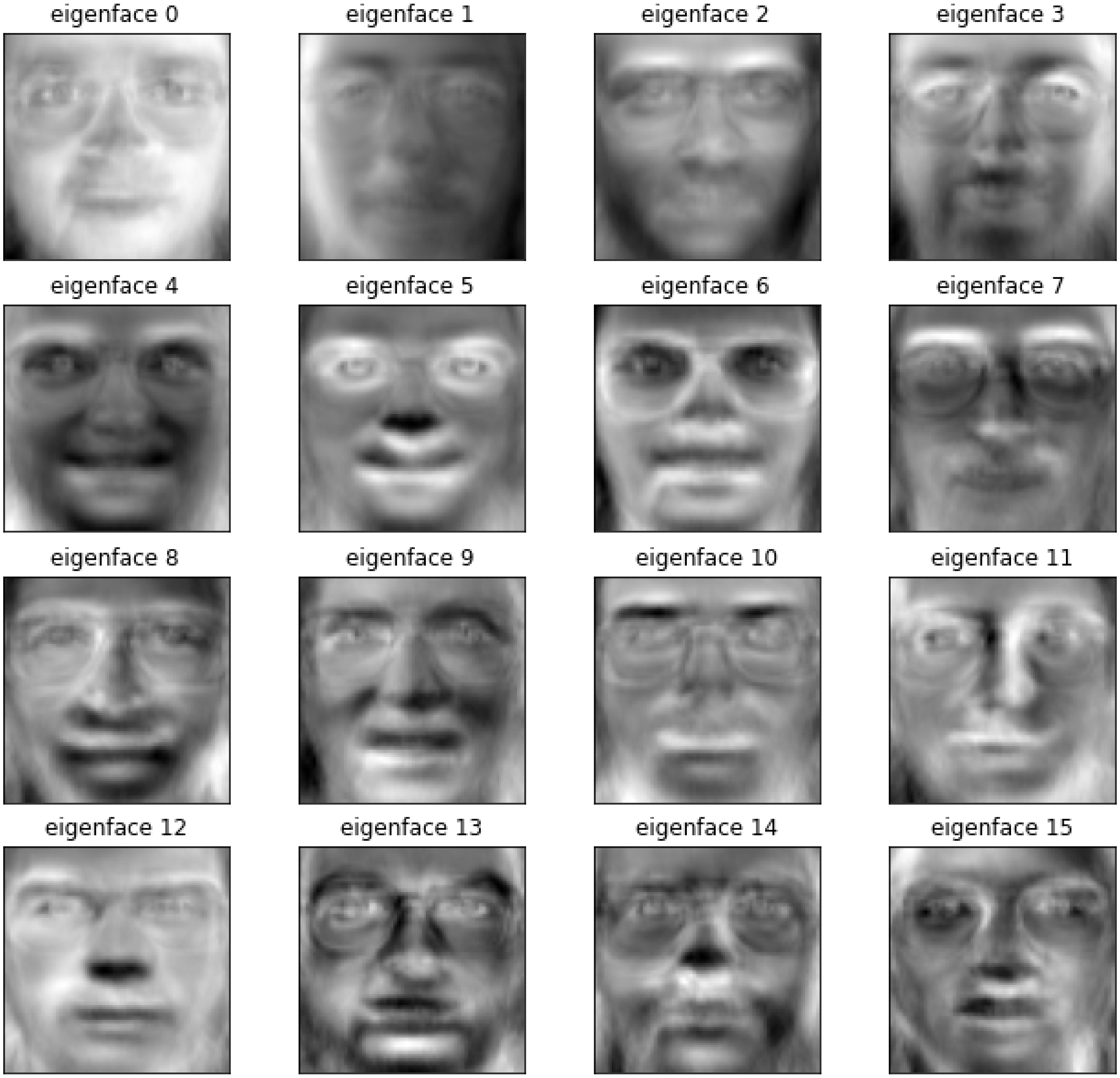
3 Principal Component Analysis on Faces

3.1 Part 1: PCA with SVD and Eigenfaces

1. Report the first five singular values here.

   [86.702 66.4653 50.1552 39.7225 33.7574]

2. Paste the eigenfaces output here.

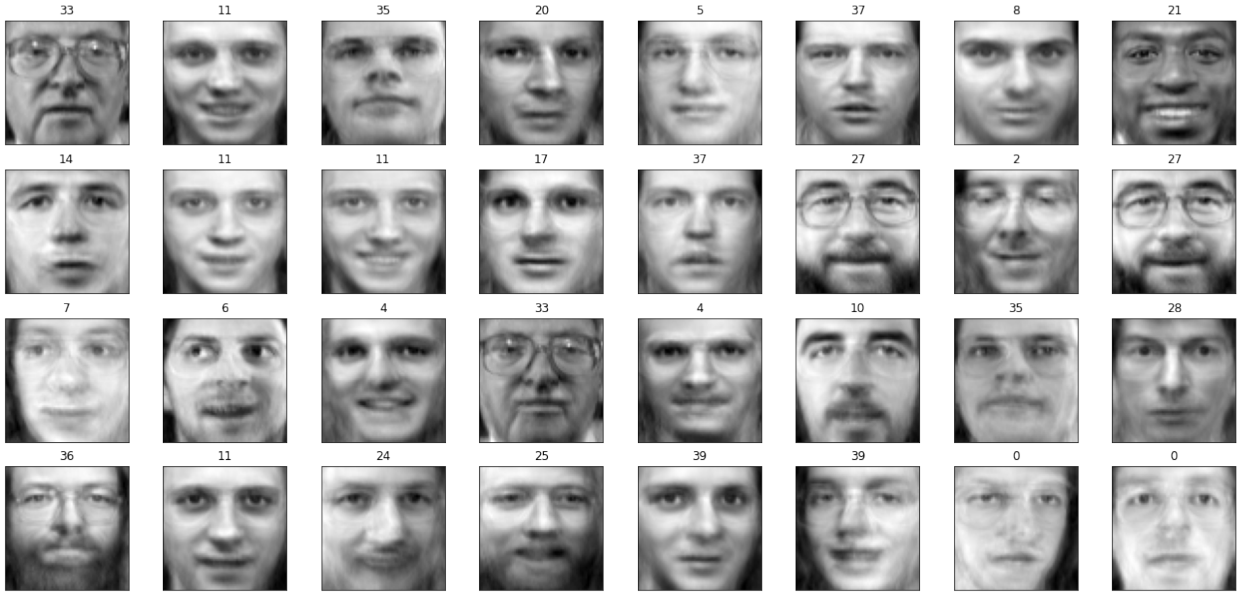


3. Describe what the eigenfaces look like. What do you expect to observe with the eigenfaces associated with larger eigenvalues?

   Each principal vector is associated with certain feature like eyebrow, nose etc. Eigenfaces associated with larger eigenvalues is more blur and has higher variance.

3.2 Part 2: Reconstructing Faces

1. Paste the portrait reconstructions here.



2. Compare the reconstructed images to the original images. How are they similar and how are they different? Shortly explain why they are different.

   The reconstructed images still have most features, like shape of eyes and noses, color of skins, from the original images, but they are less clear. Since a mean of features is added to each image, the less frequent features, like mustache, is less distinguishable, the frequent features, like glasses, introduce a shade around eyes of each face. And the female face (7) is much more masculine.

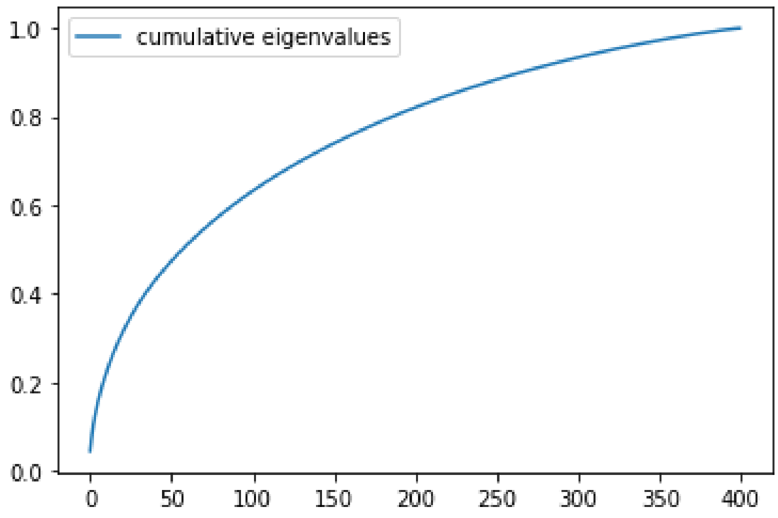
3. What do you expect to see from the reconstructed images as the number of principal components chosen for PCA increases? Please explain why.

   The lost information of reconstructed image should decrease, makeing the reconstructed images more clear and similar to the original ones. Because the more principal component we use, the reconstruction will be more similar to transforming the image to original ones using SVD.

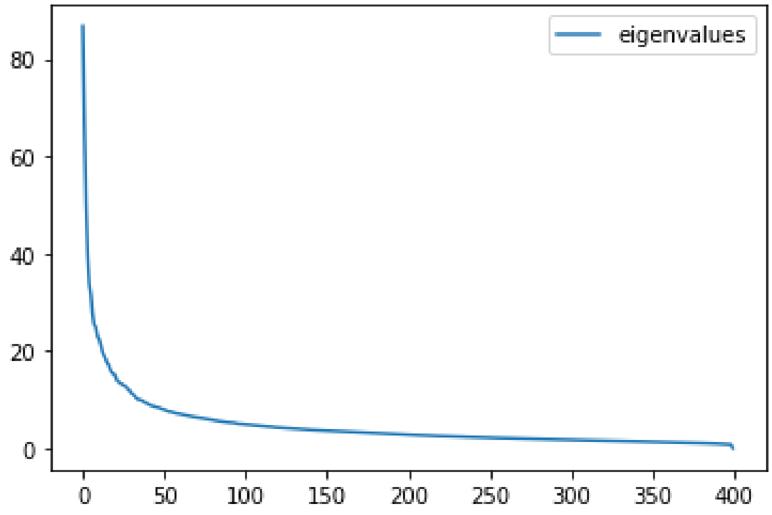
3.3 Part 3: Variance Explanation

1. Paste the plots here.

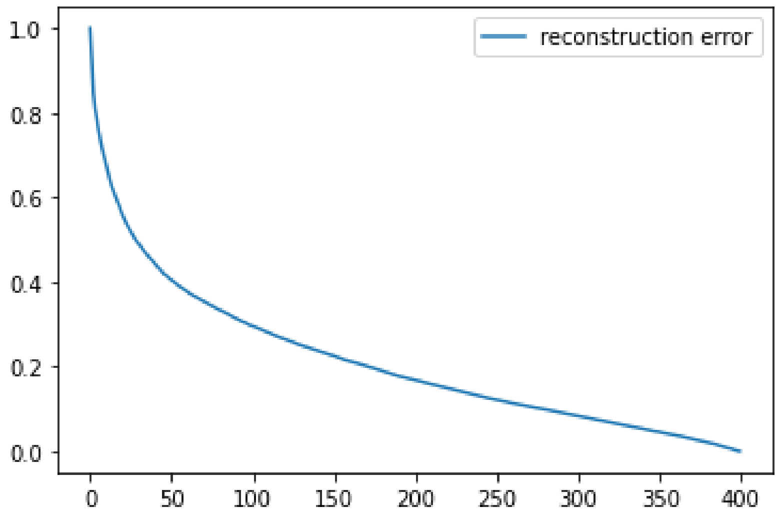
   Explanation vs. number of components plot



Eigenvalue vs. number of components plot



Reconstruction error vs. number of components plot



2. How do you expect (based on theory; please be precise!) the plot of variance explained as the number of components to relate to the eigenvalues of the corresponding components?

   The variance explained is determined by the eigenvalue corresponding to the eigenvector (principal components).

3. What is the relation between reconstruction error and the variance explained?

   As the variance explained increases, the reconstruction error decreases, the reconstructed image become more similar to the original image.