CS589: Homework 5 Report

Author: Quoc Anh Bui

May 4, 2020

PCA

(a) **Show that the direction that maximizes variance (minimizes reconstruction error) is the eigenvector corresponding to the largest eigenvalue of the Covariance matrix of the data**
Answer: We have an optimization problem (reconstruction error) below:

$$\min_w \frac{1}{N}\sum_{n=1}^{N} \|x^{(n)} - \hat{x}^{(n)}\| \qquad \text{s.t } \|w\| = 1$$

Since this is constrained optimization, we need to convert $f(x)$ to the Lagrangian form $L(x, \lambda)$. Thus:

$$\begin{aligned} L(w, \lambda) &= \frac{1}{N}\sum_{n=1}^{N} \|x^{(n)} - (w^T x^{(n)})w\|^2 + \lambda(\|w\|^2 - 1) \\ &= \frac{1}{N}\sum_{n=1}^{N} (x^{(n)} - (w^T x^{(n)})w)^T (x^{(n)} - (w^T x^{(n)})w) + \lambda(w^T w - 1) \end{aligned} \tag{1}$$

Then, we find $\frac{\partial L}{\partial w} = 0$. Note that, derivative of a summation is the summation of derivative with each components. So right now we can ignore the part $\frac{1}{N}\sum_{n=1}^{N}$. Therefore, the job is simplified down to find $\frac{\partial}{\partial w}(x - (w^T x)w)^T (x - (w^T x)w) + \lambda(w^T w - 1)$.

$$\begin{aligned} \frac{\partial}{\partial w}(x - (w^T x)w)^T (x - (w^T x)w) + \lambda(w^T w - 1) &= \frac{\partial}{\partial w}\ x^T x - 2(w^T x)^2 + (w^T x)^2 w^T w + \lambda(w^T w - 1) \\ &= \frac{\partial}{\partial w}\ x^T x - 2(w^T x)^2 + (w^T x)^2 + \lambda(w^T w - 1) \\ &= -2xx^T w + 2\lambda w \end{aligned} \tag{2}$$

Now we can introduce the summation back and set the whole thing to 0, we obtain:

$$\frac{1}{N}\sum_{n=1}^{N} x^{(n)} x^{(n)T} w = \lambda w$$

$$Cw = \lambda w$$

Thus, this suggests that w must be the eigenvector of C and the Lagrangian term is the corresponding eigenvalue.
By duality property of Lagrangian Method: $q(\lambda) \leq \inf_x L(x, \lambda) \leq f(x)$ for all x. Then the dual problem is to maximize $q(\lambda)$ and since we deduce λ is the eigenvalue of C. Thus:

$$\underset{\lambda}{\text{argmax}}\ q(\lambda) = \lambda_1$$

Where λ_1 is the largest eigenvalue of C and therefore w^* is the corresponding eigenvector (**Q.E.D**)

(b) **Show that the subspace of 2 dimensions that maximizes variance are the 2 eigenvectors corresponding to the largest 2 eigenvalues of the Covariance matrix**

Answer: Note that D components of the data are all pairwise orthogonal. Thus, the subspace of dimensions 2 that maximizes the variance consists of 2 vectors w_1 and w_2 s.t $w_1 \neq w_2$ and $w_1 \perp w_2$.

The optimization problem (reconstruction error) is defined as following:

$$\min_{w_1, w_2} \frac{1}{N} \sum_{n=1}^{N} \|x^{(n)} - (w_1^T x^{(n)}) w_1 - (w_2^T x^{(n)}) w_2\|^2 \qquad \text{s.t } \|w_1\| = 1, \quad \|w_2\| = 1$$

Similar to part (a), we construct the Lagrangian form $L(w_1, w_2, \lambda_1, \lambda_2)$:

$$L(w_1, w_2, \lambda_1, \lambda_2) = \frac{1}{N} \sum_{n=1}^{N} \|x^{(n)} - (w_1^T x^{(n)}) w_1 - (w_2^T x^{(n)}) w_2\|^2 + \lambda_1(\|w_1\|^2 - 1) + \lambda_2(\|w_2\|^2 - 1)$$

Again, we simplify the part inside the sum for a particular x:

$$\begin{aligned} \|x - (w_1^T x) w_1 - (w_2^T x) w_2\|^2 &= (x - (w_1^T x) w_1 - (w_2^T x) w_2)^T \ (x - (w_1^T x) w_1 - (w_2^T x) w_2) \\ &= x^T x - (w_1^T x)^2 - (w_2^T x)^2 + 2(w_1^T x)(w_2^T x) w_1^T w_2 \ (= 0 \text{ since } w_1 \perp w_2) \end{aligned} \tag{3}$$

Introduce back the summation:

$$\min_{w_1, w_2} \frac{1}{N} \sum_{n=1}^{N} \left(x^{(n)T} x^{(n)} - (w_1^T x^{(n)})^2 - (w_2^T x^{(n)})^2 \right) + \lambda_1(w_1^T w_1 - 1) + \lambda_2(w_2^T w_2 - 1)$$

Collect the all the same terms and we get the equivalent problem:

$$\min_{w_1} \frac{1}{N} \sum_{n=1}^{N} x^{(n)T} x^{(n)} - (w_1^T x^{(n)})^2 + \lambda_1(w_1^T w_1 - 1) \quad - \quad \min_{w_2} \frac{1}{N} \sum_{n=1}^{N} -(w_2^T x^{(n)})^2 + \lambda_2(w_2^T w_2 - 1)$$

From (a), we know the solution w_1^* is the eigenvector corresponding to the largest eigenvalue. Follow the same logic and steps in part (a), we also find that w_2^* is the eigenvector that corresponding to the second largest eigenvalue follows the assumption $w_1 \neq w_2$ (Note that $x^{(n)T} x^{(n)}$ part does not contribute to the gradient, thus does not affect the final solution). (**Q.E.D**)

(c) **Minimum eigenvectors to store X**

Answer: Since there exists a set of constants $a_1, a_2, ..., a_{D-1}$ such that the last component for every x is $x_D = \sum_{i=1}^{D-1} a_i x_i$, the last column of the dataset X is linear dependent. Thus, this would mean that X has $D-1$ rank and the Covariance matrix $C = 1/N \cdot X^T X = 1/N \cdot (U \Sigma V^T)^T (U \Sigma V^T) = 1/N \cdot (V \Sigma^2 V^T)$. There are $D - 1$ singular values that make up C that corresponds to $D - 1$ eigenvalues. Therefore, it would need $D - 1$ eigenvectors to store X perfectly.

(d) **For each k, show the projected image and plot the MSE of the reconstruction error for the dataset X as a function of k**

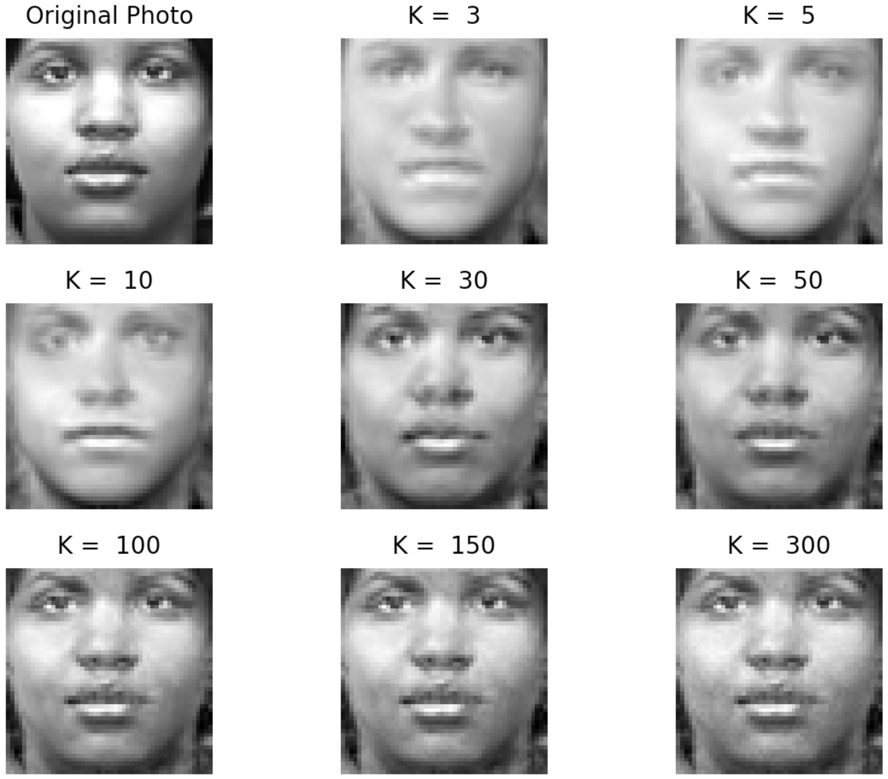


Figure 1: Projecting New Face to the subspace of k eigenvectors

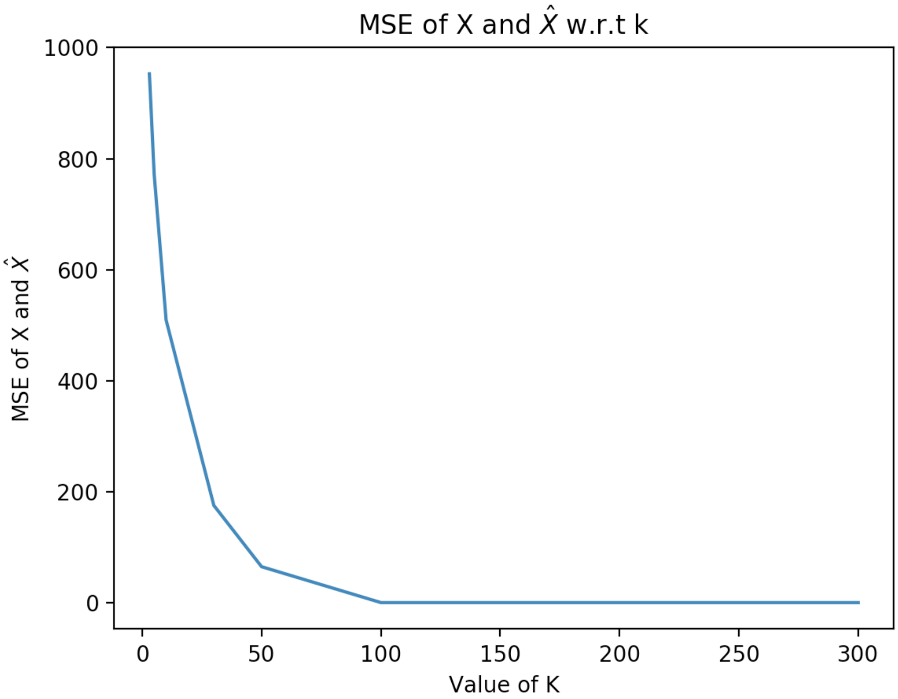


Figure 2: $\frac{1}{N}\sum_{n=1}^{N} \|x^{(n)} - \hat{x}^{(n)}\|^2$

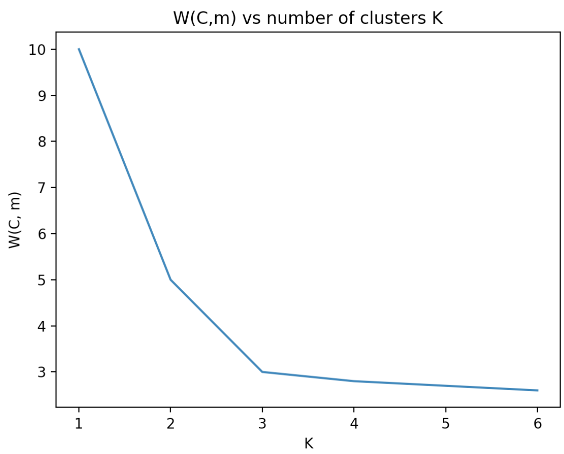
Note: The reconstruction error (MSE) takes account of mean pixels between $x^{(n)}$ and $\hat{x}^{(n)}$.

(e) **Compression rate of compressed images for different values of k:**

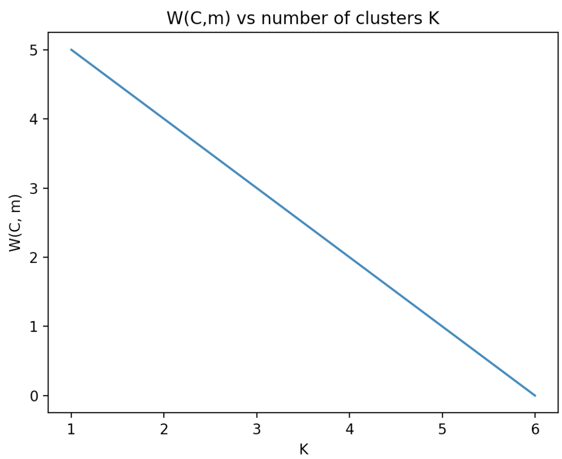
	Compression Rate
3	0.031
5	0.052
10	0.104
30	0.312
50	0.52
100	1.04
150	1.56
300	3.12

K-Means

(a) **Explain the Elbow rule for determining the "optimal" number of clusters**
Answer: The Elbow Method determines K to be the "optimal" number of clusters by making sure that $W(C_{K+1}, m_1, ..., m_{K+1})$, i.e adding 1 more cluster, is not much better than $W(C_K, m_1, ..., m_K)$. To do this, plot W over variety numbers of K and the point where the curve starts to flatten out will be the "optimal" number of clusters (Figure 1) that Elbow method suggests. A drawback of Elbow method is sometimes ambiguous; e.g the curve $W(C, m)$ is linear so that $|W(C_k) - W(C_{k+1}|$ is determined by the slop of W, implying no "flatten" point mentioned earlier (Figure 2).



(a) Figure 1: Optimal $k = 5$, $|W(C_5) - W(C_4)| \approx 0$



(b) Figure 2: $|W(C_k) - W(C_{k+1})|$ is large $\forall k$

(b) **Explain the idea behind K-means++**
Answer: Instead of randomized initialization of the centroids, K-means++ greedily initialized the centroids such that they are far apart or evenly spaced between each others. K-means++ first choose a centroid randomly and then select another centroid so that they are apart, and repeat. This could help to avoid bad initialization of randomized centroids where the centroids are packed; leading to computional inefficiency by which we have to reallocate $m_1, ..., m_K$ and potentially poor clusterings (where a group of clusters fall into local optimal).

(c) **Show the original image and report the reconstructed images for each value of k**

(d) **For each k, show the reconstruction error and the compression rate**

Answer: Using the Root Mean Squared Error provided in the code, we obtain:

	Reconstruction Error
2	70.15
5	44.5
10	31.33
25	22.41
50	17.88
100	14.15
200	11.15

	Compression Rate
2	0.042
5	0.097
10	0.139
25	0.194
50	0.237
100	0.28
200	0.325