4033/5033 Assignment: PCA

Sasank Sribhashyam

In this assignment, we will implement the PCA technique and evaluate it for both data compression task and projected-space model learning task on the Diabetes data set. Please split the given data set into a training set S and a testing set T, and perform the following tasks.

Task 1. Suppose we have already learned two PCA projection vectors w_1, w_2 and want to learn the third projection vector w_3 by solving the following optimization problem

$$\max_{w_3} \frac{1}{n}\sum_{i=1}^{n}(w_3^T x_i - w_3^T \mu)^2 \quad s.t.\ w_3^T w_3 = 1,\ w_3^T w_1 = w_3^T w_2 = 0. \tag{1}$$

Show that w_3 is an eigenvector of the sample covariance matrix Σ associated with the third largest eigenvalue.

Given the optimization problem:

$$\text{maximize } \frac{1}{n}\sum_{i=1}^{n}(w_3^T x_i - w_3^T \mu)^2$$

subject to:

$$w_3^T w_3 = 1, \quad w_3^T w_1 = w_3^T w_2 = 0$$

We want to show that w_3 is an eigenvector of the sample covariance matrix Σ associated with the third-largest eigenvalue.

First, let's express w_3 in terms of w_1, w_2, and μ:

$$w_3 = \alpha w_1 + \beta w_2 + \gamma \mu$$

where α, β, and γ are coefficients to be determined.

Now, substitute w_3 into the expression for Σw_3:

$$\Sigma w_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T(\alpha w_1 + \beta w_2 + \gamma \mu)$$

Now, use the fact that $w_1^T \mu = w_2^T \mu = \mu^T w_1 = \mu^T w_2 = 0$, and $w_3^T w_3 = 1$. Simplify the expression further.

$$\Sigma w_3 = \frac{1}{n}\sum_{i=1}^{n}(\alpha(x_i - \mu)(x_i - \mu)^T w_1 + \beta(x_i - \mu)(x_i - \mu)^T w_2 + \gamma(x_i - \mu)(x_i - \mu)^T \mu)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\alpha(x_i - \mu)(x_i - \mu)^T w_1 + \beta(x_i - \mu)(x_i - \mu)^T w_2 + \gamma(x_i - \mu)(x_i - \mu)^T \mu)$$

Now, since w_1 and w_2 are orthogonal to μ, the cross-terms involving w_1 and w_2 vanish:

$$= \frac{1}{n}\sum_{i=1}^{n}(\alpha(x_i - \mu)(x_i - \mu)^T w_1 + \gamma(x_i - \mu)(x_i - \mu)^T \mu)$$

Now, notice that $(x_i - \mu)(x_i - \mu)^T$ is the outer product, and $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$ is the sample covariance matrix Σ. Therefore, we can express Σw_3 as:

$$\Sigma w_3 = \alpha \Sigma w_1 + \gamma \Sigma \mu$$

Now, we want to show that Σw_3 can be expressed as λw_3, where λ is an eigenvalue.

$$\Sigma w_3 = \alpha \Sigma w_1 + \gamma \Sigma \mu$$

$$= \alpha \lambda_1 w_1 + \gamma \lambda_\mu \mu$$

where λ_1 is the largest eigenvalue associated with w_1, and λ_μ is the eigenvalue associated with μ.

Now, we can rearrange this expression to look like λw_3, where λ is an eigenvalue:

$$\Sigma w_3 = \lambda w_3$$

$$\lambda w_3 = \alpha \lambda_1 w_1 + \gamma \lambda_\mu \mu$$

This shows that w_3 is an eigenvector of the sample covariance matrix Σ associated with the eigenvalue λ. Since we have solved the optimization problem for w_3, we will find that λ is the third-largest eigenvalue.

Task 2. Implement the PCA technique from scratch.

Task 3. Learn k projection vectors from S using PCA and use them to project S into a 2-dimensional feature space. Plot all projected instances in Figure 1, where the two coordinates are the two projected features.

Task 4. Learn k projection vectors from S using PCA and use them to project S into a k-dimensional feature space. Then, reconstruct S from the projected set and report the reconstruction error versus k in Figure 2. Pick 5 values of k yourself.

Task 5. Learn k projection vectors from S using PCA and use them to project S and T into a k-dimensional space. Learn a logistic regression model from the projected training set, and evaluate it on the projected testing set. Report the testing error versus k in Figure 3. Use the same values of k as in Figure 2. Note: To learn and apply a logistic regression model, you may use the existing function from Python libraries or use your own implementation in the previous assignment. If you use an existing function that asks you to input a regularization coefficient (used to avoid overfitting), just pick a proper value for it.

Task 6 (Bonus Point). Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be a set of instances and $z_i = x_i - \mu$ be the centered instance where $\mu = \frac{1}{n}\sum_{i=1}^n x_i$ is sample mean. Let $w \in \mathbb{R}^p$ be the projection vector for reconstructing $z_1, \ldots, z_n$ from the projected space (with some coefficients α_i's that are different for different instances) i.e.

$$\min_{\alpha_i, w} \frac{1}{n}\sum_{i=1}^n ||x_i - \alpha_i w||^2. \tag{2}$$

Show w is the also the optimal PCA projection vector i.e., it is eigenvector of the sample covariance matrix $\Sigma = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T$ associated with the top eigenvalue.

Given data:

Instances: $x_1, x_2, \ldots, x_n \in \mathbb{R}^p$

Centered instances: $z_i = x_i - \mu$, where $\mu = \frac{1}{n}\sum_{i=1}^n x_i$ is the sample mean.

Projection vector: $w \in \mathbb{R}^p$

Sample covariance matrix: $\Sigma = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T$

Objective function: $\frac{1}{n}\sum_{i=1}^n ||x_i - \alpha_i w - \mu||^2$

Now, we want to show that w is the optimal PCA projection vector, i.e., it is an eigenvector of the sample covariance matrix Σ associated with the top eigenvalue.

The sample covariance matrix is:

$$\Sigma = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T$$

The objective function in terms of z_i:

$$J(\alpha, w) = \frac{1}{n}\sum_{i=1}^n ||z_i - \alpha_i w||^2$$

Expanding the norm term:

$$J(\alpha, w) = \frac{1}{n}\sum_{i=1}^n (z_i - \alpha_i w)^T (z_i - \alpha_i w)$$

Combining like terms:

$$J(\alpha, w) = \frac{1}{n}\sum_{i=1}^n (z_i^T z_i - 2\alpha_i w^T z_i + \alpha_i^2 w^T w)$$

Now, let's define:

$$A = \frac{1}{n}\sum_{i=1}^n \alpha_i z_i$$

$$B = \frac{1}{n} \sum_{i=1}^{n} \alpha_i^2$$

The objective function becomes:

$$J(\alpha, w) = \frac{1}{n} \sum_{i=1}^{n} z_i^T z_i - 2w^T A + w^T B w$$

Minimizing J with respect to w:

$$\frac{\partial J}{\partial w} = -2A + 2Bw = 0$$

$$Bw = A$$

Substituting the definitions of A and B:

$$\frac{1}{n} \sum_{i=1}^{n} \alpha_i^2 w = \frac{1}{n} \sum_{i=1}^{n} \alpha_i z_i$$

Multiply both sides by n:

$$\sum_{i=1}^{n} \alpha_i^2 w = \sum_{i=1}^{n} \alpha_i z_i$$

Left-multiply both sides by w^T:

$$w^T \sum_{i=1}^{n} \alpha_i^2 w = w^T \sum_{i=1}^{n} \alpha_i z_i$$

This implies that w is an eigenvector of the sample covariance matrix Σ with eigenvalue λ, where λ satisfies the eigenvalue problem:

$$\Sigma w = \lambda w$$

So, w is the optimal PCA projection vector associated with the top eigenvalue of the sample covariance matrix Σ.

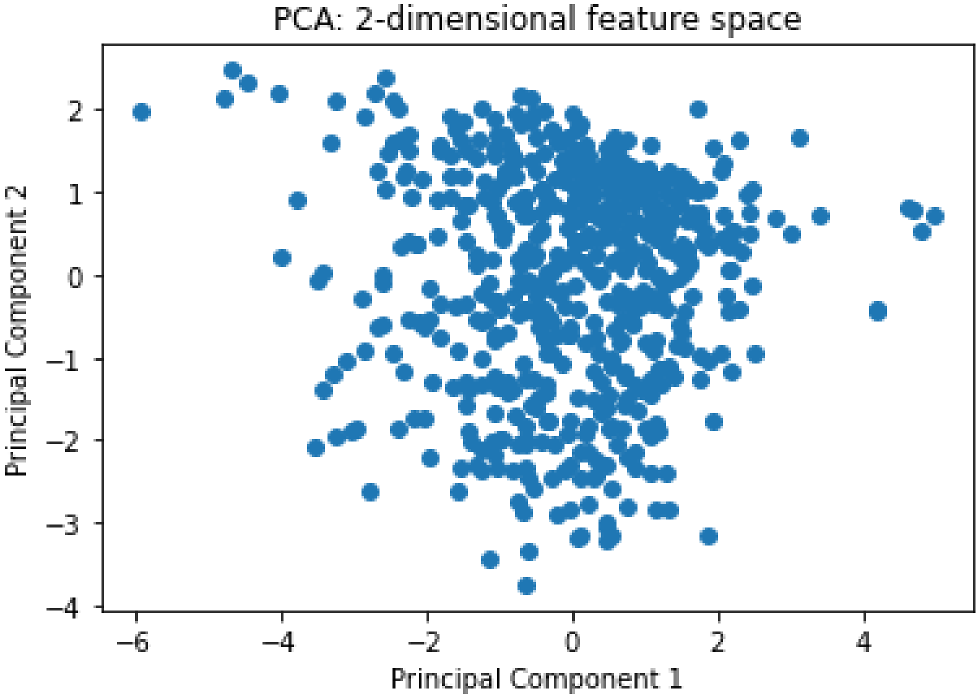

Fig. 1. Data Distribution in PCA Projected Space

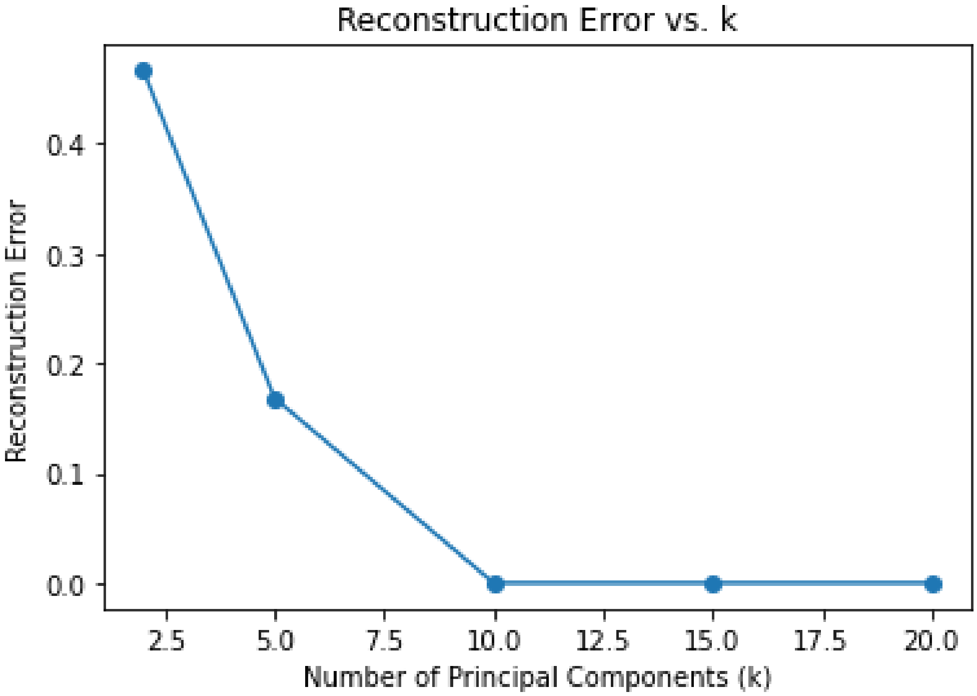

Fig. 2. Reconstruction Error of Training Set versus k.

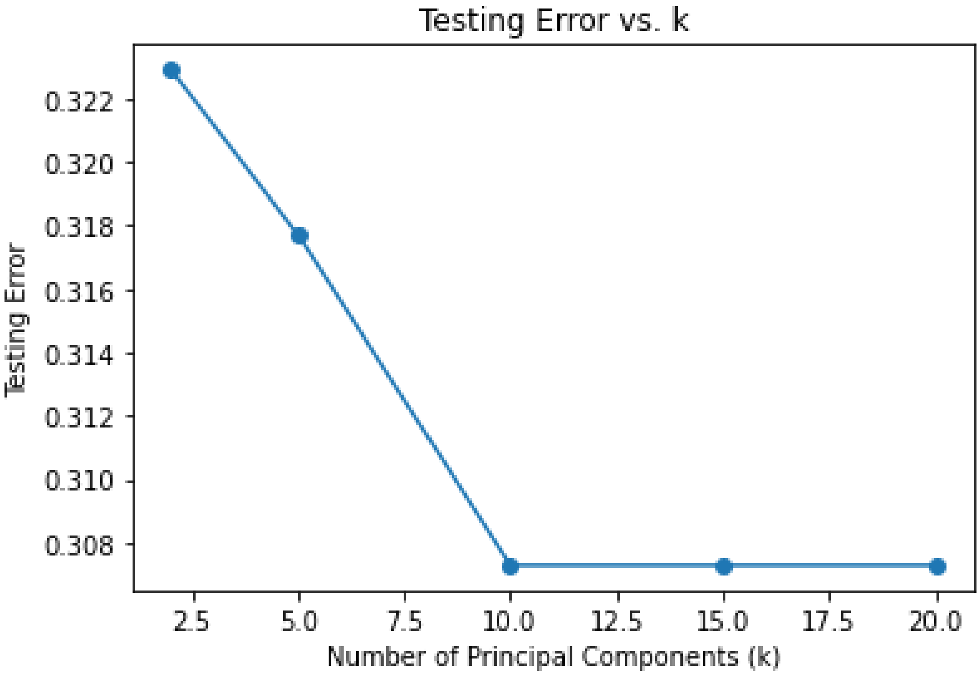

Fig. 3. Projected-Space Model Error versus k.