ECE 586 Application: Principal Component Analysis (PCA)

Henry D. Pfister
Duke University

November 14th, 2019

1 A Few Questions

1.1 What affine subspace best approximates a given set of points in $\mathbb{R}^n$?

For a given set of N data points $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N \in \mathbb{R}^n$, what p-dimensional affine subspace $W \subset \mathbb{R}^n$ best approximates these points in the sense that the error

$$\frac{1}{N} \sum_{i=1}^{N} \| \underline{x}_i - P_W(\underline{x}_i) \|^2$$

is minimized. We refer to this as the *empirical PCA problem*. Recall that an *affine subspace* $W = \text{span}\{\underline{w}_1, \dots, \underline{w}_p\} + \underline{w}_0$ is defined by linearly independent vectors $\underline{w}_0, \dots, \underline{w}_p$ where adding $\underline{w}_0$ has the effect of translating the subspace $\text{span}\{\underline{w}_1, \dots, \underline{w}_p\}$. For such an affine subspace, we define $B \in \mathbb{R}^{n \times p}$ by $B = [\underline{w}_1, \dots, \underline{w}_p]$ and the projection onto W is defined by $P_W(\underline{x}) = B(B^T B)^{-1} B^T (\underline{x} - \underline{w}_0) + \underline{w}_0$. Also, choosing $\underline{w}_0 = \frac{1}{N} \sum_{i=1}^{N} \underline{x}_i$ is optimal because this corresponds to first removing the mean and then solving the problem for a set of zero-mean vectors.

Let $A = [\underline{x}_1 - \underline{w}_0, \dots, \underline{x}_N - \underline{w}_0]$ be the mean-corrected data matrix for the given data points. Then, the solution to this problem can be computed using the SVD $A = U\Sigma V^T$ of A. In particular, we can choose $B = U_p \triangleq [\underline{u}_1, \dots, \underline{u}_p]$ to be the first p columns of U so that $P_W(\underline{x}) = U_p U_p^T (\underline{x} - \underline{w}_0) + \underline{w}_0$. When used for dimension reduction, the idea is to store $\underline{y}_i = U_p^T \underline{x}_i \in \mathbb{R}^p$ instead of $\underline{x}_i \in \mathbb{R}^n$.

1.2 What is the best p-variable approximation of n random variables?

Let $\underline{X} = (X_1, X_2, \dots, X_n)^T$ be a vector of n real random variables and let $U_p = [\underline{u}_1, \dots, \underline{u}_p]$ be a matrix with orthogonal columns whose i-th column is $\underline{u}_i \in \mathbb{R}^n$. Consider the linear transformation to $\underline{Y} = (Y_1, Y_2, \dots, Y_p)^T$ defined by

$$Y_i = [U_p^T \underline{X}]_i = \sum_{j=1}^{n} u_{j,i} X_j.$$

For the random vector $\underline{Y}$, one can show that the covariance is given by

$$\text{Cov}(Y_i, Y_k) = \sum_{j=1}^{n} \sum_{l=1}^{n} u_{j,i} \text{Cov}(X_j, X_l) u_{l,k} = \underline{u}_i^T K \underline{u}_k,$$

where $\text{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ and $[K]_{j,l} = \text{Cov}(X_j, X_l)$. For $p < n$, the key question is "How can we choose $\underline{u}_1, \dots, \underline{u}_p$ to maximize the variance in $X_1, X_2, \dots, X_n$ that is explained by $Y_1, \dots, Y_p$?". We refer to this as the *probabilistic PCA problem.*

Mathematically, this question can be formalized by the sequential computation, for $i = 1, \dots, p$, of the vectors

$$\underline{u}_i = \arg \max_{\underline{u} \in \mathbb{R}^n : \|\underline{u}\| = 1} \text{Var}(Y_i) \quad \text{subject to } \text{Cov}(Y_i, Y_k) = 0 \; \forall k \in \{1, \dots, i-1\}$$

$$= \arg \max_{\underline{u} \in \mathbb{R}^n : \|\underline{u}\|=1} \underline{u}^T K \underline{u} \quad \text{subject to } \underline{u}^T K \underline{u}_k = 0 \; \forall k \in \{1, \ldots, i-1\}.$$

Since K is symmetric and symmetric matrices have a complete set of orthogonal eigenvectors, the second line is solved by choosing $\underline{u}_i$ to be the i-th normalized eigenvector of K (i.e., associated with the i-th largest eigenvalue). Thus, the implied transformation (i.e., dimension reduction) is given by $\underline{Y} = U_p^T \underline{X}$.

1.3 How are these two problems related?

While these two problems may seem quite different (e.g., one is deterministic and the other is probabilistic), they are actually quite similar. To see this, we observe that each vector $\underline{x}_i$ in the first problem can be seen as a sample of the random vector $\underline{X} = (X_1, \ldots, X_n)^T$ in the second problem. Then, focusing on the second problem, we can assume the distribution of the random vector $\underline{X}$ is the empirical distribution defined by the data set in the first problem. Using this interpretation, the empirical covariance matrix of the data is given by

$$\hat{K} = \frac{1}{N} \sum_{i=1}^{N} \underline{x}_i \underline{x}_i^T = \frac{1}{N} A A^T,$$

where A is the mean-corrected data matrix from the first problem. Using the SVD $A = U \Sigma V^T$, it follows that

$$\hat{K} = \frac{1}{N} U \Sigma V^T V \Sigma^T U^T = U \left(\frac{1}{N} \Sigma^2 \right) U^T.$$

From this, we see that the columns of $U = [\underline{u}_1, \ldots, \underline{u}_n]$ define a set n orthogonal eigenvectors for $\hat{K}$ where $\hat{K} \underline{u}_i = \left(\frac{1}{N} \sigma_i^2\right) \underline{u}_i$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. Thus, if we solve the second problem by finding the p eigenvectors of $\hat{K}$ with largest eigenvalues, then the desired dimension reduction is given by $\underline{Y} = U_p^T \underline{X}$ where $U_p \triangleq [\underline{u}_1, \ldots, \underline{u}_p]$. But, based on this, the U_p matrix from the second problem is identical to the U_p matrix computed in the first problem! Finally, by treating $\underline{x}_i$ as a realization of $\underline{X}$ and applying the dimension reduction, one gets a sample $\underline{y}_i = U_p^T \underline{x}_i$ of the random vector $\underline{Y}$.

2 Exercises

The following exercises use the MNIST dataset to highlight PCA as an application of linear algebra for dimension reduction.

Exercise 2.1. Use the steps outlined in Section 1.1 to solve the empirical PCA problem for the MNIST digit 2. Plot the first 30 reconstruction vectors (i.e., the first 30 columns of V in the SVD) as small MNIST images. How many of these look a lot like the digit 2?

Exercise 2.2. Do the following (separately) for each of the 10 digits in the given MNIST dataset:

1. Use the given samples to solve the empirical PCA problem *with* mean removal as outlined in Section 1.1.
2. Use the given samples to solve the empirical PCA problem *without* mean removal (i.e., set $\underline{w}_0 = \underline{0}$) using the steps in Section 1.1.
3. Plot the first 15 image samples of this digit in three rows:
   (a) the original image
   (b) the image after projection onto $p = 3$ dimensions (with mean removal)
   (c) the image after projection onto $p = 4$ dimensions (without mean removal)
4. For the case without mean removal, plot the magnitude of the first 12 singular values in decreasing order using a logarithmic scale for the y-axis.