Principal Component Analysis (PCA)

Group 5: John, Dieu-Donne, Khady, Mame Diarra Diouf

African Master in Machine Intelligence, AMMI

Supervised by : Prof. MOUSTAPHA Cisse

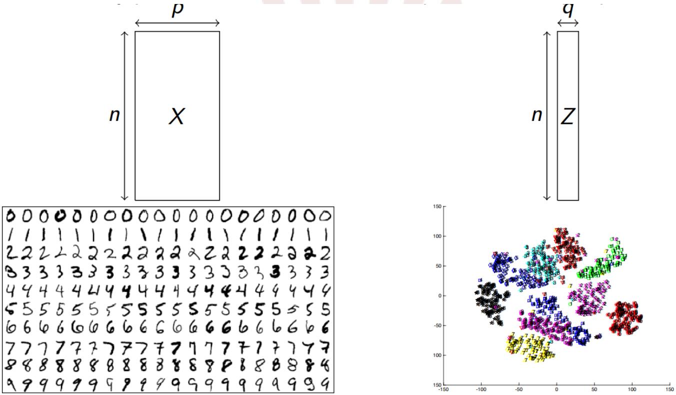
April 6, 2023

AIMS African Institute for Mathematical Sciences SENEGAL

Overview

Motivation

let X be the dataset with n individus and $p >> 2$ variable.



$p = 784$ $\quad$ $q = 2$

Crédit : CSC 2535: 2013 Lecture 11 Non-linear dimensionality reduction Geoffrey Hinton
Nonlinear dimensionality reduction JA Lee, M Verleysen, Springer, 2007

What is PCA

Principal Component Analysis (PCA) is an exploratory statistical for reducing the dimensionality of a dataset.
The goal is to find a low dimension representation that retains as much possible the information contained in our data. PCA allows the

- visualization of multidimensional and understanding of data
- extract the most valuable characteristics of the data (most useful information)
- reduce the learning time

Overview

Maximum Variance Formulation

We aim to find the principal components that maximises the variance of the projected data.
Consider a set of observation X_n where n=1,2...N and $X_n \in \mathbf{R}^D$.
PCA aims in find in an orthogonal projection of X_N on the space with dimensions $M < D$
Lets consider a projecting vector w_1 a unit vector whose norm is 1
Generally the projected data is

$$\hat{X}_n = w_1^T X_n \quad \text{and} \quad \bar{\hat{X}}_n = w_1^T \bar{X}_n \tag{2.1}$$

Maximum Variance Formulation

$$\sigma^2(\hat{X}_n) = \frac{1}{N}\sum_{n=1}^{N}(\hat{X}_n - \bar{\hat{X}}_n)^2 \quad = \quad \frac{1}{N}\sum_{n=1}^{N}(w_1^T X_n - w_1^T \bar{X}_n)^2$$

$$\sigma^2(\hat{X}_n) = \frac{1}{N}\sum_{n=1}^{N}(w_1^T X_n - w_1^T \bar{X}_n)(w_1^T X_n - w_1^T \bar{X}_n)^T$$

$$= \frac{1}{N}\sum_{n=1}^{N} w_1^T (X_n - \bar{X}_n)(X_n - \bar{X}_n)^T w_1$$

$$= w_1^T \frac{1}{N}\sum_{n=1}^{N}(X_n - \bar{X}_n)(X_n - \bar{X}_n)^T w_1$$

Maximum Variance Formulation

We observe that

$$S = \frac{1}{N}\sum_{n=1}^{N}(X_n - \bar{X}_n)(X_n - \bar{X}_n)^T \quad \text{is a covariance matrix} \tag{2.2}$$

Therefore we have

$$\sigma^2(\hat{X}_n) = w_1^T S w_1 \tag{2.3}$$

Since PCA aims to maximise the variance of the projected data , the optimization problem for PCA is

$$\begin{aligned} &\underset{w_1}{\text{maximise}} \quad (w_1^T S w_1) \\ &\text{subject to} \quad ||w_1||_2 = 1 \end{aligned} \tag{2.4}$$

Maximum Variance Formulation

Optimization objective

We introduce a new variable lagrange multiplier λ such that our maximization problem becomes

$$J(w_1, \lambda) = w_1^T S w_1 - \lambda(w_1^T w_1 - 1) \tag{2.5}$$

The partial derivatives of equation 2.5 with respect to λ and w_1 we have

$$\begin{cases} \frac{\partial J(w_1,\lambda)}{\partial w_1} = 2Sw_1 - 2\lambda w_1 \\ \frac{\partial J(w_1,\lambda)}{\partial \lambda} = w^T w - 1 \end{cases} \tag{2.6}$$

set equation 2.6 to 0 and evaluate the stationary point we have

Maximum Variance Formulation

$$\begin{cases} \frac{\partial J(w_1,\lambda)}{\partial w_1} = 0 \\ \frac{\partial J(w_1,\lambda)}{\partial \lambda} = 0 \end{cases} \Longrightarrow \begin{cases} 2Sw_1 - 2\lambda w_1 & = 0 \\ w^T w - 1 & = 0 \end{cases} \tag{2.7}$$

We therefore have

$$\begin{cases} Sw_1 & = \lambda w_1 \\ w^T w & = 1 \end{cases} \tag{2.8}$$

We are interested with the equation

$$Sw_1 = \lambda w_1 \tag{2.9}$$

Maximum Variance Formulation

Evaluated at the stationary point , this shows that w_1 must be an eigen vector of covariance matrix S and λ is the eigen value of associated with eigen vector w_1

The matrix of eigen vectors gives the principal components associated with the normalised data set with the largest eigen value corresponding to the eigen vector that leads to the principal component explaining most of the variation in the projected data.

Overview

1. Motivation
2. Background
   - Background
3. How PCA works ?
   - How PCA works ?
4. Loadings and Scores
   - Loadings and Scores
5. Conclusion

Normalization of the Data

PCA is a practice to change the direction of components to maximum variance directions. Basically, the original data will be allocated in different directions that maximize the variance. If we don't normalize our dataset, the PCA technique can be bias towards specific features.

To do this, we first of all, subtracting the respective means of the variable from every data point. This produces a dataset whose mean is zero. After that, dividing every data point to standard deviation.

Calculate the covariance matrix

The covariance matrix plays a crucial role in PCA because it provides information on the relationships among the variables and enables us to identify the most important components of variation in the data. Let S our covariance matrix of X, so:

$$S = \frac{1}{N}\sum_{n=1}^{N}(X_n - \bar{X})(X_n - \bar{X})^T = \frac{X^T X}{N-1} \tag{3.1}$$

Calculate EigenValues

Eigenvalue is a scalar that is used to transform (stretch) an Eigenvector. From our optimization problem, we obtain what we call eigenvalues by the following relation:

$$Sw_1 = \lambda w_1 \tag{3.2}$$

Where S is a square matrix (which is covariance matrix of X), w_1 is an eigenvector, λ is a scaler which is eigenvalue associated with eigenvector of S matrix.

To determine the eigenvalues, we solve the following equation :

$$det(S - \lambda I) = 0 \tag{3.3}$$

From this we determine the eigenvalues matrix.

Calculate EigenVectors

Intuitively, an eigenvector is a vector whose direction remains unchanged when a linear transformation is applied to it. After finding Eigenvalues, now is the time for finding eigenvectors. You may remember our equation in the beginning was:

$$Sw_1 = \lambda w_1$$

For every λ value, we will find different eigenvectors through the following equation:

$$(S - \lambda I)w_1 = 0 \tag{3.4}$$

Sort the eigenvectors and select

- Sort the eigenvectors by corresponding to decreasing eigenvalues and choose first eigenvectors with the largest eigenvalues, we call this eigenvectors the principal components PCs.
- Next we form a feature vector which is a matrix of vectors, in our case, the eigenvectors. In fact, only those eigenvectors which we want to proceed with.

Some rules to choose the Principal Components PCs

Empirical rules

- The best known is the **Kaiser rule**: for standardized data, principal components corresponding to eigenvalues larger than 1 are retained; this means only components which 'bring' more than original variables are of interest.

- It is also usual to employ a scree test, which consists of detecting the existence of a significant decay (biggest drop) on the eigenvalues diagram. This is not always easy in practice.

Create final data points with an PCs matrix

The end goal of Transform the real dataset onto the new subspace.

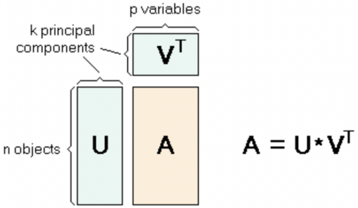
$$\textit{Transformed} \text{ Dataset} = (\textit{Normalised } X \text{ matrix}) \;.\; (\text{PCs } \textit{Matrix})$$

Overview

AIMS African Institute for Mathematical Sciences SENEGAL

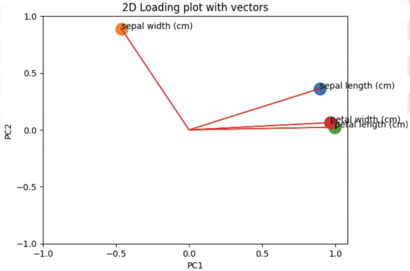
Loadings and Scores

- If we look at PCA more formally, it turns out that the PCA is based on a decomposition of the data matrix A into two matrices V and U:



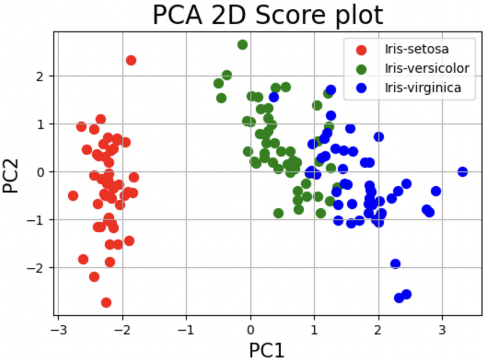
Loadings

- Suppose that after applying Principal Component Analysis (PCA) to your dataset, you are interested in understanding which is the contribution of the original variables to the principal components. How can we do that?.



Scores

- Scores describe the coordinates of each data point in the new coordinate system defined by the principal components.



Conclusion

- We can conclude that PCA is a powerful tool for reducing the dimensionality of high-dimensional data while retaining as much of the original information as possible ;
- By identifying the principal components that capture the most variation in the data, PCA allows us to represent complex data in a more simplified and interpretable form.


THANK YOU !