Software Development for Data Analysis

Principal Component Analysis (PCA)

- A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- The analyzed data consist in a table of observations, having ***n*** rows and ***m*** columns.

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \cdots & & \\ x_{n1} & .. & x_{nm} \end{bmatrix}$$

, where x_{ij} is the value taken by variable j for the observation i.

- The variable described by table X are also known as *initial, causal or observed variables*.

Principal Component Analysis (PCA)

- $\boldsymbol{X_j}$ is the column vector containing the values of variable $\boldsymbol{j}$ for $\boldsymbol{n}$ observations;
- The goal of the procedure is to describe table X through a reduced number of noncorelated variables: C_1, C_2, ..., C_s.

Phase 1

Determine a new variable C_1, the first principal component, as linear combination of variables X_j:

$$C_1 = a_{11}X_1 + ... + a_{j1}X_j + ... + a_{m1}X_m$$

The value taken by C_1 for a given observation i :

$$c_{i1} = a_{11}x_{i1} + ... + a_{j1}x_{ij} + ... + a_{m1}x_{im}$$

where $a_{j1}, j = \overline{1,m}$

Principal Component Analysis (PCA)

Phase k

Determine a new variable C_k, the k principal component, as linear combination of variables X:

$$C_k = a_{1k}X_1 + ... + a_{jk}X_j + ... + a_{mk}X_m \quad ,$$

where a_k is the vector containing the multipliers $a_{jk}, j = \overline{1,m}$

The link between the causal variables (X) and the principal components (C) is given by:

$C_k = X \cdot a_k,\ \ k=1,s$, where $\boldsymbol{s}$ is the number of principal components.

Principal Component Analysis (PCA)

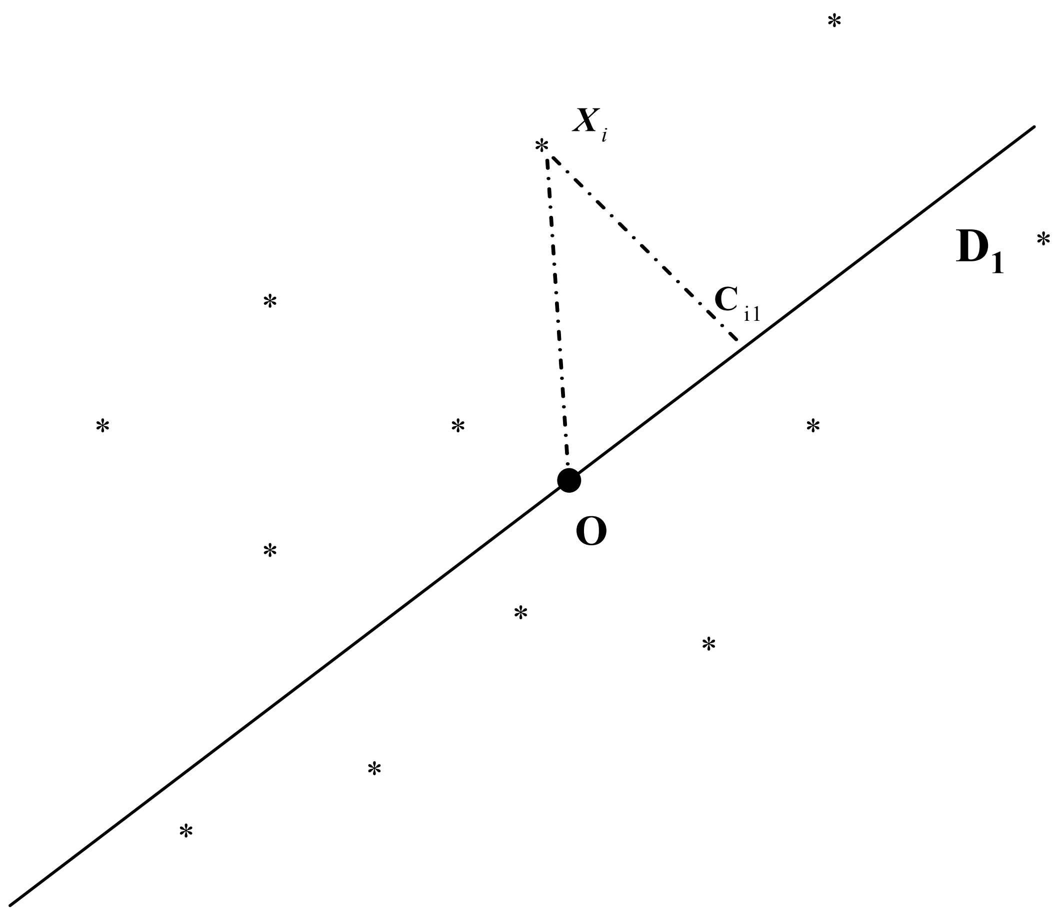
Observation driven approach

- The cloud of observations has ***n*** points within a ***m***-dimensional space;
- Those ***m*** variables determine the ***m*** axis of coordinates;
- If the data is standardized, then the variables have the mean 0, and the standard deviation 1;
- Consider a system of orthonormal axes (it is orthogonal and having the norm 1) for those ***n*** points;
- Each axis corresponds to one principal component, and the vectors a_k are unit vectors (in a normed vector space, it is a vector, often a spatial vector, of length 1):

$$\sum_{j=1}^{m} {a_{kj}}^2 = 1, k = \overline{1,s}$$, where ***s*** is the maximum number of axes

Principal Component Analysis (PCA)

Observation-driven approach: projection on D_1 axis



Principal Component Analysis (PCA)

Observation driven approach

Step 1

- Determine first axis, corresponding to the first principal component, so the component's variance is maxim;
- **O** is the center of gravity for the cloud of points;
- The distance from the point (observation) X_i to the D_1 axis, corresponding to the first principal component is $d(i, D_1)$;
- The distance from X_i to origin **O** is $d(i, \mathbf{O})$.

Then we have the following relation between distances in the corresponding right-triangle:

$d(i, \mathbf{O})^2 = d(i, D_1)^2 + c_{i1}{}^2$, where c_{i1} is the projection of X_i on D_1 axis.

Principal Component Analysis (PCA)

Observation driven approach

- Therefore, for all the points in the cloud we have the following equality of sums of square distances (SSD):

$$\frac{1}{n}\sum_{i=1}^{n} d(i,O)^2 = \frac{1}{n}\sum_{i=1}^{n} d(i,D_1)^2 + \frac{1}{n}\sum_{i=1}^{n} c_{i1}^2$$

Principal Component Analysis (PCA)

Observation driven approach

- The sum of square distances toward the center of gravity (*barycenter*) does not depend on the chosen axis;
- The *variance explained* through axis 1 (D_1) is $\frac{1}{n}\sum_{i=1}^{n} c_{i1}^2$
- Which in terms of matrixes, knowing that $(Xa)^t = a^t X^t$, we then have:

$C_1 = X \cdot a_1$, then square the equality and divide by *n* (no. of observations)

$$\frac{1}{n}(C_1)^t C_1 = \frac{1}{n}(a_1)^t X^t X a_1$$

The problem is to dually (complementary) reach the same goal:

1. Maximize the explained variance on axis 1;
2. Minimize the sum point distances to axis 1.

Principal Component Analysis (PCA)

Observation driven approach

$$\begin{cases} \underset{a^1}{Max} \dfrac{1}{n}(a_1)^t X^t X a_1 \\ subject\ of\ (a_1)^t a_1 = 1 \end{cases}$$

Lagrange function (or Lagrangean) associated to the problem is defined by:

$$L(a_1, \lambda) = \frac{1}{n}(a_1)^t X^t X a_1 - \lambda((a_1)^t a_1 - 1)$$

where λ is a Lagrange multiplier.

Principal Component Analysis (PCA)

Observation driven approach

Partial derivatives:

$$\frac{\partial L}{\partial a_1} = 2\frac{1}{n}X^t X a_1 - 2\lambda a_1 = 0 \qquad \frac{\partial L}{\partial \lambda} = (a_1)^t a_1 - 1 = 0$$

Having then $\frac{1}{n}X^t X a_1 = \lambda a_1$.

Therefore a_1 is an *eigenvector* of the matrix $\frac{1}{n}X^t X$, corresponding to the *eigenvalue* (*characteristic value*) λ.

Multiplying on the left with $(a_1)^t$ we have:

$$\frac{1}{n}(a_1)^t X^t X a_1 = \lambda$$

Principal Component Analysis (PCA)

Then

$\frac{1}{n}(a_1)^t X^t X a_1$ is the quantity we need to maximize.

Therefore:

- λ is the greatest characteristic value (eigenvalue), and a_1 is the corresponding characteristic vector (eigenvector);
- we shall assign λ to α_1.

Principal Component Analysis (PCA)

Step 2

- Determine axis 2 described by vector a_2 so axis 2 is orthogonal with axis 1;
- Maximize the explained variance (the points are more scattered, disperse on the axis);
- The applied optimization is:

$$\begin{cases} \underset{a_2}{Max} \frac{1}{n}(a_2)^t X^t X a_2 \\ (a_2)^t a_2 = 1 \\ (a_2)^t a_1 = 0 \end{cases}$$

$$L(a_2, \lambda_1, \lambda_2) = \frac{1}{n}(a_2)^t X^t X a_2 - \lambda_1((a_2)^t a_2 - 1) - \lambda_2 (a_2)^t a_1$$

Principal Component Analysis (PCA)

Step 2

Set the partial derivative on a_2 to zero:

$$\frac{\partial L}{\partial a_2} = 2\frac{1}{n} X^t X a_2 - 2\lambda_1 a_2 - \lambda_2 a_1 = 0$$

Multiplying on the left with $(a_1)^t$ we obtain:

$$2\frac{1}{n}(a_1)^t X^t X a_2 - 2\lambda_1 (a_1)^t a_2 - \lambda_2 (a_1)^t a_1 = 0$$

Principal Component Analysis (PCA)

Step 2

The 2 axis being orthogonal, we have: $(a_1)^t a_2 = 0$. Then, since:

$\frac{1}{n} X^t X a_1 = \alpha_1 a_1$ through transposition, it implies that

$$(a_1)^t \frac{1}{n} X^t X = \alpha_1 (a_1)^t$$

since the matrix $X^t X$ is symmetrical. Then, multiplying with 2 and a_2 on the right-hand side:

$$2\frac{1}{n}(a_1)^t X^t X a_2 = 2\frac{1}{n}\alpha_1 (a_1)^t a_2 = 0$$

Therefore, we must have $\lambda_2 = 0$.

Principal Component Analysis (PCA)

Step 2

Making the substitution in the derivative we obtain that

$$\frac{1}{n} X^t X a_2 = \lambda_1 a_2$$

and therefore a_2 is the eigenvector corresponding to the eigenvalue λ_1 , and this eigenvalue is maximal having given the equality:

$$\frac{1}{n} (a_2)^t X^t X a_2 = \lambda_1$$

Since $\frac{1}{n} X^t X a_2 = \lambda_1 a_2$ it is maximized at this step, we shall assign λ_1 to α_2

Principal Component Analysis (PCA)

Step *k*

- Determine k axis of a_k vector, orthogonal on the previous axis and to maximize the explained variance;
- The optimum problem is as follows:

$$\begin{cases} \underset{a^k}{Max} \dfrac{1}{n}(a_k)^t X^t X a_k \\ (a_k)^t a_k = 1 \\ (a_k)^t a_j = 0, j = \overline{1, k-1} \end{cases}$$

Principal Component Analysis (PCA)

Step *k*

The associated Lagrange function L(a_k, λ_1, λ_2,..., λ_k) is as follows:

$$L(a_k, \lambda_1, \lambda_2, \ldots, \lambda_k) = \frac{1}{n}(a_k)X^t X a_k - \lambda_1((a_k)^t a_k - 1) - \lambda_2 (a_k)^t a_1 - \ldots - \lambda_k (a_k)^t a_{k-1}$$

Setting the derivative on zero:

$$\frac{\partial L}{\partial a_k} = 2\frac{1}{n} X^t X a_k - 2\lambda_1 a_k - \lambda_2 a_1 - \ldots - \lambda_k a_{k-1} = 0$$

Then multiply the first relation successively with $(a_1)^t$, $(a_2)^t$,..., $(a_{k-1})^t$, and obtain $\lambda_2 = 0$, $\lambda_3 = 0$, ..., $\lambda_k = 0$. Returning with these results to the first partial derivative we have:

$$\frac{1}{n} X^t X a_k = \lambda_1 a_k$$

Principal Component Analysis (PCA)

Step *k*

Therefore a_k is the eigenvector of matrix $\frac{1}{n} X^t X$, corresponding to
The eigenvalue λ_1, and since the quantity

$$\frac{1}{n}(a_k)^t X^t X a_k$$

it is the one maximized at this step, then λ_1 is eigenvalue of *k* order.

We shall assign λ_1 to α_k.

Principal Component Analysis (PCA)

PCA in variable spaces

Phase 1

Determine the first principal component C_1 so, it is maximally correlated with initial, causal variables:

$$\sum_{j=1}^{m} R^2(C_1, X_j) \quad \text{to be maximum}$$

$$R^2(C_1, X_j) = \frac{Cov(C_1, X_j)^2}{Var(C_1)Var(X_j)} = \frac{1}{n}\frac{(C_1)^t X_j (X_j)^t C_1}{(C_1)^t C_1}$$

$$\sum_{j=1}^{m} R^2(C_1, X_j) = \frac{1}{n}\sum_{j=1}^{m}\frac{(C_1)^t X_j (X_j)^t C_1}{(C_1)^t C_1} = \frac{1}{n}\frac{(C_1)^t XX^t C_1}{(C_1)^t C_1}$$

Principal Component Analysis (PCA)

PCA in variable spaces

Phase 1

Solve the following problem:

$$\underset{C_1}{Maxim}\frac{1}{n}\frac{(C_1)^t XX^t C_1}{(C_1)^t C_1}$$

The solution is the eigenvector of matrix $\frac{1}{n}XX^t$, corresponding to the greatest eigenvalue β_1:

$$\frac{1}{n}XX^t \cdot C_1 = \beta_1 \cdot C_1$$

Principal Component Analysis (PCA)

PCA in variable spaces

Phase 2

Determine the second principal component C_2, maximally correlated with initial variables and not correlated at all with the first principal component C_1.

$$\begin{cases} \underset{C^2}{Maxim} \dfrac{1}{n} \dfrac{(C_2)^t XX^t C_2}{(C_2)^t C_2} \\ R(C_1, C_2) = 0 \end{cases}$$

The solution is the eigenvector of the matrix $\frac{1}{n} XX^t$, corresponding to the second greatest eigenvalue β_2:

$$\frac{1}{n} XX^t \cdot C_2 = \beta_2 \cdot C_2$$

Principal Component Analysis (PCA)

PCA in variable spaces

Phase *k*

Determine the principal component C_k, maximally correlated with initial variables and not correlated at all with the components previously determined, C_i, $i=1,k-1$.

$$\begin{cases} \underset{C^1}{Maxim} \dfrac{1}{n} \dfrac{(C_k)^t XX^t C_k}{(C_k)^t C_k} \\ R(C_k, C_i) = 0, i = \overline{1, k-1} \end{cases}$$

The solution is the eigenvector of the matrix $\frac{1}{n} XX^t$, corresponding to the *k* eigenvalue β_k:

$$\frac{1}{n} XX^t \cdot C_k = \beta_k \cdot C_k$$

Principal Component Analysis (PCA)

The link between the two approaches

In the observation spaces, at step k it is determined the eigenvector a_k, which is the unit vector of k axis, corresponding to C_k component:

$$\frac{1}{n} X^t X \cdot a_k = \alpha_k a_k$$

Multiplying this equation on the left with X we obtain:

$$\frac{1}{n} X X^t X a_k = X \alpha_k a_k \quad \Rightarrow \quad \frac{1}{n} X X^t C_k = \alpha_k C_k$$

Principal Component Analysis (PCA)

The link between the two approaches

It is the same equality obtained in the variable spaces approach, if considered $\beta_k = \alpha_k$

$$\frac{1}{n} XX^t C_k = \beta_k C_k$$

The maximum number of steps in the observation spaces may be $\boldsymbol{m}$ (the rank of matrix $\frac{1}{n} X^t X$), while in the variable spaces, the maximum number of steps may be $\boldsymbol{n}$ (the rank of matrix $\frac{1}{n} XX^t$).

The number of non-zero eigenvalues is ***min*(*m*, *n*).**