

# Software Development for Data Analysis

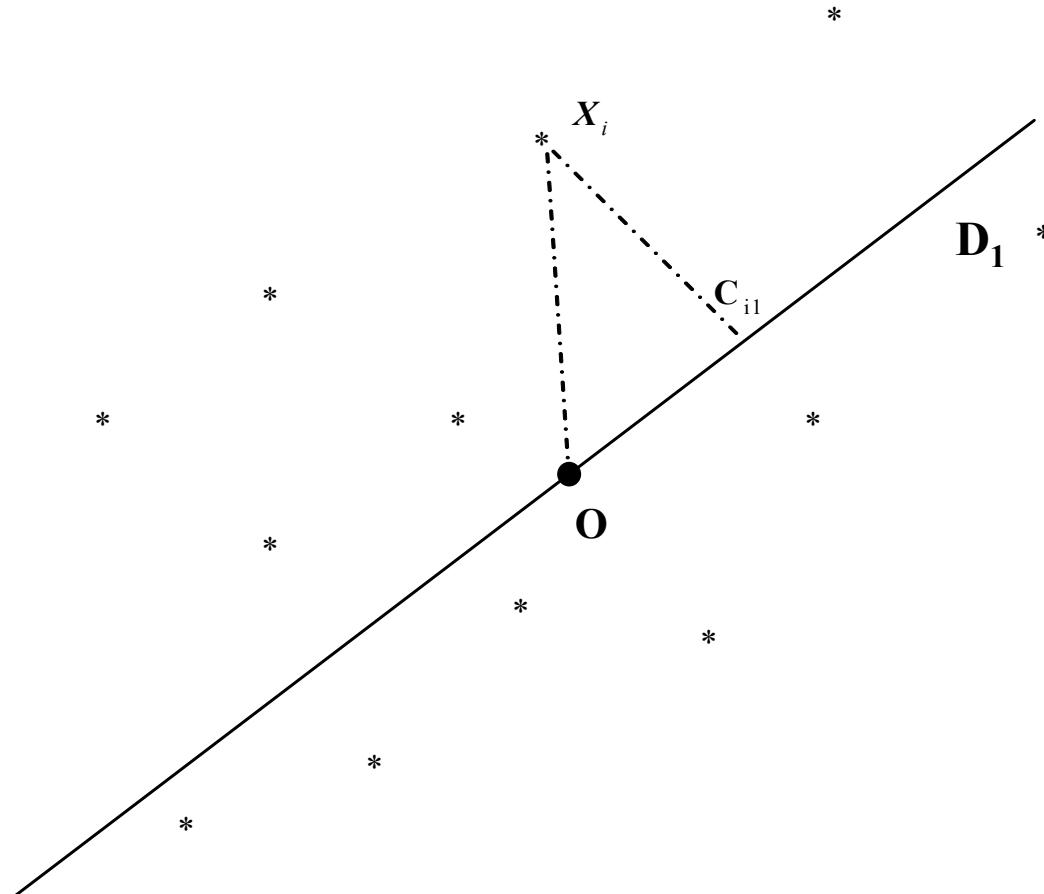
# Principal Component Analysis (PCA)

## Rationales, criteria upon choosing the axis numbers

- The main goal of PCA: to highlight the significant information regarding the overall data set.
- Hence, the first component agglutinates the most important information type because it contains the maximum variance.
- The question is: how many types of information deserve to be thoroughly, exhaustively investigated?
- Geometrically, it is all about determining the number of axis to be chosen for a multidimensional representation in order to obtain a satisfactory informational coverage.

# Principal Component Analysis (PCA)

Observation driven approach: projection on  $D_1$  axis



# Principal Component Analysis (PCA)

## Criteria upon choosing the number of axis

### 1. Coverage percentage criterion:

- Determining the variance quantity explained on each axis.
- Since the optimum criteria in choosing axis  $k$  is to maximize the variance on that axis, then:

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

- Therefore, the explained variance on axis  $k$  is the eigenvalue  $\alpha_k$ .
- Table  $X$  being standardized, the overall variance is  $m$ , the number of variables.
- Consequently, the explained variance percentage on  $k$  axis is  $\alpha_k/m$ .

# Principal Component Analysis (PCA)

## Criteria upon choosing the number of axis

- Hence, the percentage of variance explained by the  $k$  axis is:

$$\frac{\sum_{j=1}^k \alpha_j}{\sum_{i=1}^m \alpha_i}$$

- If the variables  $X$  are standardized then:  $\frac{\sum_{j=1}^{k-1} \alpha_j}{m}$
- Similarly, approaching the problem from the variable spaces, at the  $k$  step (phase), the coefficient of determination (square of Pearson correlation coefficient) between the new  $C_k$  component and the initial, causal variables is:

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

# Principal Component Analysis (PCA)

## Criteria upon choosing the number of axis

- The eigenvalue  $\alpha_k$  (or the characteristic value) is the sum between the determined coefficients of determination of the new component and the previously determined component coefficients.

$$\sum_{j=1}^m R^2(C_k, X_j) = \frac{1}{n} \frac{(C_k)^t X X^t C_k}{(C_k)^t C_k} = \frac{(C_k)^t \alpha_k C_k}{(C_k)^t C_k} = \alpha_k.$$

- If  $s$  is the number of significant axis, then, according to the coverage percentage criteria,  $s$  is the first value for which  $\alpha_s > P$ , where  $P$  is the chosen coverage percentage.

# Principal Component Analysis (PCA)

## Criteria upon choosing the number of axis

### 1. Variance explained criteria:

- Some researchers simply use the rule of keeping enough factors to account for 90% (sometimes 80%) of the variation.
- If research goal is to emphasizes *parsimony* (explaining variance with as few factors as possible), then the percentage for the coverage criterion could be as low as 50%.

# Principal Component Analysis (PCA)

## Criteria upon choosing the number of axis

### 2. Kaiser criterion:

- The criterion is applicable only if the causal variables  $X_j, j = 1, m$  are standardized (mean is 0 and variance is 1).
- In such a case it makes sense that the new variables, the principal components, to be considered important, significant, if they agglutinate more variance than an initial variable  $X_j$ , which may have the maximum variance equal to 1.
- The Kaiser rule recommends to keep those principal components which have a variance (eigenvalue) greater than 1.

# Principal Component Analysis (PCA)

## Criteria upon choosing the number of axis

### 3. Cattell criterion:

- The criterion may be applied in both graphical and analytical approaches.
- Graphically, beginning with the third principal component, is to detect the first turn, an angle of less than  $180^\circ$ .
- Only the eigenvalues up to that point, inclusive, are to be retained.
- In the analytical approach, there are to be computed the second order differences between the eigenvalues, starting with the 3<sup>rd</sup> eigenvalue:

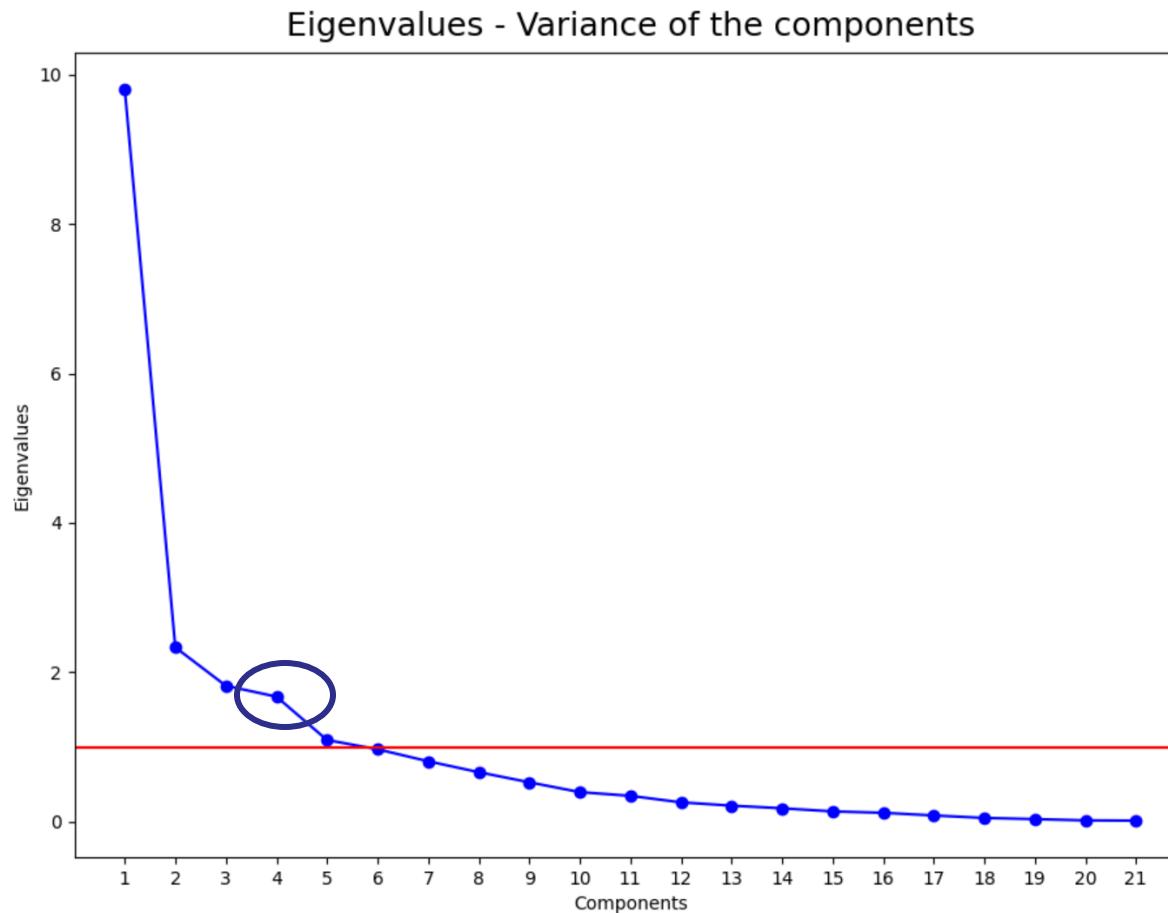
$$\varepsilon_k = \alpha_k - \alpha_{k+1}, k = 3, m-1$$

$$\delta_k = \varepsilon_k - \varepsilon_{k+1}, k = 3, m-2$$

- The value for  $s$  is determined such as deltas  $\delta_1, \delta_2, \dots, \delta_{s-1}$  to be greater or equal to 0 (zero).
- The following axis are retained:  $a_1, a_2, \dots, a_s$ .

# Principal Component Analysis (PCA)

**Criteria upon choosing the number of axis**



# Principal Component Analysis (PCA)

## Scores

- Are standardized values of the principal components:

$$C_{ik}^s = \frac{C_{ik}}{\sqrt{\alpha_k}}, \quad i = \overline{1, n}, k = \overline{1, s}$$

- Where  $\sqrt{\alpha_k}$  is the standard deviation of component  $C_k$ .

# Principal Component Analysis (PCA)

## The quality of points representation

- The principal components represents a new space of the observations – *the principal space*, as oppose to the initially observed, causal, space.
- The basis for this new space, the unit vector of its axes, is constituted by the eigenvectors  $a_k$ ,  $k = 1, m$ .
- The coordinates of the observations within these new axes are given by the vectors  $C_k$ ,  $k = 1, m$ .
- As we mentioned earlier, an observation is geometrically represented by a point in a *m-dimensional* space.
- The sum of square distances, from the projection of point  $i$  on axis  $a_k$ ,  $k = 1, m$  to the barycenter of the data cloud is given by:

$$\sum_{k=1}^m c_{ik}^2$$

# Principal Component Analysis (PCA)

## The quality of points representation

- An observation is better represented on a given axis  $a_j$  as  $c_{ij}^2$  has a greater value in relation to  $\sum_{k=1}^m c_{ik}^2$
- The quality of representing the  $i$  observation on  $a_j$  axis, is determined by the ratio: 
$$\frac{c_{ij}^2}{\sum_{k=1}^m c_{ik}^2}$$
- The value of the ratio is equal with square cosine of the angle between the vector associated to point  $i$  and  $a_j$  axis.

# Principal Component Analysis (PCA)

## The observation contributions to axis variances

- The explained variance on  $a_j$  axis is:  $\frac{1}{n} \sum_{i=1}^n c_{ij}^2 = \alpha_j$
- The contribution of  $i$  observation to this variance is:  $\frac{c_{ij}^2}{n}$
- Therefore, the contribution of  $i$  observation to the variance of  $a_j$  axis is:

$$\frac{c_{ij}^2}{n \cdot \alpha_j}$$

# Principal Component Analysis (PCA)

## Correlation coefficients between principal components and the initial variables

- The degree (coefficient) of determination between a causal variable  $X_j$  and the principal component  $C_r$  are computed as squared Pearson correlation coefficient:

$$R^2(C_r, X_j) = \frac{\text{Cov}(C_r, X_j)^2}{\text{Var}(C_r)\text{Var}(X_j)} = \frac{\text{Cov}(C_r, X_j)^2}{\alpha_r}$$

since  $\text{Var}(C_r) = \alpha_r$ , and  $\text{Var}(X_j) = 1$ , being standardized unit vector.

# Principal Component Analysis (PCA)

**Correlation coefficients between principal components and the initial variables**

- In terms of matrices, the correlation coefficients vector between the initial (causal) variables and the principal component  $\mathbf{C}_r$  is given by:

$$R_r = \frac{\frac{1}{n} X^t C_r}{\sqrt{\alpha_r}} = \frac{\frac{1}{n} X^t X a_r}{\sqrt{\alpha_r}} = \frac{\alpha_r a_r}{\sqrt{\alpha_r}} = a_r \sqrt{\alpha_r}$$

These correlations are labeled as *factor loadings*.

# Principal Component Analysis (PCA)

## Commonalities in PCA

- The commonality of an initial variable  $X_j$  in relation to **the first  $s$  principal components** is the sum of correlation coefficients between the causal variable and the principal components.
- Represent the proportion of each observed variable's variance that can be explained by the principal components (e.g., the underlying latent continua).
- It can be defined as the sum of squared factor loadings:

$$h^2 = \sum_{k=1}^s R(X_j, C_k)^2$$

# Principal Component Analysis (PCA)

## Commonalities in PCA

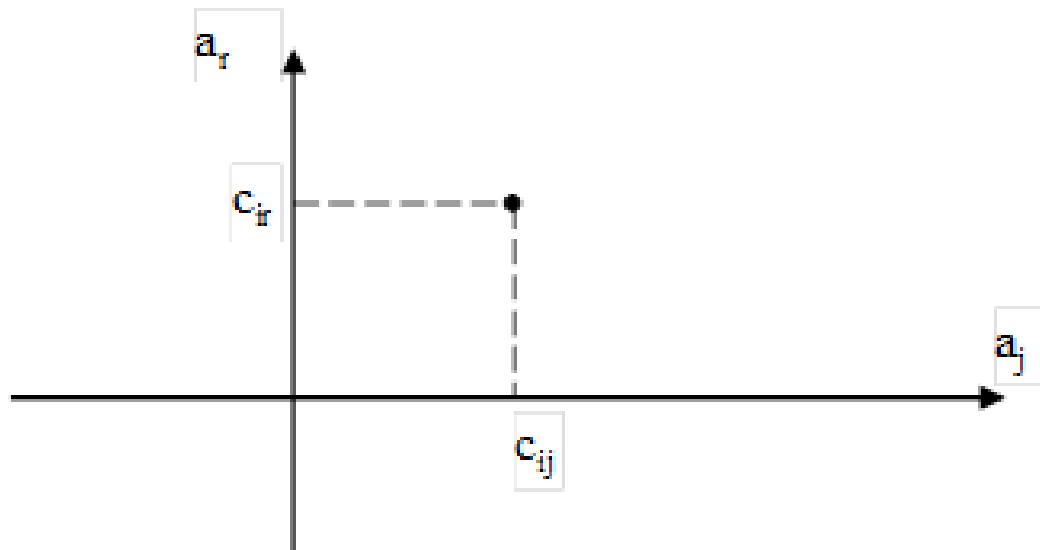
- Principal component  $C_k$  contains a variance quantity given by  $\alpha_k$ , and the sum of the correlation coefficients between this component and the causal variables is equal to  $\alpha_k$  as well.
- For  $s = m$ ,  $h^2 = \sum_{k=1}^s R(X_j, C_k)^2$

becomes equal to 1, meaning that those  $m$  principal components explain entirely the information from the initial data table  $X$ .

# Principal Component Analysis (PCA)

## Observation graphical representations

- In order to analyze the results obtained at 2 phases  $j$  and  $r$ , any given observation  $i$  can be represented by projecting it on the plane created by  $a_j$  and  $a_r$  axes.

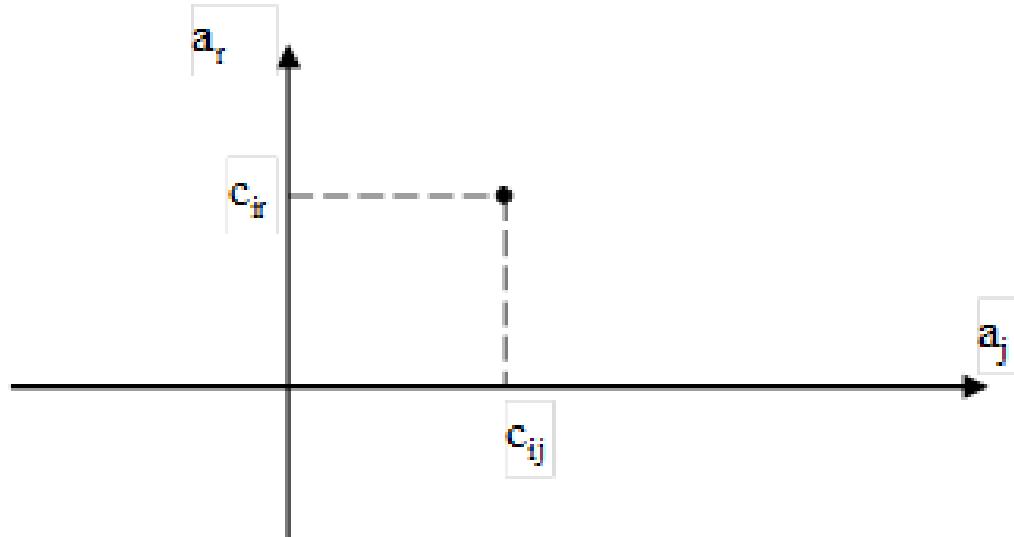


# Principal Component Analysis (PCA)

## Observation graphical representations

- Then the entire cloud of observation points can be represented by projecting it on the plane created by the  $a_j$  and  $a_r$  vectors.
- The coordinates of given observation (point in the cloud) are:

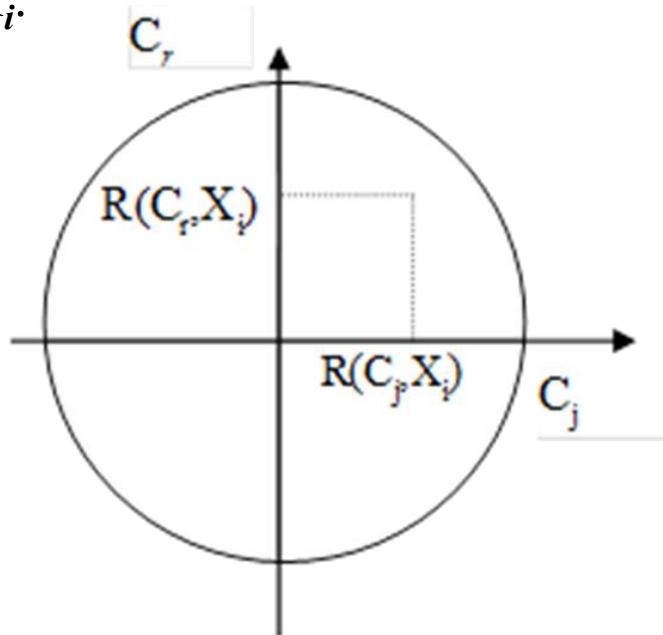
$c_{ij}$  and  $c_{ir}$



# Principal Component Analysis (PCA)

## Variable graphical representations

- Accomplished by using the correlation circle between the initial, causal variables and the principal components.
- Those 2 axis correspond to the chosen principal components  $C_j$  and  $C_r$  in relation to a given causal variable  $X_i$ .



# Principal Component Analysis (PCA)

## Non-standard PCA

- Having given the hypothesis that, the initial variables are only centered, but not normalized.
- The initial variables' variance is no longer 1.

- The analysis is conducted on covariance matrix, since:

$$\frac{1}{n} X^t X$$

is the covariance matrix of the observation tables.

- In the observation space, the optimum criterion at any given phase remains the same, but it applies to a different cloud of points.
- The vectors  $a_k, k = \overline{1, m}$ , are eigenvectors of the covariance matrix.

# Principal Component Analysis (PCA)

## Non-standard PCA

- In the variable spaces, the optimum criterion at a given phase  $k$ ,

$$\underset{C_k}{\text{Maxim}} \sum_{j=1}^m R^2(C_k, X_j)$$

becomes:

$$\underset{C_k}{\text{Maxim}} \sum_{j=1}^m \text{Cov}^2(C_k, X_j)$$

# Principal Component Analysis (PCA)

## Weighted PCA

- The assumption is that the weight of each observation is different than  $\frac{1}{n}$ .
- Let's  $p_i$ , be the weights associated to the  $i$  observation,  $0 < p_i < 1$ ,

$$\sum_{i=1}^n p_i = 1$$

Then there can be defined the square matrix  $P$  of weights as being:

$$P = \begin{bmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & p_n \end{bmatrix}$$

# Principal Component Analysis (PCA)

## Weighted PCA

- The optimum criterion in the observation spaces becomes a sum of explained variance on each axis  $a_k, k = \overline{1, m}$  multiplied with the corresponding weight associated to each observation

$$Maxim \sum_{i=1}^n p_i c_{ik}^2$$

# Principal Component Analysis (PCA)

## Weighted PCA

- The optimum criterion remains unchanged in the variable spaces.
- The correlation between 2 variables is computed taking into account the observation weights.
- The covariance between 2 centered variable  $X$  and  $Y$  is:

$$Cov(X, Y) = \sum_{i=1}^n p_i x_i y_i$$

And the variance of  $X$  is:  $Var(X) = \sum_{i=1}^n p_i x_i^2$

- The vectors  $a_k$  are computed as successive eigenvectors of matrix  $X^t P \cdot X$
- And the principal components as successive eigenvectors of matrix  $X \cdot X^t P$