Data Analysis

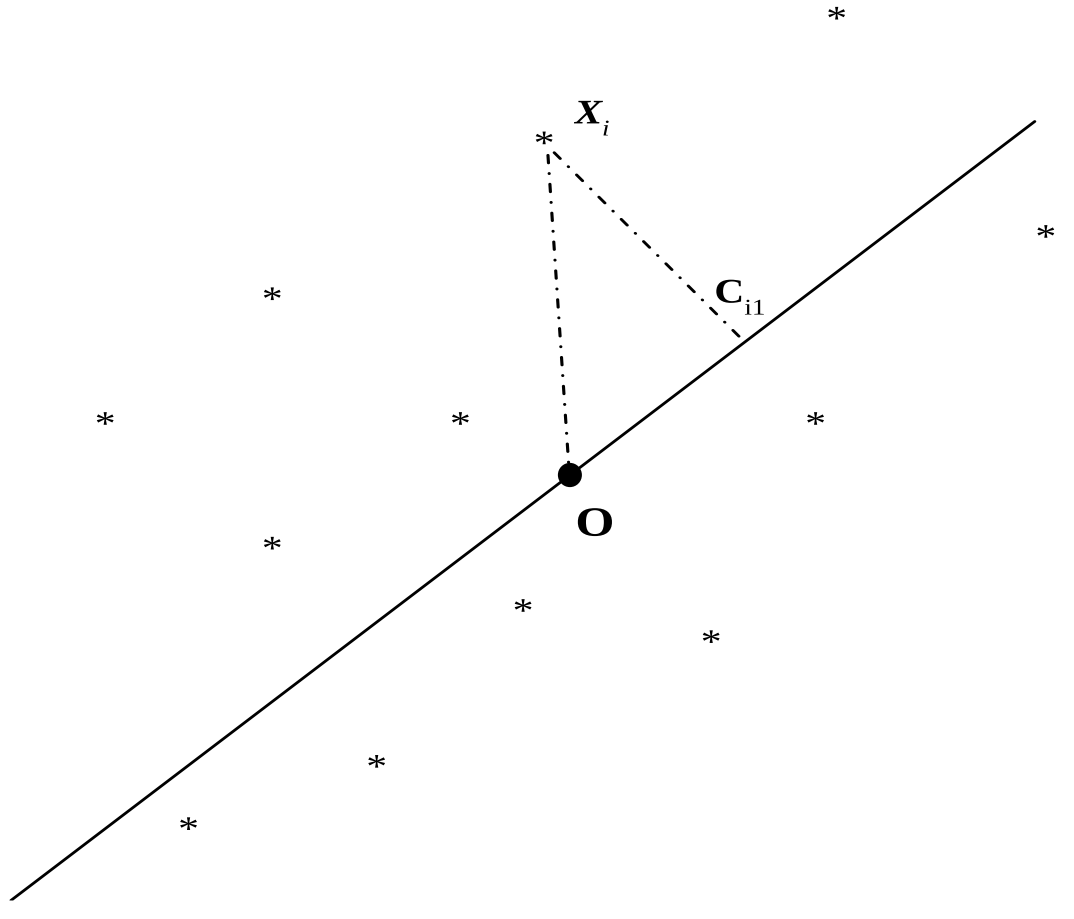
Principal Component Analysis (PCA)

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 axis numbers

1. <u>**Coverage percentage criterion**</u>:

- 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 α_k.
- Table *X* being standardized, the overall variance is ***m***.
- Consequently, the explained variance percentage on ***k*** axis is este α_k / m.

Principal Component Analysis (PCA)

Criteria upon choosing the axis numbers

- Hence, the variance percentage 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:

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

Principal Component Analysis (PCA)

Criteria upon choosing the axis numbers

- Similarly, approaching the problem from the variable spaces, at the ***k*** step (phase), the correlation between the new $\boldsymbol{C_k}$ component and the initial, causal variables is:

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

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

- The eigenvalue (characteristic value) α_k is the sum between the determined coefficients of the new component and the previously determined component coefficients.

- 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 axis numbers

1. **<u>Kaiser criterion</u>**:

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

Principal Component Analysis (PCA)

Criteria upon choosing the axis numbers

2. <u>Cattell criterion</u>:

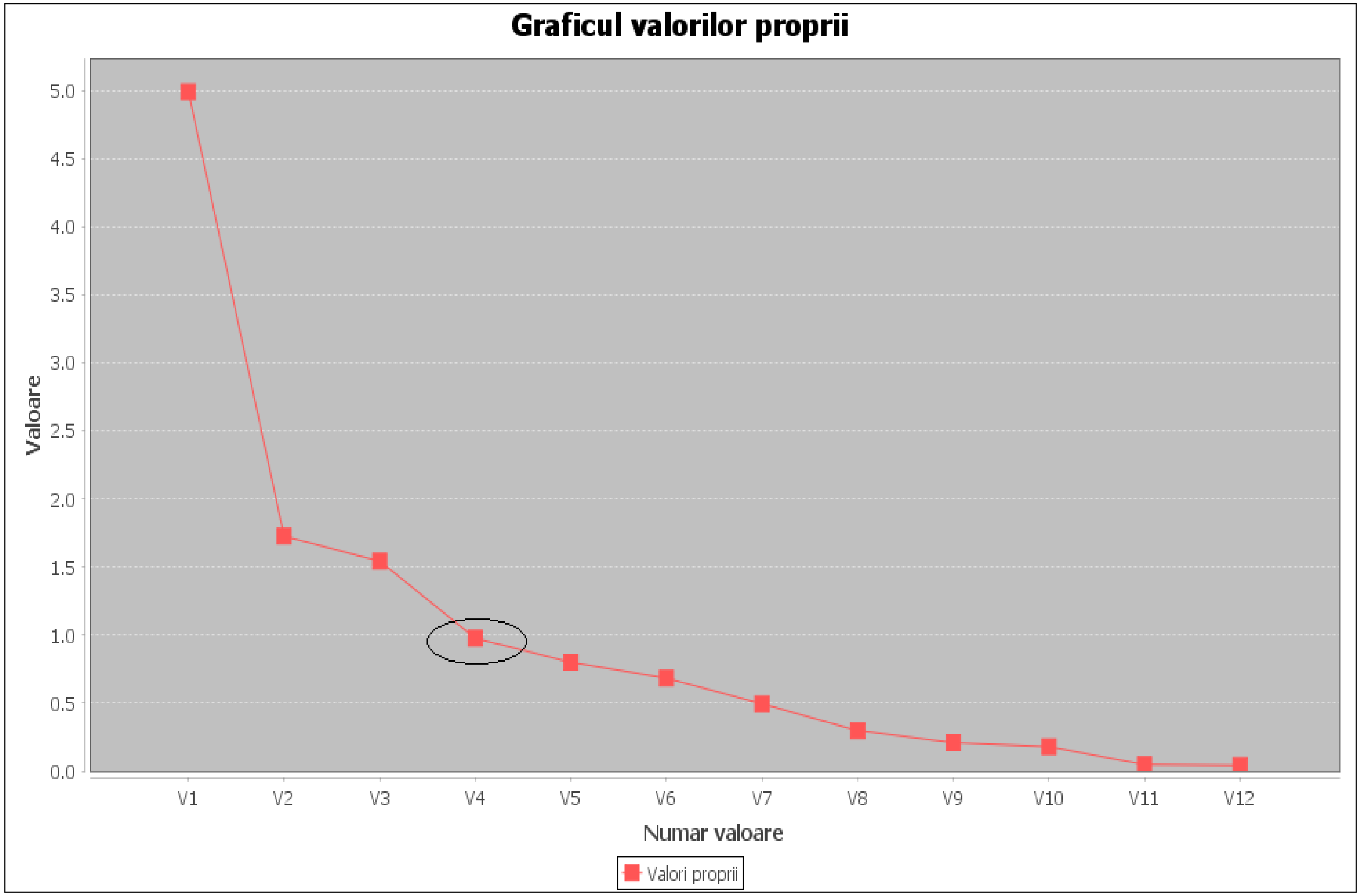
- The criterion may be applied in both graphical and analytical approaches.
- Graphically, beginning with the third principal component, is to detect the first turn, an angle of less than 180°.
- Only the eigenvalues up to that point, inclusive, are to be retained.
- In the analytical approach, there are to be computed the second order differences between the eigenvalues

  $\varepsilon_{\mathrm{k}} = \alpha_{\mathrm{k}} - \alpha_{k+1}$, $k = 1, m\text{-}1$

  $\delta_{\mathrm{k}} = \varepsilon_{\mathrm{k}} - \varepsilon_{\mathrm{k+1}}$, $k=1, m\text{-}2$
- The value for ***s*** is determined such as $\delta_1, \delta_2, \ldots, \delta_{s\text{-}1}$ to be greater or equal to 0 (zero).
- The following axis are retained: $a_1, a_2, ..., a_{s+2}$.

Principal Component Analysis (PCA)

Criteria upon choosing the axis numbers



Principal Component Analysis (PCA)

Criteria upon choosing the axis numbers

3. Variance explained criteria:

- Some researchers simply use the rule of keeping enough factors to account for 90% (sometimes 80%) of the variation.
- Where the researcher's goal emphasizes parsimony (explaining variance with as few factors as possible), the criterion could be as low as 50%.

Principal Component Analysis (PCA)

Criteria upon choosing the axis numbers

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 point representations

- The principals components represents a new space of the observations – *the principal space*.
- The basis for this new space, the unit vector of its axis, it is constituted by the eigenvectors a_k, $k = 1,m$.
- The coordinates of the observations within these new axis 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 square distances from an index point ***i*** to the barycenter of the data cloud is:

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

Principal Component Analysis (PCA)

The quality of point representations

- An observation is better represented on a given axis $\boldsymbol{a_j}$ 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: $\dfrac{c_{ij}^2}{\sum_{k=1}^{m} c_{ik}^2}$
- The value of the ratio is equal with square cosine of the angle between the vector associated to point ***i*** and $\boldsymbol{a_j}$ vector.

Principal Component Analysis (PCA)

The observation contributions to axis variances

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

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

Principal Component Analysis (PCA)

Communalities in PCA

- The communality of an initial variable $\boldsymbol{X_j}$ in relation to the first $\boldsymbol{s}$ principal components is the sum of correlation coefficients between the causal variable and the principal components.
- Represent the proportion of each 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:

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

Principal Component Analysis (PCA)

Communalities in PCA

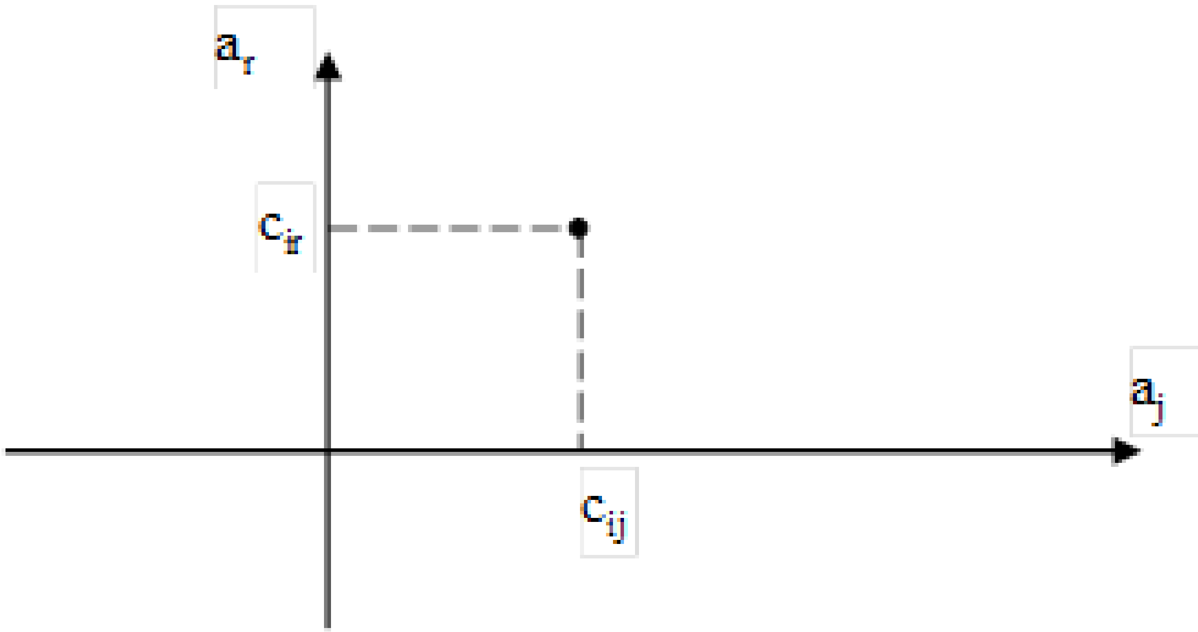
- Principal component C_k contains a variance quantity given by α_k, and the sum of the correlation coefficients between this component and the causal variables is equal to α_k as well.
- For $s = m$, $h^2 = \sum_{k=1}^{s} R\left(X_j, C_k\right)^2$

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 $\boldsymbol{j}$ and $\boldsymbol{r}$, any given observation $\boldsymbol{i}$ can be represented by projecting it on a plane created by $\boldsymbol{a_j}$ and $\boldsymbol{a_r}$ axis.

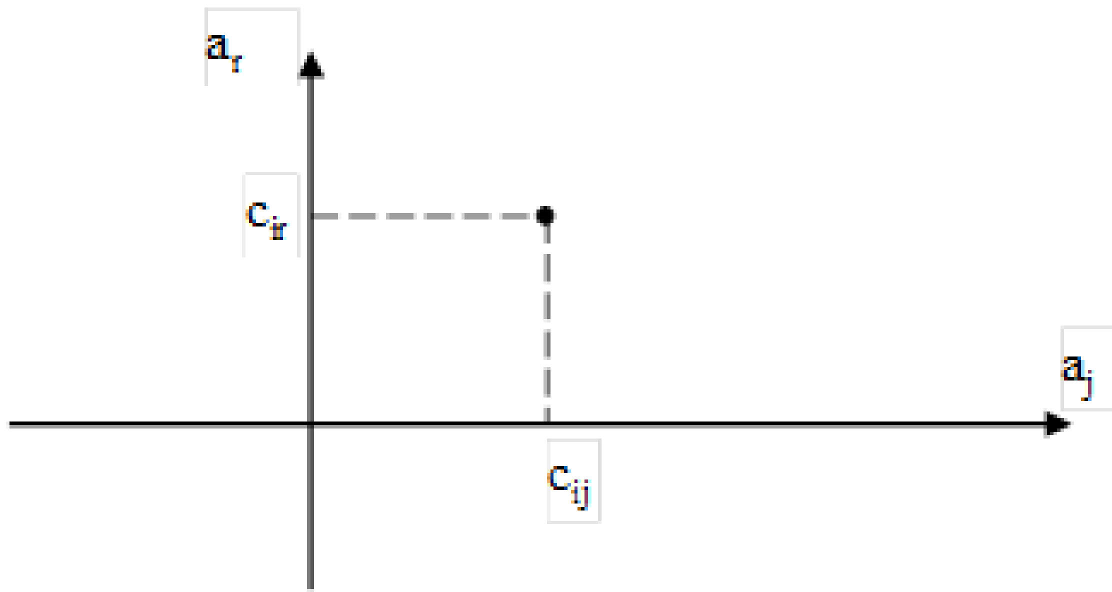


Principal Component Analysis (PCA)

Observation graphical representations

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

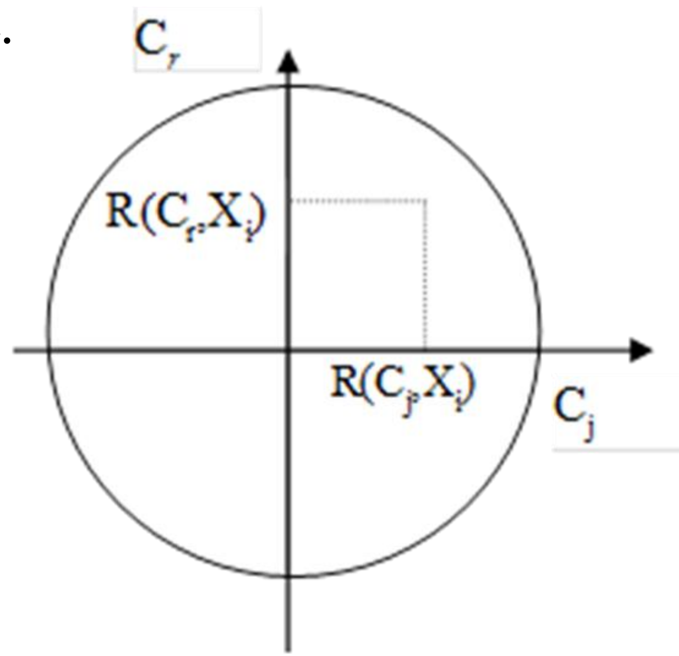
c_{ij} and c_{ir}



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 $\boldsymbol{C_j}$ and in relation to a given causal variable $X_{\boldsymbol{i}}$.



Principal Component Analysis (PCA)

Correlation coefficients

- The degree of determination between a causal variable $\boldsymbol{X_j}$ and the principal component $\boldsymbol{C_r}$ are computed as Pearson correlation coefficient:

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

since Var(C_r)=α_r, and Var(X_j) =1, being standardized unit vector.

Principal Component Analysis (PCA)

Correlation coefficients

- In terms of matrixes, the correlation coefficients vector between the initial (causal) variables and the principal component $\mathbf{C}_{\boldsymbol{r}}$ 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)

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*,

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

becomes:

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

Principal Component Analysis (PCA)

Weighted PCA

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

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

Then there can be defined the weight matrixes P 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:

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

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

$$X \cdot X^t P$$