Statistical tools

DATA SCIENCE

Nicolas Sutton-Charani (Euromov - DHM)

Covariance

Let X and Y be 2 random variables such that $(X, Y) \in \mathbb{R}^2$, the covariance is defined as follows :

$$\begin{aligned} Cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \in \mathbb{R} \end{aligned}$$

Estimation from data $(x_i, y_i)_{i=1,\ldots,n}$

$$\begin{aligned} Cov(X, Y) &\approx \overline{xy} - \overline{x}.\overline{y} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \end{aligned}$$

Correlation

Let X and Y be 2 random variables such that $(X, Y) \in \mathbb{R}^2$, the **Pearson** correlation index ρ_{XY} is defined as follows :

$$\rho_{XY} = Corr(X, Y) = \frac{Cov(X, Y)}{\rho_X \rho_Y} \in [-1, +1]$$

Estimation from data $(x_i, y_i)_{i=1,\ldots,n}$

$$\rho_{XY} \approx \frac{\bar{xy} - \bar{x}\bar{y}}{\sqrt{\overline{(x - \bar{x})^2}}.\sqrt{\overline{(y - \bar{y})^2}}}$$

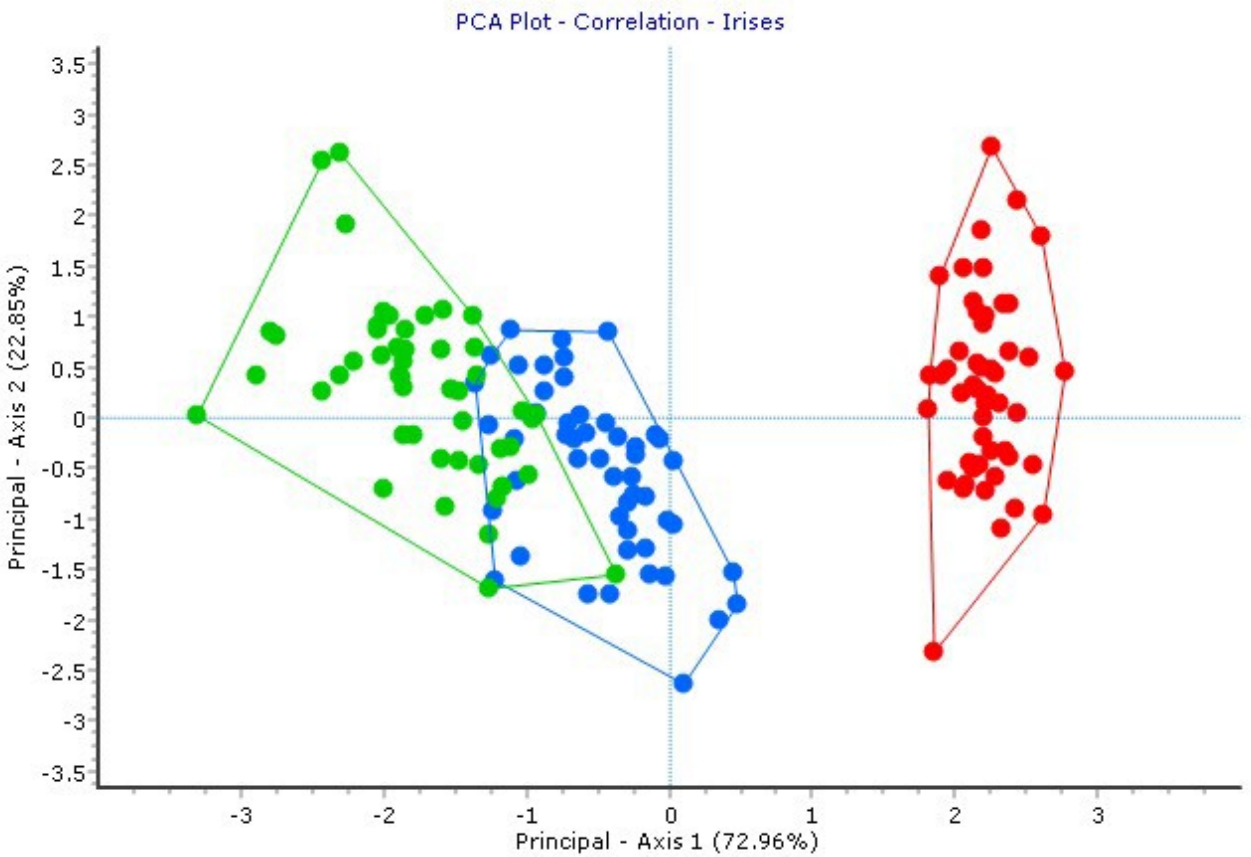
Spearman correlation = Pearson correlation computed on x, y values **ranks** (not on x, y values)
$\rightarrow$ extension from *linear* to *monotonic* correlations

PCA

Common data analysis tool

- **matrix diagonalisation** $\rightarrow$ **dimension reduction** from j to p
- for $p = 2 \rightarrow$ 2D- or 3D-graphical representations
  - $\rightarrow$ clustering relevance (examples plot)
  - $\rightarrow$ correlations insights (variables plot)
- PCA limits curse of the dimensionnality and increase models predictive power
- limited to linear transformation
- many extensions

PCA examples plot



PCA variables plot