SPLEX TME 8

Partial Least Squares and Canonical Correlation Analysis

The goal of the TME is to understand and get skills in Partial Least Squares (PLS) and Canonical Correlation Analysis (PCA).

Data

- Data provided during the TME

Analysis

- **Canonical correlation**.

  Short description: Canonical correlations analysis (CCA) is an exploratory statistical method to highlight correlations between two data sets acquired on the same experimental units. CCA is most appropriate when a researcher desires to examine the relationship between two variable set.

  X and Y are matrices of order $n \times p$ and $n \times q$. The columns correspond to variables and the rows correspond to experimental units (patients). Find two vectors a and b that maximize the correlation between the linear combinations

  $$U = a_1 X^1 + a_2 X^2 + \cdots + a_p X^p$$
  $$V = b_1 Y^1 + b_2 Y^2 + \cdots + b_q Y^q$$

  The problem consists in solving

  $$\rho = cor(U, V) = \max_{a,b} cor(Xa, Yb)$$

  Canonical correlations ρ are the positive square roots of the eigenvalues λ of $P_X P_Y$ ($\rho = \sqrt{\lambda}$), where

  $$P_X = X(X^T X)^{-1} X^T$$
  $$P_Y = Y(Y^T Y)^{-1} Y^T$$

  The canonical correlation coefficient is the Pearson relationship between the two synthetic variables on a given canonical function. Because of the scaling created by the standardized weights in the linear equations, this value cannot be negative and only ranges from 0 to 1. Visualization of the results of canonical correlation is usually through bar plots of the coefficients of the two sets of variables for the pairs of canonical variates showing significant correlation.

- **Partial least squares**. *Short description:* PLS regression is a recent technique that generalizes and combines features from principal component analysis and multiple regression. It is particularly useful when we need to predict a set of dependent variables from a (very) large set of independent variables (i.e., predictors). SPLS is its sparse version.

1. Explore the canonical correlation analysis of Python: find relations between groups of variables

   `http://scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.CCA.html`

2. `http://scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.PLSCanonical.html`

3. Estimate accuracy of the methods on the heterogeneous data