Software Development for Data Analysis

Canonical Correlation Analysis

The explained variance and informational redundancy

- The quantity of variance explained by each pair of canonical variables, in connection with each of the initial data set, is given by the sum of correlations between the canonical variables and the causal variables of the sets:

$$VX_k = \sum_{j=1}^{p} R(z_k, X_j)^2, \text{ k} = 1{,}m$$

$$VY_k = \sum_{j=1}^{q} R(u_k, Y_j)^2, \text{ k} = 1{,}m \ ,$$

- where $R(z_k, X_j)^2$ is the determination (correlation) coefficient between the canonical variable z_k of the pair k, and the variable X_j, belonging to the first data set (the j column of matrix X),

Canonical Correlation Analysis

The explained variance and informational redundancy

- and $R(u_k, Y_j)^2$ is the correlation coefficient between the canonical variable u_k of the pair k, and the causal variable Y_j, belonging to the second data set (the j column of matrix Y)
- Proportionally, the values are: $\frac{VX_k}{p}$ and $\frac{VY_k}{q}$, p and q being the number of causal variables, columns, of matrices X, and Y, respectively.
- The overall variance explained by all m canonical roots is:
  $VX = \sum_{k=1}^{m} VX_k$, for the first data set X, and
  $VY = \sum_{k=1}^{m} VY_k$, for the second data set Y.

Canonical Correlation Analysis

The explained variance and informational redundancy

- The redundancy is given by the common information existent in both data sets, and extracted by the canonical roots (pairs).
- The common information is given by the canonical correlation.
- If there is a certain quantity of information extracted by a canonical variable from one of the sets, then the part of this information found in the other set it is retrieved by using the canonical correlation, as follows:

  $SX_k = VX_k \cdot \alpha_k \ , \ \ k = 1, m$

  $SY_k = VY_k \cdot \alpha_k \ , \ \ k = 1, m$

  where the eigenvalue α_k is the correlation coefficient between the canonical variables z_k and u_k .

Canonical Correlation Analysis

The explained variance and informational redundancy

- The redundancy of all *m* canonical roots is:

$$SX = \sum_{k=1}^{m} SX_k ,$$

$$SY = \sum_{k=1}^{m} SY_k .$$

Canonical Correlation Analysis

Standardizing canonical factors

- Canonical factors are easier to interpret if standardized. Standardizing canonical factors implies to relate them to the standard deviation of initial and canonical variables.

$$as_{ik} = a_{ik} \cdot \frac{\sigma_{X_i}}{\sigma_{z_k}} \quad , \; i = 1, p \; ; k = 1, m$$

$$bs_{ik} = b_{ik} \cdot \frac{\sigma_{Y_i}}{\sigma_{u_k}} \quad , \; i = 1, q \; ; k = 1, m$$

- The interpretation of standardized canonical factors is similar to multiple regression: the increase with one unit of variables X_i or Y_i standard deviation, generates an increase with as_{ik} or bs_{ik} of canonical variables z_k or u_k standard deviation.

Canonical Correlation Analysis

Canonical roots - Bartlett χ^2 relevance test

- Bartlett χ^2 is the most employed test to evaluate canonical correlations.
- For any given canonical root, the result of the test indicates if there is any dependency between the two sets of variables or, on the contrary, the two sets of variables are independent.
- H0 hypothesis: correlation coefficient $R(z_k, u_k)$ indicates the existence of a linear correlation between the two sets of initial (causal) variables.
- H1, the alternative hypothesis: correlation coefficient $R(z_k, u_k)$ indicates a lack of connection.

Canonical Correlation Analysis

Canonical roots - Bartlett χ^2 relevance test

- For a canonical root (z_k, u_k), the test is applied as follows:

1. The number of degrees of freedom is computed, associated to each canonical root of rank k:

$$df_k = (p-k+1)(q-k+1)$$

where p and q are the number of initial variables of the first and second sets, respectively.

2. The statistics of the test is computed as follows:

$$\chi_k^2 = \left(-n+1+\frac{p+q+1}{2}\right)\log\left(1-\lambda_k\right)$$

where n is the number of observations, and λ_k is an indicator named lambda Wilks.

Canonical Correlation Analysis

Canonical roots - Bartlett χ^2 relevance test

Lambda Wilks indicator is computed in the following manner:

$\lambda_k = \prod_{i=k}^{m} \left(1 - R(z_i, u_i)^2\right)$, where m is the number of canonical roots.

3. Using χ^2 distribution, it is then determined the critical value for the test:

$\chi c_k^2 (1-\alpha, df_k)$, for a significance threshold α.

4. Then the test is applied:

If $\chi_k^2 \geq \chi c_k^2 (1-\alpha, df_k)$,

the H0 hypothesis is accepted, with a level of confidence 1 - α, otherwise it is rejected.

Canonical Correlation Analysis

Generalized Canonical Analysis (gCCA)

Having given q data sets X_1, X_2, ..., X_q which describe the same n observations, m_i the number of columns of matrix X_i and W_i the subspace in $\mathbf{R}^n$ generated by the columns, we make the following notations:

- P_i is the orthogonal projection X_i on subspace W_i.
- The total number of causal variables is $m = \sum_{i=1}^{q} m_i$.
- Make the assumption that $n > m$.

Canonical Correlation Analysis

Generalized Canonical Analysis (gCCA)

Generalized canonical analysis is to determine in the first phase an auxiliary variable Z_1, as a linear combination of causal variables and q canonical variables z_{i1} ($i = 1, q$), such that:

- $\sum_{i=1}^{q} R^2(Z_1, z_{i1})$ to be maximal, under the restriction of having:
- To ensure their unicity, the restriction of normality has to be imposed: $(Z_1)^t Z_1 = 1$

Canonical Correlation Analysis

Generalized Canonical Analysis (gCCA)

- In order to have the sum of the correlation maximal, the vectors z_{i1} are selected such that to be orthogonal projection of Z_1 vector on subspaces W_i: $z_{i1} = P_i \cdot Z_1$.
- Returning to the correlation sums, it can be rewritten as follows:

$$\sum_{i=1}^{q} R^2(Z_1, z_{i1}) = \sum_{i=1}^{q} \frac{Cov\left(Z_1, z_{i1}\right)^2}{Var\left(Z_1\right)Var\left(z_{i1}\right)} =$$

$$= \sum_{i=1}^{q} \frac{\left(\frac{1}{n}\left(Z_1\right)^t z_{i1}\right)^2}{\frac{1}{n}\left(Z_1\right)^t Z_1 \ \frac{1}{n}\left(z_{i1}\right)^t z_{i1}} = \sum_{i=1}^{q} \frac{\left(\frac{1}{n}\left(Z_1\right)^t z_{i1}\right)^2}{\frac{1}{n^2}\left(z_{i1}\right)^t z_{i1}} = \sum_{i=1}^{q} \frac{\left(\left(Z_1\right)^t z_{i1}\right)^2}{\left(z_{i1}\right)^t z_{i1}}$$

Canonical Correlation Analysis

Generalized Canonical Analysis (gCCA)

- Replacing z_{i1} with $P_i \cdot Z_1$, we obtain:

$$\sum_{i=1}^{q} R^2(Z_1, z_{i1}) = \sum_{i=1}^{q} \frac{\left(\left(Z_1\right)^t P_i Z_1\right)^2}{\left(Z_1\right)^t \left(P_i\right)^t P_i Z_1} = \sum_{i=1}^{q} \left(Z_1\right)^t P_i Z_1 = \left(Z_1\right)^t \left(\sum_{i=1}^{q} P_i\right) Z_1 ,$$

because $\left(P_i\right)^t P_i = P_i^2 = P_i$.

- Therefore, the optimum problems becomes:

$$\begin{cases} \underset{Z_1}{Maxim} \left(Z_1\right)^t \left(\sum_{i=1}^{q} P_i\right) Z_1 \\ \left(Z_1\right)^t Z_1 = 1 \end{cases}$$

Canonical Correlation Analysis

Generalized Canonical Analysis (gCCA)

- The solution of this problem (see PCA) implies that the variable Z_1,

  is the eigenvector of the matrix $\sum_{i=1}^{q} P_i$

  corresponding to the greatest eigenvalue, while the canonical variables of the sets, $z_{\mathrm{i}1}$, are determined using the relation:

  $z_{\mathrm{i}1} = P_i \cdot Z_1$.

- At the k phase there is to be determined the auxiliary variable Z_k and the canonical variables z_{ik} (i = 1, q) such that:

  $$\sum_{i=1}^{q} R^2(Z_k, z_{ik})$$

  to be maximal.

Canonical Correlation Analysis

Generalized Canonical Analysis (gCCA)

- Under the following conditions:

  1) $\left(Z_k\right)^t Z_k = 1$,

  2) $\left(Z_k\right)^t Z_j = 0, \quad j = \overline{1, k-1}$

- The variable Z_k is the eigenvector of matrix $\sum_{i=1}^{q} P_i$,

  corresponding to the eigenvalue of rank k, while the canonical variables of the sets are: $z_{ik} = P_i \cdot Z_k$.

Canonical Correlation Analysis

Generalized Canonical Analysis (gCCA)

- The orthogonal projection on space W_i is determined using the following relation:

  $P_i = X_i(X_i^t X_i)^{-1} X_i^t$, i=1, q (see the CCA).

- If the sum of the projection matrices is developed then:

  $$\sum_{i=1}^{q} P_i = \sum_{i=1}^{q} X_i \cdot \left(X_i^t X_i\right)^{-1} X_i^t = X \cdot D_{XX}^{-1} X^t ,$$

  where D_{XX} is a block-diagonal matrix of the following format:

  $$\begin{bmatrix} V_{11} & 0 & \ldots & 0 \\ 0 & V_{22} & \ldots & 0 \\ \ldots & & \ldots & \\ 0 & 0 & \ldots & V_{qq} \end{bmatrix},$$

Canonical Correlation Analysis

Generalized Canonical Analysis (gCCA)

with $V_{jj} = X_j^t X_j$

The matrix $X \cdot D_{XX}^{-1} X^t$ has *n* rows and *n* columns.

- The number non-zero eigenvalue of the matrix $X \cdot D_{XX}^{-1} X^t$, and the number of implicit steps (phases) of the algorithm is *m*, the total number of the initial (causal) variables.
- An auxiliary variable Z_k is eigenvector of the matrix

  $X \cdot D_{XX}^{-1} X^t$

  if :

  $$X \cdot D_{XX}^{-1} X^t Z_k = \alpha_k Z_k \qquad (1)$$

Canonical Correlation Analysis

Generalized Canonical Analysis (gCCA)

- Since Z_k is a linear combination of causal variables, it can be written as:

  $$Z_k = X \cdot b_k \quad ,$$

  where b_k is the vector of the linear combination, or *the factor*.
- Replacing Z_k in relation (1) it is obtained:

  $$X \cdot D_{XX}^{-1} X^t X \cdot b_k = \alpha_k X \cdot b_k \qquad (2)$$

- Multiplying relation (2) at the left with the matrix $\left(X^t X\right)^{-1} X^t$ it is obtained:

  $$\left(X^t X\right)^{-1} X^t X \cdot D_{XX}^{-1} X^t X \cdot b_k = \alpha_k \left(X^t X\right)^{-1} X^t X \cdot b_k$$

Canonical Correlation Analysis

Generalized Canonical Analysis (gCCA)

- Therefore: $D_{XX}^{-1} X^t X \cdot b_k = \alpha_k b_k$.
- Hence the factors b_k are obtained as eigenvectors of the matrix $D_{XX}^{-1} X^t X$
- The eigenvalues of the matrix $D_{XX}^{-1} X^t X$ coincide with the m non-zero eigenvalues of the matrix $X \cdot D_{XX}^{-1} X^t$.
- The eigenvalues represent the sum of the determination (correlation) coefficients between the auxiliary variables Z_k and the canonical variables of the sets:

$$\alpha_k = \sum_{j=1}^{q} R\left(Z_k, z_{jk}\right)^2 , \quad k = \overline{1,m} .$$