Software Development for Data Analysis

Canonical Correlation Analysis

Canonical Correlation Analysis (CCA)

- Describes the linear relations between 2 sets of observed or causal variables, concerning the same group of individuals (observations).
- It was first introduced and formalized by Harold Hotelling in 1936.
- In statistics, canonical analysis (from Ancient Greek: κανων bar, measuring rod, ruler) belongs to the family of regression methods employed in data analysis. It is a generalization of multiple linear regression analysis.

Canonical Correlation Analysis

Canonical Correlation Analysis (CCA)

- In linear regression, there is a dependent (explained) variable and a set of independent (explanatory) variables.
- Contrary to linear regression, in canonical correlation analysis both set of variables play the same role, as explanatory variables.
- Canonical correlation analysis determines to what degree, 2 sets of variables reflect or not the same reality.

Canonical Correlation Analysis

CCA data

- The data could be presented in 2 matrices X and Y, with ***n*** rows, ***p*** and ***q*** columns, respectively, having the following format:

$$X_{n\times p} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ & & \vdots & & \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ & & \vdots & & \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} \quad Y_{n\times q} = \begin{pmatrix} y_{11} & \cdots & y_{1j} & \cdots & y_{1q} \\ & & \vdots & & \\ y_{i1} & \cdots & y_{ij} & \cdots & y_{iq} \\ & & \vdots & & \\ y_{n1} & \cdots & y_{nj} & \cdots & y_{nq} \end{pmatrix}$$

Canonical Correlation Analysis

CCA data

- The data can be centered or standardized. We will discussed the centered data variant of analysis.
- The columns of table X define $\boldsymbol{p}$ quantitative variables, while the columns of table Y define $\boldsymbol{q}$ quantitative variables. The assumption is that the matrix X is a $\boldsymbol{p}$ rank one, and matrix Y is a $\boldsymbol{q}$ rank one.
- CA is a iterative process, in k steps, having as result the extractions of k pairs of new variables, named canonical variables, (z_i, u_i), $i = 1,k$.

Canonical Correlation Analysis

CCA data

- The z_i variables belong to W_1 space, generated by the columns of matrix X, while the u_i variables belong to W_2 space, generated by the columns of matrix Y.
- The variables making up a canonical pair are maximum correlated between each other, and completely uncorrelated toward all the other canonical variable belonging to the same space.

Canonical Correlation Analysis

CCA phases (steps)

1. Determine a pair of canonical variables (z_1, u_1) as a linear combination of causal variables:

- z_1 is a linear combination or variables X_1, ... , X_p
- while u_1 is linear combination of variables Y_1, ... , Y_q.

$$z_1 = X \cdot a_1$$

$$u_1 = Y \cdot b_1$$

Canonical Correlation Analysis

CCA phases (steps)

$$z_1 = a_{11} \cdot X_1 + a_{21} \cdot X_2 + ... + a_{p1} \cdot X_p = X \cdot a_1$$

where $a_1 = \begin{bmatrix} a_{11} \\ ... \\ a_{p1} \end{bmatrix}$,

$$u_1 = b_{11} \cdot Y_1 + b_{21} \cdot Y_2 + ... + b_{q1} \cdot Y_q = Y \cdot b_1$$

where $b_1 = \begin{bmatrix} b_{11} \\ ... \\ b_{q1} \end{bmatrix}$

Canonical Correlation Analysis

CCA phases (steps)

- The canonical variables are maximum correlated between each other.
- Therefore, multiplying them with scalars, the correlation is maintained: $R(z_1, u_1) = R(\alpha \cdot z_1, \beta \cdot u_1)$.
- To ensure their unicity, the restriction of normality has to be imposed: $(z_1)^t z_1 = 1$ and $(u_1)^t u_1 = 1$.

Canonical Correlation Analysis

CCA phases (steps)

- At the first step, the solution of the problem is the following:
  - z_1 is the first eigenvector of the P_1P_2 matrix, corresponding to the greatest eigenvalue,
  - while u_1 is the first eigenvector of P_2P_1 matrix, corresponding to the same eigenvalue.
- P_1 and P_2 are the linear orthogonal projectors on W_1 and W_2 spaces, generated by the columns of the matrices X and Y, respectively.
- The eigenvalue α_1 is the correlation coefficient between the canonical variables z_1 and u_1.

Canonical Correlation Analysis

CCA phases (steps)

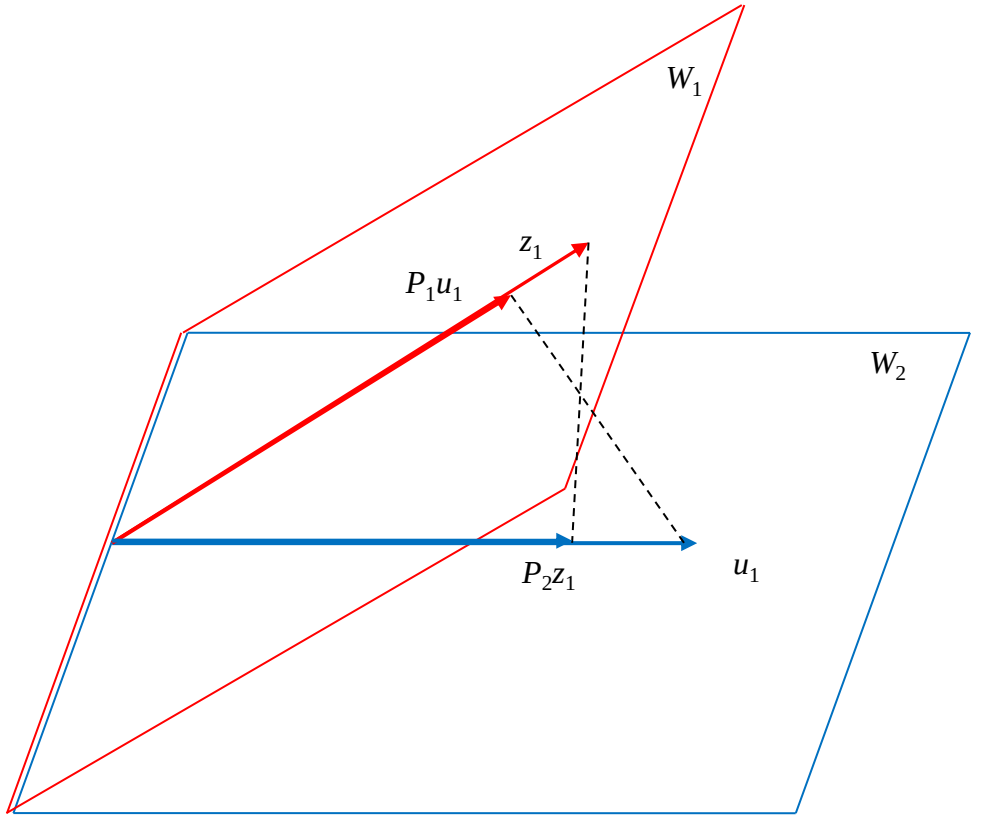
- For a given vector z_1, $z_1 \in R^n$, the vector belonging to W_1 space which makes the smallest angle with z_1 is the orthogonal projection of z_1 on W_2 space.
- Hence, $R^2(z_1,u_1)$ is maximal if u_1 is collinear with the orthogonal projection of z_1 on W_2 space.
- The projection of z_1 vector on W_2 space is, at the same time, the projection on the u_1 vector axis.
- Since the vectors are normalized, and their correlation is given by the cosine of the angle between them, we have:

$$P_2 z_1 = R^2(z_1, u_1)\, u_1 \qquad (1)$$

see the following figure.

Canonical Correlation Analysis

CCA phases (steps) – a graphical perspective



Canonical Correlation Analysis

CCA phases (steps)

- Symmetrically, for a given vector u_1 in in the space W_2, $R^2(z_1,u_1)$ is maximal if the vector z_1 is collinear with the orthogonal projection of u_1 on W_1 space. Therefore:

  $P_1u_1=R^2(z_1, u_1)z_1$ (2), then multiply (1) with P_1 on the left, and use (2)

  $P_1P_2z_1 = P_1R^2(z_1, u_1)u_1 = R^2(z_1, u_1)P_1u_1 = R^2(z_1, u_1)z_1$

  $\Rightarrow$ $\boldsymbol{P_1P_2z_1=R^2(z_1,u_1)z_1}$, where $R^2(z_1,u_1)$ is maximal.
- z_1 is the eigenvector of the matrix P_1P_2, corresponding to the greatest eigenvalue $\alpha_1 = R^2(z_1, u_1)$

  $\Rightarrow$ $\boldsymbol{P_2P_1u_1} = P_2R^2(z_1, u_1)z_1 = R^2(z_1, u_1)P_2z_1 = \mathbf{R^2(z_1, u_1)u_1}$
- u_1 is the eigenvector of the matrix P_2P_1, corresponding to the greatest eigenvalue $\alpha_1 = R^2(z_1,u_1)$

Canonical Correlation Analysis

CCA phases (steps)

2. Determine the second pair of canonical variables, z_2 and u_2 such that to obtain the maximum correlation coefficient $R^2(z_2, u_2)$, along with the following restrictions:

$$\begin{cases} (z_2)^t z_2 = 1, \quad (u_2)^t u_2 = 1 \\ (z_2)^t z_1 = 0, \quad (u_2)^t u_1 = 0 \end{cases}$$

- Similarly to the first step, the problem's solutions are the eigenvectors of the matrices P_1P_2 and P_2P_1, corresponding to the second biggest eigenvalue.
- The eigenvalue α_2 is $R^2(z_2,u_2)$, the correlation coefficient between the canonical variables z_2 and u_2:

  $P_1P_2z_2 = \alpha_2 z_2, \;\; P_2P_1u_2 = \alpha_2 u_2.$

Canonical Correlation Analysis

CCA phases (steps)

- In addition to the first step, there has to be proved that z_2 and z_1, u_2 and u_1 respectively, are not correlated at all, meaning that: $(z_2)^t z_1 = 0$ and $(u_2)^t u_1 = 0$
- Developing $\alpha_2 (z_2)^t z_1$, we have:

  $\alpha_2 (z_2)^t z_1 = (z_2)^t P_2 P_1 z_1$, because $(P_1 P_2 z_2)^t = (\alpha_2 \cdot z_2)^t$
- Then $\alpha_2 (z_2)^t z_1 = (z_2)^t P_1 P_2 P_1 z_1$, since the projection of vector on its own space is the vector itself, $P_1 \cdot z_2 = z_2$, or $(z_2)^t \cdot P_1 = (z_2)^t$.
- Therefore:

  $\alpha_2 (z_2)^t z_1 = (z_2)^t P_1 P_2 z_1$, because $P_1 \cdot z_1 = z_1$

Canonical Correlation Analysis

CCA phases (steps)

- Consequently:

$\alpha_2 (z_2)^t z_1 = \alpha_1 (z_2)^t z_1$, because $P_1 P_2 z_1 = \alpha_1 \cdot z_1$

But $\alpha_1 \neq \alpha_2$, implying that $(z_2)^t z_1 = 0$

- Similarly:

$$\alpha_2 (u_2)^t u_1 = (u_2)^t P_1 P_2 u_1 = (u_2)^t P_2 P_1 P_2 u_1 = (u_2)^t P_2 P_1 u_1 =$$

$$= \alpha_1 (u_2)^t u_1$$

- Therefore: $(u_2)^t u_1 = 0$

Canonical Correlation Analysis

CCA phases (steps)

k. Determine the pair *k* of canonical variables (z_k, u_k) such that they are maximally correlated, their variance to be 1, and the correlation coefficient against the canonical variables determined at the previous steps to be 0 (zero).

$$\begin{cases} \underset{z^k, u^k}{Max}(R^2(z_k, u_k)) \\ (z_k)^t z_k = 1, \quad (u_k)^t u_k = 1 \\ (z_k)^t z_i = 0, (u_k)^t u_i = 0, \quad i = \overline{1, k-1} \end{cases}$$

Canonical Correlation Analysis

CCA phases (steps)

- The problem solutions are the eigenvectors of the matrices P_1P_2 and P_2P_1 , corresponding to the eigenvalue of rank *k*.
- Such a pair it is called ***canonical root***.
- The canonical variables of the same set are not correlated to each other.
- It can be proved that canonical variables of different ranks from different groups are uncorrelated as well, used (1):

$$(z_r)^t u_k = (P_1 z_r)^t u_k = (z_r)^t P_1 u_k = (z_r)^t R(z_k, u_k) z_k = R(z_k, u_k)(z_r)^t z_k = 0$$

Canonical Correlation Analysis

Canonical factors

- Canonical variables are linear combinations of the initial variables from those 2 sets, so:

  $z_i = X \cdot a_i,\ u_i = Y \cdot b_i,\ i=1,k$

  where unde a_i and b_i are the corresponding canonical factors.
- It has to be notice that the matrices P_1P_2 and P_2P_1 are of rank n.
- In most of the cases the number of individuals (observations) is greater than the number of variables.
- Consequently, determining the eigenvectors and the eigenvalues are intensively resource consuming operations on such a matrix.

Canonical Analysis

Canonical factors

- At the step k we determine the eigenvector z_k of the matrix P_1P_2. Therefore:

  $P_1P_2z_k = R^2(z_k, u_k)z_k$.
- Since $P_1 = X(X^tX)^{-1}X^t$ and $P_2 = Y(Y^tY)^{-1}Y^t$ (see the multiple variable regression), and replacing $z_k = Xa_k$, we obtain that:

  $X(X^tX)^{-1}X^tY(Y^tY)^{-1}Y^t\ Xa_k = R^2(z_k, u_k)\ Xa_k$.
- Multiplying this relation at the left with $(X^tX)^{-1}X^t$, then:

  $(X^tX)^{-1}X^tY(Y^tY)^{-1}Y^t\ Xa_k = R^2(z_k, u_k)\ a_k$.

Canonical Analysis

Canonical factors

- We make the following notations:
  $V_{11} = \frac{1}{n} X^t X$, $V_{22} = \frac{1}{n} Y^t Y$, $V_{12} = \frac{1}{n} X^t Y$, $V_{21} = \frac{1}{n} Y^t X$,
- Then we have:
  $V_{11}^{-1} V_{12} V_{22}^{-1} V_{21} a_k = R^2(z_k, u_k) a_k$.
- Similarly, for the factors corresponding to the second data set:
  $V_{22}^{-1} V_{21} V_{11}^{-1} V_{12} b_k = R^2(z_k, u_k) b_k$.
- Should be noticed that V_{11} is the covariance matrix between the variables from X set and V_{22} is the covariance matrix between the variables from Y set, while V_{12} is the covariance matrix between the variables from X and Y sets, and V_{21} is the covariance matrix between the variables frome Y and X sets.

Canonical Analysis

Canonical factors

- Therefore a_k is the eigenvector of rank k of the matrix $V_{11}^{-1}V_{12}V_{22}^{-1}V_{21}$, corresponding to the eigenvalue
  $$\alpha_k = R^2(z_k, u_k),$$
- And b_k is the eigenvector of rank k of the matrix $V_{22}^{-1}V_{21}V_{11}^{-1}V_{12}$, corresponding to the same eigenvalue α_k.
- Since these 2 matrices have p and q columns respectively, it is much more convenient to determine the canonical variables in this manner.
- The number of steps is given by $m = \min(p, q)$.

Canonical Analysis

The connection between the canonical factors

- We have the following relation:
  $P_2 z_k = R(z_k, u_k) u_k$
- Making in the above relation the following substitutions:
  $P_2 = Y(Y^t Y) Y^t$

  $z_k = X a_k \,, \;\; u_k = Y b_k$
- Then we obtain that:
  $Y(Y^t Y)^{-1} Y^t X a_k = R(z_k, u_k) Y b_k$
- Multiplying at the left with $(Y^t Y)^{-1} Y^t$ then:

  $V_{22}^{-1} V_{21} a_k = R(z_k, u_k) b_k \,.$

Canonical Analysis

The connection between the canonical factors

- Therefore, the connection between the canonical factors are given by the followings relations:

$$b_k = \frac{1}{R(z_k, u_k)} V_{22}^{-1} V_{21} a_k$$

$$a_k = \frac{1}{R(z_k, u_k)} V_{11}^{-1} V_{12} b_k$$