Photo: Vladimir Batagelj, ***UNI-LJ***

Canonical Correlation Analysis

Anuška Ferligoj

UL, Ljubljana, Slovenia

NRU HSE, Moscow, Russia

Content

Introduction

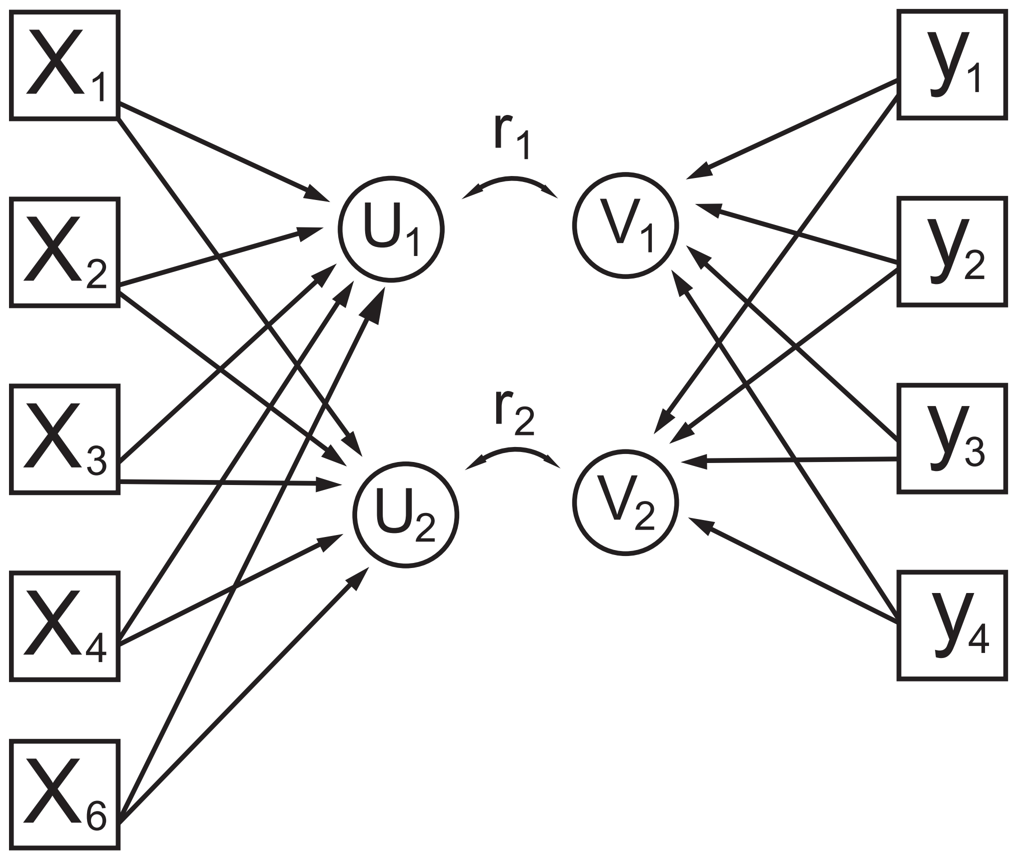
Canonical correlation analysis describes a multivariate technique that investigates the relationship between **two sets of variables**.

Hotelling (1936) introduced canonical correlation analysis by the research question whether the reading ability (measured by reading speed and reading power) is related to the arithmetic ability (measured by arithmetic speed and arithmetic power).

To study this relationship he searched for a linear combination of two variables measuring reading ability and another linear combination of two variables measuring arithmetic ability in such a way that two linear combinations correlated as much as possible.

Some definitions

- The linear combination of the first set of variables is called **canonical variable of the first set of variables (U)**.
- Similarly the linear combination of the second set of variables is called **canonical variable of the second set of variables (V)**.
- The maximal correlation between two canonical variables is called **canonical correlation coefficient ($r(U, V)$)**.
- The triplet U, V and $r(U, V)$ is called **canonical solution**.


X_1
X_2
X_3
X_4
X_6
U_1
U_2
r_1
r_2
V_1
V_2
y_1
y_2
y_3
y_4

The first set of variables: X_1, X_2, ... , X_p
The second set of variables: Y_1, Y_2, ... , Y_q

$\min(p, q)$ canonical solutions can be obtained.

First canonical solution:

$$U_1 = c_{11}X_1 + c_{12}X_2 + ... + c_{1p}X_p$$

$$V_1 = d_{11}Y_1 + d_{12}Y_2 + ... + d_{1p}Y_q$$

We search for such canonical loadings c_{ij} and d_{ij} that

$$r(U_1, V_1) = max$$

The second canonical solution is

$$U_2 = c_{21}X_1 + c_{22}X_2 + ... + c_{2p}X_p$$

$$V_2 = d_{21}Y_1 + d_{22}Y_2 + ... + d_{2p}Y_q$$

$$r(U_2, V_2) = max$$

U_2 and V_2 have to be uncorrelated with U_1 and V_1.

The third canonical solution is

$$U_3 = c_{31}X_1 + c_{32}X_2 + ... + c_{3p}X_p$$

$$V_3 = d_{31}Y_1 + d_{32}Y_2 + ... + d_{3p}Y_q$$

$$r(U_3, V_3) = max$$

U_3 and V_3 have to be uncorrelated with U_1, V_1, U_2, and V_2; and so on.

The first canonical solution has the highest possible correlation and is the most important; the second solution has the second highest correlation and is therefore the second most important; etc.

Estimation

$$\Sigma = \begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_p \\ Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{array} \begin{array}{c} \begin{array}{cccccccc} X_1 & X_2 & \cdots & X_p & Y_1 & Y_2 & \cdots & Y_q \end{array} \\ \left[\begin{array}{c|c} \Sigma_{XX} & \Sigma_{XY} \\ \hline \Sigma_{YX} & \Sigma_{YY} \end{array} \right] \end{array}$$

$|\Sigma_{XX}| \neq 0, |\Sigma_{YY}| \neq 0, \Sigma_{XY} = \Sigma'_{YX}$

We have to estimate c_{ij}, d_{ij}, and r_i by solving the optimization problem:

$$r_i(U_i, V_i) = \max$$

The solution is obtained by determining of the eigenvalues and eigenvectors of the following matrices:

$$Q_1 = \Sigma_{XX}^{-1} \cdot \Sigma_{XY} \cdot \Sigma_{YY}^{-1} \cdot \Sigma_{YX}$$

$$Q_2 = \Sigma_{YY}^{-1} \cdot \Sigma_{YX} \cdot \Sigma_{XX}^{-1} \cdot \Sigma_{XY}$$

The eigenvalues λ_i of Q_1 and Q_2 are the same and equal to

$$\lambda_i = r_i^2$$

Eigenvectors of the matrix Q_1 are the canonical loadings of the canonical variables U_i and eigenvectors of the matrix Q_2 are the canonical loadings of the canonical variables V_i.

There is a relationship between loadings c and d:

$$c = \frac{\Sigma_{XX}^{-1} \cdot \Sigma_{XY} \cdot d}{\sqrt{\lambda}}$$

$$d = \frac{\Sigma_{YY}^{-1} \cdot \Sigma_{YX} \cdot c}{\sqrt{\lambda}}$$

These loadings are **regression coefficients**.

Canonical structure loadings c^* and d^* are:

$$c_j^* = \Sigma_{XX} \cdot c_j$$

$$d_j^* = \Sigma_{YY} \cdot d_j$$

Canonical structure loadings are **correlation coefficients**.

Example

1st set of variables: Extraversion
(from Big Five, International Personality Item Pool
http://ipip.ori.org/ipip/)

EXTA	Am the life of the party.
EXTB	Don't mind being the center of attention.
EXTE	Talk to a lot of different people at parties.
EXTH	Start conversations.
EXTIR	Don't like to draw attention to myself. (*)
EXTLR	Don't talk a lot. (*)
EXTNR	Am quiet around strangers. (*)
EXTOR	Have little to say. (*)
EXTP	Feel comfortable around people.
EXTRR	Keep in the background. (*)

Scale:
1 Very Inaccurate
2 Moderately Inaccurate
3 Neither Accurate Nor Inaccurate
4 Moderately Accurate
5 Very Accurate

The statements marked with (*) were negative statements and were in prior to the analysis recoded (1=5) (2=4) (3=3) (4=2) (5=1).

2nd set of variables: education and age

EDU education (ordinal scale)

1 – uncompleted primary school
2 – completed primary school
3 – vocational school
4 – four year secondary school
5 – non-university collage
6 – university collage
7 – masters
8 – PhD

AGE (ratio scale)

Root No.	Eigenvalue	Canon Cor.
1	,227	,430
2	,046	,210

Standardized canonical coefficients

Variable	1	2
EKSTA	-,326	,781
EKSTB	,035	,003
EKSTE	,048	,268
EKSTH	-,013	-,548
EKSTIR	,282	,252
EKSTLR	-,095	,164
EKSTNR	,081	,053
EKSTOR	,417	-,192
EKSTP	-,270	-,147
EKSTRR	,639	-,012

Correlations between variables and canonical variables

Variable	1	2
EKSTA	-,090	,762
EKSTB	,242	,152
EKSTE	,170	,326
EKSTH	,117	-,310
EKSTIR	,528	,340
EKSTLR	,301	,236
EKSTNR	,466	,128
EKSTOR	,636	-,036
EKSTP	-,084	-,076
EKSTRR	,798	,210

Standardized canonical coefficients

VARIABLE	1	2
AGE	-,893	-,457
EDU	,524	-,855

Correlations between variables and canonical variables

VARIABLE	1	2
AGE	-,852	-,523
EDU	,456	-,890