

Chapter XV**Canonical Correlation Analysis****Siripuram Haripriya¹, B. Nirmala² and Santosha Rathod²**

¹Professor Jayashankar Telangana State Agricultural University,
Hyderabad-500030

²ICAR-Indian Institute of Rice Research, Hyderabad.
Email: haripriyasiripuram1532@gmail.com

Canonical correlation analysis (CCA) Originally defined by Hotelling in 1936 (Abdi *et al.*, 2018). It is a statistical method whose goal is to extract the information common to two data tables that measure quantitative variables on the same set of observations. Canonical correlation analysis creates pairs of linear combinations of the variables (one per table) that have maximal correlation. CCA generalizes many standard statistical techniques (*e.g.*, multiple regression, analysis of variance, discriminant analysis). After principal component analysis (PCA), CCA is one of the oldest multivariate techniques. CCA has been really feasible only with the advent of modern computers. Recent developments involve the generalization of CCA to the case of more than two tables and cross-validation approaches to select important variables and the stability and reliability of the solution obtained on a given sample. For each data table, CCA computes a set of linear combinations of the variables of this table called latent variables or canonical variates with the constraints that a latent variable from one table has a maximal correlation with one latent variable of the other table and no correlation with the remaining latent variables of the other table.

Canonical correlation analysis can be used in a wide field of study (plant science, biology, chemistry, meteorology, demography, artificial intelligence, cognitive science, political science, sociology, psychometrics, educational research, economics, and management science) to analyse multidimensional relationships between multiple independent and dependent variable sets (Akbas and Takma, 2005). In agriculture, CCA analysis is widely used as an indirect selection method for the genetic improvement of plants.

Linear combination of two set of variables

The association between two sets of variables may be quantified by canonical correlation analysis (CCA). Given a set of variables $X \in R^q$ and another set $Y \in R^p$, one asks for the linear combination $a^T X$ That “best matches” a linear combination $b^T Y$. The best match in CCA is defined through maximal correlation. The task of CCA is, therefore, to find $a \in R^q$ and $b \in R^p$ so that the correlation $\rho(a, b) = \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a} \sqrt{b^T \Sigma_{YY} b}}$ is maximized. The best-matching linear combination $a^T X$ and $b^T Y$ are then called canonical correlation variables their correlation is the canonical correlation coefficient. The coefficients a and b of the canonical correlation variable are the vectors.

Let us assume that the two random vectors under investigation, X and Y have the following covariance structure.

$$\text{Var} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}$$

The algorithm of CCA consists of calculating the matrix

$$\kappa = \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2}$$

and it's SVD

$$\kappa = \Gamma \Lambda \Delta^T$$

The diagonal elements of the matrix Λ are the canonical correlation coefficients. The canonical correlation vectors can be obtained as

$$a_i = \Sigma_{XX}^{-1/2} \gamma_i$$

$$b_i = \Sigma_{YY}^{-1/2} \delta_i$$

and the canonical correlation variables are

$$\eta_i = a_i^T X,$$

$$\varphi_i = b_i^T Y.$$

It can be easily verified that

$$\text{Var} \begin{pmatrix} \eta \\ \varphi \end{pmatrix} = \begin{pmatrix} I_k & \Lambda \\ \Lambda & I_k \end{pmatrix}$$

The assumptions for conducting canonical correlation.

1. It is assumed that the interval type of data is used to carry out canonical correlation.
2. It is assumed in canonical correlation that the relationships should be linear in nature.
3. It is assumed that there should be low multicollinearity in the data while performing canonical correlation. If the two sets of data are highly inter-correlated, then the coefficient of the canonical correlation is unstable.
4. There should be unrestricted variance in canonical correlation. If the variance is not unrestricted, then this might make the canonical correlation look unstable.

The interpretation of canonical correlation analysis: can be performed in different ways: canonical weights or standardized coefficients, canonical loadings, and canonical cross-loadings (Hair Junior *et al.*, 1998). Canonical loadings measure the simple linear correlation between an original observed variable in the dependent or independent set and the canonical statistical variable of the set; and the higher the coefficient, the more important the variable for the canonical statistical variable (Hair Júnior *et al.*, 1998). Canonical loadings have advantages over canonical weights, such as lower standard error and greater stability in replicate samples (Meredith and William, 1964). The use of canonical cross-loadings consists of the correlation between the observed dependent variable and the independent canonical statistical variable, and vice versa, providing a direct measure of the relations of dependent-independent variables by eliminating an intermediate step involved in conventional loadings (Hair *et al.*, 2017).

R codes to perform canonical correlation analysis

```
require(ggplot2)
require(GGally)
require(CCA)
require(CCP)
#import the data
colnames(mm) <- c()
summary(mm)
psych <- mm[, 1:3]
acad <- mm[, 4:8]
```

```
ggpairs(psych)
# correlations
matcor(psych, acad)
cc1 <- cc(psych, acad)
# display the canonical correlations
cc1$cor
# raw canonical coefficients
cc1[3:4]
# compute canonical loadings
cc2 <- comput(psych, acad, cc1)
# display canonical loadings
cc2[3:6]
# tests of canonical dimensions
rho <- cc1$cor
## Define number of observations, number of variables in first set,
and number of variables in the second set.
n <- dim(psych)[1]
p <- length(psych)
q <- length(acad)
## Calculate p-values using the F-approximations of different test
statistics:
p.asym(rho, n, p, q, tstat = "Wilks")
p.asym(rho, n, p, q, tstat = "Hotelling")
p.asym(rho, n, p, q, tstat = "Pillai")
p.asym(rho, n, p, q, tstat = "Roy")
# standardized psych canonical coefficients diagonal matrix of psych
sd's
s1 <- diag(sqrt(diag(cov(psych))))
s1 %*% cc1$xcoef
#standardized acad canonical coefficients diagonal matrix of acad
sd's
s2 <- diag(sqrt(diag(cov(acad))))
s2 %*% cc1$ycoef
```

Canonical correlation analysis is used to describe the relationship between the two sets of variables. And independence between the two sets of variables.

References

- Hair, Joseph F., William C. Black, Barry J. Babin, and Rolph E. Anderson. (1998). *Multivariate Data Analysis*. v.
- Abdi, H., Guillemot, V., Eslami, A., & Beaton, D. (2018). Canonical Correlation Analysis.
- Akbas, Y. and Takma, C. H. (2005). Canonical correlation analysis for studying the relationship between egg production traits and body weight, egg weight and age at sexual maturity in layers. *Czech Journal of Animal Science*, 50(4), 163-168.
- Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution* (pp. 162-190). New York, NY: Springer New York.
- Hair Jr, J. F., Babin, B. J., & Krey, N. (2017). Covariance-based structural equation modeling in the Journal of Advertising: Review and recommendations. *Journal of Advertising*, 46(1), 163-177.
- Hair, Joseph F., William C. Black, Barry J. Babin, and Rolph E. Anderson. 1998. *Multivariate Data Analysis*. v.
- Meredith, W. (1964). Canonical correlations with fallible data. *Psychometrika*, 29(1), 55-65.

How to Cite: Haripriya, S., Nirmala, B., and Rathod, S. (2023). Canonical Correlation Analysis (CCA). In S. Rathod, B. Sailaja, N. Bandumula, S. Arun Kumar, P. A. Lakshmi Prasanna, P. Jeyakumar, A. Waris, P. Muthuraman, and R. M. Sundaram (Eds.), *Statistical Procedures for Analysing Agricultural Data using R* (pp. 191-195). ICAR - Indian Institute of Rice Research, Hyderabad.