# COMPARISON OF CANONICAL CORRELATION AND DISCRIMINANT ANALYSIS

Alvin F. Terry[1] , Etikan I[2]

1. **PhD Student,** Near East University Faculty of Medicine, Biostatistics Department Near East Avenue, Postcode: 99138, Nicosia, North Cyprus, **Mersin 10, Turkey**
2. Near East University Faculty of Medicine Head of Biostatistics Department Near East Avenue, Postcode: 99138, Nicosia, North Cyprus, **Mersin 10, Turkey**

ARTICLE INFO

**Keywords:**
Biostatistics, Canonical correlation, Discriminant analysis

**Corresponding author:**
**Prof. Dr. Etikan I**
**Email:** ilker.etikan@neu.edu.tr

Vol 01 Issue 02
APR-JUNE 2023

**ISSN Online:** 2960-2599
**ISSN Print:** 2960-2580

REVIEW ARTICLE

## ABSTRACT

Learning from data is at the heart of the statistical discipline known as statistics. Understanding statistics, you will be able to use the right techniques to obtain the data, carry out the relevant analysis, and present the results in an effective way if you have a working knowledge of statistics. The process of creating scientific discoveries, as well as decisions and projections that are founded on data, involves the use of statistics, which is a key element in the process.

## INTRODUCTION

Learning from data is at the heart of the statistical discipline known as statistics. Understanding statistics, you will be able to use the right techniques to obtain the data, carry out the relevant analysis, and present the results in an effective way if you have a working knowledge of statistics. The process of creating scientific discoveries, as well as decisions and projections that are founded on data, involves the use of statistics, which is a key element in the process. By making use of statistics, one is able to have a much better understanding of a subject.[1] Which statistical test is used to analyze research data depends on a variety of factors, including the study's premise, the data's type, the number of measurements, and whether or not the data are paired or unpaired. Several errors in the use of descriptive and inferential statistics were found in assessments of publications in the medical domains of family medicine, cytopathology, and pain.[2-5] Those are a few of the many studies that have highlighted that many wrong conclusions have been derived by using the wrong statistical test and wrong inferences have been made. This article is intended to compare statistical tests which show an association between data and they are canonical correlation and discriminant analysis.

### Canonical Correlation

Canonical Correlation Analysis is a statistical method for finding pairs of random variables that optimize their correlation by linear combination. The canonical correlation analysis, often known as CCA, is one of the potential methods for elucidating these combined multivariate correlations between the various modalities.[6] CCA is able to detect the cause of statistical fluctuations that are consistent across several modalities without supposing any specific sort of directionality. Multimodal data fusion relies heavily on canonical correlation analysis (CCA) since it has largely replaced the univariate general linear model (GLM) in connecting different modalities.[7] Finding and quantifying relationships between groups of data is the job of canonical correlation analysis. In cases when many outcome variables are highly associated with one another, canonical correlation may be used instead of multiple regression. Canonical correlation analysis identifies the canonical variates, or orthogonal linear combinations of the variables in each set, that best explain the variability within and across sets.[8] Each set of variables must be collapsed into a single variable, and then their variables must be determined, in order to perform a canonical correlation. And the two variables are determined by performing linear combinations of the variables in each set while holding all other factors constant. Canonical Correlation refers to the relationship between canonical variables, which are the results of a linear combination.[9]

### Reason for using canonical Correlation

The goal of the approach known as canonical correlation is to discover and quantify the nature of the connection that exists between two distinct groups of data. In addition to this, it is a well-known statistical method that is used extensively in a variety of subfields within the social sciences, psychology research, and marketing analytics. The researchers are able to monitor the link between a large number of dependent and independent variables, in contrast to regression analysis[9]

### Assumptions of Canonical Correlation

**1.** When carrying out canonical correlation, it is presumed that the interval kind of data will be used.

**2.** Pre-supposes that there is a linear connection between the variables being studied (the dependent and the independent ones).

**3.** When using canonical correlation, it is expected that there would be little multi-collinearity in the data. When there is a great deal of connection between the two data sets, the canonical correlation coefficient tends to fluctuate.

**4.** Variability that is constant throughout the data set.

**5.** It requires a multivariate normality test.

### Terms associated with Canonical Correlation

A canonical variable or variate - is the linear combination of the initial set of variables. Latent variables may be described as this kind of variable. Eigenvalues - Canonical correlation uses the approximation that the square of an Eigenvalue is equal to that value. The canonical correlation between the two sets of variables accounts for some of the variations in the canonical variate, which is reflected in the Eigenvalues. Canonical Weight - Canonical weight is often referred to as the canonical coefficient. Canonical correlation requires initial standard-ization of the canonical weight. This information is then utilized to evaluate the variable's significance in the context of the whole. Redundancy coefficient, d - This canonical correlation coefficient shows how much one set of data may be predicted by looking at the other sets. Likelihood ratio test - This canonical correlation significance test is designed to check the validity of the linear connection between the two canonical variables.

### Discriminant Analysis

Discriminant analysis (DA) is a multivariate method used to categorize data into distinct groups according to the values of one or more variables assessed for each subject in a study's sample. Assigning each person to preexisting groups using a linear or quadratic function may also be studied for purposes of prediction or allocation. This may be accomplished by classifying people into the appropriate categories.[10] Using simply the linear combination of inde-pendent variables, discriminant analysis seeks to generate discriminant functions that reliably separate categories of the dependent variable. Researchers may use this method to check for differences in predictor factors across sets of subjects. The accuracy of the categorization is measured as well.[11]

## Assumptions

**1.** It's important for samples to be autonomous and unrelated to one another.

**2.** Each group's variance-covariance matrix should be the same, and the predictor variables should follow a multivariate normal distribution.

**3.** Since group membership is presumed to be exclusive, it follows that no case may correspond to more than one group and therefore all cases must belong to a group.

**4.** It assumes that group membership is a true categorical variable.

## TYPES OF DISCRIMINANT ANALYSIS

### Linear Discriminant Analysis

This method, also known as linear discriminant analysis (LDA), takes the independent variables and uses their linear combination to make a prediction about the class of the dependent variable. It assumes that the variance and covariance of each group are the same and that the independent variables follow a normal distribution (continuous and numerical). This technique may be used for both categorization and conditionality reduction.

### Quadratic Discriminant Analysis

Using quadratic combinations of independent variables, this variant of Linear Discriminant Analysis (LDA) may be used to predict the category of the dependent variable of interest. We continue to assume a normal distribution. Although it does not assume that each class has the same covariance. A quadratic decision boundary is generated by the QDA.

## Comparison of Canonical Correlation & Discriminant Analysis

There are many similarities share between the two types of statistical analysis in terms of the tactics that are used for analysis. The main similarities are:

**1.** They are both multivariate analysis that is used in statistics. They consider the relationships and interactions between multiple independent variables and their association with dependent or categorical variables.

**2.** They both result in the reduction of the data - The goal of both methods is to minimize the dimensionality of the data by developing new variables or functions that are able to isolate the most relevant aspects of the data or differentiate between different categories. In canonical variate analysis (CCA), linear combinations of the original variables are used to produce canonical variates. In discriminant analysis, on the other hand, discriminant functions are developed.

**3.** They both examine the relationship between variables - The purpose of the canonical correlation analysis (CCA) is to locate linear combinations of variables from two distinct sets (X and Y) that have the highest possible correlation. The goal of discriminant analysis is to find linear combinations of independent variables that provide the greatest degree of differentiation or separation between the different groups of people.

**4.** They both share assumptions in common- they both assume a linear relationship, they both have equal variances assumed and they both check for normality assumptions.

While they both have some similarities as a multivariate analysis, they also have dissimilarities that set them apart as a unique form of multivariate analysis. They can be distinguished based on the following:

**1.** Kind of analysis - Canonical correlation analysis (CCA) is a method of exploratory analysis that is used to get a better understanding of the links that exist between two different sets of data. On the other hand, discriminant analysis is a method of supervised classification that may be used for a variety of problems such as to carry-on forecasting and categorization.

**2.** Aim - The objective of canonical correlation analysis (CCA) is to investigate the connection between two sets of data, often referred to as X and Y, and determine the straight-line connections that are most significant between them. It is an approach that is descriptive. Discriminant analysis, on the other hand, is centered on the goal of predicted classification by the maximization of the separation between previously specified groups or classes.

**3.** Measures variable - Both sets of variables are considered dependent in the CCA analysis. The highest correlating canonical variates are sought. The discriminant analysis makes use of a categorical dependent variable to examine relationships between sets of independent factors. As we have seen from the various discussion of the two types of statistical analysis, they both are unique and well-defined to a specific aim or objective that the researcher has in mind. Both canonical correlation analysis and discriminant analysis are multivariate approaches that seek to uncover correlations between variables, their goals, areas of attention, and applications are somewhat different from one another. The canonical correlation analysis (CCA) investigates the link between two sets of data, while the discriminant analysis concentrates on classifying and predicting based on the differences between groups.

## AUTHORS CONTRIBUTION

**AFT**: Main author, **EI:** Co-author

## REFERENCES

1. Stärk K, Kidd E, Frost RL. Close encounters of the word kind: Attested distributional information boosts statistical learning. Language Learning. 2023;73(2):341-73.

2. Hasanzadeh Kiabi F, Alipour A, Darvishi-Khezri H, Aliasgharian A, Emami Zeydi A. Zinc Supplementation in Adult Mechanically Ventilated Trauma Patients is Associated with Decreased Occurrence of Ventilator-associated Pneumonia: A Secondary Analysis of a Prospective, Observational Study. Indian J Crit Care Med. 2017;21(1):34-9.

3. Yim KH, Nahm FS, Han KA, Park SY. Analysis of statistical methods and errors in the articles published in the Korean journal of pain. Korean J Pain. 2010;23(1):35-41.

4. Bahar B, Pambuccian SE, Barkan GA, Akdaş Y. The use and misuse of statistical methods in cytopathology studies: review of 6 journals. Lab Med. 2019;50(1):8-15.

5. Nour-Eldein H. Statistical methods and errors in family medicine articles between 2010 and 2014-Suez Canal University, Egypt: A cross-sectional study. J Family Med Prim Care. 2016;5(1):24-33.

6. Harold H. Relations between two sets of variables. Biometrika. 1936;28(3):321-77.

7. Zhuang X, Yang Z, Cordes D. A technical review of canonical correlation analysis for neuroscience applications. Hum Brain Mapp. 2020;41(13):3807-33.

8. UCLA. Canonical Correlation Analysis. Statistical Consulting Group. from https://stats.oarc.ucla.edu/sas/modules/introduction-to-the-features-of-sas/ (accessed June 9, 2023). 2023.

9. Hessing T. Canonical Correlation Analysis. https://sixsigmastudyguide.com/canonical-correlation-analysis/( accessed June 9,

2023). 2023.

10. Dheeraj V. Discriminant Analysis. https://www.wallstreetmojo.com/discriminant-analysis/ (retrieved June 9, 2023). 2023.

11. Statistics solution (2023). Discriminant Analysis.https://www.statisticssolutions.com/discriminant-analysis/ (retrieved June 9, 2023).