UNIVERSITY OF SCIENCE

FACULTY OF INFORMATION TECHNOLOGY

MULTIVARIATE STATISTICAL ANALYSIS

Report Practice 05
Canonical Correlation Analysis

Lecturers Lý Quốc Ngọc
Nguyễn Mạnh Hùng
Phạm Thanh Tùng

Student Võ Nguyễn Hoàng Kim
21127090

CONTENTS

I. SELF-ASSESSMENT FORM

Features		Level of completion
Source	Implement code	100%
Requests	Request 1	100%
	Request 2	100%
	Request 3	100%
	Request 4	100%
	Request 5	100%
Research questions	Question 1	100%
	Question 2	100%
	Question 3	100%

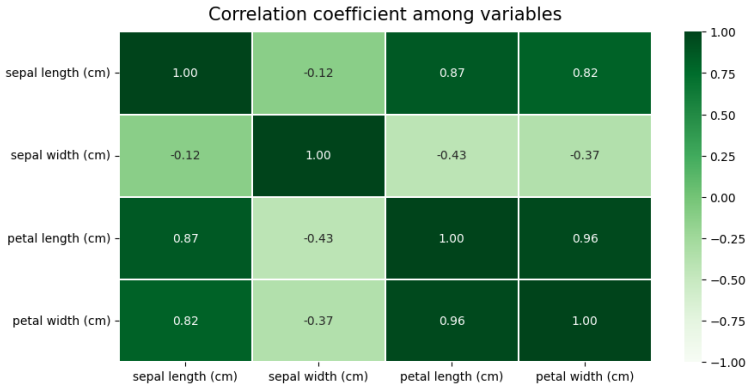
II. REQUESTS

1. Request 1

Students explain differences of correlation between sepal width / sepal length with petal related value

a. Correlation coefficient

- The correlation coefficient is a statistical measure of the strength and the direction of the relationship between two variables.
- The values range from -1.0 to 1.0. A correlation of -1.0 indicates a strong negative correlation, while a correlation of 1.0 indicates a strong positive correlation. A correlation of 0.0 indicates no linear relationship between the movements of the two variables.
- A positive correlation indicates that both variables change in the same directions, in inverse, a negative one indicates that the variables change in opposite directions. The zero correlation indicates that there is no relationship between the variables.
- In this scope, we use a threshold of 0.5 as a standard. Correlation coefficients with values greater than or equal to 0.5 are considered strong correlations, while those below this threshold are typically seen as weak correlations. The sign of the correlation coefficient indicates the direction of the relationship (negative for opposite directions and positive for same directions).
- With the Iris dataset, we can represent the correlation between the attributes of Sepal and Petal to examine the strength of the relationship among them. Below is a heatmap used for visualization:



b. Differences of correlation between sepal width / sepal length with petal related value

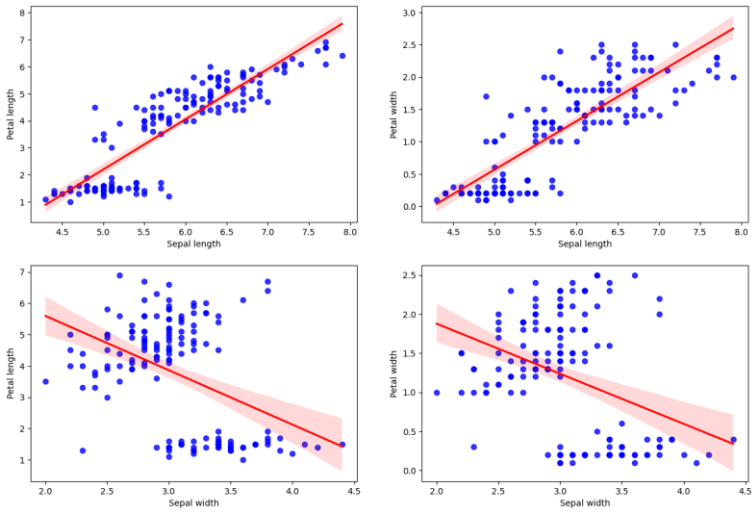
Sepal width

- For petal length, the correlation coefficient between them is -0.43. This indicates that they change in different directions with a correlation of 0.43. With this value, it shows that the correlation between these two variables is not very strong, meaning that if one variable changes, it doesn't affect the other variable too much.
- For petal width, the correlation coefficient between them is -0.37. Similar to petal length, this is also an inverse relationship with a correlation coefficient of 0.37. With this value, the correlation between petal width and sepal width is assessed to be low, meaning that the change in one variable will not have a significant impact on the other variable.

Sepal length

- For petal length, the correlation coefficient between them is 0.87. This indicates that they change in the same direction with a correlation of 0.87. With this value, it shows that the correlation between petal length and sepal length is very strong, it is meaningful and needs to be considered. This means that the change in petal length significantly affects the value of sepal length (either increasing or decreasing together).
- For petal width, the correlation coefficient between them is 0.82. Similar to petal length, this is a relationship in the same direction with a correlation of 0.82. With this value, it indicates that the correlation between petal width and sepal length is very strong, meaningful, and needs to be considered. This implies that changes in petal width significantly influence the value of sepal length (either increasing or decreasing together).

From the analysis above, it can be observed that the correlation between sepal width and the related values of petal is not strong, and these values are inversely related, while sepal length has a strong correlation and is positively correlated with the related values of petal. This indicates that changes in the values of petal significantly affect the value of sepal length (either increasing or decreasing together), contrary to sepal width (an increase in one leads to a decrease in the other).



2. Request 2

Students explain the important of scaling data.

- Scaling data is the process of adjusting the scale of variables or standardlize data within a dataset. In this case, we use the StandardScaler function from the scikit-learn library to scale the data. This is done by subtracting the mean of the dataset from the value of each data point and then dividing by the standard deviation, as below:

$$X' = \frac{X - \mu}{\sigma}$$

  Where
  - X' is the value after scaling
  - μ is the mean
  - σ is the standard deviation
- After scaling the data, the final result will have an expected value (mean) close to 0 and a standard deviation equal to 1.
- Scaling data helps balance and standardize data within a dataset. This normalization is particularly important for datasets with features that are not measured on the same scale.
  - Consider an example with a dataset measuring a person's health, with features such as eye prescription and height (with values like 0 for eye prescription and 150cm for height). Although they are in the same dataset, their measurements are different. If we increase the value of both features by 5 units, the eye prescription would be 5 degrees, while the height would only be 155cm. This indicates that even though the same amount of value change is applied, the impact differs between eye prescription and height due to their specific measurement units. This could affect data analysis or other tasks. Therefore, scaling data is necessary to ensure uniformity among features in a dataset.
- Scaling data also helps to reduce the range of variables with large values. This is crucial because variables with larger scales might overshadow the impact of variables with smaller scales. By scaling the data, we ensure that all variables contribute equally.

3. Request 3

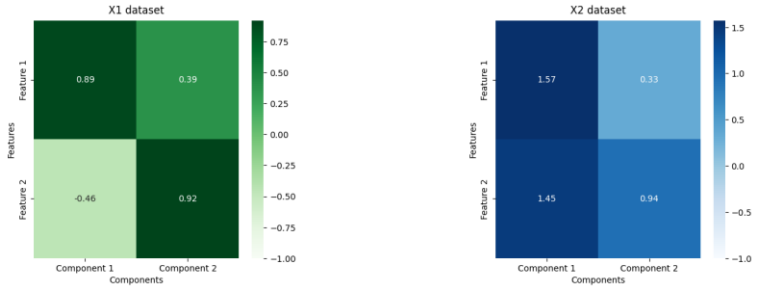
Compare the first canonical variates pair and the second canonical variates pair. Which one should be analyze?

- Pair of canonical variables is a pair of new variables generated from the results of Canonical Correlation Analysis (CCA). In each pair of canonical variables, each variable from the first dataset corresponds to one variable from the second dataset, and both variables are created to have the highest correlation with each other.
- When comparing the values of the first pair of canonical variables (0.94) and the second pair of canonical variables (0.12), we can make the following observations:
    - The first pair of canonical variables (0.94). This value is close to 1, indicating a strong relationship between the corresponding canonical variables from the two datasets. This demonstrates that there is a correlation, and this is very high, indicating a strong relationship between the canonical variables in this dataset.
    - The second pair of canonical variables (0.12): It has a low correlation coefficient (close to 0), indicating a weaker relationship between the corresponding canonical variables from the two datasets. Although there is a correlation, the degree of correlation is lower than that of the first pair of canonical variables, indicating that the relationship between the canonical variables in the second dataset is less strong than in the first dataset.In the loading table represented in the source code, with a threshold value of 0.3 for high loadings, it can be observed that all factors have coefficients greater than this value.
- The main goal of CCA is to find corresponding canonical variables from two datasets such that the correlation between them is maximized. Therefore, we should choose the first pair of canonical variables with a high correlation coefficient (in this case, 0.94). This pair of variables could provide important and reliable information about the relationship between variables from the two datasets.

4. Request 4

Student draw conclusions based on the loadings table.

- Canonical loadings are coefficients used to describe the relationship between the original variables and the canonical variables in Canonical Correlation Analysis (CCA). They are commonly used to represent the relationship between the original variables and the canonical variables. High values of canonical loadings indicate a strong influence of an original variable on the corresponding canonical variable, while values close to 0 indicate a weaker relationship.
- Based on the loadings table, we can visualize it as a heatmap to clearly see the influence of features on the canonical variables in the two datasets (X1 and X2):

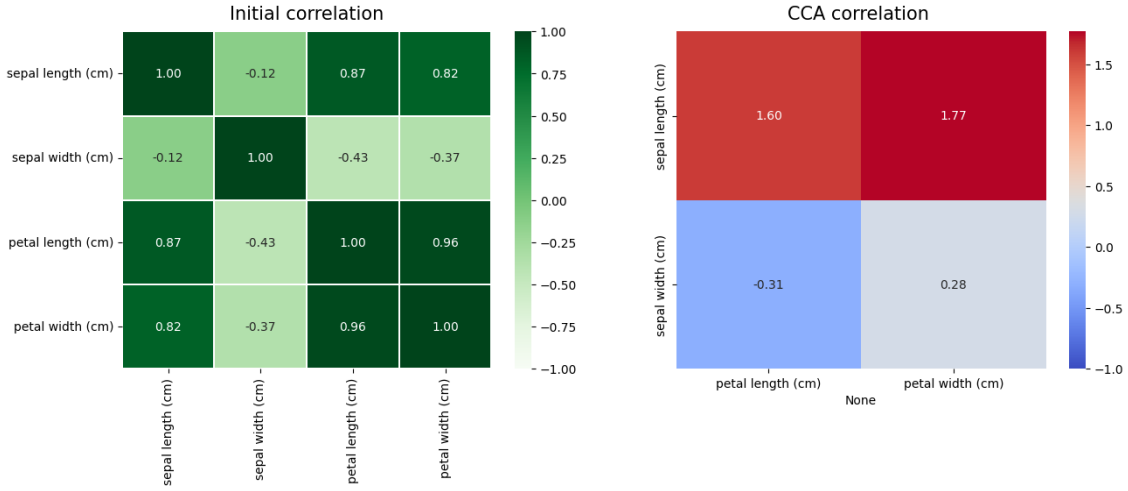


- In dataset X1, the first feature has a strong and positive influence on the first canonical variate, with a value of 0.89. Meanwhile, the second feature has a strong positive influence on the second canonical variate, with a value of 0.92.
- In dataset X2, both features have a strong and positive influence on the first canonical variate, with values of 1.57 and 1.45 respectively. Meanwhile, for the second canonical variate, only the second feature has a strong impact on it, with a value of 0.94.

5. Request 5

Compare the heatmap at step 2 with this CCA coefficients

- We have visualized two heatmaps side by side to get an overview of the differences between them. The chart on the right represents the heatmap shown in step 2 (temperately called 1) , and the other chart represents the heatmap shown in step 6 (temperately called 2).



- If (1) is used to represent the correlation between variable pairs in the Iris dataset, then (2) is used to measure the degree of correlation between canonical variables from two datasets in CCA.
- It can be seen that the original data dimension of (1) from 4×4 has been reduced to 2×2 in (2), which makes it simpler to extract general information from the dataset compared to (1). Specifically:
    - Based on (2), it can quickly be inferred that sepal length has a strong and positive correlation with both petal length and petal width. Meanwhile, sepal width has a weak negative correlation with petal length and an insignificant correlation with petal width.
- The goal of Canonical Correlation Analysis (CCA) is to find canonical variables from the original datasets so that the correlation between them is maximized. Therefore, the values represented on the heatmap in (6) may differ from (1), but it still provides general and important information about the relationship between variables in the two datasets.

III. REASEARCHING QUESTIONS

1. What is Canonical Correlation Analysis, and how does it differ from traditional correlation analysis?

- In statistics, canonical-correlation analysis (CCA), also called canonical variates analysis, is a way of inferring information from cross-covariance matrices [1].
- According to [2], Canonical correlation analysis seeks to identify and quantify the associations between two sets of variables. CCA focuses on the correlation between a linar combination of the variables in one set and a linear combination of the variables in another set.
    - The idea is first to determine the pair of linear combinations having the largets correlation. Next, we determine the pair of linear combinations having the largets correlation among all pairs uncorrelated with the initially selected pair and so on.
    - The pairs on linear combinations are called the canonical variables, and their correlations are called canonical correlations.

- The canonical correlations measure the strength of association between the two set of variables. The maximization aspect the technique represents an attempt to concentrate a high-dimensional relationship between two sets of variables into few pairs of canonical variables.

- The table below shows the difference between the CCA and the traditional correlation analysis that we can summarize:

Features	CCA	Traditional correlation analysis
Objective	Find canonical variables to optimize the correlation between two different datasets.	Measure the degree of correlation between variables within the same dataset.
Calculation method	Uses canonical correlation analysis.	Uses Pearson or Spearman correlation coefficients.
Ability to consider relationships	Has the ability to consider complex or nonlinear relationships between variables.	Typically only considers linear relationships between variables.
Scope of application	Often applied when wanting to understand relationships between variables from two different datasets.	Typically used to measure the degree of correlation between variables within the same dataset or between variables from the same research field.

2. Explain the concept of canonical variables and their significance in CCA

- Canonical variables are the linear combinations of the original variables in each dataset that maximize the correlation between the two sets. Canonical variables are what CCA aims to find [3].
- To assess the importance of canonical variables in CCA, we can rely on several aspects, as follows:
    - **Maximizing correlation:** Canonical variables are constructed to maximize the correlation between pairs of variables from different datasets. This allows CCA to identify the strongest relationships between the two sets of variables.
    - **Reducing dimensionality:** CCA can be used to reduce the dimensionality of the data by transforming the original variables into a smaller number of canonical variables. These canonical variables capture the most important information from the original datasets while minimizing redundancy.
    - **Interpretability:** Canonical variables provide a way to interpret the relationships between variables from different datasets. By examining the weights (coefficients) associated with each original variable in the canonical variables, we can understand how variables from one dataset relate to variables from another dataset.

3. Do datasets required to have the same dimensionality for CCA

- Based on a comprehensive understanding of CCA principles and its application in the literature referenced from [1], the datasets do not need to be of the same dimension for Canonical Correlation Analysis (CCA).
- CCA is designed to find linear combinations of variables from each data set that maximize the correlation between these combinations, regardless of the number of variables in each data set. This means that CCA can effectively identify shared information or underlying relationships between variables from datasets with different dimensions.
- When applying CCA to data sets with many different dimensions, the algorithm automatically adjusts to ensure that the resulting standard variables capture the most prominent information from each data set while maximizing the correlation. correlation between them. This flexibility makes CCA a versatile

tool for analyzing and integrating data from a variety of sources, even when the data sets have different numbers of variables or different levels of complexity.

IV. REFFERENCES

[1] Wikipedia, "Canonical correlation," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Canonical_correlation#cite_note-1.

[2] R. A. JOHNSON and D. W. WICHERN, "Canonical correlation analysis," in *Applied Multivariate Statistical Analysis*, 2007.

[3] GeeksForGeeks, "Canonical Correlation Analysis (CCA) using Sklearn," 09 December 2023. [Online]. Available: https://www.geeksforgeeks.org/canonical-correlation-analysis-cca-using-sklearn/.