Assignment 02
Canonical Correlation Analysis of a real-world Problem (S17403)

1. Introduction

Canonical correlation analysis (CCA) is a way of measuring the linear relationship between two multidimensional variables. It finds two bases, one for each variable, that are optimal with respect to correlations and, at the same time, it finds the corresponding correlations. In other words, it finds the two bases in which the correlation matrix between the variables is diagonal and the correlations on the diagonal are maximized. The dimensionality of these new bases is equal to or less than the smallest dimensionality of the two variables. An important property of canonical correlations is that they are invariant with respect to affine transformations of the variables. This is the most important difference between CCA and ordinary correlation analysis which highly depend on the basis in which the variables are described. CCA allows us to summarize the relationship into a lesser number of statistics while preserving the main facts of the relationships. CCA is a dimension-reduction technique.

Objectives

The objectives of a Canonical Correlation Analysis (CCA) project are typically focused on understanding the relationship between two sets of variables and identifying the underlying patterns and associations between them. Here are the main objectives of this project:

- Exploring Associations (determine if there are any significant relationships between two sets of variables)
- Dimension Reduction
- Finding a common structure (discover the common underlying structure between the two sets of variables)
- Identifying key variables (identify which variables from each set contribute the most to the canonical correlations)

2. Methodology

Dataset Description: - Algerian Forest Fires Dataset

The dataset includes 244 instances that regroup a data of two regions of Algeria, namely the Bejaia region located in the northeast of Algeria and the Sidi Bel-abbes region located in the northwest of Algeria. 122 instances for each region. The period from June 2012 to September 2012. The dataset includes 11 attributes and 1 output attribute (class) The 244 instances have been classified into fire (138 classes) and not fire (106 classes) classes.

Dataset Link: - https://archive.ics.uci.edu/dataset/547/algerian+forest+fires+dataset

Attribute information

1. Date : (DD/MM/YYYY) Day, month ('June to 'september'), year (2012)

Weather data observations

2. Temp : temperature noon (temperature max) in Celsius degrees: 22 to 42
3. RH : Relative Humidity in %: 21 to 90
4. Ws :Wind speed in km/h: 6 to 29
5. Rain: total day in mm: 0 to 16.8 FWI Components

FWI_compo data observations

6. Fine Fuel Moisture Code (FFMC) index from the FWI system: 28.6 to 92.5
7. Duff Moisture Code (DMC) index from the FWI system: 1.1 to 65.9
8. Drought Code (DC) index from the FWI system: 7 to 220.4
9. Initial Spread Index (ISI) index from the FWI system: 0 to 18.5
10. Buildup Index (BUI) index from the FWI system: 1.1 to 68
11. Fire Weather Index (FWI) Index: 0 to 31.1
12. Classes: two classes, namely fire and not fire

Problem Statement:

The objective of this project is to investigate the relationship and patterns between weather variables and Fire Weather Index (FWI) components in the context of forest fires.

Key Steps:

- Data Preparation: Clean and preprocess data to ensure it is suitable for apply Canonical Correlation Analysis. This involves handling missing values, checking for outliers, and appropriately transforming variables if necessary and get the idea about each variables.
- Partitioning the data: Split the dataset into two sets, X and Y, where X contains the variables for the first set and Y contains the variables for the second set. Here, two sets are Weather and FWI_compo
- Standardization (optional): If the variables within X and Y have different scales, it is recommended to standardize them (subtract mean and divide by standard deviation) to give each variable equal importance in the analysis.
- Perform Canonical Correlation Analysis: Set up the CCA model to explore the relationships between Set X and Set Y. The CCA model aims to find linear combinations of the variables in Set X and Set Y that are maximally correlated.
- Select the Canonical Correlations: They indicate how well the two sets of variables are related to each other.
- Compute the Canonical Variates: The canonical variates are the linear combinations of the original variables that maximize the correlation between the two sets. The canonical variates for X are computed as X_c = X * U, and for Y, Y_c = Y * V. (U will correspond to the linear combinations from the first set of variables X, and V will correspond to the linear combinations from the second set of variables, Y)

- Interpretation and Analysis: Examine the canonical correlations and canonical loadings to understand the strength and direction of the relationships between the two sets of variables. The canonical loadings provide insights into which variables contribute most to each canonical variate.
- Evaluate the Significance (optional): Conduct statistical tests (e.g., Wilks' Lambda) to determine the significance of the canonical correlations.

Statistical methods used in CCA:

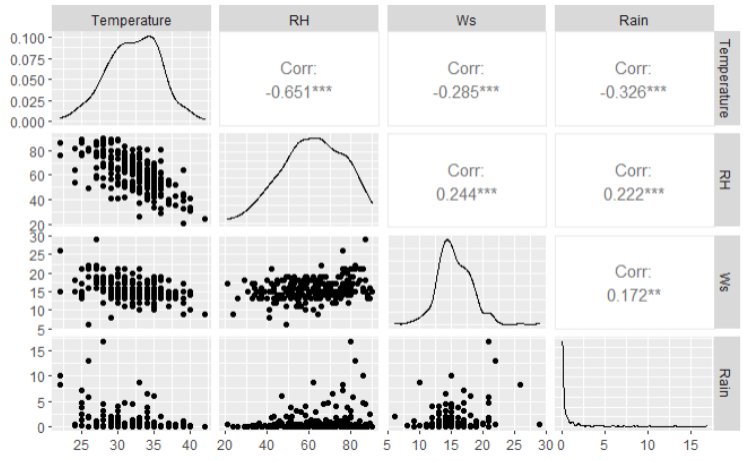
- **Canonical Correlation Analysis**: The primary statistical method used is CCA itself, which involves finding the canonical correlations and the canonical variates.
- **Multivariate Hypothesis Testing**: Optional statistical tests can be performed to assess the significance of the canonical correlations.
- **Dimension Reduction Techniques:** CCA can be considered a dimension reduction technique as it reduces the dimensionality of the two sets of variables to a smaller number of canonical variates.

3. Results and discussion

Correlations

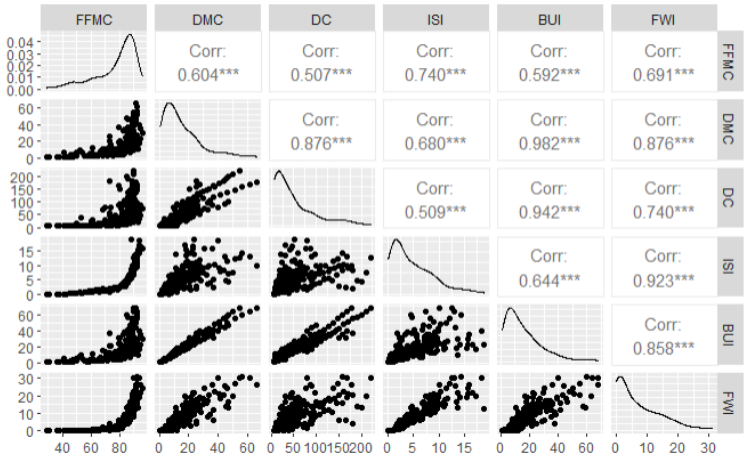
After importing the data, the first step is the EDA and data cleaning after that the next step is to spilt the x variables (Weather) and the y variables (FWI_compo)

```
{r}
weather <- new_data[, c("Temperature", "RH","Ws","Rain")]
FWI_compo <- new_data[, c("FFMC", "DMC","DC","ISI","BUI","FWI")]
```



in weather multivariate set:-

RH and temperature variables show a highest negative correlation between variables.



in FWI_compo multivariate set :-

the highest positive correlation (0.942) shows between BUI and DC variables. BUI and FWI show a 0.858 positive correlation. In this variable set, all the variables show positive correlation values.

Model Fitting

We can then fit the canonical correlation model using the cc function which is built-in in the R CCA library.

```{r}
#Canonical correlation
cc1 <- cc(Weather, FWI_compo)

# display the canonical correlations
cc1$cor

# raw canonical coefficients
cc1[3:4]
```

```
[1] 0.8399173 0.3984355 0.3112653 0.1344204
$xcoef
                [,1]        [,2]        [,3]        [,4]
Temperature -0.10315053  0.12792789 -0.13175043 -0.31483591
RH           0.03914075  0.06353169 -0.02597180 -0.04134852
WS          -0.05990063 -0.11178467 -0.34726856  0.05402229
Rain         0.18212156 -0.32917360  0.06639048 -0.36760036

$ycoef
            [,1]        [,2]        [,3]        [,4]
FFMC -0.046469484  0.07486498  0.05677798  0.02356368
DMC  -0.060439405 -0.04485567  0.22563936 -0.12473490
DC   -0.006339360  0.01314634 -0.01848025  0.02735178
ISI  -0.115751724 -0.29407436 -0.25209715  0.12069574
BUI   0.078758418  0.02663832 -0.18309962 -0.07538721
FWI   0.002871459 -0.01597301  0.06038108  0.03522416
```

The raw canonical coefficients are interpreted in a manner analogous to interpreting regression coefficients i.e., for the variable FFMC, a one-unit increase in FFMC leads to a 0.0464 decrease in the first canonical variate of set 2 when all of the other variables are held constant. FWI leads to a 0.002871 increase in dimension 1 for the FWI_Compo set with the other predictors held constant. Next, we will use compute to compute the loadings of the variables on the canonical dimensions (variates). These loadings are correlations between variables and the canonical variates.

The first canonical correlation has a high correlation (0.8399173), therefore, enough evidence to say the first canonical variate pair is highly correlated.

```r
# compute canonical loadings
cc2 <- comput(weather,FWI_compo,cc1)

# display canonical loadings
cc2[3:6]
```

```
$corr.X.xscores
                [,1]       [,2]         [,3]         [,4]
Temperature -0.8234979  0.1551730  0.007214859 -0.545637621
RH           0.8641844  0.4164061 -0.282441272  0.004258191
Ws           0.1422818 -0.3295025 -0.911488404  0.200930036
Rain         0.5871815 -0.6553635  0.035980706 -0.473731929

$corr.Y.xscores
           [,1]        [,2]        [,3]         [,4]
FFMC -0.7978544  0.11749991  0.01521896 -0.002188395
DMC  -0.5240900  0.03121606 -0.11253096 -0.091760375
DC   -0.3945849  0.13249945 -0.20935845 -0.059185475
ISI  -0.7527288 -0.14015675 -0.07874522 -0.011576566
BUI  -0.4921211  0.06790534 -0.15408038 -0.082661709
FWI  -0.6730546 -0.08053571 -0.12187813 -0.047260416

$corr.X.yscores
                [,1]        [,2]         [,3]          [,4]
Temperature -0.6916702  0.06182643  0.002245735 -0.0733448208
RH           0.7258435  0.16591097 -0.087914171  0.0005723876
Ws           0.1195049 -0.13128550 -0.283714724  0.0270090934
Rain         0.4931839 -0.26112009  0.011199546 -0.0636792297

$corr.Y.yscores
           [,1]        [,2]        [,3]        [,4]
FFMC -0.9499202  0.29490320  0.04889385 -0.01628023
DMC  -0.6239781  0.07834659 -0.36152746 -0.68263733
DC   -0.4697902  0.33254931 -0.67260449 -0.44030133
ISI  -0.8961940 -0.35176771 -0.25298426 -0.08612210
BUI  -0.5859162  0.17042994 -0.49501300 -0.61494919
FWI  -0.8013344 -0.20212986 -0.39155704 -0.35158667
```

These correlations are between observed variables and canonical variables which are known as the canonical loadings. These canonical variates are actually a type of latent variable.

In general, the number of canonical dimensions is equal to the number of variables in the smaller set; however, the number of significant dimensions maybe even smaller. Canonical dimensions, also known as canonical variates, are latent variables that are analogous to factors obtained in factor analysis. For this particular model, there are four canonical dimensions of which all are statistically significant. For the statistical test, we use the R package “CCP”.

Check whether are the canonical correlations significant or not

```r
library(CCP)
# 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(weather)[1]
p <- length(weather)
q <- length(FWI_compo)

## Calculate p-values using the F-approximations of different test
statistics:
p.asym(rho, n, p, q, tstat = "Wilks")
```

```
Wilks' Lambda, using F-approximation (Rao's F):
              stat    approx df1      df2      p.value
1 to 4:  0.2197308 18.451704  24 814.0499 0.000000e+00
2 to 4:  0.7460161  4.825319  15 646.3724 5.978211e-09
3 to 4:  0.8867957  3.637285   8 470.0000 4.019205e-04
4 to 4:  0.9819312  1.447571   3 236.0000 2.296535e-01
```

Wilk lambda =0.2197308, F=18.452, d.f=24 : p<0.0001.Here we reject the null hypothesis that there is no relationship between the two sets of variables and can conclude that two sets of variables are dependent.

same as the second, third, and fourth canonical variate pair is correlated (All the P values are lower than 0.001). Therefore, all four canonical variate pairs are significantly correlated and dependent on one another.

Estimates of Canonical Correlation

```
{r}
# Get the squared canonical correlations
squared_canonical_correlations <- cc1$cor^2

# Print the squared canonical correlations
print(squared_canonical_correlations)
```

```
[1] 0.70546105 0.15875085 0.09688610 0.01806884
```

70.54% of the variation in U1 is explained by the variation in V1 ,15.87% of the variation in U2 is explained by the variation in V2 and 9.68% of the variation in U3 is explained by the variation in V3 but only 1.80% of the variation in U4 is explained by V4.

This first one is a very high canonical correlation and implies that only the first one canonical correlation is important.

Canonical coefficients

	[,1]	[,2]	[,3]	[,4]
Temperature	-0.10315053	0.12792789	-0.13175043	-0.31483591
RH	0.03914075	0.06353169	-0.02597180	-0.04134852
Ws	-0.05990063	-0.11178467	-0.34726856	0.05402229
Rain	0.18212156	-0.32917360	0.06639048	-0.36760036

These are the estimated canonical coefficients a_{ij} for the Weather variables.

The first canonical variable for weather:

U_1= -0.10315 (Temperature) + 0.03914 (RH) – 0.0599 (Ws) +0.18212 (Rain)

	[,1]	[,2]	[,3]	[,4]
FFMC	-0.046469484	0.07486498	0.05677798	0.02356368
DMC	-0.060439405	-0.04485567	0.22563936	-0.12473490
DC	-0.006339360	0.01314634	-0.01848025	0.02735178
ISI	-0.115751724	-0.29407436	-0.25209715	0.12069574
BUI	0.078758418	0.02663832	-0.18309962	-0.07538721
FWI	0.002871459	-0.01597301	0.06038108	0.03522416

These are the estimated canonical coefficients b_{ij} for the FWI_compo variables.

The first canonical variable for FWI_compo:

V_1= -0.04646 (FFMC) - 0.0604 (DMC) – 0.00634 (DC) - 0.11575 (ISI) + 0.07876 (BUI) + 0.00287 (FWI)

The magnitude of the coefficients gives the contributions of the individual variables to the corresponding canonical variables. These magnitudes also depend on the variances of the corresponding variables.

4. Conclusion and Recommendation

- The canonical correlation analysis revealed a significant relationship between Set 1 Weather variables (e.g., X1, X2, X3, X4) and Set 2 FWI_compo variables (e.g., Y1, Y2, Y3, Y4, Y5, Y6). The analysis identified four canonical covariate pairs with squared canonical correlations of 0.8399, 0.3984, 0.3112, and 0.1344, indicating substantial shared variance between the two sets of variables. These findings suggest a strong association between the variables in Set 1 and Set 2.

- The first canonical covariate pair (X1, Y1) exhibited the highest squared canonical correlation 83.99% (0.8399), indicating a strong relationship between X1 and Y1. This suggests that changes in X1 are associated with corresponding changes in Y1, implying a significant linkage between these variables. In the context of our study, this finding may indicate that X1 could be a critical predictor for Y1's behavior, and further investigation into their causal relationship is warranted.

- Based on the significant relationship between weather and FWI_compo variables, we recommend that forest fire management agencies closely monitor weather conditions, especially during periods of elevated fire danger. Implementing an early warning system that integrates real-time weather data with fire weather indices would enable timely responses to potential fire outbreaks

- Limitations: Although the canonical correlation analysis provided valuable insights into the relationship between weather and fire weather indices, our study has some limitations. The dataset might not capture all relevant weather and fire behavior variables, and additional factors such as land cover and fire suppression efforts could influence the results.

5. References

RPuBS - CANONICAL CORRELATION ANALYSIS IN R. (n.d.). https://rpubs.com/Devy/902673

Lesson 13: Canonical Correlation Analysis | STAT 505. (n.d.). PennState: Statistics Online Courses. https://online.stat.psu.edu/stat505/lesson/13

Korstanje, J. (2022, January 6). Canonical Correlation Analysis | towards Data Science. *Medium*. https://towardsdatascience.com/canonical-correlation-analysis-b1a38847219d

6. Glossary

Canonical correlation: Correlation between two canonical variates of the same pair. This is the criterion optimized by CCA

Canonical loadings: Correlation between the original variables and the canonical variates. Sometimes used as a synonym for canonical vectors (because these quantities differ only by their normalization)

Canonical variates: The latent variables (one per data table) are computed in CCA (also called canonical variables, canonical variable scores, or canonical factor scores). The canonical variates have a maximal correlation

Canonical vectors: The set of coefficients of the linear combinations used to compute the canonical variates, also called canonical weights. Canonical vectors are also sometimes called canonical loadings

Latent variable: A linear combination of the variables of one data table. In general, a latent variable is computed to satisfy some predefined criterion

7. Appendices

Data Set

	day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes
1	01	06	2012	29	57	18	0.0	65.7	3.4	7.6	1.3	3.4	0.5	not fire
2	02	06	2012	29	61	13	1.3	64.4	4.1	7.6	1.0	3.9	0.4	not fire
3	03	06	2012	26	82	22	13.1	47.1	2.5	7.1	0.3	2.7	0.1	not fire
4	04	06	2012	25	89	13	2.5	28.6	1.3	6.9	0.0	1.7	0.0	not fire
5	05	06	2012	27	77	16	0.0	64.8	3.0	14.2	1.2	3.9	0.5	not fire
6	06	06	2012	31	67	14	0.0	82.6	5.8	22.2	3.1	7.0	2.5	fire

R markdown PDF