Canonical Correlation Analysis (CCA)

ST405 MULTIVARIATE METHODS II
ASSIGNMENT II

S/18840
Dinuka Weerasooriya

1 Introduction

In multivariate statistics, Canonical Correlation Analysis (CCA) technique is used to identify and measure the associations between two sets of variables. Unlike simple correlation which deals with the relationship between two variables, CCA extends this concept to explore relationships between multiple interrelated quantitative variables, thus providing deeper understanding about the data set.

In this study, I'll analyze mentioned dataset using canonical correlation analysis. Let's explore how it applies to our selected data.

1.1 Purpose of the study

The purpose of canonical correlation analysis is dimension reduction. I am going to try to determine how many numbers of canonical variate pairs should we consider and the values of canonical variate scores, and relationships between variables in the first set and those in the second set.

1.2 Hypothesis

In this study, we need to test the hypothesis that at least the first canonical correlation is statistically significant. We will repeat this process until we find a non-significant canonical correlation. Additionally, we need to conduct a hypothesis test to determine the relationship between the first and second data sets.

1.3 Importance of the study

In lots of studies, we have large number of individual variables, and handling big datasets can be tough. That's why finding a efficient way to reduce the data set's dimensions is important. Canonical correlation analysis helps with issue, making it easier to make predictions and conclusions.

2 Methodology

In this study, we applied Canonical Correlation Analysis (CCA) to a dataset from the World Health Organization (WHO) and the United Nations, which includes health and economic indicators for 193 countries from the years 2000 to 2015. The dataset consists of 22 variables, capturing various aspects of health and economic status.

Here is the description of each variables,

Variable name	Description
"Country"	Country
"Year"	Year
"Status"	Developed or Developing status
"Life expectancy"	Life Expectancy in age
"Adult Mortality"	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
"infant deaths"	Number of Infant Deaths per 1000 population
"Alcohol"	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
"percentage expenditure"	Expenditure on health as a percentage of Gross Domestic Product per capita(%)
"Hepatitis B"	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
"Measles"	Measles - number of reported cases per 1000 population
"BMI"	Average Body Mass Index of entire population
"under-five deaths"	Number of under-five deaths per 1000 population
"Polio"	Polio (Pol3) immunization coverage among 1-year-olds (%)
"Total expenditure"	General government expenditure on health as a percentage of total government expenditure (%)
"Diphtheria"	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
"HIV/AIDS"	Deaths per 1 000 live births HIV/AIDS (0-4 years)
"GDP"	Gross Domestic Product per capita (in USD)
"Population"	Population of the country
"thinness 1-19 years"	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
"thinness 5-9 years"	Prevalence of thinness among children for Age 5 to 9(%)
"Income composition of resources"	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
"Schooling"	Number of years of Schooling(years)

For the canonical correlation analysis purpose, I selected only two sets of variables. From here I called my first variable set as “Economic and social” variables and the second variable set as “Health related” variables.

Since there are large number of variables in the data set and to make it easier for the analysis I have chosen only 3 most important variables for each set. So I have included the variables of each set as follows

Economic and social	Health related
GDP	Life expectancy
Population	Adult Mortality
percentage expenditure	Infant deaths

Since different variables have different measurement units, I standardized the whole data set

In this study I’m going to apply statistical methods to perform the canonical correlation analysis on above mentioned two variable sets using R software. For the convenience, the variables in the first set are referred as “eco_social” variables and the variables in the second set are called “health” variables in the R

3 Results and Discussions

3.1 Pairwise Scatter plots

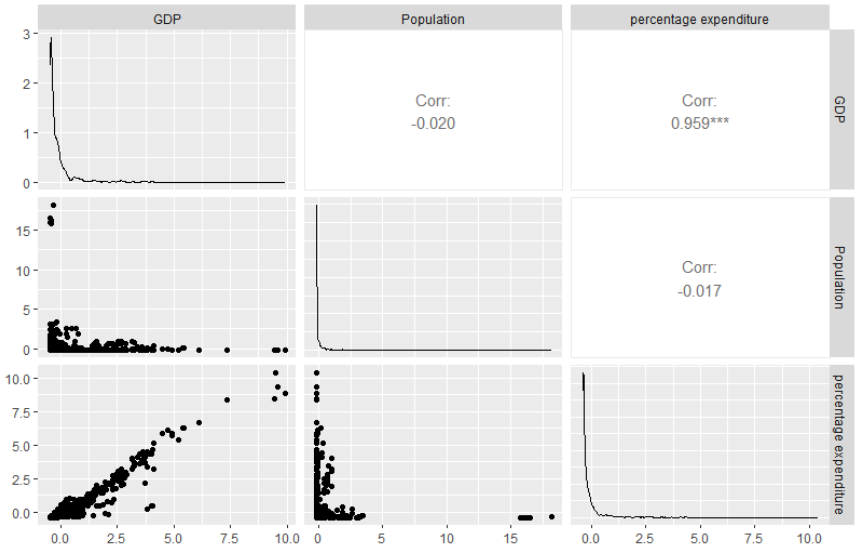


Figure 1 Pairwise scatter plot for the eco_social variables

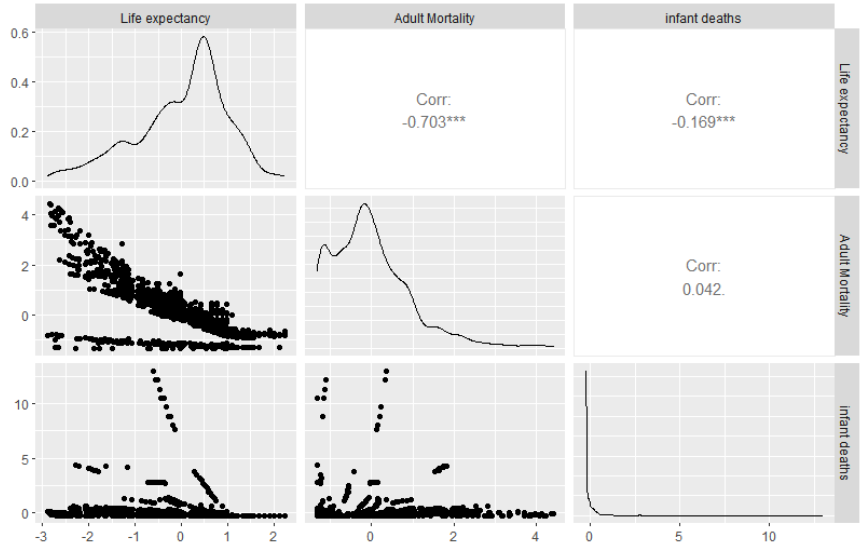


Figure 2 Pairwise Scatter plot for the health variables

3.2Correlation Matrix

	GDP	Population	% Expenditure	Life Expectancy	Adult Mortality	Infant Deaths
GDP	1.0000	−0.0204	0.9593	0.4413	−0.2550	−0.0981
Population	−0.0204	1.0000	−0.0168	−0.0223	−0.0150	0.6718
% Expenditure	0.9593	−0.0168	1.0000	0.4096	−0.2376	−0.0908
Life Expectancy	0.4413	−0.0223	0.4096	1.0000	−0.7025	−0.1691
Adult Mortality	−0.2550	−0.0150	−0.2376	−0.7025	1.0000	0.0425
Infant Deaths	−0.0981	0.6718	−0.0908	−0.1691	0.0425	1.0000

Figure 3 correlation matrix of six variables chosen to CCA

3.3 Canonical correlations and squared canonical correlations

Canonical correlation	Squared canonical correlation
$\rho_1^* = 0.679386701$	4.615663×10^{-01}
$\rho_2^* = 0.449462031$	2.020161×10^{-01}
$\rho_3^* = 0.004294201$	1.844016×10^{-05}

3.4 Estimated canonical coefficients

3.4.1. Estimated canonical coefficients(aij) for the Economic and social variables

	U1	U2	U3
GDP	0.0538882166	1.37110284	-3.264948571
Population	-0.9973990118	00.07552349	-0.001117374
Expenditure percentage	0.0007388668	-0.39464660	3.519275778

3.4.2. Estimated canonical coefficients(bij) for the health variables

	V1	V2	V3
Life expectancy	-0.14249491	1.16951292	-0.81558187
Adult Mortality	-0.05543304	0.24630249	-1.39083005
infant deaths	-1.01581814	0.08061585	-0.05812126

3.5. The correlations between each variable and the canonical variables

3.5.1. The correlations between the square Economic and social variables and the canonical variables of itself

	1	2	3
GDP	0.07491300	0.9909805	0.11111142
Population	-0.99850907	0.0542225	0.00629007
percentage expenditure	0.06918224	0.9193826	0.38723311

3.5.2. *The correlations between the health variables and the canonical variables of itself*

	1	2	3
Life expectancy	0.068196317	0.9828497	0.1713351
Adult Mortality	0.001551198	-0.5718851	-0.8203322
infant deaths	-0.994079131	-0.1066625	0.0207312

3.6. Testing the relationship between the canonical variate pairs

In here we wish to test for the independence between the two sets of variables.

Hypothesis;

H_{01} : First set of variables is independent from the second set of variables

vs

H_{11} : First set of variables is dependent from the second set of variables

For this purpose, we can use likelihood ratio test. It is carried out using Wilk's lambda.

	Likelihood	Approximate F value	Num df	P-value
1	0.4296535	184.37638404	9	0.0000000
2	0.7979692	98.19265377	4	0.0000000
3	0.9999816	0.03033463	1	0.8617546

1. First Canonical Correlation:

- Wilk's lambda= 0.4297; F= 184.376; p-value = 0.0000
- Therefore, we reject the null hypothesis that there is no relationship between the two sets of variables.
- We can conclude that the two sets of variables are dependent.
- The above hypothesis is also equivalent to testing all the canonical variate pairs are uncorrelated. (Hypothesis; $H_{02} : \rho_1^* = \rho_2^* = \rho_3^* = 0$ vs. H_{12} : At least $\rho_1^* \neq 0$)
- Therefore, by considering this hypothesis, we can also say that at least the first canonical correlation is significant.

- Next we have to check whether the second and third canonical variate pairs are correlated ornot. (Hypothesis; H_{03} : $\rho_2^* = \rho_3^* = 0$ vs. H_{13} : At least $\rho_2^* \neq 0$).

2. Second Canonical Correlation:

- Wilk's lambda = 0.7980; F= 98.193; p-value = 0.0000
- We reject H_{03} and this implies that at least the second canonical correlation is significant.

- Finally, we have to check whether the third canonical variate pair is correlated or not.(Hypothesis; H_{04} : $\rho_3^* = 0$ vs. H_{14} : $\rho_3^* \neq 0$).

3. Third Canonical Correlation:

- Wilk's lambda= 0.99998; F = 0.0303; p-value = 0.8618
- We fail to reject H_{04} at the 5% significance level, implying that the third canonical correlation is not significant.

- We can test above each and every hypothesis using Pillai's test, Lawley-Hotelling test andRoy's largest root test. Those are also provided similar results.

4 Conclusions and recommendations

•By considering squared values of the canonical variate pairs we can conclude that 46.15% of the total variation in U1 is explained by the variation in V1, and 20.20% of the variation in U2 is explained by V2. But only 0.000018% of the variation in U3 is explained by V3.

•The first one is a moderate high canonical correlation and implies that first canonical correlation is very important rather than the others.

•Using 3.4.1, the first canonical variable for square "Economic and social" is,

U1 = - 0.053882GDP – 0.99736Population +0.0007Expenditure

•Using 3.4.2, the first canonical variable for "Health" is,

V1 = -0.14249491Life_expectancy– 0.05543Adult_mortality – 1.0158infant_deaths

• The magnitudes of the above coefficients give the contributions of the individual variables to the corresponding canonical variable.

• By considering the correlations between "**Economic and social**" variables and its canonical variables, for the first canonical variable, "**Population**" variable has a large negative correlation (-0.9985). Other two variables have moderately low positve correlations.Therefore we can conclude that there is a very strong inverse relationship between Population and first canonical variable

• Second and third canonical variables for "**Economic and social**", all correlations have a positive values. **GPD** and **Percentage expenditure** is Strongly associated with second canonical variable while third canonical variable does not show any strong associations.

• By considering the correlations between "**Health**" variables and its canonical variables, for the first canonical variable, **Infant Deaths** has a very high negative correlation (-0.9941) with the first canonical variable, indicating a very strong inverse relationship between Infant Deaths and this canonical variable. While other two variables have a very weak associations

• When you consider Second and third canonical variables for "**Health**", **Adult Mortality** is most strongly (and negatively) associated with the third canonical variable**.Infant Deaths** is most strongly (and negatively) associated with the first canonical variable

•As a summary, we can say that both variable sets are dependent. Not All three canonical correlations are significant but first two are Significant. Canonical correlation analysis technique can be easily applicable to this analysis.

5 References

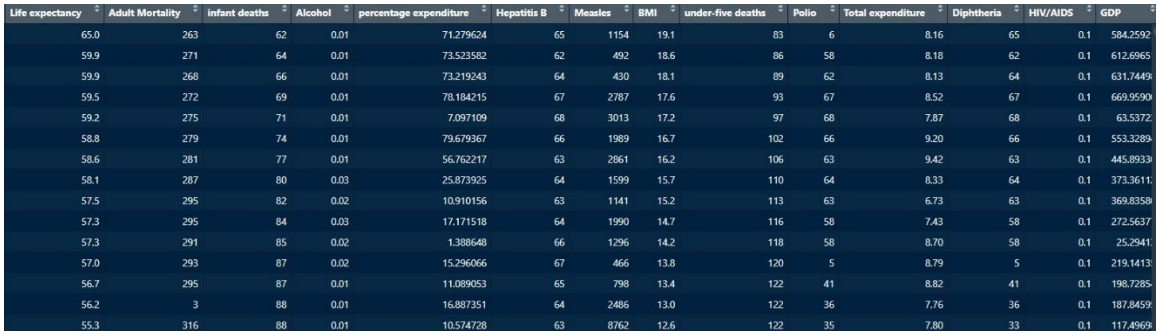
Kumar, A. (n.d.). Life expectancy (WHO). Kaggle. Retrieved May 16, 2024, from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data

González, I., Déjean, S., Martin, P. G. P., & Baccini, A. (2023). CCA: Canonical correlation analysis. R package version 1.2. Retrieved from https://cran.r-project.org/web/packages/CCA/CCA.pdf

Devy. (2022). Canonical Correlation Analysis with R. RPubs. Retrieved May 20, 2024, from https://rpubs.com/Devy/902673

6 Appendices

#head of the data



Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP
65.0	263	62	0.01	71.279624	65	1154	19.1	83	6	8.16	65	0.1	584.2592
59.9	271	64	0.01	73.523582	62	492	18.6	86	58	8.18	62	0.1	612.6965
59.9	268	66	0.01	73.219243	64	430	18.1	89	62	8.13	64	0.1	631.7449
59.5	272	69	0.01	78.184215	67	2787	17.6	93	67	8.52	67	0.1	669.9590
59.2	275	71	0.01	7.097109	68	3013	17.2	97	68	7.87	68	0.1	63.5372
58.8	279	74	0.01	79.679367	66	1989	16.7	102	66	9.20	66	0.1	553.3289
58.6	281	77	0.01	56.762217	63	2861	16.2	106	63	9.42	63	0.1	445.8933
58.1	287	80	0.03	25.873925	64	1599	15.7	110	64	8.33	64	0.1	373.3611
57.5	295	82	0.02	10.910156	63	1141	15.2	113	63	6.73	63	0.1	369.8358
57.3	295	84	0.03	17.171518	64	1990	14.7	116	58	7.43	58	0.1	272.5637
57.3	291	85	0.02	1.388648	66	1296	14.2	118	58	8.70	58	0.1	25.2941
57.0	293	87	0.02	15.296066	67	466	13.8	120	5	8.79	5	0.1	219.1413
56.7	295	87	0.01	11.089053	65	798	13.4	122	41	8.82	41	0.1	198.7285
56.2	3	88	0.01	16.887351	64	2486	13.0	122	36	7.76	36	0.1	187.8459
55.3	316	88	0.01	10.574728	63	8762	12.6	122	35	7.80	33	0.1	117.4969

#Loading Libraries that are needed

```
> library(dplyr)
> library(tidyr)
> library(Amelia)
> library(tidyverse)
> require(GGally)
> require(CCA)
> require(CCP)
```

#importing the data Set

```
> life_exp_data<-read_csv("../data/Life Expectancy Data.csv")
Rows: 2938 Columns: 22── Column specification
```

#Cleaning the dataset

```
> life_exp_data <- life_exp_data %>% select_if(~ mean(is.na(.)) < 0.5)
> life_exp_data<-life_exp_data%>%drop_na()
```

Standardize numeric columns

```
> life_exp_data <- life_exp_data%>% mutate_if(is.numeric, scale)
```

Select the variables for X and Y

```
> eco_social<- as.matrix(life_exp_data[, c("GDP", "Population", "percen
tage expenditure")])
>
> health<- as.matrix(life_exp_data[, c("Life expectancy", "Adult Mortal
ity", "infant deaths")])
```

#Pairwise scatter plots.

```
> ggpairs(eco_social)
> ggpairs(health)
```

#Find the correlations within and between the two sets of variables.

```
> matcor(eco_social,health)
```

#Apply Canonical Correlation Analysis to the two groups.

```
> cc1 <- cc(eco_social,health)
```

#The canonical correlations.

```
> cc1$cor
```

#The squared canonical correlations

```
> (cc1$cor)^2
```

#Raw canonical coefficients for the eco_social variables.

```
> cc1[3]
```

#Row canonical coefficients for the health variables.

```
> cc1[4]
```

#Compute canonical loadings.

```
> cc2 <- comput(eco_social,health, cc1)
```

#Correlations between each variable and the corresponding canonical variate.

```
> 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(X)[1]
> p <- length(eco_social)
> q <- length(health)
```

#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 square feet sizes canonical coefficients diagonal matrix of eco_social standard deviations.

```
> s1 <- diag(sqrt(diag(cov(eco_social))))
> s1 %*% cc1$xcoef
```

#Standardized facilities canonical coefficients diagonal matrix of facilities standard deviations.

```
> s2 <- diag(sqrt(diag(cov(health))))
> s2 %*% cc1$ycoef
```