Canonical Correlation Analysis Report

W.Kavinda Uthsuka

1. Introduction

- Canonical correlation analysis is a method for exploring the relationship between two multivariate sets of variables, all measured on the same individual.
- Canonical correlation analysis allows us to summarize the relationship into lesser number of statistics while preserving the main facets of the relationships. This is another dimension reduction technique. These are the main objectives of Canonical Correlation Analysis.
- Test the hypothesis that canonical variate pairs are correlated or not (canonical correlations are equal or not equal to zero) at 1% significance level.
- I use "Fat Supply Quantity" dataset and it includes percentage of fat intake from different types of food.

2. Methodology

- I use "Fat Supply Quantity Data" dataset which includes percentage of fat intake from different types of food in 170 countries around the world.
  - AnimalProducts: Percentage of fat intake from animal products.
  - AnimalFats: Percentage of fat intake from animal fats.
  - Cereals: Percentage of fat intake from cereals.
  - Eggs: Percentage of fat intake from eggs.
  - FishSeafood: Percentage of fat intake from fish, seafood.
  - Fruits: Percentage of fat intake from fruits.
  - Meat: Percentage of fat intake from meat.
  - Milk: Percentage of fat intake from milk.
  - Oilcrops: Percentage of fat intake from oilcrops.
  - Pulses: Percentages of fat intake from pulses.
  - Spices: Percentage of fat intake from spices.
  - Treenuts: Percentage of fat intake from treenuts.

- ➢ VegetalProducts: Percentage of fat intake from vegetal products.
- ➢ VegetableOils: Percentage of fat intake from vegetable oils.
- ➢ Vegetables: Percentage of fat intake from vegetables.

- Use methods;
  - ➢ Canonical correlation analysis.

3. Results and Discussion

- Split the dataset into two sets. Set 1 is “ani_res” & set 2 is “veg_res”. From this we can conclude that, there is only 6 canonical covariate pairs. (“ani_res” set has six variables and “veg_res” has only nine pairs)

ani_res;

	AnimalProducts	AnimalFats	Eggs	FishSeafood	Meat	Milk
1	0.11795823	0.633908343	-0.416161152	-0.884897933	-0.722102288	9.518471e-01
2	1.41258168	-0.219394369	1.072912801	-0.763915781	-0.161028018	3.796458e+00
3	-0.78451071	-0.985943218	0.408738299	-0.702991853	-1.199585405	8.957491e-01
4	-0.67372328	-0.859462658	-1.242302922	0.611471650	0.328390570	-1.164030e+00
5	0.87565127	0.161264056	-0.880011804	0.731371672	1.034104747	4.657243e-01
6	1.20727631	-0.252733160	0.957687765	-0.740217131	2.094604544	2.227529e-01
7	1.12068067	0.645923692	1.033779770	-0.680267120	0.283327204	1.609355e+00
8	0.42662855	0.141309451	-0.391625363	-0.431701823	0.451259493	4.233731e-01
9	0.89108354	2.650448934	0.405011344	-0.459404355	-0.296646680	1.219774e-02
10	1.42752661	1.113396816	1.655094286	-0.690655570	0.536779177	1.444063e+00
11	1.19086937	0.128229451	0.510918992	0.701937731	1.714212303	6.825056e-03

veg_res;

	Cereals	Fruits	Oilcrops	Pulses	Spices	Treenuts	VegetalProducts	VegetableOils	Vegetables
1	1.149172413	-0.140187366	-0.489360929	-0.176650092	-0.008133702	0.0642409016	-0.11696492	-0.21998548	0.235109722
2	-0.534939264	0.118912855	-0.035453794	-0.397219032	-0.626641064	0.2689466460	-1.41265625	-1.37843767	1.609268270
3	-0.054352294	0.041850819	-0.456534674	0.023817739	-0.277282439	0.1970296399	0.78493051	1.29886847	0.967994281
4	0.684070011	-0.230610003	0.142078142	0.182453528	-0.603692124	-0.8199946427	0.67437575	0.57519793	-0.880272577
5	-0.364734651	0.805910172	-0.422314750	-0.526246430	0.173452573	-0.6098889866	-0.87535285	-0.61006235	-0.295664920
6	-0.956383351	-0.397259639	-0.699740562	-0.651199278	-0.579183547	-0.6901513108	-1.20707963	-0.18575862	-0.574746604
7	-0.587266425	0.005586331	-0.571908992	-0.498811130	-0.603692124	0.3132504671	-1.12041779	-0.85108389	2.654762093
8	-1.063424747	-0.166192821	-0.436251436	-0.638432356	-0.402276182	1.1236054011	-0.42666072	0.26446649	-0.449136235
9	-0.988388979	-0.378411648	-0.469742364	-0.676461484	-0.399379714	0.1728527624	-0.89076079	-0.09448699	-0.498719276
10	0.336562381	0.061891720	-0.673432387	-0.639790539	-0.588764172	1.9124834135	-1.42760183	-1.68738091	1.623907072
11	-0.233634068	0.426325962	-0.154344444	-0.652285824	-0.020833601	-0.3016031175	-1.19093441	-1.24223959	0.415025326

Showing 1 to 12 of 170 entries, 9 total columns

- Fit the canonical correlation model and get canonical correlations. We can conclude that there is six (equal to number of variables in set "ani_res"; small set) canonical correlations in this model.

```
cc_model <- cc(ani_res,veg_res)

cc_model$cor

## [1] 0.9999999 0.5533591 0.4215768 0.3705586 0.2367454 0.1914381
```

We can see first canonical correlation has high correlation therefore enough evidence to say first canonical variate pair is highly correlated and other canonical variate pairs are not highly correlated.

- Test for independence between canonical variate pairs.

```
## Test of H0: The canonical correlations in the
## current row and all that follow are zero
##
##      CanR LR test stat approx F numDF  denDF   Pr(> F)
## 1 1.00000      0.00000   325.43    54 794.94 < 2.2e-16 ***
## 2 0.55336      0.44754     3.46    40 682.78 2.372e-11 ***
## 3 0.42158      0.64506     2.62    28 567.49 1.553e-05 ***
## 4 0.37056      0.78449     2.23    18 447.38  0.002788 **
## 5 0.23675      0.90936     1.55    10 318.00  0.121622
## 6 0.19144      0.96335     1.52     4 160.00  0.198327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First four canonical variate pairs are significantly correlated and two sets of variables are dependent on one another. Therefore we can conclude that first four canonical correlations are significant and last two are not significant at 1% significance level.

- Significant canonical correlations and squared canonical correlations

Significant canonical correlations

```
## [1] 0.9999999 0.5533591 0.4215768 0.3705586
```

Squared canonical correlations

```
## [1] 0.9999998 0.3062063 0.1777270 0.1373137
```

From squared canonical correlation we can conclude, 99.9% of the variation in first canonical variable of “ani_res” set is explained by the variation in first canonical variable of “veg_res” set. Others have low values. Therefore, only first one is very high canonical correlation and implies that only the first canonical correlation is important.

- The estimated canonical coefficients for the “ani_res” set.

```
##                        [,1]       [,2]       [,3]       [,4]
[,5]
## AnimalProducts 0.9976232826 -2.4319721  45.659204  32.900305 -
21.998206
## AnimalFats     0.0009558419  1.0097234 -18.906548 -14.072821
8.796062
## Eggs           0.0003018777  0.6603871  -4.118515  -1.942727
1.173209
## FishSeafood    0.0002717702 -0.7476856  -5.481495  -3.651426
2.629940
## Meat           0.0013725996  1.2404833 -26.289839 -19.480356
12.654152
## Milk           0.0009512173  0.8444683 -18.996125 -13.272894
10.113044
##                      [,6]
## AnimalProducts  46.395161
## AnimalFats     -18.200507
## Eggs            -3.775237
## FishSeafood     -5.178059
## Meat           -27.823843
## Milk           -19.386196
```

The magnitudes of the coefficients give the contribution of the individual variables to the corresponding canonical variable. ‘AnimalProducts’ give the best contribution to all the canonical varibales of “ani_res” set.

- Estimated canonical coefficients for the “veg_res” set

```
##                          [,1]        [,2]         [,3]       [,4]        [,5]
## Cereals          4.570047e-05  1.05494516  0.390373285  0.1506988 -3.41524936
## Fruits          -9.841973e-05  0.14345828 -0.008032296  0.3311021 -1.33116594
## Oilcrops         7.767810e-05  0.97257212 -0.370950339 -0.7350961 -4.53988448
## Pulses          -9.430399e-05  0.45090825  0.132607195 -0.5366556  0.69313194
## Spices          -4.609398e-05 -0.16917091 -0.340577118  0.1816536 -0.55261436
## Treenuts         4.275477e-05  0.09905329 -0.598279358  0.2204507 -0.78211597
## VegetalProducts -1.000072e+00 -2.90147472  0.434466565  0.9746708  7.68910949
## VegetableOils    1.251305e-04  2.75651525 -0.778604325 -0.7691170 -6.47587348
## Vegetables       4.227223e-05  0.24592643  0.104341368  0.5575298 -0.01334811
##                        [,6]
## Cereals          0.59983462
## Fruits          -0.35715100

## Oilcrops         0.16267779
## Pulses           0.31490783
## Spices          -0.43740568
## Treenuts         0.82422368
## VegetalProducts -0.30511423
## VegetableOils    0.05501768
## Vegetables      -0.36984505
```

“veg_res” variables give less contribution to first canonical variable of that set. “VegetalProducts” and “VegetableOils” give high contribution to second canonical variable of this set. “VegetableOils” gives highest contribution to third canonical variable of this set. “VegetalProducts” gives highest contribution to fourth and fifth canonical variable of this set. “Treenuts” gives highest contribution to sixth canonical variable of this set.

- The correlation between the “ani_res” variables and the canonical variables for “ani_res” set.

```
##                       [,1]          [,2]          [,3]          [,4]
## AnimalProducts  0.99999999 -1.617076e-05  8.333953e-05 -4.354264e-05
## AnimalFats      0.68682947  1.651676e-01 -2.841463e-01 -3.843286e-01
## Eggs            0.44629255  1.925868e-01 -5.340615e-01  5.570080e-01
## FishSeafood    -0.02024282 -8.957598e-01 -3.253313e-01  1.952269e-01
## Meat            0.73678527 -1.101341e-01  3.477803e-01 -1.016160e-01
## Milk            0.60941697  2.047182e-01 -4.128288e-02  3.459455e-01
##                         [,5]          [,6]
## AnimalProducts  5.743002e-05  2.647053e-05
## AnimalFats     -5.730575e-02  5.188964e-01
## Eggs           -4.055938e-01 -6.121575e-02
## FishSeafood    -2.292318e-01  2.651973e-02
## Meat           -3.335592e-01 -4.499770e-01
## Milk            6.728883e-01  1.119829e-01
```

First canonical variable correlations are uniformly large. Therefore we can think of this canonical variate as an overall measure of "ani_res" set. Rest of canonical variables for "ani_res", none of the correlations are particularly large, so those canonical variables yield little information about the data.

- The correlation between the "veg_res" and the canonical variables for "veg_res" set.

```
##                        [,1]          [,2]          [,3]          [,4]
## Cereals         -0.45645161 -1.354637e-02  6.795183e-01  3.055238e-01
## Fruits          -0.10941149 -8.498520e-02  1.993843e-01  1.947085e-01
## Oilcrops        -0.42195573 -7.506256e-01  1.197709e-01 -2.952161e-01
## Pulses          -0.42107070  6.625405e-02  3.048556e-01 -6.173674e-02
## Spices          -0.17627460 -3.552740e-01 -2.290476e-01  2.819358e-01
## Treenuts         0.13245861 -8.613161e-02 -6.489986e-01  4.438821e-01
## VegetalProducts -0.99999998  3.514645e-05 -6.960627e-05  6.160517e-06
## VegetableOils   -0.67444851  5.984756e-01 -3.528274e-01 -8.004472e-02
## Vegetables       0.04955136  1.035675e-01 -6.083604e-03  7.689409e-01
##                          [,5]          [,6]
## Cereals         -6.229153e-02  3.631788e-01
## Fruits           6.489570e-02 -3.172500e-01
## Oilcrops        -9.824484e-02 -5.461504e-02
## Pulses           5.354597e-01  5.342975e-02
## Spices           2.227011e-01 -2.656214e-01
## Treenuts         9.785763e-02  5.613001e-01

## VegetalProducts -7.008511e-05  4.871979e-05
## VegetableOils   -4.003546e-02 -1.268832e-01
## Vegetables       1.978142e-01 -2.408587e-01
```

All the correlations are large for the first canonical variable, this can be thought of as an overall measure of "veg_res" set as well. However, it is most strongly correlated with "VegetalProducts".Therefore we can conclude that, best predictor of "ani_res" is "VegetalProducts" as this indicator stands out most.

- The correlation between the "ani_res" variables and the canonical variables for "veg_res" set.

```
##                        [,1]          [,2]          [,3]          [,4]
## AnimalProducts  0.99999987 -8.948234e-06  3.513401e-05 -0.0000161351
## AnimalFats      0.68682939  9.139697e-02 -1.197895e-01 -0.1424162619
## Eggs            0.44629250  1.065697e-01 -2.251479e-01  0.2064041314
## FishSeafood    -0.02024282 -4.956768e-01 -1.371521e-01  0.0723430262
## Meat            0.73678518 -6.094371e-02  1.466161e-01 -0.0376546744
## Milk            0.60941690  1.132827e-01 -1.740390e-02   0.1281930964
##                          [,5]          [,6]
## AnimalProducts  1.359629e-05  5.067468e-06
## AnimalFats     -1.356687e-02  9.933656e-02
## Eggs           -9.602245e-02 -1.171903e-02
## FishSeafood    -5.426956e-02  5.076889e-03
## Meat           -7.896859e-02 -8.614276e-02
## Milk            1.593032e-01  2.143780e-02
```

Here we consider only first canonical variable because other canonical variables have small correlations. According to first canonical variable, all six of these correlations are strong and show a pattern similar to that with the first canonical variate for “ani_res” set. The reason for this obvious: the first canonical correlation is very high.

- The correlations between the “veg_res” variables and the canonical variables for “ani_res” set.

```
##                        [,1]          [,2]          [,3]          [,4]
## Cereals         -0.45645155 -7.496004e-03  2.864691e-01  1.132145e-01
## Fruits          -0.10941147 -4.702733e-02  8.405578e-02  7.215092e-02
## Oilcrops        -0.42195568 -4.153655e-01  5.049264e-02 -1.093949e-01
## Pulses          -0.42107065  3.666228e-02  1.285200e-01 -2.287708e-02
## Spices          -0.17627458 -1.965941e-01 -9.656115e-02  1.044737e-01
## Treenuts         0.13245860 -4.766171e-02 -2.736027e-01  1.644843e-01
## VegetalProducts -0.99999986  1.944861e-05 -2.934439e-05  2.282833e-06
## VegetableOils   -0.67444843  3.311719e-01 -1.487439e-01 -2.966126e-02
## Vegetables       0.04955135  5.731004e-02 -2.564706e-03  2.849377e-01
##                          [,5]          [,6]
## Cereals         -1.474723e-02  6.952628e-02
## Fruits           1.536376e-02 -6.073376e-02
## Oilcrops        -2.325901e-02 -1.045540e-02
## Pulses           1.267676e-01  1.022849e-02
## Spices           5.272345e-02 -5.085007e-02
## Treenuts         2.316734e-02  1.074542e-01
## VegetalProducts -1.659233e-05  9.326826e-06
## VegetableOils   -9.478210e-03 -2.429029e-02
## Vegetables       4.683160e-02 -4.610953e-02
```

Here we consider only first canonical variable because other canonical variables have small correlations. According to first canonical variable, all nine of these correlations are strong and show a pattern similar to that with the first canonical variate for “veg_res” set. Again, this is because the first canonical correlation is very high. These results confirm that “ani_res” set is best predicted by “VegetalProducts”.

4. **Conclusion and Recommendation**

- From the analysis we can conclude that don’t want to get 54 pairwise scatterplots to explain the dataset. We can do it from only four canonical variate pairs. I proved it by “Wilki’s lambda” test.

- From squared canonical correlation we can conclude, 99.9% of the variation in first canonical variable of “ani_res” set is explained by the variation in first canonical variable of “veg_res” set.
- These results confirm that “ani_res” set is best predicted by “VegetalProducts”.
- Finally, we can conclude that there is strong relationship among these two sets of variables because first canonical correlation is significant and its value is 0.99999.

5. References

References

Institute for Digital Research and Education, S. C. (n.d.). *Canonical Correlation Analysis.*

Korstanje, J. (n.d.). *Canonical Correlation Analysis.*

6. Appendices

- Part of the Dataset

AnimalProducts	AnimalFats	Cereals	Eggs	FishSeafood	Fruits	Meat	Milk	Oilcrops	Pulses	Spices	Treenuts	VegetalProducts	VegetableOils	Vegetables
21.6397	6.2224	8.0353	0.6859	0.0327	0.4246	6.1244	8.2803	1.0452	0.196	0.2776	0.7513	28.3684	17.0831	0.3593
32.0002	3.4172	2.6734	1.6448	0.1445	0.6418	8.7428	17.7576	3.1622	0.1148	0	0.9181	17.9998	9.2443	0.6503
14.4175	0.8972	4.2035	1.2171	0.2008	0.5772	3.8961	8.0934	1.1983	0.2698	0.1568	0.8595	35.5857	27.3606	0.5145
15.3041	1.313	6.5545	0.1539	1.4155	0.3488	11.0268	1.2309	3.9902	0.3282	0.0103	0.0308	34.701	22.4638	0.1231
27.7033	4.6686	3.2153	0.3872	1.5263	1.2177	14.3202	6.6607	1.3579	0.0673	0.3591	0.202	22.2995	14.4436	0.2469
30.3572	3.3076	1.3316	1.5706	0.1664	0.2091	19.2693	5.8512	0.064	0.0213	0.0213	0.1366	19.6449	17.3147	0.1878
29.6642	6.2619	2.5068	1.6196	0.2218	0.5468	10.8165	10.4709	0.6602	0.0774	0.0103	0.9542	20.3384	12.8127	0.8717
24.1099	4.603	0.9908	0.7017	0.4515	0.4028	11.6002	6.5196	1.2929	0.026	0.1007	1.6145	25.8901	20.3612	0.2144
27.8268	12.8517	1.2297	1.2147	0.4259	0.2249	8.1099	5.1497	1.1367	0.012	0.102	0.8398	22.1762	17.9323	0.2039

- Codes

```
library(ggplot2)

library(GGally)

library(CCA)

library(CCP)

library(CCA)

library(candisc)

library(skimr)
```

```r
fat <- read.csv("Fat_Supply_Quantity_Data.csv",header = TRUE)
head(fat)

skim(fat)

fat[is.na(fat)] <- 0

st_fat <- apply(fat,2,scale)

matcor(ani_res,veg_res)

cc_model <- cc(ani_res,veg_res)

cc_model$cor

rho <- cc_model$cor
n <- dim(ani_res)[1]
p <- dim(ani_res)[2]
q <- dim(veg_res)[2]

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")

Wilks(cancor(ani_res,veg_res))

cc_model$cor[1:4]

cc_model$cor[1:4]^2

cc_model$xcoef

cc_model$ycoef

loadings <- comput(ani_res,veg_res,cc_model)

loadings$corr.X.xscores

loadings$corr.Y.yscores

loadings$corr.X.yscores

loadings$corr.Y.xscores
```