Canonical Correlation Analysis Report for Travel Reviews Dataset

S/18/843

1 Introduction

Canonical Correlation Analysis (CCA) is a statistical technique that is used to examine and measure the relationships between two different sets of variables. This method is valuable when trying to comprehend how the variables in one set are connected to those in another set.

This method is useful for finding complicated connections between different variables, simplifying data by concentrating on the most important combinations of variables and improving predictive analysis by using the discovered relationships.

Travel Reviews dataset is populated by crawling TripAdvisor.com. Reviews on destinations in 10 categories mentioned across East Asia are considered. Each traveler rating is mapped as Excellent (4), Very Good (3), Average (2), Poor (1), and Terrible (0) and average rating is used against each category per user.

2 Methodology

I perform canonical correlation analysis using the Travel Reviews dataset. This data set is populated by crawling TripAdvisor.com. Reviews on destinations in 10 categories mentioned across East Asia are considered. Each traveler rating is mapped as Excellent (4), Very Good (3), Average (2), Poor (1), and Terrible (0), and the average rating is used against each category per user. There are 10 variables, all quantitative, and continuous variables.

Variable Description:

art.galleries : Average user feedback on art galleries

dance.clubs : Average user feedback on dance clubs

juice.bars : Average user feedback on juice bars

restaurants : Average user feedback on restaurants

museums : Average user feedback on museums

resorts : Average user feedback on resorts

parks.picnic.spots : Average user feedback on parks/picnic spots

beaches : Average user feedback on beaches

theaters : Average user feedback on theaters religious.institutions : Average user feedback on religious institutions

- Since different variables have different measurement units, I standardized the whole data set.
- Test the hypothesis that the canonical variate pairs are correlated (i.e., the canonical correlations are significantly different from zero) at a 1% significance level.
- For the two sets of variables (cultural artistic, recreational social), I use Canonical Correlation Analysis.

3 Results and Discussion

3.1 Summary of the dataset

```
── Data Summary ────────────────────────
                           Values
Name                       travel_reviews
Number of rows             980
Number of columns          11
_______________________
Column type frequency:
  character                1
  numeric                  10
________________________
Group variables            None
```

3.2 Split the dataset

Split the dataset into two sets

```
> cultural_artistic <- travel_reviews[,c("art.galleries","restaurants","museums","thea
ters","religious.institutions")]
> recreational_social <- travel_reviews[,c("dance.clubs","juice.bars","resorts","parks
.picnic.spots","beaches")]
```

Set 1: Cultural Artistic dataset

```
head(cultural_artistic)
     art.galleries  restaurants    museums     theaters religious.institutions
[1,]     0.1125872  0.312800143 -0.3194447  0.687166716             -1.1799885
[2,]     0.3878904  0.384297318  1.0979250  0.796867124             -1.4911467
[3,]     0.9996754 -0.008937147 -1.5996497 -0.711513490             -0.9310618
[4,]    -1.3556968  0.134057204 -1.0967120  0.001539164              0.1891080
[5,]    -1.1721613  0.134057204  1.3722547 -1.068039817             -0.8065985
[6,]     0.2961227 -0.938400429 -0.4566095  0.248365083              2.6783744
```

Set 2: Recreational Social dataset

```
> head(recreational_social)
     dance.clubs juice.bars    resorts parks.picnic.spots    beaches
[1,]   0.9354094  1.6189231  1.0696225          1.1580656 -0.3277049
[2,]   1.7717393  2.0881049  2.4782350          3.7141565 -1.4912948
[3,]  -1.1554154 -0.6001801 -0.5614024         -0.1199798 -0.2549805
[4,]   0.9354094 -0.9171948 -0.5984712         -0.1199798  0.9086094
[5,]  -0.3190855  0.2113777  0.3282476         -0.1199798 -0.4004293
[6,]  -0.1518195 -0.3719295 -1.0803649         -1.3980252  0.3995388
```

3.3 Fit the canonical correlation model

```
> cc_model <- cc(cultural_artistic,recreational_social)
> cc_model$cor
[1] 0.7761525 0.5439067 0.2991467 0.2169611 0.1430161
```

The first pair of canonical variates have a strong correlation (0.7762), indicating a substantially linear relationship between the cultural artistic, and recreational social datasets. The second pair shows a moderate correlation (0.5439), suggesting a moderately strong relationship. However, the subsequent pairs have much weaker correlations (0.2991, 0.2170, and 0.1430), indicating diminishing linear relationships between the corresponding canonical variates of the two datasets.

3.4 Test for independence between canonical variate pairs.

```
Wilks' Lambda, using F-approximation (Rao's F):
              stat   approx df1      df2      p.value
1 to 5:  0.2379455 68.03045  25 3604.892 0.000000e+00
2 to 5:  0.5984737 33.93327  16 2967.092 0.000000e+00
3 to 5:  0.8499049 18.16542   9 2365.743 0.000000e+00
4 to 5:  0.9334370 17.04737   4 1946.000 9.481305e-14
5 to 5:  0.9795464 20.33781   1  974.000 7.280554e-06
 Hotelling-Lawley Trace, using F-approximation:
              stat   approx df1  df2      p.value
1 to 5:  2.10385423 81.49490  25 4842 0.000000e+00
2 to 5:  0.58868288 35.70362  16 4852 0.000000e+00
3 to 5:  0.16856216 18.21220   9 4862 0.000000e+00
4 to 5:  0.07027809 17.11974   4 4872 5.950795e-14
5 to 5:  0.02088070 20.38792   1 4882 6.471646e-06
 Pillai-Bartlett Trace, using F-approximation:
              stat   approx df1  df2      p.value
1 to 5:  1.05526180 52.11119  25 4870 0.000000e+00
2 to 5:  0.45284902 30.37483  16 4880 0.000000e+00
3 to 5:  0.15701451 17.61542   9 4890 0.000000e+00
4 to 5:  0.06752576 16.77030   4 4900 1.161293e-13
5 to 5:  0.02045362 20.16796   1 4910 7.255744e-06
 Roy's Largest Root, using F-approximation:
              stat   approx df1 df2 p.value
1 to 1:  0.6024128 295.1554   5 974       0

 F statistic for Roy's Greatest Root is an upper bound.
```

```
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 0.77615      0.23795   68.030    25 3604.9 < 2.2e-16 ***
2 0.54391      0.59847   33.933    16 2967.1 < 2.2e-16 ***
3 0.29915      0.84990   18.165     9 2365.7 < 2.2e-16 ***
4 0.21696      0.93344   17.047     4 1946.0 9.486e-14 ***
5 0.14302      0.97955   20.338     1  974.0 7.281e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The relationships between the variable sets are significant, as indicated by the extremely small p-values ($p < 0.001$) across all statistical tests (Wilks' Lambda, Hotelling-Lawley Trace, Pillai-Bartlett Trace, and Roy's Largest Root). Therefore, we can conclude that two sets of variables are dependent on one another at a 1% significance level.

3.5 Significant canonical correlations and squared canonical correlations

```
> cc_model$cor[1:4]
[1] 0.7761525 0.5439067 0.2991467 0.2169611
> cc_model$cor[1:4]^2
[1] 0.60241277 0.29583451 0.08948875 0.04707214
```

The first pair of canonical variates explains 60.2% of the variance, demonstrating a strong relationship. The second pair explains 29.6% of the variance, indicating a moderate relationship. However, the third and fourth pairs explain only 8.9% and 4.7% of the variance, respectively, suggesting much weaker associations.

3.6 Estimated canonical coefficients for the cultural artistic dataset

```
> cc_model$xcoef
                              [,1]        [,2]        [,3]        [,4]       [,5]
art.galleries           0.13310732  0.25367158 -0.34548782 -0.76550032  0.4853870
restaurants             0.02726077 -0.01231734  0.56121497  0.40904127  0.8192104
museums                 0.31540217  0.96545073  0.09542636  0.11086420 -0.1503436
theaters                0.10792682 -0.11252324  0.75994134 -0.58141624 -0.2554495
religious.institutions -0.85345314  0.52645756  0.38063004  0.08348416  0.2286347
```

Religious institutions have the largest negative effect on the first canonical variate, while museums have the highest positive effect on the second variate. Additionally, theaters and restaurants positively impact the third and fifth variates, respectively, and art galleries have significant negative impacts on the fourth variate.

3.7 Estimated canonical coefficients for the recreational social dataset

```
> cc_model$ycoef
                         [,1]        [,2]       [,3]       [,4]       [,5]
dance.clubs       -0.06917722  0.02667484  0.6181748  0.7075859 -0.4158924
juice.bars        -0.25385844  0.69621132 -0.9173813  0.4397939 -0.9030657
resorts            0.47733163  0.89554961  0.2468120 -0.2020184  0.3498486
parks.picnic.spots  0.87875500 -1.20909386  0.3658250 -0.2664925  0.2464463
beaches           -0.10245814  0.08733309  0.3187923 -0.5182687 -0.8273824
```

Parks and picnic spots have the strongest positive impact on the first canonical variate, with resorts also contributing positively. In the second variate, juice bars and resorts have the most significant positive impacts, while parks and picnic spots exert a strong negative influence. The third variate is primarily influenced by dance clubs, which have the highest positive coefficient, whereas juice bars have the strongest negative impact. For the fourth variate, dance clubs again have the highest positive impact, and in the fifth variate, juice bars, and beaches exhibit the greatest negative effects..

3.8 Correlation between the cultural artistic variables and the canonical variables for the cultural artistic dataset

```
$corr.X.xscores
                             [,1]        [,2]       [,3]       [,4]       [,5]
art.galleries          0.05504102  0.18777105 -0.3305092 -0.7148527  0.5843473
restaurants            0.37324072 -0.08411875  0.4314985  0.3191368  0.7520480
museums                0.52050814  0.80373251  0.1248073  0.1845547 -0.1828881
theaters               0.15449399 -0.10865038  0.7778456 -0.5334805 -0.2732790
religious.institutions -0.93930801  0.30988609  0.1067249 -0.1005564  0.0130117
```

The loadings indicate how each variable correlates with the canonical variates. For the first canonical variate, religious institutions have the strongest negative correlation, while museums have the strongest positive correlation. The second variate is mainly associated with museums, explaining a strong positive correlation. In the third variate, theaters exhibit the highest positive correlation and in the fourth and fifth variates, art galleries have the strongest negative correlation and restaurant have the strongest positive correlation.

3.9 Correlation between the recreational social variables and the canonical variables for the recreational social dataset

```
$corr.Y.yscores
                        [,1]        [,2]       [,3]        [,4]        [,5]
dance.clubs        0.1035588  0.04300473  0.6045036  0.74966552 -0.24495543
juice.bars         0.5906227  0.09386709 -0.5829943  0.28822178 -0.46840110
resorts            0.7550921  0.62685925  0.1691639 -0.05507864  0.07234128
parks.picnic.spots  0.8936176 -0.30412544 -0.1715739  0.09206271 -0.26653728
beaches           -0.1112469  0.05040200  0.3528860 -0.68728086 -0.62305718
```

The loadings of the recreational and social activities on the canonical variates reveal that parks and picnic spots have the highest positive correlation with the first canonical variate, followed closely by resorts. Juice bars exhibit a strong negative correlation with the third variate, while beaches have a significant negative correlation with the fifth variate.

3.10 Correlation between the cultural artistic variables and the canonical variables for the recreational social dataset

```
> loadings$corr.X.yscores
                             [,1]        [,2]        [,3]        [,4]         [,5]
art.galleries          0.04272023  0.10212993 -0.09887073 -0.15509527  0.083571098
restaurants            0.28969173 -0.04575275  0.12908136  0.06924029  0.107555002
museums                0.40399372  0.43715551  0.03733569  0.04004119 -0.026155947
theaters               0.11991091 -0.05909567  0.23268996 -0.11574454 -0.039083309
religious.institutions -0.72904630  0.16854912  0.03192640 -0.02181682  0.001860883
```

The loadings of cultural and artistic activities on the canonical variates for the recreational and social activity scores show that religious institutions have the highest negative correlation with the first canonical variate. Museum's exhibit a moderate positive correlation with both the first and second canonical variates. Art galleries and restaurants have relatively smaller correlations, suggesting a more moderate impact on the recreational and social dimensions.

3.11 Correlation between the recreational and social variables and the canonical variables for the cultural artistic dataset

```
> loadings$corr.Y.xscores
                         [,1]        [,2]        [,3]        [,4]        [,5]
dance.clubs        0.08037739  0.02339056  0.18083526  0.16264829 -0.03503258
juice.bars         0.45841334  0.05105494 -0.17440083  0.06253293 -0.06698892
resorts            0.58606669  0.34095295  0.05060483 -0.01194993  0.01034597
parks.picnic.spots  0.69358359 -0.16541587 -0.05132576  0.01997403 -0.03811914
beaches           -0.08634460  0.02741398  0.10556469 -0.14911324 -0.08910724
```

The loadings of recreational and social activities on the canonical variates for the cultural and artistic activity scores indicate that parks and picnic spots have the highest positive correlation with the first canonical variate. Resorts also show a significant positive correlation with the first and second canonical variates, indicating their notable influence. Conversely, dance clubs and beaches have relatively low correlations across all canonical variates, implying a lesser impact on the cultural and artistic dimensions.

4 Conclusion and Recommendation

- From the analysis, we can conclude that no need to get all the pairwise scatterplots to explain the dataset. We can do it from only five canonical variate pairs. I proved it with "Wilk's lambda" test.
- we can conclude that 60.2% of the variation in the first canonical variable of the cultural artistic dataset is explained by the variation in the first canonical variable of the recreational social dataset and 29.6% of the variation in the second canonical variable of the cultural artistic dataset is explained by the variation in the second canonical variable of recreational social dataset From the squared canonical correlation. These findings confirm that the set of recreational social variables partially predicts the set of cultural artistic variables.
- The loadings indicate that parks and picnic spots, as well as resorts, are the most influential recreational activities associated with cultural and artistic activities. In contrast, dance clubs and beaches exhibit weaker associations across all canonical variates.
- Finally, we can conclude that there is a strong relationship between these two sets of variables because the first correlation is significant, and it is 0.7761525 and the second correlation is significant, and it is 0.5439067.

5 References

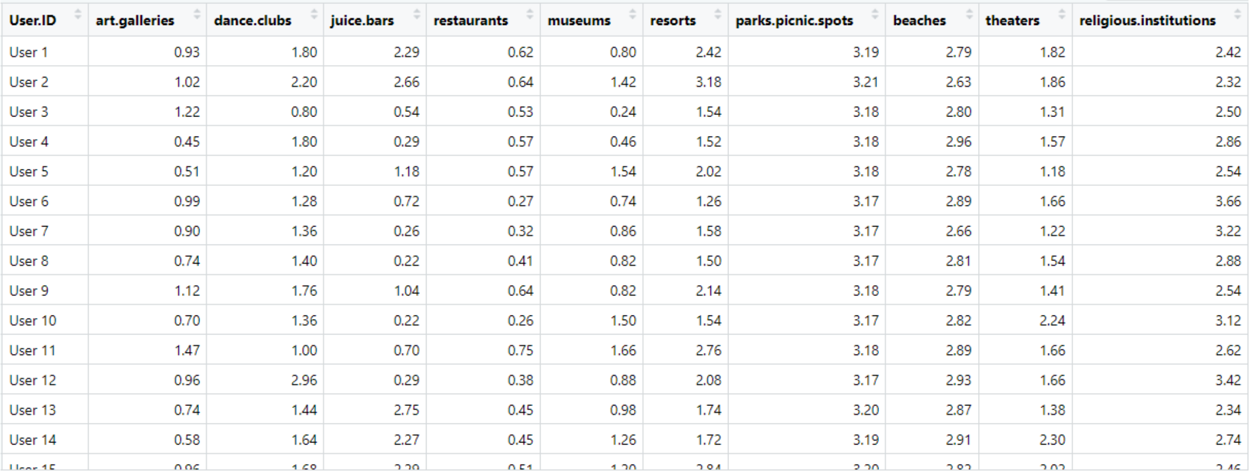
https://archive.ics.uci.edu/dataset/484/travel+reviews

https://stats.oarc.ucla.edu/r/dae/canonical-correlation-analysis/

https://online.stat.psu.edu/stat505/book/export/html/682

6 Appendices

6.1 Part of dataset



User.ID	art.galleries	dance.clubs	juice.bars	restaurants	museums	resorts	parks.picnic.spots	beaches	theaters	religious.institutions
User 1	0.93	1.80	2.29	0.62	0.80	2.42	3.19	2.79	1.82	2.42
User 2	1.02	2.20	2.66	0.64	1.42	3.18	3.21	2.63	1.86	2.32
User 3	1.22	0.80	0.54	0.53	0.24	1.54	3.18	2.80	1.31	2.50
User 4	0.45	1.80	0.29	0.57	0.46	1.52	3.18	2.96	1.57	2.86
User 5	0.51	1.20	1.18	0.57	1.54	2.02	3.18	2.78	1.18	2.54
User 6	0.99	1.28	0.72	0.27	0.74	1.26	3.17	2.89	1.66	3.66
User 7	0.90	1.36	0.26	0.32	0.86	1.58	3.17	2.66	1.22	3.22
User 8	0.74	1.40	0.22	0.41	0.82	1.50	3.17	2.81	1.54	2.88
User 9	1.12	1.76	1.04	0.64	0.82	2.14	3.18	2.79	1.41	2.54
User 10	0.70	1.36	0.22	0.26	1.50	1.54	3.17	2.82	2.24	3.12
User 11	1.47	1.00	0.70	0.75	1.66	2.76	3.18	2.89	1.66	2.62
User 12	0.96	2.96	0.29	0.38	0.88	2.08	3.17	2.93	1.66	3.42
User 13	0.74	1.44	2.75	0.45	0.98	1.74	3.20	2.87	1.38	2.34
User 14	0.58	1.64	2.27	0.45	1.26	1.72	3.19	2.91	2.30	2.74
User 15	0.96	1.68	2.29	0.51	1.20	2.84	3.20	2.82	2.02	2.46

6.2 R Codes

```
library(tidyverse)

library(CCA)

library(CCP)

library(skimr)

library(candisc)

travel_reviews <- read.csv("../Data/tripadvisor_review.csv")

view(travel_reviews)

travel_reviews<-travel_reviews[,-1]

head(travel_reviews)

skim(travel_reviews)

travel_reviews <- apply(travel_reviews,2,scale)
```

```
cultural_artistic <-
travel_reviews[,c("art.galleries","restaurants","museums","theaters","religious.institutions")]

recreational_social <-
travel_reviews[,c("dance.clubs","juice.bars","resorts","parks.picnic.spots","beaches")]

head(cultural_artistic)

head(recreational_social)

matcor(cultural_artistic,recreational_social)

cc_model <- cc(cultural_artistic,recreational_social)

cc_model$cor

rho <- cc_model$cor

n <- dim(cultural_artistic)[1]

p <- dim(recreational_social)[2]

q <- dim(cultural_artistic)[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(cultural_artistic,recreational_social))

cc_model$cor[1:4]

cc_model$cor[1:4]^2

cc_model$xcoef

cc_model$ycoef

loadings <- comput(cultural_artistic,recreational_social,cc_model)

loadings
```

```
loadings$corr.X.xscores

loadings$corr.Y.yscores

loadings$corr.X.yscores

loadings$corr.Y.xscores
```