Multivariate Procedures with R

Canonical Correlation Analysis

::::

Canonical Variables Analysis in R

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Canonical Correlation Analysis: Example

[Chapter 12, Canonical variables and canonical correlations in An Introduction to Multivariate Statistical Analysis, TW Anderson, Wiley]

The head length and head breadth of two children are collected in a sample of 25 families. Let for $\alpha = 1, 2, \ldots, 25$,

$x_{1\alpha}$: Head length of the first child in α^{th} family,

$x_{2\alpha}$: Head breadth of the first child in α^{th} family,

$x_{3\alpha}$: Head length of the second child in α^{th} family, and

$x_{4\alpha}$: Head breadth of the second child in α^{th} family.

The aim is to investigate the relations between the measurements for the first child and for the second child.

Thus $\underline{x}_{\alpha}^{(1)} = (x_{1\alpha}, x_{2\alpha})'$ and $\underline{x}_{\alpha}^{(2)} = (x_{3\alpha}, x_{4\alpha})'$.

Canonical Correlation Analysis: Example

The sample data is summarized as sample mean vector and sample covariance matrix as

$$\underline{\bar{x}} = \begin{bmatrix} 185.72 \\ 151.12 \\ 183.84 \\ 149.24 \end{bmatrix}, \quad S = \frac{1}{24}A = \left[\begin{array}{cc|cc} 95.29 & 52.86 & 69.66 & 46.11 \\ 52.86 & 54.36 & 51.31 & 35.05 \\ \hline 69.66 & 51.31 & 100.81 & 56.54 \\ 46.11 & 35.05 & 56.54 & 45.02 \end{array}\right]$$

The correlation matrix (scaled data of covariance matrix) is

$$R = \left[\begin{array}{cc|cc} 1.00 & 0.73 & 0.71 & 0.70 \\ 0.73 & 1.00 & 0.69 & 0.71 \\ \hline 0.71 & 0.69 & 1.00 & 0.84 \\ 0.70 & 0.71 & 0.84 & 1.00 \end{array}\right] = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

Canonical Correlation Analysis: Example

The characteristics roots are obtained as $\widehat{\rho_1^{*2}} = 0.622$,

$\widehat{\rho_2^{*2}} = 0.622$, so that $\widehat{\rho_1^*} = 0.789$, $\widehat{\rho_2^*} = 0.054$.

Then for

$$\widehat{\rho_1^*} = 0.789, \quad \underline{\hat{a}}_1 = \begin{pmatrix} 0.057 \\ 0.071 \end{pmatrix}, \quad \underline{\hat{b}}_1 = \begin{pmatrix} 0.050 \\ 0.080 \end{pmatrix};$$

$$\widehat{\rho_2^*} = 0.054, \quad \underline{\hat{a}}_2 = \begin{pmatrix} 0.140 \\ -0.187 \end{pmatrix}, \quad \underline{\hat{b}}_2 = \begin{pmatrix} 0.176 \\ -0.262 \end{pmatrix}.$$

Since $\widehat{\rho_1^*} = 0.789 > \widehat{\rho_2^*} = 0.054$. So we can confine our attention to the first canonical variates to study the relation between two head dimensions of first child and second child. The second canonical variates are weakly correlated.

Canonical Correlation Analysis: Example

$$\widehat{\rho_1^*} = 0.789, \quad \underline{\hat{a}}_1 = \begin{pmatrix} 0.057 \\ 0.071 \end{pmatrix}, \quad \underline{\hat{b}}_1 = \begin{pmatrix} 0.050 \\ 0.080 \end{pmatrix};$$

$$\widehat{\rho_2^*} = 0.054, \quad \underline{\hat{a}}_2 = \begin{pmatrix} 0.140 \\ -0.187 \end{pmatrix}, \quad \underline{\hat{b}}_2 = \begin{pmatrix} 0.176 \\ -0.262 \end{pmatrix}.$$

The canonical variables are as follows:

$$U_1 = 0.057X_1 + 0.071X_2, \quad V_1 = 0.050X_1 + 0.080X_2$$

$$U_2 = 0.140X_1 - 0.187X_2, \quad V_2 = 0.176X_1 - 0.262X_2$$

The first set of canonical variable s are as follows:

$$U_1 = 0.057X_1 + 0.071X_2, \quad V_1 = 0.050X_1 + 0.080X_2$$

Canonical Correlation Analysis in R:

Compute the canonical correlations between two data matrices.

Usage

```
cancor(x, y, xcenter = TRUE, ycenter = TRUE)
```

Arguments

`x` numeric matrix (n * p), containing the X coordinates.

`y` numeric matrix (n * q), containing the Y coordinates.

`xcenter` logical or numeric vector of length, describing any centering to be done on the `x` values before the analysis. If `TRUE` (default), subtract the column means. If `FALSE`, do not adjust the columns. Otherwise, a vector of values to be subtracted from the columns.

`ycenter` analogous to `xcenter`, but for the `y` values.

Canonical Correlation Analysis in R:

Details

The canonical correlation analysis seeks linear combinations of the y variables which are well explained by linear combinations of the x variables. The relationship is symmetric as ‘well explained’ is measured by correlations.

Value

A list containing the following components:

`cor` **correlations.**

`xcoef` **estimated coefficients for the `x` variables.**

`ycoef` **estimated coefficients for the `y` variables.**

`xcenter` **the values used to adjust the `x` variables.**

`ycenter` **the values used to adjust the `x` variables.**

Canonical Correlation Analysis in R: Intercountry Life-Cycle Savings Data

Intercountry Life-Cycle Savings Data
Data on the savings ratio 1960–1970.

Usage: `LifeCycleSavings`

Format: A data frame with 50 observations on 5 variables.

[,1]	**sr**	**numeric aggregate personal savings**
[,2]	**pop15**	**numeric % of population under 15**
[,3]	**pop75**	**numeric % of population over 75**
[,4]	**dpi**	**numeric real per-capita disposable income**
[,5]	**ddpi**	**numeric % growth rate of dpi**

Canonical Correlation Analysis in R: Intercountry Life-Cycle Savings Data

Intercountry Life-Cycle Savings Data
Data on the savings ratio 1960–1970.

Details

Under the life-cycle savings hypothesis as developed by Franco Modigliani, the savings ratio (aggregate personal saving divided by disposable income) is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the percentage of the population over 75 years old. The data are averaged over the decade 1960–1970 to remove the business cycle or other short-term fluctuations.

The data were obtained from Belsley, Kuh and Welsch (1980).
Belsley, D. A., Kuh. E. and Welsch, R. E. (1980) Regression Diagnostics. New York: Wiley.

Canonical Correlation Analysis in R: Intercountry Life-Cycle Savings Data

Intercountry Life-Cycle Savings Data

Data on the savings ratio 1960–1970.

```
> LifeCycleSavings
                   sr pop15 pop75     dpi  ddpi
Australia       11.43 29.35  2.87 2329.68  2.87
Austria         12.07 23.32  4.41 1507.99  3.93
Belgium         13.17 23.80  4.43 2108.47  3.82
Bolivia          5.75 41.89  1.67  189.13  0.22
Brazil          12.88 42.19  0.83  728.47  4.56
Canada           8.79 31.72  2.85 2982.88  2.43
Chile            0.60 39.74  1.34  662.86  2.67
...   ...  ...
...   ...  ...
```

Canonical Correlation Analysis in R: Intercountry Life-Cycle Savings Data

Intercountry Life-Cycle Savings Data

Data on the savings ratio 1960–1970.

```
> LifeCycleSavings
```

R Console

```
> LifeCycleSavings
                  sr pop15 pop75     dpi ddpi
Australia      11.43 29.35  2.87 2329.68 2.87
Austria        12.07 23.32  4.41 1507.99 3.93
Belgium        13.17 23.80  4.43 2108.47 3.82
Bolivia         5.75 41.89  1.67  189.13 0.22
Brazil         12.88 42.19  0.83  728.47 4.56
Canada          8.79 31.72  2.85 2982.88 2.43
Chile           0.60 39.74  1.34  662.86 2.67
China          11.90 44.75  0.67  289.52 6.51
Colombia        4.98 46.64  1.06  276.65 3.08
Costa Rica     10.78 47.64  1.14  471.24 2.80
Denmark        16.85 24.42  3.93 2496.53 3.99
Ecuador         3.59 46.31  1.19  287.77 2.19
Finland        11.24 27.84  2.37 1681.25 4.32
France         12.64 25.06  4.70 2213.82 4.52
Germany        12.55 23.31  3.35 2457.12 3.44
Greece         10.67 25.62  3.10  870.85 6.28
Guatamala       3.01 46.05  0.87  289.71 1.48
Honduras        7.70 47.32  0.58  232.44 3.19
Iceland         1.27 34.03  3.08 1900.10 1.12
India           9.00 41.31  0.96   88.94 1.54
Ireland        11.34 31.16  4.19 1139.95 2.99
Italy          14.28 24.52  3.48 1390.00 3.54
Japan          21.10 27.01  1.91 1257.28 8.21
Korea           3.98 41.74  0.91  207.68 5.81
Luxembourg     10.35 21.80  3.73 2449.39 1.57
Malta          15.48 32.54  2.47  601.05 8.12
Norway         10.25 25.95  3.67 2231.03 3.62
Netherlands    14.65 24.71  3.25 1740.70 7.66
New Zealand    10.67 32.61  3.17 1487.52 1.76
Nicaragua       7.30 45.04  1.21  325.54 2.48
Panama          4.44 43.56  1.20  568.56 3.61
Paraguay        2.02 41.18  1.05  220.56 1.03
Peru           12.70 44.19  1.28  400.06 0.67
Philippines    12.78 46.26  1.12  152.01 2.00
Portugal       12.49 28.96  2.85  579.51 7.48
South Africa   11.14 31.94  2.28  651.11 2.19
South Rhodesia 13.30 31.92  1.52  250.96 2.00
```

Canonical Correlation Analysis in R: Intercountry Life-Cycle Savings Data

```
pop = LifeCycleSavings[, 2:3]

oec = LifeCycleSavings[, -(2:3)]

cancor(pop, oec)
```

R Console

```
> pop = LifeCycleSavings[, 2:3]
> pop
             pop15 pop75
Australia    29.35  2.87
Austria      23.32  4.41
Belgium      23.80  4.43
Bolivia      41.89  1.67
Brazil       42.19  0.83
Canada       31.72  2.85
Chile        39.74  1.34
China        44.75  0.67
Colombia     46.64  1.06
Costa Rica   47.64  1.14
Denmark      24.42  3.93
```

R Console

```
> oec <- LifeCycleSavings[, -(2:3)]
> oec
                sr     dpi  ddpi
Australia    11.43 2329.68  2.87
Austria      12.07 1507.99  3.93
Belgium      13.17 2108.47  3.82
Bolivia       5.75  189.13  0.22
Brazil       12.88  728.47  4.56
Canada        8.79 2982.88  2.43
Chile         0.60  662.86  2.67
China        11.90  289.52  6.51
Colombia      4.98  276.65  3.08
Costa Rica   10.78  471.24  2.80
Denmark      16.85 2496.53  3.99
```

Canonical Correlation Analysis in R:

```
> cancor(pop, oec)
$cor
[1] 0.8247966 0.3652762

$xcoef
               [,1]        [,2]
pop15 -0.009110856 -0.03622206
pop75  0.048647514 -0.26031158

$ycoef
             [,1]          [,2]          [,3]
sr   0.0084710221  3.337936e-02 -5.157130e-03
dpi  0.0001307398 -7.588232e-05  4.543705e-06
ddpi 0.0041706000 -1.226790e-02  5.188324e-02

$xcenter
  pop15   pop75
35.0896  2.2930

$ycenter
       sr       dpi      ddpi
   9.6710 1106.7584    3.7576
```

Canonical Correlation Analysis in R:

R Console

```
> cancor(pop, oec)
$cor
[1] 0.8247966 0.3652762

$xcoef
               [,1]        [,2]
pop15 -0.009110856 -0.03622206
pop75  0.048647514 -0.26031158

$ycoef
             [,1]          [,2]          [,3]
sr   0.0084710221  3.337936e-02 -5.157130e-03
dpi  0.0001307398 -7.588232e-05  4.543705e-06
ddpi 0.0041706000 -1.226790e-02  5.188324e-02

$xcenter
  pop15   pop75
35.0896  2.2930

$ycenter
     sr       dpi      ddpi
   9.6710 1106.7584    3.7576
```

Canonical Correlation Analysis in R:

```
> cancor(pop, oec)
$cor
[1] 0.8247966 0.3652762
```

The first canonical correlation = 0.825 is much greater than the second canonical correlation = 0.365. So, we consider the first canonical correlation and the first canonical variates

```
$xcoef
                  [,1]        [,2]
pop15 -0.009110856 -0.03622206
pop75  0.048647514 -0.26031158
```

First canonical variable $U_1 = -0.009 * \mathbf{pop15} + 0.049 * \mathbf{pop75}$

Second canonical variable $U_2 = -0.036 * \mathbf{pop15} - 0.260 * \mathbf{pop75}$

Canonical Correlation Analysis in R:

```
$ycoef
             [,1]          [,2]          [,3]
sr   0.0084710221  3.337936e-02 -5.157130e-03
dpi  0.0001307398 -7.588232e-05  4.543705e-06
ddpi 0.0041706000 -1.226790e-02  5.188324e-02
```

- First canonical variable

$$V_1 = 0.008 * \mathbf{sr} + 0.0001 * \mathbf{dpi} + 0.004 * \mathbf{ddpi}$$

- Second canonical variable

$$V_1 = 0.003 * \mathbf{sr} - 0.000007 * \mathbf{dpi} - 0.001 * \mathbf{ddpi}$$

- Third canonical variable

$$V_1 = -0.0005 * \mathbf{sr} + 0.0000004 * \mathbf{dpi} + 0.005 * \mathbf{ddpi}$$

Since min(2,3) = 2, so ideally we need to consider only the first two canonical correlations.

Canonical Correlation Analysis in R:

```
$xcenter
  pop15   pop75
35.0896  2.2930
```

`$xcenter` gives the values used to adjust the `x` variables.

`xcenter` logical or numeric vector of length, describing any centering to be done on the `x` values before the analysis. If `TRUE` (default), subtract the column means.

Canonical Correlation Analysis in R:

```
$ycenter
       sr         dpi       ddpi
   9.6710 1106.7584     3.7576
```

`$ycenter` gives the values used to adjust the `y` variables.

`ycenter` logical or numeric vector of length, describing any centering to be done on the `y` values before the analysis. If `TRUE` (default), subtract the column means.

Canonical Correlation Analysis in R: Using CCA package

The function performs Canonical Correlation Analysis to highlight correlations between two data matrices.

Package CCA is required

```
install.packages("CCA")
library(CCA)
```

Usage

```
cc(X, Y)
```

Arguments

X numeric matrix (n * p), containing the X coordinates.

Y numeric matrix (n * q), containing the Y coordinates.

The canonical correlation analysis seeks linear combinations of the 'X' variables which are the most correlated with linear combinations of the 'Y' variables.

Canonical Correlation Analysis in R: Using `CCA` package

Value: A list containing the following components:

`cor` canonical correlations

`names` a list containing the names to be used for individuals and variables for graphical outputs

`xcoef` estimated coefficients for the '`X`' variables as returned by `cancor()`

`ycoef` estimated coefficients for the '`Y`' variables as returned by `cancor()`

`scores` a list returned by the internal function `comput()` containing individuals and variables coordinates on the canonical variates basis.

Canonical Correlation Analysis in R: Intercountry Life-Cycle Savings Data using CCA package

```
pop = LifeCycleSavings[, 2:3]

oec = LifeCycleSavings[, -(2:3)]

cc(pop, oec)
```

Canonical Correlation Analysis in R: Intercountry Life-Cycle Savings Data using CCA package

R Console

```
> cc(pop, oec)
$cor
[1] 0.8247966 0.3652762

$names
$names$Xnames
[1] "pop15" "pop75"

$names$Ynames
[1] "sr"   "dpi"  "ddpi"

$names$ind.names
 [1] "Australia"      "Austria"        "Belgium"        "Bolivia"
 [5] "Brazil"         "Canada"         "Chile"          "China"
 [9] "Colombia"       "Costa Rica"     "Denmark"        "Ecuador"
[13] "Finland"        "France"         "Germany"        "Greece"
[17] "Guatamala"      "Honduras"       "Iceland"        "India"
[21] "Ireland"        "Italy"          "Japan"          "Korea"
[25] "Luxembourg"     "Malta"          "Norway"         "Netherlands"
[29] "New Zealand"    "Nicaragua"      "Panama"         "Paraguay"
[33] "Peru"           "Philippines"    "Portugal"       "South Africa"
[37] "South Rhodesia" "Spain"          "Sweden"         "Switzerland"
[41] "Turkey"         "Tunisia"        "United Kingdom" "United States"
[45] "Venezuela"      "Zambia"         "Jamaica"        "Uruguay"
[49] "Libya"          "Malaysia"


$xcoef
             [,1]        [,2]
```

Canonical Correlation Analysis in R: Intercountry Life-Cycle Savings Data using CCA package

R Console

```
$xcoef
             [,1]       [,2]
pop15  0.06377599 -0.2535544
pop75 -0.34053260 -1.8221811

$ycoef
              [,1]          [,2]
sr   -0.0592971550  0.2336554912
dpi  -0.0009151786 -0.0005311762
ddpi -0.0291942000 -0.0858752749

$scores
$scores$xscores
                      [,1]        [,2]
Australia      -0.56253600  0.40390249
Austria        -1.47152544 -0.87332319
Belgium        -1.44772362 -1.03147293
Bolivia         0.64585407 -0.58905269
Brazil          0.95103425  0.86551308
Canada         -0.40457624 -0.16057787
Chile           0.62111144  0.55740907
China           1.16878601  0.50796273
Colombia        1.15651493 -0.68190575
Costa Rica      1.19304831 -1.08123466
Denmark        -1.23791620 -0.27758614
Ecuador         1.09119961 -0.83511633
Finland        -0.48857145  1.69786021
France         -1.45930966 -1.84294039
Germany        -1.11119865  1.06072429
Greece         -0.87874295  0.93055884
```

R Console

```
Greece         -0.87874295  0.93055884
Guatamala       1.18358828 -0.18609424
Honduras        1.36333825  0.02032415
Iceland        -0.33557620 -1.16539024
India           0.85064214  0.85175743
Ireland        -0.89660448 -2.46031003
Italy          -1.07829893  0.51703990
Japan          -0.38486053  2.74651367
Korea           0.89509245  0.83383808
Luxembourg     -1.33690279  0.75116267
Malta          -0.22287754  0.32393631
Norway         -1.05180046 -0.19175733
Netherlands    -0.98785900  0.88796621
New Zealand    -0.45678604 -0.96933925
Nicaragua       1.00339345 -0.54954583
Panama          0.91241030 -0.15606348
Paraguay        0.81170333  0.72072321
Peru            0.92534657 -0.46157725
Philippines     1.11184809 -0.69488593
Portugal       -0.58059799  0.53923234
South Africa   -0.19644195  0.82228337
South Rhodesia  0.06108731  2.21221207
Spain          -0.66521535  0.81212511
Sweden         -1.63569355 -0.63352441
Switzerland    -1.22912136  0.32265569
Turkey          0.94434558  0.09809587
Tunisia         1.07227152 -0.82338461
United Kingdom -1.49174087 -0.95175452
United States  -0.72389730 -0.73315394
Venezuela       1.19569390 -0.32950372
Zambia          1.23813259  0.58162543
```

Canonical Correlation Analysis in R: Intercountry Life-Cycle Savings Data using CCA package

```
R Console

Zambia            1.23813259  0.58162543
Jamaica           0.57631460 -0.50314665
Uruguay          -0.58926282  0.98656605
Libya             0.62443782 -1.77432308
Malaysia          1.32844252 -0.09502380

$scores$yscores
                       [,1]        [,2]
Australia      -1.197582618 -0.16236396
Austria        -0.514485534  0.33260994
Belgium        -1.126047497  0.28011657
Bolivia         1.175575433 -0.12494843
Brazil          0.132491457  0.88183195
Canada         -1.625987352 -1.08839364
Chile           0.975882427 -1.79030274
China           0.535391632  0.71855258
Colombia        1.057642399 -0.59695499
Costa Rica      0.543808669  0.67893036
Denmark        -1.704368254  0.91924174
Ecuador         1.155871496 -0.85121380
Finland        -0.635218480  0.01315294
France         -1.211470012  0.04020705
Germany        -1.397266488 -0.01731182
Greece          0.083021015  0.14211897
Guatamala       1.209216281 -0.92679302
Honduras        0.933602822  0.05262497
Iceland        -0.150891245 -2.15783934
India           1.036015081  0.57429510
Ireland        -0.106933726  0.43825829
Italy          -0.526164584  0.94515342
Japan          -0.945445589  2.20814404
```

```
R Console

Japan          -0.945445589  2.20814404
Korea           1.100359257 -1.02841475
Luxembourg     -1.205145263 -0.36666114
Malta          -0.009000439  1.25130272
Norway         -1.059225255 -0.45008336
Netherlands    -0.989337775  0.49151632
New Zealand    -0.349384397  0.20271478
Nicaragua       0.892846437 -0.02931829
Panama          0.807040147 -0.92369850
Paraguay        1.344342456 -1.08273725
Peru            0.557284192  1.34827236
Philippines     0.740722188  1.38450895
Portugal        0.206695290  0.61907452
South Africa    0.375656978  0.71988759
South Rhodesia  0.619330744  1.45344991
Spain           0.167542079  0.61909114
Sweden         -1.818231179 -1.75733210
Switzerland    -1.628446935  0.32307189
Turkey          0.948826794 -0.61162985
Tunisia         1.267754398 -0.92230572
United Kingdom -0.485816535 -0.66038997
United States  -2.486211894 -1.91878127
Venezuela       0.389454702  0.32762273
Zambia          0.318834488  2.47265581
Jamaica         0.591415820 -0.62589387
Uruguay         0.391732707  0.24124982
Libya           0.567959967 -0.77253487
Malaysia        1.046343696 -0.81375377

$scores$corr.X.xscores
                   [,1]        [,2]
```

Canonical Correlation Analysis in R: Intercountry Life-Cycle Savings Data using CCA package

```
R Console

Jamaica          0.591415820 -0.62589387
Uruguay          0.391732707  0.24124982
Libya            0.567959967 -0.77253487
Malaysia         1.046343696 -0.81375377

$scores$corr.X.xscores
            [,1]       [,2]
pop15  0.9829821 -0.1837015
pop75 -0.9697929 -0.2439299

$scores$corr.Y.xscores
            [,1]        [,2]
sr   -0.40500636  0.31259455
dpi  -0.78728255 -0.09633306
ddpi -0.03904398  0.05142128

$scores$corr.X.yscores
            [,1]        [,2]
pop15  0.8107603 -0.06710179
pop75 -0.7998819 -0.08910177

$scores$corr.Y.yscores
           [,1]       [,2]
sr   -0.4910379  0.8557760
dpi  -0.9545172 -0.2637266
ddpi -0.0473377  0.1407737
```