Brief look at Ch 10: Canonical Correlations

Two sets of variables relating to one group of individuals (or items):

eg. A-set: pre-college performance variables
B-set: college performance variables

Which linear combinations in A are related to linear combinations in B?

Answers in a sequence of pairs of linear combinations: to make them capture different connections between the two sets, we require a certain orthogonality in extracting the sequence of pairs.

We use covariances to summarize "relatedness".

Covariance matrix within A set
Covariance matrix within B set
Cross-covariances.

p 544 (10-1) and with one long data vector X= $\left[\frac{X^{(1)}}{X^{(2)}}\right]$ see (10-4)

$U = a' X^{(1)}$ scalar since a is p x 1 and X is p x 1
$V = b' X^{(2)}$ scalar

Then,

$Var(U) = a'\Sigma_{11}a$ scalar since a is p x 1, Σ_{11} is p x p.

$Var(V) = b'\Sigma_{22}b$ scalar

$Cov(U,V) = a'\Sigma_{12}b$ scalar since a is p x 1, Σ_{12} is p x q, b is q x 1. U

choose a,b to maximize

$$\mathbf{Corr(U,V) = a'\Sigma_{12}b \,/\, (a'\Sigma_{11}a \,.\, b'\Sigma_{22}b)^{1/2}}$$

U_1, V_1 first pair of canonical variates can be found from

$U_1 = e_1{}'\Sigma_{11}{}^{-1/2}X^{(1)}$ and $V_1 = f_1{}'\Sigma_{22}{}^{-1/2}X^{(2)}$

where e_1 and f_1 are eigenvectors of $\Sigma_{11}{}^{-1/2}\Sigma_{12}\Sigma_{22}{}^{-1}\Sigma_{21}\Sigma_{11}{}^{-1/2}$ and $\Sigma_{22}{}^{-1/2}\Sigma_{21}\Sigma_{11}{}^{-1}\Sigma_{12}\Sigma_{22}{}^{-1/2}$ resp. Moreover the maximized correlation is the largest eigenvalue of either matrix.

We need to review the meaning of $\Sigma^{-1/2}$ See (2-22) on p 67. As long as all the eigenvalues are positive, this square root matrix is defined by the eigenanalysis.

(Spectral decomposition of symmetric matrix depends only on eigenanalysis (2-16) p 62)

Once U_1 and V_1 are computed, we can seek U_2 and V_2 such that U_2 is uncorrelated with U_1 and V_2 is uncorrelated with V_1 and among such corr of U_2 and V_2 is as large as possible. It turns out this correlation is the second eigenvalue of the big matrix defined above, and U_2 and V_2 are computed from the eigenanalysis similarly to U_1 and V_1.

Note that we are not going to get an essentially different solution in this case if we use standardized variables. (p 548). This is different from PC or FA.

For sample estimates, just substitute sample estimates for the covariance matrices or correlation matrices. (Result 10.2 p 557)

Canonical Correlations given by eigenvalues.

Example 10.5 pp 559-564.

An example from R documentation:

```
> 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
Spain	11.77	27.74	2.87	768.79	4.35
Sweden	6.86	21.44	4.54	3299.49	3.01
Switzerland	14.13	23.49	3.73	2630.96	2.70
Turkey	5.13	43.42	1.08	389.66	2.96
Tunisia	2.81	46.12	1.21	249.87	1.13
United Kingdom	7.81	23.27	4.46	1813.93	2.01
United States	7.56	29.81	3.43	4001.89	2.45
Venezuela	9.22	46.40	0.90	813.39	0.53
Zambia	18.56	45.25	0.56	138.33	5.14
Jamaica	7.72	41.12	1.73	380.47	10.23
Uruguay	9.24	28.13	2.72	766.54	1.88
Libya	8.89	43.69	2.07	123.58	16.71
Malaysia	4.71	47.20	0.66	242.69	5.08

About this data set:

LifeCycleSavings {datasets}

Intercountry Life-Cycle Savings Data

Description

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

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.

Source

The data were obtained from Belsley, Kuh and Welsch (1980). They in turn obtained the data from Sterling (1977).

References

Sterling, Arnie (1977) Unpublished BS Thesis. Massachusetts Institute of Technology.

Belsley, D. A., Kuh. E. and Welsch, R. E. (1980) Regression Diagnostics. New York: Wiley.

Examples

```
require(stats)
pairs(LifeCycleSavings, panel = panel.smooth,
      main = "LifeCycleSavings data")
fm1 <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data =
LifeCycleSavings)
summary(fm1)
```

The Canonical Correlation analysis of this data:

```
>  pop <- LifeCycleSavings[, 2:3]
>  oec <- LifeCycleSavings[, -(2:3)]
>  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
```

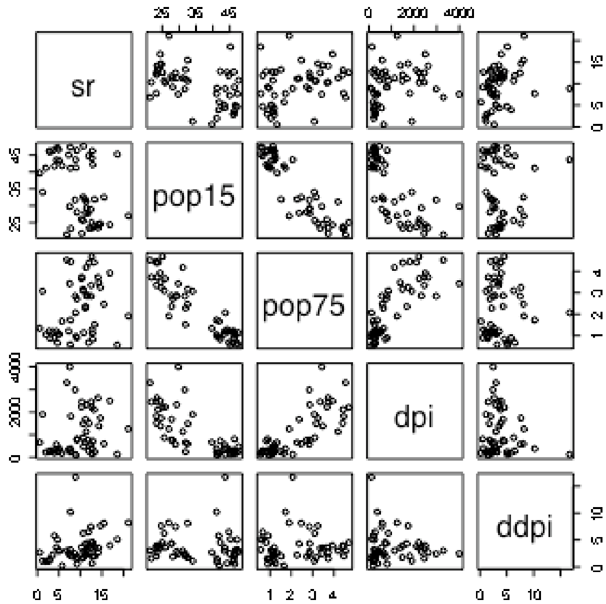
The first canonical correlation pair says that a country with lots of old people and not too many young will have high savings and disposable income!

Would we have found this from the combined correlation matrix itself?

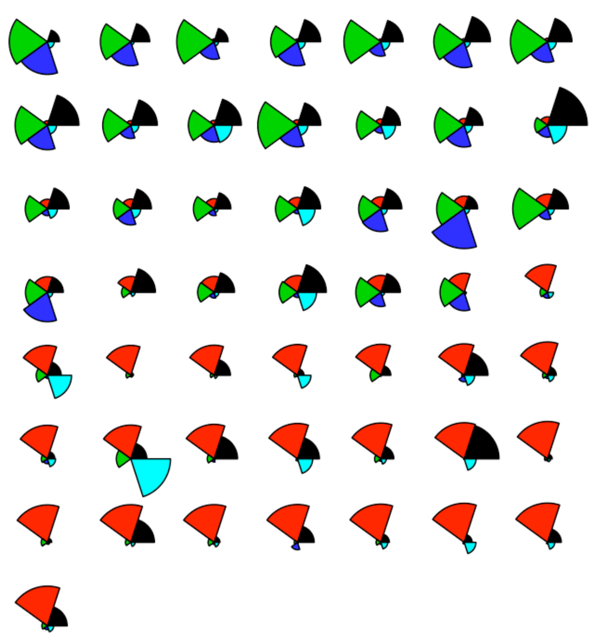
```
>  cor(LifeCycleSavings)
              sr       pop15       pop75        dpi        ddpi
sr     1.0000000 -0.45553809  0.31652112  0.2203589  0.30478716
pop15 -0.4555381  1.00000000 -0.90847871 -0.7561881 -0.04782569
pop75  0.3165211 -0.90847871  1.00000000  0.7869995  0.02532138
dpi    0.2203589 -0.75618810  0.78699951  1.0000000 -0.12948552
ddpi   0.3047872 -0.04782569  0.02532138 -0.1294855  1.00000000
```

Is it useful to test significance of CCs?

Other ways to look at this data?

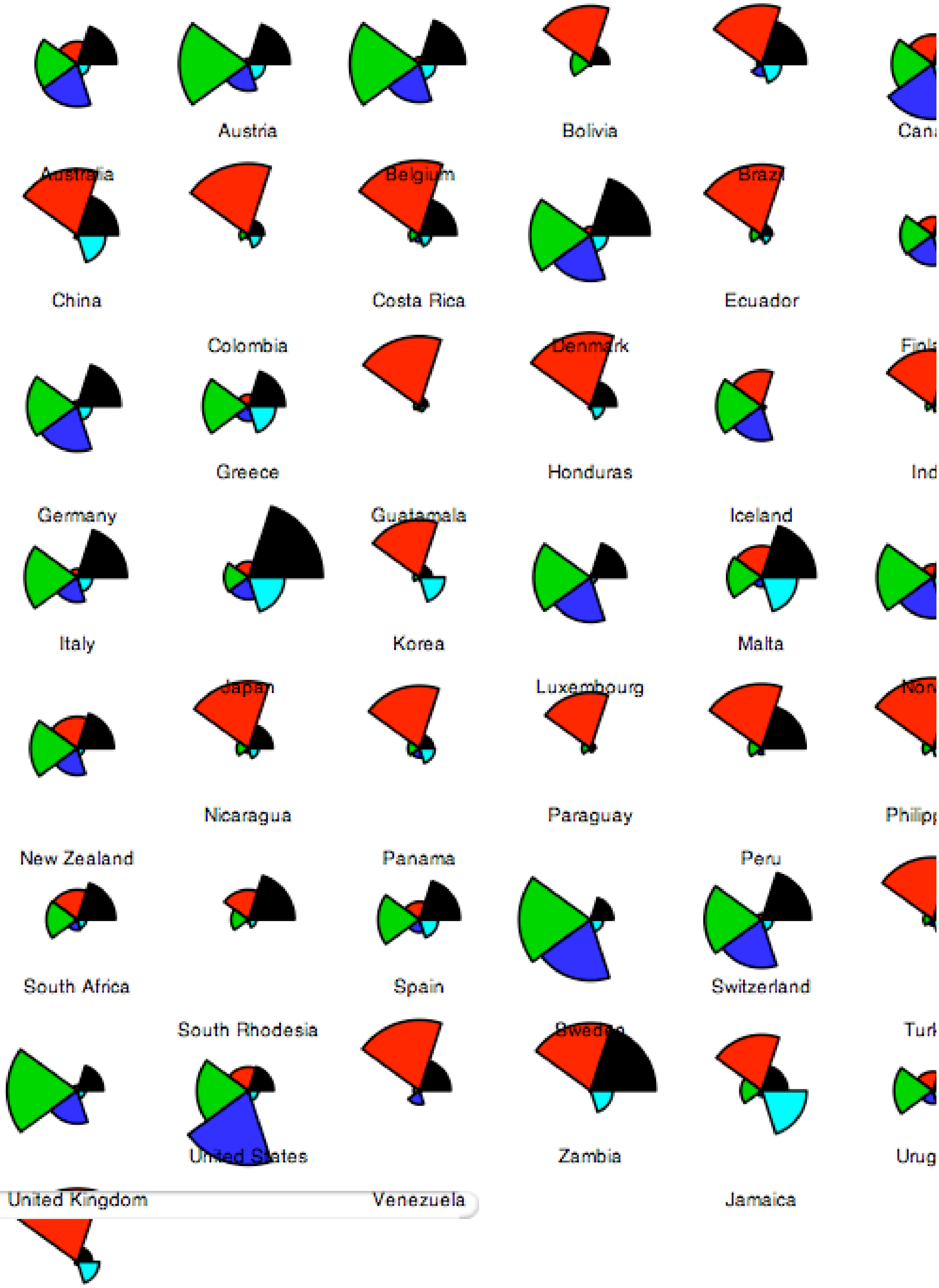


sr is Savings Rate, pop15 is % population under 15, pop75 % is population over 75, dpi is per capita disposable income, ddpi is % growth rate in dpi.
Note bimodality on "pop15" and outlier in "ddpi". Let's look at data sorted by pop15 (RED in plot below).



Black is Savings Rate, Green is % pop over 75, Blue is disposable income, Aqua is growth rate in disposable income. All data from 1960-1970.

Here is the unsorted plot with labels. Lybia appears to be the ddpi outlier.


Austria
Bolivia
Australia
Belgium
Brazil
China
Costa Rica
Ecuador
Colombia
Denmark
Greece
Honduras
Germany
Guatamala
Iceland
Italy
Korea
Malta
Japan
Luxembourg
Nicaragua
Paraguay
New Zealand
Panama
Peru
South Africa
Spain
Switzerland
South Rhodesia
Sweden
United States
Zambia
United Kingdom
Venezuela
Jamaica

Can graphical analysis tell us something that CC misses?

Note: Anyone want answers to exercises for your presentation?