Canonical Correlation Analysis II

Mengqian LU

Introduction

- Linear Regression: $Y = \beta\mathbf{X}$
  - Maximize R^2 to find a linear combination of $\mathbf{X}$ that explains the most variance in Y
  - $\mathbf{X}$ can be observed variables, or PCs from PCA, or Factors from FA etc.
  - Use PCs or Factors, we have reduced dimensionality and independence of predictors, which solves issues of collinearity. (Conventional FA with orthogonal rotation)
- What if now $Y \rightarrow \mathbf{Y}$? Multiple dependent variables
  - We need a way of measuring the linear relationship between two sets of variables → Canonical Correlation Analysis
  - An extension of multiple regression, to find linear combinations of both sets, that maximizes their correlations
  - Extraction of coefficients is similar to the process of finding PCs

Canonical Correlation Analysis

- Canonical Correlation Analysis extends the analysis on "Y~**X**" to 2+ continuous IVs with 2+ continuous DVs
- Key objective is to answer "How are the best linear combinations of predictors related to the best linear combinations of the DVs?"
  1. Linear combination of predictors
  2. Linear combination of predictants
  3. The relationship between the two linear combinations
  4. … more (coming up)

PCA finds the best combination in terms of maximum variance

- CCA is a multidimensional exploratory statistical method in the same vein as PCA, however CCA aims at the exploration of sample correlations between two sets of quantitative variables observed on the same experimental units, while PCA focus on dimension reduction of one data set through linear combination of variables.

When to use?

- Two sets of variables are related – at exploratory stage
- Your modeling assumptions rely on the theory that the predictor set as a whole inform the predictant set as a whole
- One set of two or more variables related longitudinally across two time points
- Examples: (1) An psychologist measure a number of aptitude variables and a number of achievement variables on some students and study the relation between "aptitude" and "achievement"; (2) An researcher study the relation between p measurements related to the yield of plants (e.g., height, dry weight, number of leaves) at each of n sites in a region and q variables related to the weather conditions at these sites. (e.g., average daily rainfall, humidity, hours of sunshine)

Formulation

- One pair of bases – the ones correspond to the largest CC
  - Consider the linear combinations $x = \boldsymbol{x}^T\widehat{w}_x$ and $y = \mathbf{y}^T\widehat{w}_y$ of the two variables respectively, this means we need to maximize:

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[\hat{\mathbf{w}}_x^T \mathbf{x}\mathbf{y}^T \hat{\mathbf{w}}_y]}{\sqrt{E[\hat{\mathbf{w}}_x^T \mathbf{x}\mathbf{x}^T \hat{\mathbf{w}}_x]E[\hat{\mathbf{w}}_y^T \mathbf{y}\mathbf{y}^T \hat{\mathbf{w}}_y]}}$$

$$= \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y}}.$$

  - The maximum of ρ w.r.t w_x and w_y is the maximum canonical correlation. The subsequent CCs are uncorrelated for different solutions, i.e.,

$$\begin{cases} E[x_i x_j] &= E[\mathbf{w}_{xi}^T \mathbf{x}\mathbf{x}^T \mathbf{w}_{xj}] = \mathbf{w}_{xi}^T \mathbf{C}_{xx} \mathbf{w}_{xj} = 0 \\ E[y_i y_j] &= E[\mathbf{w}_{yi}^T \mathbf{y}\mathbf{y}^T \mathbf{w}_{yj}] = \mathbf{w}_{yi}^T \mathbf{C}_{yy} \mathbf{w}_{yj} = 0 \\ E[x_i y_j] &= E[\mathbf{w}_{xi}^T \mathbf{x}\mathbf{y}^T \mathbf{w}_{yj}] = \mathbf{w}_{xi}^T \mathbf{C}_{xy} \mathbf{w}_{yj} = 0 \end{cases} \quad \text{for} \quad i \neq j.$$

  - The projections onto w_x and w_y, i.e. x and y, are the canonical variates

Layers of Analysis

1. Correlation between pairs of canonical variates
2. Loadings between predictors (IVs) and their canonical variates
3. Loadings between predictants (DVs) and their canonical variates
4. Adequacy
5. Communalities
6. Redundancy – between IVs and their canonical variates or between DVs and their canonical variates

Formulation (Cont'd)

- Let's create some single variable F that represent ***x*** and another single variable G for **y**, and F and G are linear combinations of X and Y, respectively: $F = f_1X_1 + f_2X_2 + \ldots + f_pX_p$; $G = g_1Y_1 + g_2Y_2 + \ldots + g_pY_p$.
- The first canonical correlation is the maximum correlation coefficient between F and G, for all F and G
- The correlation of our interest are now those between the canonical variates (linear combinations of **X** and **Y**)
- Terms for CCA: variables → (pairs of) canonical variates (both predictors and dependent variables) → canonical correlation
  - The maximum number of canonical variate pairs equals to the number of variables in the smaller set
  - There could be many canonical correlations, arranged in descending order, most of the time, only the first one or two are of the interest.

Formulation (Cont'd)

- Here are the reasons why Canonical Correlation Analysis is closely related to PCA: Properties of the linear combination pairs (F, G) Each $F_i = f_i^T \mathbf{X}$, $G_i = g_i^T \mathbf{Y}$ with (f_i^T, g_i^T) chosen so that they satisfy:

1. The F_i are mutually uncorrelated; i.e., Cov $(F_i,F_j) = 0$ for $i \neq j$
2. The G_i are mutually uncorrelated; i.e., $Cov(G_i,G_j) = 0$ for $i \neq j$
3. The correlation between F_i and G_i is R_i for I = 1, …s, where $R_1 > R_2 > \ldots > R_s$. The R_i are the canonical correlations
4. The F_i are uncorrelated with all G_j except G_i, i.e., $Cov(F_i,G_j) = 0$ for $i \neq j$.

Canonical Correlation Calculation

- Consider two random variables **X** and **Y** with zero mean. The total variance matrix is:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} = E\left[\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \right]$$

C is (p+q)× (p+q)

  - C_{xx} and C_{yy} are the within-sets covariance matrices
  - $C_{xy} = C_{yx}^T$ is the between-sets covariance matrix

- The canonical correlations between **X** and **Y** can be found by solving the eigenvalue equation:

$$\begin{cases} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\hat{\mathbf{w}}_x = \rho^2\hat{\mathbf{w}}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\hat{\mathbf{w}}_y = \rho^2\hat{\mathbf{w}}_y \end{cases}$$

  - The eigenvalues ρ^2 are the squared *canonical correlations* and the eigenvectors $\hat{w}_x$ and $\hat{w}_y$ are the normalized canonical correlation *basis vectors*.
  - Only one of the eigenvalue equations needs to be solved since the solutions are related by

$$\begin{cases} \mathbf{C}_{xy}\hat{\mathbf{w}}_y = \rho\lambda_x\mathbf{C}_{xx}\hat{\mathbf{w}}_x \\ \mathbf{C}_{yx}\hat{\mathbf{w}}_x = \rho\lambda_y\mathbf{C}_{yy}\hat{\mathbf{w}}_y \end{cases}$$

where

$$\lambda_x = \lambda_y^{-1} = \sqrt{\frac{\hat{\mathbf{w}}_y^T\mathbf{C}_{yy}\hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T\mathbf{C}_{xx}\hat{\mathbf{w}}_x}}$$

Canonical Correlation Calculation (Cont'd)

- Given the input correlation:

R_{xx}	R_{xy}
R_{yx}	R_{yy}

- The canonical correlation matrix (**R**): $\mathbf{R = R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy}}$
  - The product of four correlation matrices
- It also can be viewed as a product of regression coefficients for predicting **X** from **Y**, and **Y** from **X**
- Eigen-decompose the **R** to get the eigenvalues and take the square root → canonical correlations; the associated eigenvector to each eigenvalue is transformed into the coefficients that specify the linear combination making up a canonical variate

Canonical Coefficients

- Two sets of canonical coefficients (weights)
  - One set to combine the Xs
  - The other set to combine the Ys
  - Interpreted similarly to regression coefficients
- Test statistical significance of canonical correlations
  - The maximum number of canonical correlations equals to the number of variables in the smaller set
  - Not all will be statistically significant
  - Not all statistical significant CCs will be meaningful
  - This can be done using *CCP* package in R by *p.asym()*, with options for test statistics, commonly use "Wilks" stand for Wilk's Lambda to be approximated by Chi-square distribution – Bartlett's Chi Square test

Some Remarks on the tests

- Bartlett's Chi Square test (Wilk's Lambda)
  - Tests whether an eigenvalue and the ones that follow are significantly different than zero
  - It is possible this test would be significant even though a test for the correlation itself would not be
  - Is it really significant? – whether "significantly different from zero" is an interesting question?
  - It is simply a correlation coefficient, one should be more interested in the size of the effect, rather than whether it is different from zero (it always is)

Example: head lengths of 1st and 2nd sons (IAMA 3.13.1)

```
> headsize
      head1 breadth1 head2 breadth2
 [1,]   191      155   179      145
 [2,]   195      149   201      152
 [3,]   181      148   185      149
 [4,]   183      153   188      149
 [5,]   176      144   171      142
 [6,]   208      157   192      152
 [7,]   189      150   190      149
 [8,]   197      159   189      152
 [9,]   188      152   197      159
[10,]   192      150   187      151
[11,]   179      158   186      148
[12,]   183      147   174      147
[13,]   174      150   185      152
[14,]   190      159   195      157
[15,]   188      151   187      158
[16,]   163      137   161      130
[17,]   195      155   183      158
[18,]   186      153   173      148
[19,]   181      145   182      146
[20,]   175      140   165      137
[21,]   192      154   185      152
[22,]   174      143   178      147
[23,]   176      139   176      143
[24,]   197      167   200      158
[25,]   190      163   187      150
```

```
headsize.std <- sweep(headsize, 2, apply(headsize, 2, sd), FUN="/")
R <- cor(headsize.std)
R11 <- R[1:2,1:2]
R12 <- R[3:4,1:2]
R21 <- R[1:2,3:4]
R22 <- R[3:4,3:4]
R11.inv <- solve(R11)
R22.inv <- solve(R22)

# compute E1 and E2
E1 <- R11.inv %*% R12 %*% R22.inv %*% R21
E2 <- R22.inv %*% R21 %*% R11.inv %*% R12

# compute eigenvalues and eigenvectors of E1 and E2:
eigen(E1)
eigen(E2)
```

$$\mathbf{E}_1 = \mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}, \quad \mathbf{E}_2 = \mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$$

E_1 is p-by-p; E_2 is q-by-q.
p is the dimension of the predictors set
q is the dimension of the predictants set.

```
>eigen(E1)
$values
[1] 0.621781555 0.002887785
$vectors [,1] [,2]
[1,] -0.6947269 -0.7089828
[2,] -0.7192736 0.7052258
> eigen(E2)
$values
[1] 0.621781555 0.002887785
$vectors [,1] [,2]
[1,] 0.7424369 -0.7039264
[2,] 0.6699160 0.7102729
```

The canonical correlations express the association between the x and y variables after removal of the within-set correlation.

F_1 = -0.69*head1 - 0.72*breadth1;

F_2 = -0.71*head1 + 0.71*breadth1;

G_1 = +0.74*head2 + 0.70*breadth2;

G_2 = -0.70*head2 + 0.71*breadth2.

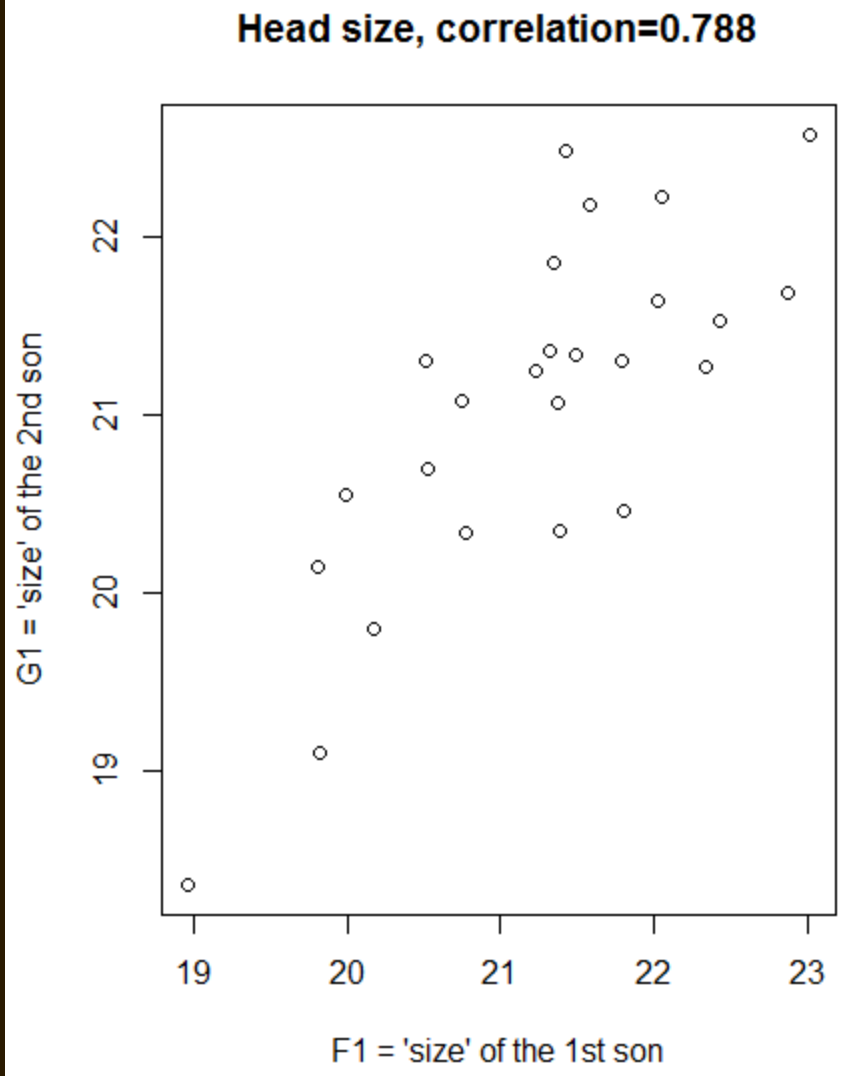
 # Since we work with standardized data from the beginning , so
 # no correction of scaling of CC vectors is needed:
 # OPTIONAL:

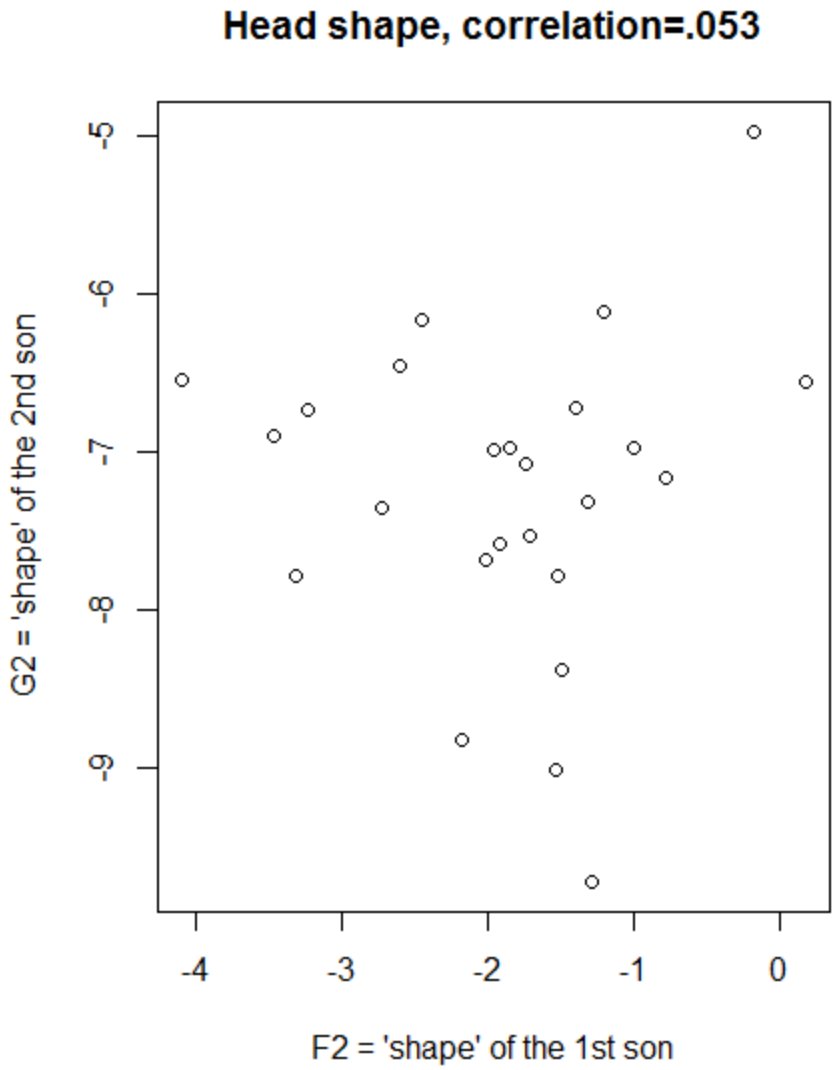
```
# compute the canonical correlation vectors:
(a1 <- eigen(E1)$vectors[,1])
(a2 <- eigen(E1)$vectors[,2])
(b1 <- eigen(E2)$vectors[,1])
(b2 <- eigen(E2)$vectors[,2])
# correct the scaling of the canonical correlation vectors:
(a1 <- -1 * a1 / sqrt(t(a1) %*% R11 %*% a1))
(a2 <- -1 * a2 / sqrt(t(a2) %*% R11 %*% a2))
(b1 <- b1 / sqrt(t(b1) %*% R22 %*% b1))
(b2 <- -1 * b2 / sqrt(t(b2) %*% R22 %*% b2))
# check scaling:
t(a1) %*% R11 %*% a1
t(a2) %*% R11 %*% a2
t(b1) %*% R22 %*% b1
t(b2) %*% R22 %*% b2
```

```
# compute canonical correlation variables:
F1 <- headsize.std[,1:2] %*% a1   # size first son
F2 <- headsize.std[,1:2] %*% a2   # shape first son
G1 <- headsize.std[,3:4] %*% b1   # size second son
G2 <- headsize.std[,3:4] %*% b2   # shape second son

# check covariance matrix:
round(var(cbind(F1,F2,G1,G2)),3)

# plot canonical correlation variables:
par(mfrow=c(1,2))
plot(F1,G1,main="Head size, correlation=0.788", xlab="F1 =
'size' of the 1st son", ylab="G1 = 'size' of the 2nd son")
plot(F2,G2, main="Head shape, correlation=.053", xlab="u2 =
'shape' of the 1st son", ylab="G2 = 'shape' of the 2nd son")
```


Head size, correlation=0.788
G1 = 'size' of the 2nd son
F1 = 'size' of the 1st son
19
20
21
22
23


Head shape, correlation=.053
G2 = 'shape' of the 2nd son
F2 = 'shape' of the 1st son
-5
-6
-7
-8
-9
-4
-3
-2
-1
0

Loadings (structure coefficients)

- Questions to assist your interpretations:
  1. How well do the variates on either side relate to their own original variables?
  2. What is the associations between a variable and its respective canonical variate? – correlations
- Recall that: the pair of canonical variates with best correlation might not exactly interpretable
- Coefficients are for the computation of the variates, loadings are more for expressing the relationships of the variables to the construct (canonical variates)
  - ❑ Canonical communality coefficient
  - ❑ Canonical variate adequacy coefficient

Coefficients

- Canonical communality coefficient
  - Sum of the squared structure coefficients (loadings) across selected* canonical variates for a given variable
  - Measures how much variance of that original variable is explained or reproduced by the canonical variates

  * If looking at all variates, it shall equal one, typically we often want to check those retained for interpretation or prediction.
- Canonical variate adequacy coefficient
  - Average of all the squared structure coefficients (loadings) for one set of variables with respect to their canonical variate
  - Measures how well a resulted canonical variate represents the variance in that set of original variables.

Redundancy

- Questions to assist checking redundancy:
  1. How strongly do the original variables in one set relate to the canonical variates on the other side?
  2. How much of the average proportion of variance of the original variables in one set may be predicted from the variables in the other set – High redundancy suggests potentially high ability to predict
     - Product of the mean squared structure coefficient, i.e., the canonical adequacy coefficient for a canonical variate times the squared canonical correlation coefficient.
     - Canonical correlation reflects the % of variance in the predictant canonical variate explained by the predictor canonical variate

➢ Redundancy has to do with assessing the effectiveness of the canonical analysis in capturing the variance of the original variables