Canonical Correlation Analysis

Max Turgeon

STAT 4690–Applied Multivariate Analysis

Introduction

- Canonical Correlation Analysis (CCA) is a dimension reduction method that is similar to PCA, but where we simultaneously reduce the dimension of **two** random vectors $\mathbf{Y}$ and $\mathbf{X}$.
- Instead of trying to explain overall variance, we try to explain the covariance $\mathrm{Cov}(\mathbf{Y}, \mathbf{X})$.
  - Note that this is a measure of **association** between $\mathbf{Y}$ and $\mathbf{X}$.
- Examples include:
  - Arithmetic speed and power $(\mathbf{Y})$ and reading speed and power $(\mathbf{X})$
  - College performance metrics $(\mathbf{Y})$ and high-school achievement metrics $(\mathbf{X})$

Population model i

- Let $\mathbf{Y}$ and $\mathbf{X}$ be p- and q-dimensional random vectors, respectively.
    - We will assume that $p \leq q$.
- Let μ_Y and μ_X be the mean of $\mathbf{Y}$ and $\mathbf{X}$, respectively.
- Let Σ_Y and Σ_X be the covariance matrix of $\mathbf{Y}$ and $\mathbf{X}$, respectively, and let $\Sigma_{XY} = \Sigma_{YX}^T$ be the covariance matrix $\mathrm{Cov}(\mathbf{Y}, \mathbf{X})$.
    - Assume Σ_Y and Σ_X are positive definite.
- Note that Σ_{YX} has pq entries, corresponding to all covariances between a component of $\mathbf{Y}$ and a component of $\mathbf{X}$.

Population model ii

- **Goal of CCA**: Summarise Σ_{YX} with p numbers.
    - These p numbers will be called the *canonical correlations*.

Dimension reduction i

- Let $U = a^T\mathbf{Y}$ and $V = b^T\mathbf{Y}$ be linear combinations of $\mathbf{Y}$ and $\mathbf{X}$, respectively.
- We have:
  - $\mathrm{Var}(U) = a^T\Sigma_Y a$
  - $\mathrm{Var}(V) = b^T\Sigma_X b$
  - $\mathrm{Cov}(U, V) = a^T\Sigma_{YX} b$.
- Therefore, we can write the correlation between U and V as follows:

$$\mathrm{Corr}(U, V) = \frac{a^T\Sigma_{YX} b}{\sqrt{a^T\Sigma_Y a}\sqrt{b^T\Sigma_X b}}.$$

Dimension reduction ii

- We are looking for vectors $a \in \mathbb{R}^p, b \in \mathbb{R}^q$ such that $\mathrm{Corr}(U, V)$ is **maximised**.

Definitions

- The *first pair of canonical variates* is the pair of linear combinations U_1, V_1 with unit variance such that $\mathrm{Corr}(U_1, V_1)$ is maximised.
- The k-**th pair of canonical variates** is the pair of linear combinations U_k, V_k with unit variance such that $\mathrm{Corr}(U_k, V_k)$ is maximised among all pairs that are uncorrelated with the previous $k - 1$ pairs.
- When U_k, V_k is the k-th pair of canonical variates, we say that $\rho_k = \mathrm{Corr}(U_k, V_k)$ is the k-th *canonical correlation*.

Derivation of canonical variates i

- Make a change of variables:
  - $\tilde{a} = \Sigma_Y^{1/2} a$
  - $\tilde{b} = \Sigma_X^{1/2} b$
- We can then rewrite the correlation:

$$\begin{aligned}\mathrm{Corr}(U, V) &= \frac{a^T \Sigma_{YX} b}{\sqrt{a^T \Sigma_Y a}\sqrt{b^T \Sigma_X b}} \\ &= \frac{\tilde{a}^T \Sigma_Y^{-1/2} \Sigma_{YX} \Sigma_X^{-1/2} \tilde{b}}{\sqrt{\tilde{a}^T \tilde{a}}\sqrt{\tilde{b}^T \tilde{b}}}.\end{aligned}$$

Derivation of canonical variates ii

- Let $M = \Sigma_Y^{-1/2}\Sigma_{YX}\Sigma_X^{-1/2}$. We have

$$\max_{a,b} \mathrm{Corr}(a^T\mathbf{Y}, b^T\mathbf{Y}) \Longleftrightarrow \max_{\tilde{a},\tilde{b}:\|\tilde{a}\|=1,\|\tilde{b}\|=1} \tilde{a}^T M \tilde{b}$$

- The solution to this maximisation problem involves the **singular value decomposition** of M.
- Equivalently, it involves the **eigendecomposition** of MM^T, where

$$MM^T = \Sigma_Y^{-1/2}\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1/2}.$$

CCA: Main theorem i

- Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of $\Sigma_Y^{-1/2}\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1/2}$.
  - Let $e_1, \ldots, e_p$ be the corresponding eigenvector with unit norm.
- Note that $\lambda_1 \geq \cdots \geq \lambda_p$ are also the p largest eigenvalues of

$$M^T M = \Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}\Sigma_X^{-1/2}.$$

  - Let $f_1, \ldots, f_p$ be the corresponding eigenvectors with unit norm.

CCA: Main theorem ii

- Then the k-th pair of canonical variates is given by

$$U_k = e_k^T \Sigma_Y^{-1/2} \mathbf{Y}, \qquad V_k = f_k^T \Sigma_X^{-1/2} \mathbf{X}.$$

- Moreover, we have

$$\rho_k = \mathrm{Corr}(U_k, V_k) = \sqrt{\lambda_k}.$$

Some vocabulary

1. **Canonical directions**: $(e_k^T \Sigma_Y^{-1/2}, f_k^T \Sigma_X^{-1/2})$
2. **Canonical variates**: $(U_k, V_k) = \left(e_k^T \Sigma_Y^{-1/2} \mathbf{Y}, f_k^T \Sigma_X^{-1/2} \mathbf{X}\right)$
3. **Canonical correlations**: $\rho_k = \sqrt{\lambda_k}$

Example i

```r
Sigma_Y <- matrix(c(1, 0.4, 0.4, 1), ncol = 2)
Sigma_X <- matrix(c(1, 0.2, 0.2, 1), ncol = 2)
Sigma_YX <- matrix(c(0.5, 0.3, 0.6, 0.4), ncol = 2)
Sigma_XY <- t(Sigma_YX)

rbind(cbind(Sigma_Y, Sigma_YX),
      cbind(Sigma_XY, Sigma_X))
```

Example ii

```
##      [,1] [,2] [,3] [,4]
## [1,]  1.0  0.4  0.5  0.6
## [2,]  0.4  1.0  0.3  0.4
## [3,]  0.5  0.3  1.0  0.2
## [4,]  0.6  0.4  0.2  1.0
```

Example iii

```
library(expm)
sqrt_Y <- sqrtm(Sigma_Y)
sqrt_X <- sqrtm(Sigma_X)
M1 <- solve(sqrt_Y) %*% Sigma_YX %*% solve(Sigma_X)%*%
  Sigma_XY %*% solve(sqrt_Y)

(decomp1 <- eigen(M1))
```

Example iv

```
## eigen() decomposition
## $values
## [1] 0.5457180317 0.0009089525
##
## $vectors
##              [,1]        [,2]
## [1,] -0.8946536  0.4467605
## [2,] -0.4467605 -0.8946536
```

```
decomp1$vectors[,1] %*% solve(sqrt_Y)
```

Example v

```
##              [,1]       [,2]
## [1,] -0.8559647 -0.2777371
```

```
M2 <- solve(sqrt_X) %*% Sigma_XY %*% solve(Sigma_Y)%*%
  Sigma_YX %*% solve(sqrt_X)

decomp2 <- eigen(M2)
decomp2$vectors[,1] %*% solve(sqrt_X)
```

```
##             [,1]      [,2]
## [1,] 0.5448119 0.7366455
```

Example vi

```r
sqrt(decomp1$values)
```

```
## [1] 0.73872731 0.03014884
```

Sample CCA

- Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ and $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be random samples, and arrange them in $n \times p$ and $n \times q$ matrices $\mathbb{Y}, \mathbb{X}$, respectively.
    - Note that both sample sizes are equal.
    - Indeed, we assume that $(\mathbf{Y}_i, \mathbf{X}_i)$ are sampled jointly, i.e. on the **same** experimental unit.
- Let $\bar{\mathbf{Y}}$ and $\bar{\mathbf{X}}$ be the sample means.
- Let S_Y and S_X be the sample covariances.
- Define

$$S_{YX} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\mathbf{Y}_i - \bar{\mathbf{Y}}\right) \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^T .$$

Sample CCA: Main theorem i

- Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ be the eigenvalues of $S_Y^{-1/2} S_{YX} S_X^{-1} S_{XY} S_Y^{-1/2}$.
    - Let $\hat{e}_1, \ldots, \hat{e}_p$ be the corresponding eigenvector with unit norm.
- Note that $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ are also the p largest eigenvalues of

$$S_X^{-1/2} S_{XY} S_Y^{-1} S_{YX} S_X^{-1/2}.$$

    - Let $\hat{f}_1, \ldots, \hat{f}_p$ be the corresponding eigenvectors with unit norm.

Sample CCA: Main theorem ii

- Then the k-th pair of *sample* canonical variates is given by

$$\hat{U}_k = \mathbb{Y}S_Y^{-1/2}\hat{e}_k, \qquad \hat{V}_k = \mathbb{X}S_X^{-1/2}\hat{f}_k.$$

- Moreover, we have that $\hat{\rho}_k = \sqrt{\hat{\lambda}_k}$ is the sample correlation of $\hat{U}_k$ and $\hat{V}_k$.

Example (cont'd) i

```r
# Let's generate data
library(mvtnorm)
Sigma <- rbind(cbind(Sigma_Y, Sigma_YX),
               cbind(Sigma_XY, Sigma_X))

YX <- rmvnorm(100, sigma = Sigma)
Y <- YX[,1:2]
X <- YX[,3:4]

decomp <- cancor(x = X, y = Y)
```

Example (cont'd) ii

```r
U <- Y %*% decomp$ycoef
V <- X %*% decomp$xcoef

diag(cor(U, V))
```

```
## [1] 0.6927462 0.1136006
```

```r
decomp$cor
```

```
## [1] 0.6927462 0.1136006
```

Example i

```r
library(tidyverse)
library(dslabs)

X <- olive %>%
  select(-area, -region) %>%
  as.matrix

Y <- olive %>%
  select(region) %>%
  model.matrix(~ region - 1, data = .)
```

Example ii

```r
head(unname(Y))
```

```
##      [,1] [,2] [,3]
## [1,]    0    0    1
## [2,]    0    0    1
## [3,]    0    0    1
## [4,]    0    0    1
## [5,]    0    0    1
## [6,]    0    0    1
```

Example iii

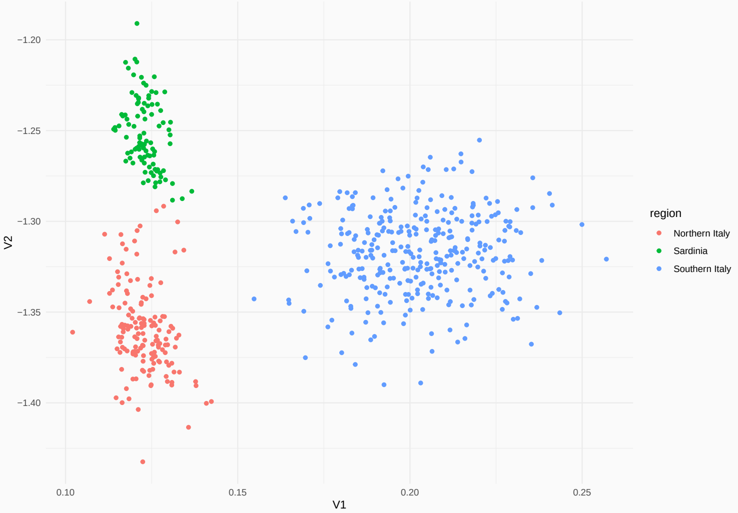
```r
decomp <- cancor(X, Y)

V <- X %*% decomp$xcoef
```

Example iv

```
data.frame(
  V1 = V[,1],
  V2 = V[,2],
  region = olive$region
) %>%
  ggplot(aes(V1, V2, colour = region)) +
  geom_point() +
  theme_minimal()
```

Example v



−1.20
−1.25
−1.30
−1.35
−1.40
V2
0.10
0.15
0.20
0.25
V1
region
Northern Italy
Sardinia
Southern Italy

Comments i

- The main difference between CCA and Multivariate Linear Regression is that CCA treats $\mathbb{Y}$ and $\mathbb{X}$ *symmetrically*.
- As with PCA, you can use CCA and the covariance matrix or the correlation matrix.
  - The latter is equivalent to performing CCA on the standardised variables.
- Note that sample CCA involves inverting the sample covariance matrices S_Y and S_X:
  - This means we need to assume $p, q < n$.
  - In general, this is what drives most of the performance (or lack thereof) of CCA.

Comments ii

    - There may be gains in efficiency by directly estimating the inverse covariance.
- When one of the two datasets $\mathbb{Y}$ or $\mathbb{X}$ represent indicators variables for a categorical variables (cf. the olive dataset), CCA is equivalent to **Linear Discriminant Analysis**.
    - To learn more about this method, see a course/textbook on Statistical Learning.

Interpreting the population canonical variates i

- To help interpretating the canonical variates, let's go back to the population model.
- Define

$$
\begin{aligned}
A &= \begin{pmatrix} e_1^T \Sigma_Y^{-1/2} & \cdots & e_p^T \Sigma_Y^{-1/2} \end{pmatrix}^T, \\
B &= \begin{pmatrix} f_1^T \Sigma_X^{-1/2} & \cdots & f_p^T \Sigma_X^{-1/2} \end{pmatrix}^T.
\end{aligned}
$$

- In other words, both A and B are $p \times p$, and their *rows* are the canonical directions.

Interpreting the population canonical variates ii

- Using this notation, we can get all canonical variates using one linear transformation:

$$\mathbf{U} = A\mathbf{Y}, \qquad \mathbf{Y} = B\mathbf{X}.$$

- We then have

$$\mathrm{Cov}(\mathbf{U}, \mathbf{Y}) = \mathrm{Cov}(A\mathbf{Y}, \mathbf{Y}) = A\Sigma_Y.$$

- Since $\mathrm{Cov}(\mathbf{U}) = I_p$, we have

$$\mathrm{Corr}(U_k, Y_i) = \mathrm{Cov}(U_k, \sigma_i^{-1} Y_i),$$

where σ_i^2 is the variance of Y_i.

Interpreting the population canonical variates iii

- If we let D_Y be the diagonal matrix whose i-th diagonal element is $\sigma_i = \sqrt{\mathrm{Var}(Y_i)}$, we can write

$$\mathrm{Corr}(\mathbf{U}, \mathbf{Y}) = A\Sigma_Y D_Y^{-1}.$$

- Using similar computations, we get

$$\begin{aligned} \mathrm{Corr}(\mathbf{U}, \mathbf{Y}) &= A\Sigma_Y D_Y^{-1}, & \mathrm{Corr}(\mathbf{V}, \mathbf{Y}) &= B\Sigma_{XY} D_Y^{-1}, \\ \mathrm{Corr}(\mathbf{U}, \mathbf{X}) &= A\Sigma_{YX} D_X^{-1}, & \mathrm{Corr}(\mathbf{V}, \mathbf{X}) &= B\Sigma_X D_X^{-1}. \end{aligned}$$

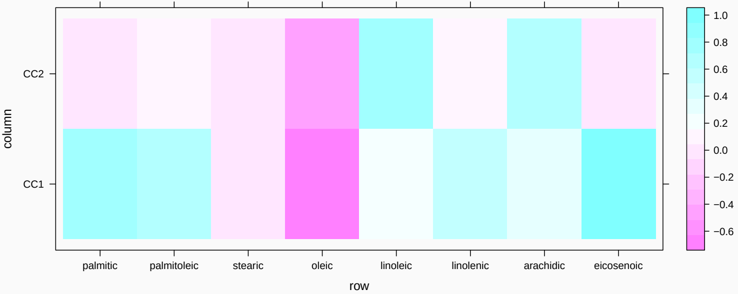
- **These quantities** (and their sample counterparts) **give us information about the contribution of the original variables to the canonical variates**.

Example i

```r
# Let's go back to the olive data
decomp <- cancor(X, Y)
V <- X %*% decomp$xcoef
colnames(V) <- paste0("CC", seq_len(8))

library(lattice)
levelplot(cor(X, V[,1:2]))
```

Example ii

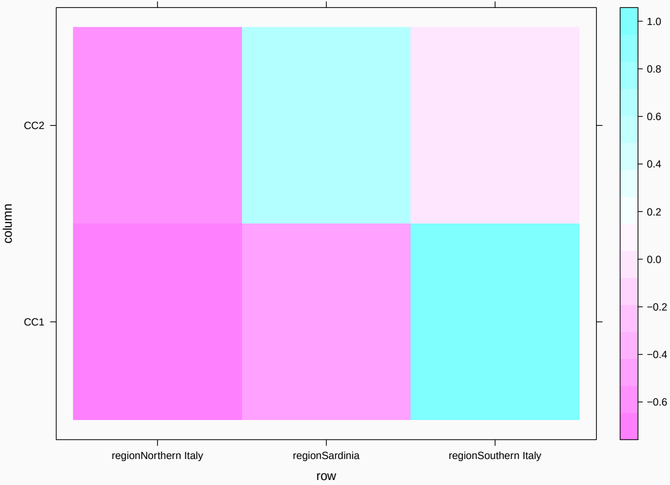


CC2
CC1
column
palmitic
palmitoleic
stearic
oleic
linoleic
linolenic
arachidic
eicosenoic
row
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Example iii

```r
levelplot(cor(Y, V[,1:2]))
```

Example iv



CC2
CC1
column
regionNorthern Italy
regionSardinia
regionSouthern Italy
row
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6