STAT 3690 Lecture 29

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Apr 08, 2022

CCA (Con'd)

- Sample version
  - $(\mathbf{Y}_1, \mathbf{X}_1), \ldots, (\mathbf{Y}_n, \mathbf{X}_n) \overset{\text{iid}}{\sim} (\mathbf{Y}, \mathbf{X})$
    * $\mathbf{Y}_i$ and $\mathbf{X}_i$ jointly sampled
    * $p \leq q < n$
  - $n \times p$ matrix $\mathbb{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_n]^\top$ and $n \times q$ matrix $\mathbb{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]^\top$
  - Sample covariance matrices
    * $\mathbf{S}_\mathbf{Y} = (n-1)^{-1} \sum_i (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^\top$
    * $\mathbf{S}_\mathbf{X} = (n-1)^{-1} \sum_i (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$
    * $\mathbf{S}_{\mathbf{YX}} = \mathbf{S}_{\mathbf{XY}}^\top = (n-1)^{-1} \sum_i (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$
  - Vocabulary
    * (The kth pair of) sample canonical directions: $(\hat{\boldsymbol{a}}_k \in \mathbb{R}^p, \hat{\boldsymbol{b}}_k \in \mathbb{R}^q)$
    * (The kth pair of) sample canonical variates: $(\mathbb{Y}_C \hat{\boldsymbol{a}}_k, \mathbb{X}_C \hat{\boldsymbol{b}}_k)$
    * (The kth) canonical correlation: $\hat{\rho}_k$
  - Goal: find $\hat{\boldsymbol{a}}_k$ and $\hat{\boldsymbol{b}}_k$, $k = 1, \ldots, r \leq p$, to maximize

$$\hat{\rho}_k = \frac{\hat{\boldsymbol{a}}_k^\top \mathbf{S}_{\mathbf{YX}} \hat{\boldsymbol{b}}_k}{\sqrt{\hat{\boldsymbol{a}}_k^\top \mathbf{S}_\mathbf{Y} \hat{\boldsymbol{a}}_k} \sqrt{\hat{\boldsymbol{b}}_k^\top \mathbf{S}_\mathbf{X} \hat{\boldsymbol{b}}_k}}$$

    subject to
    * $\hat{\boldsymbol{a}}_k^\top \mathbf{S}_\mathbf{Y} \hat{\boldsymbol{a}}_k = 1$
    * $\hat{\boldsymbol{b}}_k^\top \mathbf{S}_\mathbf{X} \hat{\boldsymbol{b}}_k = 1$
    * $\hat{\boldsymbol{a}}_k^\top \mathbf{S} \hat{\boldsymbol{a}}_\ell = 0$, $\ell = 1, \ldots, k-1$
    * $\hat{\boldsymbol{a}}_k^\top \mathbf{S}_{\mathbf{YX}} \hat{\boldsymbol{b}}_\ell = 0$, $\ell = 1, \ldots, k-1$
    * $\hat{\boldsymbol{b}}_k^\top \mathbf{S}_\mathbf{X} \hat{\boldsymbol{b}}_\ell = 0$, $\ell = 1, \ldots, k-1$
    * $\hat{\boldsymbol{b}}_k^\top \mathbf{S}_{\mathbf{XY}} \hat{\boldsymbol{a}}_\ell = 0$, $\ell = 1, \ldots, k-1$
  - Solution: Let $\widehat{\mathbf{M}} = \mathbf{S}_\mathbf{Y}^{-1/2} \mathbf{S}_{\mathbf{YX}} \mathbf{S}_\mathbf{X}^{-1/2}$
    * $\hat{\rho}_k = \sqrt{\hat{\lambda}_k}$ is the kth largest singular value of $\widehat{\mathbf{M}}$
      · $\hat{\lambda}_k$: the kth largest eigenvalue of $\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\top$ (or $\widehat{\mathbf{M}}^\top\widehat{\mathbf{M}}$)
    * $\hat{\boldsymbol{a}}_k = \mathbf{S}_\mathbf{Y}^{-1/2} \hat{\boldsymbol{e}}_k$
      · $\hat{\boldsymbol{e}}_k$: the left-singular vector corresponding to the kth largest singular value of $\widehat{\mathbf{M}}$, i.e., the eigenvector corresponding to the kth largest eigenvalue of $\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\top$
    * $\hat{\boldsymbol{b}}_k = \mathbf{S}_\mathbf{X}^{-1/2} \hat{\boldsymbol{f}}_k$
      · $\hat{\boldsymbol{f}}_k$: the right-singular vector corresponding to the kth largest singular value of $\widehat{\mathbf{M}}$, i.e., the eigenvector corresponding to the kth largest eigenvalue of $\widehat{\mathbf{M}}^\top\widehat{\mathbf{M}}$

- Example: olive oil data
  - 572 olive oils
  - 10 features
    * `region` indicates the general region (in Italy) of origin.
    * `area` details the area of Italy.
    * Remaining variables are continuous valued and measure the percentage composition of 8 different fatty acids
  - Interested in the correlations between the region of origin and the fatty acid measurements
    * $\mathbb{Y} \in \mathbb{R}^{572\times 3}$ an indicator matrix, i.e., each row of Y indicates the region with a 1 and otherwise has 0
    * $\mathbb{X} \in \mathbb{R}^{572\times 8}$ contains the 8 fatty acid measurements

```r
options(digits=4)
Y = as.matrix(dslabs::olive[,3:4])
X = as.matrix(dslabs::olive[,5:10])
p = ncol(Y)
q = ncol(X)

# by definition
S_Y = cov(Y)
S_X = cov(X)
S_YX = cov(Y, X)
S_Y_sqrt = expm::sqrtm(S_Y)
S_X_sqrt = expm::sqrtm(S_X)
M = solve(S_Y_sqrt) %*% S_YX %*% solve(S_X_sqrt)
decomp1 = svd(M)
decomp1$d
A1 = solve(S_Y_sqrt) %*% decomp1$u
B1 = solve(S_X_sqrt) %*% decomp1$v
YA1 = scale(Y, scale=F) %*% A1 # canonical variates
XB1 = scale(X, scale=F) %*% B1 # canonical variates

# by cancor (not recommended if you try to know the specific values of canonical directions/variates)
decomp2 = cancor(x=X, y=Y)
decomp2$cor
A2 = decomp2$ycoef # identical to A1 up to a constant
B2 = decomp2$xcoef[, 1:min(p,q)] # identical to B1 up to a constant
YA2 = scale(Y, scale=F) %*% A2 # canonical variates
XB2 = scale(X, scale=F) %*% B2 # canonical variates

# comparison
A1/A2
B1/B2
t(A1) %*% S_Y %*% A1
t(B1) %*% S_X %*% B1
t(A2) %*% S_Y %*% A2
t(B2) %*% S_X %*% B2
```

- Proportion of explained correlation
  - Determining r, the number of pairs of canonical variates to retain
  - $p \times r$ matrix $\text{corr}(\mathbf{Y}, \mathbf{A}_r^\top \mathbf{Y})$ and $q \times r$ matrix $\text{corr}(\mathbf{X}, \mathbf{B}_r^\top \mathbf{X})$
    * The correlation matrix between $\mathbf{Y}$ (or $\mathbf{X}$) and canonical variates
    * $\mathbf{A}_r = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r]$ and $\mathbf{B}_r = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_r]$

  - $\|\text{corr}(\mathbf{Y}, \mathbf{A}_r^\top \mathbf{Y})\|_F^2/p$ and $\|\text{corr}(\mathbf{X}, \mathbf{B}_r^\top \mathbf{X})\|_F^2/q$
    * Proportion of explained correlation of $\mathbf{Y}$ (or $\mathbf{X}$)
    * $\|\cdot\|_F^2$: squared Frobenius norm, i.e., sum of squared entries

```
Y = as.matrix(dslabs::olive[,3:6])
X = as.matrix(dslabs::olive[,7:10])
p = ncol(Y)
q = ncol(X)

S_Y = cov(Y)
S_X = cov(X)
S_YX = cov(Y, X)
S_Y_sqrt = expm::sqrtm(S_Y)
S_X_sqrt = expm::sqrtm(S_X)
M = solve(S_Y_sqrt) %*% S_YX %*% solve(S_X_sqrt)
decomp1 = svd(M)
A1 = solve(S_Y_sqrt) %*% decomp1$u
B1 = solve(S_X_sqrt) %*% decomp1$v
YA1 = scale(Y, scale=F) %*% A1
XB1 = scale(X, scale=F) %*% B1

cor(Y, YA1)
cor(X, XB1)

cumsum(colMeans(cor(Y, YA1)^2))
cumsum(colMeans(cor(X, XB1)^2))
```

- Interpreting canonical variates
  - $\text{corr}(\mathbf{Y}, \mathbf{A}_r^\top \mathbf{Y})$
  - $\text{corr}(\mathbf{X}, \mathbf{B}_r^\top \mathbf{X})$
  - $\text{corr}(\mathbf{Y}, \mathbf{B}_r^\top \mathbf{X})$
  - $\text{corr}(\mathbf{X}, \mathbf{A}_r^\top \mathbf{Y})$

```
colnames(U1) = paste0("Y'a", seq_len(q))
colnames(V1) = paste0("X'b", seq_len(q))
cor(Y, YA1)
cor(Y, XB1)
cor(X, YA1)
cor(X, XB1)
```