STAT 730 Chapter 10: Canonical correlation analysis

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 730: Multivariate Data Analysis

Basic idea

Start with

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right) = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Want to know how correlated a linear combination of $\mathbf{x}$ can be with a linear combination of $\mathbf{y}$. Original paper:

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28, 321-377.

Useful to see what aspects are common in two sets of variables. Often used in psychological testing, as is factor analysis (coming up).

Maximizing correlation

The idea, to maximize correlation among subsets of variables, is similar to PCA (maximizing variability among all variables), but the motivation and math is a bit different.

Recall Cauchy-Schwartz for vectors is $|\mathbf{a}'\mathbf{b}| \leq ||\mathbf{a}||\ ||\mathbf{b}||$ with equality only when $\mathbf{b} = k\mathbf{a}$ for some $k \in \mathbb{R}$.

Assume $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$. Let $\mathbf{x} \sim (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, $\mathbf{y} \sim (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, and $C(\mathbf{x}, \mathbf{y}) = \boldsymbol{\Sigma}_{12}$. Let $\eta = \mathbf{a}'\mathbf{x}$ and $\phi = \mathbf{b}'\mathbf{y}$. The correlation between η and ϕ is

$$\rho = \frac{\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a}\mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b}}}.$$

Define $\mathbf{c} = \boldsymbol{\Sigma}_{11}^{1/2}\mathbf{a}$ and $\mathbf{d} = \boldsymbol{\Sigma}_{22}^{1/2}\mathbf{b}$ yielding

$$\rho = \frac{\mathbf{c}'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\mathbf{d}}{||\mathbf{c}||\ ||\mathbf{d}||}.$$

Obtaining the first CCA vectors

Our old friend Cauchy-Schwartz bounds the numerator

$$\left|[\mathbf{c}'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}]\mathbf{d}\right| \leq ||\mathbf{c}'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}||\ ||\mathbf{d}||,$$

with equality only when $\mathbf{d}$ and $\mathbf{c}'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$ coincide. Coupled with the correlation formula, the inequality implies

$$\rho^2 = \frac{\mathbf{c}'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{c}}{\mathbf{c}'\mathbf{c}}.$$

The maximization result from Chapter 4 gives that the maximum occurs when $\mathbf{c}$ is the e-vector corresponding to the largest e-value of $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$. The maximized correlation is the largest e-value of this matrix.

k CCA vectors

So we have that $\mathbf{a} = \mathbf{\Sigma}_{11}^{-1/2}\gamma_{(1)}$ where $\gamma_{(1)}$ is the e-vector associated with the largest e-value of $\mathbf{\Sigma}_{11}^{-1/2}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1/2}$, and $\mathbf{b} = \mathbf{\Sigma}_{22}^{-1/2}\mathbf{d} = \mathbf{\Sigma}_{22}^{-1/2}\gamma'_{(1)}\mathbf{\Sigma}_{11}^{-1/2}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1/2}$ maximizes the correlation. One can show (The. A.6.2) that this implies $\mathbf{a}$ is the "largest" e-vector of $\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ and $\mathbf{b}$ is the "largest" e-vector of $\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}$. Note that *these matrices have the same non-zero e-values* $\lambda_1 \geq \cdots \geq \lambda_k > 0$ where $k = \text{rank}(\mathbf{\Sigma}_{12})$; $k = \min\{p, q\}$ if $\mathbf{\Sigma} > 0$.

Write $\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} = [\boldsymbol{\alpha}_{(1)} \cdots \boldsymbol{\alpha}_{(q)}]\mathbf{\Lambda}_1[\boldsymbol{\alpha}_{(1)} \cdots \boldsymbol{\alpha}_{(q)}]'$ and $\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} = [\boldsymbol{\beta}_{(1)} \cdots \boldsymbol{\beta}_{(p)}]\mathbf{\Lambda}_2[\boldsymbol{\beta}_{(1)} \cdots \boldsymbol{\beta}_{(p)}]'$. Then $(\boldsymbol{\alpha}_{(i)}, \boldsymbol{\beta}_{(i)})$ are the ith raw canonical correlation vectors for $\mathbf{x}$ and $\mathbf{y}$, where $i = 1, \ldots, k$, s.t. $||\boldsymbol{\alpha}_{(i)}|| = ||\boldsymbol{\beta}_{(i)}|| = 1$.

Normalized CCA vectors & data sample versions

Often the CCA vectors are normalized so that $V(\eta_i) = V(\phi_i) = 1$. They are no longer SLC's.

Then $(\mathbf{a}_i, \mathbf{b}_i) = \left(\frac{\boldsymbol{\alpha}_{(i)}}{\sqrt{\boldsymbol{\alpha}'_{(i)} \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}_{(i)}}}, \frac{\boldsymbol{\beta}_{(i)}}{\sqrt{\boldsymbol{\beta}'_{(i)} \boldsymbol{\Sigma}_{22} \boldsymbol{\beta}_{(i)}}} \right)$ are the normalized versions. These are what are used in MKB but not what R produces with the `cancor` function.

Also $\eta_i = \mathbf{a}'_i \mathbf{x}$ and $\phi_i = \mathbf{b}'_i \mathbf{y}$ are the ith canonical correlation variables, and $\rho_i = \sqrt{\lambda_i}$ is the ith canonical correlation coefficient. The empirical version replaces $\boldsymbol{\Sigma}$ with $\mathbf{S}$ and puts hats on everything else, e.g.
$\mathbf{S}_{11}^{-1} \mathbf{S}_{12} \mathbf{S}_{22}^{-1} \mathbf{S}_{21} = [\hat{\boldsymbol{\alpha}}_{(1)} \cdots \hat{\boldsymbol{\alpha}}_{(q)}] \hat{\boldsymbol{\Lambda}}_1 [\hat{\boldsymbol{\alpha}}_{(1)} \cdots \hat{\boldsymbol{\alpha}}_{(q)}]'$ and
$\mathbf{S}_{22}^{-1} \mathbf{S}_{21} \mathbf{S}_{11}^{-1} \mathbf{S}_{12} = [\hat{\boldsymbol{\beta}}_{(1)} \cdots \hat{\boldsymbol{\beta}}_{(p)}] \hat{\boldsymbol{\Lambda}}_2 [\hat{\boldsymbol{\beta}}_{(1)} \cdots \hat{\boldsymbol{\beta}}_{(p)}]'$. Also $\hat{\rho}_i = \sqrt{\hat{\lambda}_i}$.

Open/closed book exams

```
library(bootstrap)
data(scor)
S=cov(scor)
S11=S[1:2,1:2]; S22=S[3:5,3:5]; S12=S[1:2,3:5]; S21=S[3:5,1:2]
e1=eigen(solve(S11)%*%S12%*%solve(S22)%*%S21)
e2=eigen(solve(S22)%*%S21%*%solve(S11)%*%S12)
# canonical vectors for 2 closed book exams
a=-1*e1$vectors[,1]/sqrt(e1$vectors[,1]%*%S11%*%e1$vectors[,1])
a # weighted average
e1$vectors[,2]/sqrt(e1$vectors[,2]%*%S11%*%e1$vectors[,2])
# canonical vectors for 3 open book exams
b=e2$vectors[,1]/sqrt(e2$vectors[,1]%*%S22%*%e2$vectors[,1])
b # weighted average
e2$vectors[,2]/sqrt(e2$vectors[,2]%*%S22%*%e2$vectors[,2])

eta=a%*%t(scor[,1:2]); phi=b%*%t(scor[,3:5])
plot(eta,phi)
cor(t(rbind(eta,phi)))
sqrt(e2$values[1])
```

Comments

- CCA finds the two linear combinations that maximize correlation, then two more linear combinations orthogonal to the first that maximize correlation, etc.
- $C(\eta_i, \eta_j) = \delta_{ij}$ and $C(\phi_i, \phi_j) = \delta_{ij}$.
- Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)'$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)'$. Then $C(\boldsymbol{\eta}, \boldsymbol{\phi}) = \boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k})$.
- As in the canonical variables in MANOVA and the principal loadings in PCA, the $(\boldsymbol{\alpha}_{(i)}, \boldsymbol{\beta}_{(i)})$ are often interpretable SLCs.

Comments

- Recall that $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ appeared in Chapter 5 when we wanted to test $\mathbf{x}$ ind. $\mathbf{y}$ assuming normality, i.e. $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ (which implies $\rho_1 = 0$). The LRT is $-2\log\lambda = -n\log\prod_{i=1}^{k}(1-\hat{\lambda}_i) = -n\log\prod_{i=1}^{k}(1-\hat{\rho}_i^2)$.
- CCA is carried out in R using the `cancor` function; it's not much harder to do it from scratch as in the last slide. The `cancor` functions normalizes the $(\mathbf{a}_i, \mathbf{b}_i)$ differently than MKB and I cannot figure out how. There is also the CCA package on CRAN.
- This discussion is much shorter and a bit different than MKB; more details are in Chapter 10, but we've hit the highlights.

Categorical predictors

In Section 10.4 (pp. 293–295) MKB consider categorical predictors. For a categorical measurement with g levels, they advocate placing $g-1$ zero-one dummy variables into either **x** or **y** are proceeding as usual.

I will point out one caveat of this approach: the correlation between two binary variables is bounded away from unity. Let (x, y) be jointly distributed $P(x = i, y = j) = \pi_{ij}$ where $\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11} = 1$. Then $x \sim \text{Bern}(\pi_{1+})$ and $y \sim \text{Bern}(\pi_{+1})$. One can show

$$\rho(x, y) = \frac{\pi_{11}\pi_{00} - \pi_{10}\pi_{01}}{\sqrt{\pi_{1+}\pi_{0+}\pi_{+1}\pi_{+0}}} \leq \frac{\min\{\pi_{+1}\pi_{0+},\ \pi_{1+}\pi_{+0}\}}{\sqrt{\pi_{1+}\pi_{0+}\pi_{+1}\pi_{+0}}}.$$

An interesting question is "Is the correlation between linear combinations of binary variables bounded away from unity?"

The correlation between two dichotomous variables is called the ϕ-coefficient.