3/4/2010 36-402/608 ADA-II H. Seltman

Handout #15: Canonical Correlation Analysis

1. Overview

   (a) CCA (canonical correlation analysis) is a technique for finding relationships (correlations) between two high dimensional sets of variables (e.g., an explanatory set and an outcome set). It finds linear combinations in both sets (generically called X and Y) that are cross-correlated. The linear combinations are analogous to those in PCA. If X can be thought to *cause* Y, the first canonical X variable is the "best predictor" and the first canonical Y variable is the "most predictable criterion."

   (b) CCA is useful if the new variables are simpler to understand and work with than the original variables and are meaningful.

   (c) Number of cases should be at least 5 or 10 time the number of variables.

   (d) p-values depend on multivariate normality.

2. Simulated example

```
X = matrix(rnorm(500),100,5)
Y = cbind(3*X[,1]-3*X[,4]+2*X[,5]+rnorm(100),
          rnorm(100),
          X[,2]+X[,3]+rnorm(100))

# Cross-correlation of the original data:
round(cor(X,Y),2)
#       [,1]  [,2]  [,3]
# [1,]  0.70  0.02  0.00
# [2,]  0.01  0.01  0.49
# [3,] -0.05 -0.02  0.68
# [4,] -0.71 -0.15 -0.06
# [5,]  0.41  0.06 -0.08

# Perform the canonical correlation analysis:
CCAfake = cancor(X,Y)
names(CCAfake)
# [1] "cor"     "xcoef"   "ycoef"   "xcenter" "ycenter"

# Correlation of the (new) canonical variables:
CCAfake$cor
# [1] 0.98113389 0.83166627 0.09960887
```

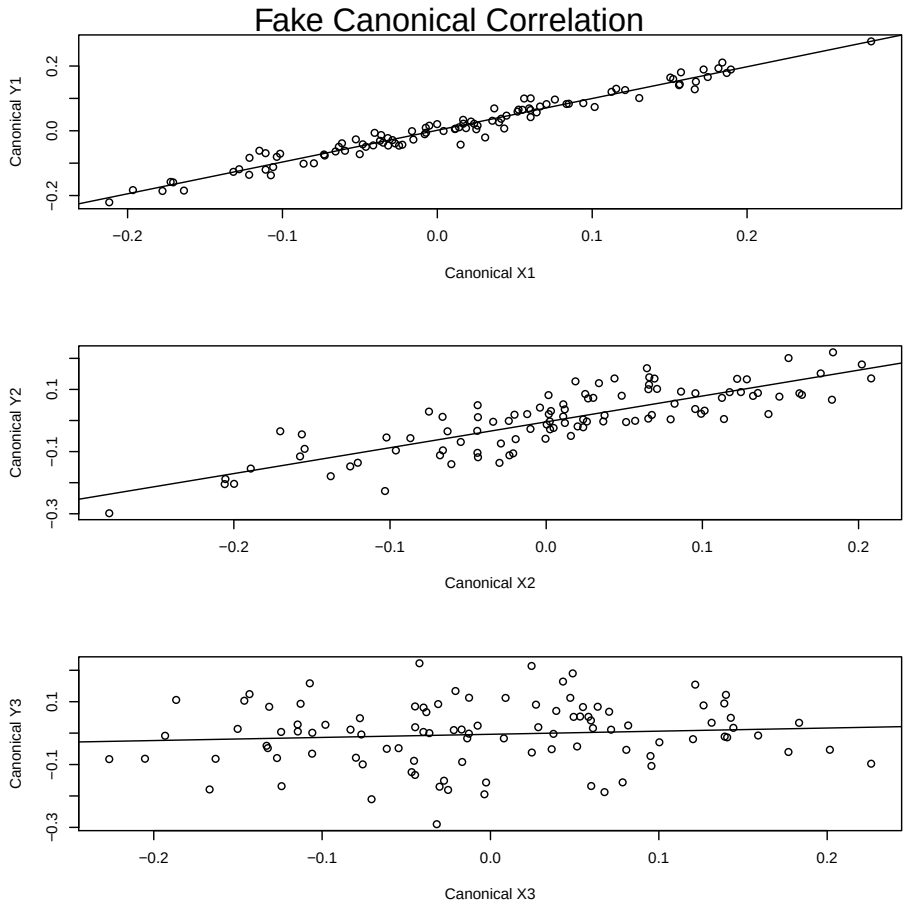
```
library(CCP) # needed for p-value (p.asym)
p.asym(CCAfake$cor, nrow(X), ncol(X), ncol(Y))
# Wilks' Lambda, using F-approximation (Rao's F):
#                 stat     approx df1      df2   p.value
# 1 to 3:  0.01140993 68.7692402  15 254.3729 0.0000000
# 2 to 3:  0.30527197 18.8303645   8 186.0000 0.0000000
# 3 to 3:  0.99007807  0.3140026   3  94.0000 0.8152149

# Loadings (weights) for the ``x'' variables:
round(CCAfake$xcoef,3)
#        [,1]   [,2]   [,3]   [,4]   [,5]
# [1,]  0.057  0.009  0.077 -0.006  0.024
# [2,] -0.003  0.055 -0.003 -0.085  0.000
# [3,] -0.006  0.079  0.008  0.056 -0.013
# [4,] -0.063 -0.006  0.078 -0.022 -0.031
# [5,]  0.039 -0.006 -0.004 -0.019 -0.106

# Loadings (weights) for the ``y'' variables:
round(CCAfake$ycoef,3)
#        [,1]  [,2]   [,3]
# [1,]  0.020 0.001  0.003
# [2,]  0.003 0.002 -0.101
# [3,] -0.004 0.055 -0.001

# Cross-correlation of the canonical variables:
round(cor(X%*%CCAfake$xcoef, Y%*%CCAfake$ycoef),3)
#       [,1]  [,2] [,3]
# [1,] 0.981 0.000  0.0
# [2,] 0.000 0.832  0.0
# [3,] 0.000 0.000  0.1
# [4,] 0.000 0.000  0.0
# [5,] 0.000 0.000  0.0
```

```
# Plot the canonical variables against each other:
par(mfrow=c(3,1), oma=c(0,0,1.5,0), mai=c(1,0.8,0,0.2))
for (i in 1:3) {
  XX = X%*%CCAfake$xcoef[,i]
  YY = Y%*%CCAfake$ycoef[,i]
  plot(XX, YY, xlab=paste("Canonical X",i,sep=""),
               ylab=paste("Canonical Y",i,sep=""))
  abline(lm(YY~XX))
}
mtext("Fake Canonical Correlation", outer=T, cex=1.4)
```



3. Breakout and Discussion