Canonical Correlations

Assignments:
Use the data set `Lab3_close.csv` from the course web site; it consists of daily closing prices for ten stocks. (Alternatively, you can use a data set of your choice.) The goal of this assignment is to find the best possible pair of “duplicate portfolios,” that is two groups of stocks with the maximal possible first canonical correlation. In the lab we found two groups with canonical correlation (CC) of 0.8. Try to beat this result or find another pair of groups with a comparable CC.

Reports: Printed reports are due on Thursday, April 23, 2009.

Report preparation: Consider each report as a mini-paper. It should NOT be long, but it should provide a reader with all background information about the data, problem, and methods you are using. Review the necessary theoretical material (use formulas if needed), describe the data. Do not insert the R-output in your report; instead, summarize it in tables or text in a nice readable form. If you feel some parts of the output should be included, put them in Appendix. Put your name on the title page. Illustrations should support your conclusions and make it easier to read a report.

- Codes used for class presentations are available on the course Web page: http://wolfweb.unr.edu/homepage/zal/STAT755_Lab.htm
#=================================================#
#                   STAT 755                      #
#             Canonical Correlations              #
#=================================================#

# Install libraries ... 
library(MASS)     # ... for Multivariate Normal Distribution
library(car)      # ... for ellipse plots

# Example 1: X=N1+N2+2*N3+2*N4
Sigma=diag(c(1,1,1,1))
N<-mvrnorm(n=1000,c(0,0,0,0),Sigma)
a=as.matrix(c(1,1,2,2))
X<-N%*%a
cor(X,N)
cancor(X,N)

# Example 2: X=N1+2*N2+3*N3+4*N4, Y=4*N1+3*N2+2*N3+N4
Sigma=diag(c(1,1,1,1))
N<-mvrnorm(n=1000,c(0,0,0,0),Sigma)
a=matrix(c(1,2,3,4,4,3,2,1),4,2)
X<-N%*%a
Y<-N[,,(1:2)]
cor(X,Y)
cc<-cancor(X,Y)
cc
X1<-X%*%cc$xcoef[,1]
Y1<-Y%*%cc$ycoef[,1]
X2<-X%*%cc$xcoef[,2]
Y2<-Y%*%cc$ycoef[,2]
SA(cbind(X1,Y1))
SA(cbind(X2,Y2))
# Real data example
# (Closing prices on 10 stocks)

# read the data table
T<-read.table('Lab3_close.csv',sep=',',header=TRUE)
len<-dim(T)[1]          # length
Itime<-seq(len,1,by=-1) # inverse index for plots

time<-1989+(31+28+31+30+4)/365.25+seq(1,len)/250

names(T)     # names of variables
P<-T[,seq(2,11)]    # remove the dates
P<-log10(P)

# Remove long-term trend
#=========================  
PD<-P*0
for (i in seq(1,10))
{
  t<-ts(P[,i],frequency=365)
  s<-stl(t,s.window=250)
  PD[,i]=t-s$time.series[,2]
}

# CC
#============================
X<-PD[,c(1,2,3,4,5)]
Y<-PD[,c(6,7,8,9,10)]

cor(X,Y)
cancor(X,Y)
# Function that illustrates spectral decomposition
# and statistical distance ellipses

SA <- function(X, add=FALSE, data.plot=TRUE) {
  # Vector of means
  n <- dim(X)[1]
  ones <- matrix(rep(1, n), ncol=1)
  mu <- as.vector(t(X) %*% ones / n)

  # Variance
  Sigma <- var(X)
  e <- eigen(Sigma)
  par(bg='yellow')
  ellipse(mu, Sigma, 3, add=add, xlim=range(X), ylim=range(X))
  ellipse(mu, Sigma, 2, add=TRUE)
  ellipse(mu, Sigma, 1, add=TRUE)
  if (data.plot)
    points(X[,1], X[,2], pch=20, col=4)
    arrows(mu[1], mu[2],mu[1]+e$vectors[1,1]*sqrt(e$values[1]),
           mu[2]+e$vectors[2,1]*sqrt(e$values[1]), length=.1, col='green', lwd=2)
    arrows(mu[1], mu[2],mu[1]+e$vectors[1,2]*sqrt(e$values[2]),
           mu[2]+e$vectors[2,2]*sqrt(e$values[2]), length=.1, col='green', lwd=2)
  e
}