Factor Analysis (FA)

Assignments:
Use the table `salespeople.txt` from the course website or a data set of your choice. Data description can be found in Chapter 9, Ex. 9.19, p. 538.

1) Find the correlation matrix for the data, study it, and comment.
2) Check data normality (univariate and multivariate).
3) Perform FA with different number of common factors (try 1 to 5) and different rotations, analyze loadings and scores, find main groups of variables, find single-factor variables (if any), check scores' normality. Find the optimal number of factors (consider the easiness of factor interpretation, number of factors, the proportion of the total variance explained, the residual matrix, etc.) Discuss and justify your choice. [This assignment will be evaluated on the basis of your discussion and justification. Support each of your statements by figures or numerical summaries of analysis.]
4) Perform FA with two factors and promax rotation. Report the loadings, estimated communalities, specific variances, and proportion of the variance explained by each factor. Interpret the factor solution. Use factor score scatterplot to discuss possible outliers. Report the residual matrix \( \rho - LL' - \Psi \).

Reports: Printed reports are due on Thursday, April 18, 2013.

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

Simple example, with a block covariance matrix
Sigma<-matrix(c(10,4,5,4,3,2,5,2,3),3,3)
X<-mvrnorm(n=1000,c(0,0,0),Sigma)
Y<-mvrnorm(n=1000,c(0,0,0),2*Sigma)
Z<-cbind(X,Y)

FA
fa<-factanal(Z,factors=2,rotation='promax',scores='Bart')

Residuals
L<-fa$loadings
U<-fa$uniquenesses
res<-cor(Z)-(L%*%t(L)+diag(U))
res*(res>.01)

Plot loadings
# (notice two clusters)
par(bg='yellow')
plot(loadings(fa),pch=19,col='blue',
     xlab='Loadings for Factor 1',
     ylab='Loadings for Factor 2')
grid(col='black')

Check factor scores for Normality
# (scores are Normal)
qqnorm(fa$scores[,1])
qqnorm(fa$scores[,2])

Plot scores
plot(fa$scores,pch=19,col='blue',
     xlab='Factor 1 score',
     ylab='Factor 2 score')
grid(col='black')
SA(fa$scores)
#=======================================================
# 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
It ime<-seq(len,1,by=-1)  # inverse index for plots

# time in years (start date is 5/4/89)
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]
}

# Detrended FA
#==========================================================
fa<-factanal(PD,factors=2,rotation='promax',scores='Bart')
fa

# Plot loadings
#===============================================
par(bg='yellow')
plot(loadings(fa)[,1],loadings(fa)[,2],pch=19,col='blue',
xlab='Loadings for Factor 1',
ylab='Loadings for Factor 2',
xlim=c(-.2,1),ylim=c(-.2,1))
grid(col='black')

# Scores scatterplot
#===============================================
plot(fa$scores,pch=19,col='blue',
xlab='Factor 1 score',
ylab='Factor 2 score')
grid(col='black')

SA(fa$scores)

# Check factor scores for Normality
# (scores are Normal)
#===============================================
qqnorm(fa$scores[,1])
qqnorm(fa$scores[,2])
# Residuals
#==============================================
L<-fa$loadings
U<-fa$uniquenesses
res<- cor(PD)-(L%*%t(L)+diag(U))
res*(res>.5)

# 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)
}

Page 4