Assignment 2 (3 pages)

Statistics 32950-24620 (Spring 2024)

Due 9 am Tuesday, April 2nd.

Requirements

- Your answers should be typed (allowing clear handwritten between typed texts for complex math formulas).
  Started with your name, Assignment 1, STAT 32950 or 24620; saved as `LastnameFirstnamePset2.pdf (or ...hw2.pdf)`,
  and uploaded to Gradescope under either 329Pset2 or 246Pset2.
  Make sure to **submit to the correct course number** you registered, and tag the pages for each question.
- When you use R (or others) to solve problems such as Question 1 in this assignment, select only relevant parts of the output, edit, then insert in your writing.
- You may discuss approaches with others. However the assignment should be devised and written by yourself.
  Capturing contents from other sources then pasting as your answers are not allowed.

Exercise assignments: (Topics correspond to parts in chapters 3, 4, 8 and 9 in Johnson and Wichern)

1. (*Properties of multivariate normal distribution*)
   Assume random vector $\boldsymbol{X} \sim N_3(\boldsymbol{\mu}, \Sigma)$, with $\boldsymbol{\mu} = \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 4 & 0 & -1 \\ 0 & 7 & 0 \\ -1 & 0 & 2 \end{bmatrix}$.

   (a) Answer the following questions, show the steps of your derivation.

   i. $Y_1 = X_1 + 3X_2 - 2X_3$. Are X_1 and Y_1 independent?

   ii. $Y_2 = X_1 - X_2 + 4X_3$. Are X_1 and Y_2 independent?

   (b) Show the formula you use in the following derivation.

   i. Derive $\mathbb{E}(X_3|X_1 = x_1)$.

   ii. Derive $Var(X_3|X_1 = x_1)$.

   (c) Specify the conditional distribution of X_1 given that $X_3 = x_3$, write the density function $f_{X_1|X_3=x_3}(x_1)$.

   (d) Specify the conditional distribution of X_1 given that $(X_2, X_3) = (x_2, x_3)$. What is $f_{X_1|(X_2,X_3)=(x_2,x_3)}(x_1)$?

2. (*Basic Principal Component Analysis*)

   Use the same dataset ladyrun24.dat on national track records for women as used in Assignment 1. The nominal variable "`Country`" should be excluded in numerical calculations.

   (a) Determine the first two principal components for the scaled data with variable variance = 1, express the two principal components as linear combination of the scaled variables. State the meaning of the (scaled) variables that the principal components consist of. (Recall that a PC variable is unique up to a multiple of ± 1.)

   (b) Compare the two principal components PC1 and PC2 with the eigenvectors of the sample correlation matrix (which you obtained in Question 1(f) in Assignment 1).

   (c) What are the percentages of total (scaled data) sample variation explained by the first and second principal components you obtained in (a)?

   (d) Plots and interpretations.

   i. Now every observation has its coordinates in the space of principle components (PC1, ..., PC7) obtained using the scaled data (of variable variance 1). Construct a two-dimensional scatterplot of the 54 observations in the (PC1, PC2) plane.
   Compare the values of PC1 scores of the countries of the atheletes with their approximate ranking (i.e. goodness or levels of overall performance) in track records.

ii. Now every variable has its loading coefficients on each of the principle components (PC1, ..., PC7) obtained using the scaled data (of variable variance 1). Construct a two-dimensional scatterplot of the 7 original variables (all but the "Country" variable) in the (PC1, PC2) plane.
Comment on the pattern of the variable loadings (coefficients) in PC2.

3. (*Scaling effects in Principal Component Analysis*)

Download data Harman5.txt (automatic download when clicked, also available next to the link of this p-set in Canvas).

The measurements are on five socioeconomic variables for each of 12 census tracts (years ago).

Data can be input and examined using R commands

```
mydata = read.table("Harman5.txt")
```

Use the data to conduct Principal Component Analysis, using the original data as well as scaled data (by making each variable of variance 1, equivalent to using correlation matrix).

(a) In each of the two cases, how many principal components are needed to effectively summarize at least 75% of the variability in the data?

(b) Plot two scree plots, one from PCA based on the original data, one based on the standardized data.

(c) Compare and comment, based on the coefficients (loadings) of the first principal component in each case, and using the results in (a) and (b). Which analysis result is better? Why?

4. (*Simulation on consistency and sparsity of high dimensional PCA*)

Let $\boldsymbol{X} \sim N_p(\mathbf{0}, \Sigma)$ be of multivariate normal, and $\boldsymbol{v}_1$ be an eigenvector of the largest eigenvalue of Σ.
(Assuming the largest eigenvalue of Σ is unique, strictly larger than other eigenvalues.)

(a) Consider a sample of size n drawn from $N_p(\mathbf{0}, \Sigma)$ with $n > p$. Suppose you conducted a principal component analysis on the sample data and obtained the first sample principal component $Y = \hat{\boldsymbol{e}}_1' \boldsymbol{X}$. According to your reasoning, what should be the approximate value of the correlation (i.e. Pearson correlation coefficient) $\rho = corr(\hat{\boldsymbol{e}}_1, \boldsymbol{v}_1)$? And, what should be the approximate value of $|\rho|$?

(b) (Notes: A sample R code for smaller p and n is provided below.)
Construct a non-trivial covariance matrix $\Sigma \neq cI_p$ (and non-diagonal) for $p = 10$.
Draw a sample of size $n = 50$ from $N_p(\mathbf{0}, \Sigma)$ using your Σ. Compute ρ defined in (a).
Repeat the above construction and sample draw for 100 times, obtain ρ values $\{\rho_1, \cdots, \rho_{100}\}$.
Plot a histogram for $|\rho_i|$ using the 100 ρ_i's.

Does the distribution of the values of ρ's agree with your hypothesis in (a)?

```
# Sample code for Q4(b)
library(MASS)
Dim = 5
sampN = 30
C = replicate(Dim, rnorm(Dim))
myCov = C%*%t(C)      # constructed correlation matrix of the variables
M = 100             # number of samples drawn
Rhos = rep(0,M)
for (i in 1:M)
{
  Data = mvrnorm(n=sampN, mu=rep(0,Dim), Sigma=myCov)
  samS=cov(Data)
  Rhos[i]=cor(eigen(samS)$vector[,1],eigen(myCov)$vector[,1])
}
hist(abs(Rhos),nclass=15,main=paste("Distribution of PC1 corr's, p =",Dim, ", n =",sampN),xlim=c(0,1))
# dev.copy2pdf()    # save to pdf
```

(c) Conduct a similar simulation as in (b), now with $p = 100$ (keep $n = 50$).
Plot the histogram of the $|\rho_i|$'s. Is the histogram similar to that in (b)?
Does the distribution of $|\rho|$'s agree with your hypothesis in (a) this time?
What does your result mean in terms of the goodness of the first sample principal component?

(d) Try larger p in (c) without breaking your laptop(keep $n = 50$). How far could you go? Plot the histogram(s) and comment.

(e) A researcher acquired data of gene expression levels of 4000 genes for each of the 100 patients. To find which genes are important, the researcher conducted PCA on the data and obtained a nice L shape in the scree plot. Subsequently the genes with larger loadings on the first couple of principal components were reported as important. Why is the researcher's conclusion problematic?

5. (*Factor analysis using PC and ML methods*)

Utilize the same data Harman5.txt as in Exercise 2 in this assignment.

(a) Obtain the principal component solution to the factor model $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{LF} + \boldsymbol{\varepsilon}$ with the number of factors $m = 2$,

i. Using original data.

ii. Using normalized (variance = 1) data.

(b) Find the maximum likelihood estimates of $\boldsymbol{L}$ and $\boldsymbol{\Psi}$ for $m = 2$. What happens if you try $m = 3$?

(c) Compare the factors obtained by principal component methods and by maximum likelihood, especially on their estimates of the covariance or correlation matrix. Compare the entries in the residual matrices. Which method is better in estimating correlation matrix?

6. (*Population factor Model with $m = 1$*) [4+6+5=15pts]

A researcher believes that the variables in the study form a factor model with $m = 1$ factor.

The covariance matrix of the variables Σ (possibly an estimate) is given below, and it is decomposed (factored) based on the factor model assumption.

$$\Sigma = \begin{bmatrix} 5 & 2 & 3 \\ 2 & 6 & 6 \\ 3 & 6 & 10 \end{bmatrix} = \begin{bmatrix} \ell_1 \\ \ell_2 \\ \ell_3 \end{bmatrix} [\ell_1 \quad \ell_2 \quad \ell_3] + \begin{bmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{bmatrix}$$

(a) How many variables are there in this study? Write the population factor model in vector-matrix notations ($\boldsymbol{X} = \cdots$), indicate the dimensions of every term in the model.

(b) Setup a system of equations and solve for $\ell_i, \psi_i, i = 1, 2, 3$.

(c) For each variable X_i in the model, calculate the percentage of $Var(X_i)$ explained by the common factor.

(d) (*Haywood case*) (Required for 32950, optional for 24620)

Do the same as in parts (a), (b), (c), now for the covariance matrix $\Sigma = \begin{bmatrix} 5 & 2 & 3 \\ 2 & 6 & 6 \\ 3 & 6 & 8 \end{bmatrix}$.

Comment on your results in part (c) for this Σ. Is the factor model reasonable?

Useful R commands for Q2, 3, and 5 (modify variable names to fit your specific need):

```
eigen(mydata)
ranking1 = order(variable, decreasing=T)
mydata[ranking1,]
mydata$variable[ranking1]
PC=princomp(mydata, cor=T)    # or ... = princomp(mydata)
summary(PC, loading=T)
PC$scores
screeplot(PC, type="l")
factanal(mydata,#factors)    # ML method
?princomp                    # help for the command "princomp", likewise for other commands
```