STAT 315 / 463 Multivariate Statistical Methods
Assignment 3
Principal Components & Factor Analysis

Out of Max of 10 marks.

Set: Thursday 29th April

Due: Thursday 13th May, 5:00pm

Maximum page limit 6 A4 pages, late assignment will be penalized.

Make sure you explain your answer. This assignment contributes 10% for STAT315 and 5% for STAT463 to the course grade.Include your R script as an Appendix (**not** counted in the page limit).

Submit your script on Learn. You must submit the script as a SINGLE PDF, containing all the answers. A commonly used software to convert documents, or take a photo using a camera and convert it, is Microsoft Office Lens which is free and available for a lot of devices, but there are many alternatives. This will prepare you for online exams mid-year, so we will practice creating such PDF's this term. Submissions that are not single PDF files will have to be re-submitted.

A) Nordic Combination (max. 4 marks)

The file Nordic.txt contains the result of the Sochi 2014 Nordic Combined 10k/Normal Hill event. The competition is decided by who performs the best in a combination of ski jumping and cross-country skiing. The variable SkiJump is the ski jump score and CrossCountry is the cross-country time in seconds.

Source: `http://www.sochi2014.com/en/nordic-combined-ind-gund-nh-10-km-cross-c-free-race`

1. Perform the principal component analysis on the correlation matrix.
2. One way of combining the scores is to use the first principal component. Why might this be a good idea?
3. If the competitors were ranked based on the first principal component, who would have won the bronze medal?
4. What do you think the second principal component represents?
5. Are the data adequately summarized by one principal component?
6. The IOC wants to introduce a new snowmobile half-pipe event and is considering dropping the Nordic combined on the grounds that ability in cross-country skiing and ski jumping are more or less equivalent. Do you think this is reasonable in terms of correlation?
7. Would it be better to run a PCA on the covariance matrix instead of the correlation matrix in this example? Who would be the gold medallist in that case (first PCA component on the covariance matrix) mean?

B) Police Applicants (max. 3 marks)

The police.csv file contains measurements relevant to assessing the health of police applicants:

ID Applicant ID number

HEIGHT height [cm]

REACT reaction time to visual stimulus

WEIGHT weight [kg]

SHLDR shoulder width [cm]

PELVIC width of pelvic [cm]

CHEST min chest circumference [cm]

THIGH thigh skinfold thickness [mm]

PULSE resting pulse rate

DIAST diastolic blood pressure

CHNUP No. of chin-ups completed in 1 min

BREATH max breathing capacity [l]

RECVR pulse rate after 5 min recovery from treadmill

SPEED maximum treadmill speed

ENDUR treadmill endurance time [min]

FAT total body fat

To obtain a simplified rating scheme of police applicants, the variables should be categorised into groups that characterise different aspects of the applicants abilities.

Perform Factor Analysis to allocate the variables into several groups:

1. How many factors can be found? (using hypotheses testing with $p \leq 0.05$)
2. Which variables are grouped by the first two factors? (e.g. threshold $|loading| \geq 0.5$)
3. To reduce the time and effort of obtaining so many variables, we would rather not measure the diastolic blood pressure. Just measuring the resting pulse rate should be sufficient. Do you agree? (Why or why not...)
4. When we want to separate huge athletic applicants from huge non-athletic applicants, which factor scores can be used?

C) Singular Value Decomposition (max. 3 marks)

Let $\boldsymbol{A}$ be a (3 observations by 2 variables) matrix:

$$\boldsymbol{A} = \begin{bmatrix} 10 & 10 \\ 3 & 0 \\ 4 & -5 \end{bmatrix}$$

Find the Singular Value Decomposition of $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V'}$:

1. Compute $\boldsymbol{A'A}$, please show the working by hand.
2. Take this matrix $\boldsymbol{A'A}$, compute its eigenvalue $[\boldsymbol{\Lambda}^2]$ and eigenvector $[\boldsymbol{V}]$ by hand. You could used R or other software to help check your answer.
3. Compute $\boldsymbol{U} = \boldsymbol{A}\boldsymbol{V}\boldsymbol{\Lambda}^{-1}$.
4. What does the the eigenvectors $\boldsymbol{V}$ and eigenvalues $\boldsymbol{\Lambda}$ tell you about the variation in the 3 observation in $\boldsymbol{X}$ (think about the principal components).