HW3

ST 537: Q1-Q5 ST 437: Q1-Q3 and Q5

1. The academic score data (OECD_PISA.txt in the data folder) gives the outcome of a standardized test of students taken in different countries in 2009. The various scores that went into this table have been standardized across countries to adjust for cultural differences.

(a) Perform a principal components analysis of this data. Notice that the first principal component accounts for much of the variance.What do the factor loadings suggest? Do the loadings differ if you use the covariance matrix or the correlation matrix?

(b) Look at the correlation matrix of this data. What does the correlation matrix say about this data?

2. In factor analysis we saw that the $p \times p$ covariance of the data $X = (X_1, \dots, X_p)^T$, Σ, is represented as

$$\Sigma = LL^T + \Psi.$$

Pre and post multiply this equation by a diagonal matrix containing the inverse variances of the X_j's for $j = 1, \dots, p$. Hence obtain an equivalent decomposition for the correlation matrix. Discuss your observation..

3. Generate $n = 200$ observations of the three variates X_1, X_2, X_3 according to

$$X_1 \sim Z_1 \qquad Z_1 \sim N(0,1)$$

$$X_2 = X_1 + .001 Z_2, \qquad Z_2 \sim N(0,1)$$

$$X_3 = 10 Z_3, \qquad Z_3 \sim N(0,1)$$

where Z_1, Z_2, Z_3 are independent variables. Compute the leading principal component and factor analysis directions. In generating the data specify the seed so that we can replicate your results; R syntax is {set.seed (1257)} Discuss the results.

4. The dataset {Harmon23.cor} in the {datasets} package is a list with the first element equal to the correlation matrix of eight physical measurements made on 305 girls between 7 and 17.

(a) Perform a factor analysis of this data using the command

```
# factanal(factors=m, covmat = Harman23.cor)
```

(b) Vary the number of factors to find an adequate fit of the model and interpret the resulting factor loadings.

(c) Does the principal component analysis produce different conclusions when the correlation matrix {(cor=TRUE)} option is used ? Which analysis do you prefer?

5. Six different tests of intelligence and ability were administered to 112 people. The covariance (but not the original data) of the test results is given in {ability.cov} in the {datasets} library. The six tests are called *general, picture, blocks, maze, reading, vocabulary*, and *reading*. More information is given in the R help file.

(a) Perform a factor analysis on the covariance matrix with

```
#factanal(factors = 2, covmat=ability.cov)
```

Use the loadings to identify those variables that group together within the first two factors. Interpret these factors.

(b) Perform a principal components analysis using the covariance matrix

```
# summary(pc <- princomp(ability.cor))
# pc$loadings
```

and identify the variables making the largest contributions
to the first two principal components. How do you interpret these principal components?

(c) Do you think it is more appropriate to examine the covariance or the correlation in a principal components analysis of this data?

(d) The {cov2cor} function efficiently converts covariances into correlation matrices. That is

```
# ability.cor <- cov2cor(ability.cov$cov)
# princomp(ability.cor)
# princomp(ability.cor)$loadings
```

obtains the correlation and performs the principal components analysis. Examine the loadings and interpret the first two principal components. Compare this data summary with parts (a) and (b) How do these differ? How are they similar?