

Assignment 1

1. Six variables are measured on 100 genuine and 100 forged old swiss 1000-franc bills. The observations 1-100 are genuine, the other 100 observations are forged banknotes. You can obtain the data via

```
X = read.table(file=url(
"http://www.maths.usyd.edu.au/u/UG/SM/STAT3014/r/Data/SwissBankNotes.dat"),
header=TRUE, sep=";")
```

The variables, in column order, are:

LEN	Length of the bill
LH	Height of the bill, measured on the left
RH	Height of the bill, measured on the right
LM	Distance of inner frame to the lower border
UM	Distance of inner frame to the upper border
DG	Length of the diagonal

- (a) Use R to perform a principal component analysis on X based in the correlation matrix. Obtain the variability explained by the first three principal components.
 - (b) Calculate the standard deviations and principal component loadings for the first two principal components.
 - (c) Find the correlations of the first and second principal components with all the original variables.
 - (d) Which variables are highly positively or highly negatively correlated with the first and second principal components?
 - (e) Plot the first two principal components and label the points with the 0-1 code denoting genuine or forged notes. Use the plot to relate the principal components to whether the notes are genuine or forged.
2. Samples of effluent were sent to two laboratories for testing. One half of each sample was sent to the Wisconsin State Laboratory of Hygiene and one half sent to a private laboratory. Measurements of biochemical oxygen demand (BOD) and suspended solids (SS) were obtained for 11 samples. These data are stored in a text file and can be downloaded via

```
effluent = read.table(file=url("http://
www.maths.usyd.edu.au/u/UG/SM/STAT3014/r/Data/effluent", header=TRUE)).
```

The first two columns, $\mathbf{effluent}[1 : 2]$, contains the BOD and SS readings from the private laboratory and the last two columns, $\mathbf{effluent}[3 : 4]$, contain the BOD and SS readings from the state laboratory.

The BOD and SS measurements for the 11 samples from private and state laboratory given in the file $\mathbf{effluent}$ are in the same order. Suppose that we wish to test whether the laboratory readings for BOD and SS are different for each of the laboratories.

- (a) State an appropriate null hypothesis. Rewrite the hypothesis in the form $H_0: \mathbf{A}\boldsymbol{\mu} = \mathbf{r}_0$ for an appropriate matrix \mathbf{A} and vector \mathbf{r}_0 .
 - (b) Calculate the means and covariance of the original data. Using these and the matrix \mathbf{A} calculate the means and covariance of the differences in BOD and SS readings between the two laboratories.
 - (c) Calculate an appropriate test statistic, determine the appropriate p -value for this test and form a conclusion.
3. The Wisconsin Department of Health and Social Services reimburses nursing homes in the state for the services provided. Nursing homes can be classified on the basis of ownership (private, nonprofit or government). One recent study investigated the effects of ownership on costs. Four costs were selected for the analysis:

X_1	Cost of nursing labour
X_2	Cost of dietary labour
X_3	Cost of plant operation
X_4	Cost of housekeeping and laundry labour

A total of $n = 516$ observations on each of the $p = 4$ cost variables were separated according to ownership type. For private owners

$$n_1 = 271, \quad \bar{x}_1 = \begin{bmatrix} 2.066 \\ 0.480 \\ 0.082 \\ 0.360 \end{bmatrix} \quad \mathbf{S}_1 = \begin{bmatrix} 0.291 & -0.001 & 0.002 & 0.010 \\ -0.001 & 0.011 & 0.000 & 0.003 \\ 0.002 & 0.000 & 0.001 & 0.000 \\ 0.010 & 0.003 & 0.000 & 0.010 \end{bmatrix}.$$

For nonprofit owners

$$n_2 = 138, \quad \bar{x}_2 = \begin{bmatrix} 2.167 \\ 0.596 \\ 0.124 \\ 0.418 \end{bmatrix} \quad \mathbf{S}_2 = \begin{bmatrix} 0.561 & 0.011 & 0.001 & 0.037 \\ 0.011 & 0.025 & 0.004 & 0.007 \\ 0.001 & 0.004 & 0.005 & 0.002 \\ 0.037 & 0.007 & 0.002 & 0.019 \end{bmatrix}.$$

For government owners

$$n_3 = 107, \quad \bar{x}_3 = \begin{bmatrix} 2.273 \\ 0.521 \\ 0.125 \\ 0.383 \end{bmatrix} \quad \mathbf{S}_3 = \begin{bmatrix} 0.261 & 0.030 & 0.003 & 0.018 \\ 0.030 & 0.017 & 0.000 & 0.006 \\ 0.003 & 0.000 & 0.004 & 0.001 \\ 0.018 & 0.006 & 0.001 & 0.013 \end{bmatrix}.$$

Suppose we wish to test that there is no difference in average costs among the three different type of owner types.

- (a) State an appropriate null hypothesis. What assumptions will you make?
 - (b) Calculate the pooled covariance \mathbf{S}_p , within-group sum of squares matrix \mathbf{W} and between-group sum of squares matrix \mathbf{B} .
 - (c) Use these values to calculate the trace test statistic, using this statistic calculate the appropriate corresponding p -value for the test statistic and form a conclusion.
 - (d) Use the test of dimensionality to determine an appropriate dimension for the group means.
4. In a combined dataset from several US hospitals it was attempted to determine the risk factors associated with diabetes. These data are stored in a text file and can be downloaded via

```
effluent = read.table(file=url("http://
www.maths.usyd.edu.au/u/UG/SM/STAT3014/r/Data/pid.dat", header=TRUE)).
```

From these hospitals various measurements were obtained from 392 patients:

pregnant	Number of times pregnant
glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
pressure	Diastolic blood pressure (mm Hg)
triceps	Triceps skin fold thickness (mm)
insulin	2-Hour serum insulin (mu U/ml)
mass	Body mass index (weight in kg/(height in m) ²)
pedigree	Diabetes pedigree function
age	Age (years)
diabetes	Class variable ("pos" or "neg")

- (a) Calculate Fisher linear discriminant function based on these data. Use classifier to determine how to classify a patient with the measurements (1, 89, 66, 23, 94, 28.1, 0.167, 21).
 - (b) Use the function `rpart` from the `rpart` package to build and plot a classification tree. Use the classification tree to describe how to classify a patient with the measurements (1, 89, 66, 23, 94, 28.1, 0.167, 21).
 - (c) Use 49-fold cross-validation to determine the best k nearest neighbour classifier using the function `disc.knn` from the `class` package where $k \in (3, 5, 7, \dots, 37, 39)$ (Also, use the R command `set.seed(1)`).
5. For any symmetric positive definite matrix Σ and appropriately size vectors $\bar{\mathbf{X}}$ and $\boldsymbol{\mu}_0$ show that

$$\max_{\mathbf{a} \neq \mathbf{0}} \frac{n(\mathbf{a}^T(\bar{\mathbf{X}} - \boldsymbol{\mu}_0))^2}{\mathbf{a}^T \Sigma \mathbf{a}} = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^T \Sigma^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0).$$

and find the value of the vector \mathbf{a} which achieves this maximum.