Assignment #1 STA437H1S/2005H1S

due Friday February 5, 2016

Instructions: Students in STA437S do problems 1 through 3; those in STA2005S do all 4 problems.

1. Suppose that $\boldsymbol{X} = (X_1, \cdots, X_5)^T \sim \mathcal{N}_5(\boldsymbol{\mu}, C)$ where

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} 4.5 & -2.0 & -1.5 & -1 & -0.5 \\ -2.0 & 4.0 & 3.0 & 2.0 & 1.0 \\ -1.5 & 3.0 & 7.5 & 5.0 & 2.5 \\ -1.0 & 2.0 & 5.0 & 8.0 & 4.0 \\ -0.5 & 1.0 & 2.5 & 4.0 & 5.5 \end{pmatrix}$$

(a) What are the marginal distributions of $X_1, \cdots, X_5$?
(b) What is the conditional distribution of (X_1, X_2) given $X_3 = 2$, $X_4 = 3$ and $X_5 = -1$? (You should use R or some other software to compute whatever matrix inverses you need.)
(c) Using the inverse of C, give the graph structure of the dependence in $\boldsymbol{X}$. (Again use R to compute C^{-1} but note that the numerical computation is subject to roundoff error!)

2. The file `marks.txt` on Blackboard contains the exam marks data considered in lecture. In that analysis, we assumed that the data came from a multivariate normal distribution. The data can be read into R as follows:

```
> exam <- scan("marks.txt",what=list(0,0,0,0,0))
> mec <- exam[[1]]
> vec <- exam[[2]]
> alg <- exam[[3]]
> ana <- exam[[4]]
> sta <- exam[[5]]
```

(a) Look at normal quantile-quantile plots of the 5 variables separately using `qqnorm`. You can judge the "goodness-of-fit" using the Shapiro-Wilk test, which can be implemented in R using `shapiro.test`. Comment on the results.
(b) Use the function `qqmultinorm` (which is in the file `qqmultinorm.txt` on Blackboard) to assess the multivariate normality of the data. The following R code will look at 100 normal quantile-quantile plots of 100 randomly chosen projections and compute p-values for the Shapiro-Wilk test for each projection:

```
> r <- qqmultinorm(cbind(mec,vec,alg,ana,sta),nproj=100,plot.edf=T)
```

The plot produced will be the empirical distribution function of the 100 p-values compared to the distribution function of a uniform distribution on $[0, 1]$. Based on this plot, do the data seem to be (at least approximately) multivariate normal?

3. The file `crabs.txt` on Blackboard contains data on two species of rock crabs, which are distinguished by their colour (blue or orange); the columns of the file are species (B or O), sex (M or F),

index (1-50 within each species-sex combination), width of the frontal lip (LP), the rear width of the shell (RW), length along the midline of the shell (CL), the maximum width of the shell (CW), and the body depth (BD). Ultimately, we would like to use the latter 5 variables to classify the species and sex of a crab but at this stage, we will simply look at the structure of the data to see which variables might be useful in classifying the species and sex of a rock crab.
The data can ne read into R using the following code:

```
> x <- scan("crabs.txt",skip=1,what=list("c","c",0,0,0,0,0,0))
> colour <- ifelse(x[[1]]=="B","blue","orange")
> sex <- x[[2]]
> FL <- x[[4]]
> RW <- x[[5]]
> CL <- x[[6]]
> CW <- x[[7]]
> BD <- x[[8]]
```

Use the following code to look at pairwise scatterplots of the 5 variables:

```
> pairs(cbind(FL,RW,CL,CW,BD),pch=sex,col=colour)
```

The males and females are indicated on the plots by M and F respectively with the two species being indicated by the colour of the points.
(a) Which pairwise scatterplots are particularly effective for “separating” the two species?
(b) Which pairwise scatterplots are effective for “separating” the two sexes?
(c) Use the `scatterplot3d` function to look at various 3 dimensional (3D) scatterplots of the data. Is there a particular 3D scatterplot that seems to effectively separate the 4 species/sex groups?

4. In class, we stated that we can assess whether multivariate data can be modeled by a multivariate normal distribution by checking the normality of a collection of one dimensional projections (for example, by using normal quantile-quantile plots). When p is very large, this procedure breaks down – almost every one dimensional projection appears to be normal.
(a) Consider n=100 observations of a 1000-variate distribution whose components are independent exponential random variables with mean 1; the joint density of this distribution is

$$f(x_1, \cdots, x_{1000}) = \exp\left(-\sum_{i=1}^{1000} x_i\right) \quad \text{for } x_1, \cdots, x_{1000} \geq 0$$

We can simulate the 100 observations in R as follows:

```
> x <- matrix(rgamma(100000,1),ncol=1000)
```

Now use the function `qqmultinorm` (available on Blackboard) to look at normal quantile-quantile plots of 20 one dimensional projections:

```
> r <- qqmultinorm(x,nproj=20,plot.qq=T,plot.edf=T)
```

How do these quantile-quantile plots compare to the quantile-quantile plots of each variable (for example, `qqnorm(x[,1])`)?
(b) Can you explain this phenomenon? (Hint: What is the approximate distribution of $\boldsymbol{a}^T \boldsymbol{X} = \sum_{j=1}^{p} a_j X_j$ when p is large if $\max_{1 \leq j \leq p} a_j^2 / \boldsymbol{a}^T \boldsymbol{a}$ is small?)