Prelims Data Analysis TT 2019
Sheet 6

At the end of this exercise sheet there are Optional Practical Exercises in `R` and `Matlab`. It is **strongly recommended** that students do these exercises, but students should ask their college tutor whether to use `R` or `Matlab`. The course website has an Introduction to `R`, which students should work through before starting the `R` exercises.

1. Let X, Y and Z be random variables and let $\text{cov}(A, B)$ denote the covariance of any two random variables A and B. Show that

   (a) $\text{cov}(aX, Y) = a\text{cov}(X, Y)$

   (b) $\text{cov}(X, Y + Z) = \text{cov}(X, Y) + \text{cov}(X, Z)$

   (c) If $X_1, \ldots, X_p$ is a set of random variables then show that

$$\text{cov}\left(\sum_{i=1}^{p} \alpha_i X_i, \sum_{j=1}^{p} \beta_j X_j\right) = \sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_i \beta_j \text{cov}(X_i, X_j)$$

2. Suppose $X = (X_1, \ldots, X_p)^T$ is a p-vector of random variables with covariance matrix $\mathbf{\Sigma}$ where

$$\begin{aligned} \text{var}(X_i) &= \mathbf{\Sigma}_{ii} \quad \text{for } i \in 1, \ldots, p \\ \text{cov}(X_i, X_j) &= \mathbf{\Sigma}_{ij} \quad \text{for } i \neq j \in 1, \ldots, p \end{aligned}$$

   Define new random variables Z and W to be linear combinations of $X = (X_1, \ldots, X_p)$ such that

$$Z = \alpha^T X = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_p X_p$$

$$W = \beta^T X = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

   where

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$$
$$\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$$

   then show that

   (a) $\text{var}(Z) = \alpha^T \mathbf{\Sigma} \alpha$

   (b) $\text{cov}(Z, W) = \alpha^T \mathbf{\Sigma} \beta$

3. If $X = (X_1, \ldots, X_p)^T$ is a p-dimensional random column vector such that $X \sim N_p(\mu, \mathbf{\Sigma})$ and $\boldsymbol{B}$ is a $m \times p$ matrix then

   (i) Show that the m-dimensional random column vector $Y = \boldsymbol{B}X$ has covariance matrix $\text{cov}(Y) = \boldsymbol{B}\mathbf{\Sigma}\boldsymbol{B}^T$.

   (ii) If $m = p$ how can we choose the matrix $\boldsymbol{B}$ so that the transformed variable $\boldsymbol{B}X$ has a covariance matrix that is the identity matrix.

4. Let $X = (X_1, X_2)^T$ be a 2-dimensional random column vector such that $X \sim N_p(\mu, \mathbf{\Sigma})$ with $\mu = (0, 0)^T$, $\mathbf{\Sigma}_{11} = \mathbf{\Sigma}_{22} = 1$ and $\mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{21} = \rho$.

(i) Show that pdf of X is given by

$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)}\right) \quad x \in \mathbb{R}^2$$

(ii) Describe the shape of the pdf of X at any fixed value of the first random variable $X_1 = a \in \mathbb{R}$.

5. Let $X = (X_1, X_2)^T$ be a 2-dimensional random column vector such that $X \sim N_p(\mu, \mathbf{\Sigma})$ with $\mu = (120, 80)^T$, $\mathbf{\Sigma}_{11} = 25$, $\mathbf{\Sigma}_{22} = 16$ and $\mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{21} = 12$. Define the new random variable $Y = 2X_1 - 3X_2$.

   (i) Calculate $P(Y > 20)$

   (ii) If $\Sigma_{21} = 0$ what is $P(Y > 20)$.

   Hint : you may assume the result stated in the notes that if $X = (X_1, \ldots, X_p)^T$ is a p-dimensional random column vector such that $X \sim N_p(\mu, \mathbf{\Sigma})$ and $\boldsymbol{B}$ is a $m \times p$ matrix then $Y \sim N_m(\boldsymbol{B}\mu, \boldsymbol{B}\mathbf{\Sigma}\boldsymbol{B}^T)$

6. Let $x_1, \ldots, x_n$ be iid realizations of a p-dimensional random column vector $X = (X_1, \ldots, X_p)^T$ such that $X \sim N_p(\mu, \mathbf{\Sigma})$.

   (i) Prove that the maximum likelihood estimator of μ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

   (ii) Show that the log-likelihood can be expressed as

$$\ell(\mu, \mathbf{\Sigma}) = \frac{n}{2}\log|\mathbf{\Sigma}^{-1}| - \frac{1}{2}tr(\mathbf{\Sigma}^{-1}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T)$$

   (iii) Using Hints (c) and (d) prove that the maximum likelihood estimator of $\mathbf{\Sigma}$ is

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

   (iv) Show that the sample covariance $\boldsymbol{S}$ is unbiased for $\mathbf{\Sigma}$ where

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

   Hint You may find it helpful to use the following results

   (a) If a and x are p-column vectors, $\boldsymbol{B}$ is a $p \times p$ symmetric matrix and y and z are scalars such that $y = a^T x$ and $z = x^T \boldsymbol{B} x$ then

$$\nabla y = a \text{ and } \nabla z = 2\boldsymbol{B}x$$

   (b) If $\boldsymbol{C}$ is a $n \times m$ matrix and $\boldsymbol{D}$ is $m \times n$ matrix then $tr[\boldsymbol{CD}] = tr[\boldsymbol{DC}]$

   (c) The matrix of partial derivatives of a scalar y function of an $n \times n$ matrix $\boldsymbol{X}$ of independent variables, with respect to the matrix $\boldsymbol{X}$, is defined as

$$\nabla y = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1n}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{2n}} & \frac{\partial y}{\partial x_{n2}} & \cdots & \frac{\partial y}{\partial x_{nn}} \end{bmatrix}.$$

   If $\boldsymbol{E}$ and $\boldsymbol{F}$ are $p \times p$ matrices then

$$\begin{aligned} \nabla \log|\boldsymbol{E}| &= (\boldsymbol{E}^{-1})^T \\ \nabla tr[\boldsymbol{EF}] &= \boldsymbol{F}^T \end{aligned}$$

(d) If $L(\theta)$ is a likelihood function with Maximum Likelihood Estimator (MLE) of $\hat{\theta}$ and $g(\theta)$ is a function of θ then the MLE of $g(\theta)$ is $g(\hat{\theta})$. *Note : this is known as the invariance property of MLEs. This is covered formally in Part A Statistics course, but it is straightforward to prove, and could be covered in tutorials if you ask your tutor nicely.*

Optional Practical Exercises using R

Students should carry out these practical exercises and produce a report summarizing the results of their analysis i.e. produce a document that contains the plots produced and hand this in to your tutor.

NOTE To run these exercises in R you will need to install a few packages called MASS, rgl and car. To do this in RStudio click

Tools − > Install Packages

and then type in the names of the packages and install them. Make sure to click the box that says "install dependencies"

1. Use the following R code to simulate and plot 200 points from a bivariate normal distribution with mean $\mu = (0, 0)^T$, both variances equal to 1 and covariance equal to 0.5.

```
library(MASS)
S = matrix(c(1,0.5,0.5,1),2,2)
x = mvrnorm(200, mu = c(0,0), Sigma = S)
plot(x)
```

   Change the code and make a new plot for the situation where the covariance is equal to -0.7.

2. The Crabs dataset is in the MASS library which can be loaded using

```
library(MASS)
```

   To look at the dataset just type

```
crabs
```

   We can create a new dataset with the 5 main variables as follows

```
varnames = c('FL','RW','CL','CW','BD')
Crabs = crabs[,varnames]
```

   Then we can create boxplots of the 5 variables as follows

```
boxplot(Crabs)
```

   and a pairs plot of the variables as follows

```
pairs(Crabs)
```

   Histograms of the variables can be created as follows

```
par(mfrow=c(2,3))
hist(Crabs$FL)
hist(Crabs$RW)
hist(Crabs$CL)
hist(Crabs$CW)
hist(Crabs$BD)
```

   To explore the dataset in 3D using triples of variables we can use the following code

```
library(rgl)
library(car)
rgl.open()
scatter3d(x=Crabs$RW, y=Crabs$CW, z=Crabs$CL, surface = F)
```

Optional Practical Exercises using `Matlab`

Students should carry out these practical exercises and produce a report summarizing the results of their analysis i.e. produce a document that contains the plots produced and hand this in to your tutor.

1. Use the following `Matlab` code to simulate and plot 200 points from a bivariate normal distribution with mean $\mu = (0, 0)^T$, both variances equal to 1 and covariance equal to 0.5.

   ```
   S = [1 0.5; 0.5 1];
   x = mvnrnd([0 0], S, 200);
   plot(x);
   ```

   Change the code and make a new plot for the situation where the covariance is equal to -0.7.

2. Download a copy of the Crabs dataset from this link

   http://www.stats.ox.ac.uk/~sejdinov/teaching/data/crabs.txt

   and save it in a new folder, for example `P:\\Downloads` (this will depend on your own machine/OS.)

   Change to that directory using something like

   ```
   cd P:\\Downloads;
   ```

   Read the dataset into `Matlab` using

   ```
   crabs = readtable('crabs.txt', 'Delimiter', 'space');
   ```

   To look at the dataset just type

   ```
   crabs
   ```

   We can create a new dataset with the 5 main variables as follows

   ```
   varnames =  'FL' 'RW' 'CL' 'CW' 'BD' ;
   Crabs = crabs(:, varnames);
   ```

   Then we can create boxplots of the 5 variables as follows

   ```
   boxplot(table2array(Crabs), 'Labels', varnames);
   ```

   and a pairs plot of the variables as follows

   ```
   corrplot(Crabs);
   ```

   Histograms of the variables can be created as follows

   ```
   hist(Crabs.FL);
   xlabel('FL');
   hist(Crabs.RW);
   xlabel('RW');
   hist(Crabs.CL);
   xlabel('CL');
   hist(Crabs.CW);
   xlabel('CW');
   hist(Crabs.BD);
   xlabel('BD');
   ```

   To explore the dataset in 3D using triples of variables we can use the following code. To explore the 3D projection click the "rotate" tool i.e. the button with the circular arrow.

   ```
   scatter3(Crabs.RW, Crabs.CW, Crabs.CL);
   xlabel('RW');
   ylabel('CW');
   zlabel('CL');
   ```