The University of New South Wales

Department of Statistics
MATH5855 - Multivariate Analysis I

Assignment 2
Due Tuesday, 25th September 2018, 5pm

1. i) You are asked to write a subroutine (module) within SAS/IML with an input:

- an arbitrary data matrix with n datapoints, each containing p dimensions ($p < n$)
- a vector a containing 2 integers among the set $\{1, 2, \ldots, p\}$

If the integers are i and j, say, the module should calculate an estimate of the partial correlation of the ith and jth component when the remaining ones have been fixed.

Head Length, First Son	Head Breadth, First Son	Head Length, Second Son	Head Breadth, Second Son
191	155	179	145
195	149	201	152
181	148	185	149
...	...	...	...
190	163	187	150

The complete file `brothers.dat` (available in moodle) contains the head lengths and breadths of brothers (first and second son in a sample of 25 families). Enter the 25 × 4 matrix within IML, call the module and calculate the partial correlation $r_{34.12}$. Verify your calculation using the CORR procedure (study its help first) or by hand calculation. Calculate also the partial correlation $r_{21.34}$.

Hint. You may consult the file imlregress1.sas in moodle for a hint in organising subroutines in SAS/IML. Operators and control structures you may possibly need, include: `DO..END, IF..THEN, START..FINISH`, comparison operators, subsetting of matrices can be found in the help of the SAS/IML procedure. If you face a difficulty writing the module in its complete generality (that is, arbitrary indices i, j), write a simpler version with $(i = 1, j = 2)$ which could then be used after the columns of the original data matrix have been reshuffled.

ii) Compare $r_{12.34}$ to r_{12} and explain the differences having in mind the meaning of the four variables.

iii) Use Fisher's z to find a confidence interval (CI) for $\rho_{12.34}$ with a level of confidence 0.95.

iv) Estimate the multiple correlation between x_3 and (x_1, x_2), and test its significance at 5% level.

v) Test the significance of the correlation coefficient ρ_{34}, i.e., test $H_0 : \rho_{34} = 0$ against a two-sided alternative, using level of significance $\alpha = 0.05$.

2. Soil samples were taken at $n = 45$ randomly selected locations in South Queensland. Measurements of nitrogen concentration in the soil were made at depths of 1, 3, 5 and 7 feet from the surface. The four measurements from the i-th location can be arranged in a vector as $\mathbf{X_i} = (X_{1i}, X_{2i}, X_{3i}, X_{4i})'$. Let

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X_i} - \bar{\mathbf{X}})(\mathbf{X_i} - \bar{\mathbf{X}})'$$

where $\bar{\mathbf{X}}$ is the sample mean. The data is in the file `soil.dat` on moodle. Multivariate normality can be assumed.

i) Perform a test of the hypothesis that the mean nitrogen concentration is the same at all 4 depths. Report the relevant statistic. State your conclusions.

Hint Transform the four-dimensional data vector X into a three-dimensional vector $Y = CX$ with $C = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$ and reformulate the hypothesis.

ii) Test the null hypothesis that the mean nitrogen concentration decreases in such a way that the mean at one depth is half of the mean at the previous depth, i.e., $H_0 : (\mu_i/\mu_{i-1}) = 1/2, i = 2, 3, 4$. Report and comment.

3. Consider identifying the neurotic state of an individual referred for psychiatric examination. Three measurements A, B, C are made on each individual. The mean scores for each of 3 groups are given as:

Group	A	B	C
Anxiety State	2.9	1.2	0.75
Obsession	4.6	1.6	1.2
Normal	0.6	0.15	0.25

The pooled within group covariance matrix is: $\hat{\Sigma} = \begin{pmatrix} 2.30 & 0.25 & 0.47 \\ 0.25 & 0.60 & 0.03 \\ 0.47 & 0.03 & 0.59 \end{pmatrix}$. Assume equal misclassification costs and equal priors for the three groups.

a) Calculate the linear discriminant scores for classifying into one of the three groups.

b) Classify the following newly observed individuals:

	A	B	C
Mary:	3.000	1.200	1.000
Fred:	4.000	1.400	1.320
Giselda:	1.000	0.500	0.330

c) Consider classifying individuals from the "Anxiety state" and "Obsession" groups **only**. Determine the linear discriminant function and estimate the probabilities of misclassification P(1|2) and P(2|1).

4. The vectors $x_1, x_2, \ldots, x_n$ are a sample from $N_p(0, \lambda D)$, where $\lambda > 0$ is an unknown scalar and D is a known symmetric positive definite matrix. Show that the Maximum Likelihood Estimator of λ is $\hat{\lambda} = \frac{1}{np}\text{tr}(D^{-1}B)$ where $B = \sum_{i=1}^{n} x_i x_i'$. Show also that $\frac{np\hat{\lambda}}{\lambda} \sim \chi^2_{np}$. Hence suggest a two-sided confidence interval for λ at level $(1 - \alpha)$.

(**Hint:** You may find it useful to consider vectors $Y_i = D^{-1/2}X_i$)

5. For a random vector $(X,Y)'$ of continuous random variables with marginal distributions F and G, the coefficient of upper dependence is defined as

$$\lambda_{upper} = \lim_{u \to 1} P(Y > G^{-1}(u) | X > F^{-1}(u))$$

provided that the limit exists. In the context of copulae, this results in the investigation of

$$\lambda_{upper} = \lim_{u \to 1} (1 - 2u + C(u,u))/(1-u).$$

When $\lambda_{upper} \in (0,1]$ we say that there exists an asymptotic dependence in the upper tail; when $\lambda_{upper} = 0$ the random variables are said to be asymptotically independent in the upper tail.

Show that the Gumbell-Hougaard copula

$$C_\theta(u,v) = exp(-[(-logu)^\theta + (-logv)^\theta]^{1/\theta}), u \in [0,1], v \in [0,1]$$

with a parameter $\theta \in [1,\infty)$ exhibits upper tail dependence when $\theta > 1$).

(Reminder: as we know when $\theta = 1$ the above copula coincides with the independence copula).