STOCKHOLM UNIVERSITY
Department of Statistics
Mahmood Ul Hassan

EXAM IN MULTIVARIATE METHODS
November 04 2020

Time: 6 hours

The exam is for individual solving. It is an open-book exam, but you are not allowed to use the help of other students, friends, family, or similar. In case you need clarification, the teacher is available at Zoom:

Join Zoom Meeting from
https://stockholmuniversity.zoom.us/j/9255236581
Meeting ID: 925 523 6581

The exam consists of five questions. To score maximum points on a question solutions need to be clear, detailed and well-motivated.

Question. 1 (6+2+3+3+2=16 Points)

For a data set with observations on two variables x_1 and x_2 the sample correlation matrix was found to be

$$R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}.$$

a) Find Eigen values and construct two principal components that are orthogonal to each other.
b) What proportion of variance is accounted by these principal components? Assuming r=0.4.
c) Compute the loadings of the variables by assuming r=0.6.
d) Find the covariance matrix (S) by assuming r=0.8, $Var(x_1) = 65.41$ and $Var(x_2) = 1.27$.
e) Find the generalized variance using S matrix in part d.

Question. 2 (3+2+4+4+3=16 Points)

The sample correlation matrix given below arises from the scores of 220 boys in six school subjects: (1) French, (2) English, (3) History, (4) Arithmetic, (5) Algebra, and (6) Geometry.

$$R = \begin{array}{r} \textit{French} \\ \textit{English} \\ \textit{History} \\ \text{Arithmetic} \\ \text{Algebra} \\ \text{Geometry} \end{array} \begin{bmatrix} 1 & & & & & \\ 0.50 & 1 & & & & \\ 0.55 & 0.45 & 1 & & & \\ 0.29 & 0.35 & 0.16 & 1 & & \\ 0.33 & 0.32 & 0.19 & 0.59 & 1 & \\ 0.25 & 0.33 & 0.18 & 0.50 & 0.60 & 1 \end{bmatrix}$$

A factor analysis was performed to analyze the correlation matrix from the scores of boys in six school subjects by the Principal component method where two factors were extracted. The two factors are assumed uncorrelated. The un-rotated two-factor solution is given below

Variable	F1	F2
French	0.69	0.48
English	0.71	0.33
History	0.58	0.63
Arithmetic	0.71	-0.42
Algebra	0.75	-0.43
Geometry	0.70	-0.44

Based on these reported results obtain:

a) The communalities.
b) The proportion of variance explained by each factor.
c) The estimated/reproduced correlation matrix.
d) The residual correlation matrix.
e) RMSR.

Question. 3 (8+2+6=16 Points)

For the following data

Observation	Y_1	Y_2	Gender
1	3	6	Male
2	6	2	Female
3	5	8	Male
4	4	12	Female
5	6	10	Male
6	12	8	Female
7	10	4	Female
8	5	6	Female
9	14	5	Male

a) Compute the $\mathbf{SSCP_b}$, $\mathbf{SSCP_w}$ and $\mathbf{SSCP_t}$ matrices.

b) Compute the statistical distance between observations 8 and 9.

c) Suppose n1=6 and n2=7 are observations in group-1 and group-2, respectively and

Within-group covariance matrix for group-I = $S_1 = \begin{bmatrix} 9.70 & -3.45 \\ -3.45 & 4.70 \end{bmatrix}$

Within-group covariance matrix for group-II = $S_2 = \begin{bmatrix} 11 & -3.0 \\ -3.0 & 5.4 \end{bmatrix}$

$\bar{X}_1 = \begin{bmatrix} 5.1 \\ 4.8 \end{bmatrix}$ and $\bar{X}_2 = \begin{bmatrix} 9 \\ 4 \end{bmatrix}$

Calculate Fisher's linear discriminant function for this data set.

Question. 4 (4+4+4+4=16 Points)

Observations on two variables were made for five subjects according to the following table.

Subject	Variable-1	Variable-2
1	2	4
2	6	5
3	9	7
4	8	1
5	10	12

a) Construct a similarity matrix containing squared Euclidean distances
b) Use the similarity matrix in part (a) and perform a cluster analysis with the following method
   I. Farthest neighbor method.
   II. Centroid method.
   III. Average linkage method.

Question. 5 (2+3+3+4+4=16 Points)

A company that manufactures riding mowers wants to identify the best sales prospects for an intensive sales campaign. In particular, the manufacturer is interested in classifying households as prospective owners or nonowners on the basis of Income (in $1000s) and Lot Size (in 1000 ft2). Data were collected and a logistic regression was fitted:

```
Coefficients:
             Estimate   Std. Error z value  Pr(>|z|)
(Intercept) -33.72805   15.75557   -2.141    0.0323 *
Income        0.15374    0.07617    2.018    0.0436 *
Lot_Size      1.24452    0.61295    2.030    0.0423 *
```

The following table displays observations on 12 riding-mower owners and 12 nonowners as well as the estimated probability to be an owner based on the logistic regression.

Ownership	Income	Lot Size	$\hat{P}$	Ownership	Income	Lot Size	$\hat{P}$
Owner	60	18.4	0.17	Owner	75	19.6	0.89
Owner	85.5	16.8	0.58	Nonowner	52.8	20.8	0.56
Owner	64.8	21.6	0.96	Nonowner	64.8	17.2	0.09
Owner	61.5	20.8	0.83	Nonowner	43.2	20.4	0.15
Owner	87	23.6	0.99	Nonowner	84	17.6	0.74
Owner	110.1	19.2	0.99	Nonowner	49.2	17.6	0.01
Owner	108	17.6	0.99	Nonowner	59.4	16	0.01
Owner	82.8	22.4	0.99	Nonowner	66	18.4	0.34
Owner	69	20	0.85	Nonowner	47.4	16.4	0.002
Owner	93	20.8	0.99	Nonowner	33	18.8	0.01
Owner	51	22	0.72	Nonowner	51	14	0.0002
Owner	81	20	0.97	Nonowner	63	14.8	0.003

a) Interpret the parameter $\hat{\beta}_1$.
b) What are the odds that a household with a $70000 income and a lot size of 21000 ft^2 is an owner?

c) What is the classification of a household with a \$66,000 income and a lot size of 25,000 ft^2?

d) What is the minimum income that a household with 10,000 ft^2 lot size should have before it is classified as an owner?

e) Classify the observations given in the table and compute the sensitivity and specificity.