STOCKHOLM UNIVERSITY
Department of Statistics
Mahmood Ul Hassan

EXAM IN MULTIVARIATE METHODS
September 29 2021

Time: 5 hours
Aids allowed: Pocket calculator, language dictionary.
The exam consists of five questions. To score maximum points on a question, solution 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) Compute the Mahalanobis distance between observations 4 and 5.
c) Use the similarity matrix in part (a) and perform a cluster analysis with the following method
   I. Centroid method.
   II. 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:

Model-1

```
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 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance:     33.104  on 23  degrees of freedom
Residual deviance: 11.966  on 21  degrees of freedom
AIC: 17.966
```

The following table displays observations on 13 riding-mower owners and 11 nonowners as well as the estimated probability to be an owner based on the above logistic regression model.

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

Model-2

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.52042	2.76448	-2.359	0.0183 *
Income	0.10119	0.04231	2.392	0.0168 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 33.104 on 23 degrees of freedom

Residual deviance: 22.279 on 22 degrees of freedom

AIC: 26.279

a) Using Model-1: Interpret the value of $\exp(\hat{\beta}_1)$.
b) Formulate the null and alternative hypothesis and perform a test if model-1 is significantly better than the model-2 using deviance statistic. Use $\alpha = 0.10$.?
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.

Formula Sheet for the Exam in Multivariate Methods

Vectors and matrices

- Length of a vector $\mathbf{a} = (a_1, a_2, \ldots, a_p)$

$$\|\mathbf{a}\| = \sqrt{a_1^2 + a_2^2 + \cdots + a_p^2}$$

- Determinant of a 2×2 matrix $\mathbf{A}$

$$\det(\mathbf{A}) = |\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}$$

- Inverse of a 2×2 matrix $\mathbf{A}$

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

- Eigenvalues are the roots of the characteristic equation

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{0}$$

For each eigenvalue the solution to

$$(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{x} = \mathbf{0}$$

gives the associated eigenvector $\mathbf{x}$

Distances

- Euclidean

$$D_{ik} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2}$$

- Statistical

$$SD_{ik} = \sqrt{\sum_{j=1}^{p} \left(\frac{x_{ij} - x_{kj}}{s_j} \right)^2}$$

- Mahalanobis

$$MD_{ik} = \sqrt{(\mathbf{x}_i - \mathbf{x}_k)^T \, \mathbf{S}^{-1} \, (\mathbf{x}_i - \mathbf{x}_k)}$$

For $p = 2$

$$MD_{ik} = \sqrt{\frac{1}{1 - r^2} \left[\frac{(x_{i1} - x_{k1})^2}{s_1^2} + \frac{(x_{i2} - x_{k2})^2}{s_2^2} - \frac{2r \, (x_{i1} - x_{k1}) \, (x_{i2} - x_{k2})}{s_1 s_2} \right]}$$

Mean-correction and covariance

- Mean-corrected data

$$\underset{(n \times p)}{\mathbf{X}_m} = \{x_{ij}\} = \left\{ X_{ij} - \bar{X}_j \right\}$$

- Covariance

$$\underset{(p \times p)}{\mathbf{S}} = \{s_{ij}\} = \left\{ \frac{\sum_{i=1}^{n} x_{ij} x_{ik}}{n - 1} \right\} = \frac{\mathbf{SSCP}}{df} = \frac{1}{n - 1} \mathbf{X}_m^T \mathbf{X}_m$$

Group Analysis

- Total sum of squares and cross products

$$\mathbf{SSCP}_{\text{total}} = \mathbf{SSCP}_{\text{within}} + \mathbf{SSCP}_{\text{between}}$$

- Pooled within-group sum of squares and cross products

$$\mathbf{SSCP}_{\text{within}} = \sum_{\ell=1}^{g} \mathbf{SSCP}_{\ell}$$

- Pooled covariance matrix

$$\mathbf{S}_{\text{pooled}} = \frac{\mathbf{SSCP}_{\text{within}}}{n-g}$$

- Between-group sum of squares and cross products

$$\mathbf{SSCP}_{\text{between}} = \mathbf{SSCP}_{\text{total}} - \mathbf{SSCP}_{\text{within}}$$

For $g = 2$ groups

$$\mathbf{SSCP}_{\text{between}} = \frac{n_1 n_2}{n_1 + n_2} \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^T$$

Factor Analysis

- For the two-factor model

$$Var(x) = \lambda_1^2 + \lambda_2^2 + Var(\epsilon) + 2\lambda_1\lambda_2\phi$$

$$Cor(x, \xi_1) = \lambda_1 + \lambda_2\phi$$

$$Cor(x_j, x_k) = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} + \left(\lambda_{j1}\lambda_{k2} + \lambda_{j2}\lambda_{k1}\right)\phi$$

- RMSR for EFA

$$RMSR = \sqrt{\frac{\sum_{i=1}^{p} \sum_{j=i+1}^{p} res_{ij}^2}{p(p-1)/2}}$$

- RMSR for CFA

$$RMSR = \sqrt{\frac{\sum_{i=1}^{p} \sum_{j=i}^{p} \left(s_{ij} - \hat{\sigma}_{ij}\right)^2}{p(p+1)/2}}$$

Two-Group Discriminant Analysis

- Maximize

$$\lambda = \frac{\gamma^T \mathbf{B} \gamma}{\gamma^T \mathbf{W} \gamma}$$

- Fisher's linear discriminant function

$$\gamma^T = (\mu_1 - \mu_2)^T \, \mathbf{\Sigma^{-1}}$$

- Wilks' Λ

$$\Lambda = \frac{|\mathbf{SSCP}_w|}{|\mathbf{SSCP}_t|}$$

$$F = \left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n_1+n_2-p-1}{p}\right) \sim F\left(p, n_1+n_2-p-1\right)$$

- Classification based on decision theory: assign the observation to group 1 if

$$Z \geq \frac{\bar{Z}_1 + \bar{Z}_2}{2} + \ln\left[\frac{p_2 C(1|2)}{p_1 C(2|1)}\right]$$

Logistic regression

- Odds of the event $Y = 1$

$$odds = \frac{p}{1-p}$$

  where

$$p = P(Y = 1)$$

- Probability of the event $Y = 1$ as a function of the explanatory variables

$$P(Y = 1 | X_1, X_2, \ldots, X_k) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$$

Quadratic equation

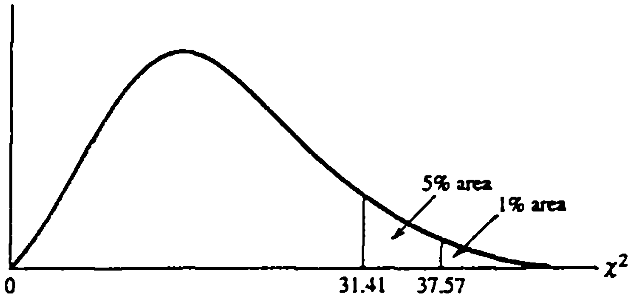
- The roots of the quadratic equation $ax^2 + bx + c$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Table T.3 χ^2 Critical Points

Example

$\Pr(\chi^2 > 23.8277) = 0.25$
$\Pr(\chi^2 > 31.4104) = 0.05$
for $df = 20$
$\Pr(\chi^2 > 37.5662) = 0.01$



df \ Pr	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944	10.828
2	2.77259	4.60517	5.99146	7.37776	9.21034	10.5966	13.816
3	4.10834	6.25139	7.81473	9.34840	11.3449	12.8382	16.266
4	5.38527	7.77944	9.48773	11.1433	13.2767	14.8603	18.467
5	6.62568	9.23636	11.0705	12.8325	15.0863	16.7496	20.515
6	7.84080	10.6446	12.5916	14.4494	16.8119	18.5476	22.458
7	9.03715	12.0170	14.0671	16.0128	18.4753	20.2777	24.322
8	10.2189	13.3616	15.5073	17.5345	20.0902	21.9550	26.125
9	11.3888	14.6837	16.9190	19.0228	21.6660	23.5894	27.877
10	12.5489	15.9872	18.3070	20.4832	23.2093	25.1882	29.588
11	13.7007	17.2750	19.6751	21.9200	24.7250	26.7568	31.264
12	14.8454	18.5493	21.0261	23.3367	26.2170	28.2995	32.909
13	15.9839	19.8119	22.3620	24.7356	27.6882	29.8195	34.528
14	17.1169	21.0641	23.6848	26.1189	29.1412	31.3194	36.123
15	18.2451	22.3071	24.9958	27.4884	30.5779	32.8013	37.697
16	19.3689	23.5418	26.2962	28.8454	31.9999	34.2672	39.252
17	20.4887	24.7690	27.5871	30.1910	33.4087	35.7185	40.790
18	21.6049	25.9894	28.8693	31.5264	34.8053	37.1565	42.312
19	22.7178	27.2036	30.1435	32.8523	36.1909	38.5823	43.820
20	23.8277	28.4120	31.4104	34.1696	37.5662	39.9968	45.315
21	24.9348	29.6151	32.6706	35.4789	38.9322	41.4011	46.797
22	26.0393	30.8133	33.9244	36.7807	40.2894	42.7957	48.268
23	27.1413	32.0069	35.1725	38.0756	41.6384	44.1813	49.728
24	28.2412	33.1962	36.4150	39.3641	42.9798	45.5585	51.179
25	29.3389	34.3816	37.6525	40.6465	44.3141	46.9279	52.618
26	30.4346	35.5632	38.8851	41.9232	45.6417	48.2899	54.052
27	31.5284	36.7412	40.1133	43.1945	46.9629	49.6449	55.476
28	32.6205	37.9159	41.3371	44.4608	48.2782	50.9934	56.892
29	33.7109	39.0875	42.5570	45.7223	49.5879	52.3356	58.301
30	34.7997	40.2560	43.7730	46.9792	50.8922	53.6720	59.703
40	45.6160	51.8051	55.7585	59.3417	63.6907	66.7660	73.402
50	56.3336	63.1671	67.5048	71.4202	76.1539	79.4900	86.661
60	66.9815	74.3970	79.0819	83.2977	88.3794	91.9517	99.607
70	77.5767	85.5270	90.5312	95.0232	100.425	104.215	112.317
80	88.1303	96.5782	101.879	106.629	112.329	116.321	124.839
90	98.6499	107.565	113.145	118.136	124.116	128.299	137.208
100	109.141	118.498	124.342	129.561	135.807	140.169	149.449
$Z^{\dagger}$	+0.6745	+1.2816	+1.6449	+1.9600	+2.3263	+2.5758	+3.0902

†For df greater than 100, the expression

$$\sqrt{2\chi^2} - \sqrt{(2k-1)} = Z$$

follows the standardized normal distribution, where k represents the degrees of freedom.

Source: From E. S. Pearson and H. O. Hartley, eds., *Biometrika Tables for Statisticians*, vol. 1, 3d ed., table 8, Cambridge University Press, New York, 1966. Reproduced by permission of the editors and trustees of *Biometrika*.