Manipal Academy of Higher Education, Manipal
Department of Statistics, PSPH
III Semester M.Sc Biostatistics
Subject: Applied Multivariate Analysis (MBS 701)
Sessional I

Dur: 120mins | 17.09.2018 | Max: 50 marks

Answer all the Questions:

1. A sample data of $n = 5$ has been collected on the variables 'Age', 'Height' and 'Weight' as given in the following table.

Age (in years)	Height (in cm)	Weight (in kg)
33	173	54
36	175	53
41	165	63
50	164	56
30	151	68

   (a) What is the sample mean vector of the data set? (3)

   (b) Find the sample covariance matrix of the data set. (4)

   (c) Define generalized sample variance and total sample variance. Also find the generalized sample variance and total variance for the given data set. (3)

2. Consider the data set collected on the variables x_1, x_2.

$$X = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 3 & 2 \end{bmatrix}.$$

   The variance-covariance matrix is given by $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. Let $z_1 = -x_1 + 2x_2$ and $z_2 = 2x_1 + 3x_2$.

   (a) Evaluate the sample mean and variance of z_1. (2)

   (b) Evaluate the sample mean and variance of z_2. (2)

   (c) Evaluate the covariance between z_1 and z_2. (2)

   (d) Mr. Nayan had found that 1.5 and 0.5 are the eigenvalues and, $(0.70, 0.70)$ and $(-0.70, 0.70)$ are the eigenvectors of the sample covariance matrix of the given data set respectively to help you. Now sketch the 95% confidence region for the given data. (Assume that the random vector (x_1, x_2) follows bivariate normal distribution) (4)

3. Suppose the random variables x_1, x_2 and x_3 have the covariance matrix

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

This time Ms. Nupur found the eigenvalues and corresponding normalized eigenvectors for you (You are so lucky!!!).

$$\lambda_1 = 5.83, \quad e_1 = (0.383, -0.924, 0)$$
$$\lambda_2 = 2.00, \quad e_2 = (0, 0, 1)$$
$$\lambda_3 = 0.17, \quad e_3 = (0.924, 0.383, 9)$$

(a) What are the principal components? (2)

(b) Is the covariance between any two principal components zero? Justify your answer. (3)

(c) I say the variance of the first principal component is 10. Do you agree? If not, justify your answer. (2)

(d) Vasanth Anna had come and told me that the second principal component and the second variable x_2 are positively correlated. Is that true? Justify your answer. (3)

4. A consultee from dental college is sitting in our round table and waiting for you. Ms. Lintu, who is invigilating now, had collected the following informations for you.

(i) Sample size is 101

(ii) There are 5 continuous variables

(iii) The correlation matrix of the 5 variables is

$$R = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{pmatrix} 1 & & & & \\ 0.58 & 1 & & & \\ 0.51 & 0.60 & 1 & & \\ 0.39 & 0.39 & 0.44 & 1 & \\ 0.46 & 0.32 & 0.43 & 0.52 & 1 \end{pmatrix} \end{array}$$

(iv) Eigenvalues of R are 2.86, 0.81, 0.54, 0.45, 0.34

(v) $\det(R) = 0.19$ (and $\log_e 0.19 = -1.6$)

(a) Test the correlation matrix for the principal component analysis using an appropriate test at 5% level of significance. (5)

(b) Use 2 different criteria to decide how many principal components to be retained? (5)

Note: If you use total variance to decide how many principal components to be retained, the consultee wants 90% of the total variability or more to be explained.

5. A 12-year-old girl made five ratings on a 9-point semantic differential scale for each of seven of her acquaintances. The ratings were based on the five adjectives kind, intelligent, happy, likeable, and just. The correlation matrix for the five variables (adjectives) is as follows:

$$\begin{array}{c} \\ \text{kind} \\ \text{intelligent} \\ \text{happy} \\ \text{likeable} \\ \text{just} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{kind} & \text{intelligent} & \text{happy} & \text{likeable} & \text{just} \end{array} \\ \begin{pmatrix} 1 & & & & \\ 0.300 & 1 & & & \\ 0.880 & -0.022 & 1 & & \\ 0.990 & 0.330 & 0.870 & 1 & \\ 0.540 & 0.840 & 0.130 & 0.540 & 1 \end{pmatrix} \end{array}$$

and the eigenvalues and eigenvectors of the correlation matrix is given below.

Eigenvalues	Eigenvectors
3.300	$(-0.540, -0.190, -0.190, -0.120, 0.790)^T$
1.500	$(-0.290, 0.650, 0.680, -0.120, 0.100)^T$
0.170	$(-0.430, -0.470, 0.410, 0.610, -0.210)^T$
0.030	$(-0.540, -0.170, -0.095, -0.630, -0.530)^T$
0.0002	$(-0.390, 0.540, -0.570, 0.440, -0.200)^T$

(a) What is an orthogonal factor model? (3)

(b) How many factors to be retained if one wants to retain the factors if the cummulative proportion of variance explained is more than 90% of the total variance? (3)

(c) Using the part (b), tabularize the loadings, cummunalities and specific variances, and also report the eigenvalues and cumulative proportion of variance explained by the factors. (4)

MANIPAL ACADEMY OF HIGHER EDUCATION, MANIPAL
MBS 701 Answer key
Sessional I

1. (a) $\begin{pmatrix} 38 \\ 166 \\ 59 \end{pmatrix}$

1. (b)

	Age	Height	Weight
Age	62	10	−15
Height	10	90	−56
Weight	−15	−56	42

1. (c) Generalized sample variance is defined as the determinant of the variance-covariance matrix and the total sample variance is defined as the trace of the sample variance-covariance matrix.
For the given data,

$$\text{Generalized sample variance } = \det(S) = 32873$$
$$\text{Total sample variance } = \text{Trace}(S) = 194$$

2. (a) $z_1 = a^T x$ where $a = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. Therefore,

$$\text{sample Mean of } z_1 = E(a^T x) = a^T E(x) = \frac{1}{3} a^T X^T e = \frac{1}{3}\begin{pmatrix} -1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \frac{6}{3} = 2$$

and

$$\text{Sample Variance of } z_1 = \text{cov}\left(a^T x, a^T x\right) = a^T \text{cov}(x)\ a = 3$$

2. (b) $z_2 = b^T x$ where $b = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. Therefore,

$$\text{sample Mean of } z_2 = E(b^T x) = b^T E(x) = \frac{1}{3} b^T X^T e = \frac{1}{3}\begin{pmatrix} 2 & 3 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{3} \times 30 = 10$$

and

$$\text{Sample Variance of } z_2 = \text{cov}\left(b^T x, b^T x\right) = b^T \text{cov}(x)\ b = 19.$$

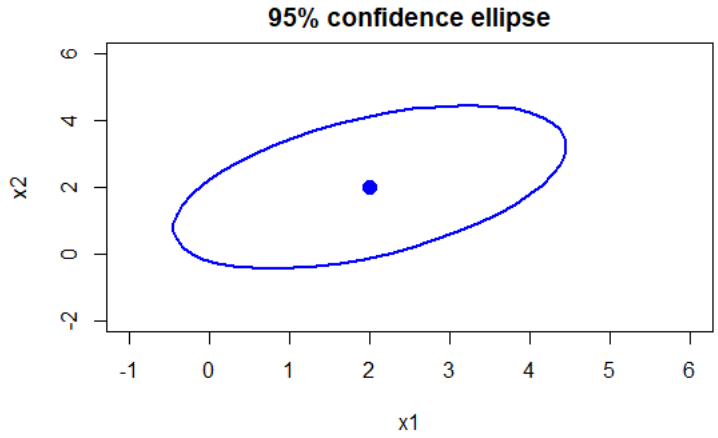
2. (c)

$$\text{cov}(z_1, z_2) = \text{cov}(a^T x, b^T x) = a^T \text{cov}(x) b = \begin{pmatrix} -1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 3 \end{pmatrix} = 30$$

and

$$\text{Sample Variance of } z_2 = \text{cov}\left(b^T x, b^T x\right) = b^T \text{cov}(x)\ b = 4.5.$$

2. (d)



3.(a) The principal components are

$$z_1 = 0.383x_1 - 0.924x_2$$
$$z_2 = x_3$$
$$z_3 = 0.924x_1 + 0.383x_2 + 9x_3$$

3.(b) Yes. For any i and j, $i \neq j$,

$$\text{cov}(z_i, z_j) = \text{cov}(e_i^T x, e_j^T x)$$

where e_k is the eigenvector of Σ corresponding to λ_k for $k = i, j$. Therefore,

$$\text{cov}(z_i, z_j) = e_i^T \text{cov}(x, x) e_j = e_i^T \Sigma e_j = e_i^T \lambda_j e_j = 0$$

since any two distinct eigenvectors of Σ are orthogonal to each other. Hence the proof.

3.(c) No. Variance of the first principal component is 5.83 since

$$\begin{aligned}
\text{var}(z_1) &= \text{cov}(e_1^T x, e_1^T x) \text{ where } e_1 \text{ is the eigenvector of } \Sigma \\
&= e_1^T \text{cov}(x) e_1 \\
&= e_1^T \Sigma e_1 \\
&= e_1^T \lambda_1 e_1 \text{ since } e_1 \text{ is the eigenvector of } \Sigma \text{ corresponding to } \lambda_1 \\
&= \lambda_1 \\
&= 5.83
\end{aligned}$$

Note: If you have proved for general i, dont worry, you will be getting marks.

3.(d) No. They are uncorrelated since

$$\text{cor}(x_2, z_2) = \frac{\sqrt{\lambda_2} e_{22}}{\sqrt{s_{22}}} = \frac{\sqrt{2} \times 0}{\sqrt{5}} = 0.$$

4.(a) **Bartlett's test:**

Hypothesis: $H_0 : R = I$ vs $H_1 : R \neq I$

Level of significance: 5%

Test statistic: $-\left[(n-1)-\left(\frac{2p+5}{6}\right)\right]\log_e|R| \sim \chi^2_{p(p-1)/2}(0.05)$

$$-\left[(n-1)-\left(\frac{2p+5}{6}\right)\right]\log_e|R| = -\left[100-\left(\frac{2\times 5+5}{6}\right)\right]\log_e 0.19$$
$$= -\left[100-\frac{15}{6}\right]\times -1.6$$
$$= 156$$

and

$$\chi^2_{p(p-1)/2}(0.05) = \chi^2_{\frac{7\times 6}{2}}(0.05) = \chi^2_{21}(0.05) = 32.67.$$

Decision: Since the computed value = 156 > 32.67 = Table value of $\chi^2_{21}(0.05)$, we reject the null hypothesis.

4.(b) Using any two is suffice.

5. (a) Consider $y_1,\ldots,y_n$ be the observations collected on p random variables $x_1,\ldots,x_p$. Let $x = (x_1,\ldots,x_p)$ and $\mu = (\mu_1,\ldots,\mu_p)$ be the random vector and the mean vector respectively. Let Σ be the variance-covariance matrix of x. A factor model

$$x - \mu = Lf + \epsilon$$

where L is the loading matrix, $f_{m\times 1}$ is the factor vector and ϵ is the error vector, is said to be orthogonal factor model if the following assumptions are hold.

(i) $E(f) = 0$, $\text{cov}(f) = I$

(ii) $E(\epsilon) = 0$, $\text{cov}(\epsilon) = I$

(iii) $\text{cov}(f,\epsilon) = 0$.

5. (b)

Eigenvalue	Proportion of variance explained	Cummulative proportion of variance explained
3.263	0.653	0.653
1.538	0.308	0.960
0.168	0.034	0.994
0.030	0.006	1.000
0.0002	0.00005	1

From the above table, it is clear that only the first two factors explain more than 90% of the variance and hence we retain the two factors.

5. (c)

Variables	Loadings		Communalities h_i^2	Specific Variances ψ_i
	f_1	f_2		
Kind	0.969	−0.231	0.993	0.007
Intelligent	0.519	0.807	0.921	0.079
Happy	0.785	−0.587	0.960	0.040
Likeable	0.971	−0.210	0.987	0.013
Just	0.704	0.667	0.940	0.060
Eigenvalues	3.263	1.538		
Proportion of total variance	0.653	0.308		
Cumulative proportion	0.653	0.960		