STOCKHOLM UNIVERSITY
Department of Statistics
Mahmood Ul Hassan

# EXAM IN MULTIVARIATE METHODS
# February 17 2020

---

Time: 5 hours
Aids allowed: Pocket calculator, language dictionary.
The exam consists of five questions. To score maximum points on a question solutions need to be clear, detailed and well-motivated.

---

**Question. 1** (4+4+4+4=16 Points)

Define and describe the following:

a) Centroid method
b) Scree plot
c) Deviance
d) Statistical distance

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

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

$$S = \begin{bmatrix} 65.41 & 4.57 \\ 4.57 & 1.27 \end{bmatrix}$$

a) Find the correlation matrix (R).
b) Using R, construct two principal components that are orthogonal to each other.
c) What proportion of variance is accounted by these principal components?
d) Compute the loadings of the variables.
e) What is the difference between principal component analysis and exploratory factor analysis?

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

A two-factor model represented by the following equations.

$$x_1 = 0.104F_1 + 0.824F_2 + U_1$$
$$x_2 = 0.065F_1 + 0.959F_2 + U_2$$
$$x_3 = 0.065F_1 + 0.725F_2 + U_3$$
$$x_4 = 0.906F_1 + 0.134F_2 + U_4$$
$$x_5 = 0.977F_1 + 0.116F_2 + U_5$$
$$x_6 = 0.827F_1 + 0.016F_2 + U_6$$

The usual assumptions hold for the above model

a) What are the pattern loading of indicators $x_1, x_4$ and $x_6$ on the factors $F_1$ and $F_2$ when
   (i) $cor(F1, F2) = \Phi_{12} = 0$ (ii) $cor(F1, F2) = \Phi_{12} = -0.3$
b) What are the degrees of freedom for the model when
   (i) $cor(F1, F2) = \Phi_{12} = 0$ (ii) $cor(F1, F2) = \Phi_{12} = -0.3$
c) Compute the correlation between $x_3$ and $x_6$ when
   (i) $cor(F1, F2) = \Phi_{12} = 0$ (ii) $cor(F1, F2) = \Phi_{12} = -0.3$
d) What percentage of the variance of indicators $x_3$ and $x_5$ is not accounted by the common factors $F_1$ and $F_2$ when
   (i) $cor(F1, F2) = \Phi_{12} = 0$ (ii) $cor(F1, F2) = \Phi_{12} = -0.3$
e) Find the structural loading of indicators $x_2, x_4$ and $x_5$ on the factors $F_1$ and $F_2$ when
   (i) $cor(F1, F2) = \Phi_{12} = 0$ (ii) $cor(F1, F2) = \Phi_{12} = -0.3$

**Question. 4** (8+8=16 Points)
For the following data

| Group-I | | Group-II | |
|---|---|---|---|
| $Y_1$ | $Y_2$ | $Y_1$ | $Y_2$ |
| 1 | 5 | 6 | 10 |
| 2 | 4.7 | 7 | 9.7 |
| 5 | 1 | 10 | 6 |
| 4 | 3.2 | 9 | 8.2 |
| 5 | 1 | 10 | 6 |
| 3 | 4.1 | 8 | 4.1 |

Within-group covariance matrix for group-II $= S_2 = \begin{bmatrix} 2.667 & -2.433 \\ -2.433 & 5.495 \end{bmatrix}$

Total sample covariance matrix $= S_t = \begin{bmatrix} 9.242 & 3.318 \\ 3.318 & 8.685 \end{bmatrix}$

a) Compute the $\mathbf{SSCP_b}$ and $\mathbf{SSCP_w}$ matrices.
b) Calculate Fisher's linear discriminant function for this data set.

**Question. 5** (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 | 1 | 1 |
| 2 | 2 | 2 |
| 3 | 6 | 3 |
| 4 | 8 | 1 |
| 5 | 10 | 1 |

a) Construct a similarity matrix containing squared Euclidean distances
b) Use the similarity matrix in part (a) and perform a cluster analysis with the complete linkage method.
c) A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution (4 categories), effect admission into graduate school. The response variable, admit/don't admit, is a binary variable. The model fit is summarized below. Formulate the null and alternative hypothesis and perform testing of hypothesis using deviance statistic to compare the fitted model with the null model. Use $\alpha = 0.05$.

```
Coefficients:
                Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -3.989979   1.139951  -3.500   0.000465 ***
gre            0.002264   0.001094   2.070   0.038465 *
gpa            0.804038   0.331819   2.423   0.015388 *
factor(rank)2 -0.675443   0.316490  -2.134   0.032829 *
factor(rank)3 -1.340204   0.345306  -3.881   0.000104 ***
factor(rank)4 -1.551464   0.417832  -3.713   0.000205 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.52
```

d) Using the model in part (c), classification was performed and resulted in the confusion matrix

| | | Predicted | | |
|---|---|---|---|---|
| | | Admitted | Not admitted | Total |
| Observed | Admitted | 30 | --- | --- |
| | Not admitted | ---- | 254 | --- |
| | Total | 127 | --- | 400 |

Compute the missing values in the table and calculate the false positive, false negative, 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 $2 \times 2$ matrix $\mathbf{A}$

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

- Inverse of a $2 \times 2$ 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

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

- 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

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

## Factor Analysis

- For the two-factor model

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

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

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

- 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

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

**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' $\Lambda$

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

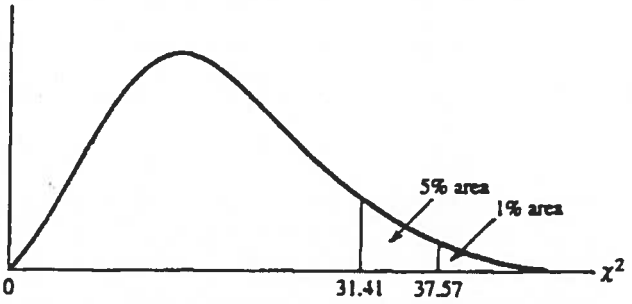
$$F = \left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n_1 + n_2 - p - 1}{p}\right) \sim F\left(p, n_1 + n_2 - p - 1\right)$$

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

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

## *Table T.3* $\chi^2$ 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*.

Department of Statistics

Stockholms universitet

# Correction sheet

**Date:** 200217

**Room:** Laduvikssalen

**Exam:** Multivariate Methods

**Course:** Multivariate Methods

**Anonymous code:** 0008-EYM

☒ I authorise the anonymous posting of my exam, in whole or in part, on the department homepage as a sample student answer.

**NOTE! ALSO WRITE ON THE BACK OF THE ANSWER SHEET**

**Mark answered questions**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Total number of pages |
|---|---|---|---|---|---|---|---|---|---|
| X | X | X | X | X | | | | H.A | 8 |
| Teacher's notes 16 | 15.5 | 14 | 06 | 15.5 | | | | 15 | |

| Points | Grade | Teacher's sign. |
|---|---|---|
| 82 | B | |

**SU, DEPARTMENT OF STATISTICS**

Room: Ladunit Anonymous code: 0008-EYM Sheet number: 1

1. Define and describe the following:

a) Centroid method

A hierarchical clustering method that follows the steps

1. Identify the smallest distance in the similarity matrix, and merge those subjects into a cluster.
2. For the new cluster, calculate the centroid

$$m_{ij} = \left( \frac{x_{i1}+x_{j1}}{2}, \frac{x_{i2}+x_{j2}}{2} \right)$$

4/4

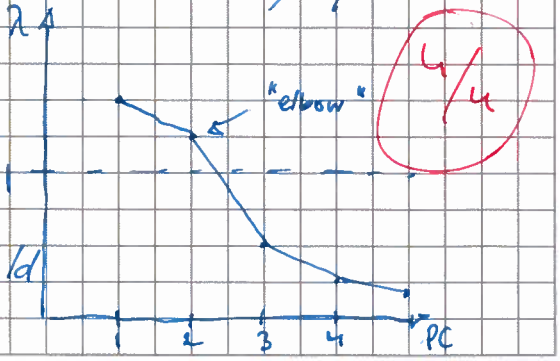
and use the centroid values to recalculate the distance between the cluster and the other subjects.

3. Go back to step 1 and keep on clustering subjects/clusters, calculating a new centroid for each new cluster, and the respective distances between subjects/clusters.

For each new cluster you will make a new similarity matrix of the recalculated distances. By using this method, the distance between clusters will be measured b/w the "cluster means".

b) Scree plot

A scree plot is a plot of the principal components (x-axis) versus its eigen values (y-axis). It's used in principal component analysis to visually determine how many PC:s to keep for explaining the original variables. Usually you look for the "elbow", this is where there is a big decrease in how much variance the next PC will be able to explain. In my example, we would keep 2 PC:s.




4/4

c) Deviance

A statistic used in logistic regression to measure the goodness-of-fit of a model using the likelihood functions. It's defined as

$$-2\ln\left(\frac{L(\text{fitted model})}{L(\text{null model})}\right) \overset{\text{approx}}{\sim} \chi^2(n-q)$$

and it's approximately $\chi^2$-distributed with $df = n-q$, where $n$ = df in null model and $q$ = df in the fitted model. By computing the deviance one can assess a models statistical significance, or by calculating the difference of the deviances for two models one can determine which model has a better fit.

d) Statistical distance

The statistical distance is a special case of the Mahalanobis distance, where we do not account for correlation between variables. It's defined as:

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

What we do is we scale the distances by dividing them with the standard deviation $s_j$. This makes it possible to compare distances between observations on variables with different sized variance/standard deviation because we have adjusted the distances to have the same variance/standard deviation.

**SU, DEPARTMENT OF STATISTICS**

Room: dadunik Anonymous code: 0008-EYM Sheet number: 2

2. For a data set with obervrations on two variables, $x_1$ and $x_2$

Sample covariance matrix

$$S = \begin{pmatrix} 65.41 & 4.57 \\ 4.57 & 1.27 \end{pmatrix} \qquad s_{11} = s_1^2 = 65.41 \qquad s_{22} = s_2^2 = 1.27 \qquad s_{12} = s_{21} = 4.57$$

a) Find the correlation matrix (R)

This is the "standardized version" of S and we need to calculate $r_{12}$

$$r_{12} = \frac{s_{12}}{s_1 s_2} = \frac{4.57}{\sqrt{65.41 \cdot 1.27}} = 0.50$$ 2/2

We then get the correlation matrix

$$R = \begin{pmatrix} 1 & 0.50 \\ 0.50 & 1 \end{pmatrix}$$

b) Construct two principal components (orthogonal to each other)

$$\mathcal{E}_1 = w_{11} x_1 + w_{12} x_2$$
$$\mathcal{E}_2 = w_{21} x_1 + w_{22} x_2$$

We start by finding the eigen values, which are the roots of the charactenistic equation $\det(R - \lambda I) = 0$

$$R - \lambda I = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1-\lambda & 0.5 \\ 0.5 & 1-\lambda \end{pmatrix}$$

$$\det(R-\lambda I) = 0 \iff (1-\lambda)(1-\lambda) - 0.5 \cdot 0.5 = 0$$
$$(1-\lambda)^2 - 0.25 = 0$$
$$1 - 2\lambda + \lambda^2 - 0.25 = 0$$
$$\lambda^2 - 2\lambda + 0.75 = 0$$

We get the roots of the equation

$$\lambda_1 = \frac{-(-2) + \sqrt{(-2)^2 - 4 \cdot 0.75}}{2 \cdot 1} = \frac{2 + \sqrt{1}}{2} = 1.5$$

$$\lambda_2 = \frac{-(-2) - \sqrt{(-2)^2 - 4 \cdot 0.75}}{2 \cdot 1} = \frac{2 - \sqrt{1}}{2} = 0.5$$

We then find the associated eigen vector $r$ for each eigen value, by solving $(R - \lambda I)x = 0$ for $x$.

$$\begin{pmatrix} 1-\lambda_1 & 0.5 \\ 0.5 & 1-\lambda_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} -0.5 & 0.5 \\ 0.5 & -0.5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$-0.5x_1 + 0.5x_2 = 0$$
$$0.5x_1 - 0.5x_2 = 0$$

Set $x_1 = 1$

$$0.5 + 0.5x_2 = 0$$
$$0.5x_2 = -0.5$$
$$x_2 = -1$$

The associated eigen vector is $r_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

Normalizing it $r_1 = \begin{pmatrix} \frac{1}{\sqrt{1^2+1^2}} \\ \frac{-1}{\sqrt{1^2+1^2}} \end{pmatrix} = \begin{pmatrix} 0.7071 \\ -0.7071 \end{pmatrix}$

**SU, DEPARTMENT OF STATISTICS**

Room: Dadunik Anonymous code: 0008-EYM Sheet number: 3

2.b) cont'd Same thing for $\lambda_2$

$$\begin{pmatrix} 1-\lambda_2 & 0.5 \\ 0.5 & 1-\lambda_2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$0.5x_1 + 0.5x_2 = 0$ 6/6

Set $x_1 = 1$

$0.5 + 0.5x_2 = 0$

$0.5x_2 = 0.5$

$x_2 = -1$

The associated eigen vector is $v_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

Normalizing it $v_2 = \begin{pmatrix} \frac{1}{\sqrt{1^2+1^2}} \\ \frac{1}{\sqrt{1^2+1^2}} \end{pmatrix} = \begin{pmatrix} 0.7071 \\ 0.7071 \end{pmatrix}$ ✓

Then we have the principal components:

$\mathcal{E}_1 = 0.7071x_1 \mp 0.7071x_2$

$\mathcal{E}_2 = +0.7071x_1 + 0.7071x_2$ ✓ ← take w.r.n sign (+,-)

Based on the weights from the eigen vectors.

c) Calculate the proportion of variance accounted for by each principal component.

$Var(\varepsilon_i) = \lambda_i$

Total variance $= \lambda_1 + \lambda_2 = 2$ (because standardized data)

| | Variance | Proportion of variance |
|---|---|---|
| PC1 | $\lambda_1 = 1.5$ | $\lambda_1 / \lambda_1 + \lambda_2 = 1.5/2 = 0.75$ |
| PC2 | $\lambda_2 = 0.5$ | $\lambda_2 / \lambda_1 + \lambda_2 = 0.5/2 = 0.25$ |

✓ 2/2

d) Compute the loadings of the variables

$\left(\begin{matrix} i\text{th variable} \\ j\text{th PC}\end{matrix}\right) l_{ij} = \frac{w_{ij}\sqrt{\lambda_j}}{s_j}$, where $s_j = 1$ (because standardized data)

$l_{11} = w_{11}\sqrt{\lambda_1} = 0.7071 \cdot \sqrt{1.5} = 0.8660$

$l_{12} = w_{12}\sqrt{\lambda_2} = -0.7071 \cdot \sqrt{0.5} = -0.5000$

$l_{21} = w_{21}\sqrt{\lambda_1} = +0.7071 \cdot \sqrt{1.5} = +0.8660$

$l_{22} = w_{22}\sqrt{\lambda_2} = 0.7071 \cdot \sqrt{0.5} = 0.5000$

2.5/3

e) What is the difference between principal component analysis (PCA) and exploratory factor analysis (EFA)?

PCA forms linear combinations of the original variables, in such a way that they account for the greatest amount of variance in the original data, with the first component accounting for the biggest part of variance, the second component accounting for the second biggest part of the variance, and so on. That is: $Var(\varepsilon_1) > Var(\varepsilon_2) > \dots > Var(\varepsilon_p)$

This way it is possible to use a few principal components in analysis, instead of many original variables, while still accounting for most of the variance in the original data.

2.e) cont'd. EFA is also a data reduction techniques, but here the objective is to find a factor structure that explains the covariation among variables. That is, we want to determine how a number of indicator variables best can work to describe some common factors. In EFA we also know that there will be some variation we will not be able to account for, and this is held in the unique factor U.

3/3

15.5/16

3. Assume the following two-factor model (all usual assumptions hold)

$$x_1 = 0.104F_1 + 0.824F_2 + U_1$$
$$x_2 = 0.065F_1 + 0.959F_2 + U_2$$
$$x_3 = 0.065F_1 + 0.725F_2 + U_3$$
$$x_4 = 0.906F_1 + 0.134F_2 + U_4$$
$$x_5 = 0.977F_1 + 0.116F_2 + U_5$$
$$x_6 = 0.827F_1 + 0.016F_2 + U_6$$

a) What are the pattern loadings of indicators $x_1, x_4, x_6$ on $F_1$ and $F_2$?

Since pattern loading $= \lambda_{ij}$, answers will be the same for (i)(ii)

(i)(ii)

| | $F_1(\lambda_{i1})$ | $F_2(\lambda_{i2})$ |
|---|---|---|
| $x_1$ | 0.104 | 0.824 |
| $x_4$ | 0.906 | 0.134 |
| $x_6$ | 0.827 | 0.016 |

✓ 2/2

b) What are the degrees of freedom for the model? $df = \frac{p(p+1)}{2} - q$

(i) When $Cor(F_1, F_2) = \phi_{12} = 0$

$p$ = # of parameters to estimate = 21 $\quad q$ = # of parameters in model = 12

$$df = \frac{p(p+1)}{2} - q = \frac{21 \cdot 22}{2} - 12 = 219$$

18

(ii) When $Cor(F_1, F_2) = \phi_{12} = -0.3$ 0/2

$p = 22 \quad q = 12$

$$df = \frac{p(p+1)}{2} - q = \frac{22 \cdot 23}{2} - 12 = 241$$

19

3. c) (cont'd) Compute the correlations between $x_3$ and $x_6$

(i) When $\phi_{12} = 0$

$$Cor(x_3, x_6) = \lambda_{31}\lambda_{61} + \lambda_{32}\lambda_{62} + (\lambda_{31}\lambda_{62} + \lambda_{32}\lambda_{61})\phi_{12}$$
$$= 0.065 \cdot 0.827 + 0.725 \cdot 0.016 + 0$$
$$= 0.0654$$ ✓

(ii) When $\phi_{12} = -0.3$

$$Cor(x_3, x_6) = \lambda_{31}\lambda_{61} + \lambda_{32}\lambda_{62} + (\lambda_{31}\lambda_{62} + \lambda_{32}\lambda_{61})\phi_{12}$$
$$= 0.065 \cdot 0.827 + 0.725 \cdot 0.016 + (0.065 \cdot 0.016 + 0.725 \cdot 0.827)(-0.3)$$
$$= -0.1148$$ ✓

4/4

d) What percentage of the variance of indicators $x_3$ and $x_5$ is not accounted for by $F_1$ and $F_2$

The variance not accounted for by the factors is the unique variance $Var(U_j) = Var(x_j) - (\lambda_1^2 + \lambda_2^2 + 2\lambda_1\lambda_2\phi_{12})$

Where $Var(x_j) = 1$ (because standardized data)

(i) When $\phi_{12} = 0$

$$Var(U_3) = 1 - (0.065^2 + 0.725^2 + 0)$$
$$= 0.4702 \Rightarrow 47.02\%$$ ✓

$$Var(U_5) = 1 - (0.977^2 + 0.116^2 + 0)$$
$$= 0.0320 \Rightarrow 3.20\%$$ ✓

4/4

(ii) When $\phi_{12} = -0.3$

$$Var(U_3) = 0.4702 - (2 \cdot 0.065 \cdot 0.725 \cdot (-0.3))$$
$$= 0.4984 \Rightarrow 49.84\%$$ ✓

$$Var(U_5) = 0.032 - (2 \cdot 0.977 \cdot 0.116 \cdot (-0.3))$$
$$= 0.1000 \Rightarrow 10.00\%$$ ✓

e) Find the structural loading of indicators $x_2$, $x_4$ and $x_5$ on $F_1$ and $F_2$

Since the structural loading is given by $Cor(x_j, F_i) = \lambda_{j1} + \lambda_{j2}\phi_{12}$ the structure loading will be the same as the pattern loading when $\phi_{12} = 0$. That is

(i)(ii)

| | $Cor(x_j, F_1)$ $\phi_{12}=0$ | $Cor(x_j, F_2)$ $\phi_{12}=0$ | $Cor(x_j, F_1)$ $\phi_{12}=-0.3$ | $Cor(x_j, F_2)$ $\phi_{12}=-0.3$ |
|---|---|---|---|---|
| $X_2$ | $\lambda_{21}=0.065$ | $\lambda_{22}=0.959$ | $0.065+0.959(-0.3)=-0.2227$ | $0.959+0.065(-0.3)=0.9395$ |
| $X_4$ | $\lambda_{41}=0.906$ | $\lambda_{42}=0.134$ | $0.906+0.134(-0.3)=0.8658$ | $0.134+0.906(-0.3)=-0.1378$ |
| $X_5$ | $\lambda_{51}=0.977$ | $\lambda_{52}=0.116$ | $0.977+0.116(-0.3)=0.9412$ | $0.116+0.977(-0.3)=-0.1771$ |

4/4

**SU, DEPARTMENT OF STATISTICS**

Room: Radunik Anonymous code: 0008-EYM Sheet number: 6

4. Data set

| Group I | | Group II | |
|---|---|---|---|
| $Y_1$ | $Y_2$ | $Y_1$ | $Y_2$ |
| 1 | 5 | 6 | 10 |
| 2 | 4.7 | 7 | 9.7 |
| 5 | 1 | 10 | 6 |
| 4 | 3.2 | 9 | 8.2 |
| 5 | 1 | 10 | 6 |
| 3 | 4.1 | 8 | 4.1 |

$n_1 = 6$

$n_2 = 6$

$\bar{X}_1 = \begin{pmatrix} 3.3333 \\ 8.3333 \end{pmatrix}$

$\bar{X}_2 = \begin{pmatrix} 3.1667 \\ 7.3333 \end{pmatrix}$

$\bar{X} = \begin{bmatrix} \quad \end{bmatrix}$ Grand Mean Vector

a) Compute the $SSCP_B$ and $SSCP_W$ matrices

$$SSCP_B = \frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)(\bar{X}_1 - \bar{X}_2)^T$$

$$= \frac{36}{12} \left( \begin{pmatrix} 3.3333 \\ 8.3333 \end{pmatrix} - \begin{pmatrix} 3.1667 \\ 7.3333 \end{pmatrix} \right) \left( \begin{pmatrix} 3.3333 \\ 8.3333 \end{pmatrix} - \begin{pmatrix} 3.1667 \\ 7.3333 \end{pmatrix} \right)^T$$

$$= 3 \begin{pmatrix} 0.1666 \\ 1.0000 \end{pmatrix} \begin{pmatrix} 0.1666 & 1.0000 \end{pmatrix}$$

$$= \begin{pmatrix} 0.0833 & 0.4998 \\ 0.4998 & 3.0006 \end{pmatrix}$$ X 0/8

$$SSSC_W = SSCP_T - SSCP_B$$

$$= \begin{pmatrix} 9.242 & 3.318 \\ 3.318 & 8.685 \end{pmatrix} - \begin{pmatrix} 0.0833 & 0.4998 \\ 0.4998 & 3.0000 \end{pmatrix}$$

$$= \begin{pmatrix} 9.1587 & 2.8182 \\ 2.8182 & 5.6850 \end{pmatrix}$$

b) Calculate Fisher's linear discriminant function, which is given by: $\gamma^T = (\mu_1 - \mu_2)^T \Sigma^{-1}$, and estimated as

$$\gamma^T = (\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1}$$

Here $\bar{x}_1 = 5.8333$

$\bar{x}_2 = 5.25$

$$S_{pooled} = \frac{SSCP_W}{n-g} = \frac{1}{12-2}\begin{pmatrix} 9.1587 & 2.8182 \\ 2.8182 & 5.6850 \end{pmatrix}$$

$$= \begin{pmatrix} 0.91587 & 0.28182 \\ 0.28182 & 0.56850 \end{pmatrix}$$

$$\gamma^T = (5.8333 - 5.25)^T \frac{1}{0.91587 \cdot 0.56850 - 0.28182 \cdot 0.28182} \begin{pmatrix} 0.56850 & -0.28182 \\ -0.28182 & 0.91587 \end{pmatrix}$$

$$= \begin{pmatrix} 5.8333 \\ 5.25 \end{pmatrix} \begin{pmatrix} 1.0919 & -0.5413 \\ -0.5413 & 1.7590 \end{pmatrix}$$

$$= \begin{pmatrix} 5.8333 \cdot 1.0919 + 5.25(-0.5413) & 5.8333(-0.5413) + 5.25 \cdot 1.7590 \end{pmatrix}$$

$$= \begin{pmatrix} 3.5276 & 6.0772 \end{pmatrix}$$

Which gives

$$\gamma = \begin{pmatrix} 3.5276 \\ 6.0772 \end{pmatrix}$$

5/8

**SU, DEPARTMENT OF STATISTICS**

Room: dadunik Anonymous code: 0008-EYM Sheet number: 7

5. a) Construct a similarity matrix containing squared euclidean distances.

$D_{12}^2 = (1-2)^2 + (1-2)^2 = 2$ ✓

$D_{13}^2 = (1-6)^2 + (1-3)^2 = 29$ ✓

$D_{14}^2 = (1-8)^2 + (1-1)^2 = 49$ ✓

$D_{15}^2 = (1-10)^2 + (1-1)^2 = 81$ ✓

$D_{23}^2 = (2-6)^2 + (2-3)^2 = 17$ ✓

$D_{24}^2 = (2-8)^2 + (2-1)^2 = 37$ ✓

$D_{25}^2 = (2-10)^2 + (2-1)^2 = 65$ ✓

$D_{34}^2 = (6-8)^2 + (3-1)^2 = 8$

$D_{35}^2 = (6-10)^2 + (3-1)^2 = 20$ ✓

$D_{45}^2 = (8-10)^2 + (1-1)^2 = 4$ ✓ ✓

4/4

Giving the matrix

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | | | | |
| 2 | 2 | 0 | | | |
| 3 | 29 | 17 | 0 | | |
| 4 | 49 | 37 | 8 | 0 | |
| 5 | 81 | 65 | 20 | 4 | 0 |

Symmetric

✓

b) Use the similarity matrix in a) to perform a cluster analysis with the complete linkage method.

Starting by searching for the smallest distance in the matrix, which is (1 2). We'll merge them into a cluster.

5b) contd. Updated similarity matrix:

| | (12) | 3 | 4 | 5 |
|---|---|---|---|---|
| (12) | 0 | | | |
| 3 | 29 | 0 | | |
| 4 | 49 | 8 | 0 | |
| 5 | 81 | 20 | 4 | 0 |

Calculating new distances:

$d_{(12)3} = \max(d_{13}, d_{23}) = 29$

$d_{(12)4} = \max(d_{14}, d_{24}) = 49$

$d_{(12)5} = \max(d_{15}, d_{25}) = 81$

Now, the smallest distance is that between (4,5), which will be merged into a new cluster

Updated similarity matrix:

| | (12) | 3 | (45) |
|---|---|---|---|
| (12) | 0 | | |
| 3 | 29 | 0 | |
| (45) | 81 | 20 | 0 |

Calculating new distances:

$d_{(12)(45)} = \max(d_{(12)4}, d_{(12)5}) = 81$

$d_{(45)3} = \max(d_{43}, d_{53}) = 20$

The smallest distance makes cluster (345)

Updated similarity matrix:

| | (12) | (345) |
|---|---|---|
| (12) | 0 | |
| (345) | 81 | 0 |

Calculating new distance

$d_{(12)(345)} = \max(d_{(12)3}, d_{(12)(45)}) = 81$

✓

Next step is to merge all subjects into one cluster (12345).

4/4

5.c) contd. The hypotheses to test the fitted model against the null model is

$H_0$: The model is a good fit to the data

$H_0$: The model is not a good fit to the data

The deviance statistic is defined as

$$D = -2 \ln \left( \frac{L(\text{fitted model})}{L(\text{null model})} \right)$$

The deviance statistic is approximately $\chi^2$ - distributed with df = # of parameters in model.

In this case we have $df = 5$, and the deviance statistic is already calculated for the null vs. the residual.

We can take the difference of the two to compare them ✓

$$D = 499.98 - 458.52 = 41.46$$

This is a big enough value to reject the null hypothesis on an $\alpha$-level of 0.05 and $df = 5$.

That is, the model is not a good fit. 3.5/4

↓ why not look at table value and conclude it

$D \geq \chi^2_{5,0.05} = 11.07$

Reject $H_0$

d) Classification table : Admitted = "event"

| Observed \ Predicted | Admitted | Not Admitted | Total |
|---|---|---|---|
| Admitted | 30 (TP) | 19 (FN) | 49 |
| Not Admitted | 97 (FP) | 254 (TN) | 351 |
| Total | 127 | 273 | 400 |

$$\text{False positive rate} = \frac{FP}{FP+TN} = \frac{97}{351} = 0.2764$$

$$\text{False negative rate} = \frac{FN}{TN+TP} = \frac{19}{49} = 0.3878$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} = \frac{30}{49} = 0.6122$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{254}{351} = 0.7236$$