Solutions

1. (30 Points). Please round your answers to two decimal places if applicable.

a) (2 Pts) Consider the following **R** output obtained from running PCA on a data set with four variables:

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	9	4	?	2
Proportion of Variance	0.736	0.145	?	0.036
Cumulative Proportion	0.736	0.881	?	?

Fill in the gaps in the output summary.

Answer:

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	9	4	3	2
Proportion of Variance	0.736	0.145	0.083	0.036
Cumulative Proportion	0.736	0.881	0.964	1

(0.5P.) for each right answer.

b) (4 Pts) Find the 1st Principal Component for the covariance matrix $\Sigma = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$. Compute the proportion of total variance explained by the 1st PC.

Answer: Find the eigenvalues of the matrix and sort them in the decreasing order: $\lambda_1 = 1.2$, $\lambda_2 = 0.8$. We need to compute the 1st PC. Hence, find the 1st eigenvector by solving $\begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix} \times \begin{bmatrix} a_{1,1} \\ a_{1,2} \end{bmatrix} = 1.2 \cdot \begin{bmatrix} a_{1,1} \\ a_{1,2} \end{bmatrix}$, get $a_{1,1} = a_{1,2}$, and consequently $a_1 \sim \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Want $||a_1|| = 1$, so choose $\begin{bmatrix} 0.7 \\ 0.7 \end{bmatrix}$. Then the 1st PC equals $\mathbf{X} \times \begin{bmatrix} 0.7 \\ 0.7 \end{bmatrix} = 0.7X_1 + 0.7X_2$.

$\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{1.2}{1.2+0.8} = 0.6$ or 60%

(3P.) for the 1st task, **(1P.)** for the 2nd. **NOTE: (4P) for the task for the correct computation of the 1st eigenvector (only if normalized!), even though the 1st PC wasn't given explicitly as ($PC1 = 0.7X_1 + 0.7X_2$). (-1P) if the 1st eigenvector is computed but not normalized. From (-0.5P) up to (-2P) for calculative errors depending on the severity.**

c) (4 Pts) Consider data with covariance matrix $\mathbf{\Sigma} = \begin{bmatrix} 5 & 2 \\ 2 & 1 \end{bmatrix}$ and mean vector $\mu = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$. Compute the Mahalanobis distance between the point $\mathbf{x} = \begin{bmatrix} 5 \\ 6 \end{bmatrix}$ and the data.

Assuming the data is q-variate normally distributed, which distribution does the squared Mahalanobis distance follow? Specify all parameters of this distribution.

Which diagnostic plot can be used to detect outliers via the squared Mahalanobis distance? In brief describe the procedure.

Answer: First, invert the covariance matrix: the determinant equals $5 - 4 = 1$. Hence, $\mathbf{\Sigma}^{-1} = \begin{bmatrix} 1 & -2 \\ -2 & 5 \end{bmatrix}$ and the squared Mahalanobis distance $\mathbf{D_M}^2 = \left(\begin{bmatrix} 5 \\ 6 \end{bmatrix} - \begin{bmatrix} 4 \\ 3 \end{bmatrix} \right)^T \begin{bmatrix} 1 & -2 \\ -2 & 5 \end{bmatrix} \left(\begin{bmatrix} 5 \\ 6 \end{bmatrix} - \begin{bmatrix} 4 \\ 3 \end{bmatrix} \right) = 34$. Take square root: $\mathbf{D_M} = \sqrt{34}$.

$\mathbf{D_M}^2 \sim \chi^2(q)$, with q degrees of freedom (equal the number of variables). In this case $q = 2$.

QQ-plot. Compute squared Mahalanobis distance and plot it versus the theoretical quantiles of chi-square distribution (in this case $\chi^2(2)$). If there are no outliers, the data will follow a straight line. **(2P.)** for the 1st question, **(1P.)** for the 2nd and **(1P.)** for the 3d.

d) **(2 Pts)** Write down the multifactor FA model in matrix-vector form. Specify what each of the matrices/vectors means. What are the assumptions of this model?

Answer:

$$\mathbf{X} = \mathbf{\Lambda f} + \mathbf{u},$$

where $\mathbf{X} = (X_1, \ldots, X_q)^T$ are the observed variables, $\mathbf{f} = (f_1, \ldots, f_k)^T$ are the latent/common factors, $\mathbf{u} = (u_1, \ldots, u_q)^T$ are the specific factors, $\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \ldots & \lambda_{1k} \\ \vdots & \ddots & \vdots \\ \lambda_{q1} & \ldots & \lambda_{qk} \end{bmatrix}$ matrix of factor loadings.

The assumptions of the model are:

1. $E(u) = 0$ and $Cov(u) = \Psi$ is a diagonal matrix,
2. $Cov(f_i, u_j) = 0$. Additionally,
3. $E(X) = 0$,
4. $E(f) = 0$, $Cov(f) = I \; \forall i, j$

(0.5P.) if only the main ones are mentioned.

e) **(3 Pts)** Consider a matrix $\mathbf{D}$ of Euclidean distances derived from the $n \times q$ data matrix $\mathbf{X}$, with $n > q$. $\mathbf{X}$ has a full rank. Let $\mathbf{B}$ be the $n \times n$ inner products matrix, i.e. $\mathbf{B} = \mathbf{XX}^T$. Describe in brief how to obtain the low dimensional representation of the coordinates $\mathbf{X}$ of points based on the matrix $\mathbf{D}$. You do not need to derive the elements of $\mathbf{B}$ in terms of Euclidean distances. You only need to show how the data matrix $\mathbf{X}$ can be approximated.

Answer: Consider the spectral decomposition of $\mathbf{B}$

$$\mathbf{B} = \mathbf{V\Lambda V}^T,$$

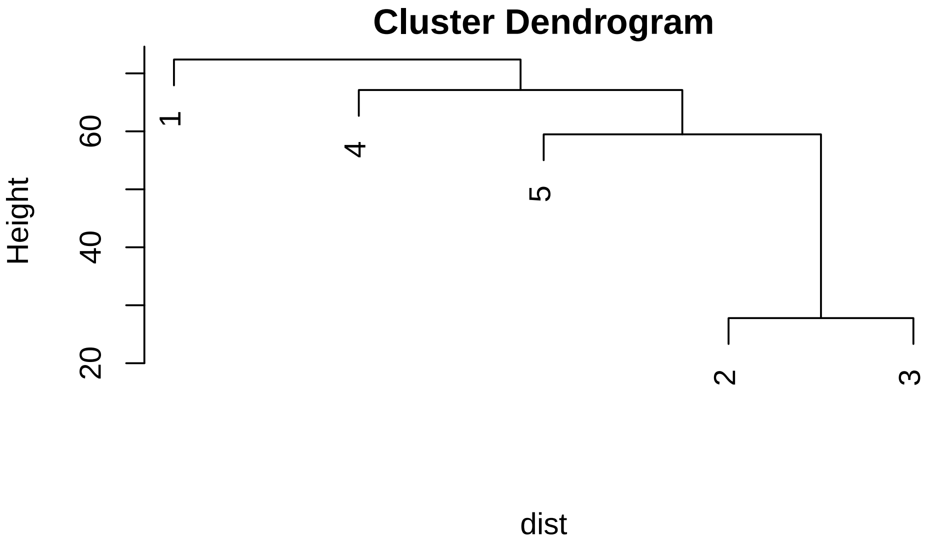
where $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix of eigenvalues of $\mathbf{B}$ with $\lambda_1 \geq \cdots \geq \lambda_n$ and $\mathbf{V} = (\mathbf{V_1}, \ldots, \mathbf{V_n})$ the corresponding matrix of eigenvectors, normalised s.t. $\mathbf{V}_i\mathbf{V}_i^T = 1$, i.e. $\mathbf{V}$ orthogonal. The rank of $\mathbf{B}$ is q, so that the last $n - q$ of its eigenvalues will be zero. Hence, can write $\mathbf{B} = \tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^T$, where $\tilde{\mathbf{V}}$ contains the 1st q eigenvectors and $\tilde{\mathbf{\Lambda}}$ the q non-zero eigenvalues. Then $\tilde{\mathbf{X}}$ is given by $\mathbf{X} = \tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}}^{\frac{1}{2}}$.

f) **(4 Pts)** Five cities are labelled by numbers 1, 2, 3, 4 and 5. The distances between the cities are given in the following matrix:

```
       1      2      3      4
2 102.68
3  86.33  27.78
4  72.39  75.01  67.13
5 104.17  59.48  62.41  79.61
```

Apply single-linkage hierarchical clustering to this data and draw a dendrogram of the clustering.

Answer: Correct procedure of choosing the minimal distances **(2P.)**



Correct dendogram **2P.**

g) **(4 Pts in total)** We perform statistical tests for several null hypotheses and obtain the following p-values: 0.001, 0.004, 0.03, 0.02, 0.01, 0.025 and 0.035.

a) **(2 Pt)** How many hypotheses are rejected at the 5% level after we apply the classical Bonferroni correction?

Answer: 2 hypotheses are rejected **(2P.)**

b) **(2 Pt)** How many hypotheses are rejected at the 5% level after we apply the Holm-Bonferroni correction?

Answer: 3 hypotheses are rejected **(2P.)**

h) **(4 Pts in total)** After training an LDA classifier on a dataset, we obtain a confusion matrix $\begin{pmatrix} 300 & 45 \\ 10 & 100 \end{pmatrix}$.

a) **(2 Pt)** What is the error rate?

Answer: (45+10)/(300+45+10+100)= 0.12. **(2P.)**

b) **(2 Pt)** Assume that we have 15 new samples that we need to classify. If we suppose that the misclassification probability for a single sample is $p = 0.15$, what is the probability that we misclassify 2 or fewer of the 15 samples?

Answer: Probability of misclassifying k samples is $\binom{15}{k} \cdot (1-p)^(15-k) \cdot p^k$ **(1P.)**. When summed up for 0, 1, 2 the probability is 0.60 **(1P.)**

i) **(3 Pts)** Assume that the variable y is binary, $y \in \{0, 1\}$ and that x_1, x_2 are the predictor variables. After fitting a logistic regression model in R, the output below is obtained. What is the expectation of y if we know that $x_1 = 1$ and $x_2 = -0.75$?

```
> glm(formula = y ~ x1 + x2, family = "binomial", data = training_data)

Coefficients:
              Estimate  Std. Error  z value   Pr(>|z|)
(Intercept)     1.00       ---      -3.183    0.00146 **
x1             -0.50       ---       7.511    5.87e-14 ***
x2              2.50       ---      -2.405    0.01616 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: The log odds are $1 - 0.5 \cdot 1 + 2.5 \cdot -0.75$ **(1P.)**. Therefore, the probability that y equals 1 is $e^{-1.375}/(1 + e^{-1.375}) = 0.2$ **(1P.)**. This equals the expectation. **(1P.)**

2. (30 Points). Please tick the corresponding boxes in the answer sheet provided.

1) The Squared Mahalanobis distance for a vector x of measurements can be interpreted as squared distance measured in standard deviations from a set of observations with the mean μ in the direction of x. **TRUE.**

2) The Mahalanobis distance is useful for detecting multivariate outliers and is also used in discriminant analysis and Gaussian mixture model clustering. **TRUE.**

3) First k principal components represent a linear subspace of dimension k with smallest orthogonal distance to all points. **TRUE.**

4) When applying PCA to unscaled data, the 1st principal component is likely to have a high contribution from the variable that has the highest variance in the dataset. **TRUE.**

5) The 2nd PC is the linear combination of the original variables that has the largest variance subject to being uncorrelated with the 1st PC. **TRUE.**

6) Before applying PCA, the data matrix $\mathbf{X}$ needs to be mean centered. **TRUE.**

7) The principal components calculated using the covariance matrix of a data set can differ from the principal components extracted from the data's correlation matrix. **TRUE.**

8) PCA relies on the assumption that the data is generated by the multivariate normal distribution. **FALSE.**

9) Principal components of a matrix X are unique up to rotations. **FALSE.**

10) Non-metric MDS can only identify points up to shift, rotation and reflection, while classical MDS results in a unique set of coordinate values. **FALSE.**

11) Performing classical MDS using Manhattan distances is equivalent to applying PCA to the covariance matrix of the original data. **FALSE.**

12) Assume that we are performing MDS on a matrix of distances $\mathbf{D}$, derived from an $n \times q$ data matrix, $\mathbf{X}$. Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues obtained in result of the spectral decomposition of $\mathbf{B} = \mathbf{X}\mathbf{X}^T$. If some of these eigenvalues are negative, we can conclude that non-Euclidean distances have been used. **TRUE.**

13) For non-metric MDS the exact numerical values of the observed proximities are not taken into account, only the rank order of the proximities is used to produce the spatial representation. **TRUE.**

14) FA can be applied to either the covariance matrix or the correlation matrix. **TRUE.**

15) An oblique rotation does not allow for correlated factors, which implies that the matrix of correlations between factors after rotation is the diagonal matrix. **FALSE.**

16) If the observed variables are almost uncorrelated, PCA will be useless while FA will be useful to explain the specific variances in the observed factors. **FALSE.**

17) A varimax rotation is applied to obtain many factor loadings which are large and non-zero in order to improve interpretability. **FALSE.**

18) For a fixed number of clusters k, the k-means algorithm always produces the same clustering. **FALSE.**

19) The difference between partitioning around medoids (PAM) and k-means is that for PAM cluster centers are points that are actually observed in the data. **TRUE.**

20) When applying model based clustering (for instance Bayesian Mixture Model Clustering) it is advised to pick the clustering which has the smallest value of the Bayesian information criterion (BIC). **TRUE.**

21) A dendrogram obtained from applying the single-linkage clustering method always has a depth of $\frac{n}{2}$. **FALSE.**

22) If an observation has a negative $S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$ in the silhoutte plot then the point has been assigned to the wrong cluster. **TRUE.**

23) After fitting a logistic regression model to binary data $y \in \{0, 1\}$, we observe that all the coefficients apart from the intercept β_0 are equal to 0. If β_0 is large and positive then most of the observations in the data are from class 0. **FALSE.**

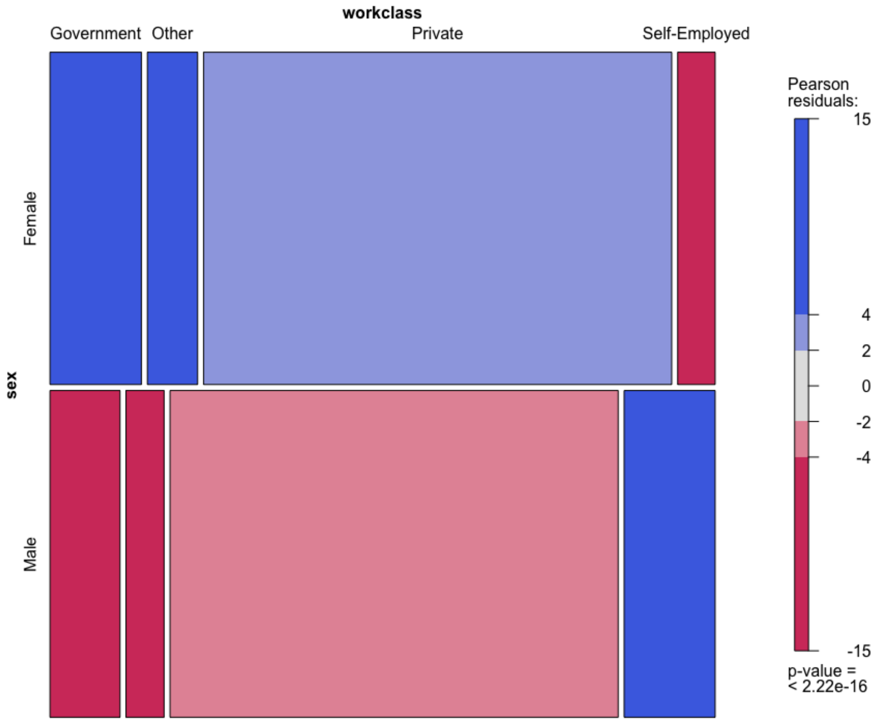
24) Logistic regression models are commonly used when the outcome variable y is continuous. **FALSE.**

25) If the data comes from a multidimensional Gaussian distribution and the data points are labelled into 2 classes, then principal component analysis (PCA) is generally more useful than linear discriminant analysis (LDA) to classify possible new observations. **FALSE.**

26) The benefit of using a random forest regressor compared to a single regression tree is that it reduces the variance of the predictions. **TRUE.**

27) A random forest consisting of 10 trees is grown. The split chosen in the root node for every tree must be the same. **FALSE.**

28) The Adult dataset contains information on categorical variables `workclass` (work class : Government, Private sector, Self-employed or Other) and `sex` (gender : male or female). The following plot was obtained by using the R command mosaic(∼ workclass + sex, data = adult, shade = TRUE). Based on the plot, determine whether the following statements are true:



a) The majority of the people in the dataset work in the private sector. **TRUE.**

b) The hypothesis that variables `workclass` and `sex` are independent cannot be rejected since the p-value is smaller than $2.22 * 10^{-16}$. **FALSE.**

c) The number of self-employed females is surprisingly small compared to the number of self-employed males. **TRUE.**