YOU CAN PREVIEW THIS QUIZ, BUT IF THIS WERE A REAL ATTEMPT, YOU WOULD BE BLOCKED BECAUSE:

This quiz is not currently available

QUESTION 1

Not yet answered Marked out of 20.00

A centered dataset with $n = 85$ observations and $p = 6$ variables was analysed to reduce its dimensionality. As part of Principal Component Analysis, the following variance-covariance matrix $\mathbf{\Sigma}$ was generated

$$\begin{pmatrix} 399.019 & 49.66 & -1.793 & 1.7 & 8.333 & 16.583 \\ 49.66 & 38.529 & -6.952 & 1.733 & 2.409 & 7.873 \\ -1.793 & -6.952 & 36.051 & 12.142 & -4.986 & 2.841 \\ 1.7 & 1.733 & 12.142 & 37.132 & -4.093 & -0.04 \\ 8.333 & 2.409 & -4.986 & -4.093 & 41.277 & -3.757 \\ 16.583 & 7.873 & 2.841 & -0.04 & -3.757 & 45.062 \end{pmatrix}$$

A) Compute and write the numerical value of the eigenvalue λ_4 of $\mathbf{\Sigma}$. This eigenvalue is located in the position $(4, 4)$ of the matrix $\mathbf{\Lambda}$ and is simultaneously the sample variance of the score PC4:

B) Compute and write the percentage of total variability explained by the Principal component PC4. The number you write should be between 0 and 100 and you should include decimals in your answer.

C) As seen in lectures, the eigenvalue λ_4 is related to d_4, one singular eigenvalue of the data matrix $\mathbf{X}$. Compute and write the value of d_4.

D) A threshold of total variability explained has been set at 80\%. How many principal components must you select? Write your answer.

QUESTION 2

Not yet answered Marked out of 20.00

Consider the following data set with $n = 9$ observations and $p = 4$ variables. The data set is given next

	$V1$	$V2$	$V3$	$V4$
A	1.9	1.3	3.1	4.9
B	5.2	2	3.2	4.9
C	1.3	5.2	4.2	4.1
D	1.2	1	5.1	4
E	1.7	3.3	1.2	1.1
F	2.3	3.2	3.3	1.1
G	4.3	3.2	2.8	5.2
H	1.3	1.9	2.2	1
I	4.2	2	1.9	1.9

as well as the distance matrix using the "Euclidean" metric. The symbol x in the matrix below is to be calculated later.

	A	B	C	D	E	F	G	H	I
A	0	3.375	4.174	2.322	4.7	4.272	3.09	4.091	4.027
B	3.375	0	5.205	4.628	5.69	4.93	1.581	5.606	3.419
C	4.174	5.205	0	4.298	x	3.848	4.021	4.95	5.365
D	2.322	4.628	4.298	0	5.4	4.207	4.602	4.27	4.965
E	4.7	5.69	x	5.4	0	2.186	5.113	1.769	3.012
F	4.272	4.93	3.848	4.207	2.186	0	4.589	1.977	2.766
G	3.09	1.581	4.021	4.602	5.113	4.589	0	5.356	3.626
H	4.091	5.606	4.95	4.27	1.769	1.977	5.356	0	3.053
I	4.027	3.419	5.365	4.965	3.012	2.766	3.626	3.053	0

A) In the distance matrix there is a missing distance x. Compute its value and write it.

B) Consider two arbitrary clusters GH and ABCDEFI. Compute and write the dissimilarity between these clusters under "average" linkage.

C) Using the above data **X**, the `R` command `KM<-kmeans(x=X,centers=3)` was run, with the following output

```
> KM$cluster
[1] 2, 1, 2, 2, 3, 3, 1, 3, 3
```

There is interest in determining the center of the cluster identified with the label `1`. By computing this center manually or otherwise, identify which of the following is the correct centroid of this cluster:

D) Still using the above data **X**, the `R` command `pam(x=X,k=3)->PM` was run, with the following output:

```
> PM$id.med
[1] 1, 7, 6
```

Identify correctly the medoids yielded by this cluster analysis.

The following data are the results of a classification analysis. The output includes the validation output `Ytrue` and predicted classifications obtained with three trained classification algorithms termed `Y1`, `Y2` and `Y3`.

	Ytrue	Y1	Y2	Y3
1	1	1	0	0
2	1	1	0	1
3	1	1	0	0
4	0	0	1	0
5	1	1	0	1
6	1	1	0	1
7	0	0	1	0
8	1	1	0	0
9	0	0	1	1
10	1	1	0	0
11	0	0	1	0
12	0	0	1	0
13	0	0	1	1
14	0	0	1	0

Analyze the performance of the classifier `Y1`. To this end and using the given data, compute the usual figures TN, FP, FN and TP for the confusion matrix as well as the performance measures TPR and FPR. Report the figures you have obtained and briefly comment on the performance of this classifier.

QUESTION 4

Not yet answered Marked out of 20.00

The following table contains output from a lasso fit to a linear model with $d = 5$ variables and $n = 50$ observations. Starting from the left, the columns are λ, and $\beta_1, \ldots, \beta_5$, i.e. each row has λ and the transposed column vector $\beta(\lambda)$.

0.00000	1.20706	−0.66487	0.46392	0.19746	−0.38526
6.18304	1.05715	−0.50740	0.27952	0.00000	−0.17280
12.48795	0.91734	−0.36985	0.11063	0.00000	0.00000
16.89171	0.82829	−0.28890	0.00000	0.00000	0.00000
33.19002	0.53551	0.00000	0.00000	0.00000	0.00000
59.28000	0.00000	0.00000	0.00000	0.00000	0.00000

For each of the required computations below, briefly report your procedure and the required quantity.

a) For each row in the table, compute s, the proportion of shrinkage defined as $s = s(\lambda) = ||\beta(\lambda)||_1 / \max_\lambda ||\beta(\lambda)||_1$.

b) Consider $\lambda' = 25.040865$. Note that λ' is the intermediate value between $\lambda = 16.89171$ and $\lambda = 33.19002$ of the 4th and 5th rows above. Using this value of λ', compute and report the shrunk estimator $\beta(\lambda')$.

c) Give the proportion of shrinkage $s(\lambda')$ for the shrunk estimator $\beta(\lambda')$.

Given a data set `X`, the following `R` commands have been run:

```
library(cluster);
agnes(x=X)->AG;
K<-3; kmeans(x=X,centers=K)->KM
```

Match the following objects with what you expect the `R` output to be.

`AG$height`	Choose...
`AG$order`	Choose...
`KM$cluster`	Choose...
`KM$betweenss`	Choose...
`KM$tot.withinss`	Choose...
`KM$totss`	Choose...
`KM$withinss`	Choose...

QUESTION 6

Not yet answered Marked out of 5.00

Examine carefully the following lines of `R` code.

```
X<-scale(x=X,center=TRUE,scale=FALSE)
S<-svd(x=X)
S$d^2/(nrow(X)-1)
S$v
pairs(S$u\%*\%diag(S$d))
```

Briefly explain what the code is about, and what each line of code is doing. If there is output, say what would the output be.

QUESTION 7

Not yet answered Marked out of 5.00

In clustering

Select one:

- a. it is not possible to use cross-validation to select a good number of clusters.
- b. in some cases we can validate with data to determine number of clusters.
- c. the objective is to reduce dimensionality of the data.
- d. it is possible to use cross-validation to select a good number of clusters.

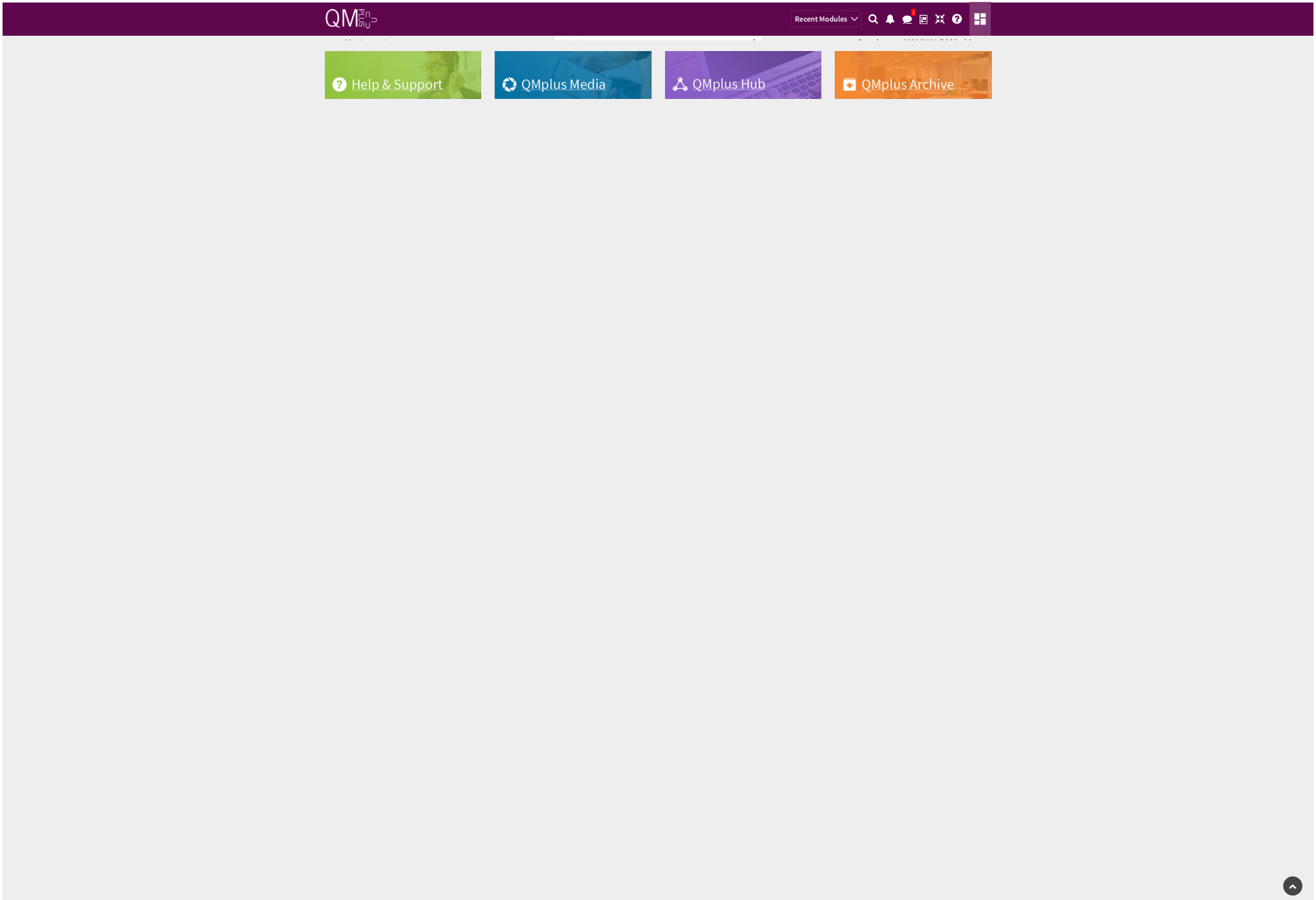
QUESTION 8

Not yet answered Marked out of 5.00

Consider that you have performed Principal Component Analysis of a centered and unscaled data set, that is, the variance-covariance matrix to be analysed is not equal to the correlation matrix. To do the PCA for the same set, but now centered and scaled,

Select one:

- a. reuse the eigenvalues, with the only change is to rescale them to add to the number of components. The eigenvectors are the same.
- b. it is not possible to reuse eigenvalues nor eigenvectors.
- c. in some selected instances we can reuse eigenvalues and eigenvectors of the analysis of centered and unscaled data.


Recent Modules
Help & Support
QMplus Media
QMplus Hub
QMplus Archive