ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐỀ THI CUỐI KỲ II
NĂM HỌC 2020 - 2021

***Môn thi:* PHÂN TÍCH THỐNG KÊ NHIỀU CHIỀU**

Mã môn học: **MAT3452** Số tín chỉ: **3** Đề số: **1**
Dành cho sinh viên hệ: Chính quy *Ngành:* Máy tính và Khoa học thông tin
Thời gian làm bài: **60 phút** *(không kể thời gian phát đề)*

Trong dự án MS305 của mình, Michael Larner đã đo cân nặng và các chỉ số thể chất khác của 22 đối tượng nam trong độ tuổi $16-30$. Đối tượng là những tình nguyện viên được chọn ngẫu nhiên, tất cả đều có sức khỏe tốt. Các đối tượng được yêu cầu căng từng cơ được đo để đảm bảo tính nhất quán của phép đo. Ngoài Cân nặng (Mass), tất cả các phép đo đều được tính bằng cm.

Mô tả biến:

- Mass: Cân nặng tính bằng kg
- Fore: Chu vi tối đa của cẳng tay
- Bicep: Chu vi tối đa của bắp tay
- Chest: Khoảng cách xung quanh ngực ngay dưới nách
- Neck: Khoảng cách quanh cổ
- Shoulder: Khoảng cách xung quanh vai, được đo xung quanh đỉnh của bả vai
- Waist: Khoảng cách quanh eo
- Height: Chiều cao từ đầu đến chân
- Calf: Chu vi tối đa của bắp chân
- Thigh: Chu vi của đùi
- Head: Chu vi vòng đầu

Gọi $X =$ (Mass, Fore, Bicep, Chest, Neck,Shoulder, Waist, Height, Calf, Thigh, Head)^T là vectơ ngẫu nhiên $11-$chiều.
Sử dụng phần mềm thống kê R/RStudio, hãy:

(i) Tìm trung bình mẫu, ma trận hiệp phương sai và ma trận tương quan.

(ii) Tìm giá trị riêng, vectơ riêng của ma trận hiệp phương sai.

(iii) Vẽ đồ thị xác suất chuẩn của các biến trong cùng một khung hình với 3 cột và 4 hàng. Từng biến có phân phối chuẩn $1-$chiều không? X có phân bố chuẩn $11-$chiều không?

(iv) Ước lượng các hệ số trong mô hình hồi quy tuyến tính của Mass theo Fore, Bicep, Chest, Neck, Shoulder. Viết phương trình hồi quy tuyến tính.

(v) Đưa ra dự đoán về giá trị của Mass khi Fore = 28, Bicep = 35, Chest = 105, Neck = 38.5, Shoulder = 116. Tìm khoảng tin cậy 95% cho ước lượng về Mass của tất cả nam giới có số đo về 5 chỉ số như trên.

(vi) Thực hiện phân tích thành phần chính dựa trên ma trận tương quan mẫu. Tỷ lệ biến sai tổng cộng của X do thành phần chính thứ 3 gây ra là bao nhiêu?

(vii) Biểu diễn các thành phần chính theo các biến ban đầu.

(viii) Cần bao nhiêu thành phần chính để thu được 90% thông tin về tập dữ liệu ban đầu?

(ix) Thực hiện phân tích nhân tố với số nhân tố bằng 5. Tìm ma trận tải trọng.

(x) Với mức ý nghĩa 5%, số nhân tố bằng 5 có phù hợp với số liệu không?

———————*Hết*———————

Ghi chú: *Sinh viên được dùng tài liệu, cán bộ coi thi không giải thích gì thêm.*

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐỀ THI CUỐI KỲ II
NĂM HỌC 2021 - 2022

***Môn thi:* PHÂN TÍCH THỐNG KÊ NHIỀU CHIỀU**

Mã môn học: **MAT3452**
Dành cho sinh viên hệ: Chính quy

Số tín chỉ: **3** Đề số: **1** (gồm 3 trang)
Ngành: Máy tính và Khoa học thông tin, Toán tin ứng dụng

Thời gian làm bài: **90 phút** *(không kể thời gian phát đề)*

Câu 1. (*8 điểm*) Cho bộ dữ liệu về 5 chỉ số sức khỏe của 30 vận động viên nữ tại Viện Thể thao Úc được thu thập bởi Richard Telford và Ross Cickyham. Ký hiệu:

- RCC - Số lượng hồng cầu (triệu tế bào/cm^3);
- Hc - Chỉ số các tế bào hồng cầu trong máu (%);
- Hg - Nồng độ huyết sắc tố trong các tế bào hồng cầu (gm/dL);
- BMI - Chỉ số thể trọng (kg/m^2);
- X.Bfat - Tỷ lệ mỡ cơ thể (%).

(Nguồn: http://www.statsci.org/data/oz/ais.html)

Gọi $X = (\text{RCC, Hc, Hg, BMI, X.Bfat})^T$ là vectơ ngẫu nhiên 5−chiều. Sử dụng phần mềm thống kê R/RStudio, ta thu được một số kết quả.

Ma trận hiệp phương sai mẫu của X

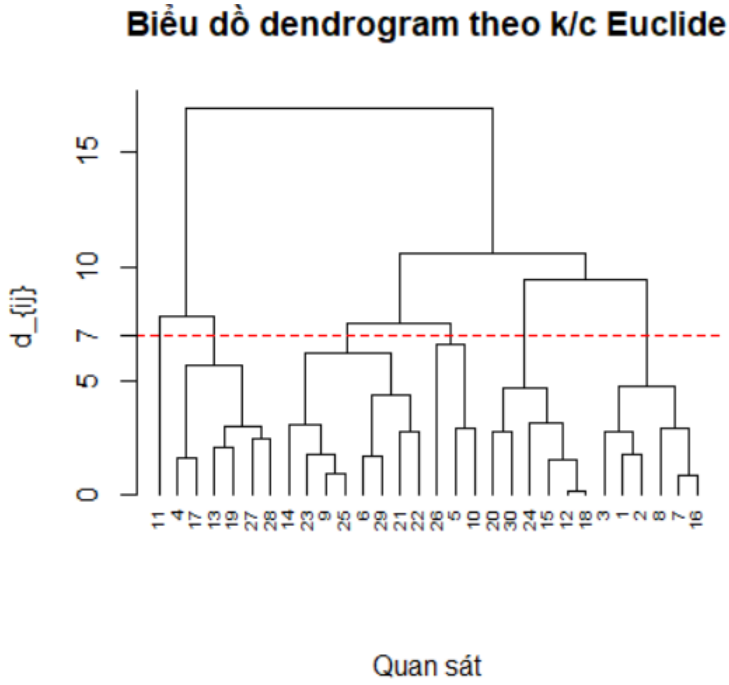
	RCC	Hc	Hg	BMI	X.Bfat
RCC	0.06646448	0.51157586	0.1725862	-0.04243276	0.05358690
Hc	0.51157586	5.41702299	1.9420690	-0.06943103	0.01066207
Hg	0.17258621	1.94206897	0.8234483	0.18734483	0.17610345
BMI	-0.04243276	-0.06943103	0.1873448	2.86781207	1.87293793
X.Bfat	0.05358690	0.01066207	0.1761034	1.87293793	12.23729379

Ma trận tương quan mẫu của X

	RCC	Hc	Hg	BMI	X.Bfat
RCC	1.00000000	0.852579509	0.73772263	-0.09719213	0.059418420
Hc	0.85257951	1.000000000	0.91953052	-0.01761562	0.001309539
Hg	0.73772263	0.919530516	1.00000000	0.12191249	0.055476238
BMI	-0.09719213	-0.017615621	0.12191249	1.00000000	0.316158853
X.Bfat	0.05941842	0.001309539	0.05547624	0.31615885	1.000000000

Giá trị riêng và vectơ riêng của ma trận tương quan mẫu của X

```
eigen() decomposition
$values
[1] 2.67820292 1.32703421 0.71082067 0.23232769 0.05161451

$vectors
            [,1]        [,2]        [,3]       [,4]        [,5]
[1,] -0.55719011  0.07641253  0.17358728  0.7645089  0.26285721
[2,] -0.59787256  0.05203142 -0.04471111 -0.1618958 -0.78206891
[3,] -0.57468666 -0.06507666 -0.16049863 -0.5691246  0.56199437
[4,] -0.01097582 -0.71587430 -0.65741887  0.2297438 -0.04921107
[5,] -0.04125952 -0.68901587  0.71407913 -0.1123872 -0.03185754
```



(i) (*1 điểm*) Biểu diễn thành phần chính thứ nhất và thành phần chính thứ ba theo các biến ban đầu.

(ii) (*1 điểm*) Tỷ lệ biến sai tổng cộng của X do thành phần chính thứ hai gây ra là bao nhiêu?

(iii) (*1 điểm*) Để thu được 90% thông tin về tập dữ liệu ban đầu thì cần m thành phần chính. Tìm m.

(iv) (*1 điểm*) Sai số của rút gọn từ 5 chiều về m chiều bằng bao nhiêu?

(v) (*1 điểm*) Tìm ma trận tải trọng $L = (l_{ij})$ khi phân tích nhân tố với số nhân tố bằng 2.

(vi) (*1 điểm*) Khi hệ số tải trọng $|l_{ij}| < 0.1$ thì ta cho rằng thành phần X_i không bị ảnh hưởng (chi phối) bởi nhân tố F_j. Từ ma trận tải trọng ở câu (v), hãy chỉ ra nhân tố F_1 và F_2 lần lượt chi phối các chỉ số sức khỏe nào? Từ đó, hãy đưa ra tên của F_1 và F_2.

(vii) (*1 điểm*) Gọi d_{ij} là khoảng cách giữa quan sát thứ i và quan sát thứ j. Hai quan sát thứ i và j được gọi là thuộc cùng một nhóm nếu $d_{ij} < 7$ và không thuộc cùng một nhóm nếu $d_{ij} \geq 7$. Dựa vào *Biểu đồ dendrogram theo k/c Euclide*, hãy cho biết bộ dữ liệu ban đầu được phân thành bao nhiêu nhóm? Mỗi nhóm gồm các quan sát nào?

(viii) (*1 điểm*) Sử dụng phương pháp k-trung bình với $k = 5$ thu được kết quả sau. Xác định tâm của mỗi nhóm. Quan sát thứ 20 thuộc nhóm nào?

Phương pháp k-trung bình

```
K-means clustering with 5 clusters of sizes 6, 6, 6, 6, 6

Cluster means:
       RCC       Hc       Hg      BMI   X.Bfat
1 4.496667 41.70000 14.10000 23.81500 24.09333
2 4.158333 37.81667 12.55000 21.31000 22.03000
3 4.633333 43.38333 14.63333 20.77167 18.86667
4 4.281667 39.10000 13.03333 21.88667 15.34333
5 4.325000 40.78333 13.68333 22.99167 18.78667

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
 2  2  2  2  3  4  5  2  5  4  1  3  1  5  3  5  2  3  1  1  4  4  5
24 25 26 27 28 29 30
 3  5  4  1  1  4  3

Within cluster sum of squares by cluster:
[1] 50.45522 34.64362 25.12122 52.53368 19.09223
 (between_SS / total_SS =  70.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Câu 2. (*2 điểm*) Cho X là vectơ ngẫu nhiên có phân phối chuẩn 2−chiều với vectơ giá trị trung bình $\mu = (3, 4)$ và ma trận hiệp phương sai $\sum = \begin{pmatrix} 4 & 0 \\ 0 & 9 \end{pmatrix}$. Hãy viết phương trình mặt mức $c^2 = 9$. Tìm toạ độ các điểm nằm trong mặt mức này. Mặt mức dùng để làm gì? Nêu tính chất và ý nghĩa của mặt mức.

———————*Hết*———————

Ghi chú: *Sinh viên được dùng tài liệu, cán bộ coi thi không giải thích gì thêm.*

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐỀ THI CUỐI KỲ II
NĂM HỌC 2022 - 2023

***Môn thi:* PHÂN TÍCH THỐNG KÊ NHIỀU CHIỀU**

Mã môn học: **MAT3452**
Dành cho sinh viên hệ: Chính quy

Số tín chỉ: **3** Đề số: **1** (gồm 08 trang)
Ngành: Máy tính và Khoa học thông tin, Toán tin ứng dụng

Thời gian làm bài: **90 phút** *(không kể thời gian phát đề)*

Câu 1. Cho bộ dữ liệu về 10 chỉ số sức khỏe của 100 vận động viên tại Viện Thể thao Úc được thu thập bởi Richard Telford và Ross Cunningham. Ký hiệu:

- WCC - Số lượng bạch cầu (triệu tế bào/cm^3);
- Hc - Chỉ số các tế bào hồng cầu trong máu (%);
- Hg - Nồng độ huyết sắc tố trong các tế bào hồng cầu (mg/dL);
- Ferr - Nồng độ ferritin huyết tương (mg/dL);
- BMI - Chỉ số thể trọng (kg/m^2);
- SSF - Tổng số nếp gấp da;
- XBfat - Tỷ lệ mỡ cơ thể (%);
- LBM - Khối lượng nạc (kg);
- Ht - Chiều cao (cm);
- Wt - Cân nặng (kg).

(Nguồn: http://www.statsci.org/data/oz/ais.html)
Gọi X là vectơ ngẫu nhiên $10-$chiều gồm các biến ở trên. Sử dụng phần mềm RStudio, thu được một số kết quả sau.

ANOVA của mô hình hồi quy tuyến tính

```
 Step Df Deviance Resid. Df Resid. Dev     AIC
1      NA       NA         90   242.4074 108.545
```

Tóm tắt mô hình hồi quy tuyến tính

```
Call:
lm(formula = WCC ~ Hc + Hg + Ferr + BMI + SSF + XBfat + LBM +
Ht + Wt, data = X)

Residuals:
Min      1Q  Median      3Q     Max
-3.1120 -1.0456 -0.2452  0.7048  6.0411

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -91.381518  35.321507  -2.587  0.01128 *
Hc            0.040322   0.161444   0.250  0.80334
Hg            0.279759   0.459705   0.609  0.54435
Ferr         -0.003011   0.005618  -0.536  0.59333
BMI           2.456997   0.903867   2.718  0.00787 **
SSF          -0.005354   0.023749  -0.225  0.82216
XBfat        -0.408350   0.252880  -1.615  0.10985
LBM          -0.745446   0.409388  -1.821  0.07195 .
Ht            0.582962   0.224750   2.594  0.01108 *
Wt           -0.208075   0.259992  -0.800  0.42564
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.641 on 90 degrees of freedom
Multiple R-squared:  0.1482, Adjusted R-squared:  0.063
F-statistic:  1.74 on 9 and 90 DF,  p-value: 0.09141
```

Ma trận tương quan mẫu của X

	WCC	Hc	Hg	Ferr	BMI
WCC	1.000000000	0.198980182	0.20142185	-0.02067485	0.145820313
Hc	0.198980182	1.000000000	0.90343235	-0.12714188	0.008696347
Hg	0.201421854	0.903432347	1.00000000	-0.03582085	0.131101890
Ferr	-0.020674850	-0.127141877	-0.03582085	1.00000000	0.135065945
BMI	0.145820313	0.008696347	0.13110189	0.13506594	1.000000000
SSF	0.119704763	-0.224654389	-0.15803790	0.15664559	0.678488013
XBfat	0.118672968	-0.194048115	-0.13299330	0.13231937	0.660492175
LBM	0.048301041	0.119843999	0.16347139	-0.04997854	0.747491455
Ht	-0.006855086	0.020647228	-0.03799253	-0.14188959	0.231664850
Wt	0.088732592	0.010431487	0.07258187	0.02128948	0.847033451

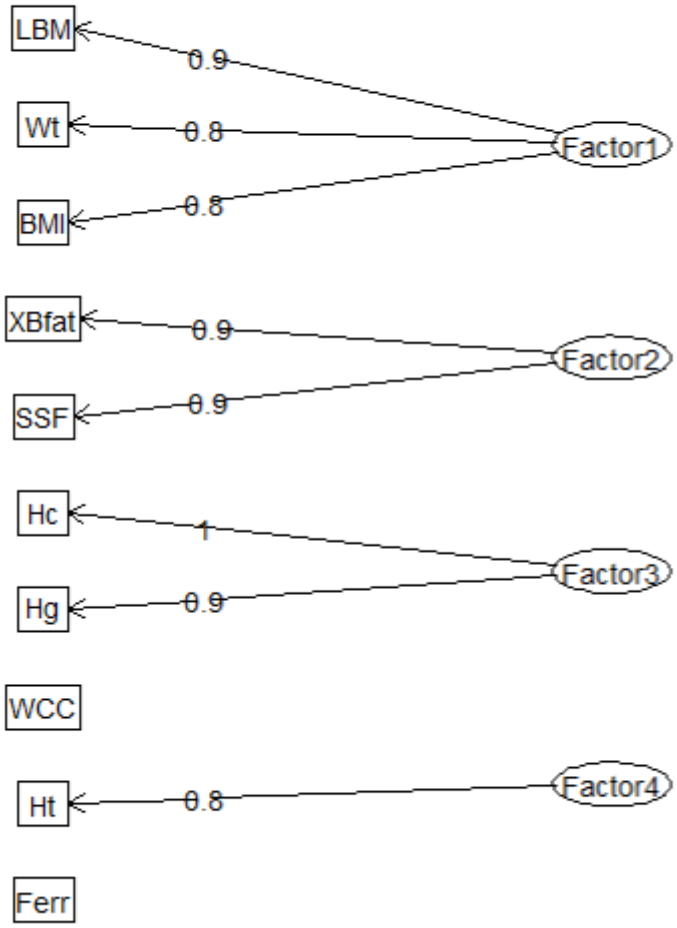
	SSF	XBfat	LBM	Ht	Wt
WCC	0.1197048	0.1186730	0.04830104	-0.006855086	0.08873259
Hc	-0.2246544	-0.1940481	0.11984400	0.020647228	0.01043149
Hg	-0.1580379	-0.1329933	0.16347139	-0.037992530	0.07258187

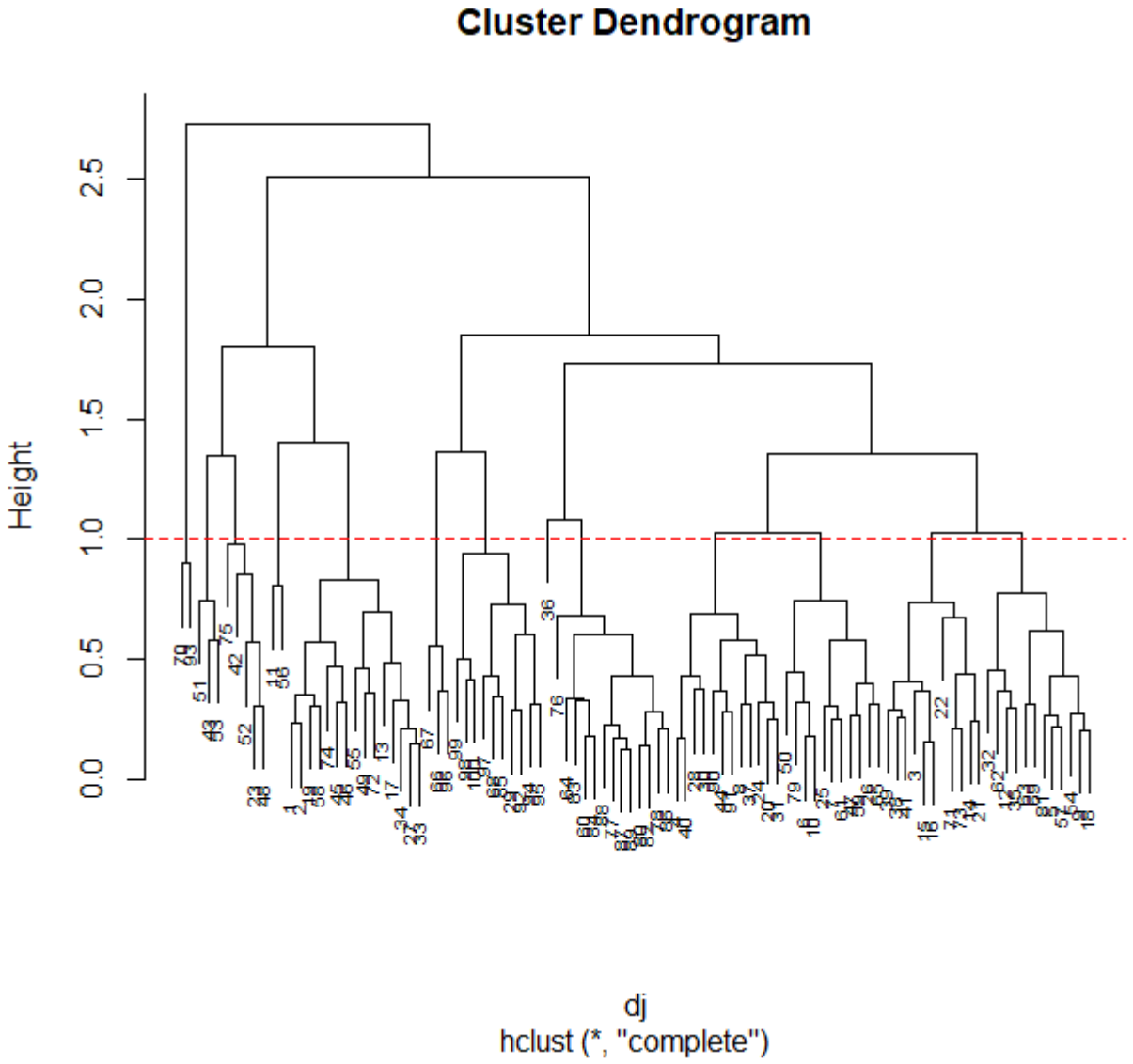
```
Ferr   0.1566456  0.1323194 -0.04997854 -0.141889586 0.02128948
BMI    0.6784880  0.6604922  0.74749146  0.231664850 0.84703345
SSF    1.0000000  0.9695352  0.40649120  0.406515525 0.71966485
XBfat  0.9695352  1.0000000  0.40618230  0.443053911 0.72487638
LBM    0.4064912  0.4061823  1.00000000  0.708293376 0.92079759
Ht     0.4065155  0.4430539  0.70829338  1.000000000 0.70873995
Wt     0.7196649  0.7248764  0.92079759  0.708739954 1.00000000
```

Giá trị riêng và vectơ riêng của ma trận tương quan mẫu của X

```
eigen() decomposition
$values
[1] 4.257642569 2.137716078 1.257911499 0.951846926 0.676702128 0.602968643
[7] 0.081820660 0.029280350 0.002520393 0.001590754

$vectors
            [,1]        [,2]       [,3]        [,4]        [,5]
[1,]  0.064548597 -0.21399674 -0.3905983  0.74633643 -0.48660325
[2,] -0.030144867 -0.64651240 -0.0675537 -0.08672252  0.15529558
[3,]  0.002050524 -0.63806844 -0.1697737 -0.15723212  0.18128158
[4,]  0.036919566  0.14582103 -0.5878887 -0.57621525 -0.49846465
[5,]  0.410749448 -0.06834530 -0.2093752 -0.10613147  0.13747779
[6,]  0.408740030  0.18407339 -0.2413253  0.12522086  0.32437917
[7,]  0.409913581  0.16506954 -0.2167684  0.13515563  0.33389325
[8,]  0.400164256 -0.18519943  0.2790499 -0.16721181 -0.29970994
[9,]  0.324112039 -0.05107036  0.4791309  0.02471835 -0.35546283
[10,]  0.473882683 -0.07157472  0.1092409 -0.07023673 -0.07735621
           [,6]         [,7]         [,8]         [,9]        [,10]
[1,]  0.05768856  0.013184752 -0.002618281  0.007981911  0.009932078
[2,] -0.23531187 -0.699229435  0.004460954  0.018927986  0.015502792
[3,] -0.10969621  0.702236235  0.024836973 -0.023504524 -0.010866518
[4,] -0.22175882 -0.044197869 -0.011300211  0.003111847  0.001101244
[5,]  0.56290743 -0.102604061 -0.102576739 -0.390425200 -0.512118752
[6,] -0.26053819 -0.015471192  0.736826817 -0.056575627  0.075329314
[7,] -0.31949011  0.015766074 -0.658238011 -0.077803812  0.293818697
[8,]  0.31119502 -0.013766575  0.100140379 -0.216198452  0.675361623
[9,] -0.53322302  0.067388156 -0.026808886 -0.294970157 -0.399021211
[10,]  0.11000971  0.008305192 -0.042057546  0.838797138 -0.173034464
```

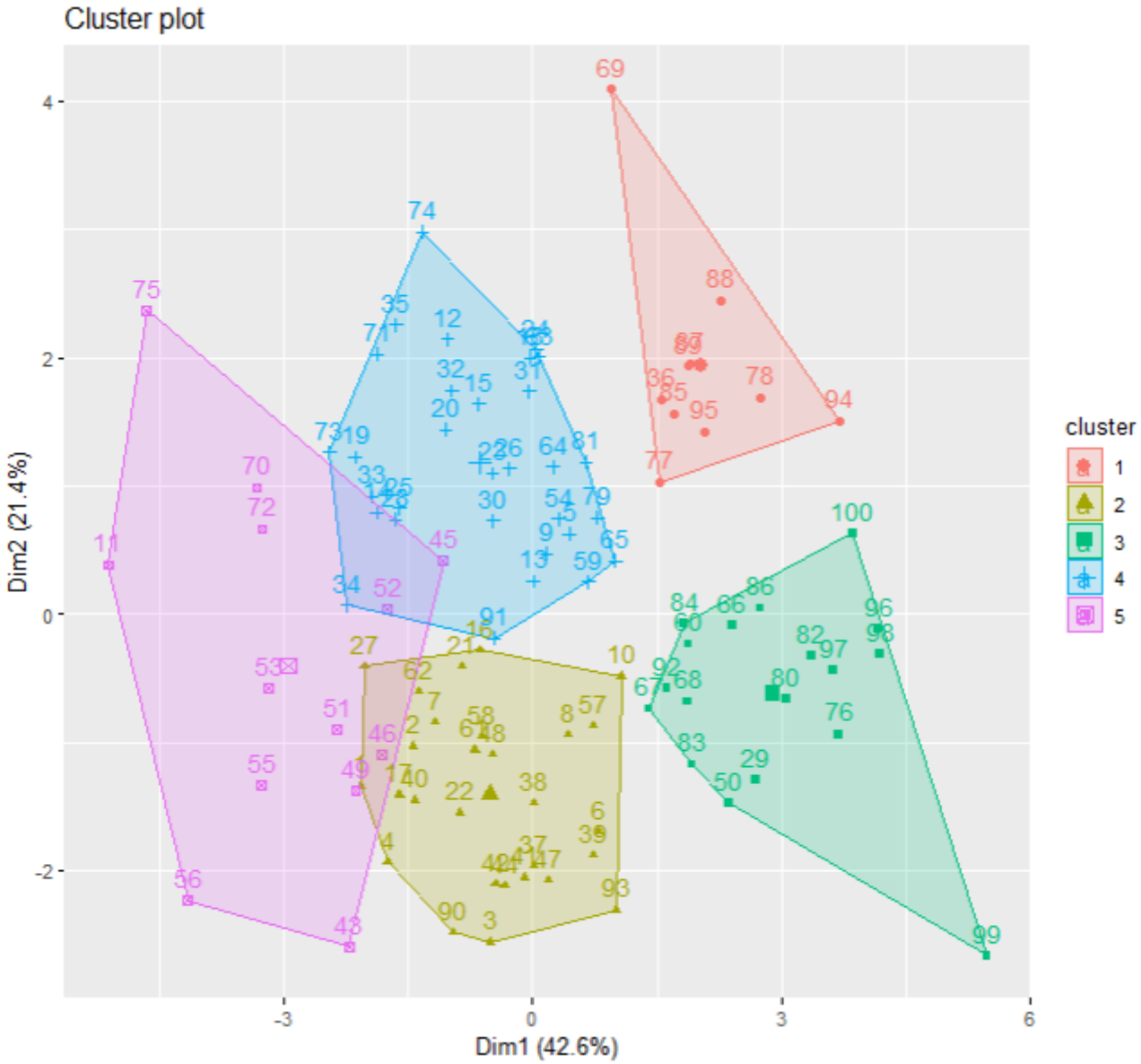
Factor Analysis




Cluster Dendrogram
Height
0.0
0.5
1.0
1.5
2.0
2.5
dj
hclust (*, "complete")

Phương pháp K-means

```
K-means clustering with 5 clusters of sizes 18, 9, 20, 24, 29

Cluster means:
WCC       Hc       Hg      Ferr      BMI       SSF     XBfat
1 0.6627778 3.514881 3.103359 0.3529412 1.272544 0.3220559 0.4390407
2 0.9188889 3.466270 3.025840 0.4549020 1.621505 0.8944777 0.9824732
3 0.7235000 3.791964 3.361628 0.3332353 1.629677 0.6289820 0.7703461
4 0.7425000 3.850074 3.323643 0.2683824 1.368687 0.3810130 0.5153461
5 0.6017241 3.404865 2.939856 0.3432049 1.445346 0.5692546 0.7070285
LBM       Ht        Wt
1 1.179038 3.481442 0.8859449
2 1.436907 3.740662 1.3012346
3 1.611380 3.807979 1.3513675
4 1.401131 3.708333 1.0787749
5 1.452812 3.792590 1.1910109

Clustering vector:
1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
5   5   5   5   4   5   5   5   4   5   3   3   4   3   3   5   5   4   3
20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38
3   5   5   3   4   3   4   5   3   1   3   4   3   3   3   3   4   5   5
39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57
5   5   5   5   2   5   2   2   5   5   2   1   2   2   2   4   2   2   5
58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76
5   4   1   5   5   4   4   4   1   1   1   4   3   3   3   3   3   3   1
77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95
4   4   4   1   4   1   1   1   4   1   4   4   4   5   5   1   5   4   4
96  97  98  99 100
1   1   1   1   1

Within cluster sum of squares by cluster:
[1] 2.990930 1.344283 3.843469 3.875258 4.403992
(between_SS / total_SS =  55.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
> rge
WCC     Hc     Hg   Ferr    BMI    SSF  XBfat    LBM     Ht     Wt
10.00  11.20   4.30 170.00  15.18 167.00  27.45  38.62  47.00  58.50
```



(i) Khi thực hiện giải thuật từng bước *forward* để tìm mô hình hồi quy tuyến tính biểu diễn WCC theo các biến còn lại, ta thu được kết quả tóm tắt về ***ANOVA của mô hình hồi quy tuyến tính***. Kết quả ấy cho ta biết điều gì?

(ii) Dựa vào ***Tóm tắt mô hình hồi quy tuyến tính***, biểu diễn WCC theo mô hình này. Nhận xét về các biến tham gia mô hình. Mô hình có cần cải tiến không?

(iii) Biểu diễn thành phần chính thứ ba và thành phần chính thứ tư theo các biến ban đầu.

(iv) Tỷ lệ biến sai tổng cộng của X do thành phần chính thứ ba gây ra là bao nhiêu?

(v) Sai số của rút gọn từ 10 chiều về 6 chiều bằng bao nhiêu?

(vi) Dựa vào đồ thị trực quan hóa mô hình ***Factor Analysis***, tìm ma trận tải trọng $L = (l_{ij})$ khi phân tích nhân tố với số nhân tố bằng 4.

(vii) Trong biểu đồ ***Factor Analysis***, một số biến không được nối với các nhân tố, lý do là gì? Đặt tên cho các nhân tố.

(viii) Chuẩn hóa bộ dữ liệu. Gọi d_{ij} là khoảng cách Euclide giữa quan sát thứ i và quan sát thứ j. Hai quan sát thứ i và j được gọi là thuộc cùng một nhóm nếu $d_{ij} < 1.0$ và không thuộc cùng một nhóm nếu $d_{ij} \geq 1.0$. Dựa vào biểu đồ ***Cluster dendrogram***, hãy cho biết bộ dữ liệu ban đầu được phân thành bao nhiêu nhóm? Nhóm thứ 9 gồm bao nhiêu quan sát?

(ix) Chuẩn hóa bộ dữ liệu với *rge* là vectơ gồm độ rộng khoảng giá trị của các biến. Sử dụng phương pháp k-trung bình với $k = 5$ thu được kết quả sau. Mỗi nhóm gồm bao nhiêu quan sát? Xác định tâm của mỗi nhóm.

(x) Quan sát thứ 20 thuộc nhóm nào? Dựa vào biểu đồ ***Cluster plot*** về các giá trị thành phần chính thứ nhất và thứ hai của các quan sát, nêu nhận xét.

Câu 2. Cho X là vectơ ngẫu nhiên có phân phối chuẩn $3-$chiều với vectơ giá trị trung bình $\mu = (1, 2, 3)$ và ma trận hiệp phương sai là ma trận đơn vị. Tìm tọa độ các điểm trong mặt mức $c^2 = 9$. Nêu đặc điểm và ý nghĩa của thể tích bao bởi mặt mức này. Giải thích.

———————*Hết*———————

Ghi chú: *Sinh viên không được dùng tài liệu, cán bộ coi thi không giải thích gì thêm.*

CHỮA ĐỀ THI NĂM 2022-2023

Câu 1.

(i)
- Mô hình khởi đầu với WCC là biến phụ thuộc, 9 biến còn lại là biến giải thích (Do $df = 90 = 100 - 10$ nên có 9 biến giải thích).
- Tổng bình phương phần dư là 242.4074.
- AIC mô hình là 108.545.
- Số bước đã chạy là 1.

(ii)
- Ở mức ý nghĩa 5% chỉ biến MBI và Ht là có ảnh hưởng đến mô hình.
- Mô hình cần cải tiến do nhiều biến không có ý nghĩa.

(iii) Hệ số cho TPC thứ i là $\lambda_i e_i$ (trong đó λ_i là giá trị riêng thứ i, e_i là véc-tơ riêng thứ i). Ví dụ:

$$\lambda_3 e_3 = (-0.491; -0.085; -0.214; -0.740; -0.263; -0.304; -0.273; 0.351; 0.603; 0.137).$$

Khi đó,

$$PC3 = -0.491 * WCC - 0.085 * Hc + \cdots + 0.137 * Wt.$$

(iv) Tỉ lệ

$$\frac{\lambda_3}{\sum \lambda_i} = 0.1258 \sim 12.58\%.$$

(v) Tỉ lệ

$$\frac{\lambda_1 + \cdots + \lambda_6}{\sum \lambda_i} = 0.9885 \sim 98.85\%.$$

Suy ra, sai số là 1.15%.

(vi)

$$L = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0.9 & 0 \\ 0.8 & 0 & 0 & 0 \\ 0 & 0.9 & 0 & 0 \\ 0 & 0.9 & 0 & 0 \\ 0.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 \\ 0.8 & 0 & 0 & 0 \end{bmatrix}$$

(vii) Đặt tên

- F_1 : Lượng cơ.
- F_2 : Khí động học cơ thể.

- F_3 : Khả năng hấp thu oxi.
- F_4 : Chiều cao.

Lưu ý: đặt tên khác cũng được.

(viii) Dữ liệu chia thành 13 nhóm và nhóm 9 có 13 quan sát.

(ix)
- Nhóm 1: 18 quan sát với các tâm như sau:

WCC	Hc	Hg	Ferr	BMI	SSF	XBfat	LBM	Ht	Wt
0.663	3.515	3.103	0.353	1.273	0.322	0.439	1.179	3.481	0.886

  Xác định tâm cho các nhóm 2-5 cũng chính là các dòng còn lại ở phương pháp k-means.
- Nhóm 2: 9 quan sát.
- Nhóm 3: 20 quan sát.
- Nhóm 4: 24 quan sát.
- Nhóm 5: 29 quan sát.

(x) Nhóm 4 (gần tâm hình xd).

Câu 2.

$$X \sim N\left(\begin{pmatrix}1\\2\\3\end{pmatrix}, I\right).$$

Mặt mức $c^2 = 9$ trong toạ độ Oxyz:

$$(x-1)^2 + (y-2)^2 + (z-3)^2 = c^2 = 9.$$

Phần thể tích nằm trong mặt mức này là:

$$P((x_1-1)^2 + (x_2-2)^2 + (x_3-3)^2 \leq 9).$$

Đặc điểm là một hình cầu do phương sai các chiều bằng nhau và không có tương quan.

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐỀ THI CUỐI KỲ II
NĂM HỌC 2023 - 2024

***Môn thi:* PHÂN TÍCH THỐNG KÊ NHIỀU CHIỀU**

Mã môn học: **MAT3452**
Dành cho sinh viên hệ: Chính quy

Số tín chỉ: **3** Đề số: **1** (gồm 5 trang)
Ngành: Máy tính và Khoa học thông tin, Toán tin ứng dụng

Thời gian làm bài: **90 phút** *(không kể thời gian phát đề)*

Câu 1. Xét một bộ dữ liệu con của **College** trong thư viện **ISLR** đưa ra 5 trường thông tin về 300 trường đại học tại Mỹ từ ấn bản năm 1995 của US News and World Report, đặt tên là **data**. Ký hiệu:

- Enroll - Số lượng sinh viên mới;
- F.Undergrad - Số lượng sinh viên chính quy;
- P.Undergrad - Số lượng sinh viên bán thời gian;
- S.F.Ratio - Tỷ lệ sinh viên/giảng viên;
- Grad.Rate - Tỷ lệ tốt nghiệp đúng hạn.

Phân tích dữ liệu bằng phần mềm thống kê RStudio.

(1) Xây dựng mô hình hồi quy tuyến tính biểu diễn tỷ lệ tốt nghiệp theo các biến còn lại. Viết phương trình hồi quy tuyến tính tương ứng. Những biến nào có ý nghĩa thống kê?

(2) Thành phần chính thứ nhất và thành phần chính thứ hai bị chi phối bởi những biến nào?

(3) Tỷ lệ biến sai tổng cộng của **data** do thành phần chính thứ hai gây ra là bao nhiêu?

(4) Để thu được 90% thông tin về tập dữ liệu ban đầu thì cần m thành phần chính. Tìm m.

(5) Sai số của rút gọn từ 5 chiều về m chiều bằng bao nhiêu?

(6) Tìm ma trận tải trọng $L = (l_{ij})$ khi phân tích nhân tố với số nhân tố bằng 2.

(7) Khi hệ số tải trọng $|l_{ij}| < 0.1$ thì ta cho rằng thành phần X_i không bị ảnh hưởng (chi phối) bởi nhân tố F_j. Từ ma trận tải trọng ở câu (6), hãy chỉ ra nhân tố F_1 và F_2 lần lượt chi phối các chỉ số sức khỏe nào? Từ đó, hãy đưa ra tên của F_1 và F_2.

(8) Xét các quan sát 1 đến 30 của bộ dữ liệu **data** và chuẩn hóa bộ dữ liệu con đó . Gọi d_{ij} là khoảng cách Euclide giữa quan sát thứ i và quan sát thứ j. Hai quan sát thứ i và j được gọi là thuộc cùng một nhóm nếu $d_{ij} < 2.1$ và không thuộc cùng một nhóm nếu $d_{ij} \geq 2.1$. Dựa vào biểu đồ dendrogram, hãy cho biết bộ dữ liệu được phân thành bao nhiêu nhóm? Nhóm thứ 6 gồm những quan sát?

Mô hình hồi quy tuyến tính biểu diễn tỷ lệ tốt nghiệp theo các biến còn lại

```
Call:
lm(formula = Grad.Rate ~ ., data = data)

Residuals:
Min      1Q  Median      3Q     Max
-54.813 -10.126   0.882  10.715  50.521

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 86.0642712  3.5092904  24.525  < 2e-16 ***
Enroll       0.0077697  0.0040482   1.919  0.05591 .
F.Undergrad -0.0006185  0.0008137  -0.760  0.44780
P.Undergrad -0.0028848  0.0010369  -2.782  0.00575 **
S.F.Ratio   -1.5399196  0.2570344  -5.991 6.06e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.23 on 295 degrees of freedom
Multiple R-squared:  0.1682, Adjusted R-squared:  0.1569
F-statistic: 14.91 on 4 and 295 DF,  p-value: 4.112e-11
```

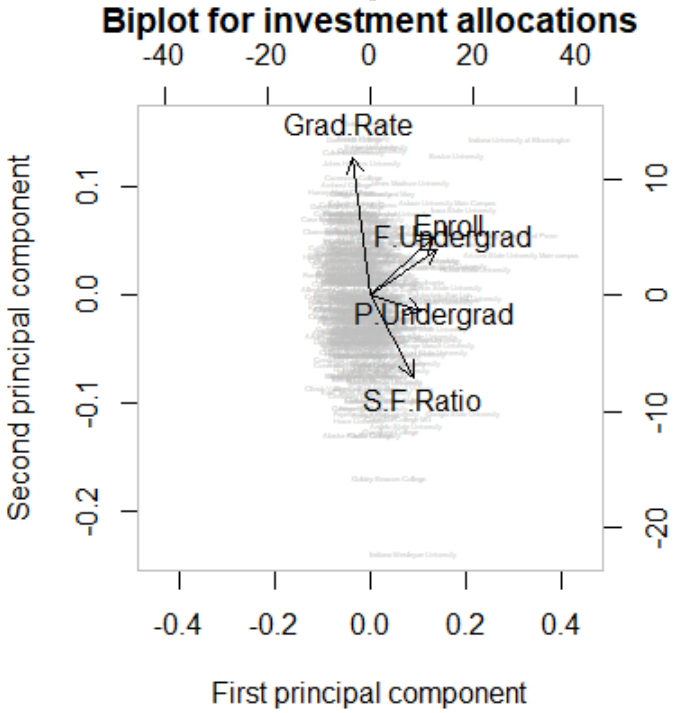
Ma trận tương quan mẫu của ***data***

	Enroll	F.Undergrad	P.Undergrad	S.F.Ratio	Grad.Rate
Enroll	1.000000000	0.9481969	0.4485566	0.3112805	0.009235172
F.Undergrad	0.948196934	1.0000000	0.5237141	0.3682697	-0.048037903
P.Undergrad	0.448556635	0.5237141	1.0000000	0.3030000	-0.207971041
S.F.Ratio	0.311280487	0.3682697	0.3030000	1.0000000	-0.350042330
Grad.Rate	0.009235172	-0.0480379	-0.2079710	-0.3500423	1.000000000

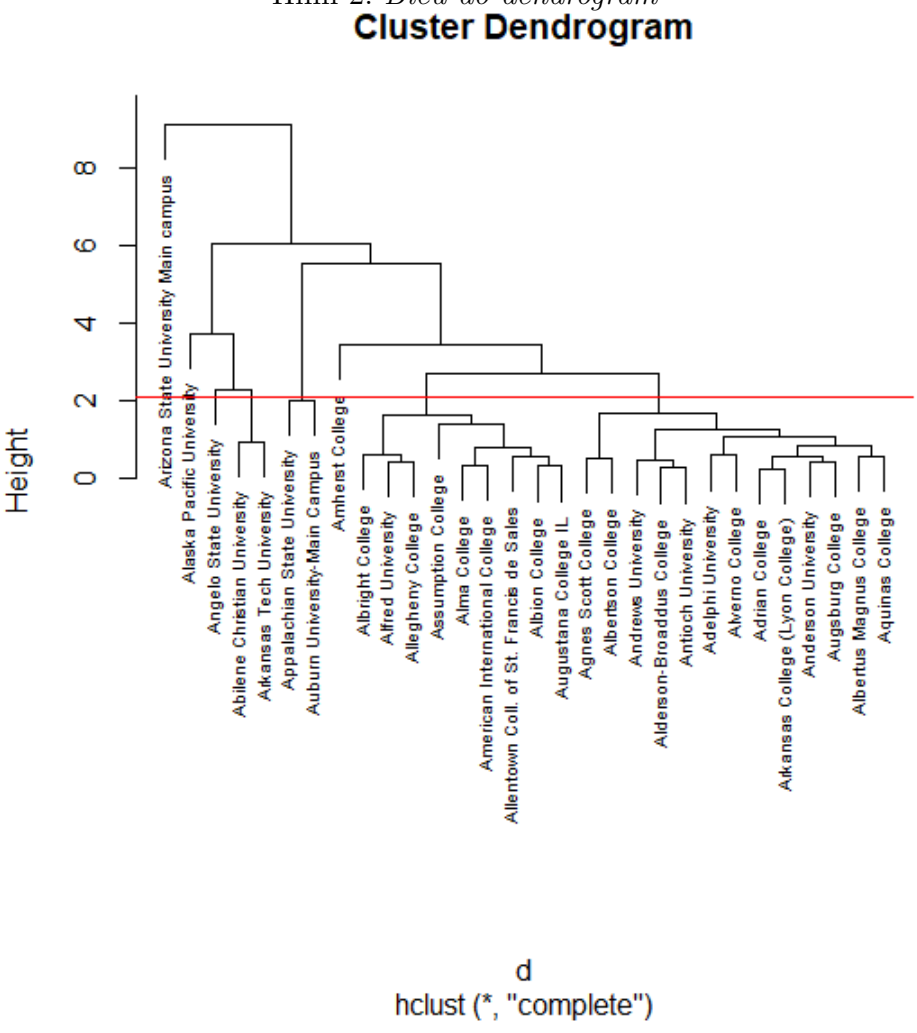
Giá trị riêng và vectơ riêng của ma trận tương quan mẫu của ***data***

```
eigen() decomposition
$values
[1] 2.55868096 1.21872178 0.65176259 0.52444191 0.04639276

$vectors
        [,1]       [,2]        [,3]       [,4]        [,5]
[1,] -0.5529447  0.3265264  0.13590057 -0.3211572  0.68265788
[2,] -0.5782941  0.2602319  0.08481007 -0.2492223 -0.72701499
[3,] -0.4461102 -0.0864730 -0.72817288  0.5090362  0.06445573
[4,] -0.3690286 -0.4617309  0.63271649  0.4993371  0.03089995
[5,]  0.1569450  0.7778090  0.20921852  0.5712188 -0.01783493
```

Hình 1: *Biểu đồ biplot của* ***data***



Hình 2: *Biểu đồ dendrogram*



Câu 2. Dữ liệu được thu thập trên hai loài côn trùng thuộc chi Chaetocnema gồm: Ch. concinna (loài a) và Ch. heikertlingeri (loài b). mỗi loài côn trùng được đánh giá qua ba biến đo sau:

- X_1: Chiều rộng của khớp thứ nhất của xương cổ chân
- X_1: Chiều rộng của khớp thứ hai của xương cổ chân
- X_1: Chiều rộng của aedeagus (cơ quan sinh sản)

Bộ dữ liệu như sau:

Loài	X_1	X_2	$X3$
a	191	131	53
a	185	134	50
a	200	137	52
a	173	127	50
a	171	128	49
a	160	118	47
a	188	134	54
a	186	129	51
a	174	131	52
a	163	115	47
b	186	107	49
b	211	122	49
b	201	144	47
b	242	131	54
b	184	108	43
b	211	118	51
b	217	122	49
b	223	127	51
b	208	125	50
b	199	124	46

Trong trường hợp này, do không có bất kỳ thông tin nào liên quan đến mức độ phong phú tương đối của hai loài nên phân phối xác suất xuất hiện hai loài là như nhau. Kết quả phân tích dữ liệu trên R như sau:

```
> mvn(data, subset="L")
$multivariateNormality
$multivariateNormality$a
Test        HZ   p value MVN
1 Henze-Zirkler 0.4346568 0.6623007 YES

$multivariateNormality$b
Test        HZ   p value MVN
```

```
1 Henze-Zirkler 0.5489465 0.3156677 YES


$univariateNormality
$univariateNormality$a
Test  Variable Statistic   p value Normality
1 Anderson-Darling    X1        0.2419    0.6929    YES
2 Anderson-Darling    X2        0.5084    0.1497    YES
3 Anderson-Darling    X3        0.2324    0.7281    YES

$univariateNormality$b
Test  Variable Statistic   p value Normality
1 Anderson-Darling    X1        0.2160    0.7858    YES
2 Anderson-Darling    X2        0.3394    0.4200    YES
3 Anderson-Darling    X3        0.2826    0.5554    YES


$Descriptives
$Descriptives$a
n  Mean   Std.Dev Median Min Max   25th   75th          Skew
X1 10 179.1 12.879355  179.5 160 200 171.50 187.50  0.0009885795
X2 10 128.4  6.995236  130.0 115 137 127.25 133.25 -0.7336015930
X3 10  50.5  2.368778   50.5  47  54  49.25  52.00 -0.1805668154
Kurtosis
X1 -1.4472317
X2 -0.8503498
X3 -1.3972307

$Descriptives$b
n  Mean   Std.Dev Median Min Max  25th   75th       Skew   Kurtosis
X1 10 208.2 17.222724  209.5 184 242 199.5 215.50  0.3193835 -0.7780416
X2 10 122.8 10.716550  123.0 107 144 119.0 126.50  0.2330507 -0.5913369
X3 10  48.9  3.034981   49.0  43  54  47.5  50.75 -0.3009063 -0.6281605
```

Ngoài ra, sử dụng kiểm định Barlett để kiểm tra tính đồng nhất của ma trận phương sai-hiệp phương sai không tìm thấy sự khác biệt đáng kể giữa ma trận phương sai-hiệp phương sai của hai loài (L' = 9,83; d.f. = 6; p = 0,132). Giả sử tổn thất khi xếp loại là $r_{ii} = 0,\ r_{ij} = 1\ \forall i \neq j$. Thực hiện quan sát một con côn trùng với các chỉ số đo được $X_1 = 194, X_2 = 124, X_3 = 49$. Hỏi con côn trùng này thuộc loài nào?

Hết

Ghi chú: *Sinh viên được dùng tài liệu, cán bộ coi thi không giải thích gì thêm.*

CHỮA ĐỀ THI NĂM 2022-2023

Câu 1. Thầy đã chữa trên lớp.
Câu 2. Có 2 nhóm cùng số mẫu, có phân bố chuẩn cùng phương sai.

$$Y_1 \sim N(\mu_1; \Sigma), Y_2 \sim N(\mu_2; \Sigma).$$

Loài \ Biến	X_1	X_2	X_3
a	179.1	128.4	50.5
b	208.2	122.8	48.9

$$\mu_1 = \begin{pmatrix} 179.1 \\ 128.4 \\ 50.5 \end{pmatrix}, \mu_2 = \begin{pmatrix} 208.2 \\ 122.8 \\ 48.9 \end{pmatrix}$$

Ta sẽ sử dụng phương pháp phân biệt Gauss để xác định con trùng thuộc loài nào. Trước tiên ta cần tính ma trận hiệp phương sai của từng nhóm:

$$S_a = \begin{pmatrix} 165.87778 & 78.95556 & 24.833333 \\ 78.95556 & 48.93333 & 13.888889 \\ 24.833333 & 13.888889 & 5.611111 \end{pmatrix}$$

$$S_a = \begin{pmatrix} 296.62222 & 95.711111 & 43.466667 \\ 95.711111 & 114.844444 & 9.422222 \\ 43.466667 & 9.422222 & 9.211111 \end{pmatrix}$$

$$S = \frac{9 \cdot S_a + 9 \cdot S_b}{18} = \begin{pmatrix} 231.25 & 87.3333 & 34.15 \\ 87.3333 & 81.8889 & 11.6556 \\ 34.15 & 11.6556 & 7.4111 \end{pmatrix}$$

$$S^{-1} = \begin{pmatrix} 0.0178 & -0.0094 & 0.0673 \\ -0.0094 & 0.0207 & 0.0108 \\ 0.0673 & 0.0108 & 0.428 \end{pmatrix}$$

Cá thể X:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 194 \\ 124 \\ 49 \end{pmatrix}$$

Khoảng cách đến nhóm a:

$$S_a = \mu_1^T S^{-1} \begin{pmatrix} 194 \\ 124 \\ 49 \end{pmatrix} - \frac{1}{2} \mu_1^T S^{-1} \mu_1 + \ln\left(\frac{10}{20}\right) = 202.3435.$$

Tương tự, $S_b = 205.1268$.
Xếp loại nào lớn nhất thì thuộc vào loại đó nên cá thể thuộc vào nhóm b.