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

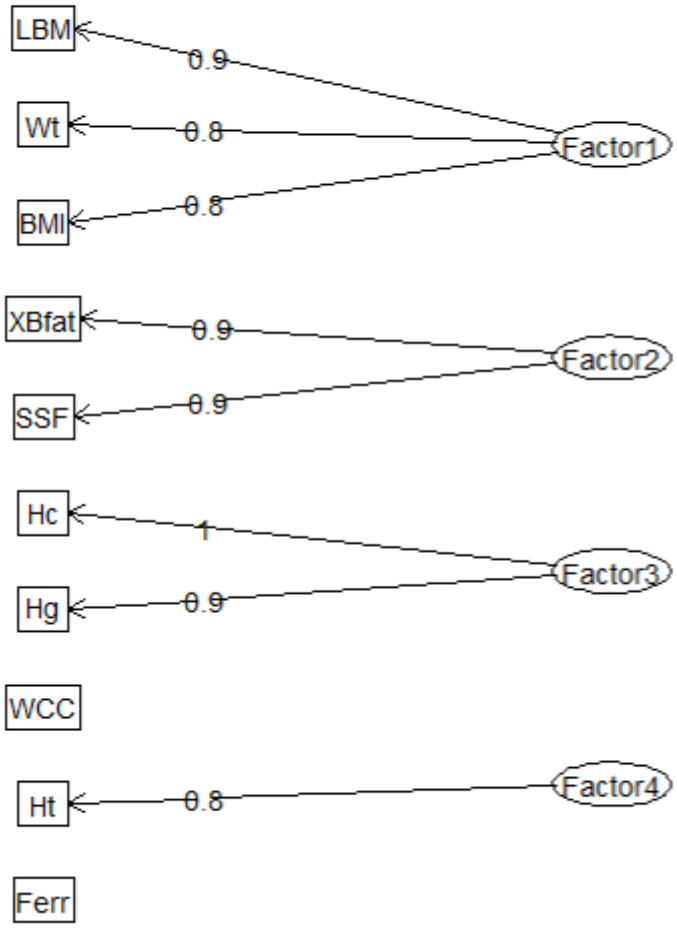
	SSF	XBfat	LBM	Ht	Wt
WCC	0.1197048	0.1186730	0.04830104	-0.006855086	0.08873259
Hc	-0.2246544	-0.1940481	0.11984400	0.020647228	0.01043149
Hg	-0.1580379	-0.1329933	0.16347139	-0.037992530	0.07258187
Ferr	0.1566456	0.1323194	-0.04997854	-0.141889586	0.02128948
BMI	0.6784880	0.6604922	0.74749146	0.231664850	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

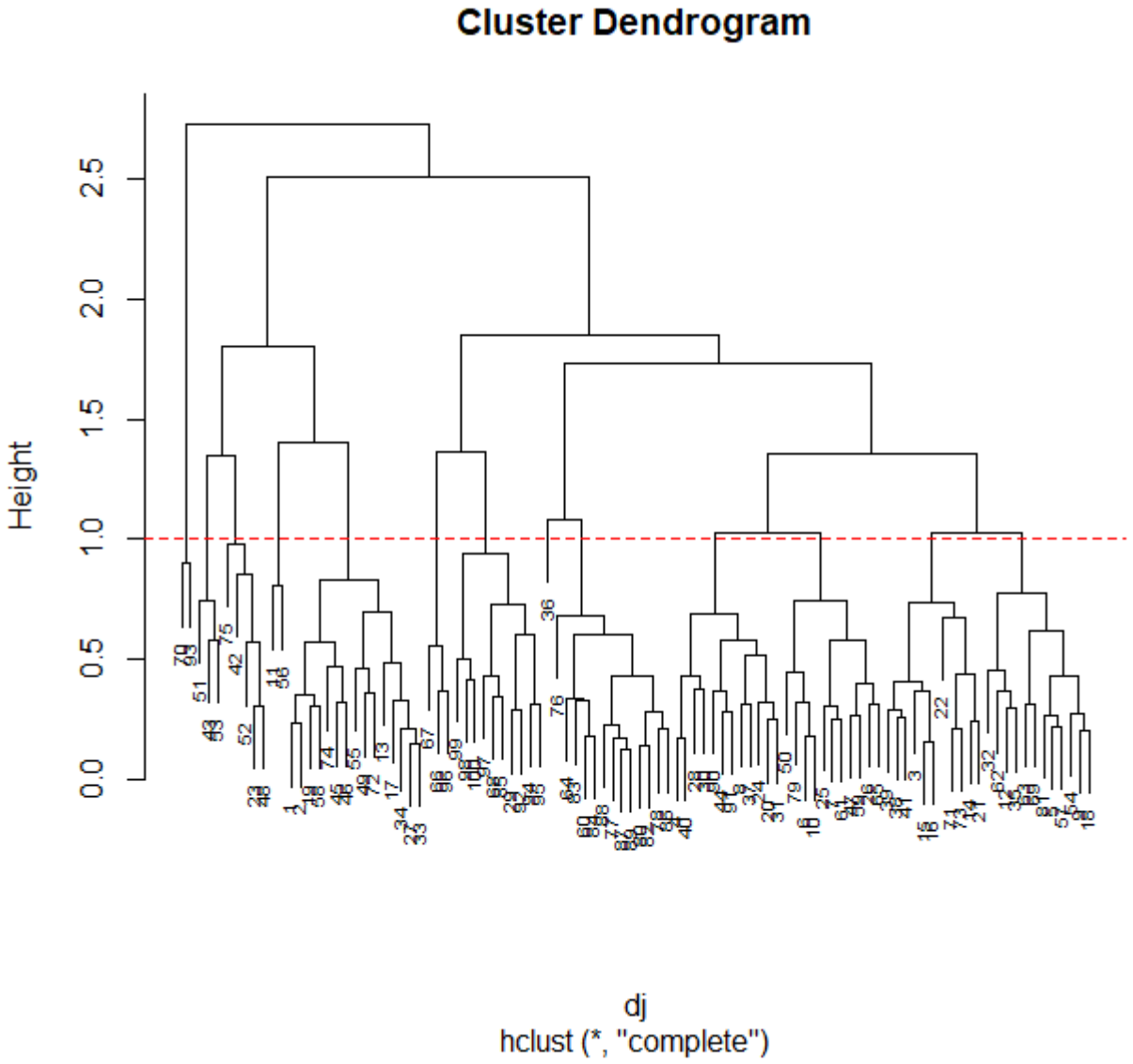
Giá trị riêng và vectơ riêng của ma trận tương quan mẫu của X

```
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
```

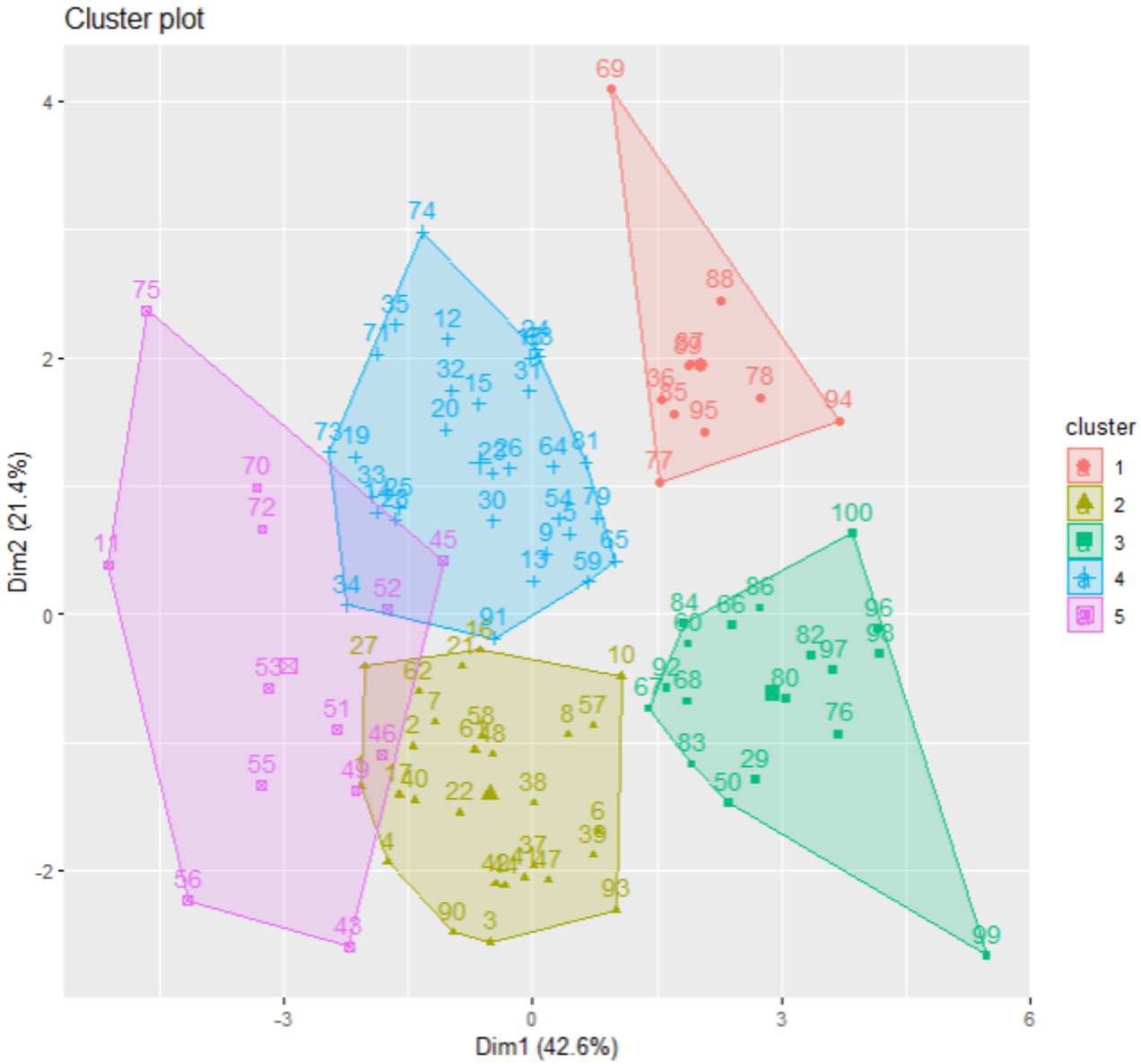
Factor Analysis




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.*