REPORT

수강과목	:	다변량통계학(I)
담당교수	:	최용석
학 과	:	통계학과
학 번	:	201611531
이 름	:	정호재
제출일자	:	2020.07.03.

HW3 for Multivariate Statistics I

June 25, 2020

Chapter 5. Cluster Analysis

Solve the problems (1) ~ (6) in Exercise 5.7.

5.7 [자료 5.8.1] (kellogg.txt)은 켈로그에 의해서 제조된 총 23종류의 시리얼에 대한 것으로 10가지 변수 X_1 ~ X_{10}으로 측정한 결과를 표준화한 자료이다. 여기서 변수들은 다음과 같다.

X_1 : 칼로리, X_2: 단백질(g), X_3: 지방(g), X_4: 나트륨(mg),
X_5: 다이어트 식이섬유(g), X_6: 복합탄수화물(g), X_7: 당분(g), X_8: 칼륨(mg),
X_9: 비타민과 무기물(하루 권장량 (%) : 0, 25, 100), X_{10}: 유형(온 또는 냉)

(1) 주관적 기준에 따라 이진수자료를 만들어 보라.

자료의 데이터가 평균보다 작으면 0 평균보다 크면 1으로 두고 X_{10}은 데이터를 그대로 두어 이진수 자료를 만들었다.

```
> X
      new_X1 new_X2 new_X3 new_X4 new_X5 new_X6 new_X7 new_X8 new_X9 new_X10
AllB       0      1      1      1      1      0      0      1      1       0
AllF       0      1      0      0      1      0      0      1      1       0
AppJ       1      0      0      0      0      0      1      0      0       0
CorF       0      0      0      1      0      1      0      0      0       0
CorP       1      0      0      0      0      0      1      0      0       0
Crac       1      1      1      0      1      0      0      1      1       0
Cris       1      0      0      1      0      1      0      1      0       0
Froo       1      0      1      0      0      0      1      0      0       0
FroF       1      0      0      1      0      0      1      0      0       0
FrMW       0      1      0      0      1      0      0      0      0       0
FruB       1      1      0      1      1      0      1      1      1       0
JRCN       1      0      1      0      0      1      0      1      0       1
JRFN       1      1      1      0      0      1      1      1      0       1
MuCB       1      1      1      0      1      1      1      1      1       0
Nut&       1      0      1      1      0      0      1      0      0       0
NGAR       1      1      1      1      1      1      0      1      1       0
NutW       0      1      0      0      1      1      0      1      0       0
Prod       0      1      0      1      0      1      0      1      0       1
RaBr       1      1      1      1      1      0      1      0      1       0
Rais       0      0      0      0      0      0      0      1      1       0
RiKr       1      0      0      1      0      1      0      0      0       0
Smac       1      0      1      0      0      0      1      0      0       0
Spec       1      1      0      1      0      1      0      0      0       0
```

(2) 유사성계수로 단순매칭계수를 이용하여 비유사성행렬을 구하라.

유클리드 거리에 의한 비유사성 d_{rs}와 단순매칭계수에 의한 유사성 c_{rs}와의 관계는 $d_{rs} = \sqrt{p(1-c_{rs})}$이 된다.
이를 이용하여 비유사성 행렬 d_{rs}를 구하면 다음과 같다.
proxy 패키지를 이용하여 simple matching으로 c_{rs}를 유클리드 거리를 사용하여 d_{rs}를 구하고 이들 간의 관계를 비교해보았다.

단순매칭계수에 의한 유사성 c_{rs}

```
> crs
     AllB AllF AppJ CorF CorP Crac Cris Froo FroF FrMW FruB JRCN JRFN MuCB Nut& NGAR NutW Prod RaBr Rais RiKr Smac
AllF  0.8
AppJ  0.2  0.4
CorF  0.4  0.4  0.6
CorP  0.2  0.4  1.0  0.6
Crac  0.8  0.8  0.4  0.2  0.4
Cris  0.4  0.4  0.6  0.8  0.6  0.4
Froo  0.3  0.3  0.9  0.5  0.9  0.5  0.5
FroF  0.3  0.3  0.9  0.7  0.9  0.3  0.7  0.8
FrMW  0.6  0.8  0.6  0.6  0.6  0.6  0.4  0.5  0.5
FruB  0.7  0.7  0.5  0.3  0.5  0.7  0.5  0.4  0.6  0.5
JRCN  0.3  0.3  0.5  0.5  0.5  0.5  0.7  0.6  0.4  0.3  0.2
JRFN  0.3  0.3  0.5  0.3  0.5  0.5  0.5  0.6  0.4  0.3  0.4  0.8
MuCB  0.6  0.6  0.4  0.2  0.4  0.8  0.4  0.5  0.3  0.4  0.7  0.5  0.7
Nut&  0.4  0.2  0.8  0.6  0.8  0.4  0.6  0.9  0.9  0.4  0.5  0.5  0.5  0.4
NGAR  0.8  0.6  0.2  0.4  0.2  0.8  0.6  0.3  0.3  0.4  0.7  0.5  0.5  0.8  0.4
NutW  0.6  0.8  0.4  0.6  0.4  0.6  0.6  0.3  0.3  0.8  0.5  0.5  0.5  0.6  0.2  0.6
Prod  0.5  0.5  0.3  0.7  0.3  0.3  0.7  0.2  0.4  0.5  0.4  0.6  0.6  0.3  0.3  0.5  0.7
RaBr  0.7  0.5  0.5  0.3  0.5  0.7  0.3  0.6  0.6  0.5  0.8  0.2  0.4  0.7  0.7  0.7  0.3  0.2
Rais  0.6  0.8  0.6  0.6  0.6  0.6  0.6  0.5  0.5  0.6  0.5  0.5  0.3  0.4  0.4  0.4  0.6  0.5  0.3
RiKr  0.3  0.3  0.7  0.9  0.7  0.3  0.9  0.6  0.8  0.5  0.4  0.6  0.4  0.3  0.7  0.5  0.5  0.6  0.4  0.5
Smac  0.3  0.3  0.9  0.5  0.9  0.5  0.5  1.0  0.8  0.5  0.4  0.6  0.6  0.5  0.9  0.3  0.3  0.2  0.6  0.5  0.6
Spec  0.4  0.4  0.6  0.8  0.6  0.4  0.8  0.5  0.7  0.6  0.5  0.5  0.5  0.4  0.6  0.6  0.6  0.7  0.5  0.4  0.9  0.5
```

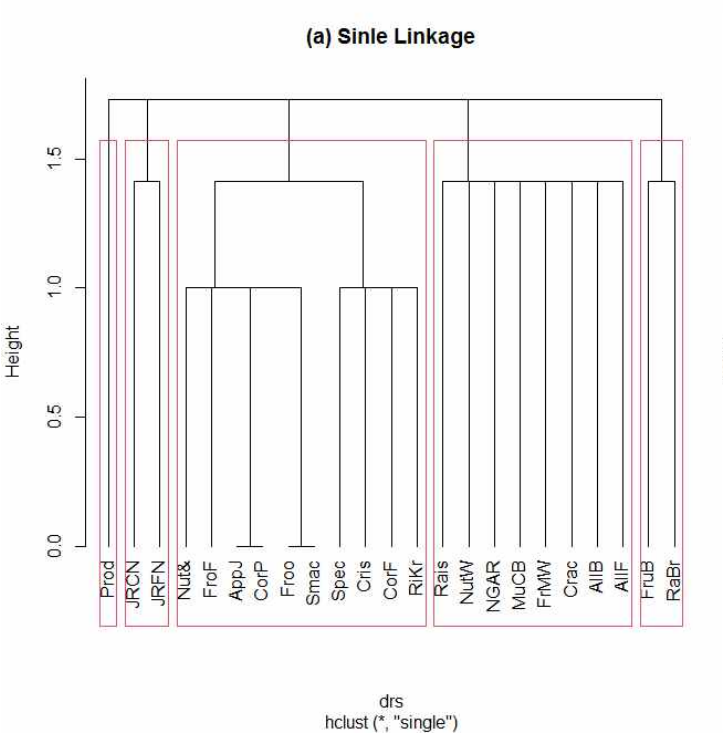
유클리드 거리에 의한 비유사성 d_{rs}

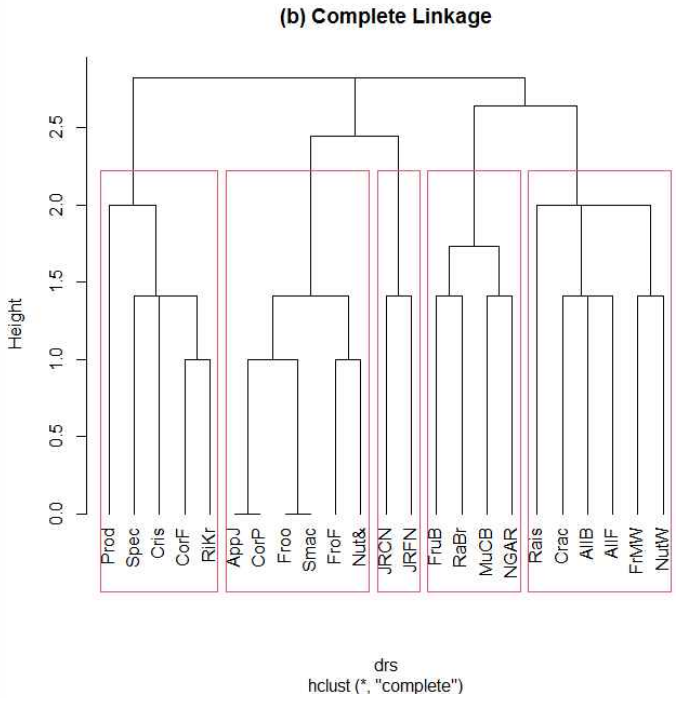
```
> round(drs,2)
     AllB AllF AppJ CorF CorP Crac Cris Froo FroF FrMW FruB JRCN JRFN MuCB Nut& NGAR NutW Prod RaBr Rais RiKr Smac
AllF 1.41
AppJ 2.83 2.45
CorF 2.45 2.45 2.00
CorP 2.83 2.45 0.00 2.00
Crac 1.41 1.41 2.45 2.83 2.45
Cris 2.45 2.45 2.00 1.41 2.00 2.45
Froo 2.65 2.65 1.00 2.24 1.00 2.24 2.24
FroF 2.65 2.65 1.00 1.73 1.00 2.65 1.73 1.41
FrMW 2.00 1.41 2.00 2.00 2.00 2.00 2.45 2.24 2.24
FruB 1.73 1.73 2.24 2.65 2.24 1.73 2.24 2.45 2.00 2.24
JRCN 2.65 2.65 2.24 2.24 2.24 2.24 1.73 2.00 2.45 2.65 2.83
JRFN 2.65 2.65 2.24 2.65 2.24 2.24 2.24 2.00 2.45 2.65 2.45 1.41
MuCB 2.00 2.00 2.45 2.83 2.45 1.41 2.45 2.24 2.65 2.45 1.73 2.24 1.73
Nut& 2.45 2.83 1.41 2.00 1.41 2.45 2.00 1.00 1.00 2.45 2.24 2.24 2.24 2.45
NGAR 1.41 2.00 2.83 2.45 2.83 1.41 2.00 2.65 2.65 2.45 1.73 2.24 2.24 1.41 2.45
NutW 2.00 1.41 2.45 2.00 2.45 2.00 2.00 2.65 2.65 1.41 2.24 2.24 2.24 2.00 2.83 2.00
Prod 2.24 2.24 2.65 1.73 2.65 2.65 1.73 2.83 2.45 2.24 2.45 2.00 2.00 2.65 2.65 2.24 1.73
RaBr 1.73 2.24 2.24 2.65 2.24 1.73 2.65 2.00 2.00 2.24 1.41 2.83 2.45 1.73 1.73 1.73 2.65 2.83
Rais 2.00 1.41 2.00 2.00 2.00 2.00 2.00 2.24 2.24 2.00 2.24 2.24 2.65 2.45 2.45 2.45 2.00 2.24 2.65
RiKr 2.65 2.65 1.73 1.00 1.73 2.65 1.00 2.00 1.41 2.24 2.45 2.00 2.45 2.65 1.73 2.24 2.24 2.00 2.45 2.24
Smac 2.65 2.65 1.00 2.24 1.00 2.24 2.24 0.00 1.41 2.24 2.45 2.00 2.00 2.24 1.00 2.65 2.65 2.83 2.00 2.24 2.00
Spec 2.45 2.45 2.00 1.41 2.00 2.45 1.41 2.24 1.73 2.00 2.24 2.24 2.24 2.45 2.00 2.00 2.00 1.73 2.24 2.45 1.00 2.24
```

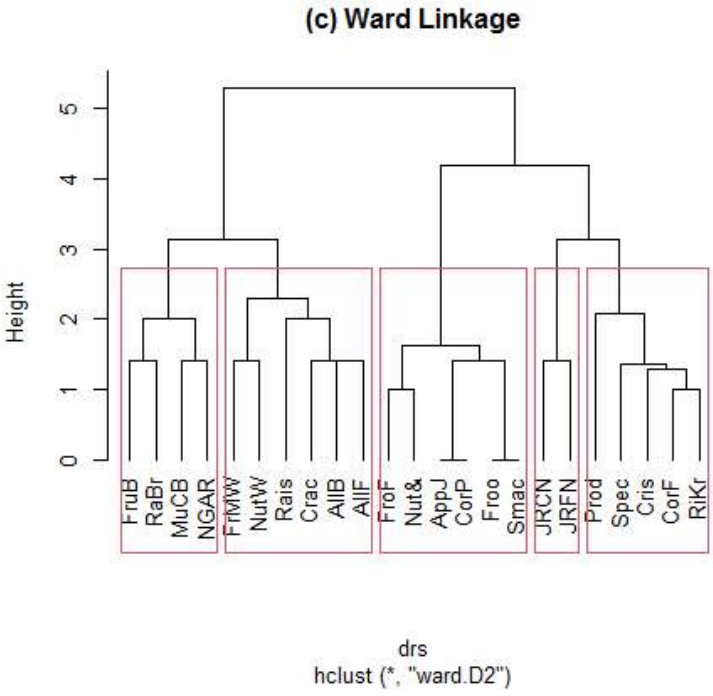
$d_{rs} = \sqrt{p(1-c_{rs})}$ 관계식에 의해 이들의 값은 모두 같음을 알 수 있다.

```
> round(sqrt(p*(1-crs)),2)==round(drs,2)
  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [45] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [89] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[133] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[177] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[221] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

(3) 계층 군집분석인 단일연결법, 평균연결법과 와드법의 덴드로그램을 통해 군집을 얻고 각 군집의 특성을 논하라.





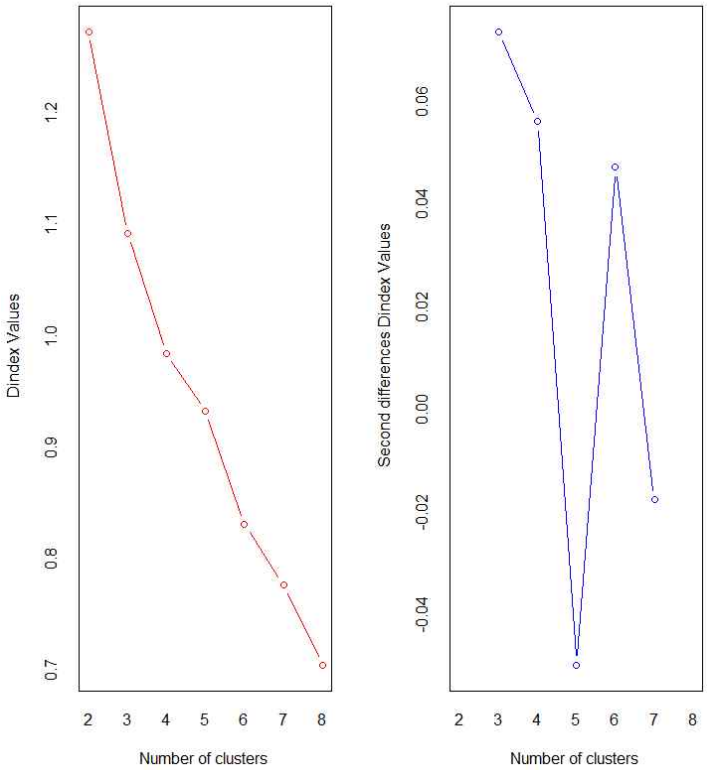


덴드로그램을 통하여 각 5개의 군집을 얻었다. 이를 표로 나타내면 다음과 같다.

군집방법	C1	C2	C3	C4	C5
단일연결법	Prod	JRCN JRFN	Nut&, FroF AppJ, CorP Froo, Smac Spec, Cris CorF, RiKr	Rais, NutW NGAR, MuCB FrMW, Crac AIIB, AIIF	FruB RaBr
단일연결법 군집특성	나트륨과 지방이 낮은 냉시리얼	칼로리가 높은 냉 시리얼	칼로리가 높은 시리얼	칼로리가 낮은 시리얼	다이어트 식이섬유가 높은 시리얼
평균연결법	Prod, Spec Cris, CorF RiKr	Nut&, FroF AppJ, CorP Froo, Smac	JRCN JRFN	FruB, RaBr NGAR, MuCB	Rais, Crac AIIB, AIIF FrMW, NutW
평균연결법 군집특성	나트륨과 복합탄수화물이 높은 시리얼	칼로리와 칼륨이 높은 온시리얼	칼로리가 높은 냉 시리얼	단백질과 다이어트 식이섬유가 높은 시리얼	칼로리가 낮은 시리얼
와드연결법	FruB, RaBr NGAR, MuCB	Rais, Crac AIIB, AIIF FrMW, NutW	Nut&, FroF AppJ, CorP Froo, Smac	JRCN JRFN	Prod, Spec Cris, CorF RiKr
와드연결법 군집특성	단백질과 다이어트 식이섬유가 높은 시리얼	칼로리가 낮은 시리얼	칼로리와 칼륨이 높은 온시리얼	칼로리가 높은 냉 시리얼	나트륨과 복합탄수화물이 높은 시리얼

각 분석방법별로 군집이 조금씩 다르게 뽑혔지만 전체적으로 칼로리와 단백질, 지방, 나트륨, 유형으로 군집이 분류가 되어있다. 단일연결법에서는 군집의 특성을 명확하게 나누지 않았지만 평균연결법과 와드연결법의 경우에서는 조금 더 군집을 세분화 시켜서 분류되어있음을 알 수 있다.

(4) K-평균법과 K-대표개체법에 의한 군집분석을 서로 비교하라.




```
> dindex
$All.index
     2      3      4      5      6      7      8
1.2736 1.0922 0.9844 0.9326 0.8309 0.7764 0.7044
```

Dindex에서는 급격하게 감소하는 지점이 군집의 수 5에 해당하므로 5개의 군집을 분석한다.

K-평균법

```
> C1;C2;C3;C4;C5;C5
     Cereals cluster
JRCN    JRCN       1
JRFN    JRFN       1
     Cereals cluster
CorF    CorF       2
Cris    Cris       2
Prod    Prod       2
RiKr    RiKr       2
Spec    Spec       2
     Cereals cluster
AllF    AllF       3
FrMW    FrMW       3
NutW    NutW       3
Rais    Rais       3
     Cereals cluster
AllB    AllB       4
Crac    Crac       4
FruB    FruB       4
MuCB    MuCB       4
NGAR    NGAR       4
RaBr    RaBr       4
     Cereals cluster
AppJ    AppJ       5
CorP    CorP       5
Froo    Froo       5
FroF    FroF       5
Nut&    Nut&       5
Smac    Smac       5
```

K-대표개체법

```
> C1;C2;C3;C4;C5
     Cereals cluster
AllB    AllB       1
AllF    AllF       1
Crac    Crac       1
FrMW    FrMW       1
FruB    FruB       1
NutW    NutW       1
Rais    Rais       1
     Cereals cluster
AppJ    AppJ       2
CorP    CorP       2
FroF    FroF       2
     Cereals cluster
CorF    CorF       3
Cris    Cris       3
JRCN    JRCN       3
Prod    Prod       3
RiKr    RiKr       3
Spec    Spec       3
     Cereals cluster
Froo    Froo       4
Nut&    Nut&       4
Smac    Smac       4
     Cereals cluster
JRFN    JRFN       5
MuCB    MuCB       5
NGAR    NGAR       5
RaBr    RaBr       5
```

K-평균법의 군집별 특성

```
> aggregate(X, by=list(kmeans$cluster),FUN=mean)
  Group.1    new_X1 new_X2    new_X3    new_X4 new_X5    new_X6 new_X7    new_X8 new_X9 new_X10
1       1 1.0000000   0.50 1.0000000 0.0000000   0.00 1.0000000    0.5 1.0000000    0.0     1.0
2       2 0.6000000   0.40 0.0000000 1.0000000   0.00 1.0000000    0.0 0.4000000    0.0     0.2
3       3 0.0000000   0.75 0.0000000 0.0000000   0.75 0.2500000    0.0 0.7500000    0.5     0.0
4       4 0.8333333   1.00 0.8333333 0.6666667   1.00 0.3333333    0.5 0.8333333    1.0     0.0
5       5 1.0000000   0.00 0.5000000 0.3333333   0.00 0.0000000    1.0 0.0000000    0.0     0.0
```

K-대표개체법의 군집별 특성

```
> aggregate(X, by=list(kmedoids$cluster),FUN=mean)
  Group.1    new_X1    new_X2    new_X3    new_X4    new_X5    new_X6    new_X7    new_X8    new_X9   new_X10
1       1 0.2857143 0.8571429 0.2857143 0.2857143 0.8571429 0.1428571 0.1428571 0.8571429 0.7142857 0.0000000
2       2 1.0000000 0.0000000 0.0000000 0.3333333 0.0000000 0.0000000 1.0000000 0.0000000 0.0000000 0.0000000
3       3 0.6666667 0.3333333 0.1666667 0.8333333 0.0000000 1.0000000 0.0000000 0.5000000 0.0000000 0.3333333
4       4 1.0000000 0.0000000 1.0000000 0.3333333 0.0000000 0.0000000 1.0000000 0.0000000 0.0000000 0.0000000
5       5 1.0000000 1.0000000 1.0000000 0.5000000 0.7500000 0.7500000 0.7500000 0.7500000 0.7500000 0.2500000
```

K-평균법과 K-대표개체법 각 5개의 군집을 얻었다. 이를 표로 나타내면 다음과 같다.

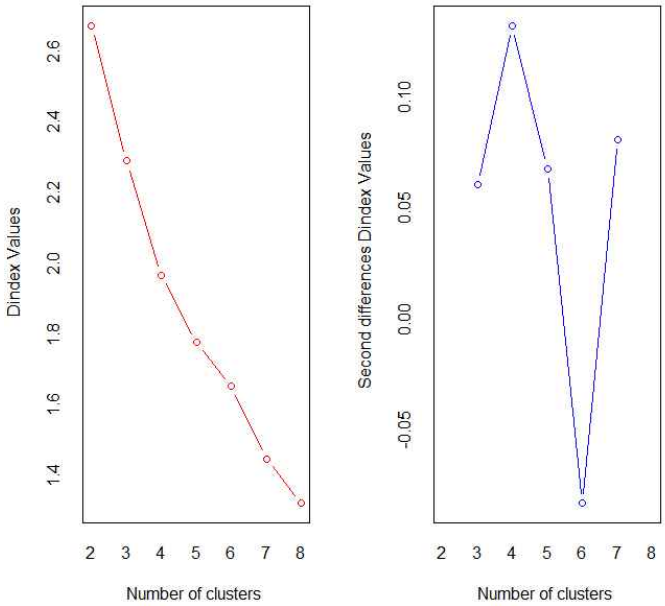
군집방법	C1	C2	C3	C4	C5
K-평균법	JRCN JRFN	CorF, Cris Prod, RiKr Spec	AllF, FrMW NutW, Rais	AllB, Crac FruB, MuCB NGAR, RaBr	AppJ, CorP Froo, FroF Nut&, Smac
K-평균법 군집특성	칼로리가 높은 냉 시리얼	나트륨과 복합탄수화물이 높은 시리얼	칼로리와 지방, 나트륨이 낮은 시리얼	단백질과 다이어트 식이섬유가 높은 시리얼	칼로리와 칼륨이 높은 온시리얼
K-대표개체법	AllB, AllF Crac, FrMW FruB, NutW Rais	AppJ CorP FroF	CorF, Cris3 JRCN, Prod RiKr, Spec	Froo Nut& Smac	JRFN MuCB NGAR RaBr
K-대표개체법 군집특성	칼로리가 낮은 시리얼	칼로리와 당이 높지만 지방은 낮은 시리얼	나트륨과 복합탄수화물이 높은 시리얼	칼로리와 당이 높은 시리얼	칼로리, 단백질, 지방이 높은 시리얼

K-평균법과 K-대표개체법을 비교하였을 때 군집이 조금씩 다르게 뽑혔지만 전체적으로 비슷한 경향으로 군집이 나누어져있다.

(5) (3)과 (4)의 결과를 서로 비교하라.

(3)과 (4)의 결과를 서로 비교해보았을 때 평균연결법, 와드연결법, K-평균법, K-대표개체법은 모두 크게는 칼로리와 에너지원, 칼륨과 다이어트 식이섬유 등으로 군집을 나누었다. 평균연결법과 와드연결법에 비해 와드연결법, K-평균법, K-대표개체법은 칼로리가 높고 낮은 집단에 추가적으로 다이어트 식이섬유, 지방, 나트륨 등의 변수를 추가적으로 더 비교해주었다.

(6) 시리얼 간의 유클리드 거리에 대한 K-평균법을 실시하고 (4)의 결과와 비교하라.




```
> dindex
$All.index
     2      3      4      5      6      7      8
2.6740 2.2907 1.9678 1.7770 1.6539 1.4478 1.3226
```

Dindex에서는 급격하게 감소하는 지점이 군집의 수 6에 해당하므로 6개의 군집을 분석한다.

K-평균법

```
> C1;C2;C3;C4;C5;C6
     Creals cluster
CorF   CorF       1
Cris   Cris       1
NutW   NutW       1
RiKr   RiKr       1
     Creals cluster
Spec   Spec       2
     Creals cluster
AllB   AllB       3
AllF   AllF       3
     Creals cluster
Crac   Crac       4
FruB   FruB       4
MuCB   MuCB       4
NGAR   NGAR       4
RaBr   RaBr       4
     Creals cluster
AppJ   AppJ       5
CorP   CorP       5
Froo   Froo       5
FroF   FroF       5
FrMW   FrMW       5
Nut&   Nut&       5
Rais   Rais       5
Smac   Smac       5
     Creals cluster
JRCN   JRCN       6
JRFN   JRFN       6
Prod   Prod       6
```

K-대표개체법

```
> C1;C2;C3;C4;C5;C6
     Creals cluster
AllB   AllB       1
AllF   AllF       1
     Creals cluster
AppJ   AppJ       2
CorP   CorP       2
Froo   Froo       2
FroF   FroF       2
Nut&   Nut&       2
Smac   Smac       2
     Creals cluster
CorF   CorF       3
Cris   Cris       3
RiKr   RiKr       3
Spec   Spec       3
     Creals cluster
Crac   Crac       4
FruB   FruB       4
MuCB   MuCB       4
NGAR   NGAR       4
RaBr   RaBr       4
     Creals cluster
FrMW   FrMW       5
NutW   NutW       5
Rais   Rais       5
     Creals cluster
JRCN   JRCN       6
JRFN   JRFN       6
Prod   Prod       6
```

K-평균법의 군집별 특성

```
> aggregate(X, by=list(kmeans$cluster),FUN=mean)
  Group.1        X1        X2        X3        X4         X5        X6      X7  X8        X9 X10
1       1 0.4772750 0.2500000 0.0000000 0.7578500 0.08927500 0.8999750 0.16665 0.5 0.0887250   0
2       2 0.5455000 1.0000000 0.0000000 0.7188000 0.07140000 0.6000000 0.20000 0.0 0.1129000   0
3       3 0.0909000 0.6000000 0.1666500 0.6250000 0.82145000 0.0333500 0.16665 1.0 0.9838500   0
4       4 0.7273000 0.4000000 0.5333400 0.6000200 0.28570000 0.5466800 0.68002 0.9 0.5032200   0
5       5 0.5227500 0.1750000 0.1249875 0.3125125 0.08927500 0.3833375 0.72500 0.5 0.0947625   0
6       6 0.6060667 0.3333333 0.2222000 0.6875333 0.09523333 0.8000333 0.40000 1.0 0.1505000   1
```

K-대표개체법의 군집별 특성

```
> aggregate(X, by=list(kmedoids$cluster),FUN=mean)
  Group.1        X1        X2      X3        X4         X5        X6        X7        X8        X9 X10
1       1 0.0909000 0.6000000 0.16665 0.6250000 0.82145000 0.0333500 0.1666500 1.0000000 0.9838500   0
2       2 0.5606500 0.1333333 0.16665 0.4166833 0.05950000 0.3444500 0.8222167 0.4166667 0.0349500   0
3       3 0.5227500 0.4000000 0.00000 0.8047250 0.05355000 0.8666500 0.1833250 0.2500000 0.0605000   0
4       4 0.7273000 0.4000000 0.53334 0.6000200 0.28570000 0.5466800 0.6800200 0.9000000 0.5032200   0
5       5 0.3939000 0.3333333 0.00000 0.1771000 0.19050000 0.5777667 0.3333333 0.8333333 0.2580667   0
6       6 0.6060667 0.3333333 0.22220 0.6875333 0.09523333 0.8000333 0.4000000 1.0000000 0.1505000   1
```

군집방법	C1	C2	C3	C4	C5	C6
단일연결법	Corf RiKr Cris NutW	Spec	AIIB AIIF	Crac, FruB RaBr, MuCB NGAR	FroF, Nut& Froo, Smac AppJ, CorP FrMW, Rais	Prod JRCN JRFN
단일연결법 군집특성	칼로리와 복합탄수화물이 낮은 시리얼	단백질이 높은 시리얼	칼로리와 복합탄수화물이 낮은 시리얼	칼로리와 지방이 높은 시리얼	나트륨이 적고 당 높은 시리얼	냉 시리얼
평균연결법	AIIB AIIF	FroF, Nut& Froo, Smac AppJ, CorP	Corf RiKr Cris Spec	Crac, FruB RaBr, MuCB NGAR	Prod JRCN JRFN	Prod JRCN JRFN
평균연결법 군집특성	칼로리와 복합탄수화물이 낮은 시리얼	단백질과 다이어트 식이섬유가 낮고 당분이 높은 시리얼	나트륨과 복합탄수화물이 높은 시리얼	칼로리와 지방이 높은 시리얼	나트륨과 지방이 낮은 시리얼	냉 시리얼

위의 군집을 (4)의 결과와 비교해보면 전체적인 데이터 군집의 경향은 비슷하나 (4)의 결과에 비하여 군집의 수가 5개에서 6개로 늘어났다. 또한 지금의 데이터는 유형과 칼로리, 단백질에 대하여 더 민감하게 반응하고 있다.

```
#data
setwd("D:/2020 1학기 정호재/다변량통계학(1)/200625 다변량 실습 6/data")
Data5.8.1<-read.table("kellogg.txt", header=T)
X<-Data5.8.1[,-1]
cereal<-Data5.8.1[,1]
rownames(X)<-cereal
X
#5.7.1
dummy_var <- transform(X,
         new_X1 = ifelse(X1 >mean(X1),1,0),
         new_X2 = ifelse(X2 >mean(X2),1,0),
         new_X3 = ifelse(X3 >mean(X3),1,0),
         new_X4 = ifelse(X4 >mean(X4),1,0),
         new_X5 = ifelse(X5 >mean(X5),1,0),
         new_X6 = ifelse(X6 >mean(X6),1,0),
         new_X7 = ifelse(X7 >mean(X7),1,0),
         new_X8 = ifelse(X8 >mean(X8),1,0),
         new_X9 = ifelse(X9 >mean(X9),1,0),
         new_X10= X10)
X<-dummy_var[,-c(1:10)]
X<-as.matrix(X)
X

#5.7.2
n<-nrow(X)
p<-ncol(X)
drs<-dist(X, method="euclidean")
round(drs,2)
crs<-1-drs^2/p
crs

install.packages("proxy")
library(proxy)
summary(pr_DB)
crs<-1-dist(X, method = "simple matching")
drs<-sqrt(p*(1-crs))
drs
crs
round(sqrt(p*(1-crs)),2)==round(drs,2)
```

```
#5.7.3
par(mfrow=c(2,2))
#단일연결법
single=hclust(drs, method="single")
plot(single, hang=-1, main="(a) Sinle Linkage")
rect.hclust(single,k=5)
#완전연결법
complete=hclust(drs, method="complete")
plot(complete,hang=-1, main="(b) Complete Linkage")
rect.hclust(complete,k=5)
#와드연결법
ward=hclust(drs, method="ward.D2")
plot(ward, hang=-1, main="(c) Ward Linkage")
rect.hclust(ward,k=5)

#5.7.4
install.packages("NbClust") #군집의 개수 정해줌
library(NbClust)
#Dindex Index
dindex<-NbClust(X, distance="euclidean", min.nc = 2, max.nc = 8,
                method = "kmeans", index = "dindex")
dindex
Cereals=Data5.8.1[,1]
kmeans <- kmeans(X, 5)
kmeans
cluster=data.frame(Cereals,cluster=kmeans$cluster)
C1=cluster[(cluster[,2]==1),]
C2=cluster[(cluster[,2]==2),]
C3=cluster[(cluster[,2]==3),]
C4=cluster[(cluster[,2]==4),]
C5=cluster[(cluster[,2]==5),]
C1;C2;C3;C4;C5
# Get cluster means
aggregate(X, by=list(kmeans$cluster),FUN=mean)

library(cluster)
kmedoids <- pam(X, 5, metric="euclidean")
cluster=data.frame(Cereals,cluster=kmedoids$cluster)
C1=cluster[(cluster[,2]==1),]
C2=cluster[(cluster[,2]==2),]
C3=cluster[(cluster[,2]==3),]
C4=cluster[(cluster[,2]==4),]
C5=cluster[(cluster[,2]==5),]
C1;C2;C3;C4;C5
```

```
# Get cluster means
aggregate(X, by=list(kmedoids$cluster),FUN=mean)
#5.7.6
# Get cluster means
aggregate(X, by=list(kmedoids$cluster),FUN=mean)
Data5.8.1<-read.table("kellogg.txt", header=T)
X<-Data5.8.1[,-1]
cereal<-Data5.8.1[,1]
rownames(X)<-cereal
n<-nrow(X)
xbar<-t(X)%*%matrix(1,n,1)/n # 평균벡터
I<-diag(n)
J<-matrix(1,n,n)
H<-I-1/n*J                    # 중심화행렬
Y<-H%*%as.matrix(X)           # 중심화 자료행렬
S<-t(Y)%*%Y/(n-1)             # 공분산행렬
D<-diag(1/sqrt(diag(S)))      # 표준편차행렬의 역
Z<-Y%*%D                      # 표준화자료행렬
colnames(Z)<-colnames(X)
rownames(Z)<-rownames(X)

install.packages("NbClust") #군집의 개수 정해줌
library(NbClust)
Cereals=Data5.8.1[,1]

#Dindex Index
dindex<-NbClust(Z, distance="euclidean", min.nc = 2, max.nc = 8,
                method = "kmeans", index = "dindex")
dindex

kmeans <- kmeans(Z, 6) # 6 cluster solution
kmeans
cluster=data.frame(Cereals,cluster=kmeans$cluster)
C1=cluster[(cluster[,2]==1),]
C2=cluster[(cluster[,2]==2),]
C3=cluster[(cluster[,2]==3),]
C4=cluster[(cluster[,2]==4),]
C5=cluster[(cluster[,2]==5),]
C6=cluster[(cluster[,2]==6),]
C1;C2;C3;C4;C5;C6

# Get cluster means
aggregate(X, by=list(kmeans$cluster),FUN=mean)
```

```
library(cluster)
kmedoids <- pam(Z, 6, metric="euclidean") # 6 cluster solution
cluster=data.frame(Cereals,cluster=kmedoids$cluster)
C1=cluster[(cluster[,2]==1),]
C2=cluster[(cluster[,2]==2),]
C3=cluster[(cluster[,2]==3),]
C4=cluster[(cluster[,2]==4),]
C5=cluster[(cluster[,2]==5),]
C6=cluster[(cluster[,2]==6),]
C1;C2;C3;C4;C5;C6

# Get cluster means
aggregate(X, by=list(kmedoids$cluster),FUN=mean)
```