

# 다변량

## 5장 과제

과목	다변량데이터분석
담당교수	임태진 교수님
전공	산업정보시스템공학과
학번	20201368
이름	한채원
제출일	2022.10.31

1. 환경 변수들 고도, 연 강수량, 연 최고 기온, 연 최저 기온을 평균 0과 분산 1이 되도록 표준화하고 환경 거리행렬을 얻기 위해 수식 (5.1)을 이용하여 모든 쌍의 콜로니에 대하여 유클리디안 거리를 계산하라.

```
> setwd("c:/temp/data")
> bc = read.csv('Butterfly colonies.csv', row.names=1)
> str(bc)
'data.frame': 16 obs. of 10 variables:
 $ Altitude : num 500 808 570 550 550 380 930 650 600 1500 ...
 $ Ann.Prec : num 43 20 28 28 28 15 21 10 10 19 ...
 $ Max.Temp : num 98 92 98 98 98 99 99 101 101 99 ...
 $ Min.Temp : num 17 32 26 26 26 28 28 27 27 23 ...
 $ PFPGi0.40: num 0 0 0 0 0 0 0 10 14 0 ...
 $ PFPGi0.60: num 3 16 6 4 1 2 0 21 26 1 ...
 $ PFPGi0.80: num 22 20 28 19 8 19 15 40 32 6 ...
 $ PFPGi1.00: num 57 38 46 47 50 44 50 25 28 80 ...
 $ PFPGi1.16: num 17 13 17 27 35 32 27 4 0 12 ...
 $ PFPGi1.30: num 1 13 3 3 6 3 8 0 0 1 ...
```

scale 작업을 해준다.

```
> bc.std = scale(bc[1:4]); round(bc.std,2)
      Altitude Ann.Prec Max.Temp Min.Temp
SS      -0.56      1.07      0.12     -0.35
SB      -0.45     -0.58     -0.82      1.02
WSB     -0.53      0.00      0.12      0.47
JRC     -0.54      0.00      0.12      0.47
JRH     -0.54      0.00      0.12      0.47
SJ      -0.60     -0.93      0.27      0.65
CR      -0.41     -0.50      0.27      0.65
UO      -0.50     -1.29      0.59      0.56
LO      -0.52     -1.29      0.59      0.56
DP      -0.21     -0.65      0.27      0.19
PZ      -0.12     -0.43      0.59      0.56
MC      -0.04      2.14      0.43     -0.26
IF       0.14      0.42      0.74     -0.44
AF      -0.04     -0.50      1.21     -0.08
GH       2.00      1.00     -2.08     -1.45
GL       2.92      1.57     -2.55     -3.00
attr(,"scaled:center")
      Altitude Ann.Prec Max.Temp Min.Temp
```

```
2102.3750  28.0625  97.2500  20.8750
attr(,"scaled:scale")
      Altitude  Ann.Prec  Max.Temp  Min.Temp
2879.118774    13.992706    6.382267    10.959775
```

```
> bc.d1 = dist(bc.std, method = 'euclidean')
> round(bc.d1,3)
```

```
      SS  SB  WSB  JRC  JRH  SJ  CR  UO  LO  DP  PZ  MC  IF  AF  GH
SB  2.339
WSB  1.351 1.232
JRC  1.350 1.232 0.007
JRH  1.350 1.232 0.007 0.000
SJ   2.245 1.219 0.962 0.961 0.961
CR   1.878 1.159 0.569 0.571 0.571 0.469
UO   2.573 1.646 1.373 1.373 1.373 0.493 0.857
LO   2.572 1.647 1.373 1.373 1.373 0.490 0.859 0.017
DP   1.840 1.393 0.786 0.789 0.789 0.664 0.517 0.856 0.862
PZ   1.869 1.525 0.762 0.766 0.766 0.764 0.439 0.939 0.946 0.534
MC   1.236 3.279 2.340 2.341 2.341 3.258 2.826 3.562 3.564 2.834 2.707
IF   1.139 2.436 1.363 1.367 1.367 1.951 1.606 2.094 2.100 1.378 1.355 1.762
AF   2.005 2.350 1.414 1.417 1.417 1.385 1.247 1.280 1.287 1.005 0.902 2.764 1.117
GH   3.540 4.012 3.985 3.989 3.989 4.515 4.238 4.755 4.764 3.973 4.202 3.623 3.570 4.368
GL   5.139 5.918 5.787 5.791 5.791 6.316 6.051 6.508 6.518 5.730 5.977 5.040 5.136 5.973 1.950
```

위의 결과에서 JRC와 WSB, JRH와 WSB가 굉장히 유사하다. 이유는 이 그룹 간의 거리 (0.007)가 표전체 다른 값들보다 훨씬 작기 때문이다.

2. 비율로 전환된 Pgi 유전자 빈도를 이용하여 콜로니들 사이의 유전적 거리를 수식(5.5)로 계산하라.

```
> bc.std2=(bc[5:10])
> bc.d2 = dist(bc.std2, method = 'manhattan')
```

5.5에 제시된 식을 만족하기 위해 모든 값마다 0.5와 %단위인 0.01을 곱해준다.

```
> bc.d2 = 0.5*0.01*bc.d2
> bc.d2
```

```
      SS  SB  WSB  JRC  JRH  SJ  CR  UO  LO  DP  PZ  MC  IF  AF  GH
SB  0.25
WSB  0.11 0.20
JRC  0.13 0.23 0.11
JRH  0.23 0.34 0.25 0.14
SJ   0.17 0.25 0.15 0.05 0.12
CR   0.17 0.26 0.19 0.08 0.09 0.11
UO   0.46 0.35 0.37 0.48 0.62 0.50 0.56
LO   0.47 0.36 0.38 0.49 0.63 0.51 0.57 0.12
DP   0.23 0.42 0.34 0.33 0.30 0.36 0.31 0.64 0.65
PZ   0.24 0.24 0.15 0.19 0.30 0.21 0.24 0.32 0.35 0.47
```

```
MC 0.13 0.28 0.21 0.22 0.28 0.27 0.23 0.50 0.51 0.15 0.36
IF 0.17 0.17 0.13 0.10 0.17 0.15 0.09 0.47 0.48 0.33 0.21 0.19
AF 0.30 0.18 0.25 0.17 0.26 0.19 0.18 0.44 0.45 0.48 0.18 0.34 0.17
GH 0.29 0.46 0.38 0.38 0.38 0.43 0.39 0.59 0.60 0.09 0.52 0.18 0.37 0.52
GL 0.35 0.54 0.46 0.45 0.44 0.49 0.45 0.67 0.68 0.14 0.59 0.26 0.45 0.60 0.08
```

0에 가까울수록 유사하고, 1에 가까울수록 유사하지 않는데 GL과 GH간의 유전적 거리가 작으므로 상당히 유사하다고 볼 수 있다. 값이 0.08로 표 전체 다른 값들보다 훨씬 작기 때문이다.

3. 환경 그리고 유전적 거리행렬 사이에 유의미한 양의 상관관계가 있는지를 결정하기 위해 Mantel의 행렬 임의화 검정을 수행하고 결론에 대해 논하라.

1번에서 환경 거리 행렬을 구했고, 2번에서 유전적 거리행렬을 구했다.

```
> bc.dm1 = as.matrix(bc.d1)
> bc.dm2 = as.matrix(bc.d2)
> library(vegan)
> (mantel(bc.dm1,bc.dm2))
```

Mantel statistic based on Pearson's product-moment correlation

Call:

```
mantel(xdis = bc.dm1, ydis = bc.dm2)
```

Mantel statistic r: 0.4354

Significance 0.021

Upper quantiles of permutations (null model):

90% 95% 97.5% 99%

0.251 0.319 0.403 0.511

Permutation: free

Number of permutations: 999

만델의 상관계수는 0.4354로 유의수준은 0.021로 매우 작으므로 환경 그리고 유전적 거리 행렬 사이에 유의미한 양의 상관관계가 있다고 볼 수 있다.

4. 이와 같이 임의화 검정의 유의미한 양의 상관관계가 콜로니들의 유전적 구성에 영향을 미친 환경적 요인의 결과라기보다 근접한 콜로니들의 공간적 상관관계의 결과일 수 있는 이유를 설명하라.

우선 공간적 상관관계를 구하기 위해 주교재 9쪽에 있는 그림을 참조하여 거리를 재봤다.

```
>S.dist=matrix(c(0,4.15,4.75,4.4,4.5,4.6,4.8,9.9,10.05,4.7,6.75,2.4,4.9,5.1,6.2,4.75,
4.15,0,0.1,0.2,0.3,0.5,0.8,6.7,6.85,0.85,2.8,2.1,2.05,1.9,2.95,2.2,
4.75,0.1,0,0.1,0.2,0.4,0.7,6.1,6.25,0.8,2.6,2.25,2.05,1.9,2.9,2.15,
4.4,0.2,0.1,0,0.1,0.3,0.6,6.05,6.1,0.7,2.5,2.3,2,1.9,2.9,2.2,
4.5,0.3,0.2,0.1,0,0.3,0.5,5.95,6.05,0.65,2.4,2.4,1.95,1.85,2.8,2.2,
4.6,0.5,0.4,0.3,0.3,0,0.3,5.7,5.85,0.4,2.25,2.4,1.7,1.55,2.5,1.9,
4.8,0.8,0.7,0.6,0.5,0.3,0,5.4,5.55,0.25,1.95,2.6,1.55,1.85,2.25,2.75,
9.9,6.7,6.1,6.05,5.95,5.7,5.4,0,0.1,5.45,3.55,7.5,5.05,4.85,3.8,5.2,
10.05,6.85,6.25,6.1,6.05,5.85,5.55,0.1,0,5.65,3.7,7.6,5.15,4.95,4.9,5.3,
4.7,0.85,0.8,0.7,0.65,0.4,0.25,5.45,5.65,0,2.1,2.4,1.3,1.15,2.15,1.55,
6.75,2.8,2.6,2.5,2.4,2.25,1.95,3.55,3.7,2.1,0,4.4,2.35,2.05,1.8,2.6,
2.4,2.1,2.25,2.3,2.4,2.4,2.6,7.5,7.6,2.4,4.4,0,2.5,2.7,3.75,2.9,
4.9,2.05,2.05,2,1.95,1.7,1.55,5.05,5.15,1.3,2.35,2.5,0,0.3,1.8,0.3,
5.1,1.9,1.9,1.9,1.85,1.55,1.85,4.85,4.95,1.15,2.05,2.7,0.3,0,1.15,0.6,
6.2,2.95,2.9,2.9,2.8,2.5,2.25,3.8,4.9,2.15,1.8,3.75,1.8,1.15,0,1.95,
4.75,2.2,2.15,2.2,2.2,1.9,2.75,5.2,5.3,1.55,2.6,2.9,0.3,0.6,1.95,0), nrow=16,ncol=16)
> S.dist = (76.923*S.dist) # scale에 해당하는 mile 값을 변환해준 값 76.923을 곱한 것이
다.
```

책에 제시된 표와 똑같이 작성하기 위해 Colony의 순서를 변경해준다.

```
> bc2 = read.csv('Butterfly colonies.csv', string=T)
> colony = bc2[[1]]
> a = levels(colony)
> a
[1] "AF" "CR" "DP" "GH" "GL" "IF" "JRC" "JRH" "LO" "MC" "PZ" "SB" "SJ" "SS"
"UO" "WSB"
> colony = factor(colony,a[c(14,12,16,7,8,13,2,15,9,3,11,10,6,1,4,5)])
> levels(colony)
[1] "SS" "SB" "WSB" "JRC" "JRH" "SJ" "CR" "UO" "LO" "DP" "PZ" "MC" "IF" "AF"
"GH" "GL"
> rownames(S.dist)=colnames(S.dist)=levels(colony)
> (S.dist=as.dist(S.dist))
```

	SS	SB	WSB	JRC	JRH	SJ	CR	UO	LO	DP	PZ
SB	319.23045										
WSB	365.38425	7.69230									
JRC	338.46120	15.38460	7.69230								
JRH	346.15350	23.07690	15.38460	7.69230							
SJ	353.84580	38.46150	30.76920	23.07690	23.07690						
CR	369.23040	61.53840	53.84610	46.15380	38.46150	23.07690					
UO	761.53770	515.38410	469.23030	465.38415	457.69185	438.46110	415.38420				
LO	773.07615	526.92255	480.76875	469.23030	465.38415	449.99955	426.92265	7.69230			
DP	361.53810	65.38455	61.53840	53.84610	49.99995	30.76920	19.23075	419.23035	434.61495		
PZ	519.23025	215.38440	199.99980	192.30750	184.61520	173.07675	149.99985	273.07665	284.61510	161.53830	

```

MC 184.61520 161.53830 173.07675 176.92290 184.61520 184.61520 199.99980 576.92250 584.61480 184.61520 338.46120
IF 376.92270 157.69215 157.69215 153.84600 149.99985 130.76910 119.23065 388.46115 396.15345 99.99990 180.76905
AF 392.30730 146.15370 146.15370 146.15370 142.30755 119.23065 142.30755 373.07655 380.76885 88.46145 157.69215
GH 476.92260 226.92285 223.07670 223.07670 215.38440 192.30750 173.07675 292.30740 376.92270 165.38445 138.46140
GL 365.38425 169.23060 165.38445 169.23060 169.23060 146.15370 211.53825 399.99960 407.69190 119.23065 199.99980
      MC      IF      AF      GH
SB
WSB
JRC
JRH
SJ
CR
UO
LO
DP
PZ
MC
IF 192.30750
AF 207.69210 23.07690
GH 288.46125 138.46140 88.46145
GL 223.07670 23.07690 46.15380 149.99985

```

만델의 임의화 검정을 실시해본다.

```
> (mantel(S.dist, bc.dm2))
```

Mantel statistic based on Pearson's product-moment correlation

Call:

```
mantel(xdis = S.dist, ydis = bc.dm2)
```

Mantel statistic r: 0.4698

Significance: 0.003

Upper quantiles of permutations (null model):

```
90% 95% 97.5% 99%
```

```
0.206 0.274 0.333 0.407
```

Permutation: free

Number of permutations: 999

3번에서 구한 환경 그리고 유전적 거리 행렬의 상관계수와 유의수준 값을 확인해보면

Mantel statistic r: 0.4354

Significance: 0.021

다음과 같이 나왔고, 공간과 유전적 거리 행렬의 상관계수와 유의수준 값을 확인해보면

Mantel statistic r: 0.4698

Significance: 0.003

다음과 같다. 상관계수의 값은 더 크고, 유의수준은 더 작으므로 공간의 거리행렬과 유전적 거리 행렬과의 연관성이 환경의 거리행렬과 유전적 거리 행렬과의 연관성보다 더 크다고 볼 수 있다. 따라서 콜로니들의 유전적 구성에 영향을 미친 환경적 요인의 결과라기보다 근접한 콜로니들의 공간적 상관관계의 결과임을 알 수 있다.