HW5 stat461

Ronald Barry

2/11/2022

Problem One

Here is some data, where the variables are divided into a 'y' group and an 'x' group.

```
data_mat <-
structure(c(4, 5, 9, 10, 6, 4, 13, 4, 16, 14, 9, 2, 7, 6, 9,
16, 7, 7, 7, 9, 1, 15, 11, 20, 8, 5, 19, 16, 17, 18, 18, 18,
17, 20, 15, 29, 16, 14, 17, 17, 25, 17, 1, 12, 33, 26, 16, 3,
25, 7, 2, 30, 16, 24, 6, 3, 23, 24, 24, 15, 2.02, 3.54, 2.27,
4.97, 4.2, 3.15, 4.07, 3.3, 4.2, 4.55, 5.3, 3.86, 4.44, 5.65,
3.23, 6.4, 3.61, 4.63, 4.34, 3.58, 2.13, 5.1, 4.43, 5, 2.46,
2.77, 3.67, 5.67, 5.26, 5.03, 3.25, 5.21, 2.85, 3.87, 4.57, 4.79,
4.17, 4.91, 3.5, 3.42, 3.86, 3.45, 4.73, 2.77, 2.76, 2.67, 4.09,
3.46, 4.03, 3.23, 2.64, 5.92, 4.54, 3.55, 3.43, 4.2, 3.88, 3.59,
3.48, 5.66, 2.48, 6.57, 7.52, 11.47, 6.26, 4.83, 7.6, 9.54, 11.35,
9.96, 10, 7.4, 5.58, 8.91, 8.64, 10.77, 6.04, 9.28, 8.47, 5.72
), .Dim = c(20L, 7L), .Dimnames = list(NULL, c("y1", "y2", "y3",
"x1", "x2", "x3", "x4")))
#
data_mat
```

```
##       y1 y2 y3   x1   x2   x3    x4
##  [1,]  4  1 25 2.02 2.13 3.86  2.48
##  [2,]  5 15 17 3.54 5.10 3.45  6.57
##  [3,]  9 11  1 2.27 4.43 4.73  7.52
##  [4,] 10 20 12 4.97 5.00 2.77 11.47
##  [5,]  6  8 33 4.20 2.46 2.76  6.26
##  [6,]  4  5 26 3.15 2.77 2.67  4.83
##  [7,] 13 19 16 4.07 3.67 4.09  7.60
##  [8,]  4 16  3 3.30 5.67 3.46  9.54
##  [9,] 16 17 25 4.20 5.26 4.03 11.35
## [10,] 14 18  7 4.55 5.03 3.23  9.96
## [11,]  9 18  2 5.30 3.25 2.64 10.00
## [12,]  2 18 30 3.86 5.21 5.92  7.40
## [13,]  7 17 16 4.44 2.85 4.54  5.58
## [14,]  6 20 24 5.65 3.87 3.55  8.91
## [15,]  9 15  6 3.23 4.57 3.43  8.64
## [16,] 16 29  3 6.40 4.79 4.20 10.77
## [17,]  7 16 23 3.61 4.17 3.88  6.04
## [18,]  7 14 24 4.63 4.91 3.59  9.28
## [19,]  7 17 24 4.34 3.50 3.48  8.47
## [20,]  9 17 15 3.58 3.42 5.66  5.72
```

a. Using R, run the Mantel test to see if the distance structure in y1, y2, y3 (count data) is the same as the distance structure for the x1, x2, x3, x4 data (quantitative).

b. **What are your conclusions? What does the Mantel test actually test? What is the null hypothesis?**

(You can choose any reasonable distance measure, but note that the 'y' variables are counts and the 'x' variables are quantitative.)

Problem Two

```
D <-
structure(list(c(0, 0.089, 0.104, 0.161, 0.182, 0.232, 0.233,
0.249, 0.256, 0.273, 0.322, 0.308), c(0.089, 0, 0.106, 0.171,
0.189, 0.243, 0.251, 0.268, 0.249, 0.284, 0.321, 0.309), c(0.104,
0.106, 0, 0.166, 0.189, 0.237, 0.235, 0.262, 0.244, 0.271, 0.314,
0.293), c(0.161, 0.171, 0.166, 0, 0.188, 0.244, 0.247, 0.262,
0.241, 0.284, 0.303, 0.293), c(0.182, 0.189, 0.189, 0.188, 0,
0.247, 0.239, 0.257, 0.242, 0.269, 0.309, 0.296), c(0.232, 0.243,
0.237, 0.244, 0.247, 0, 0.036, 0.084, 0.124, 0.289, 0.314, 0.282
), c(0.233, 0.251, 0.235, 0.247, 0.239, 0.036, 0, 0.093, 0.12,
0.293, 0.316, 0.289), c(0.249, 0.268, 0.262, 0.262, 0.257, 0.084,
0.093, 0, 0.123, 0.287, 0.311, 0.298), c(0.256, 0.249, 0.244,
0.241, 0.242, 0.124, 0.12, 0.123, 0, 0.287, 0.319, 0.287), c(0.273,
0.284, 0.271, 0.284, 0.269, 0.289, 0.293, 0.287, 0.287, 0, 0.32,
0.285), c(0.322, 0.321, 0.314, 0.303, 0.309, 0.314, 0.316, 0.311,
0.319, 0.32, 0, 0.252), c(0.308, 0.309, 0.293, 0.293, 0.296,
0.282, 0.289, 0.298, 0.287, 0.285, 0.252, 0)), class = "data.frame", row.names = c("HomoSapiens",
"Pan", "Gorilla", "Pongo", "Hylobates", "MacacaFuscata", "MacacaMulatta",
"MacacaFascicular", "MacacaSylvanus", "SaimiriSciureus", "TarsiusSyrichta",
"LemurCatta"))
#
D
```

```
##
## HomoSapiens      0.000 0.089 0.104 0.161 0.182 0.232 0.233 0.249 0.256 0.273
## Pan              0.089 0.000 0.106 0.171 0.189 0.243 0.251 0.268 0.249 0.284
## Gorilla          0.104 0.106 0.000 0.166 0.189 0.237 0.235 0.262 0.244 0.271
## Pongo            0.161 0.171 0.166 0.000 0.188 0.244 0.247 0.262 0.241 0.284
## Hylobates        0.182 0.189 0.189 0.188 0.000 0.247 0.239 0.257 0.242 0.269
## MacacaFuscata    0.232 0.243 0.237 0.244 0.247 0.000 0.036 0.084 0.124 0.289
## MacacaMulatta    0.233 0.251 0.235 0.247 0.239 0.036 0.000 0.093 0.120 0.293
## MacacaFascicular 0.249 0.268 0.262 0.262 0.257 0.084 0.093 0.000 0.123 0.287
## MacacaSylvanus   0.256 0.249 0.244 0.241 0.242 0.124 0.120 0.123 0.000 0.287
## SaimiriSciureus  0.273 0.284 0.271 0.284 0.269 0.289 0.293 0.287 0.287 0.000
## TarsiusSyrichta  0.322 0.321 0.314 0.303 0.309 0.314 0.316 0.311 0.319 0.320
## LemurCatta       0.308 0.309 0.293 0.293 0.296 0.282 0.289 0.298 0.287 0.285
##
## HomoSapiens      0.322 0.308
## Pan              0.321 0.309
## Gorilla          0.314 0.293
## Pongo            0.303 0.293
## Hylobates        0.309 0.296
## MacacaFuscata    0.314 0.282
## MacacaMulatta    0.316 0.289
## MacacaFascicular 0.311 0.298
## MacacaSylvanus   0.319 0.287
## SaimiriSciureus  0.320 0.285
## TarsiusSyrichta  0.000 0.252
## LemurCatta       0.252 0.000
```

This is a phylogenetic distance matrix [from Hayasaka, K., T. Gojobori, and S. Horai. (1988) Molecular phylogeny and evolution of primate mitochondrial DNA. Molecular Biology and Evolution 5, 626-644]. Use multidimensional scaling (metric, that is, cmdscale) to make a map of these species. What does it seem to indicate?

Problem Three

The R base function **princomp()** does principal component analysis. Run the following.

```
X <- data_mat[,4:7]
summary(X)
pairs(X)
#
#
#
tmp <- princomp(X)
plot(tmp)
tmp
summary(tmp)
#
tmp2 <- princomp(X, cor=TRUE)
plot(tmp2)
tmp2
summary(tmp2)
```

a. Does using **cor=TRUE** make a difference? In general, when should you NOT use **cor=TRUE**? When should you use **cor=TRUE**? Try multiplying column x1 by 100 and run princomp both ways. Does this support your thoughts about correlation vs covariance matrices in principal components?

b. Looking at the variances (for instance, when **cor=TRUE**), how many principal components do you think you should use? Why? Does it look like principal components was a good approach to follow in this case? Why or why not?

c. What does the following set of loadings and scores tell you (in general, what ARE loadings and scores and what are they used for)?

```
X <- data_mat[,4:7]
hold = princomp(X, cor=TRUE, scores=TRUE)
names(hold)
hold$scores
hold$loadings
```

d. Repeat a (correlation-based) principal components analysis on the following. What conclusions do you draw? Do the variances of the principal components add up to 4 (which is the number of variables)?

```
M <-
structure(c(-1.4, -1.09, -0.19, -0.09, 1.04, -0.5, -0.22, -1.65,
-1.02, 0.71, 0.23, -0.48, -1.4, -1.2, -0.5, 0.19, 0.9, -0.59,
-0.23, -1.9, -0.73, 0.84, 0.21, -0.46, -1.62, -1.02, 0.03, 0.18,
1.11, -0.62, -0.12, -1.45, -1.13, 0.93, 0.4, -0.52, -1.52, -1.11,
-0.24, -0.9, 0.99, -0.38, -0.04, -1.54, -1.15, 0.37, 0.48, -0.35
), .Dim = c(12L, 4L), .Dimnames = list(NULL, c("X1", "X2", "X3",
"X4")))
#
#
M
```

```
##          X1    X2    X3    X4
##  [1,] -1.40 -1.40 -1.62 -1.52
##  [2,] -1.09 -1.20 -1.02 -1.11
##  [3,] -0.19 -0.50  0.03 -0.24
##  [4,] -0.09  0.19  0.18 -0.90
##  [5,]  1.04  0.90  1.11  0.99
##  [6,] -0.50 -0.59 -0.62 -0.38
##  [7,] -0.22 -0.23 -0.12 -0.04
##  [8,] -1.65 -1.90 -1.45 -1.54
##  [9,] -1.02 -0.73 -1.13 -1.15
## [10,]  0.71  0.84  0.93  0.37
## [11,]  0.23  0.21  0.40  0.48
## [12,] -0.48 -0.46 -0.52 -0.35
```

Problem Four

Find any multivariate data set of interest to you and perform PCA. The data should have at least three columns. Did PCA help to reduce dimensionality? Looking at the loadings of PC1, what does the PC seem to represent? Try to find a dataset that (as far as you can tell) has not already had PCA performed on it online.