Additional Multivariate Exercise 14

Brief solution, for both datasets

Prehistoric dogs data (expanded): We analyzed these data by MANOVA in the previous exercise, and the interest here is in comparing those results to what we can obtain from `ANOSIM` and `PERMANOVA`. We start by showing the `ANOSIM` results for comparing all species without accounting for sex. The analysis uses Euclidean distance for the standardized variables (very similar results were obtained for Manhattan (city-block, or L_1) distance).

```
Call:
anosim(x = prehdog2std, grouping = prehdog2$species, distance = "euclidean")
Dissimilarity: euclidean

ANOSIM statistic R: 0.7815
      Significance: 0.001

Permutation: free
Number of permutations: 999

Upper quantiles of permutations (null model):
   90%    95%  97.5%    99%
0.0362 0.0488 0.0566 0.0749

Dissimilarity ranks between and within classes:
                0%     25%    50%     75% 100%    N
Between         22 1078.50 1726.5 2338.25 2926 2344
Cuon             2  183.25  345.5  649.50 1518  136
Golden jackal    1   81.25  177.5  335.25 1217  190
Indian wolf     39  670.00 1081.0 1418.00 2403   91
Modern dog      24  268.50  465.0  836.50 2247  120
Prehistoric dog 31  375.00  849.0 1122.00 1860   45
```

The rank permutation test also shows strong significance against the hypothesis of no difference between the species. Note that the P-value is the most extreme possible with a total of 1000 values ($N = 999$ plus the data itself). The table of dissimilarity ranks shows how the ranks ($77 \cdot (77-1)/2 = 2926$ in total) are distributed as within-group ranks for each of the groups and the between-group ranks. It is seen that the between-group ranks are much higher for all quartiles than the within-group ranks, as one would expect from a significant test. Among the species, the Indian wolf has the largest median rank, whereas the golden jackal has the lowest median rank. The same pattern can for example be seen in the within-group sum of squared distances (for the standardized data) that the K-means clustering method is based on, or for each of the variables $x_1 - x_9$ in the within-group standard deviations. Because the above listing only shows details for within-group ranks, it cannot be used to infer which of the groups might be closer to each other than others. But the `ANOSIM` analysis can be repeated for pairwise comparisons, if desired; this would be the analog of Hotelling's T^2-tests between groups.

We next try the `PERMANOVA` method for the one-way ANOVA layout, again with Euclidean distances for standardized variables.

```
Call:
adonis(formula = prehdog2std ~ as.factor(species), data = prehdog2, method = "euclidean")
```

```
Permutation: free
Number of permutations: 999
Terms added sequentially (first to last)

                   Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
as.factor(species)  4    526.73 131.682  60.285 0.77007  0.001 ***
Residuals          72    157.27   2.184         0.22993
Total              76    684.00                 1.00000
```

The sum of squared distances, both total and within groups, can be computed directly by adding up the usual sum of squares values for 9 (standardized) variables. Note how the "Df" values are the same as in any one-way ANOVA for a single outcome; this is per construction. The "R2" values are simply the ratios out of the total. It is the same formula as for an ordinary ANOVA, but it is not obvious that the values mean anything else than implied by the formula. The significance is unsurprising at this point. The output gives no information about how the groups differ. Results for the Manhattan distance look similar, but cannot be compared directly because these distances are on a different scale; the significance is still the lowest possible.

We next try the two-way factorial including sex, for the subdataset without the prehistoric dogs. The `PERMANOVA` table can be computed with either sequential or partial (marginal) sums of squared distances. As is the usual practice for ANOVA tables the sequential terms are added from the top down, or removed from the bottom up.

```
Call:
adonis(formula = prehdog2sexstd ~ as.factor(species) * as.factor(sex), data = prehdog2sex,
       method = "euclidean")

Permutation: free
Number of permutations: 999
Terms added sequentially (first to last)

                                  Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
as.factor(species)                 3    472.62 157.540  96.398 0.79566  0.001 ***
as.factor(sex)                     1     16.03  16.026   9.806 0.02698  0.001 ***
as.factor(species):as.factor(sex)  3      8.93   2.978   1.822 0.01504  0.115
Residuals                         59     96.42   1.634         0.16233
Total                             66    594.00                 1.00000
```

As for the MANOVA (of the previous exercise), the tests for both species and sex are strongly significant. One could try with reverse order of the two terms or partial tests to confirm the significance of both terms in an additive model, if desired. For Euclidean distances, the assessment of the interaction term is remarkably similar to the one from MANOVA (where 3 out of 4 tests gave P-values around 0.09). For the Manhattan distance, the interaction term shows a substantially higher P-value of 0.3; the interpretation of this difference in significance is not obvious. In summary, the simulation-based ANOVA methods seems to give quite similar assessments of significance as the MANOVA for these data.

Steneryd data (for plots): The dataset does not contain any grouping variables, but we will explore effects of the four environmental variables (light, moisture, reaction, nitrogen). We previously noted that the variables are very strongly correlated, with positive correlations among the last three variables and negative correlations of those with light. With quantitative predictors ("covariates"), only `PERMANOVA` is an option for analysis. To match the analysis with MANOVA, we should again use Euclidean distances for standardized variables. Because the MANOVA tests are all based on the full model, the closest analog in the `PERMANOVA` setup is with partial tests (using the `adonis2` function).

```
Permutation test for adonis under reduced model
Marginal effects of terms
Permutation: free
Number of permutations: 999

adonis2(formula = sten.specstd ~ light + moisture + reaction + nitrogen, data = steneryd,
        method = "euclidean", by = "margin")

          Df SumOfSqs      R2      F Pr(>F)
light      1    48.84 0.12210 3.4847  0.004 **
moisture   1    55.97 0.13993 3.9935  0.001 ***
reaction   1    36.13 0.09033 2.5778  0.015 *
nitrogen   1    42.40 0.10599 3.0248  0.010 **
Residual  12   168.19 0.42048
Total     16   400.00 1.00000
```

The permutation tests give all predictors as significant; note that for partial sum of squares, the "R2"-values will no longer add up to 1 (just as with partial tests in ordinary ANOVA tables). These results contrast the tests by MANOVA with its P-values of (in the same order): 0.007, 0.031, 0.046, and 0.147. For the MANOVA, the dimension of the multivariate outcome needs to be reduced, because one cannot have more variables than observations. The above P-values were obtained after selecting the first 7 components of a PCA (see do-file for details). Considering the strong dependence between the four predictors, the significance for all of them seems a bit surprising. The estimated slopes for the 4 predictors by the MANOVA are not all particularly intuitive across the variables. Without any formal test to back this up, my sense would be that one should reduce the number of predictors for a dataset consisting of only 17 plots (and with so strongly correlated predictors; a PCA for the predictors shows that one component explains 93% of the variation/correlation).

We finally also explore the results from a totally different dissimilarity measure, namely the default Bray-Curtis index (applied to the non-standardized data). The results are now no longer comparable to MANOVA and not so easy to interpret (because the method only gives P-values, no effects estimates, in terms of regression coefficients or similar).

```
Permutation test for adonis under reduced model
Marginal effects of terms
Permutation: free
Number of permutations: 999

adonis2(formula = sten.spec ~ light + moisture + reaction + nitrogen, data = steneryd,
        by = "margin")

          Df SumOfSqs      R2      F Pr(>F)
light      1   0.3949 0.10465 3.9736  0.016 *
moisture   1   0.4327 0.11467 4.3540  0.005 **
reaction   1   0.1641 0.04349 1.6514  0.168
nitrogen   1   0.1754 0.04649 1.7652  0.137
Residual  12   1.1924 0.31604
Total     16   3.7731 1.00000
```

The scale for the dissimilarities is obviously different. Only the first two environmental variables show significance by the permutation test. These were indeed the variables with the strongest significance for Euclidean distances. It is not obvious from looking at the P-values alone which analysis to prefer.