UCLA Department of Statistics
STATS M254 / BIOINFO M271 Homework 3

Instructor: Jingyi Jessica Li

Due date: Wednesday, Mar 16, 2022 at 11:59 pm on CCLE

Please staple your homework and write down your name and UID clearly.

Problem 1: Dimension Reduction Methods (40 pts)

In this problem, please perform principal component analysis (PCA), multidimensional scaling (MDS), and t-SNE on the `USArrests` data set, which is part of the base `R` package. The rows of the data set contain the fifty states, in alphabetical order. The columns of the data set contain the four variables.

```
> row.names(USArrests)
> names(USArrests)
```

1. Perform PCA using the `prcomp()` function to find principal components as linear combinations of the four columns, with each column centered at 0 and scaled to have standard deviation 1. Print the loadings of principal components 1-4.

2. Obtain a summary of the proportion of variance explained (PVE) of the principal components using the `summary()` function on the `prcomp` object.

3. Use the `biplot()` function on the `prcomp` object to plot the scores of the 50 states in the first and second principal components along with the loadings of these two principal components shown as arrows in one plot. Explain the plot. What messages do you learn from this plot?

4. Calculate the pairwise Euclidean distance between the 50 states. Use the R function `cmdscale()` to perform MDS and visualize the result. Compare the result with the PCA result you visualize in Problem 5.3.

5. Apply t-SNE and UMAP to this data set using the R package `M3C`. Visualize the result, and compare it with the PCA and MDS results.

Problem 2: Compare Measures of Association (40 pts)

Please use the following simulation study to compare different measures of association we covered in the class.

Assuming that X and Y are $N(0, 1)$ variables, we would like to compare the power of different measures of association as test statistics for testing the following hypotheses.

Null hypothesis H_0: X and Y are independent;

vs.

Alternative hypothesis H_1: $Y = X + \epsilon$;

or

Alternative hypothesis H_2: $Y = X^2 + \epsilon$;

or

Alternative hypothesis H_3: $Y = \sin(X) + \epsilon$;

or

Alternative hypothesis H_4: $Y = \text{sign}(X) + \epsilon$, where $\text{sign}(x) = 1$ if $x > 0$, $\text{sign}(x) = 0$ if $x = 0$ and $\text{sign}(x) = -1$ if $x < 0$;

or

Alternative hypothesis H_5: $Y = (-1)^Z X + \epsilon$, where $Z \sim$ Bernoulli(0.5) and is independent of X and ϵ.

Suppose the error term $\epsilon \sim N(0, \sigma^2)$.

Calculate the power of measures of association for (1) H_0 vs. H_1, (2) H_0 vs. H_2, (3) H_0 vs. H_3, (4) H_0 vs. H_4, and (5) H_0 vs. H_5, using the following approach:

For a fixed σ^2,

1. For $B = 1,000$ times, simulate a sample of size $n = 100$ pairs of (x_i, y_i), $i = 1, \ldots, n$, under H_0. This gives B null data sets each with n data points.

2. For $B = 1,000$ times, simulate a sample of size $n = 100$ pairs of (x_i, y_i), $i = 1, \ldots, n$, under the alternative hypothesis. This gives B alternative data sets each with n data points.

3. Apply the measure of association (e.g., Pearson correlation) to each of the B null data sets, resulting in B null values.

4. Apply the measure of association to each of the B alternative data sets, resulting in B alternative values.

5. At the significance level $\alpha = 0.05$, find the 95% percentile of the null values, using this percentile as the threshold.

6. Apply the threshold to the alternative values, calculate the percentage of the alternative values that are equal or above the threshold. Record this percentage as the power estimate.

Using the above approach, you will have one power estimate per σ^2 per measure. Vary σ^2 using values $\{.1, .3, .5, .7, .9, 1.1, 1.3, 1.5\}$ and compare the following measures:

- Pearson correlation (function `cor()`)
- Spearman rank correlation (function `cor()`)
- Kendell's tau (function `cor()`)
- maximal correlation (package `acepack`)
- distance correlation (package `energy`)
- maximal information coefficient (package `minerva`)
- Chatterjee's correlation (package `XICOR`)

- generalized Pearson correlation squares (package `gR2` available at `https://github.com/lijy03/gR2`)

For each of tests (1)-(4), please summarize the power estimates using a line plot, whose horizontal axis is the σ^2 value (i.e, noise level) and vertical axis is the power estimate. This will give you a total of four plots. Use different colors for different measures. Label your plots clearly.

Problem 3: Galton's Peas (20 pts)

Following Section 3 of Chatterjee's paper (`https://doi.org/10.1080/01621459.2020.1758115`) and implement the eight correlations in Problem 2 on the `peas` dataset in `R` package `psych`. Can you reproduce the results in Chatterjee's paper? What conclusions can you draw from the other seven measures?