COS 424
Homework #4
Due Tuesday, April 13th

See the course website for important information about collaboration and late policies, as well as where and when to turn in assignments.

FactoMineR

The R package `FactoMineR` implements a number of exploratory methods, including Principal Component Analysis (PCA) and Correspondence Analysis (CA). You need a fairly recent version of R (it works with version 2.9.) See `http://factominer.free.fr` for more details.

You can quickly install `FactoMineR` with the command

```
> install.packages('FactoMineR')
```

and you can load the package with

```
> library(FactoMineR)
```

Question 1: PCA

This question uses the UCI "Wine" dataset available from `http://archive.ics.uci.edu/ml/datasets/Wine`. For your convenience, a local copy of the dataset is available from the homework web page. This dataset describes the results of a chemical analysis of wines originating from a same region of Italy but grown by three different producers.

- Edit the file `wine.data` by inserting a first row containing the column name described in file `wine.names`. Note that the first column is the producer number. It is convenient to change these numbers into producer names such as "Producer1" to help R identify this column as a categorical variable (factors in R speak). Save the file as `wine.csv`.
- Load the dataset into R using function `read.csv`. Then perform the PCA using the class as a supplementary categorical (qualitative) variable. Use `help(PCA)` to find out how to use the `PCA` function. This function usually plots both the row PCA and the column PCA. Then use function `plot.PCA` to replot the row PCA using different color for each producer.
- Provide a printout of both the row and column PCA plots. Comment the plots and explain what we can see.

Question 2: CA

This question uses the UCI "Plants" dataset available from `http://archive.ics.uci.edu/ml/datasets/Plants`. For your convenience, a local copy of the dataset is available from the homework web page.

Each line of the file `plants.data` contains a plant name and a comma separated list of abbreviations representing the places where the plant grows. The file `stateabbr.txt` lists abbreviations for the US states, the Canadian provinces, and a couple islands. Note that these abbreviations are not always the standard abbreviations...

Although the dataset documents nearly 35,000 species, the first word of the scientific name of each species describes its genus, that is, a broader category.

- Construct a contingency table whose rows are the genera or genuses and whose columns are the contiguous US states. Each entry in the table counts the number of species of that genus growing in that state. Make sure that each row and each column contains at least a non zero entry, and save the result into a comma separated file with the state abbreviations listed in the first row.
- Load this into R using `read.csv`, making sure that the row names and column names are properly identified. Then perform the correspondence analysis using function `CA`. Since plotting a thousand genus names does not make the graph easy to read, use the function `plot.CA` to replot the graph showing only the state abbreviations.
- Provide a printout of the resulting plot. Comment on the relative position of the states. The function `plot.CA` forces an aspect ratio of 1:1. You could use `fix(plot.CA)` to modify it and produce a more informative plot.
- The state `ab` (the abbreviation for Alabama) is placed very strangely. Could there be a confusion with Alberta? Correct the data and print the new plot.