Final HW

For this homework, you will use a dataset used to try to understand Parkinson's disease. The dataset includes 31 people, 23 of whom have Parkinson's. However, there are multiple observations per patient, which means that the observations are not independent. In the first column, a value like `phon_R01_S13_3` indicates that this is the third observation from subject 13. A more thorough description of the data is given at the end of the homework. Note that except for the first column, which gives the subject, and the column for `status`, all other variables are quantitative. For the `status` variable, a 1 indicates that the subject has Parkinson's disease, and a 0 indicates otherwise.

To read in the data, you can download it from the class webpage as `parkinsons.csv`. Note that R will modify some of the variable names, for example changing underscores to periods. Just let R do this automatically.

1. For problem 1, you'll use principal components to analyze the data. Create a data set that has only quantitative variables. Run principal components on the data.
(a) Give a table of the variables and the first three principal components.

(b) What proportion of the variance is explained by the first two principal components? The first three principal components?

(c) How many principal components is it reasonable to retain for this data? Justify your answer.

(d) Plot the data using the first two principal components. Describe the plot.

(e) Plot the data again using the first two principal components, but this time label the points based on whether the patient had Parkinson's or not. Use shape of the plotting character (e.g., square versus circle) to indicate Parkinson's status, not just color. If you use color, choose colors that are distinguishable when printed in black and white (not red and green). Make sure plots are well labeled and easy to read, and include a legend for Parkinson's status. Describe what you see.

(f) Also make a plots of Principal compoents 1 versus 3 and 2 versus 3 with points labeled by Parkison's status. Describe what you see.

2. For this part, you'll do cluster analysis. Consider the original data (not the principal components). You will cluster the observations using only the variables `MDVP.Fo.Hz.`, `MDVP.Fhi.Hz.`, and `MDVP.Flo.Hz.`.
(a) Make dendrograms of your clustering using average linkage, single linkage, and complete linkage. Describe similarities or differences between the different clustering methods.

(b). Comment on the dendrograms. Do observations within the same individual tend to cluster together or not?

(b) Do cases with Parkinson's tend to cluster together or not?

3. For this problem, you'll do a little bit of clustering by hand instead of using data. Given the following dissimilarity matrix, which is 5×5, (a) use single linkage clustering to determine C_1 and C_2 and C_3 Write out the 4×4, 3×3,and 2×2 matrices that result from these mergers. (b) Draw the resulting tree diagram with branch lengths (edge weights).

	A	B	C	D	E
A	0	5	3	3	2
B	5	0	4	2	1
C	3	4	0	2.5	4
D	3	2	2.5	0	4
E	2	1	4	4	0

Title: Parkinsons Disease Data Set

Abstract: Oxford Parkinson's Disease Detection Dataset

Data Set Characteristics: Multivariate
Number of Instances: 197
Area: Life
Attribute Characteristics: Real
Number of Attributes: 23
Date Donated: 2008-06-26
Associated Tasks: Classification
Missing Values? N/A

Source:

The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders.

Data Set Information:

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient, the name of the patient is identified in the first column.For further information or to pass on comments, please contact Max Little (littlem '@' robots.ox.ac.uk).

Further details are contained in the following reference -- if you use this dataset, please cite:
Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering (to appear).

Attribute Information:

Matrix column entries (attributes):
name - ASCII subject name and recording number
MDVP:Fo(Hz) - Average vocal fundamental frequency

MDVP:Fhi(Hz) - Maximum vocal fundamental frequency
MDVP:Flo(Hz) - Minimum vocal fundamental frequency
MDVP:Jitter(\%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP - Several
measures of variation in fundamental frequency
MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA - Several measures of varia
NHR,HNR - Two measures of ratio of noise to tonal components in the voice
status - Health status of the subject (one) - Parkinson's, (zero) - healthy
RPDE,D2 - Two nonlinear dynamical complexity measures
DFA - Signal fractal scaling exponent
spread1,spread2,PPE - Three nonlinear measures of fundamental frequency variation

---

Citation Request:

If you use this dataset, please cite the following paper:
'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection',
Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM.
BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)