ISE 5103 Intelligent Data Analytics
Homework #3

Instructor: Charles Nicholson

See course website for due date

Learning objective: Principal Component Analysis.

Submission notes:

1. Team assignment! Include all team member names on the submitted work.
2. You will submit a PDF file with your solutions. Additionally, you will provide the R code you created to address the problems. The PDF is primarily what will be graded. The grader *may* view your R code, but should never *have* to in order to find your solutions.
3. In the PDF, clearly identify each problem (e.g. Problem 1a, Problem 2b, etc.) Also, note that only *relevant* and informative computer output should be provided.
4. Make sure to *provide comments* on what your R code is doing. Keep it clean and clear!
5. You will submit your complete R script. Note: include `library` commands to load *all* packages that are used in the completion of the assignment. Place these statements at the top of your script.
6. Do not zip your files for submission. Submit exactly two files. Name the files "LastName-HW1" with the appropriate file extension (that is, .pdf for the write-up and .R for the script)

1 Glass data

The study of classification of types of glass is motivated by criminological investigations. At the scene of a crime, the glass left can be used as evidence... if it is correctly identified.

The data set we consider consists of 213 unique glass samples labeled as one of six class categories[1]:

type	description
1	building windows float processed
2	building windows non-float processed
3	vehicle windows float processed
5	containers
6	tableware
7	headlamps

There are nine predictors, including the refractive index and percentages of the following eight elements found in the glass: Na (Sodium), Mg (Magnesium), Al (Aluminum), Si (Silicon), K (Potassium), Ca (Calcium), Ba (Barium), and Fe (Iron).

[1] I do not know why they skipped class "4" in the data.

The data is available here: `http://archive.ics.uci.edu/ml/datasets/Glass+Identification` and is also available in the `mlbench` package as the dataset `Glass`.

Note: There is one duplicate row in the mlbench data. Please find the duplicate row and remove it. See the R function `duplicated` to help you find it.

(a) (15 points) Mathematics of PCA

i. Create the correlation matrix of all the numerical attributes in the `Glass` data and store the results in a new object `corMat`.

ii. Compute the eigenvalues and eigenvectors of `corMat`.

iii. Use `prcomp` to compute the principal components of the `Glass` attributes (make sure to use the `scale` option).

iv. Compare the results from (ii) and (iii) – Are they the same? Different? Why?

v. Using R demonstrate that principal components 1 and 2 from (iii) are orthogonal. (Hint: the inner product between two vectors is useful in determining the angle between the two vectors)

(b) (15 points) Application of PCA

i. Create a visualization of the `corMat` correlation matrix (i.e., a heatmap or variant) If you are interested and have time, consider the `corrplot` package for very nice options, `https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html`.

ii. Provide visualizations of the principal component analysis results from the `Glass` data. Consider incorporating the glass type to group and color your biplot.

iii. Provide an interpretation of the first two prinicpal components the `Glass` data.

iv. Based on the PCA results, do you believe that you can effectively reduce the dimension of the data? If so, to what degree? If not, why?

(c) (15 points) Application of LDA

i. Since the `Glass` data is grouped into various labeled glass types we can consider linear discriminant analysis (LDA) as another form of dimension reduction. Use the `lda` method from the `MASS` package to reduce the `Glass` data dimensionality.

ii. How would you interpret the first discriminant function, LD1?

iii. Use the `ldahist` function from the `MASS` package to visualize the results for LD1 and LD2. Comment on the results.

2 Principal components for dimension reduction

The `HSAUR2` package contains the data `heptathlon` which are the results of the women's olympic heptathlon competition in Seoul, Korea from 1988. A scoring system is used to assign points to the results from each of the seven events and the winner is the woman who accumulates the most points over the two days.

(a) (10 points) Examine the event results using the Grubb's test. According to this test there is one competitor who is an outlier multiple events: Who is the competitor? And for which events is there statistical evidence that she is an outlier? Remove her from the data.

(b) (5 points) As is, some event results are "good" if the values are large (e.g. highjump), but some are "bad" if the value is large (e.g. time to run the 200 meter dash). Transform the running events (`hurdles, run200m, run800m`) so that large values are good. An easy way to do this is to subtract values from the max value for the event, i.e. $x_i \leftarrow x_{\max} - x_i$.

(c) (5 points) Perform a principal component analysis on the 7 event results and save the results of the `prcomp` function to a new variable `Hpca`.

(d) (10 points) Use `ggibiplot` to visualize the first two principal components. Provide a concise interpretation of the results.

(e) (10 points) The PCA projections onto principal components $1, 2, 3, \ldots$ for each competitor can now be accessed as `Hpca$x[,1]`, `Hpca$x[,2]`, `Hpca$x[,3]`, ... Plot the heptathlon `score` against the principal component 1 projections. Briefly discuss these results.

3 Housing data dimension reduction and exploration

The `housingData.csv` file in the course website is real data associated with 1,000 residential homes sold in Ames, Iowa between 2006 and 2010. The data set includes over 70 explanatory variables – many of which are factors with several levels. The file `housingVariables.pdf` provides a concise explanation of the variables and the factor levels in the data.

To clean up and transform the data a little bit for this problem, please load the housing data into a data frame (or tibble) named `housingData` and then run the code listed below which performs the following tasks:

1. selects only numeric columns
2. creates new variables `age`, `ageSinceRemodel`, and `ageofGarage`, and
3. removes a few columns that are not needed

```
library(tidyverse)

hd <- housingData %>%
  select_if(is.numeric) %>%
      dplyr::mutate(age = YrSold - YearBuilt,
                    ageSinceRemodel = YrSold - YearRemodAdd,
                    ageofGarage = ifelse(is.na(GarageYrBlt), age, YrSold - GarageYrBlt)) %>%
      dplyr::select(!c(Id,MSSubClass, LotFrontage, GarageYrBlt,
                       MiscVal, YrSold  , MoSold, YearBuilt,
                       YearRemodAdd, MasVnrArea))
```

(15 points) Using this newly created data set `hd`, perform PCA and correlation analysis. Did you find anything worthwhile? Make sure to respond with visualizations and interpretations of at least the most important principal components.

Note: Installation of the `ggbiplot` is slightly more involved than many R packages.
The following steps should help:

```
install.packages("devtools")
library(devtools)
install_github("vqv/ggbiplot")
library(ggbiplot)
```