Principal Components Analysis

STA 325: Assignment 7, Fall 2018

Due Thursday October 25, 10 PM, Sakai

Agenda: Principal components analysis on fmri data

Team Members

Please list all team members here: Team Leader, all other team members. Please also list out each team members task. (See below for more details).

General instructions for homeworks: Change the name of your file to `hw-pca-solutions-group-name.Rmd`.

1. Make sure that you work on all the files in your Datathon group repository. (Team capitans, please make sure that each team member is contributing to each portion of a problem and break these up evenly within your groups.) Each team has 4 team members, so please make sure each team member does at least 2 problems. Suggestion: all do the first problem together. Also, it's fine if you do problems collaboratively (just mark this in your report).
2. When submitting your assignments on Sakai, please submit the url of your github repository and not your zipped code. Please make sure that all of your code in the repository is completely reproducible and that you do not change it after the deadline.
3. Please do not and start on this assignment at the last minute as your could have issues regarding github, so please start early. We will be working with this in class as well to prepare for datathon, so you should have plenty of time to complete the assignment.
4. Optional/Recommended: If you would like to turn in the assignment in the format of the final assignment for the Datathon report and make it more open ended, this is welcomed too, and Ben and I are happy to give you all feedback regarding this.

Advice: Start early on the homeworks and it is advised that you not wait until the day of. While the professor and the TA's check emails, they will be answered in the order they are received and last minute help will not be given unless we happen to be free.

Commenting code Code should be commented. See the Google style guide for questions regarding commenting or how to write code https://google.github.io/styleguide/Rguide.xml. No late homework's will be accepted.

R Markdown Test

0. Open a new R Markdown file; set the output to HTML mode and "Knit". This should produce a web page with the knitting procedure executing your code blocks. You can edit this new file to produce your homework submission.

Information about the assignment Please download the first patient data (data from subject P1) from http://www.cs.cmu.edu/afs/cs/project/theo-73/www/science2008/data.html.

Division of work Please mark the division of work for the homework assignment by each team member and please note their respective github handles. Team leaders, if you had group members that did not participate, please email myself and Ben Feder to let me know. Group members, if you had team members that did not participate, please email me and let me know. (Please be specific in your email.)

The data for this assignment comes from a brain imaging (fMRI) experiment on reading. The participants (9 adults from the Carnegie Mellon community) were shown line drawings and noun labels of 60 concrete objects from 12 semantic categories with 5 exemplars per category. The entire set of 60 stimuli was presented 6 times during the experiment, in a different random order each time. Participants silently viewed the stimuli

and were asked to think of the same item properties consistently across the 6 presentations.[1] For imaging purposes, the participants brains were divided into "volume elements" or "voxels", and functional magnetic resonance imaging (fMRI) measured the brain activity in each voxel during each stimulus presentation. More information on the study can be found on the listed webpage.

1. Using the dataset you have downloaded, let's read the data set into R. Note: make sure that you install the packages `R.matlab` and `scatterplot3d` before trying to compile the file and that you have downloaded the data set!

```
library(R.matlab)
```

```
## R.matlab v3.6.1 (2016-10-19) successfully loaded. See ?R.matlab for help.
```

```
##
## Attaching package: 'R.matlab'
```

```
## The following objects are masked from 'package:base':
##
##     getOption, isOpen
```

```
fmri.p1 <- readMat("data/data-science-P1.mat")
fmri <- do.call(rbind,lapply(fmri.p1$data,unlist))
# grab the columns of the fmri matrix
colnames(fmri) <- 1:dim(fmri)[2]
```

The command `fmri <- do.call(rbind, fmri.p1$data,unlist)` should return a matrix of dimension 360 by 21764, where the columns are voxels in the participants's brain and rows represents trials. Each trial is one exposure to one of the 60 stimuli. (Note in this assignment, $p = 21764 >> n = 360$, so keep this in mind).

Verify that you can read in the data and that you have the proper dimensions. What is the value of `fmri[172,2014]`? (Report to two decimals).

2. Next, we will find the three hundred voxels with the highest average values across trials (i.e. the 300 most active voxels) from fMRI and plot them in 3-dimensional space using the code below. Submit the scatterplot as part of your write up. Also, submit a scatterplot of the 650 most active voxels.

Note: The fmri.p1 object contains "metadata" about the experiment, and the 8th piece of metadata is a matrix giving the coordinates of all the voxels.

There is some code below to get you started with this part of the problem.

```
library(scatterplot3d)
col2coord <- fmri.p1$meta[[8]]
```

3. Perform PCA on the data. Report the fraction of variance captured by the first principal component.
4. Explain why there are only 359 principal components, even though there are 21764 variables recorded.
5. Explain why a biplot would be a very bad idea here. (Refer to ISLR, Chapter 10 for the definition of a biplot).
6. Make a 3-dimensional scatterplot of the 300 voxels with the 300 most extreme loadings on the first principal component, whether the loadings are positive or negative.
7. Does your scatterplot from 6 match your scatterplot from 2? Briefly explain whether they should or should not match.
8. Make a scree plot and based on what you find, report the optimal number of principal components (if possible; if not, explain why).

[1] Each stimulus was presented for 3s, followed by a 7s rest period, during which the participants were instructed to fixate on an X displayed in the center of the screen. There were two additional presentations of the fixation, 31s each, at the beginning and at the end of each session, to provide a baseline measure of activity.

9. Plot the first principal component with color indicating the loadings *for every voxel*. Hint: Use the `cut()` function and take the first PC and convert the colors into 10 bins. Explain what your 3D object looks like and be sure to attach the plot.