CSE40625/60625, Spring 2021

Mini-Project #1

Due 11:45 AM on Wednesday, 02/25.

Instructions

You can discuss the problem with other students, but you must complete the work individually.

Submit the work on Sakai with your code and writeup zipped into a single file, named as netID_1.zip.

Make sure your code is properly documented.

Use 12pt or higher font for your writeup.

Make sure the plots you submit are easy to read at a normal zoom level.

Reminder: No late assignments will be accepted unless alternative arrangement has been made before the deadline with the Tas or the instructor.

Problem: Principal Component Analysis (PCA) of Genomes

Goal: In this mini-project, you will run PCA on a real data set, and interpret the output.

Description: Read the paper on PCA first before you start this assignment. Download the p1dataset2021.txt attached with this assignment. The data represented there is from the 1000 genomes project. Each of the 995 lines in the file represents an individual. The first three columns represent respectively the individual's unique identifier, his/her sex (1=male, 2=female) and the population he or she belongs to[1]. The subsequent 10101 columns of each line are a subsample of nucleobases from the individual's genome.

We will be looking at the output of PCA on this dataset. PCA can refer to a number of related things, so to be explicit, in this section when we say "PCA" we mean

The data should be centered (i.e., the sample mean subtracted out) but not normalized.

The output should be the normalized principal components (i.e., unit-length eigenvectors).

Feel free to use a library implementation of PCA for the following questions. For python users, we recommend scikit learn's implementation. Matlab's built-in pca function can also be used. Note that with both python scikit and Matlab, you can specify how many principal components you want (this can save on computation time).

To start this project, first convert the data from the text file of nucleobases to a real-valued matrix (PCA needs a real-valued matrix). Specifically, convert the genetic data into a ***binary*** matrix X such that $X_{i,j}$ = 0 if the i^{th} individual has column j's mode nucleobase[2] for his or her j^{th} nucleobase, and $X_{i,j}$ = 1 otherwise. Note that all mutations appear as a 1, even if they are different mutations, so if the mode for column j is "G", then if individual i has an "A","T", or "C", then $X_{i,j}$ would still be 1.

[1] See http://www.1000genomes.org/faq/which-populations-are-part-your-study/ for decodings

[2] By "mode nucleobase", we just mean the most frequently occurring nucleobase in that position/column (across the 995 data points).

The first 3 columns of the data file provide meta-data, and should be ignored when creating the binary matrix X. We will examine genotypes to extract phenotype information. Answer the following questions:

(a) (10 points) Say we ran PCA on the binary matrix X above. What would be the dimension of the returned vectors?

(b) (20 points) We will examine the first 2 principal components of X. These components contain lots of information about our data set. Create a scatter plot with each of the 995 rows of X projected onto the first two principal components. In other words, the horizontal axis should be V_1, the vertical axis V_2, and each individual should be projected onto the subspace spanned by V_1 and V_2. Your plot must use a different color for each population and include a legend.

(c) (25 points) In two sentences, list 1 or 2 basic facts about the plot created in part (b). Can you interpret the first two principal components? What aspects of the data do the first two principal components capture? Hint: think about history and geography.

(d) (20 points) We will now examine the third principal component of X. Create another scatter plot with each individual projected onto the subspace spanned by the first and third principal components. After plotting, play with different labeling schemes (with labels derived from the meta-data) to explain the clusters that you see. Your plot must include a legend.

(e) (10 points) Something should have popped out at you in the plot above. In one sentence, what information does the third principal component capture?

(f) (20 points) In this part, you will inspect the third principal component. Plot the nucleobase index vs the absolute value of the third principal component. What do you notice? What's a possible explanation? Hint: think about chromosomes.

Bonus questions for 40625; mandatory for 60625 (10 points for each question): These questions suggest ways to explore the dataset in more detail. These are open-ended questions so think hard. Your work here could be aided with this mapping from each unique identifier to more data about the individual

(g) For this problem we simplified our dataset by capturing all deviations from the mode value with an indicator variable. This loses some information relative to the original data set. How would you create a real-valued matrix Y suitable for PCA analysis such that there is a bijection between our input data (minus the first three columns) and Y? The matrix Y should be a useful input to PCA. Explain the reasoning behind your choice of Y. Your answer to this question should not take more than a few sentences.

(h) Perform PCA on the matrix Y from part (g). Recreate the plot from part (b). What added value (if any) does this more complex representation add?

(i) Can you uncover what information is captured in the fourth principal component of X?

(j) The provided dataset represents approximately 0.6% of the dataset found at the top of this directory: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/. What other information can you uncover from this much larger file? For this question we encourage you to look at populations beyond Africa. You will notice that the file is a compressed .vcf le, you may want to look into installing vcftools to work with the file.

What to include in the write-up: One sentence answer for part (a). Scatter plot for part (b). Short discussion for part (c). Scatter plots for parts (d) and (f). One sentence answers for (e) and (f). Optinal for 40625: your answers including any figures/data analysis for parts (g)-(j)