Homework 5: Principal Component Analysis (PCA) of Genomes*

Machine Learning and Data Mining (Fall 2025)

Student Name: Student ID:

Lectured by: Shangsong Liang
Sun Yat-sen University

Your assignment should be submitted to the email that will be provided by the TA
Deadline of your submission is: 23:59PM, November 30, 2025
Do NOT Distributed This Document and the Associated Datasets

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

Description: Download the *p4dataset2024.txt* file from the folder of the assignment. The data represented there is from the *International Genome Sample Resource*[1]. Each of the 995 lines in the file represents an individual. The first three columns represent respectively the individual's unique identifier, their sex (M=male, F=female) and the population they belong to—see the file *p4dataset2024_decoding.txt* for the decodings of these population tags. 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, so its alright if some dimensions have different variance than other dimensions.
- The output should be the normalized principal components (i.e., unit-length vectors).

Feel free to use a library implementation of PCA for the following questions. For python users, we recommend *scikit learn's implementation*[2]. 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).

Exercises: 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[3] for their 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 be 1.

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.

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

(b) We will examine the first 2 principal components of X. 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 there should be 995 dots, one for each individual, with each individual'

*This assignment is taken from https://web.stanford.edu/class/cs168/p4.pdf.
[1]https://www.internationalgenome.org/
[2]http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.
[3]By "mode nucleobase", we just mean the most frequently occurring nucleobase in that position (across the 995 genomes).

s x-coordinate their genome projected onto v_1 and y coordinate the projection onto $v2$. Your plot must use a different color for each population and include a legend.

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

(f) In this part, you will inspect the third principal component. Plot the nucleobase index vs the absolute value of the corresponding value of the third principal component in that index. (The x-axis of your plot should go from 1 to 10101—you' re literally just plotting the 10101 values in the third principal component.) What do you notice? What' s a possible explanation? Hint: think about chromosomes (and if you don' t know much biology, feel free to look through the wikipedia page on chromosomes....)

(g) How much of the variance in the data is explained by the first principal component? Namely, on average over the 995 genomes, by what percentage does the squared length of each genome decrease when projected onto the top principal component? (A zero percent decrease would correspond to all the genomes lying in the one-dimensional subspace corresponding to the top principal component. A 99% decrease would correspond to saying that almost none of the variation in the genomes is captured in this single direction, even though this is the single "best" direction for capturing that variation.) What about if we project onto the top 3-dimensional subspace? In one or two sentences, discuss whether you are surprised by this or not.

Deliverables: Scatter plot for part (b). Short discussion for part (c). Scatter plots for parts (d) and (f). One sentence answers for (e) and (f). Percentages decrease for top, and top 3 principal components, and one or two sentence discussion for (g). Code for the whole section in the Appendix.