PADL Practical: Descriptive Statistics, Data Visualisation and PCA (100 marks)

Dimitar Kazakov
dlk2@york.ac.uk

29 Feb 2024

1 Modelling Migration Waves from Genetic Data

This practical assignment is based on a part of a genuine past open assessment. Individual parts of the work have marks assigned (out of 100) in order to describe their relative difficulty, rather than suggest that the amount of work was equivalent to an entire exam. A sample solution will be released at the end of Week 5 to allow each student to assess their own performance. Unlike a real exam, checking each other's work would also be an excellent idea.

1.1 Data and Deliverables

All data and any other additional files are available from the PADL VLE site in the **Practicals/Week 4** section. Your solution should be contained in single Jupyter notebook `Surname-name.ipynb` (combining code and explanations). Your code should assume all data files are in the same folder as the notebook from which they are accessed.

All Python code should be in Python3. Your Jupyter notebook must run correctly on Google Colab. There are no word limits on your text comments.

1.2 Preliminaries

The population of Square Island (Fig. 1), a territory perfectly aligned with the four cardinal directions (North, East, South, West), was established in three migration waves, which are reflected in the genetic makeup of the inhabitants.

The earliest wave preceded the other two by a long margin. It consisted of hunter-gatherers whose genetic makeup had been distributed uniformly across the whole island before the next two waves arrived. The second migration wave entered the island through an isthmus, that is, a narrow strip of land, which temporarily connected the South-Western corner of the island with the nearest

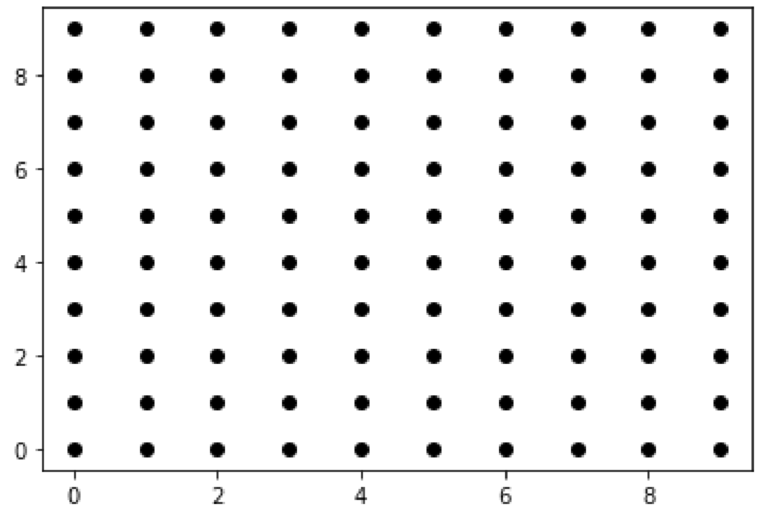

Figure 1: Map of Square Island (North is up)

continent during a mini ice age period when the sea levels dropped. The new arrivals were farmers who started to spread slowly, breaking new ground for farming and advancing by about one mile with each generation. The third and last migration wave brought a population of seafarers to the island's shores.

We have data on the relative frequency of 7 genes (proportion of population with a given gene expressed as a number between 0 and 1) as measured at various locations on the island. These locations are spaced at equal intervals along the X coordinate (West to East) and Y coordinate (South to North) over the entire area. Genes 1 and 2 are mutually exclusive alternatives (*alleles*) that can appear at one specific position in the genome known as Locus 1. The same is valid for genes 3 and 4, which are the only 2 alternatives for Locus 2, and genes 5, 6 and 7, which are the three alleles that can appear in Locus 3. This means that in any given location (x,y) the relative frequencies of Gene 1 and Gene 2 add up to 1, and so do the relative frequencies of Genes 3 and 4, and Genes 5, 6 and 7. The data is available as a CSV table (see file `sqisland.csv`) with a header row and 9 columns, representing the attributes listed in Table 1.

So, we may know, for instance, that in a given location 70% of the population have Gene 1, and the remaining 30% have Gene 2. Similarly, the proportion of the population in that location with Gene 3 may be, say, 40%, which leaves the remaining 60% carrying Gene 4. Finally, the individuals in that location that carry one of the genes 5–7 may be split as 25% : 40% : 35%.

Remember, each of the 3 migration waves brought a new population with its own specific genetic makeup (relative gene frequencies) that mixed with the existing population over time.

Column	Name	Range
1	X coordinate	$x \in \mathcal{N}, 0 \leq x \leq 9$, grows from left to right on the map
2	Y coordinate	$y \in \mathcal{N}, 0 \leq y \leq 9$, grows from bottom to top on the map
3	Gene 1	$v \in \mathcal{R}, 0 \leq v \leq 1$
4	Gene 2	$v \in \mathcal{R}, 0 \leq v \leq 1$
5	Gene 3	$v \in \mathcal{R}, 0 \leq v \leq 1$
6	Gene 4	$v \in \mathcal{R}, 0 \leq v \leq 1$
7	Gene 5	$v \in \mathcal{R}, 0 \leq v \leq 1$
8	Gene 6	$v \in \mathcal{R}, 0 \leq v \leq 1$
9	Gene 7	$v \in \mathcal{R}, 0 \leq v \leq 1$

Table 1: Data file attributes

1.3 To Do

14 marks For each of the 7 genes, produce a contour plot visualising how its relative frequency varies across the whole island. (Consider using `matplotlib.pyplot.contourf`.)

12 marks Study the contour plots to form a hypothesis about the most common alleles for Locus 1 and Locus 2 in: (a) the hunter-gatherers' population; (b) in the farmers' population.

8 marks Describe any significant characteristics of the genetic makeup of the population of seafarers.

10 marks Calculate and display the variance of each of the 7 gene attributes.

16 marks Calculate the Pearson correlation between (a) Gene 1 and Gene 4; (b) Gene 1 and Gene 5. State if the null hypothesis of non-correlation can be rejected for either pair at the 95% significance level. Do these results agree with your hypothesis about the genetic makeup of the farmers from the second wave?

16 marks Apply principal component analysis (PCA) to the data consisting of the relative frequencies of Genes 1–7. Transform the data using all 7 principal components and calculate and display the variance for each of them.

8 marks Compare the sums of variances of all 7 attributes before and after transforming the data via PCA. Comment briefly whether the result can be expected or not and why.

16 marks Plot the first two PCA components as contour plots visualising the relative frequencies of each component across the island. Compare the result to the contour plots for Gene 1 and Gene 3 (data before PCA). Which of the two pairs of plots do you find more helpful for the task of reconstructing the waves of migration? Do you expect the same result for a realistic data set with hundreds of genes and why?