Practical 04 SG: Population substructure

Write here your names and surnames

Hand-in: 10/12/2022

Resolve the following exercise in groups of two students. Perform the computations and make the graphics that are asked for in the practical below. Take care to give each graph a title, and clearly label x and y axes, and to answer all questions asked. You can write your solution in a word or Latex document and generate a pdf file with your solution. Alternatively, you may generate a solution pdf file with Markdown. You can use R packages **genetics**, **MASS**, **data.table** and others for the computations. Take care to number your answer exactly as in this exercise. Upload your solution in **pdf format** to the web page of the course at raco.fib.upc.edu no later than the hand-in date.

1. The file Chr21.dat contains genotype information of a set of individuals of unknown background. Load this data into the R environment with the **fread** instruction. The first six columns of the data matrix contain identifiers, sex and phenotype and are not needed. The remaining columns contain the allele counts (0, 1 or 2) for over 138.000 SNPs for one of the alleles of each SNP.
2. (1p) Compute the *Manhattan distance* matrix between the individuals (this may take a few minutes) using R function `dist`. Include a submatrix of dimension 5 by 5 with the distances between the first 5 individuals in your report.
3. (1p) The Manhattan distance (also known as the *taxicab metric*) is identical to the Minkowsky distance with parameter $\lambda = 1$. How does the Manhattan distance relate to the allele sharing distance, where the latter is calculated as two minus the number of shared alleles?
4. (2p) Apply metric multidimensional scaling using the Manhattan distance matrix to obtain a map of the individuals, and include your map in your report. Do you think the data come from one homogeneous human population? If not, how many subpopulations do you think the data might come from, and how many individuals pertain to each suppopulation?
5. (1p) Report the first 10 eigenvalues of the solution.
6. (1p) Does a perfect representation of this $n \times n$ distance matrix exist, in n or fewer dimensions? Why so or not?
7. (1p) What is the goodness-of-fit of a two-dimensional approximation to your distance matrix? Explain which criterium you have used.
8. (2p) Make a plot of the estimated distances (according to your map of individuals) versus the observed distances. What do you observe? Regress estimated distances on observed distances and report the coefficient of determination of the regression.
9. (1p) We now try non-metric multidimensional scaling using the `isoMDs` instruction. We use a random initial configuration. For the sake of reproducibility, make this random initial configuration with the instructions:

where n represents the sample size. Make a plot of the two-dimensional solution. Do the results support that the data come from one homogeneous population?

Try some additional runs of `isoMDS` with different initial configurations, or eventually using the classical metric solution as the initial solution. What do you observe?

10. (1p) Set the seed of the random number generator to 123. Then run isoMDS a hundred times, each time using a different random initial configuration using the instructions above. Save the final stress-value and the coordinates of each run. Report the stress of the best run, and plot the corresponding map.

11. (1p) Make again a plot of the estimated distances (according to your map of individuals of the best run) versus the observed distances, now for the two-dimensional solution of non-metric MDS. Regress estimated distances on observed distances and report the coefficient of determination of the regression.

12. (1p) Compute the stress for a $1, 2, 3, 4, \ldots, n$-dimensional solution, always using the classical MDS solution as an initial configuration. How many dimensions are necessary to obtain a good representation with a stress below 5? Make a plot of the stress against the number of dimensions

13. (2p) Compute the correlation matrix between the first two dimensions of a metric MDS and the two-dimensional solution of your best non-metric MDS. Make a scatterplot matrix of the 4 variables. Comment on your findings.