BSG-MDS practical 4 Population Substructure

Name Surname Name Surname

28/11/2023, submission deadline 05/12/2023

Resolve the following exercise in groups of two students. Write the R scripts, 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, or generate a solution pdf file with R Markdown. Take care to number your answers 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 submission deadline.

You can use of the R-package `MASS` and `data.table` to compute your answers. The dataset can be downloaded from the web page of the course at raco.fib.upc.edu.

Population substructure (10p)

The file `Chr21.dat` contains genotype information of a set of individuals of unknown background for over 138.000 SNPs.

Load this data into the R environment, with the `fread` instruction of the package `data.table`, which is more efficient for reading large datafiles. The first six columns of the data matrix contain identifiers, sex and phenotype and are not needed. Notice that the genetic variants are identifed by an "rs" identifier. The genetic data is coded in the (0, 1, 2) format with 0=AA, 1=AB, 2=BB.

1. (0.5p) How many variants are there in this database? What percentage of the data is missing?
2. (0.5p) Compute the Manhattan distance matrix between the individuals (which is identical to the Minkowsky distance with parameter $\lambda = 1$) 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) How does the Manhattan distance relate to the allele sharing distance?
4. (1p) Apply metric multidimensional scaling (`cmdscale`) with two dimensions, $k = 2$, using the Manhattan distance matrix and include the 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) What is the goodness-of-fit of the two-dimensional approximation to your distance matrix? Explain which criterium you have used.
6. (1p) Make a plot of the estimated distances (according to your two-dimensional 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 (you can use the function `lm`).
7. (1p) We now try a (two-dimensional) non-metric multidimensional scaling using the `isoMDs` function that you will find in `MASS` library. We use a random initial configuration and, for the sake of reproducibility, make this random initial configuration with the instructions: `set.seed(12345)` and `init <- scale(matrix(runif(m*n),ncol=m),scale=FALSE)` where `n` represents the sample size and `m` represents the dimensionality of the solution. Make a plot of the two-dimensional solution. Do the results support that the data come from one homogeneous population?
8. (1p) Try some additional runs of the two-dimensional isoMDS with different initial configurations. Make a plot of the solutions and report the STRESS for each of them. What do you observe?

9. (1p) Compute the stress for a 1, 2, 3, . . . , 50-dimensional solution. How many dimensions are necessary to obtain a good representation with a stress below 10? Make a plot of the stress against the number of dimensions.

10. (1p) Run the two-dimensional isoMDS a hundred times, each time using a different random initial configuration using the instructions above. Report the stress of the best and the worse run, and plot the corresponding maps. Compare your results to the metric MDS and comment on your findings.

11. (1p) Compute the correlation matrix between the first two dimensions of the metric MDS and the two-dimensional solution of your best non-metric MDS. Comment your findings.