Bioinformatics and Statistical Genetics	Population substructure
Jan Graffelman	December 4, 2018

Name: .....

Name: .....

You can make use of the R-package **genetics** (and other packages) to compute your answers. Prepare a .pdf file with all your answers and figures. Send your work by email to the course instructor (`jan.graffelman@upc.edu`) no later than the 15^{th} of December 2018.

1. The file SNPChr20.rda contains genotype information of 310 individuals of unknown background. The genotype information concerns 50.000 SNPs on chromosome 20. Load this data into the R environment. The data file contains a matrix `Y` containing the allele counts (0,1 or 2) for 50.000 SNPs for one of the alleles of each SNP.

2. (1p) Compute the Manhattan distance matrix between the 310 individuals (this may take a few minutes) Include a submatrix of dimension 5 by 5 with the distances between the first 5 individuals in your report .....

   .....

3. (1p) Use metric multidimensional scaling to obtain a map of the individuals, and include your map in your report. Do you think the data come from one homogeneous population? .....

   .....

4. (1p) Report the first 10 eigenvalues of the solution. .....

   .....

5. (1p) Does a perfect representation of this $n \times n$ distance matrix exist, in n or fewer dimensions?

   .....

6. (1p) What is the goodness-of-fit of a two-dimensional approximation to your distance matrix? ..

   .....

7. (1p) Make a plot of the estimated distances (according to your map of individuals) versus the observed distances. Regress estimated distances on observed distances and report the coefficient of determination of the regression. .....

   .....

8. (1p) Try now non-metric multidimensional scaling with your distance matrix. Use both a random initial configuration as well as the classical metric solution as an initial solution. Make a plot of

the two-dimensional solution. Do the results support that the data come from one homogeneous population? .....

.....

9. (1p) Make again a plot of the estimated distances (according to your map of individuals) 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. Is the fit better or worse than with metric MDS? .....

.....

10. (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? Make a plot of the stress against the number of dimensions .....

.....

11. (1p) Compute the correlation matrix between the first two dimensions of a metric MDS and the two-dimensional solution of a non-metric MDS. Make a scatterplot matrix of the 4 variables. Comment on your findings .....

.....

.....