Multidimensional scaling for relatedness research: an application of the Aitchison distance in the GCAT population based cohort

I. Galván-Femenía[1,2,3], J. Graffelman[5,6], C. Barceló-i-Vidal[1], L.Sumoy[4], V. Moreno[7], and R. de Cid[2,3]

[1]Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona; Girona, Spain; *ivan.galvan@udg.edu*

[2]Disease Genomics group. [3]GCAT group. [4]Genomics and Bioinformatics lab group. Germans Trias Health Research Institute (IGTP)-Program of Predictive and Personalized Medicine of Cancer (PMPPC), Can Ruti Campus; Badalona, Barcelona, Spain

[5]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya; Barcelona, Spain
[6]Department of Biostatistics, University of Washington; Seattle, WA, USA

[7]Catalan Institute of Oncology, IDIBELL, Epidemiology and Cancer Registry; L'Hospitalet, Barcelona, Spain

Abstract

Multidimensional scaling is a classical multivariate technique used for analysing similarities or distances (Mardia et al., 1979). This method is commonly used, in a genetic context, for population structure investigations, i.e. to distinguish individuals from different human populations (Nelis et al., 2009). For this purpose, genetic markers like single nucleotide polymorphisms (SNPs) are useful. These markers are bi-allelic and only three genotypes (or categories) exist for each marker. In this contribution, we use multidimensional scaling as a tool for relatedness research with SNP data. Family-relatedness investigations are crucial in gene-disease association studies. If there are related individuals in the population under study, then the statistical models can generate misleading conclusions (Shibata et al., 2013). Here, we analyze a dataset from the GCAT study, a population based cohort from Catalonia (north-east Spain, http://www.genomesforlife.com). Currently, the GCAT dataset contains roughly 5,000 individuals and a SNP-array of 2 million genetic markers. To define genetic distances between the GCAT individuals we recode each SNP by following an additive model. SNPs are coded as 0, 1 or 2 for the genotypes AA, AB and BB respectively, where A represents the minor allele. For each pair of individuals, we calculate a 3 x 3 contingency table, cross-classifying all SNPs with respect to the genotypes of the individuals. These contingency tables were converted into nine part compositions. Under these assumptions, we apply the Aitchison distance between contingency tables and plot the distances between pairs of individuals of a two-dimensional solution using classical metric multidimensional scaling. Using this method, we replicate all the family relationship clusters uncovered with the popular methods used for relatedness research, furthermore being able to identify unambiguously all unrelated individuals in the sample. Calculations were carried out using PLINK and R software.

References

Mardia K.V. et al. (1979). *Multivariate analysis.* Chapter 14. Academic press.

Nelis M. et al. (2009) Genetic structure of Europeans: a view from the North-East. *PLoS ONE 4*(5). doi:10.1371/journal.pone.0005472

Shibata K. et al. (2013) The confounding effect of cryptic relatedness for environmental risks of systolic blood pressure on cohort studies *Molecular Genetics* & *Genomic Medicine 1*(1), pp. 45-53.