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, R. de Cid 2,3

1 Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona; Girona, Spain;
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
3 Department of Statistics and Operations Research, Universitat Politècnica de Catalunya; Barcelona, Spain
4 Department of Biostatistics, University of Washington; Seattle, WA, USA
5 Department of Biostatistics, University of Washington; Seattle, WA, USA
6 Catalan Institute of Oncology, IDIBELL, Epidemiology and Cancer Registry; L'Hospitalet, Barcelona, Spain
7 Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona; Girona, Spain;

Introduction

Multidimensional scaling (MDS) 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 (Sabatti et al., 2009), to distinguish individuals from different human populations (Figure 1). For this purpose, genetic markers like single nucleotide polymorphisms (SNPs) are useful.

Figure 1: Three-dimensional MDS solution for the GCAT sample and worldwide human populations from the 1000 Genomes Project (www.internationalgenome.org) for a subset of 99,873 SNPs.

SNPs are mostly 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 genome-disease association studies. If there are related individuals in any population, like in our GCAT study, then the statistical models used in future association studies can generate misleading conclusions (Shibata et al., 2013).

Materials and methods

For this reason, we are interested in analyzing the unknown family relationships in the GCAT, a population based cohort from Catalonia (northeast Spain, https://www.genomesforlife.org). 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 the genotypes α, β and γ as 0, 1 and 2 respectively, where α represents the minor allele. For each pair of individuals, we calculate a 3 x 3 contingency table (Table 1), cross-classifying all SNPs with respect to the genotypes of the individuals.

<table>
<thead>
<tr>
<th>Individual A</th>
<th>Individual B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Contingency table of the pair of individuals A and B. The counts n(αjβk = 0,1,2) are the number of SNPs for which the individuals A and B have the i,j alleles respectively. Note that the same situation occurs when classifying the genotypes for the counts nαβ, nαγ, nβα, nβγ, nγα, nγβ, nαα, nαγ, nββ and nγγ.

By creating the vector \( \mathbf{n} \), we represent each pair of individuals as a 6 part-composition and calculate the Aitchison distance (\( d \)) between two pairs as follows:

\[
d(x_{ij}, y_{jk}) = \left( \sum_{s=1}^{3} \frac{1}{\mathbf{n}^{s}} \left( \frac{x_{ij} - y_{jk}}{x_{ij}^{s} + y_{jk}^{s}} \right) \right)^{1/2}
\]

where \( p_{x} \) and \( p_{y} \) represent the 6-part compositions for the pairs formed by the individuals A and B, C and D respectively.

Results

We consider SNPs on the autosomal chromosomes with minor allele frequency ≥ 0.4 and p-value > 0.05 for the Hardy-Weinberg Chi-square test. Correlated SNPs with Linkage Disequilibrium r² statistic > 0.2 in a sliding window of size 500 are discarded. A total of 21,618 SNPs remained after these inclusion/exclusion criteria. We estimate the identity by descent (IBD) probabilities for all the pairs of individuals from the GCAT sample. We randomly select a subset of 60 potentially unrelated individuals with IBD probability of sharing 0 alleles larger than 0.95. Starting from these unrelated individuals, we generate artificial children (Figure 2A) and plot the identity by state (IBS) proportions of all possible pairs (\( n = 1,275 \)) in the ternary diagram (Figure 2B).

Figure 2: A. Family relationships generated from unrelated individuals. B. Ternary plot of the IBS proportions of a random subset of pairs of unrelated individuals from the GCAT and simulated pairs of given family relationships.

After the simulations, we estimate potentially related pairs of individuals from the GCAT sample with IBD probability of sharing 0 alleles smaller than 0.90 (\( n = 1,236 \)). We summarize all related, unrelated and simulated pairs (\( n = n_{1} + n_{2} = 2,511 \)) in contingency tables like Table 1 and calculate the Aitchison distance between them. We consider the a x n matrix containing the Aitchison distances as input for classical metric multidimensional scaling. Finally, we plot the two-dimensional solution (Figure 3) to classify the related individuals from the GCAT (gray color) into the different family relationships categories (MZ, PO, FS, GC/GV/H, FC, UN).

Figure 3: Multidimensional scaling solution of the potentially related pairs of individuals from the GCAT (gray color) and simulated pairs of related individuals.

Conclusions

We replicate all the family relationship clusters uncovered with the popular methods used for relatedness research (Figure 2B, plot of identity by state proportions), furthermore being able to identify unrelated individuals in the sample by regarding the negative values in the 1st dimension of the two dimensional MDS solution.

References


Acknowledgments

This study makes use of data generated by the “GCAT” Genomes for Life Project, Cohort study of the Genomes of Catalonia”. Funding for the project was provided by Acción de diminución MINECO ADIE18/00026. I. Galván-Femenía and C. Barceló-i-Vidal are supported by the Ministerio de Economía y Competitividad under the project “CODA-RETOS” (Ref. MTM2015-65016-C2-1-R). J. Graffelman is supported by MTM2015-65016-C2-2-R. R. de Cid is supported by RYC-2011-07822 and AGAUR (SGR 1269). IGTP is part of the CERCA Program / Generalitat de Catalunya.