Journal of Statistics on Health Decision

EXTENDED ABSTRACT

Assessment of the stability of a procedure for variables selection in high dimensionality data: an application to genomic data - Alzheimer's Disease

Leonor Rodrigues[1], Jorge Cabral[1], Ana H. Tavares[2,3], Vera Enes[1], Miguel Pinheiro[4,5], Gabriela Moura[4,5], Vera Afreixo[1,3]

[1]Department of Mathematics, University of Aveiro, Aveiro, Portugal
[2]ESTGA – Águeda School of Technology and Management, Águeda, Portugal
[3]CIDMA – Center for Research & Development in Mathematics and Applications, University of Aveiro, Aveiro, Portugal
[4]Department of Medical Sciences, University of Aveiro, Aveiro, Portugal
[5]iBiMED – Institute of Biomedicine, University of Aveiro, Aveiro, Portugal

Introduction:

Alzheimer's disease (AD) is a neurodegenerative disease and a complex disorder caused by a combination of environmental and genetic factors [1]. One of the main goals of modern genetics has been unraveling the genetic background of common complex disorders, so Genome-Wide Association Studies have been conducted with large-scale data sets of genetic variants (Single Nucleotide Polymorphisms - SNPs). Most of these studies have relied on approaches that consist of a univariate analysis of the association of each SNP with the phenotype. Consequently, the possibility of a correlational and interactional structure between SNPs is not considered [2]. The challenge in finding a plausible method to apply to genetic data is due to its high dimensionality.

This work aims to assess the stability of a procedure that identifies association between relevant SNPs and AD in a structure where the number of SNPs (p) is much more than the number of individuals (n), ($p >>> n$ problem).

Methods:

The genotypic data used in this study was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) – 1 study (adni.loni.usc.edu). In this work, were considered 451 individuals (163 cognitively normal - 36.1%; 288 AD – 63.9%) and 518257 SNPs [3]. A hundred prediction models were constructed using Least Absolute Shrinkage and Selection Operator (LASSO) and applied a feature selection and explanation procedure [3]. A weight based on Akaike's Information Criterion (AIC) was calculated for each SNP. The criteria for classifying a SNP as important was its weight being at least 0.8 and only these SNPs were part of the final model proposed (original model). We re-run the analysis excluding one selected variable at a time to measure how sensitive the findings were. In addition, we also re-run the analysis after removing from the data SNPs selected by LASSO but not considered important according to our procedure (weight < 0.8). For this purpose, three SNPs, classified as non-important, were randomly chosen. The sensitivity of the resulting models to the removal of these SNPs was also evaluated.

Results:

The procedure that gave rise to the original model led to the choice of 11 SNPs (the first row of the heatmap represented in Fig. 1). The removal of the SNPs rs12054808, rs4391167 and rs1052242 from the data had no impact on the selected variables, but the weights of the remaining SNPs increased. The removal of the SNPs rs11625567 and rs6090754 also had no considerable impact: only the SNP rs4391167 became unimportant, but the removal of the SNP rs6090754 caused the weight of the remaining SNPs to increase. Contrariwise, removing the SNPs rs486512, rs2075650, rs6427160, rs4982401 and rs11906462 made all the SNPs unimportant. The removal of the SNP rs573399 also made changes: only three SNPs remained important. Regarding the SNPs that are not part of the original model, their removal did not have a considerable impact on the selected variables: only the SNP rs4391167 became unimportant with the removal of the SNPs rs1387089 and rs1582317. A summary of the results presented is in Fig. 1.

Keywords:
Alzheimer's Disease, Single Nucleotide Polymorphisms, Akaike's Information Criterion, Least Absolute Shrinkage and Selection Operator, Stability

Corresponding author:
Leonor Rodrigues
leonorcrodrigues@ua.pt

Conflict of interest:
The authors declare no conflict of interests.

First published: 20JUL2022

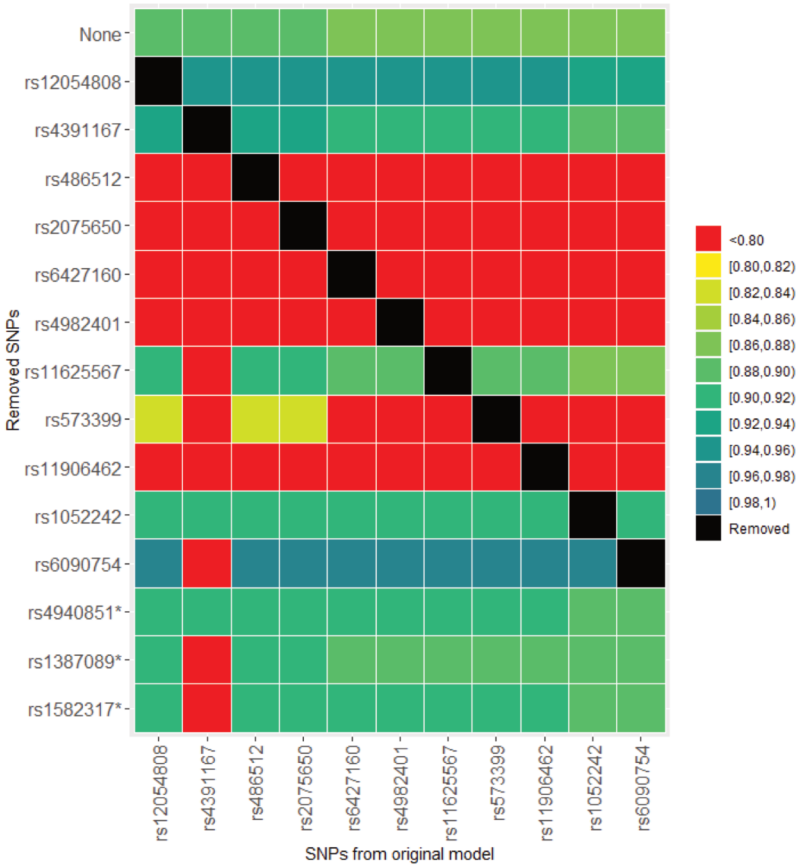

Figure 1 - Scheme of the SNPs selected in the original model (row "None") and after removal of these SNPs from the data one by one. The SNPs assigned with * are not part of the original model. The SNPs are colored according to their weight.

Discussion:

It is possible that the SNPs that when removed cause substantial changes in the selection of SNPs compared to the original model are the really important ones. An argument that can support this hypothesis is that the removal from the data of almost all SNPs tested makes the SNP rs4391167 unimportant and the removal of this SNP from the data had no impact on the selection of the SNPs. Another argument is that the removal of the SNPs that are not included in the original model did not have a substantial considerable impact on the selection of SNPs. In the future, testing the sensitivity of the model to the removal of more SNPs like this will be important.

Acknowledgements:

We thank ADNI for supplying us with their databases.

References:

1. Ridge PG, Mukherjee S, Crane PK, Kauwe JSK. Alzheimer's disease: Analyzing the missing heritability. PLoS One. 2013;8(11):1–10. https://doi.org/10.1371/journal.pone.0079771
2. Cho S, Kim K, Kim YJ, et al. Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association Analysis. Ann Hum Genet. 2010;74(5):416–428. https://doi.org/10.1111/j.1469-1809.2010.00597.x
3. Rodrigues L, Tavares AH, Enes V, Pinheiro M, Moura G, Afreixo V. Consistent variable selection in shrinkage regression: Alzheimer's Disease and genomic data. Submitted.