The SgenoLasso, a new Lasso method dedicated to extreme observations in genomics

Charles-Elie Rabier[1*], Céline Delmas[2]

[1] *IMAG, Université de Montpellier, CNRS, France; charles-elie.rabier@umontpellier.fr.*
[2] *Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France; celine.delmas.toulouse@inrae.fr*
[*] *Presenting author*

Keywords. *High-dimensional linear model; Variable selection; RALasso; Extreme data; Genomics*

In genomics, selective genotyping consists in genotyping (collecting DNA information at specific positions) only the individuals with extreme traits (i.e. with the largest or smallest trait values). This famous concept was introduced by [Lebowitz et al. , 1987]: the authors noticed that the highest or the lowest observations contain most of the signal on Quantitative Trait Loci (QTL), i.e. genes with quantitative effect on a trait. Later, [Lander & Botstein , 1989] elaborated this concept.

Nowadays, although the genotyping costs have drastically dropped, selective genotyping is still heavily used since we can optimize the statistical experiment by focusing on extreme individuals instead of random individuals. There is still a lack of tools to analyze properly this kind of data since classical penalized regressions (e.g. Lasso [Tibshirani , 1996]) are not dedicated to extreme observations.

From a statistical point of view, the linear model we are dealing with, presents the particularity of incorporating some correlation between the errors ε and the regressors, due to selective genotyping. As a consequence, we introduce the SgenoLasso [Rabier & Delmas , 2021], a new L1 penalized regression that models explicitly this correlation. The SgenoLasso relies on the “Interval Mapping” [Lander & Botstein , 1989], a famous concept in genetics that consists in scanning the genome by testing the presence of a QTL at each location. SgenoLasso is based on new limiting results on stochastic processes along the genome. SgenoLasso enjoys all known statistical properties of Lasso since the problem has been replaced in a classical L1 penalized regression framework. Typically, it is not the case for Lasso in presence of extreme data.

We compared the SgenoLasso with the “Robust Approximate quadratic Lasso” (RALasso) of Fan et al. [2017], which incorporates the Huber loss and a L1 penalty.

The RALasso can be viewed as a more flexible method than the Lasso: the loss function can be either quadratic or linear, depending on the error values. The tuning parameter helps to handle errors with different shapes and tails. Recall the [Huber , 1964] loss considered in the R package hqreg :
$\text{loss}(t) = \frac{t^2}{2M} 1_{|t| \leq M} + (|t| - M/2)\, 1_{|t| \geq M}$, where M is a tuning parameter. As soon as we multiply by $2M$ and that we replace M by α^{-1}, we obtain the RALasso loss described in formula (2.2) of Fan et al. [2017].

On the basis of a simulation study where only the largest individuals were selected, the RALasso, that models heavy tails and asymmetry, gave better results than classical methods such as Lasso, GroupLasso [Yuan & Lin , 2006] and Bayesian Lasso [Park et al. , 2008]. However, in that case, our SgenoLasso performed better than the RALasso. Last, we will show the superiority of the Adaptive version of the SgenoLasso, called the AdaptSgenoLasso, that allows to put more weights on some regressors of interests (e.g. well known genes).

References

Lander, E.S. & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **138**, 235–240 .

Lebowitz, R.J., Soller, M., & Beckmann, J.S. (1987). Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Genetics*, **73**, 556–562.

Fan, J., Li, Q. & Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society B*, **79(1)**, 247–265.

Huber, P.J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**, 73–101.

Park, T. & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103(482)**, 681–686.

Rabier, C.E. & Delmas, C. (2021). The SgenoLasso and its cousins for selective genotyping and extreme sampling: application to association studies and genomic selection. *Statistics*, **55(1)**, 18–44.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58(1)**, 267–288.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, **68(1)**, 49–67.