The SgenoLasso and its cousins for selective genotyping and extreme sampling

Charles-Elie Rabier[1,2] and Céline Delmas[3]

[1] Institut Alexander Grothendieck Montpellier Institute (IMAG), Université de Montpellier, CNRS, France

[2] Institut des Sciences de l'Evolution (ISEM), Université de Montpellier, CNRS, IRD, EPHE, Montpellier, France

[3] INRAE, UR MIAT, Université de Toulouse, Castanet-Tolosan, France

Corresponding author: charles-elie.rabier@umontpellier.fr

Reference paper:* Rabier *et al.* (2021). The SgenoLasso and its cousins for selective genotyping and extreme sampling: application to association studies and genomic selection. *Statistics: A Journal of Theoretical and Applied Statistics, 55(1).
https://www.tandfonline.com/doi/full/10.1080/02331888.2021.1881785

Keywords: Selective Genotyping, Genomic Selection, Variable Selection, Prediction Accuracy, High Dimension, Lasso, Rice data

Context: In a seminal paper, Lebowitz et al. (1987) showed that the extreme observations of a given trait (i.e. the highest or the lowest observations) contain most of the signal on Quantitative Trait Loci, so-called QTL (genes influencing a quantitative trait which is able to be measured). As a consequence, the authors suggested to genotype only the individuals with extreme phenotypes. This concept is called selective genotyping and it was formalized later by Lander and Bostein (1989). Genome Wide Association Study (GWAS) and Genomic Selection (GS) are today two research topics using the selective genotyping methodology.
We denote some recent association studies using selective genotyping in plants (e.g. sugarcane, Gutierrez et al. 2018; tomatoes, Ohlson et al. 2018) in animals (e.g. dairy cattle, Kurz et al. 2019), and in humans (e.g. on intelligence, Zabaneh et al. 2018). Selective genotyping is particularly rewarding for finding QTLs: by considering the extremes, the signal is significantly increased. The second application field of selective genotyping is Genomic Selection (GS) (Hayes et al., 2001), which is nowadays a very popular topic in genomics (e.g. strawberry, Gezan et al. 2017; banana, Nyine et al. 2018). The main goal of GS is to select individuals (i.e. candidates) by means of genomic predictions. Since predictions can be performed as soon as the DNA is available, GS accelerates significantly the genetic gain. In GS, the learning model has to be recalibrated over time, otherwise it leads to unreliable predictions (see Goddard et al. 2009). As a result, when updating the model, candidates selected at the previous steps are used to train the model. This way, the model is learned on extreme individuals, which is highly linked to selective genotyping.

Results: We introduce here a new variable selection method, called SgenoLasso (for Selective genotyping Lasso), that handles extreme data. SgenoLasso allows to estimate the number of QTLs, their positions and their effects. It differs from the classical Lasso (Tibshirani 1996) since it models explicitly the extremes. SgenoLasso enjoys all known statistical properties of Lasso since the problem has been replaced in a L1 penalized regression framework. As its famous ancestor Lasso, SgenoLasso has multiple cousins: we can cite for instance SgenoElasticNet (a mixture of L1 and L2 penalties) and SgenoGroupLasso (penalty by group).
We propose a comparison with existing methods in a GWAS context, on simulated data and on rice data. SgenoLasso and its cousins outperformed existing methods (Lasso, Group Lasso, Yuan and Lin 2006, Elastic Net, Zhou and Hastie 2005, RaLasso, Fan et al. 2017, and BayesianLasso, Park and Casella 2008), specially when a unidirectional selective genotyping was performed (i.e. we genotype only the so-called best individuals with the largest phenotypes).
In GS, Zhao et al. (2012) highlighted the "drastic reduction" in terms of predictive ability when only the best individuals were used in the learning model in GS. Interestingly, Brandariz and Bernardo (2018) have shown recently that it is crucial to include a few worst individuals in the training set, to keep GS efficient. However, keeping the poorest lines in a breeding program has a non negligible cost. In this context, we show on simulated data that SgenoLasso and its cousins do not suffer from this drawback: they give satisfactory results even when only best individuals are considered.