The SgenoLasso and its cousins for selective genotyping and extreme sampling

Charles-Elie Rabier, Céline Delmas

IMAG, Institut Montpelliérain Alexander Grothendieck
ISEM, Institut des Sciences de l'Evolution de Montpellier
Université de Toulouse, INRAE, UR MIAT

07/09/2021

Outline

“The SgenoLasso and its cousins for selective genotyping and extreme sampling”

R. and Delmas, Statistics, Volume 55, 2021

Genomic Selection (GS)

Training population

$$Y = \mu + \underbrace{f(M,\theta)}_{g} + \varepsilon$$

Model

$$\hat{\theta}$$

Estimation method

Test population (selection candidates)

$$\underbrace{f(M,\hat{\theta})}_{\hat{g}}$$

Prediction

GS motivated by Meuwissen et al (2001)

Genomic Selection

Predictions can be performed as soon as the DNA is available
⇒ GS accelerates significantly the genetic gain

We do not have to wait to observe the phenotype
of the candidate at adult age ...

For instance,

- in bananas (Nyine et al., 2018) : 8 months before having an idea on the production capacity
- in citrus (Minamikawa et al, 2017) : 25 years before obtaining the fruits of interest

Genomic Selection

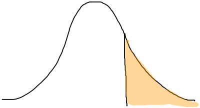
- At the first generation
  - individuals are phenotyped and genotyped
  - the model is learnt
- next, at each generation
  - no need to phenotype the individuals
  - only need to genotype individuals
  - individuals selected on the basis of genomic predictions
- After a large number of generations
  - calibration model not reliable anymore
  - need to genotype and to phenotype again
  - a new model is learnt

How can we learn a model using selected individuals ?

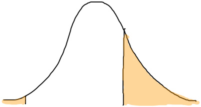
Can we learn a model using selected individuals ?

“Maintaining the accuracy of genomewide predictions when selection has occurred in the training population”

by Brandariz SP and Bernardo R, Crop Science, 58(3), 2018



it does not work



it works

In order to obtain a reliable model, we need to keep
a few worst individuals in the breeding programs

Selective Genotyping is highly linked to Genomic Selection

Genotyping was expensive in the past

⇒ Selective Genotyping : we genotype only individuals who present extreme phenotypes Y

At a given power, a large increase of the number of individuals

leads to a decrease of the number of individuals genotyped

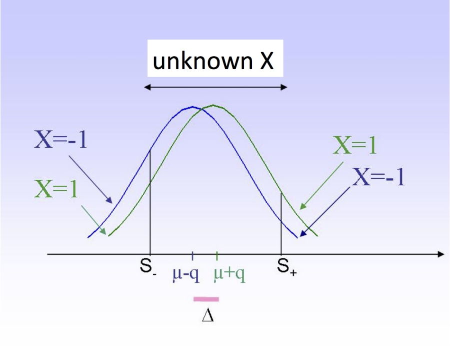
Lebowitz et al. (Theoretical and Applied Genetics, 1987)
Darvasi and Soller (Theoretical and Applied Genetics, 1992)

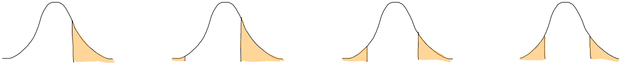
To go further in the statistical theory :

R. (Journal of Statistical Planning and Inference, 2014)

Model corresponding to selective genotyping

Probability distribution of the phenotypes *Y*





Worst scenario

Best scenario

Can we elaborate a method able to learn
a model based on extreme individuals ?

Model

- $X(.)$: genome of one individual
- $t_1^\star, \ldots, t_m^\star$: QTL (i.e. Quantitative Trait Loci) locations
- Assuming a linear model for the phenotype Y

$$Y = \mu + \sum_{s=1}^{m} X(t_s^\star) q_s + \sigma\varepsilon \qquad \text{with} \quad \varepsilon \sim N(0,1)$$

- Genome information X(.) available :
  - only at genetic markers t_1, . . ., t_K
  - only if Y is extreme (i.e. $Y > S_+$ or $Y < S_-$)

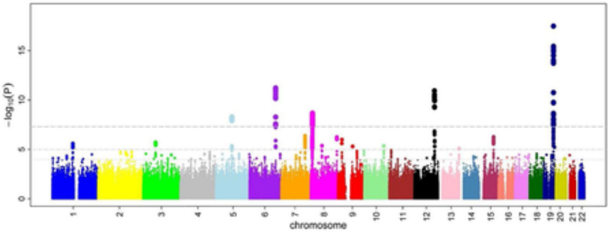
$\Rightarrow$ Dependency between the alleles at the markers and the extreme phenotypes Y

- The LASSO (Tibshirani, 1996) is unable to handle this dependency

A new approach is needed ...

Our starting point

ManhattanPlot in association studies



source Wikipedia

The Interval Mapping of Lander and Botstein (1989) :

- The chromosome is represented by a segment $[0, T]$
- $\Lambda_n(t)$: Likelihood Ratio Test at a given location $t \in [0, T]$, for testing $q_1 = 0$ vs $q_1 \neq 0$
- $\Lambda_n(.)$: Likelihood Ratio Test process on $[0, T]$

Score statistic and LRT statistic

- $\theta^1 = (q_1,\ \mu,\ \sigma)$ parameter of the model at t fixed
- $\theta_0^1 = (0,\ \mu,\ \sigma)$ stands for H_0

Score statistic at t

$$S_n(t) = \frac{\frac{\partial l_t^n}{\partial q_1}\ |_{\theta_0^1}}{\sqrt{\mathrm{Var}\left(\frac{\partial l_t^n}{\partial q_1}\ |_{\theta_0^1}\right)}}\ ,$$

with $l_t^n(\theta^1)$ log likelihood at t, associated to n observations.

LRT statistic at t

$$\Lambda_n(t) = 2\left\{l_t^n(\widehat{\theta_1}) - l_t^n(\widehat{\theta}_{1\,|H_0})\right\}\quad ,$$

with $\widehat{\theta_1}$ MLE, and $\widehat{\theta}_{1\,|H_0}$ MLE under H_0.

About the hypotheses studied

H_0 : “there is no QTL on $[0, T]$”

$H_{at^\star}$: “there are m QTL located at $t_1^\star$, ..., $t_m^\star$ with effects $q_1 = a_1/\sqrt{n}, \ldots, q_m = a_m/\sqrt{n}$ where $a_1 \neq 0, \ldots, a_m \neq 0$” .

Theorem

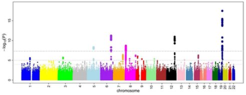
$$S_n(.) \Rightarrow Z(.) \quad , \quad \Lambda_n(.) \overset{F.d.}{\to} Z^2(.) \quad , \quad \sup \Lambda_n(.) \xrightarrow{\mathcal{L}} \sup Z^2(.)$$

- *$Z(.)$ is a Gaussian process perfectly known (i.e. the covariance function and the mean function are known)*

Introducing the SgenoLasso

1) we discretize the process at marker locations

$$\vec{S}_n = \vec{m}_{Z,t^\star} + \vec{\varepsilon} + o_P(1)$$



where $\vec{S}_n = (S_n(t_1)\ ,\ S_n(t_2)\ ,\ \dots\ ,\ S_n(t_K))'$

$$\vec{m}_{Z,t^\star} = (m_{Z,t^\star}(t_1)\ ,\ m_{Z,t^\star}(t_2)\ ,\ \dots, m_{Z,t^\star}(t_K))'$$

$$\vec{\varepsilon} \sim N(0,\Sigma) \text{ with } \Sigma_{kk'} = \mathrm{Cov}\,(Z(t_k), Z(t_{k'}))$$

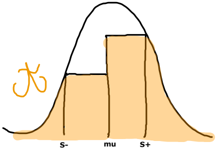
2) we decorrelate the process

Let $\mathbb{T}_K^\star := \{t_1^\star, \dots, t_m^\star\}$ and $\Sigma := BB'$, we have

$$B^{-1}\vec{S}_n = B'\Delta + B^{-1}\vec{\varepsilon} + o_P(1)$$

where $\Delta := (\Delta_1, \dots, \Delta_K)'$

$$\text{and} \quad \Delta_k = \begin{cases} 0 & \text{if} \quad t_k \notin \mathbb{T}_K^\star \\ \frac{a_s}{\sigma}\ \frac{\sqrt{\mathcal{A}}}{\sigma} & \text{if } t_k \in \mathbb{T}_K^\star \text{ with } s \mid t_s^\star = t_k \end{cases}$$



Introducing the SgenoLasso

In fact, non null Δ_k are unknown
$\Rightarrow$ L1 penalized regression Lasso (Tibshirani, 1996)

$$\hat{\Delta}_{\text{SgenoLasso}}(\lambda, \alpha) = \arg\min_{\Delta} \left(\left\| B^{-1}\vec{S}_n - B'\Delta \right\|_2^2 + \lambda \left\| \Delta \right\|_1 \right)$$

SgenoLasso presents all the properties of the classical Lasso !

Its β-min condition :

$$\min_{s|t_s^\star \in \mathbb{T}_K} \frac{|a_s|\sqrt{\mathcal{A}}}{\sigma^2\sqrt{K}} >> \Phi^{-2}\sqrt{\frac{m\log(K)}{K}}$$

Its irrepresentable condition :

$$\left\| \Sigma^{(\cdot,\star)}(\Sigma^{(\star,\star)})^{-1}\text{Sign}(a_1, \ldots, a_m) \right\|_\infty \leq C < 1$$

where $\|x\|_\infty = \max_j |x_j|$, $\text{Sign}(a_1, \ldots, a_m) = (\text{Sign}(a_1), \ldots, \text{Sign}(a_m))^\top$

β-min condition + irrep cond $\Rightarrow$ consistent variable selection

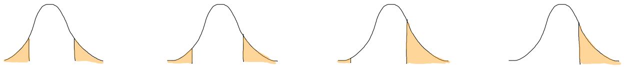
Applications to association studies (Simulated data)

10,000 markers on $[0, 10\text{M}]$ / 1,000 markers on $[0, 1\text{M}]$

16 QTLs located only on $[0, 1\text{M}]$

L1 ratio $\sum_{i=1}^{1000} |\hat{\Delta}_i| / \sum_{i=1}^{10000} |\hat{\Delta}_i|$

γ	γ^+/γ	SgenoLasso	Lasso	Group Lasso	EN	RaLasso
0.2	1/2	94.19%	91.69%	97.46%	97.44%	98.09%
	3/4	91.52%	84.75%	95.88%	96.02%	95.08%
	7/8	92.38%	75.46%	94.67%	95.23%	89.33%
	1	85.03%	21.14%	21.86%	27.37%	44.93%
0.3	1/2	91.62%	83.45%	92.87%	93.67%	95.36%
	3/4	90.88%	76.18%	89.59%	91.10%	91.13%
	7/8	86.22%	65.03%	78.00%	82.84%	80.32%
	1	78.00%	20.92%	20.82%	24.92%	48.25%



$\gamma^+/\gamma = 1/2$ 3/4 7/8 1

The SgenoLasso has several cousins

SgenoLasso is built on the L1 penalty of Lasso (Tibshirani, 1996)

$$\hat{\Delta}_{\text{SgenoLasso}}(\lambda, \alpha) = \arg\min_{\Delta} \left(\left\| B^{-1}\vec{S}_n - B'\Delta \right\|_2^2 + \lambda \left\| \Delta \right\|_1 \right)$$

SgenoElasticNet is built on the mixture of L1 and L2 penalties of Elastic Net (Zou and Hastie, 2005)

$$\hat{\Delta}_{\text{SgenoEN}}(\lambda, \alpha) = \arg\min_{\Delta} \left(\left\| B^{-1}\vec{S}_n - B'\Delta \right\|_2^2 + \frac{1-\alpha}{2} \left\| \Delta \right\|_2^2 + \alpha \left\| \Delta \right\|_1 \right)$$

SgenoGroupLasso is built on the Group Lasso penalty (Yuan and Lin, 2006)

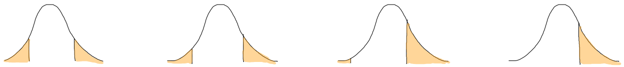
$$\hat{\Delta}_{\text{SgenoGroupLasso}}(\lambda) = \arg\min_{\Delta} \left(\left\| B^{-1}\vec{S}_n - B'\Delta \right\|_2^2 + \lambda \sum_{i=1}^{\text{nbGroup}} \sqrt{L_i} \left\| \vec{\Delta}_i \right\|_2 \right)$$

The SgenoLasso has several cousins

10,000 markers on [0, 10M] / 1,000 markers on [0, 1M]
16 QTLs located only on [0, 1M]

L1 ratio $\sum_{i=1}^{1000} |\hat{\Delta}_i| / \sum_{i=1}^{10000} |\hat{\Delta}_i|$

		SgenoLasso		SgenoGroupLasso		SgenoEN	
γ	γ^+/γ	L1 ratio	$\hat{m}$	L1 ratio	$\hat{m}$	L1 ratio	$\hat{m}$
0.2	1/2	94.19%	17.39	98.33%	24.9	96.03%	16.90
	3/4	91.52%	16.3	95.38%	24.3	92.59%	17.41
	7/8	92.38%	16.29	96.83%	24.6	93.19%	17.13
	1	85.03%	17.09	90.53%	22.8	84.93%	17.67
0.3	1/2	91.62%	17.55	92.35%	24.6	86.53%	17.87
	3/4	90.88%	17.59	94.84%	30.9	91.84%	15.43
	7/8	86.22%	16.82	89.96%	29.3	86.68%	17.30
	1	78.00%	17.28	82.61%	28.6	77.23%	17.89



$\gamma^+/\gamma = 1/2$ 3/4 7/8 1

Rice data

Data from Spindel et al. (2015) and Begum et al. (2015)

- Trait of interest : flowering date during the dry season 2012
- K =13,101 markers, randomly chosen by the authors from their 73,147 collected markers
- $n = 312$ in total (i.e. under complete genotyping)
- only 93 extreme individuals when $\gamma = 0.3$

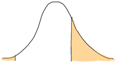
Rice data
(selective genotyping performed symmetrically)

γ	Method	Selected genes
1	Begum et al.	S3-1125848, S3-1165376, S3-1221494, S3-1269941, S3-1394477
0.3	SgenoLasso	4 genes matching those of Begum et al.
0.3	SgenoEN	5 genes matching . . .
0.3	SgenoGroupLasso	5 genes matching . . .
0.3	Lasso	2 genes matching . . .
0.3	EN	5 genes matching . . .
0.3	Group Lasso	3 genes matching . . .

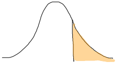
The predictive ability of the SgenoLasso (simulated data, 10000 markers)

Accuracy criterion Cor($\hat{y}, y$)

γ	γ^+/γ	SgenoLasso	Lasso	Group Lasso	EN	RaLasso
0.1	1	30.97%	6.49%	3.17%	4.38%	10.43%
	7/8	31.25%	30.55%	29.87%	29.74%	28.78%
0.2	1	27.88%	7.12%	4.05%	5.41%	11.08%
	7/8	28.26%	27.98%	27.86%	28.09%	26.28%
0.3	1	26.79%	9.02%	6.89%	7.48%	11.96%
	7/8	28.13%	27.85%	26.59%	28.25%	26.05%



$\gamma^+/\gamma = 7/8$



$\gamma^+/\gamma = 1$

Our answer to Brandariz and Bernardo (Crop Science, 2018) :
no need to keep the worst individuals in the breeding programs

Thank you for listening

A few references :

- S.P. Brandariz and R. Bernardo. *Maintaining the Accuracy of Genomewide Predictions when Selection Has Occurred in the Training Population*, Crop Science (2018)
- D. Darvasi, M. Soller, *Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus*, Theor. Appl. Genet. (1992).
- J. Fan, Q. Li, Y. Wang. *Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions*, Journal of the Royal Statistical Society : Series B (Statistical Methodology) (2017)
- E.S. Lander and D. Botstein. *Mapping mendelian factors underlying quantitative traits using RFLP linkage maps*, Genetics (1989)
- R.J. Lebowitz, M. Soller, J.S. Beckmann. *Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines*, Theor. Appl. Genet. (1987).
- C.E. Rabier. *On statistical inference for selective genotyping*, J. Stat. Plan. Infer. (2014)
- C.E. Rabier. *On stochastic processes for Quantitative Trait Locus mapping under selective genotyping*, Statistics (2015)
- C.E. Rabier and C. Delmas. *The SgenoLasso and its cousins for selective genotyping and extreme sampling*, Statistics (2021)
- R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society B (1996).
- M. Yuan, Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society Series B (2006).
- Y. Zhao, M. Gowda, F.H. Longin, T. Würschum, N. Ranc, J.C. Reif, *Impact of selective genotyping in the training population on accuracy and bias of genomic selection*, Theoretical and Applied Genetics (2012).
- H. Zou, T. Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society : Series B (Statistical Methodology), (2005).