The AdaptSgenoLasso, an extended version of the SgenoLasso, for gene mapping and for genomic prediction using the extremes

Charles-Elie Rabier, Céline Delmas

Hal
Vol. , No. , , 2–44

RESEARCH ARTICLE

The AdaptSgenoLasso, an extended version of the SgenoLasso, for gene mapping and for genomic prediction using the extremes

Charles-Elie Rabier [a*], Céline Delmas [b]

[a]*IMAG, Université de Montpellier, CNRS, France*; [b]*INRA, UR875 MIAT, F-313326 Castanet-Tolosan, France*;

(*Received 00 Month 200x; in final form 00 Month 200x*)

We introduce here the AdaptSgenoLasso, a new penalized likelihood method for gene mapping and for genomic prediction, which is an extended version of the SgenoLasso. The AdaptSgenoLasso relies on the original concept of a selective genotyping that varies along the genome. The "classical" selective genotyping on which the SgenoLasso is built on, consists in genotyping only extreme individuals, in order to increase the signal from genes. However, since the same amount of selection is applied at all genome locations, the signal is increased of the same proportional factor everywhere. With the AdaptSgenoLasso, we allow geneticists to impose more weights on some loci (i.e. locations) of interest, known to be responsible for the variation of the quantitative trait. The resulting signal is now dedicated to each locus. We propose here a deep theoretical study of the AdaptSgenoLasso, and we show on simulated data the superiority of this new approach over the SgenoLasso.

Keywords: Gaussian process, Selective Genotyping, Genomic Selection, High-Dimensional Linear Model, Variable Selection, Sparsity

AMS Subject Classification: Primary 60G15; 62F03; 62F05

1. Motivation

Nowadays, more and more genomic data are available thanks to advances in molecular biology and to technology. Genomics and mathematics, two fields not expanding at the same speed, are sometimes complementary. Old-fashioned tools, studied deeply by mathematicians, may be of importance for the genomic community. In this context, we introduced recently the SgenoLasso [53], a new variable selection method that relies on an old concept called selective genotyping [31, 32]. Our goal here is to present an even more powerful method than the SgenoLasso, and still inspired by selective genotyping.

To begin with, let us briefly recall the selective genotyping concept. In a seminal paper, [32] showed that the extreme (i.e. the highest or the lowest) observations of a given trait contain most of the signal on a Quantitative Trait Locus, so-called QTL. Roughly speaking, a QTL can be viewed as a gene influencing a quantitative trait. Then, the authors suggested to genotype only the individuals with extreme phenotypes (extreme observations). This concept was called selective genotyping and [31] formalized it later.

Today, applications fields of selective genotyping lie in Genome Wide Association Study (GWAS) and in Genomic Selection (GS).

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

The aim of GWAS is to find associations between loci (i.e. locations of the genome) and a trait of interest. In the literature, there are some recent association studies using selective genotyping in plants (e.g. sugarcane [23]; soybean [44, 63, 66]; chickpea [60]; tomatoes [42]), in animals (e.g. dairy cattle [30]; drosophila [8]; sow [17]; mouse [29]), and in humans (e.g. on Kashin-Beck disease [69]; on intelligence [68]). 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) [35], which is a very popular topic in genomics (e.g. strawberry, [22]; banana, [41]). The main goal of GS is to select individuals (i.e. candidates) by means of genomic predictions (see [47]). Since predictions can be performed as soon as the DNA is available, GS accelerates significantly the genetic gain. Indeed, we do not have to wait anymore to observe the phenotype of the candidate at adult age. With GS, after having performed genomic predictions, the best individuals are selected and are crossed to produce a new generation of offsprings. This process allows to consider many generations fastly.

GS is promising but new statistical tools are now required to exploit the potential of GS. In GS, the learning model has to be recalibrated over time, otherwise it leads to unreliable predictions (see [7, 40, 45]). Typically, after a large number of generations, we can not perform genomic predictions on the basis of a model learned on the first generations. As a consequence, it is crucial to update the model with the help of candidates selected at the previous step. In orther words, in order to recalibrate the model, the model has to be fitted on extreme individuals, which is highly linked to selective genotyping.

As mentioned before, we introduced recently the SgenoLasso [53], a new L1 penalized likelihood method able to handle extreme data, which is not the case of the famous Lasso [59]. However, the SgenoLasso presents the drawback of imposing the same weights on all loci, even when a few major genes are already known by geneticists. In this context, the aim of this present paper is to propose a new version of the SgenoLasso, called AdaptSgenoLasso, that allows to give more importance on some loci of interest. We will show that AdaptSgenoLasso enjoys better performances than SgenoLasso in terms of genomic prediction and in terms of GWAS.

2. Model

In this section, we recall the stochastic model studied in the SgenoLasso context, and we introduce new notation dedicated to the AdaptSgenoLasso.

As in our previous studies, we study a backcross population, $A \times (A \times B)$, where A and B are purely homozygous lines. The trait is observed on n individuals (progenies) and we denote by $Y_j,\ j = 1, ..., n$, these observations. The chromosome is represented by the segment $[0, T]$. The distance on $[0, T]$ is called the genetic distance, it is measured in Morgans (see for instance [65] or [56]). The genome $X(t)$ of one individual takes the value $+1$ if, for example, the "recombined chromosome" (due to meiosis) is originated from A at location t and takes the value -1 if it is originated from B. $X(0)$ is an equiprobable random sign and we consider Haldane modeling, which assumes no crossover interference. In other words, $N(.)$ is a standard Poisson process on $[0, T]$ that refers to the number of crossovers due to meiosis, and we have the relationship $X(t) = X(0)(-1)^{N(t)}$. Moreover, $r(t, t')$ will denote the probability of recombination between two loci located at t and t'.

Calculations on the Poisson distribution show that

$$r(t,t') := \mathrm{P}(X(t)X(t') = -1) = \mathrm{P}(\big|N(t) - N(t')\big| \text{ odd}) = \frac{1}{2}\left(1 - e^{-2|t-t'|}\right) .$$

We set in addition

$$\bar{r}(t,t') := 1 - r(t,t'), \;\; \rho(t,t') := e^{-2|t-t'|} \;\; .$$

We assume an "analysis of variance model" for the quantitative trait:

$$Y = \mu \; + \; \sum_{s=1}^{m} X(t_s^\star)\, q_s \; + \; \sigma\varepsilon \tag{1}$$

where μ is the global mean, ε is a Gaussian white noise independent of $X(.)$, σ^2 is the environmental variance, m is the number of QTLs, and q_s and $t_s^\star$ denote respectively the effect and the location of the sth QTL. Indeed, it is well known that there is a finite number of loci underlying the variation in quantitative traits ([27]). Besides, we will consider $0 < t_1^\star < ... < t_m^\star < T$.

Usually, in the problem of QTL mapping with a classical selective genotyping ([18, 32]), the "genome information" is available only at fixed locations $t_1 = 0 < t_2 < ... < t_K = T$, called genetic markers, and only if the trait is extreme. In order to describe this model more precisely, let us consider two real thresholds S_-^1 and S_+^1, with $S_-^1 \leq S_+^1$ and the random process $\overline{X}(.)$ such as $\overline{X}(t) := X(t)1_{Y \notin \left[S_-^1 \, , \, S_+^1\right]}$. Then, usually an observation is

$$\big(Y,\, \overline{X}(t_1),\, \overline{X}(t_2),\, ...,\, \overline{X}(t_K)\big)$$

and the challenge is that the number of QTLs m and their locations $t_1^\star$, ..., $t_m^\star$ are unknown.

The originality of our present study lies in the fact that we propose to focus here on a more sophisticated selective genotyping than the classical one ([18, 32]): our selective genotyping will vary along the genome. Note that in what follows, "genotyping" will refer to observing the genome information at markers.

In order to introduce our new selective genotyping, let us define two additional real thresholds S_-^2 and S_+^2 such as $S_-^1 \leq S_-^2 \leq S_+^2 \leq S_+^1$. As in the classical selective genotyping, we collect the genome information on the dense map (i.e. at all markers) if and only if the phenotype Y is extreme, that is to say $Y \leq S_-^1$ or $Y \geq S_+^1$. However, we also consider a sparser map containing only a few markers that belong to the original (dense) map, and we genotype, at these marker locations, the individuals for which $Y \leq S_-^2$ or $Y \geq S_+^2$. In other words, we collect the genome information of extra individuals at these markers. Intuitively, it enables to put more weights on some markers matching major genes that are well known by geneticists.

In what follows, $\mathrm{T}_K^1 := \{t_1, \ldots, t_K\}$ denotes the set of marker locations that belong to the dense map, and $\mathrm{T}_K^2 := \left\{t_{\sigma(1)}, t_{\sigma(2)}, \ldots, t_{\sigma(\#\mathrm{T}_K^2)}\right\}$ is a subset of T_K^1 (i.e. $\mathrm{T}_K^2 \subseteq \mathrm{T}_K^1$), representing the marker locations of the sparse map, where $\sigma(.)$ is a one-to-one map $\sigma : \left\{1, \ldots, \#\mathrm{T}_K^2\right\} \to \{1, \ldots, K\}$. Recall that the notation $\#$ stands for the cardinality of a set. To make the reading easier, we impose that $\sigma(k) < \sigma(k')$ for $k < k'$ and we assume $\sigma(1) = 1$ and $\sigma(\#\mathrm{T}_K^2) = K$, so that the markers located at 0 and at T are also located on the sparse map.

If we call $\widetilde{X}(t)$ the random variables such as $\widetilde{X}(t) = X(t)1_{Y \in \left[S_-^1 \, , \, S_-^2\right] \cup \left[S_+^2 \, , \, S_+^1\right]}$,

then, in our problem, one observation is now

$$\left(Y,\ \overline{X}(t_1),\ \overline{X}(t_2),\ ...,\ \overline{X}(t_K),\ \widetilde{X}(t_{\sigma(1)}), \widetilde{X}(t_{\sigma(2)}),\ ..., \widetilde{X}(t_{\sigma(\#\mathrm{T}_K^2)})\right).$$

In other words, with our notations,

- when $Y \notin \left[S_-^1\ ,\ S_+^1\right]$, we have $\overline{X}(t_1) = X(t_1),\ ..., \overline{X}(t_K) = X(t_K)$, which means that the genome information is known on the dense map T_K^1
- when $Y \in \left[S_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ S_+^1\right]$, we have $\widetilde{X}(t_{\sigma(1)}) = X(t_{\sigma(1)}), \widetilde{X}(t_{\sigma(2)}) = X(t_{\sigma(2)}),\ ..., \widetilde{X}(t_{\sigma(\#\mathrm{T}_K^2)}) = X(t_{\sigma(\#\mathrm{T}_K^2)})$, which means that the genome information is known only on the sparse map T_K^2
- when $Y \in \left[S_-^2\ ,\ S_+^2\right]$, we have $\overline{X}(t_1) = 0,\ ..., \overline{X}(t_K) = 0$, and $\widetilde{X}(t_{\sigma(1)}) = 0, \widetilde{X}(t_{\sigma(2)}) = 0,\ ..., \widetilde{X}(t_{\sigma(\#\mathrm{T}_K^2)}) = 0$, which means that the genome information is missing at all markers

We observe n observations $\left(Y_j,\ \overline{X}_j(t_1),\ \overline{X}_j(t_2),\ ...,\ \overline{X}_j(t_K),\ \widetilde{X}_j(t_{\sigma(1)}), \widetilde{X}_j(t_{\sigma(2)}),\ ..., \widetilde{X}_j(t_{\sigma(\#\mathrm{T}_K^2)})\right)$ independent and identically distributed (i.i.d.).

3. Outline

3.1. *Preliminaries*

Before detailing the roadmap of this paper, we have to recall the famous concept of Interval Mapping [31] on which our new method is built. Assuming that only one QTL lies on the genome (i.e. $m = 1$), the Interval Mapping consists in computing the Likelihood Ratio Test (LRT) at each location $t \in [0, T]$ of the null hypothesis of absence of QTL H_0:"$q_1 = 0$," against the alternative "$q_1 \neq 0$,". It leads to a LRT process and to a score process. These processes have been deeply studied in the past in the complete data situation where all the genotypes are known (e.g. [2, 5, 14–16]), and later in the selective genotyping framework [51, 52]. The supremum of these processes corresponds to the LRT on the whole genome, and the asymptotic distribution of the supremum of these processes is now well known. In this paper, as in [53], we propose to study mainly the asymptotic distribution of the LRT and score processes under the general alternative of m QTLs lying on the genome. It enables to look for multiple genes along the genome thanks to a variable selection method.

3.2. *Roadmap*

In Section 4, we present our main result, Theorem 4.1, that gives the asymptotic distribution of the score process and of the LRT process under the alternative hypothesis that there exist m QTLs located at $t_1^\star$, ..., $t_m^\star$ with effects q_1, ..., q_m. The score process converges in distribution to a Gaussian process described as an interpolation of two independent Gaussian processes $V_1(.)$ and $V_2(.)$. The processes $V_1(.)$ and $V_2(.)$ are linked to the dense map and to the sparse map, respectively. The distribution of the LRT statistic on the whole genome is asymptotically that of the maximum of the square of a function of these two interpolated processes. This result is more general than previous studies under selective genotyping [51, 52] and under the complete data situation (e.g. [2, 5, 16]).

Next, Theorem 4.2 gives the Asymptotic Relative Efficiency (ARE) with respect

to the complete data situation. The ARE depends on the QTLs effects and their locations, which is a different result from the one obtained for the classical selective genotyping in [52]. Furthermore, Lemma 4.3 and Lemma 4.4 give necessary conditions to overcome classical selective genotyping.

On the other hand, Corollary 5.1 tackles the complementary experiment that consists in genotyping on the dense map (T_K^1) the individuals for which $Y \in \left[S_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ S_+^1\right]$, and in genotyping on the sparse map (T_K^2) the individuals for which $Y \geq S_+^1$ or $Y \leq S_-^1$. We show that the experiment based on the extremes (i.e. our current study) is largely more efficient than the complementary experiment. It confirms that most of the signal is contained in extreme traits, as highlighted in many studies on selective genotyping [31, 32, 49].

Section 6 is devoted to the AdaptSgenoLasso, our new penalized likelihood method relying on results of Theorem 4.1. AdaptSgenoLasso allows to estimate the QTLs location, their effects and their number. Note that its ElasticNet cousin, called the AdaptSgenoEN, is also described. The link between the AdaptSgenoLasso (resp. AdaptSgenoEN) and the SgenoLasso (resp. SgenoEN, see [53]) is also established. Finally, we present the AdaptSgenoAdaptLasso, a new cousin inspired by the Adaptive Lasso [72]: it takes into account the prior knowledge of major genes within the weighted L1 penalization. In that sense, AdaptSgenoAdaptLasso combines the advantages of AdaptSgenoLasso and those of Adaptive Lasso.

For a deeper understanding, Section 7 investigates the asymptotic theory for the AdaptSgenoLasso under complete Linkage Disequilibrium (i.e. the m QTLs are located on some markers). In particular, we give the rate of convergence for prediction and we also study the consistency of the variable selection.

At the end of the manuscript, Section 8 proposes a simulation study. First, Section 8.1 focuses on the max test in Interval Mapping. In particular, we compare the power of the classical selective genotyping approach and our new approach where the selective genotyping varies along the genome. Last, Sections 8.2 and 8.3 are dedicated to association studies and to Genomic Selection, respectively. We will show the advantage of AdaptSgenoLasso and its cousins over the ancestor SgenoLasso.

4. Some theoretical results

In what follows, we consider values of t that are distinct of marker locations, i.e. $t \in [t_1, t_K] \backslash \mathrm{T}_K^1$. For $i = 1, 2$, we define $t^{\ell,i}$ and $t^{r,i}$ in the following way:

$$t^{\ell,i} = \sup\left\{t_k \in \mathrm{T}_K^i : t_k < t\right\} \quad , \quad t^{r,i} = \inf\left\{t_k \in \mathrm{T}_K^i : t < t_k\right\}. \tag{2}$$

In other words, depending on the map, t belongs to the "Marker interval" either $(t^{\ell,1}, t^{r,1})$ or $(t^{\ell,2}, t^{r,2})$.

4.1. *Score test and Likelihood Ratio Test (LRT) at a location t of the genome*

Let us consider the case $m = 1$ (i.e. one QTL located at $t_1^\star$), and let $\theta^1 = (q_1, \mu, \sigma)$ be the parameter of the model at t fixed. At a location $t \in [t_1, t_K] \backslash \mathrm{T}_K^1$, the likelihood of $\left(Y,\ \overline{X}(t^{\ell,1}),\ \widetilde{X}(t^{\ell,2}),\ \overline{X}(t^{r,1}),\ \widetilde{X}(t^{r,2})\right)$ with respect to the measure $\lambda \otimes N \otimes N \otimes$

$N \otimes N$, λ being the Lebesgue measure, N the counting measure on N, is :

$$\begin{aligned} L_t(\theta^1) = \Big[& p_1(t)\ f_{(\mu+q_1,\sigma)}(Y)1_{Y\notin[S^1_-,S^1_+]} + \{1-p_1(t)\}\ f_{(\mu-q_1,\sigma)}(Y)1_{Y\notin[S^1_-,S^1_+]} \\ & + p_2(t)\ f_{(\mu+q_1,\sigma)}(Y)1_{Y\in[S^1_-,S^2_-]\cup[S^2_+,S^1_+]} + \{1-p_2(t)\}\ f_{(\mu-q_1,\sigma)}(Y)1_{Y\in[S^1_-,S^2_-]\cup[S^2_+,S^1_+]} \\ & + \frac{1}{2}\ f_{(\mu+q_1,\sigma)}(Y)1_{Y\in[S^2_-,S^2_+]} + \frac{1}{2}\ f_{(\mu-q_1,\sigma)}(Y)1_{Y\in[S^2_-,S^2_+]} \Big]\ g(t) \end{aligned}$$

where $f_{(\mu,\sigma)}$ is the Gaussian density with parameters (μ,σ), $p_1(t)$ and $p_2(t)$ are the probabilities $\mathrm{P}\left\{X(t)=1 \mid X(t^{\ell,1}), X(t^{r,1})\right\}$ and $\mathrm{P}\left\{X(t)=1 \mid X(t^{\ell,2}), X(t^{r,2})\right\}$,

$$\begin{aligned} p_1(t)1_{Y\notin[S^1_-,S^1_+]} &= \mathrm{P}\left\{X(t)=1 \mid X(t^{\ell,1}), X(t^{r,1})\right\} 1_{Y\notin[S^1_-,S^1_+]} \\ &= Q^{1,1}_{t,1}\ 1_{\overline{X}(t^{\ell,1})=1}1_{\overline{X}(t^{r,1})=1} + Q^{1,-1}_{t,1}\ 1_{\overline{X}(t^{\ell,1})=1}1_{\overline{X}(t^{r,1})=-1} \\ &+ Q^{-1,1}_{t,1}\ 1_{\overline{X}(t^{\ell,1})=-1}1_{\overline{X}(t^{r,1})=1} + Q^{-1,-1}_{t,1}\ 1_{\overline{X}(t^{\ell,1})=-1}1_{\overline{X}(t^{r,1})=-1} \end{aligned}$$

and

$$\begin{aligned} p_2(t)1_{Y\in[S^1_-,S^2_-]\cup[S^2_+,S^1_+]} &= \mathrm{P}\left\{X(t)=1 \mid X(t^{\ell,2}), X(t^{r,2})\right\} 1_{Y\in[S^1_-,S^2_-]\cup[S^2_+,S^1_+]} \\ &= Q^{1,1}_{t,2}\ 1_{\widetilde{X}(t^{\ell,2})=1}1_{\widetilde{X}(t^{r,2})=1} + Q^{1,-1}_{t,2}\ 1_{\widetilde{X}(t^{\ell,2})=1}1_{\widetilde{X}(t^{r,2})=-1} \\ &+ Q^{-1,1}_{t,2}\ 1_{\widetilde{X}(t^{\ell,2})=-1}1_{\widetilde{X}(t^{r,2})=1} + Q^{-1,-1}_{t,2}\ 1_{\widetilde{X}(t^{\ell,2})=-1}1_{\widetilde{X}(t^{r,2})=-1} \end{aligned}$$

with for $i=1,2$

$$\begin{aligned} Q^{1,1}_{t,i} &= \frac{\bar{r}(t^{\ell,i},t)\ \bar{r}(t,t^{r,i})}{\bar{r}(t^{\ell,i},t^{r,i})} \quad , \quad Q^{1,-1}_{t,i} = \frac{\bar{r}(t^{\ell,i},t)\ r(t,t^{r,i})}{r(t^{\ell,i},t^{r,i})} \\ Q^{-1,1}_{t,i} &= \frac{r(t^{\ell,i},t)\ \bar{r}(t,t^{r,i})}{r(t^{\ell,i},t^{r,i})} \quad , \quad Q^{-1,-1}_{t,i} = \frac{r(t^{\ell,i},t)\ r(t,t^{r,i})}{\bar{r}(t^{\ell,i},t^{r,i})}. \end{aligned}$$

We have the relationships

$$Q^{-1,-1}_{t,i} = 1 - Q^{1,1}_{t,i} \quad \text{and} \quad Q^{-1,1}_{t,i} = 1 - Q^{1,-1}_{t,i}.$$

Besides, we have

$$\begin{aligned} g(t) = \mathrm{P}\left\{X(t^{\ell,1}), X(t^{r,1})\right\}\ 1_{Y\notin[S^1_-,S^1_+]} + \mathrm{P}\left\{X(t^{\ell,2}), X(t^{r,2})\right\}\ 1_{Y\in[S^1_-,S^2_-]\cup[S^2_+,S^1_+]} \\ + 1_{Y\in[S^2_-,S^2_+]} \end{aligned}$$

with

$$\begin{aligned} \mathrm{P}\left\{X(t^{\ell,1}), X(t^{r,1})\right\}\ 1_{Y\notin[S^1_-,S^1_+]} = \frac{1}{2}\Big\{ & \overline{r}(t^{\ell,1},t^{r,1})1_{\overline{X}(t^{\ell,1})\overline{X}(t^{r,1})=1} \\ & + r(t^{\ell,1},t^{r,1})1_{\overline{X}(t^{\ell,1})\overline{X}(t^{r,1})=-1} \Big\} \end{aligned}$$

and

$$\mathrm{P}\left\{X(t^{\ell,2}), X(t^{r,2})\right\}\ 1_{Y\in[S^1_-,S^2]\cup[S^2_+,S^1_+]} = \frac{1}{2}\left\{\overline{r}(t^{\ell,2},t^{r,2})1_{\widetilde{X}(t^{\ell,2})\widetilde{X}(t^{r,2})=1} + r(t^{\ell,2},t^{r,2})1_{\widetilde{X}(t^{\ell,2})\widetilde{X}(t^{r,2})=-1}\right\}\ .$$

Note that the true probability distribution is $L_{t_1^\star}(\theta^1)$. The score statistic of the hypothesis "$q_1 = 0$" at t, for n independent observations, is defined as

$$S_n(t) = \frac{\frac{\partial l_t^n}{\partial q_1}\ |_{\theta_0^1}}{\sqrt{\mathrm{Var}\left(\frac{\partial l_t^n}{\partial q_1}\ |_{\theta_0^1}\right)}}\ , \tag{3}$$

where l_t^n denotes the log likelihood at t, associated to n observations, and $\theta_0^1 = (0, \mu, \sigma)$ refers to the parameter θ_1 under $\mathcal{H}_0$. The likelihood ratio statistic at t will be defined as

$$\Lambda_n(t) = 2\big[l_t^n(\widehat{\theta}_1) - l_t^n(\widehat{\theta}_{1|H_0}\big],$$

on n independent observations.

4.2. *Main result on the score and LRT processes*

Let us define $\forall i = 1,2,\ \xi_i(t) := \sqrt{\alpha_i^2(t) + \beta_i^2(t) + 2\alpha_i(t)\beta_i(t)\rho(t^{\ell,i}, t^{r,i})}$ where $\alpha_i(t) := Q_{t,i}^{1,1} - Q_{t,i}^{-1,1}$ and $\beta_i(t) := Q_{t,i}^{1,1} - Q_{t,i}^{1,-1}$. By continuity, we have

$$\forall t_k \in \mathrm{T}_K^1\ \ \xi_1(t_k) = 1, \alpha_1(t_k) = 1, \beta_1(t_k) = 0$$
$$\forall t_k \in \mathrm{T}_K^2\ \ \xi_2(t_k) = 1, \alpha_2(t_k) = 1, \beta_2(t_k) = 0.$$

Before giving our first main result, let us define the following quantities:

$$\gamma_1 := \mathrm{P}_{\mathcal{H}_0}\left(Y \notin \left[S_-^1,\ S_+^1\right]\right)\ ,\ \gamma_1^+ := \mathrm{P}_{\mathcal{H}_0}\left(Y > S_+^1\right)\ ,\ \gamma_1^- := \mathrm{P}_{\mathcal{H}_0}\left(Y < S_-^1\right)\ , \tag{4}$$

$$\gamma := \mathrm{P}_{\mathcal{H}_0}\left(Y \notin \left[S_-^2,\ S_+^2\right]\right)\ ,\ \gamma^+ := \mathrm{P}_{\mathcal{H}_0}\left(Y > S_+^2\right)\ ,\ \gamma^- := \mathrm{P}_{\mathcal{H}_0}\left(Y < S_-^2\right)\ , \tag{5}$$

$$\mathcal{A}_1 := \sigma^2\left\{\gamma_1 + z_{\gamma_1^+}\varphi(z_{\gamma_1^+}) - z_{1-\gamma_1^-}\varphi(z_{1-\gamma_1^-})\right\}\ , \tag{6}$$

$$\mathcal{B} := \sigma^2\left\{\gamma + z_{\gamma^+}\varphi(z_{\gamma^+}) - z_{1-\gamma^-}\varphi(z_{1-\gamma^-})\right\}\ , \tag{7}$$

$$\mathcal{A}_2 := \mathcal{B} - \mathcal{A}_1, \tag{8}$$

where $\varphi(x)$ and z_α denote respectively the density of a standard normal distribution taken at the point x, and the quantile of order $1-\alpha$ of a standard normal distribution.

Remark 1 : According to the law of large numbers, under the null hypothesis H_0 and under contiguous alternatives, $\frac{1}{n}\sum 1_{Y_j\notin\left[S_-^1,\ S_+^1\right]} \to \gamma_1$ and $\frac{1}{n}\sum 1_{Y_j\notin\left[S_-^2,\ S_+^2\right]} \to \gamma$. So, γ_1 (resp. γ) corresponds asymptotically to the percentage of individuals for which the genome information is collected on the dense map (resp. sparse). In other words, for a location t_k belonging exclusively to the dense map (i.e. $t_k \in \mathrm{T}_K^1 \backslash \mathrm{T}_K^2$), γ_1 is asymptotically the percentage of genotyped individuals and γ_1^+ (resp. γ_1^-) is

asymptotically the percentage of individuals genotyped with the largest (resp. the smallest) phenotypes.

Our main result is the following:

Theorem 4.1 : *Suppose that the parameters* $(q_1, ..., q_m, \mu, \sigma^2)$ *vary in a compact and that* σ^2 *is bounded away from zero, and also that* m *is finite. Let* $\mathcal{H}_0$ *be the null hypothesis of no QTL on* $[0, T]$*, and let define the following local alternatives* $\mathcal{H}_{a\vec{t}^\star}$*: "there are* m *QTLs located respectively at* $t_1^\star, \cdots, t_m^\star$ *with effect* $q_1 = a_1/\sqrt{n}, \cdots, q_m = a_m/\sqrt{n}$ *where* $a_1 \neq 0, \cdots, a_m \neq 0$*". Then, as* n *tends to infinity,*

$$S_n(.) \Rightarrow Z(.) \ , \ \Lambda_n(.) \overset{F.d.}{\rightarrow} Z^2(.) \ , \ \sup \Lambda_n(.) \xrightarrow{\mathcal{L}} \sup Z^2(.) \tag{9}$$

under $\mathcal{H}_0$ *and* $\mathcal{H}_{a\vec{t}^\star}$*, where* $\Rightarrow$ *and F.d. denote the weak convergence and the convergence of finite-dimensional distributions respectively and where* $Z(.)$ *is the Gaussian process with unit variance such as* $\forall t \in [t_1, t_K] \backslash \mathrm{T}_K^1$ *:*

$$Z(t) = \frac{\sqrt{\mathcal{A}_1}\ \xi_1(t) V_1(t) \ + \ \sqrt{\mathcal{A}_2}\ \xi_2(t) V_2(t)}{\sqrt{\mathcal{A}_1\ \xi_1^2(t) + \mathcal{A}_2\ \xi_2^2(t)}} \ .$$

$V_1(.)$ *et* $V_2(.)$ *are independent Gaussian processes with unit variance such as*

$$\forall i = 1, 2 \ \ V_i(t) = \left\{ \alpha_i(t)\ V_i(t^{\ell,i}) \ + \ \beta_i(t)\ V_i(t^{r,i}) \right\} / \xi_i(t)$$

$$\forall (t_k, t_{k'}) \in \mathrm{T}_K^i \times \mathrm{T}_K^i \ \ Cov\left(V_i(t_k), V_i(t_{k'})\right) = \rho(t_k, t_{k'}) \ .$$

The mean function of $Z(.)$ *is such that:*

- *under* $\mathcal{H}_0$, $m_{Z,\vec{t}^\star}(t) = 0$
- *under* $\mathcal{H}_{a\vec{t}^\star}$,

$$m_{Z,\vec{t}^\star}(t) = \frac{\sqrt{\mathcal{A}_1}\ \xi_1(t) m_{V_1,\vec{t}^\star}(t) \ + \ \sqrt{\mathcal{A}_2}\ \xi_2(t) m_{V_2,\vec{t}^\star}(t)}{\sqrt{\mathcal{A}_1\ \xi_1^2(t) + \mathcal{A}_2\ \xi_2^2(t)}}.$$

where

$$\forall i = 1, 2 \ \ m_{V_i,\vec{t}^\star}(t) = \left\{ \alpha_i(t)\ m_{V_i,\vec{t}^\star}(t^{\ell,i}) + \beta_i(t)\ m_{V_i,\vec{t}^\star}(t^{r,i}) \right\} / \xi_i(t)$$

$$\forall t_k \in \mathrm{T}_K^i \ \ m_{V_i,\vec{t}^\star}(t_k) = \frac{\sqrt{\mathcal{A}_i}}{\sigma^2} \sum_{s=1}^{m} a_s\ \rho(t_s^\star, t_k) \ .$$

The proof is given in Section 10.

According to Theorem 4.1, the score process $S_n(.)$ converges weakly to an interpolated process $Z(.)$ that contains two components: the process $V_1(.)$ that relies on the dense map (i.e. T_K^1), and the process $V_2(.)$ that relies on the sparse map (i.e. T_K^2). This result is more general than previous studies under the complete data situation (e.g. [2, 5, 16]), and under selective genotyping [51, 52]. Indeed, in all these previous studies, the limiting process was an interpolated process based only on one component. In our present study, the limiting process Z(.) is an interpolation between two interpolated processes $V_1(.)$ and $V_2(.)$.

From Theorem 4.1, we can easily recover results present in the literature. For instance, when $S_-^1 = S_-^2$ and $S_+^1 = S_+^2$, we have $\mathcal{A}_2 = 0$ and the process $Z(.)$

contains only one component, the process $V_1(.)$ that matches the process $V(.)$ of [52]. In other words, results from Theorem 4.1 are consistent with the ones obtained under the classical selective genotyping situation with parameter $\mathcal{A}_1$ only (i.e. with only the thresholds S_-^1 and S_+^1) and using the dense map as genetic map.

In the same way, when $S_-^1 = -\infty$ and $S_+^1 = +\infty$, we have $\mathcal{A}_1 = 0$ and the process $Z(.)$ matches the process $V_2(.)$. In that case, the process $V_2(.)$ matches the $V(.)$ of [52], as soon as we consider a classical selective genotyping with parameter $\mathcal{A}_2$ only (i.e. with only the thresholds S_-^2 and S_+^2) and using the sparse map as genetic map.

In what follows, when not specified, the classical selective genotyping will denote the framework with parameter $\mathcal{A}_1$ and relying on the dense map.

4.3. *About the skeleton of the limiting process $Z(.)$*

Since our variable selection method (cf. Section 6) will be based on the skeleton of the limiting process $Z(.)$, let us describe here this skeleton. By continuity, it is easy to see that when t_k belongs to T_K^2:

$$Z(t_k) = \frac{\sqrt{\mathcal{A}_1}\ V_1(t_k)\ +\ \sqrt{\mathcal{A}_2}\ V_2(t_k)}{\sqrt{\mathcal{B}}}\ , \tag{10}$$

$$m_{Z,\vec{t}^\star}(t_k) = \frac{\sqrt{\mathcal{B}}}{\sigma^2}\sum_{s=1}^{m}\rho(t_s^\star, t_k)a_s\ .$$

However, at a location t_k that belongs to $\mathrm{T}_K^1 \backslash \mathrm{T}_K^2$:

$$Z(t_k) = \frac{\sqrt{\mathcal{A}_1}\ V_1(t_k)\ +\ \sqrt{\mathcal{A}_2}\ \left\{\alpha_2(t_k)V_2(t_k^{\ell,2}) + \beta_2(t_k)V_2(t_k^{r,2})\right\}}{\sqrt{\mathcal{A}_1\ + \mathcal{A}_2\ \xi_2^2(t_k)}}\ , \tag{11}$$

$$m_{Z,\vec{t}^\star}(t_k) = \frac{\frac{\mathcal{A}_1}{\sigma^2}\sum_{s=1}^{m}\rho(t_s^\star, t_k)a_s}{\sqrt{\mathcal{A}_1 + \mathcal{A}_2\ \xi_2^2(t_k)}} + \frac{\frac{\mathcal{A}_2}{\sigma^2}\left\{\alpha_2(t_k)\sum_{s=1}^{m}\rho(t_s^\star, t_k^{\ell,2})a_s + \beta_2(t_k)\sum_{s=1}^{m}\rho(t_s^\star, t_k^{r,2})a_s\right\}}{\sqrt{\mathcal{A}_1 + \mathcal{A}_2\ \xi_2^2(t_k)}}$$

where $t_k^{\ell,2}$ and $t_k^{r,2}$ are defined according to formula (2), using a small abuse of notation.

Using formulae (11) and (10), we can easily compute the skeleton of the covari-

ance function of $Z(.)$:

$$\forall (t_k, t_{k'}) \in \mathrm{T}_K^2 \times \mathrm{T}_K^2 \quad \mathrm{Cov}\left(Z(t_k), Z(t_{k'})\right) = \rho(t_k, t_{k'}) \ , \tag{12}$$

$$\forall (t_k, t_{k'}) \in \mathrm{T}_K^1 \backslash \mathrm{T}_K^2 \times \mathrm{T}_K^1 \backslash \mathrm{T}_K^2$$

$$\mathrm{Cov}\left(Z(t_k), Z(t_{k'})\right) = \frac{\mathcal{A}_1 \rho(t_k, t_{k'}) + \mathcal{A}_2 \left\{\alpha_2(t_k)\rho(t_k^{\ell,2}, t_{k'}) + \beta_2(t_k)\rho(t_k^{r,2}, t_{k'})\right\}}{\sqrt{\left\{\mathcal{A}_1 + \mathcal{A}_2 \xi_2^2(t_k)\right\}\left\{\mathcal{A}_1 + \mathcal{A}_2 \xi_2^2(t_{k'})\right\}}} \ , \tag{13}$$

$$\forall (t_k, t_{k'}) \in \mathrm{T}_K^2 \times \mathrm{T}_K^1 \backslash \mathrm{T}_K^2 \quad \mathrm{Cov}\left(Z(t_k), Z(t_{k'})\right) = \frac{\sqrt{\mathcal{B}}\rho(t_k, t_{k'})}{\sqrt{\mathcal{A}_1 + \mathcal{A}_2 \xi_2^2(t_{k'})}} \ . \tag{14}$$

The proof is given in Section 11.

4.4. *Asymptotic Relative Efficiency*

Let us consider one location of the genome and let us focus on the Asymptotic Relative Efficiency (ARE). Recall that the ARE determines the relative sample size required to obtain the same local asymptotic power as the one of the test under the complete data situation where the genome information is known at all markers without taking into account whether the trait Y is extreme or not.

In other words, under the complete data situation, we have $S_-^1 = S_-^2 = S_+^2 = S_+^1$, so that $\gamma = \gamma_1 = 1$, $\mathcal{A}_1 = \mathcal{B} = \sigma^2$ and $\mathcal{A}_2 = 0$. Note also that the complete data situation is the one studied in [2].

Theorem 4.2 : *Let κ denote the ARE, then we have*

i) at a location $t \notin \mathrm{T}_K^1$, $\kappa = \dfrac{\sigma^2\ \Omega^2\ \xi_1^2(t)}{\left\{\alpha_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{\ell,1})a_s + \beta_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{r,1})a_s\right\}^2}$

where

$$\Omega = \frac{\sum_{i=1}^2 \mathcal{A}_i \left\{\alpha_i(t)\sum_{s=1}^m \rho(t_s^\star, t^{\ell,i})a_s + \beta_i(t)\sum_{s=1}^m \rho(t_s^\star, t^{r,i})a_s\right\}}{\sigma^2\sqrt{\mathcal{A}_1\ \xi_1^2(t) + \mathcal{A}_2\ \xi_2^2(t)}}$$

ii) at a location $t_k \in \mathrm{T}_K^1 \backslash \mathrm{T}_K^2$, $\kappa = \dfrac{\sigma^2\ \Omega'^{\ 2}}{\left\{\sum_{s=1}^m \rho(t_s^\star, t_k)a_s\right\}^2}$

$$\text{where } \Omega' = \frac{\mathcal{A}_1 \left\{\sum_{s=1}^m \rho(t_s^\star, t_k)a_s\right\}}{\sigma^2\sqrt{\mathcal{A}_1 + \mathcal{A}_2\ \xi_2^2(t_k)}} + \frac{\mathcal{A}_2 \left\{\alpha_2(t_k)\sum_{s=1}^m \rho(t_s^\star, t_k^{\ell,2})a_s + \beta_2(t_k)\sum_{s=1}^m \rho(t_s^\star, t_k^{r,2})a_s\right\}}{\sigma^2\sqrt{\mathcal{A}_1 + \mathcal{A}_2\ \xi_2^2(t_k)}}$$

iii) at a location $t_k \in \mathrm{T}_K^2$, $\kappa = \mathcal{B}/\sigma^2$.

The proof is given in Section 12. Note that ii) and iii) can be obtained from i) by continuity.

According to Theorem 4.2, when the selective genotyping varies along the genome, the ARE depends on the QTLs effects and their locations. This result is different from the one obtained regarding the classical selective genotyping (i.e.

$S^1_- = S^2_-$ and $S^1_+ = S^2_+$), for which the ARE depends only on the factor $\mathcal{A}_1$ linked to the selection intensity (see Theorem 4.2 of [52]).

The situation iii), i.e. $t_k \in \mathrm{T}^2_K$, can be viewed as a classical selective genotyping situation at one marker of the sparse map, since all the individuals with phenotypes smaller than S^2_- or greater than S^2_+ are genotyped at t_k. As a consequence, in this case, the ARE does not depend on the QTL parameters, and matches exactly the ARE presented in Theorem 1 of [49] with parameter $\mathcal{B}$.

Last, when all the QTLs do not belong to the interval $[t^{\ell,2}, t^{r,2}]$ (i.e. $\forall s\ t^\star_s \notin [t^{\ell,2}, t^{r,2}]$), we have the relationships $\forall i = 1, 2$, $\alpha_i(t)\rho(t^\star_s, t^{\ell,i})a_s + \beta_i(t)\rho(t^\star_s, t^{r,i})a_s = \rho(t^\star_s, t)a_s$. As a result, the efficiencies i) and ii) have the following expressions: i) $\kappa = \frac{\mathcal{B}^2\xi_1^2(t)}{\sigma^2\{\mathcal{A}_1\xi_1^2(t)+\mathcal{A}_2\xi_2^2(t)\}}$ and ii) $\kappa = \frac{\mathcal{B}^2}{\sigma^2\{\mathcal{A}_1+\mathcal{A}_2\xi_2^2(t_k)\}}$. In this case, the ARE does not depend on the QTLs effects and their locations. The ARE depends only on the factors $\mathcal{A}_1$ and $\mathcal{B}$, and on the tested location.

Figures 1 and 2 illustrate the efficiency κ, given in expression i) of Theorem 4.2, as a function of γ_1, and as a function of the ratios γ_1^+/γ_1 and γ^+/γ. Note that in order to concentrate on the same kind of selective genotyping on both maps, we considered the relationship $\gamma_+/\gamma = \gamma_1^+/\gamma_1$ in all cases. Different values for γ are studied: γ takes either the value 0.3, 0.5 or 1. Only one QTL is considered ($m = 1$) located at $t^\star_1 = 0.85$, and the test is performed exactly at the QTL location ($t = t^\star_1$). As a consequence, we will focus only on the markers flanking the QTL location. The constant a linked to the QTL effect is set to the value 2. The dense map is such as $t^{\ell,1} = 0.80$ and $t^{r,1} = 0.90$, and two scenarios are investigated for the sparse map that targets a few loci: either map a) $t^{\ell,2} = 0.20$ and $t^{r,2} = 1.50$, or map b) $t^{\ell,2} = 0.70$ and $t^{r,2} = 1$.

According to Figures 1 and 2, for a given value of γ, the efficiency increases much more for sparse map a) as compared to sparse map b), when γ_1 increases. It was expected since on sparse map a), markers and the QTL are far apart. When γ_1 increases, more and more individuals are genotyped at markers of the dense map, and since these markers are closer to the QTL location, it helps for the statistical test. In contrast, on sparse map b), markers are already close to the QTL location and the dense map is not as useful as previously.

Figure 3 focuses on the opposite scenario: the value of γ_1 is set to 0.3, and we let the parameter γ vary. We can observe that when γ increases, the gain in terms of power is now more substancial on sparse map b) than on sparse map a). This result was expected in view of the previous experiment.

Remark 2 : According to the figures, the efficiencies reached their maximum for $\gamma_1^+/\gamma_1 = 1/2$ and $\gamma^+/\gamma = 1/2$. In Section 12, we prove that these points are indeed zeros of the efficiency's derivative. However, other "zeros" do exist (e.g. unidirectional selective genotyping, $\gamma_1^+/\gamma_1 = 1$ and $\gamma^+/\gamma = 1$) and the optimal setting seems to highly rely on the different parameter values. Nevertheless, on simulated data, the symetrical selective genotyping was found to be the optimal setting (see Table 1).

4.5. *Conditions required to overcome classical selective genotyping*

Let us assume that phenotyping is free and let us incorporate the number of markers into account in our mathematical treatment. Indeed, the new version of the selective genotyping will be of particular interest as soon as we observe a decrease in terms of

Table 1. Comparison in terms of power between the classical selective genotyping approach and the new approach where the selective genotyping varies along the genome ($T = 1$, markers are located every 1cM on the dense map, and every 25cM on the sparse map respectively). The analysis relies on the test statistic $\sup \Lambda_n(.)$ and on 10,000 paths for the theoretical power ($+\infty$), and 1,000 samples of size n for the empirical power. The power is computed as a function of the ratio γ_+/γ ($\gamma = 0.5$, $\gamma_1 = 0.3$, $\gamma_+/\gamma = \gamma_1^+/\gamma_1$), the sample size n, and the number m of QTLs. In all cases $|a_s| = 2.828$, + refers to positive effect, − refers to negative effect. The different QTL frameworks are the following: ($m = 1$, $t_1^\star = 0.03$), ($m = 2$, $t_1^\star = 0.03$, $t_2^\star = 0.55$), ($m = 3$, $t_1^\star = 0.03$, $t_2^\star = 0.55$, $t_3^\star = 0.80$).

						QTL number			
γ^+/γ	Method	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_2/(\mathcal{A}_1+\mathcal{A}_2)$	n	1(+)	2(++)	2(+-)	3(+-+)
1/2	**selective genotyping that varies along the genome**	0.7833	0.1454	15.65%	$+\infty$	58.55%	98.93%	38.17%	46.69%
					1,000	57.26%	96.53%	36.49%	45.71%
					200	54.20%	95.82%	33.40%	43.03%
					100	51.32%	94.90%	29.22%	38.08%
1/2	**classical selective genotyping**	0.7833	0	0%	$+\infty$	48.09%	93.65%	33.21%	40.83%
					1,000	47.53%	93.68%	32.03%	39.36%
					200	44.70%	91.76%	27.58%	35.08%
					100	40.37%	89.54%	23.47%	30.20%
1/4	**selective genotyping that varies along the genome**	0.7303	0.1273	14.84%	$+\infty$	53.28%	95.19%	35.62%	42.52%
					1,000	52.59%	95.20%	34.04%	41.44%
					200	49.51%	93.84%	30.23%	36.67%
					100	45.04%	91.68%	28.35%	33.58%
1/4	**classical selective genotyping**	0.7303	0	0%	$+\infty$	45.94%	91.68%	30.92%	38.11%
					1,000	45.89%	91.41%	29.52%	37.45%
					200	41.33%	89.31%	26.26%	32.11%
					100	36.67%	85.81%	21.57%	27.91%
1	**selective genotyping that varies along the genome**	0.4823	0.0177	3.54%	$+\infty$	30.64%	78.69%	20.14%	24.99%
					1,000	30.46%	77.65%	20.02%	24.30%
					200	27.04%	72.19%	16.45%	22.07%
					100	22.09%	66.16%	13.01%	18.31%
1	**classical selective genotyping**	0.4823	0	0	$+\infty$	32.61%	77.28%	21.41%	26.10%
					1,000	32.18%	77.60%	21.20%	25.82%
					200	27.75%	72.03%	17.57%	22.07%
					100	22.74%	65.46%	12.28%	18.31%

genotyped individuals. In this context, let us present two lemmas. In what follows, κ is the efficiency described in Theorem 4.2.

Lemma 4.3: *The selective genotyping that varies along the genome is more rewarding than the complete data situation ([2]), as soon as we have the relationship*

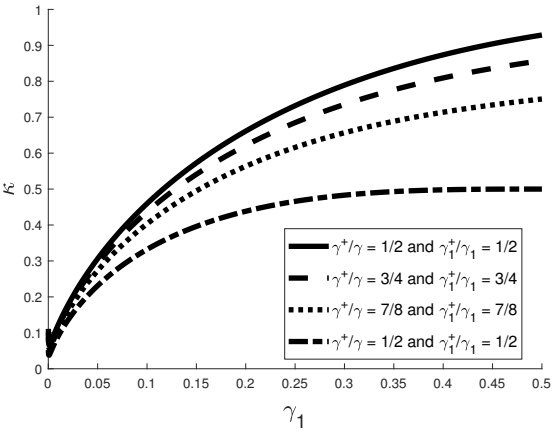
$$\kappa > \gamma_1 + (\gamma - \gamma_1)\frac{\#\mathrm{T}_K^2}{K} \ .$$

Lemma 4.4: *The selective genotyping that varies along the genome is more rewarding than the classical selective genotyping as soon as we have the relationship*

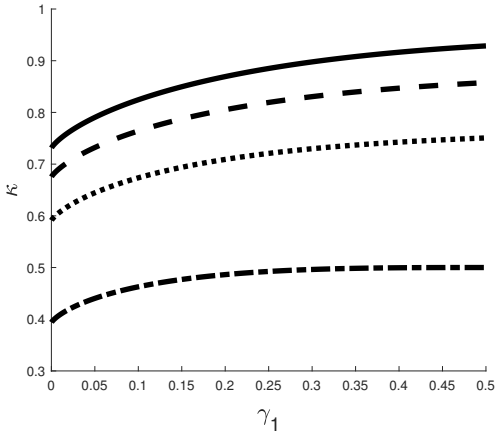
$$\Leftrightarrow \kappa > \left\{1 + \frac{z_{\gamma_1^+}\varphi(z_{\gamma_1^+}) - z_{1-\gamma_1^-}\varphi(z_{1-\gamma_1^-})}{\gamma_1}\right\}\left\{\gamma_1 + \frac{(\gamma-\gamma_1)\#\mathrm{T}_K^2}{K}\right\} \ .$$

The proofs are given in Section 13. In order to illustrate Lemma 4.4, Figure 4 proposes a comparison in terms of efficiency, between the classical selective genotyping and the selective genotyping that varies along the genome. Efficiencies with

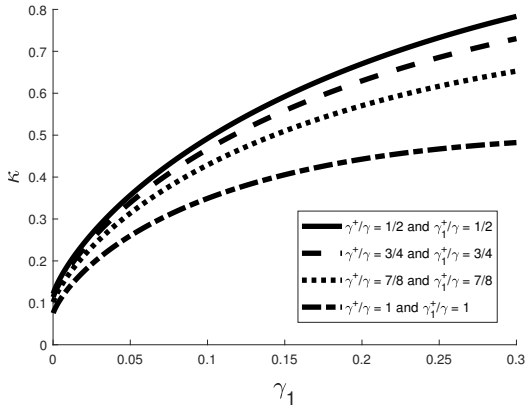
Figure 1. Efficiency κ as a function of γ_1, and as a function of the ratios γ_1^+/γ_1 and γ^+/γ. γ takes either the value 0.5 or 0.3. Only one QTL is considered ($m = 1$, $a = 2$, $\sigma = 1$) and the test is performed exactly at the QTL location ($t = t_1^\star = 0.85$). Two different sparse maps are considered, and as a dense map, we considered $t^{\ell,1} = 0.80$ and $t^{r,1} = 0.90$ in all cases.



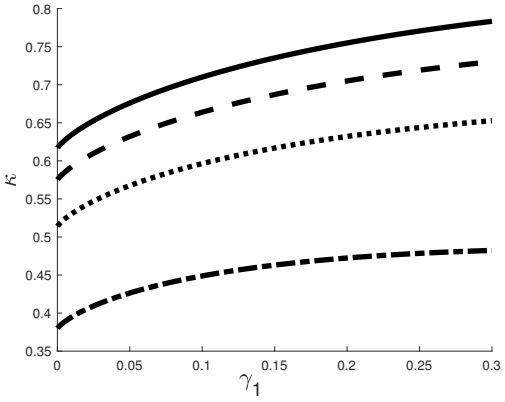
$\gamma = 0.5$
Sparse map a (i.e. $t^{\ell,2} = 0.20$, $t^{r,2} = 1.50$)



$\gamma = 0.5$
Sparse map b (i.e. $t^{\ell,2} = 0.70$, $t^{r,2} = 1$)



$\gamma = 0.3$
Sparse map a (i.e. $t^{\ell,2} = 0.20$, $t^{r,2} = 1.50$)



$\gamma = 0.3$
Sparse map b (i.e. $t^{\ell,2} = 0.70$, $t^{r,2} = 1$)

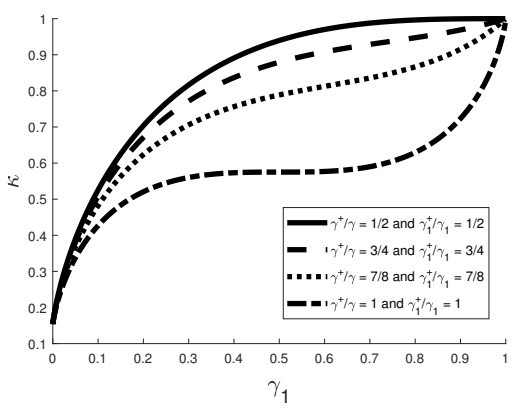
respect to the complete data situation ([2]), are illustrated as a function of γ, and as a function of the ratios γ_1^+/γ_1 and γ^+/γ. On Figure 4, is also represented the lower bound introduced in Lemma 4.4, considering as sparse map, either 5% or 10% of all markers (i.e. $\#\mathrm{T}_K^2/K = 5\%$ or 10%).

In all cases, γ_1 was set to the value 0.3, largely used in the genetic community. Indeed, this frequency has been proved to be optimal for selective genotyping experiments (cf. [18, 49]). Furthermore, we consider the same framework as in Figure 1: only one QTL is considered ($m = 1$, $a = 2$, $\sigma = 1$) and the test is performed exactly at the QTL location ($t = t_1^\star = 0.85$). The dense map is such as $t^{\ell,1} = 0.80$, $t^{r,1} = 0.90$ whereas $t^{\ell,2} = 0.70$, $t^{r,2} = 1$ for the sparse map.

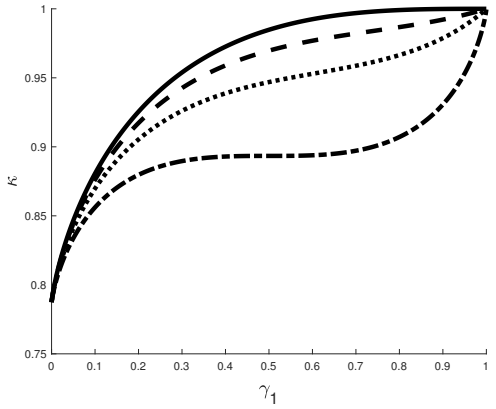
According to Figure 4, when we genotype symmetrically ($\gamma^+/\gamma = \gamma_1^+/\gamma_1 = 1/2$), our new approach is largely more rewarding than the classical genotyping, in most of cases. Indeed, the two bounds (5% or 10%) are almost always located below the efficiency curve of the selective genotyping that varies. Our new method becomes less relevant only when $\frac{\#\mathrm{T}_K^2}{K} = 10\%$ and $\gamma > 0.95$.

Note that for $\gamma^+/\gamma = \gamma_1^+/\gamma_1 = 3/4$ and $7/8$, the selective genotyping that varies was always found to be the best approach. Last, surprisingly, when the selective

Figure 2. Efficiency κ as a function of γ_1, and as a function of the ratios γ_1^+/γ_1 and γ^+/γ. In all cases, γ takes the value 1, only one QTL is considered ($m = 1$, $a = 2$, $\sigma = 1$), and the test is performed exactly at the QTL location ($t = t_1^\star = 0.85$). Two different sparse maps are considered, and as a dense map, we considered $t^{\ell,1} = 0.80$, $t^{r,1} = 0.90$, in all cases.

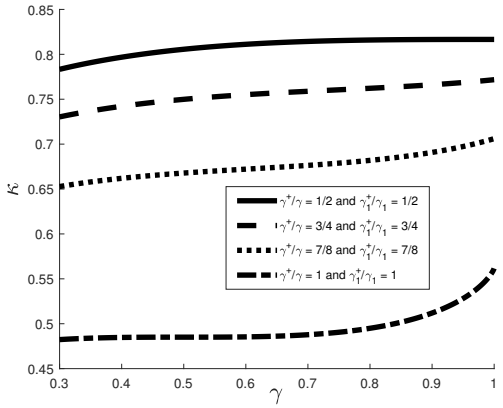


Sparse map a (i.e. $t^{\ell,2} = 0.20$, $t^{r,2} = 1.50$)

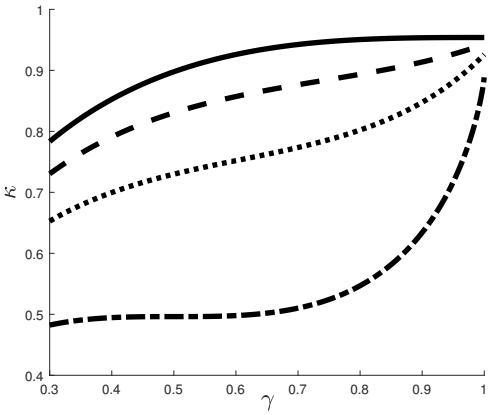


Sparse map b (i.e. $t^{\ell,2} = 0.70$, $t^{r,2} = 1$)

Figure 3. Efficiency κ as a function of γ, and as a function of the ratios γ_1^+/γ_1 and γ^+/γ. In all cases, γ_1 takes the value 0.3, only one QTL is considered ($m = 1$, $a = 2$, $\sigma = 1$) and the test is performed exactly at the QTL location ($t = t_1^\star = 0.85$). Two different sparse maps are considered, and as a dense map, we considered $t^{\ell,1} = 0.80$, $t^{r,1} = 0.90$, in all cases.



Sparse map a (i.e. $t^{\ell,2} = 0.20$, $t^{r,2} = 1.50$)



Sparse map b (i.e. $t^{\ell,2} = 0.70$, $t^{r,2} = 1$)

genotyping is performed unilaterally ($\gamma^+/\gamma = \gamma_1^+/\gamma_1 = 1$), we should choose the classical selective genotyping in some cases (e.g. $0.3 < \gamma < 0.85$ if $\frac{\#\mathrm{T}_K^2}{K} = 10\%$).

5. The complementary experiment

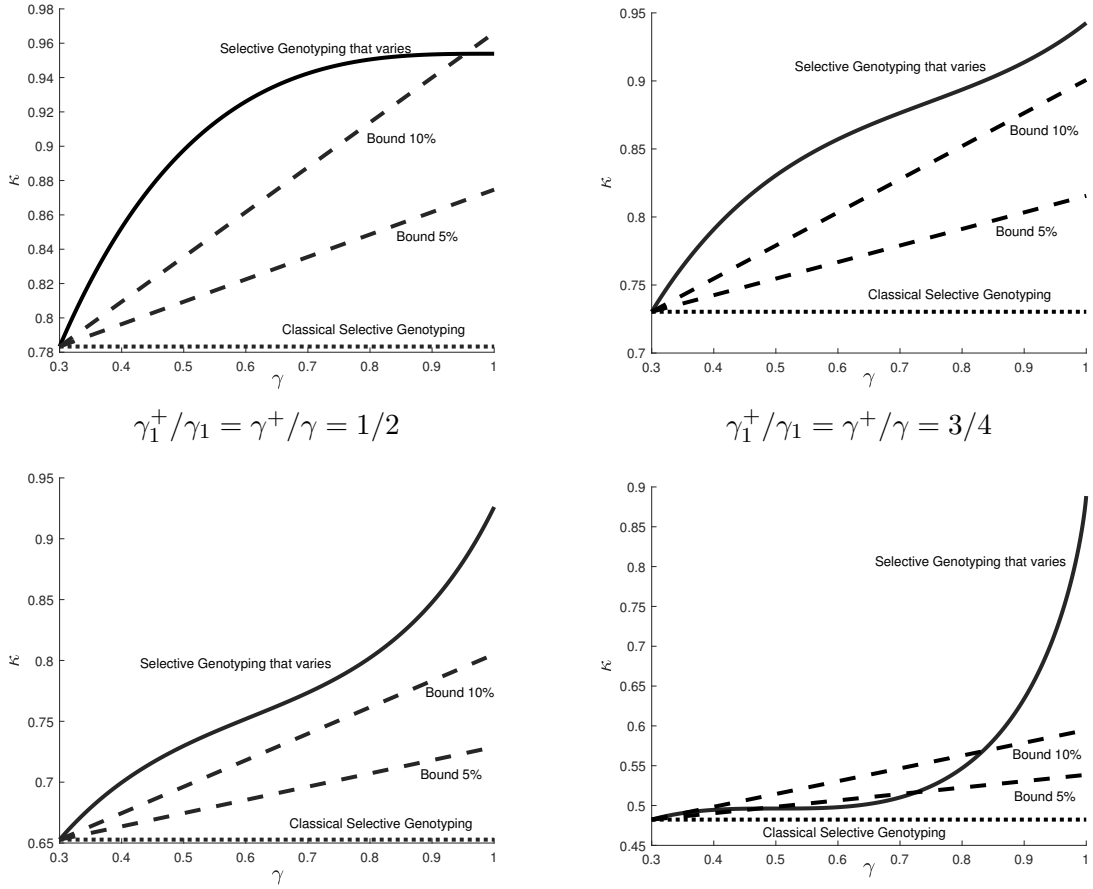
Let us now consider the complementary experiment that consists in:

- genotyping at markers belonging to the dense map (i.e. T_K^1), individuals for which $Y \in \left[S_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ S_+^1\right]$
- genotyping at markers belonging to the sparse map (i.e. T_K^2), individuals for which $Y \geq S_+^1$ or $Y \leq S_-^1$

So, in this new experiment, we observe n observations $\Big(Y_j,\ \overline{X}_j(t_{\sigma(1)}),\ \overline{X}_j(t_{\sigma(2)}),\ ...,\ \overline{X}_j(t_{\sigma(K)}),\ \widetilde{X}_j(t_1), \widetilde{X}_j(t_2),\ ..., \widetilde{X}_j(t_K)\Big)$ i.i.d.

In this context, we have the following result:

Figure 4. Comparison in terms of efficiency, between the classical selective genotyping and the approach where the selective genotyping varies along the genome. Efficiencies κ, with respect to the complete data situation ([2]), are illustrated as a function of γ, and as a function of the ratios γ_1^+/γ_1 and γ^+/γ. In all cases, γ_1 takes the value 0.3, only one QTL is considered ($m = 1$, $a = 2$, $\sigma = 1$) and the test is performed exactly at the QTL location ($t = t_1^\star = 0.85$). The dense map consists in $t^{\ell,1} = 0.80$, $t^{r,1} = 0.90$ and the sparse map consists in $t^{\ell,2} = 0.70$, $t^{r,2} = 1$. The notation Bound 10% (resp. Bound 5%) refers to the computed bound (taken from Lemma 4.4) when the ratio $\#\mathrm{T}_K^2/K$ is equal to 10% (resp. 5%).



$\gamma_1^+/\gamma_1 = \gamma^+/\gamma = 1/2$ $\gamma_1^+/\gamma_1 = \gamma^+/\gamma = 3/4$

$\gamma_1^+/\gamma_1 = \gamma^+/\gamma = 7/8$ $\gamma_1^+/\gamma_1 = \gamma^+/\gamma = 1$

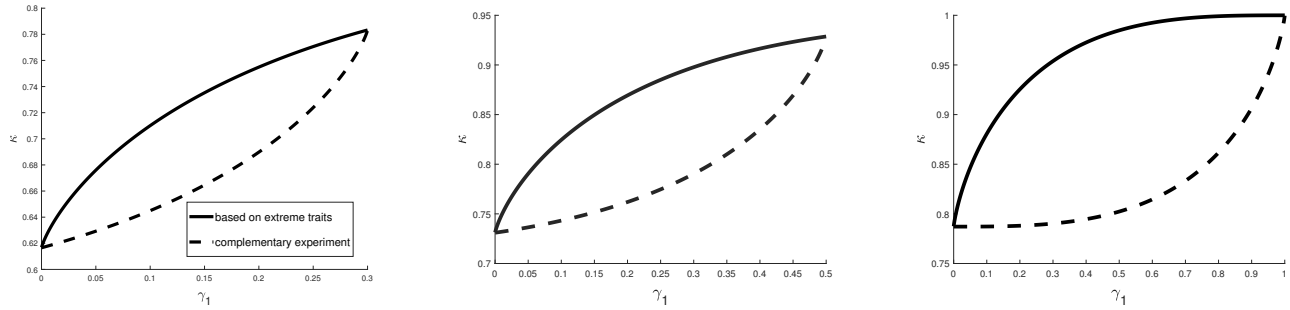
Corollary 5.1: *Under the complementary experiment, we have the same results as in Theorem 4.1 and in Theorem 4.2 provided that we swap the quantities* $\mathcal{A}_1$ *and* $\mathcal{A}_2$.

A sketch of the proof is given in the supplementary material. Figure 5 compares the efficiency obtained for the experiment based on extreme individuals on the dense map (cf. Theorem 4.2), and the efficiency of the complementary experiment (cf. Corollary 5.1). Recall that efficiencies were obtained with respect to the complete data situation ([2]). In order to fairly compare these two experiments, the percentage of individuals genotyped on the dense map has to be the same for boths experiments. Since it is equal to γ_1 in the experiment based on extreme individuals, we have to consider for the complementary experiment, two new thresholds $\tilde{S}_-^1$ and $\tilde{S}_+^1$ such as $\tilde{S}_-^1 \leq S_-^2 \leq S_+^2 \leq \tilde{S}_+^1$ and $\mathrm{P}_{\mathcal{H}_0}\left(Y \in \left[\tilde{S}_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ \tilde{S}_+^1\right]\right) = \gamma_1$. Finally, we considered the relationship $\gamma_+/\gamma = \gamma_1^+/\gamma_1 = 1/2$ and the same framework as in Section 4.4 for the marker and QTL locations.

According to Figure 5, the experiment based on the extremes is largely more efficient that the complementary experiment. It was expected since it has been shown in many studies on selective genotyping (e.g. [31, 32, 49]) that most of the

signal is contained in extreme traits. Note also that when γ_1 was set to the value 0 or to the same value as γ, we observe as expected a perfect match between the efficiencies of the two experiments.

Figure 5. Comparison of the efficiencies κ between the experiment based on extreme individuals on the dense map, and the complementary experiment based on non extreme individuals on the dense map. κ is given as a function of γ_1, γ. In all the settings, γ_1^+/γ_1 and γ^+/γ have been set to 1/2. Only one QTL is considered ($m = 1$, $a = 2$, $\sigma = 1$) and the test is performed exactly at the QTL location ($t = t_1^\star = 0.85$). The dense map consists in $t^{\ell,1} = 0.80$ and $t^{r,1} = 0.90$ whereas the sparse map consists in $t^{\ell,2} = 0.70$, $t^{r,2} = 1$.



$\gamma = 0.3$ $\gamma = 0.5$ $\gamma = 1$

6. Introducing the AdaptSgenoLasso

In this section, let us introduce a new method to estimate the number of QTLs, their effects and their positions combining results of Theorem 4.1 and a penalized likelihood method. Since our method is an extended version of the SgenoLasso ([53]) that allows to put some weights on some loci along the genome, we will call it the AdaptSgenoLasso. We will also present AdaptSgenoEN which is the Elastic Net version of our new method (see formula 19 below).

According to Theorem 4.1, as soon as we discretize the score process at markers positions, we have the following relationship when n is large:

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

where $\vec{S}_n = (S_n(t_1)\ ,\ S_n(t_2)\ ,\ \ldots\ ,\ S_n(t_K))'$,
$\vec{m}_{\vec{t}^\star} = \left(m_{\vec{t}^\star}(t_1)\ ,\ m_{\vec{t}^\star}(t_2)\ ,\ ..., m_{\vec{t}^\star}(t_K)\right)'$ and $\vec{\varepsilon} \sim N(0, \Sigma)$ with $\Sigma_{kk'} = \mathrm{Cov}\left(Z(t_k), Z(t_{k'})\right)$ given in formulae (12), (13) and (14). Since most of the penalized likelihood methods rely on i.i.d. observations, we will decorrelate the components of $\vec{S}_n$ keeping only points of the process taken at marker positions.

In what follows, we assume that we are under complete Linkage Disequilibrium, i.e. the m QTLs are located on some markers. Furthermore, we look for QTLs only at marker locations. Indeed, it will make the reading easier and is particularly appropriate with the high density of markers, thanks to new sequencing technologies. Under this context, Δ_k will denote the putative effect at location t_k.

Notation 6.1: *$\mathcal{G}_k$ denotes either $\sqrt{\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_k)}/\sigma$ or $\sqrt{\mathcal{B}}/\sigma$ depending if t_k belongs to $\mathrm{T}_K^1 \backslash \mathrm{T}_K^2$ or T_K^2, respectively.*

Using the expression of the mean function and also the Cholesky decomposition $\Sigma = AA'$, we have

$$A^{-1}\vec{S}_n = A'\left(\Delta_1\ ,\ \ldots\ ,\ \Delta_K\right)' \; + \; A^{-1}\vec{\varepsilon} \; + \; o_P(1) \tag{15}$$

where

$$\Delta_k = \begin{cases} 0 & \text{if} \quad t_k \notin \{t_1^\star, \ldots, t_m^\star\} \\ \frac{a_s \mathcal{G}_k}{\sigma} & \text{otherwise, with } s \text{ the index such as } t_s^\star = t_k. \end{cases} \tag{16}$$

We can notice that the markers are amplified of a factor $\sqrt{\mathcal{B}}/\sigma$ on the sparse map and amplified of a factor $\sqrt{\mathcal{A}_1 + \mathcal{A}_2 \xi_2^2(t_k)}/\sigma$ on the dense map.

In the sequel, we set $\Delta := (\Delta_1, ..., \Delta_K)'$. In order to find the non zero Δ_k, a natural approach is to use a penalized regression and estimate Δ by:

$$\hat{\Delta}_{\text{AdaptSgeno}}(\lambda, \alpha) = \arg\min_{\Delta} \left(\left\| A^{-1} \vec{S}_n - A' \Delta \right\|_2^2 + \lambda \ \text{pen}(\alpha) \right) \tag{17}$$

where:

$$\text{pen}(\alpha) = \frac{1-\alpha}{2} \left\| \Delta \right\|_2^2 + \alpha \left\| \Delta \right\|_1 \tag{18}$$

and $\| \ \|_2$ is the L2 norm, $\| \ \|_1$ is the L1 norm, and λ and α denote tuning parameters.

As in our previous study, we define the AdaptSgenoLasso and the AdaptSgenoEN in the following way:

$$\begin{aligned} \hat{\Delta}_{\text{AdaptSgenoLasso}}(\lambda) &= \hat{\Delta}_{\text{AdaptSgeno}}(\lambda, 1) \\ \hat{\Delta}_{\text{AdaptSgenoEN}}(\lambda, \alpha) &= \hat{\Delta}_{\text{AdaptSgeno}}(\lambda, \alpha). \end{aligned} \tag{19}$$

Note that for $S_-^1 = S_-^2$ and $S_+^1 = S_+^2$ (classical selective genotyping), since $\mathcal{A}_2 = 0$ and $\mathcal{B} = \mathcal{A}_1$, each entry of the matrix Σ is equal to $\rho(t_k, t_{k'})$ (cf. formulae 12, 13 and 14). As expected, in this case, formula (17) is identical to formula (14) of [53], and the AdaptSgenoLasso (resp. AdaptSgenoEN) matches the SgenoLasso (resp. SgenoEN) under complete Linkage Disequilibrium.

Note that by combining our results from Theorem 4.1 with the Adaptive Lasso [72], we can introduce another penalized likelihood method: we will call it AdaptSgenoAdaptLasso in what follows. In this case, it consists in considering $\alpha = 1$ in formula (18) and in imposing a penalty $\|W'\Delta\|_1$ with weights equal to $1/\sqrt{\mathcal{B}}$ on the sparse map T_K^2 (i.e. major genes) and $1/\sqrt{\mathcal{A}_1 + \mathcal{A}_2 \xi_2^2(t_k)}$ on the map $\text{T}_K^1 \backslash \text{T}_K^2$ (i.e. the dense map without loci belonging to the sparse map). The weighted L1 penalization takes now into account our prior knowledge of major genes, which is not the case of the AdaptSgenoLasso. Indeed, the AdaptSgenoLasso relies on the Lasso penalty that imposes the same Laplace prior distribution on each marker.

7. Asymptotic theory for AdaptSgenoLasso

As in the previous section, we assume that we are under complete Linkage Disequilibrium, i.e. the m QTLs are located on some markers. We have:

$$\hat{\Delta}_{\text{AdaptSgenoLasso}}(\lambda) = \arg\min_{\Delta} \left(\left\| A^{-1} \vec{S}_n - A' \Delta \right\|_2^2 + \lambda \left\| \Delta \right\|_1 \right) \ . \tag{20}$$

Let us normalize all covariables on the same scale. It will replace our problem in the classical setting where the theory for Lasso is well known (cf. [13] page 108).

Since $\hat{\sigma}_k^2 := \frac{1}{K}(AA')_{kk} = \frac{1}{K}$, let us set $A'_{\text{scal}} := \sqrt{K}A'$. Then , let us define

$$\hat{\Delta}_{\text{AdaptSgenoLasso}_{\text{scal}}}(\lambda) := \arg\min_{\Delta} \left(\frac{\left\| A^{-1}\vec{S}_n - A'_{\text{scal}}\Delta/\sqrt{K} \right\|_2^2}{K} + \lambda \left\| \frac{\Delta}{\sqrt{K}} \right\|_1 \right) .$$

As soon as we set $\tilde{\Delta} := \Delta/\sqrt{K}$, this problem can be rewritten in the following way:

$$\hat{\tilde{\Delta}}_{\text{AdaptSgenoLasso}_{\text{scal}}}(\lambda) := \arg\min_{\tilde{\Delta}} \left(\frac{\left\| A^{-1}\vec{S}_n - A'_{\text{scal}}\tilde{\Delta} \right\|_2^2}{K} + \lambda \left\| \tilde{\Delta} \right\|_1 \right) . \quad (21)$$

We can apply Corollary 6.1 of [13] with $\hat{\sigma} = 1$ (cf. our linear model in formula (15)), that establishes the slow rate of convergence

$$\begin{aligned} &\frac{\left\| A'_{\text{scal}}(\hat{\tilde{\Delta}}_{\text{AdaptSgenoLasso}_{\text{scal}}} - \tilde{\Delta}) \right\|_2^2}{K} \\ &= O_P\left(\frac{\sqrt{\log(K)}}{K} \left\{ \sum_{s|t_s^\star \in \mathrm{T}_K^2} \frac{|a_s|\sqrt{\mathcal{B}}}{\sigma^2} + \sum_{s|t_s^\star \in \mathrm{T}_K^1 \setminus \mathrm{T}_K^2} \frac{|a_s|\sqrt{\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_s^\star)}}{\sigma^2} \right\} \right) \end{aligned} \quad (22)$$

where $O_P(1)$ denotes a sequence that is bounded in probability when $K \to +\infty$.

Note also that assuming that the "compatibility condition" holds, Corollary 6.2 of [13] applies and we obtain the fast rate of convergence:

$$\frac{\left\| A'_{\text{scal}}(\hat{\tilde{\Delta}}_{\text{AdaptSgenoLasso}_{\text{scal}}} - \tilde{\Delta}) \right\|_2^2}{K} = O_P\left(\frac{\log(K)m}{K\Phi_0^2} \right) \quad (23)$$

where m is the number of QTLs (factor linked to the sparsity), and Φ_0^2 refers to a compatibility constant.

Let us state the classical Lasso conditions in the "AdaptSgenoLasso" context:

The β-min condition:

$$\min\left(\min_{s|t_s^\star \in \mathrm{T}_K^2} \frac{|a_s|\sqrt{\mathcal{B}}}{\sigma^2\sqrt{K}}, \min_{s|t_s^\star \in \mathrm{T}_K^1 \setminus \mathrm{T}_K^2} \frac{|a_s|\sqrt{\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_s^\star)}}{\sigma^2\sqrt{K}} \right) >> \Phi^{-2}\sqrt{\frac{m\log(K)}{K}}$$

where Φ^2 is a restricted eigen value of the design matrix A'_{scal}.

The irrepresentable condition:

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

where $\|x\|_\infty = \max_j |x_j|$, $\text{Sign}(a_1,\ldots,a_m) = (\text{Sign}(a_1),\ldots,\text{Sign}(a_m))'$. $\Sigma^{(\cdot,\star)}$ is a matrix of size $(K-m)\times m$: it is the submatrix of Σ where rows refers to markers not matching QTL locations, and where columns refers to QTL loci.

Table 2. Performances of the AdaptSgenoLasso as a function of γ_1, γ, n (Mean over 100 samples, $\gamma_1^+/\gamma_1 = 1/2$, $\gamma_+/\gamma = 1/2$, $\sigma = 1$). The following framework is considered : $T = 10$, K =10,001, $t_k = 0.001(k-1)$, $m = 12$, $t_1^\star = 0.65$, $t_2^\star = 1.50$, $t_3^\star = 2.35$, $t_4^\star = 2.75$, $t_5^\star = 3.10$, $t_6^\star = 3.75$, $t_7^\star = 4.15$, $t_8^\star = 4.85$, $t_9^\star = 6.30$, $t_{10}^\star = 7.90$, $t_{11}^\star = 8.10$, $t_{12}^\star = 8.60$. The sparse map consists in markers located every 0.25 Morgans. In all cases, $|q_s| = 0.1897$. The notation L1 ratio(δ) corresponds to the quantity $\sum_{k|t_1^\star-\delta\le t_k\le t_1^\star+\delta\ \cup\ldots\cup\ t_m^\star-\delta\le t_k\le t_m^\star+\delta} |\hat{\Delta}_k| / \sum_{k|t_k\in \mathrm{T}_K^1} |\hat{\Delta}_k|$.

		($T = 10$, $n = 500$, K =10,001)		($T = 10$, n =1,000, K =10,001)		($T = 10$, n =2,000, K =10,001)	
γ_1	γ	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)
0.1	0.1*	15.24%	24.48%	22.45%	34.31%	32.61%	46.91%
	0.2	16.36%	25.85%	24.49%	37.06%	35.02%	49.98%
	0.3	16.97%	26.44%	24.57%	37.61%	36.49%	51.83%
	0.4	17.75%	27.17%	24.91%	38.11%	36.97%	52.34%
	0.5	17.48%	26.85%	25.50%	38.99%	37.30%	52.77%
	1	17.92%	27.59%	25.78%	39.53%	37.38%	52.82%
0.2	0.2*	17.50%	26.89%	25.36%	37.49%	36.06%	50.44%
	0.3	17.67%	27.19%	26.19%	38.78%	37.90%	52.70%
	0.4	18.73%	28.30%	26.18%	38.91%	39.01%	53.92%
	0.5	18.85%	28.23%	26.67%	39.49%	39.28%	54.16%
	1	18.92%	28.80%	26.86%	39.89%	40.23%	55.40%
0.3	0.3*	18.36%	27.82%	26.71%	39.30%	38.19%	51.98%
	0.4	18.88%	28.56%	27.30%	40.00%	39.49%	53.47%
	0.5	19.08%	28.94%	27.35%	40.13%	40.15%	54.26%
	1	19.38%	29.49%	28.13%	41.12%	40.97%	55.07%

$*$ SgenoLasso and AdaptSgenoLasso are a perfect match.

Table 3. Same as Table 2 except that the AdaptSgenoEN is considered.

		($T = 10$, $n = 500$, K =10,001)		($T = 10$, n =1,000, K =10,001)		($T = 10$, n =2,000, K =10,001)	
γ_1	γ	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)
0.1	0.1*	15.33%	24.28%	22.31%	33.93%	32.47%	46.47%
	0.2	16.23%	25.43%	24.38%	36.83%	34.98%	49.50%
	0.3	17.03%	26.57%	24.47%	37.49%	36.03%	50.85%
	0.4	17.63%	27.05%	24.96%	38.28%	37.18%	52.20%
	0.5	17.67%	26.99%	25.38%	38.69%	37.28%	52.25%
	1	17.55%	27.16%	25.66%	39.32%	37.50%	52.44%
0.2	0.2*	17.16%	26.59%	25.07%	37.29%	35.70%	49.88%
	0.3	17.73%	27.43%	25.80%	38.15%	37.30%	51.73%
	0.4	18.35%	28.08%	26.07%	38.80%	38.56%	53.18%
	0.5	18.66%	28.41%	26.64%	39.41%	39.08%	53.74%
	1	18.50%	28.38%	26.75%	39.59%	39.42%	54.21%
0.3	0.3*	18.25%	27.68%	26.46%	38.89%	37.65%	51.65%
	0.4	18.84%	28.50%	26.71%	39.27%	38.62%	52.80%
	0.5	18.95%	28.84%	27.32%	40.09%	39.17%	53.38%
	1	19.13%	28.99%	27.76%	40.79%	39.89%	54.08%

$*$ SgenoEN and AdaptSgenoEN are a perfect match.

Table 4. Performances of the AdaptSgenoLasso as a function of γ_1, γ, n (Mean over 100 samples, $\gamma_1^+/\gamma_1 = 1/2$, $\gamma_+/\gamma = 1/2$, $\sigma = 1$). The following framework is considered : $T = 4$, K =4,001, $t_k = 0.001(k-1)$, $m = 6$, $t_1^\star = 0.65$, $t_2^\star = 1.50$, $t_3^\star = 2.35$, $t_4^\star = 2.75$, $t_5^\star = 3.10$, $t_6^\star = 3.75$. The sparse map consists in markers located every 0.25 Morgans. In all cases, $|q_s| = 0.1897$. The notation L1 ratio(δ) corresponds to the quantity $\sum_{k|t_1^\star-\delta\le t_k\le t_1^\star+\delta\ \cup\ldots\cup\ t_m^\star-\delta\le t_k\le t_m^\star+\delta} |\hat{\Delta}_k| / \sum_{k|t_k\in \mathrm{T}_K^1} |\hat{\Delta}_k|$.

		($T = 4$, $n = 500$, K =4,001)		($T = 4$, n =1,000, K =4,001)		($T = 4$, n =2,000, K =4,001)	
γ_1	γ	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)
0.1	0.1*	13.88%	23.31%	22.37%	34.51%	32.97%	47.29%
	0.2	16.51%	26.49%	24.39%	37.86%	36.08%	51.33%
	0.3	16.48%	27.21%	25.41%	38.47%	36.82%	52.32%
	0.4	17.03%	27.51%	26.13%	39.65%	37.62%	53.39%
	0.5	16.81%	27.28%	26.87%	40.16%	37.73%	53.51%
	1	16.72%	27.97%	27.38%	40.93%	39.39%	55.11%
0.2	0.2*	19.28%	29.69%	27.93%	40.29%	37.59%	50.96%
	0.3	19.89%	30.94%	28.69%	41.40%	40.18%	54.19%
	0.4	19.87%	30.96%	29.81%	42.69%	40.48%	54.39%
	0.5	20.04%	31.35%	30.26%	43.37%	41.52%	55.59%
	1	19.78%	31.33%	30.94%	43.95%	42.35%	56.12%
0.3	0.3*	20.19%	31.70%	30.28%	42.58%	40.78%	55.40%
	0.4	20.52%	31.92%	31.21%	44.03%	41.52%	56.34%
	0.5	20.28%	32.02%	31.94%	44.90%	42.06%	56.74%
	1	20.64%	32.72%	31.93%	45.39%	42.50%	56.56%

$*$ SgenoLasso and AdaptSgenoLasso are a perfect match.

Recall that according to [13], the irrepresentable condition implies the compatibility condition, that ensures the fast rate of convergence. On the other hand, the β-min condition and the irrepresentable condition, ensure consistent variable selection for AdaptiveSgenoLasso.

Note that we can easily recover the different conditions obtained for the SgenoLasso ([53]) as soon as we set $\mathrm{T}_K^2 = \emptyset$, $\mathcal{A}_2 = 0$ in the different expressions of this section.

Table 5. Same as Table 4 except that the AdaptSgenoEN is considered.

		($T=4$, $n=500$, $K=$4,001)		($T=4$, $n=$1,000, $K=$4,001)		($T=4$, $n=$2,000, $K=$4,001)	
γ_1	γ	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)
0.1	0.1*	14.45%	24.24%	22.67%	34.60%	32.44%	46.47%
	0.2	16.79%	27.02%	25.34%	37.97%	35.97%	50.81%
	0.3	16.92%	27.92%	25.72%	38.60%	37.14%	52.23%
	0.4	17.40%	28.28%	26.73%	39.62%	38.10%	53.23%
	0.5	17.09%	28.48%	27.29%	40.19%	38.55%	53.61%
	1	16.98%	28.48%	27.87%	40.81%	39.39%	54.54%
0.2	0.2*	19.37%	29.60%	27.85%	40.01%	37.64%	51.51%
	0.3	20.19%	30.74%	29.15%	41.57%	38.89%	52.96%
	0.4	20.53%	31.44%	29.97%	42.58%	40.13%	54.37%
	0.5	19.86%	30.88%	30.36%	43.00%	40.72%	55.05%
	1	20.36%	31.59%	30.56%	43.29%	41.73%	55.90%
0.3	0.3*	20.74%	32.03%	30.28%	42.40%	39.92%	54.41%
	0.4	20.86%	32.31%	31.12%	43.41%	40.10%	54.39%
	0.5	20.78%	32.35%	31.77%	44.20%	41.46%	56.10%
	1	21.35%	33.26%	31.70%	44.23%	41.76%	56.03%

∗ SgenoEN and AdaptSgenoEN are a perfect match.

Table 6. Performances of the AdaptSgenoLasso in presence of large and small effects QTLs (Mean over 100 samples, $\gamma_1 = 0.1$, $\gamma_1^+/\gamma_1 = 1/2$, $\gamma_+/\gamma = 1/2$, $\sigma = 1$). Same genetic maps as in Table 2. For the large effects, $|q_s| = 0.3794$ at locations 1.50, 2.75, and 3.75, whereas for the small effects, $|q_s| = 0.1897$ at locations 0.65, 2.35, 3.10, 4.15, 4.85, 6.30, 7.90, 8.10, 8.60 . The L1 ratio(δ) is given for the large effects QTLs, small effects QTLs, and all the QTLs.

				($T=10$, $K=$10,001)			
		Large QTLs		Small QTLs		All QTLs	
n	γ	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)
500	0.1*	12.89%	17.60%	7.32%	12.76%	20.22%	30.36%
	0.2	16.88%	22.55%	7.03%	12.34%	23.91%	34.90%
	0.3	18.16%	24.25%	6.88%	12.09%	25.05%	36.34%
	0.4	18.89%	25.21%	6.78%	11.71%	25.68%	36.92%
	0.5	19.50%	25.87%	6.91%	11.76%	26.41%	37.63%
	1	20.05%	26.67%	7.02%	11.81%	27.09%	38.48%
1,000	0.1*	16.55%	22.36%	12.98%	20.50%	29.53%	42.86%
	0.2	21.52%	28.44%	12.46%	19.59%	33.98%	48.04%
	0.3	23.25%	30.61%	12.16%	19.28%	35.41%	49.90%
	0.4	24.39%	32.02%	12.00%	19.09%	36.39%	51.12%
	0.5	24.70%	32.27%	11.81%	18.91%	36.53%	51.19%
	1	25.44%	33.45%	11.90%	18.81%	37.34%	52.26%
2,000	0.1*	22.08%	27.27%	18.68%	26.98%	40.76%	54.27%
	0.2	28.04%	32.95%	18.21%	26.38%	46.25%	59.33%
	0.3	30.27%	35.17%	18.09%	26.01%	48.35%	61.17%
	0.4	31.68%	36.51%	17.83%	25.59%	49.51%	62.10%
	0.5	32.49%	37.48%	17.89%	25.59%	50.38%	63.07%
	1	32.89%	37.84%	17.63%	25.20%	50.52%	63.04%

∗ SgenoLasso and AdaptSgenoLasso are a perfect match.

Table 7. Same as Table 6 except that the AdaptSgenoEN is considered.

				($T=10$, $K=$10,001)			
		Large QTLs		Small QTLs		All QTLs	
n	γ	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)
500	0.1*	13.03%	17.61%	7.21%	12.66%	20.25%	30.27%
	0.2	16.56%	22.06%	6.89%	12.27%	23.46%	34.33%
	0.3	17.89%	24.12%	6.81%	12.02%	24.70%	36.14%
	0.4	18.93%	25.33%	6.75%	11.66%	25.69%	36.99%
	0.5	19.28%	25.56%	6.84%	11.76%	26.12%	37.32%
	1	20.25%	26.67%	6.80%	11.64%	27.04%	38.31%
1,000	0.1*	16.23%	21.88%	12.94%	20.31%	29.17%	42.20%
	0.2	21.53%	28.10%	12.57%	19.64%	34.11%	47.75%
	0.3	23.29%	30.18%	12.22%	19.19%	35.52%	49.37%
	0.4	24.60%	31.70%	12.07%	18.97%	36.67%	50.68%
	0.5	24.97%	32.06%	11.95%	18.80%	36.92%	50.87%
	1	25.88%	33.15%	11.89%	18.60%	37.77%	51.76%
2,000	0.1*	22.03%	26.99%	18.47%	26.72%	40.50%	53.70%
	0.2	27.93%	32.70%	18.06%	26.05%	45.99%	58.75%
	0.3	30.42%	35.09%	17.93%	25.75%	48.35%	60.84%
	0.4	31.72%	36.39%	17.67%	25.36%	49.39%	61.75%
	0.5	32.20%	36.67%	17.42%	24.88%	49.62%	61.55%
	1	33.05%	37.56%	17.45%	24.91%	50.50%	62.46%

∗ SgenoEN and AdaptSgenoEN are a perfect match.

8. Simulation study

8.1. *About the Max Test*

In this section, the focus is on the max test. Recall that the max test relies on the test statistic $\sup \Lambda_n(.)$. In this context, Table 1 compares the power of the classical selective genotyping approach and our new approach where the selective genotyping varies along the genome. In order to compute the theoretical power, 10,000 paths of the asymptotic process were sampled, whereas the empirical power

Table 8. Same as Table 6 except that the AdaptSgenoAdaptLasso is considered.

		($T = 10$, K =10,001)					
		Large QTLs		Small QTLs		All QTLs	
n	γ	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)
500	0.1	13.18%	17.92%	7.30%	12.86%	20.48%	30.77%
	0.2	21.04%	25.96%	6.19%	11.06%	27.23%	37.02%
	0.3	25.82%	30.42%	5.75%	10.32%	31.57%	40.74%
	0.4	28.01%	32.53%	5.61%	9.77%	33.62%	42.30%
	0.5	28.80%	33.18%	5.34%	9.44%	34.15%	42.62%
	1	31.50%	35.86%	5.27%	9.13%	36.77%	44.99%
1,000	0.1	16.55%	22.37%	12.85%	20.39%	29.40%	42.76%
	0.2	24.46%	30.60%	11.67%	18.43%	36.12%	49.03%
	0.3	27.83%	33.90%	10.62%	17.02%	38.46%	50.92%
	0.4	29.86%	35.68%	10.15%	16.25%	40.00%	51.93%
	0.5	31.23%	36.91%	9.64%	15.81%	40.88%	52.72%
	1	31.97%	37.73%	9.53%	15.27%	41.50%	53.00%
2,000	0.1	21.92%	27.07%	18.61%	26.94%	40.52%	54.01%
	0.2	29.75%	34.10%	17.41%	25.15%	47.16%	59.24%
	0.3	33.58%	37.48%	16.78%	24.11%	50.36%	61.59%
	0.4	35.20%	38.67%	16.14%	23.03%	51.35%	61.71%
	0.5	35.98%	39.33%	15.90%	22.63%	51.88%	61.95%
	1	36.93%	40.13%	15.34%	21.84%	52.27%	61.97%

Table 9. Comparison between the AdaptSgenoLasso, the AdaptSgenoEN and the AdaptSgenoAdaptLasso in presence of large and small effects QTLs. Summary of Tables 6-8. The L1 ratio(δ) is given for all the QTLs.

		($T = 10$, K =10,001)					
		AdaptSgenoLasso		AdaptSgenoEN		AdaptSgenoAdaptLasso	
n	γ	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)
500	0.1	20.22%	30.36%	20.25%	30.27%	20.48%	30.77%
	0.2	23.91%	34.90%	23.46%	34.33%	27.23%	37.02%
	0.3	25.05%	36.34%	24.70%	36.14%	31.57%	40.74%
	0.4	25.68%	36.92%	25.69%	36.99%	33.62%	42.30%
	0.5	26.41%	37.63%	26.12%	37.32%	34.15%	42.62%
	1	27.09%	38.48%	27.04%	38.31%	36.77%	44.99%
1,000	0.1	29.53%	42.86%	29.17%	42.20%	29.40%	42.76%
	0.2	33.98%	48.04%	34.11%	47.75%	36.12%	49.03%
	0.3	35.41%	49.90%	35.52%	49.37%	38.46%	50.92%
	0.4	36.39%	51.12%	36.67%	50.68%	40.00%	51.93%
	0.5	36.53%	51.19%	36.92%	50.87%	40.88%	52.72%
	1	37.34%	52.26%	37.77%	51.76%	41.50%	53.00%
2,000	0.1	40.76%	54.27%	40.50%	53.70%	40.52%	54.01%
	0.2	46.25%	59.33%	45.99%	58.75%	47.16%	59.24%
	0.3	48.35%	61.17%	48.35%	60.84%	50.36%	61.59%
	0.4	49.51%	62.10%	49.39%	61.75%	51.35%	61.71%
	0.5	50.38%	63.07%	49.62%	61.55%	51.88%	61.95%
	1	50.52%	63.04%	50.50%	62.46%	52.27%	61.97%

Table 10. Performances of the AdaptSgenoLasso in presence of large and small effects QTLs (Mean over 100 samples, $\gamma_1 = 0.1$, $\gamma_1^+/\gamma_1 = 1/2$, $\gamma_+/\gamma = 1/2$, $\sigma = 1$). Same genetic maps as in Table 4. For the large effects, $|q_s| = 0.3794$ at locations 1.50, 2.75, and 3.75, whereas for the small effects, $|q_s| = 0.1897$ at locations 0.65, 2.35, 3.10. The L1 ratio(δ) is given for the large effects QTLs, small effects QTLs, and all the QTLs.

		($T = 4$, K =4,001)					
		Large QTLs		Small QTLs		All QTLs	
n	γ	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)
500	0.1*	19.94%	30.57%	4.70%	7.69%	24.64%	38.26%
	0.2	24.79%	36.18%	4.28%	7.09%	29.07%	43.28%
	0.3	25.49%	37.83%	4.07%	6.65%	29.56%	44.48%
	0.4	26.79%	38.72%	4.00%	6.56%	30.79%	45.28%
	0.5	27.12%	39.56%	3.88%	6.37%	31.00%	45.93%
	1	27.54%	41.07%	3.87%	6.31%	31.41%	47.38%
1,000	0.1*	32.09%	41.82%	6.55%	10.35%	38.64%	52.17%
	0.2	37.78%	47.87%	5.92%	9.56%	43.70%	57.43%
	0.3	39.87%	50.47%	5.71%	9.03%	45.58%	59.50%
	0.4	41.76%	52.25%	5.59%	8.87%	47.34%	61.12%
	0.5	41.99%	52.21%	5.46%	8.66%	47.45%	60.87%
	1	42.54%	53.13%	5.17%	8.28%	47.71%	61.41%
2,000	0.1*	38.58%	46.04%	10.58%	14.93%	49.16%	60.96%
	0.2	47.13%	53.99%	9.90%	14.00%	57.03%	67.99%
	0.3	49.19%	55.55%	9.41%	13.17%	58.60%	68.72%
	0.4	50.36%	56.35%	9.11%	12.65%	59.47%	69.00%
	0.5	51.05%	57.19%	8.82%	12.39%	59.86%	69.58%
	1	52.25%	57.84%	8.66%	12.01%	60.90%	69.85%

$*$ SgenoLasso and AdaptSgenoLasso are a perfect match.

is based on 1,000 samples of size n. n took either the value 1,000 , 200 or 100. The threshold (i.e. critical value) at the 5% level was obtained thanks to 10,000 paths of the asymptotic process $Z^2(.)$. The parameters γ and γ_1 were set to the values 0.5 and 0.3, respectively. Note that when the classical selective genotyping approach (i.e. $\gamma_1 = \gamma$) was considered, γ_1 was set to 0.3.

The chromosome is of length 1M ($T = 1$), with 101 markers ($K = 101$) equally spaced every 1cM on map 1, and 5 markers equally spaced every 25cM on map

Table 11. Same as Table 10 except that AdaptSgenoEN is considered.

		($T = 4$, K =4,001)					
		Large QTLs		Small QTLs		All QTLs	
n	γ	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)
500	0.1*	20.26%	30.69%	4.86%	7.84%	25.11%	38.53%
	0.2	24.78%	36.20%	4.42%	7.14%	29.21%	43.35%
	0.3	26.67%	38.57%	4.36%	6.88%	31.03%	45.45%
	0.4	27.30%	39.36%	4.03%	6.55%	31.32%	45.91%
	0.5	28.06%	40.65%	3.97%	6.50%	32.03%	47.14%
	1	27.84%	40.86%	3.70%	6.17%	31.54%	47.03%
1,000	0.1*	32.13%	41.36%	6.75%	10.10%	38.88%	51.45%
	0.2	37.27%	46.82%	6.11%	9.30%	43.38%	56.12%
	0.3	40.42%	50.06%	5.92%	8.89%	46.33%	58.95%
	0.4	41.95%	51.64%	5.73%	8.69%	47.68%	60.32%
	0.5	42.12%	51.63%	5.55%	8.41%	47.66%	60.04%
	1	43.59%	53.29%	5.36%	8.23%	48.95%	61.52%
2,000	0.1*	38.04%	45.76%	10.46%	14.87%	48.50%	60.63%
	0.2	45.54%	52.96%	9.81%	13.86%	55.35%	66.82%
	0.3	47.82%	54.76%	9.20%	13.01%	57.02%	67.77%
	0.4	49.54%	56.29%	9.06%	12.67%	58.60%	68.97%
	0.5	50.25%	56.81%	8.87%	12.38%	59.12%	69.19%
	1	50.87%	56.97%	8.51%	11.74%	59.37%	68.72%

∗ SgenoEN and AdaptSgenoEN are a perfect match.

Table 12. Same as Table 10 except that AdaptSgenoAdaptLasso is considered.

		($T = 4$, K =4,001)					
		Large QTLs		Small QTLs		All QTLs	
n	γ	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)
500	0.1	20.01%	30.45%	4.77%	7.66%	24.78%	38.11%
	0.2	30.15%	40.75%	3.94%	6.54%	34.09%	47.29%
	0.3	34.08%	44.70%	3.68%	5.90%	37.75%	50.60%
	0.4	36.81%	47.12%	3.28%	5.46%	40.10%	52.58%
	0.5	38.16%	48.80%	3.19%	5.21%	41.35%	54.01%
	1	39.03%	49.67%	3.03%	5.07%	42.06%	54.74%
1,000	0.1	32.28%	42.07%	6.48%	10.30%	38.75%	52.37%
	0.2	41.95%	50.93%	5.36%	8.84%	47.31%	59.77%
	0.3	45.60%	54.01%	4.89%	7.89%	50.49%	61.90%
	0.4	48.54%	56.48%	4.52%	7.48%	53.06%	63.96%
	0.5	49.41%	57.15%	4.35%	7.10%	53.75%	64.26%
	1	50.97%	58.20%	4.09%	6.79%	55.07%	65.01%
2,000	0.1	39.04%	46.56%	10.61%	15.06%	49.65%	61.62%
	0.2	50.26%	56.16%	9.59%	13.39%	59.85%	69.55%
	0.3	52.16%	57.18%	8.53%	12.02%	60.69%	69.20%
	0.4	54.94%	59.55%	8.07%	11.30%	63.02%	70.86%
	0.5	56.45%	60.68%	7.93%	10.98%	64.39%	71.66%
	1	57.82%	61.70%	7.41%	10.31%	65.23%	72.01%

Table 13. Comparison between the AdaptSgenoLasso, the AdaptSgenoEN and the AdaptSgenoAdaptLasso in presence of large and small effects QTLs. Summary of Tables 10-12. The L1 ratio(δ) is given for all the QTLs.

		($T = 4$, K =4,001)					
		AdaptSgenoLasso		AdaptSgenoEN		AdaptSgenoAdaptLasso	
n	γ	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)	L1 ratio(0.01)	L1 ratio(0.02)
500	0.1	24.64%	38.26%	25.11%	38.53%	24.78%	38.11%
	0.2	29.07%	43.28%	29.21%	43.35%	34.09%	47.29%
	0.3	29.56%	44.48%	31.03%	45.45%	37.75%	50.60%
	0.4	30.79%	45.28%	31.32%	45.91%	40.10%	52.58%
	0.5	31.00%	45.93%	32.03%	47.14%	41.35%	54.01%
	1	31.41%	47.38%	31.54%	47.03%	42.06%	54.74%
1,000	0.1	38.64%	52.17%	38.88%	51.45%	38.75%	52.37%
	0.2	43.70%	57.43%	43.38%	56.12%	47.31%	59.77%
	0.3	45.58%	59.50%	46.33%	58.95%	50.49%	61.90%
	0.4	47.34%	61.12%	47.68%	60.32%	53.06%	63.96%
	0.5	47.45%	60.87%	47.66%	60.04%	53.75%	64.26%
	1	47.71%	61.41%	48.95%	61.52%	55.07%	65.01%
2,000	0.1	49.16%	60.93%	48.50%	60.63%	49.65%	61.62%
	0.2	57.03%	67.99%	55.35%	66.82%	59.85%	69.55%
	0.3	58.60%	68.72%	57.02%	67.77%	60.69%	69.20%
	0.4	59.47%	69.00%	58.60%	68.97%	63.02%	70.86%
	0.5	59.86%	69.58%	59.12%	69.19%	64.39%	71.66%
	1	60.90%	69.85%	59.37%	68.72%	65.23%	72.01%

2. Different architectures are studied: either 1 QTL ($m = 1$) at 3cM, either 2 QTLs ($m = 2$) at 3cM and 55cM, or 3 QTLs ($m = 3$) at 3cM, 55cM and 80cM. For all cases, the absolute value of the constant linked to the QTL effect was equal to 2.8284 (i.e. $|a_s| = 2.8284$), allowing to deal with a small QTL effect of 0.2 when $n = 200$. The power is computed as a function of the ratio γ_+/γ. In order to concentrate on the same kind of selective genotyping on maps 1 and 2, we considered the relationship $\gamma_+/\gamma = \gamma_1^+/\gamma_1$ in all cases. According to Table 1, we can notice a fair agreement between the empirical power and the theoretical power for n=1,000. On the other hand, our new approach performed better than the

classical approach, when the ratios γ_+/γ took the values 1/2 or 1/4. For instance, when the selective genotyping was performed symmetrically, the asymptotic power associated to our approach was found equal to 58.55% for $m = 1$ and to 46.69% for $m = 3$. In contrast, the power associated to the classical approach was estimated to 48.09% for $m = 1$ and to 40.83% for $m = 3$.

Surprisingly, the classical approach was the best method when the selective genotyping was unidirectional ($\gamma_+/\gamma = 1$). It can be explained by the fact that in this setting, the ratio $\mathcal{A}_1/(\mathcal{A}_1 + \mathcal{A}_2)$ takes the value 3.54% , which means that the contribution to the sparse map is negligible as compared to the one of the dense map. In contrast, when γ_+/γ is set to 1/2 and 1/4, $\mathcal{A}_1/(\mathcal{A}_1 + \mathcal{A}_2)$ is equal to 14.84% and 15.65%, respectively. In this case, genotyping extra individuals on the sparse map is more rewarding.

To sum up, overall, it is clear in view of our simulation study that we should use a symmetrical selective genotyping that varies along the genome.

8.2. *Association study*

In this section, we propose to investigate the performances of the AdaptSgenoLasso and its cousins in association studies. In the different tables, performances will be reported in terms of L1 ratio which is an indicator of whether or not the detected QTLs belong to the "signal area" assuming a tolerance level of either 0.01M or 0.02M (cf. captions in tables for more details).

Tables 2-5 focus on small effects QTLs, whereas Tables 6-13 consider both small effects QTLs and large effects QTLs. n took either the value 500 , 1,000 or 2,000. The genome length was set either to 4M or to 10M. When $T = 10$ (resp. $T = 4$), 12 QTLs (resp. 6 QTLs) were placed on the genome, and the dense map relied on 10,000 equally spaced markers (resp. 4,000 markers). Besides, in both cases, the sparse map consisted in markers located every 0.25M. A symmetrical selective genotyping ($\gamma_+/\gamma = \gamma_1^+/\gamma_1 = 1/2$) was performed. We let the parameter γ vary from 0.1 to 1, and considered a few values for γ_1. Recall that under the setting $\gamma = \gamma_1$, since we have the same percentage of genotyped individuals on the two maps, the AdaptSgenoLasso and the AdaptSgenoEN match the SgenoLasso and the SgenoEN, respectively.

According to Tables 2-5, as expected, when the value of γ_1 was fixed, the L1 ratio globally increased with γ, specially for a large number of observations (see n =1,000 or 2,000). In the same way, for a given value of γ, the L1 ratio globally increased with γ_1 in most cases. Overall, the AdaptSgenoLasso and the AdaptSgenoEN presented very similar performances. Indeed, for $T = 10$ (resp. $T = 4$), the average L1 ratio assuming a tolerance level of 0.01M, was found equal to 27.21% (resp. 28.73%) for the AdaptSgenoLasso and to 26.97% (resp. 28.80%) for the AdaptSgenoEN.

Let us now focus on Tables 6-13 dealing with a mixture of small and large effects QTLs. The genetic maps were the same as before, except that 3 QTLs of large effects were considered when $T = 4$ and $T = 10$. The large effects were chosen as twice the small effects. Note that in the tables, for a given tolerance level, three kinds of L1 ratios are given: the one focusing only on large effects QTLs, the one based exclusively on small effects QTLs, and the classical one for all the QTLs. The percentage γ_1 was set to the value 0.1 in all experiments.

According to Table 6-7, the L1 ratio relying on large effects globally increased with γ whereas the one based on small effects QTLs decreased with γ. This behavior is not surprising since at loci belonging to the sparse map, QTL effects are more and more amplified (cf. formula 16) when γ increases. Then, since the denominator of the L1 ratio tends to increase whereas the numerator linked to small effect QTLs

(located on the dense map) remains the same, the L1 ratio based on small effects QTLs decreases. Last, as expected, the classical L1 ratio that considers all the QTLs, increased with n and with γ.

Table 8 describes performances of the AdaptSgenoAdaptLasso. Recall that it incorporates a weighted L1 penalty (cf. end of Section 6), in contrast to the AdaptSgenoLasso and to the AdaptSgenoEN. We can observe a more significant increase in terms of L1 ratio for large effects: more weights are imposed to the large effects thanks to the L1 penalty. In view of Table 9 that proposes a summary of the previous experiments, the AdaptSgenoAdaptLasso is clearly the most performant method. The superiority of the AdaptSgenoAdaptLasso over its cousins was found as the most significant for small number of observations (n=500): the lack of signal in the data must be compensated by the prior on large loci incorporated within the weighted L1 penalty.

Last, Tables 10-13 focus on the case $T = 4$. Same conclusions were obtained as for T=10.

8.3. *Genomic selection*

Let us illustrate now the performances of our new method, in terms of genomic prediction. As mentioned in Section 1, Genomic Selection (GS) focuses on predictions using a large number of markers, whereas GWAS looks for loci of interest. Recall that GS relies on the fact that in presence of a high density of markers, each QTL will tend to be tagged by markers located nearby. In that sense, it is likely that each QTL is in strong Linkage Disequilibrium (i.e. highly correlated) with a few markers.

In our present study, we concentrate on a backcross population which can be viewed as a population resulting from 3 generations (see for instance [65]). Since most of studies on GS rely on populations based on a large number of generations, a way to mimic a large number of generations is to increase the intensity ν of the Poisson process $N(.)$ referring to the number or recombination events along the genome. Indeed, it enables to break long stretch of Linkage Disequilibrium, that is to say it reduces correlation between markers. By default, Haldane model [24] assumes $\nu = 1$, so we will consider as ν values either 1 or 5 in what follows. Note that our theoretical results are still valid forall ν values as soon as we set $\rho(t, t') := e^{-2\nu|t-t'|}$ and $r(t, t') := \frac{1}{2}\left(1 - e^{-2\nu|t-t'|}\right)$ in all our formulas.

As in the previous section, our model was learned on extreme individuals, i.e. individuals for which $Y \notin \left[S_-^1\ ,\ S_+^1\right]$ on the dense map, and individuals for which $Y \in \left[S_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ S_+^1\right]$ on the sparse map. We considered a symmetrical selective genotyping ($\gamma_1^+/\gamma_1 = 1/2$, $\gamma_+/\gamma = 1/2$). γ_1 was always set to the value 0.1 in our experiments, whereas γ took either the value 0.1 or 0.3, in order to study the SgenoLasso and the AdaptSgenoLasso (and cousins), respectively.

The model was learned on all extreme individuals and genomic predictions were evaluated on two kinds of validation sets of size 100. The first one consists in a bootstrap sample from the individuals for which $Y \leq S_-^1$ or $Y \geq S_+^1$. Thus, the bootstrap sample contains genetic clones of a few individuals from the learning set. The second kind of validation set is largely inspired by the one studied for the SgenoLasso (see Section 6.1 of [53]): it consists in generating progenies of the extreme individuals for which $Y \leq S_-^1$ or $Y \geq S_+^1$. Moreover, in both cases, to evaluate the accuracy of the prediction, we simulated phenotypic values Y_{new} for each of the genomes present in the validation samples, using the "analysis of variance model" of formula (1) except that $X(.)$ is now replaced by the new genomes. Recall that in genomic selection, the predictive ability is evaluated according to

the accuracy criterion, that is, the correlation between predicted and true values (see [33, 62]). In our study, we considered the following genomic predictor $\sum_{k=1}^{K} \frac{\sigma}{\mathcal{G}_k\sqrt{n}} W_{\text{new}}(t_k)\hat{\Delta}_k(\nu,\alpha)$ for an individual in the validation set with genome denoted $W_{\text{new}}(.)$. Recall that α refers to the parameter linked to the penalty used (cf. formula 19).

In this context, Tables 14 and 15 focus on the first and on the second kind of validation set, respectively. A genome of length 4M was considered and the same genetic map as in our association study (cf. Section 8.2) with a mixture of small and large effects QTLs. According to Table 14, for $n = 500$ and $\nu = 1$, the accuracy increased from 73.73% for SgenoLasso to 76.00% for AdaptSgenoAdaptLasso. In the same way, for $\nu = 5$, we observed an increase from 60.03% for SgenoLasso to 62.06% for AdaptSgenoAdaptLasso. Note that when the learning set was larger (n=1,000), we observed only a slight improvement: the predictive ability rose from 75.34% to 75.88% for $\nu = 1$, and from 66.52% to 67.72% for $\nu = 5$. With such a large training set, SgenoLasso enjoyed already fair performances. So, in that case, the size of the training set was probably too large for highlighting the superiority of our new method (cf. remark at the end of Section 8.2). In the same way, on the second validation set (see Table 15), we observed good performances for $n = 500$. Indeed, the accuracy increased either from 53.63% to 54.84% for $\nu = 1$, or from 43.34% to 45.43% for $\nu = 5$. When a larger training set was under study (n =1,000), the advantage of AdaptSgenoAdaptLasso (53.01%) over SgenoLasso (51.30%) was found significant only when we reduced the correlation between markers ($\nu = 5$). Table 16 is dedicated to $T = 10$. We considered the same map as in Section 8.2, with a mixture of small and large effects QTLs. Note that we did not exclusively concentrate on 3 large effects QTLs. Indeed, we focused also on a framework with 6 large effects QTLs, in order to obtain exactly the same proportion of large effects QTLs among all QTLs as for $T = 4$. As expected, the accuracy increased with the number of large effects QTLs, and also with the size of the training set. Besides, we observed the same behavior as before: AdaptSgenoAdaptLasso is more interesting than SgenoLasso specially when the learning set is of moderate size (n =500).

9. Conclusion

In this manuscript, we introduced a new variable selection method, called AdaptSgenoLasso, that allows geneticists to give more importance to loci of interest, when the model is learned on extreme individuals. Although AdaptSgenoLasso presents better performances than the SgenoLasso, we advise potential users to choose another cousin, named AdaptSgenoAdaptLasso, that combines advantages of Adaptive Lasso and those of AdaptSgenoLasso. Indeed, according to our simulation study, when the training set was of moderate size, the AdaptSgenoAdaptLasso outperformed the AdaptSgenoLasso. We believe that our methods should be considered as a first step, in the elaboration of more sophisticated models for genomic prediction in years to come.

10. Proof of Theorem 4.1

The proof is divided into four parts:

(1) Preliminaries (i.e. computation of the Fisher Information Matrix)
(2) Study of the score process under H_0
(3) Study of the score process under the local alternative $H_{at^\star}$

Table 14. Comparison between the AdaptSgenoLasso, the AdaptSgenoEN and the AdaptSgenoAdaptLasso in terms of genomic prediction, on the basis of the first validation set (i.e. bootstrap sample). The predictive ability (i.e. accuracy) is given as a function of n, γ and the intensity ν of the Poisson process (Mean over 100 samples, $\gamma_1 = 0.1$, $\gamma_1^+/\gamma_1 = 1/2$, $\gamma_+/\gamma = 1/2$, $\sigma = 1$). Same genetic map and same QTL effects as in Table 10.

			($T = 4$, K =4,001)		
n	ν	γ	AdaptSgenoLasso	AdaptSgenoEN	AdaptSgenoAdaptLasso
500	1	0.1*	73.73%	73.75%	73.75%
		0.3	74.06%	74.05%	76.00%
	5	0.1*	60.03%	60.09%	60.08%
		0.3	60.78%	60.65%	62.06%
1,000	1	0.1*	75.34%	75.33%	75.34%
		0.3	75.77%	75.77%	75.88%
	5	0.1*	66.52%	66.57%	66.53%
		0.3	67.47%	67.48%	67.72%

∗ SgenoLasso and AdaptSgenoLasso are a perfect match.

Table 15. Same as Table 14 except that the focus is on the second validation set (i.e. new progenies).

			($T = 4$, K =4,001)		
n	ν	γ	AdaptSgenoLasso	AdaptSgenoEN	AdaptSgenoAdaptLasso
500	1	0.1*	53.63%	53.74%	53.65%
		0.3	54.54%	54.56%	54.84%
	5	0.1*	43.34%	43.47%	43.37%
		0.3	43.75%	43.63%	45.43%
1,000	1	0.1*	55.18%	55.15%	55.04%
		0.3	55.72%	55.72%	55.80%
	5	0.1*	51.30%	51.33%	51.27%
		0.3	52.29%	52.65%	53.01%

∗ SgenoLasso and AdaptSgenoLasso are a perfect match.

Table 16. Comparison between the AdaptSgenoLasso, the AdaptSgenoEN and the AdaptSgenoAdaptLasso in terms of genomic prediction, on the basis of the second validation set (i.e. new progenies). The predictive ability (i.e. accuracy) is given as a function of n, γ and the number of large effects QTLs ($\nu = 5$, Mean over 100 samples, $\gamma_1 = 0.1$, $\gamma_1^+/\gamma_1 = 1/2$, $\gamma_+/\gamma = 1/2$, $\sigma = 1$). Same genetic map as in Table 6 except that either 3 or 6 large effects QTLs are considered.

			($T = 10$, K =10,001)		
n	nb large QTLs	γ	AdaptSgenoLasso	AdaptSgenoEN	AdaptSgenoAdaptLasso
500	3	0.1*	52.56%	52.33%	52.33%
		0.3	53.85%	53.81%	54.64%
	6	0.1*	67.59%	67.64%	67.60%
		0.3	69.15%	69.20%	69.87%
1,000	3	0.1*	59.09%	59.06%	59.05%
		0.3	59.81%	59.83%	59.49%
	6	0.1*	70.80%	70.84%	70.77%
		0.3	71.35%	71.41%	71.67%

∗ SgenoLasso and AdaptSgenoLasso are a perfect match.

(4) Study of the LRT process.

Preliminaries

Let us compute the score function at a point $\theta_0^1 = (0,\ \mu,\ \sigma)$ that belongs to $\mathcal{H}_0$. We have the relationship

$$\frac{\partial l_t}{\partial q_1}\mid_{\theta_0^1} = \frac{Y-\mu}{\sigma^2}\ \{2p_1(t)-1\}\ 1_{Y\notin[S_-^1,S_+^1]} + \frac{Y-\mu}{\sigma^2}\ \{2p_2(t)-1\}\ 1_{Y\in[S_-^1,S_-^2]\cup[S_+^2,S_+^1]}$$
$$= \frac{\alpha_1(t)}{\sigma}\varepsilon\ \overline{X}(t^{\ell,1})\ +\ \frac{\beta_1(t)}{\sigma}\ \varepsilon\ \overline{X}(t^{r,1}) + \frac{\alpha_2(t)}{\sigma}\varepsilon\ \widetilde{X}(t^{\ell,2})\ +\ \frac{\beta_2(t)}{\sigma}\ \varepsilon\ \widetilde{X}(t^{r,2})$$

because of the key Lemma (Lemma 2.6 of [52]), which states that

$$\{2p_1(t)-1\}\,1_{Y\notin[S_-^1,S_+^1]} = \alpha_1(t)\overline{X}(t^{\ell,1}) + \beta_1(t)\overline{X}(t^{r,1})$$
$$\{2p_2(t)-1\}\,1_{Y\in[S_-^1,S_-^2]\cup[S_+^2,S_+^1]} = \alpha_2(t)\widetilde{X}(t^{\ell,2}) + \beta_2(t)\widetilde{X}(t^{r,2})\ .$$

Then, we have

$$\left(\frac{\partial l_t}{\partial q_1}\mid_{\theta_0^1}\right)^2 = \frac{\alpha_1^2(t)}{\sigma^2}\varepsilon^2\ 1_{Y\notin[S_-^1,S_+^1]}\ +\ \frac{\beta_1^2(t)}{\sigma^2}\ \varepsilon^2\ 1_{Y\notin[S_-^1,S_+^1]}$$
$$+\ 2\frac{\alpha_1(t)\beta_1(t)}{\sigma^2}\varepsilon^2 X(t^{\ell,1})X(t^{r,1})1_{Y\notin[S_-^1,S_+^1]} + \frac{\alpha_2^2(t)}{\sigma^2}\varepsilon^2\ 1_{Y\in[S_-^1,S_-^2]\cup[S_+^2,S_+^1]}$$
$$+\ \frac{\beta_2^2(t)}{\sigma^2}\ \varepsilon^2\ 1_{Y\in[S_-^1,S_-^2]\cup[S_+^2,S_+^1]} + 2\frac{\alpha_2(t)\beta_2(t)}{\sigma^2}\varepsilon^2 X(t^{\ell,2})X(t^{r,2})1_{Y\in[S_-^1,S_-^2]\cup[S_+^2,S_+^1]}$$

and

$$\mathrm{E}\left[\left(\frac{\partial l_t}{\partial q_1}\mid_{\theta_0^1}\right)^2\right] = \frac{\mathcal{A}_1}{\sigma^4}\xi_1^2(t) + \frac{\mathcal{A}_2}{\sigma^4}\xi_2^2(t)\ .$$

Indeed, by definition, according to [49], we have $\mathcal{A}_1 = \mathrm{E}_{\mathcal{H}_0}\left[(Y-\mu)^2 1_{Y\notin[S_-^1,S_+^1]}\right]$. In the same way, $\mathcal{A}_2 = \mathrm{E}_{\mathcal{H}_0}\left[(Y-\mu)^2 1_{Y\in[S_-^1,S_-^2]\cup[S_+^2,S_+^1]}\right]$.

To conclude, after some easy calculations, the Fisher information is the following diagonal matrix:

$$I_{\theta_0} = Diag\left[\frac{\mathcal{A}_1}{\sigma^4}\xi_1^2(t) + \frac{\mathcal{A}_2}{\sigma^4}\xi_2^2(t),\ \frac{1}{\sigma^2}\ ,\ \frac{2}{\sigma^2}\right]\ . \tag{24}$$

10.1. *Study under $\mathcal{H}_0$*

In what follows, we define the processes $V_{1,n}(.)$ and $V_{2,n}(.)$ in the following way:

$$\forall t_k \in \mathrm{T}_K^1\ \ V_{1,n}(t_k) := \frac{1}{\sqrt{n\mathcal{A}_1}}\sum_{j=1}^n (Y_j-\mu)\ \overline{X}_j(t_k)\ ,$$

$$\forall t_k \in \mathrm{T}_K^2\ \ V_{2,n}(t_k) := \frac{1}{\sqrt{n\mathcal{A}_2}}\sum_{j=1}^n (Y_j-\mu)\ \widetilde{X}_j(t_k)\ ,$$

$$V_{1,n}(t) := \left\{\alpha_1(t)V_{1,n}(t^{\ell,1}) + \beta_1(t)V_{1,n}(t^{r,1})\right\}/\xi_1(t)\ ,$$

$$V_{2,n}(t) := \left\{\alpha_2(t)V_{2,n}(t^{\ell,2}) + \beta_2(t)V_{2,n}(t^{r,2})\right\}/\xi_2(t)\ .$$

Let l_t^n denote the log likelihood at t, associated to n observations. We have

$$\frac{1}{\sqrt{n}}\frac{\partial l_t^n}{\partial q_1}\mid_{\theta_0^1} = \frac{\alpha_1(t)}{\sigma\sqrt{n}}\sum_{j=1}^n \varepsilon_j\ \overline{X}_j(t^{\ell,1})\ +\ \frac{\beta_1(t)}{\sigma\sqrt{n}}\ \sum_{j=1}^n \varepsilon_j\ \overline{X}_j(t^{r,1})$$
$$+\ \frac{\alpha_2(t)}{\sigma\sqrt{n}}\sum_{j=1}^n \varepsilon_j\ \widetilde{X}_j(t^{\ell,2})\ +\ \frac{\beta_2(t)}{\sigma\sqrt{n}}\ \sum_{j=1}^n \varepsilon_j\ \widetilde{X}(t^{r,2})$$
$$= \frac{\alpha_1(t)\sqrt{\mathcal{A}_1}}{\sigma^2}\ V_{1,n}(t^{\ell,1})\ +\ \frac{\beta_1(t)\sqrt{\mathcal{A}_1}}{\sigma^2}\ V_{1,n}(t^{r,1})$$
$$+\ \frac{\alpha_2(t)\sqrt{\mathcal{A}_2}}{\sigma^2}\ V_{2,n}(t^{\ell,2})\ +\ \frac{\beta_2(t)\sqrt{\mathcal{A}_2}}{\sigma^2}\ V_{2,n}(t^{r,2})\ . \tag{25}$$

According to formulae (3), (24) and (25), we obtain easily that

$$S_n(t) = \frac{\sqrt{\mathcal{A}_1}\ \xi_1(t)V_{1,n}(t)\ +\ \sqrt{\mathcal{A}_2}\ \xi_2(t)V_{2,n}(t)}{\sqrt{\mathcal{A}_1\ \xi_1^2(t) + \mathcal{A}_2\ \xi_2^2(t)}}\ ,$$

According to the proof of Theorem 2.5 of [52], we have:

$$\forall t_k \in \mathrm{T}_K^1\ \ V_{1,n}(t_k) \longrightarrow \mathcal{N}(0,1) \quad .$$

In the same way, we obtain easily that:

$$\forall t_k \in \mathrm{T}_K^2\ \ V_{2,n}(t_k) \longrightarrow \mathcal{N}(0,1)\ .$$

Furthermore, according to the proof of Theorem 2.5 of [52], we have:

$$\forall (t_k, t_{k'}) \in \mathrm{T}_K^1 \times \mathrm{T}_K^1\ \ \mathrm{Cov}\left(V_{1,n}(t_k), V_{1,n}(t_{k'})\right) = \rho(t_k, t_{k'})\ .$$

In the same way, we obtain easily that:

$$\forall (t_k, t_{k'}) \in \mathrm{T}_K^2 \times \mathrm{T}_K^2\ \ \mathrm{Cov}\left(V_{2,n}(t_k), V_{2,n}(t_{k'})\right) = \rho(t_k, t_{k'})\ .$$

Since $V_{1,n}(.)$ and $V_{2,n}(.)$ are interpolated processes, the convergence of $(V_{1,n}(t^{\ell,1}), V_{1,n}(t^{r,1}))$ and $(V_{2,n}(t^{\ell,2}), V_{2,n}(t^{r,2}))$, and the continuous mapping theorem, imply that

$$V_{1,n}(t) \longrightarrow \mathcal{N}(0,1)\ \ \text{and}\ \ V_{2,n}(t) \longrightarrow \mathcal{N}(0,1)\ .$$

As a consequence, according to the continuous mapping theorem

$$\forall t\ S_n(t) \longrightarrow \mathcal{N}(0,1)$$

which proves the convergence of finite-dimensional.

Let us now prove the weak convergence of the score process $S_n(.)$. Recall that the tightness and the convergence of finite-dimensional imply the weak convergence of the score process (see for instance Theorem 4.9 of [5]). Since we have already proved the convergence of finite-dimensional, let us focus on the tightness of the score process. Since $\xi_1(t)$, $\xi_2(t)$, $\alpha_1(t)$, $\alpha_2(t)$, $\beta_1(t)$ and $\beta_2(t)$ are continuous functions, each path of the process $S_n(.)$ is a continuous function on $[t_1, t_K]$.

Without loss of generalty, let us study the process $S_n(.)$ on the marker interval $[t_2, t_3]$, assuming $t_2 \notin \mathrm{T}_K^2$ and $t_3 \notin \mathrm{T}_K^2$. Besides, let us impose that $\{t_1, t_4\} \subset \mathrm{T}_K^2$. In other words, for locations t and t' that belong to $]t_2, t_3[$, we have $t'^{r,2} = t^{r,2} = t_4$, $t'^{\ell,2} = t^{\ell,2} = t_1$. and $t'^{\ell,1} = t^{\ell,1} = t_2$, $t'^{r,1} = t^{r,1} = t_3$.

Recall the modulus of continuity of a continous function $x(t)$ on $[t_2, t_3]$:

$$w_x(\delta) = \sup_{|t'-t|<\delta} \left|x(t') - x(t)\right| \quad \text{where} \quad 0 < \delta \leq t_3 - t_2.$$

According to Theorem 8.2 of Billingsley (1999), the score process is tight if and only if the two following conditions hold:

(1) the sequence $S_n(t_2)$ is tight.
(2) For each positive ε and η, there exists a δ, with $0 < \delta \leq t_3 - t_2$, and an integer n_0 such that $\mathrm{P}\left(w_{S_n}(\delta) \geq \eta\right) \leq \varepsilon \quad \forall n \geq n_0$.

According to Prohorov, the sequence $S_n(t_2)$ is tight. Then, 1) is verified. Besides, let us set

$$\forall i = 1, 2 \quad \tilde{\alpha_i}(t) = \alpha_i(t)/\sqrt{\mathcal{A}_1\xi_1^2(t) + \mathcal{A}_2\xi_2^2(t)},$$
$$\tilde{\beta_i}(t) = \beta_i(t)/\sqrt{\mathcal{A}_1\xi_1^2(t) + \mathcal{A}_2\xi_2^2(t)}.$$

First, we can notice that $\forall \delta$ such as $0 < \delta \leq t_3 - t_2$,

$$\begin{aligned} w_{S_n}(\delta) &= \sup_{|t'-t|<\delta} \left|S_n(t') - S_n(t)\right| \\ &= \sup_{|t'-t|<\delta} \left|\sqrt{\mathcal{A}_1}\left\{\tilde{\alpha_1}(t')V_{1,n}(t'^{\ell,1}) + \tilde{\beta_1}(t')V_{1,n}(t'^{r,1})\right\}\right. \\ &+ \sqrt{\mathcal{A}_2}\left\{\tilde{\alpha_2}(t')V_{2,n}(t'^{\ell,2}) + \tilde{\beta_2}(t')V_{2,n}(t'^{r,2})\right\} \\ &- \sqrt{\mathcal{A}_1}\left\{\tilde{\alpha_1}(t)V_{1,n}(t^{\ell,1}) + \tilde{\beta_1}(t)V_{1,n}(t^{r,1})\right\} \\ &- \left.\sqrt{\mathcal{A}_2}\left\{\tilde{\alpha_2}(t)V_{2,n}(t^{\ell,2}) + \tilde{\beta_2}(t)V_{2,n}(t^{r,2})\right\}\right|. \end{aligned} \tag{26}$$

Since $t'^{r,2} = t^{r,2} = t_4$, $t'^{\ell,2} = t^{\ell,2} = t_1$, $t'^{\ell,1} = t^{\ell,1} = t_2$ and $t'^{r,1} = t^{r,1} = t_3$, we have

$$\begin{aligned} w_{S_n}(\delta) &= \sup_{|t'-t|<\delta} \left|S_n(t') - S_n(t)\right| \\ &= \sup_{|t'-t|<\delta} \left|\sqrt{\mathcal{A}_1}\left\{\tilde{\alpha_1}(t') - \tilde{\alpha_1}(t)\right\}V_{1,n}(t'^{\ell,1}) + \sqrt{\mathcal{A}_2}\left\{\tilde{\alpha_2}(t') - \tilde{\alpha_2}(t)\right\}V_{2,n}(t'^{\ell,2})\right. \\ &+ \left.\sqrt{\mathcal{A}_1}\left\{\tilde{\beta_1}(t') - \tilde{\beta_1}(t)\right\}V_{1,n}(t'^{r,1}) + \sqrt{\mathcal{A}_2}\left\{\tilde{\beta_2}(t') - \tilde{\beta_2}(t)\right\}V_{2,n}(t'^{r,2})\right| \\ &\leq \sqrt{\mathcal{A}_1}\left\{w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta_1}}(\delta)\right\}\max\left(\left|V_{1,n}(t'^{\ell,1})\right|, \left|V_{1,n}(t'^{r,1})\right|\right) \\ &+ \sqrt{\mathcal{A}_2}\left\{w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta_2}}(\delta)\right\}\max\left(\left|V_{2,n}(t'^{\ell,2})\right|, \left|V_{2,n}(t'^{r,2})\right|\right) \\ &\leq \max\left\{2\sqrt{\mathcal{A}_1}\left\{w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta_1}}(\delta)\right\}\max\left(\left|V_{1,n}(t'^{\ell,1})\right|, \left|V_{1,n}(t'^{r,1})\right|\right),\right. \\ &\left.2\sqrt{\mathcal{A}_2}\left\{w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta_2}}(\delta)\right\}\max\left(\left|V_{2,n}(t'^{\ell,2})\right|, \left|V_{2,n}(t'^{r,2})\right|\right)\right\}. \end{aligned}$$

Since the events are independent,

$$\begin{aligned} &\mathrm{P}\left(\max\left\{2\sqrt{\mathcal{A}_1}\left\{w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta_1}}(\delta)\right\}\max\left(\left|V_{1,n}(t'^{\ell,1})\right|, \left|V_{1,n}(t'^{r,1})\right|\right),\right.\right. \\ &\left.\left.2\sqrt{\mathcal{A}_2}\left\{w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta_2}}(\delta)\right\}\max\left(\left|V_{2,n}(t'^{\ell,2})\right|, \left|V_{2,n}(t'^{r,2})\right|\right)\right\} \geq \eta\right) \\ &= 1 - \mathrm{P}\left(2\sqrt{\mathcal{A}_1}\left\{w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta_1}}(\delta)\right\}\max\left(\left|V_{1,n}(t'^{\ell,1})\right|, \left|V_{1,n}(t'^{r,1})\right|\right) \leq \eta\right) \\ &\times \mathrm{P}\left(2\sqrt{\mathcal{A}_2}\left\{w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta_2}}(\delta)\right\}\max\left(\left|V_{2,n}(t'^{\ell,2})\right|, \left|V_{2,n}(t'^{r,2})\right|\right) \leq \eta\right) \end{aligned}$$

Let us consider $0 < \varepsilon_1 < 1$ and $\eta > 0$. Since the sequence

$\max\left(\left|V_{1,n}(t'^{\ell,1})\right|, \left|V_{1,n}(t'^{r,1})\right|\right)$ is uniformly tight,

$$\exists M_1 > 0 \ \ \forall n \geq 1 \ \ \mathrm{P}\left(\max\left(\left|V_{1,n}(t'^{\ell,1})\right|, \left|V_{1,n}(t'^{r,1})\right|\right) \geq M_1\right) \leq \varepsilon_1 \ . \tag{27}$$

In other words,

$$\exists M_1 > 0 \ \ \forall n \geq 1 \ \ \mathrm{P}\left(\max\left(\left|V_{1,n}(t'^{\ell,1})\right|, \left|V_{1,n}(t'^{r,1})\right|\right) \leq M_1\right) \geq 1 - \varepsilon_1 \ . \tag{28}$$

In the same way, the sequence $\max\left(\left|V_{2,n}(t'^{\ell,2})\right|, \left|V_{2,n}(t'^{r,2})\right|\right)$ is uniformly tight and

$$\exists M_2 > 0 \ \ \forall n \geq 1 \ \ \mathrm{P}\left(\max\left(\left|V_{2,n}(t'^{\ell,2})\right|, \left|V_{2,n}(t'^{r,2})\right|\right) \geq M_2\right) \leq \varepsilon_1 \ . \tag{29}$$

In other words,

$$\exists M_2 > 0 \ \ \forall n \geq 1 \ \ \mathrm{P}\left(\max\left(\left|V_{2,n}(t'^{\ell,2})\right|, \left|V_{2,n}(t'^{r,2})\right|\right) \leq M_2\right) \geq 1 - \varepsilon_1 \ . \tag{30}$$

According to Heine's theorem, since $\tilde{\alpha_1}(t)$, $\tilde{\beta_1}(t)$, $\tilde{\alpha_2}(t)$ and $\tilde{\beta_2}(t)$ are continuous on the compact $[t_2, t_3]$, these functions are uniformly continuous. So,

$$\exists \delta \text{ such as } 0 < \delta < t_3 - t_2, \ w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta_1}}(\delta) < \frac{\eta}{2M_1\sqrt{\mathcal{A}_1}} \tag{31}$$

$$w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta_2}}(\delta) < \frac{\eta}{2M_2\sqrt{\mathcal{A}_2}}. \tag{32}$$

As a consequence, we have:

$$\mathrm{P}\left(2\sqrt{\mathcal{A}_1}\left\{w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta_1}}(\delta)\right\} \max\left(\left|V_{1,n}(t'^{\ell,1})\right|, \left|V_{1,n}(t'^{r,1})\right|\right) \leq \eta\right) \geq 1 - \varepsilon_1 \ .$$
$$\mathrm{P}\left(2\sqrt{\mathcal{A}_2}\left\{w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta_2}}(\delta)\right\} \max\left(\left|V_{2,n}(t'^{\ell,2})\right|, \left|V_{2,n}(t'^{r,2})\right|\right) \leq \eta\right) \geq 1 - \varepsilon_1 \ ,$$

Then,

$$\mathrm{P}\left(2\sqrt{\mathcal{A}_1}\left\{w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta_1}}(\delta)\right\} \max\left(\left|V_{1,n}(t'^{\ell,1})\right|, \left|V_{1,n}(t'^{r,1})\right|\right) \leq \eta\right)$$
$$\times \mathrm{P}\left(2\sqrt{\mathcal{A}_2}\left\{w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta_2}}(\delta)\right\} \max\left(\left|V_{2,n}(t'^{\ell,2})\right|, \left|V_{2,n}(t'^{r,2})\right|\right) \leq \eta\right) \geq (1 - \varepsilon_1)^2 \ .$$

As a result,

$$1 - \mathrm{P}\left(2\sqrt{\mathcal{A}_1}\left\{w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta_1}}(\delta)\right\} \max\left(\left|V_{1,n}(t'^{\ell,1})\right|, \left|V_{1,n}(t'^{r,1})\right|\right) \leq \eta\right)$$
$$\times \mathrm{P}\left(2\sqrt{\mathcal{A}_2}\left\{w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta_2}}(\delta)\right\} \max\left(\left|V_{2,n}(t'^{\ell,2})\right|, \left|V_{2,n}(t'^{r,2})\right|\right) \leq \eta\right) \leq 1 - (1 - \varepsilon_1)^2 \ .$$

Last, $\mathrm{P}\left(w_{S_n}(\delta) \geq \eta\right) \leq 1 - (1 - \varepsilon_1)^2$.

To conclude, we just have to set $\varepsilon := 1 - (1 - \varepsilon_1)^2$ to obtain the desired result. It concludes the proof of 2). As result, the score process is tight.

10.2. *Study under $\mathcal{H}_{a\vec{t}^\star}$*

Let us consider the local alternative $\mathcal{H}_{a\vec{t}^\star}$:

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial l_t^n}{\partial q_1}\mid_{\theta_0^1} &= \frac{\alpha_1(t)}{\sigma^2\sqrt{n}}\sum_{j=1}^{n}(Y_j-\mu)\ \overline{X}_j(t^{\ell,1}) \ +\ \frac{\beta_1(t)}{\sigma^2\sqrt{n}}\ \sum_{j=1}^{n}(Y_j-\mu)\ \overline{X}_j(t^{r,1})\\ &+\ \frac{\alpha_2(t)}{\sigma^2\sqrt{n}}\sum_{j=1}^{n}(Y_j-\mu)\ \widetilde{X}_j(t^{\ell,2}) \ +\ \frac{\beta_2(t)}{\sigma^2\sqrt{n}}\ \sum_{j=1}^{n}(Y_j-\mu)\ \widetilde{X}(t^{r,2})\\ &= \frac{\alpha_1(t)\sqrt{\mathcal{A}_1}}{\sigma^2}\ V_{1,n}(t^{\ell,1}) \ +\ \frac{\beta_1(t)\sqrt{\mathcal{A}_1}}{\sigma^2}\ V_{1,n}(t^{r,1})\\ &+\ \frac{\alpha_2(t)\sqrt{\mathcal{A}_2}}{\sigma^2}\ V_{2,n}(t^{\ell,2}) \ +\ \frac{\beta_2(t)\sqrt{\mathcal{A}_2}}{\sigma^2}\ V_{2,n}(t^{r,2})\end{aligned}$$

where

$$\forall t_k\in \mathrm{T}_K^1\ V_{1,n}(t_k) := \frac{1}{\sqrt{n\mathcal{A}_1}}\ \left\{\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\overline{X}_j(t_k) \ +\ \sum_{j=1}^{n}\sigma\varepsilon_j\ \overline{X}_j(t_k)\right\}\ ,$$

$$\forall t_k\in \mathrm{T}_K^2\ V_{2,n}(t_k) := \frac{1}{\sqrt{n\mathcal{A}_2}}\ \left\{\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\widetilde{X}_j(t_k) \ +\ \sum_{j=1}^{n}\sigma\varepsilon_j\ \widetilde{X}_j(t_k)\right\}\ .$$

By definition, we have the relationship $\mathcal{B} = \mathrm{E}_{\mathcal{H}_0}\left[(Y-\mu)^2 1_{Y\notin[S_-^2,S_+^2]}\right]$.
According to formula (2.9) of Supplement A of [53],

$$\frac{1}{\sqrt{n\mathcal{A}_1}}\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\ \overline{X}_j(t_k) \longrightarrow \sum_{s=1}^{m}\frac{a_s\rho(t_k,t_s^\star)\gamma_1}{\sqrt{\mathcal{A}_1}}\ . \tag{33}$$

In the same way, we have

$$\frac{1}{\sqrt{n\mathcal{B}}}\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\ X_j(t_k)1_{Y_j\notin[S_-^2,S_+^2]} \longrightarrow \sum_{s=1}^{m}\frac{a_s\rho(t_k,t_s^\star)\gamma}{\sqrt{\mathcal{B}}}\ .$$

As consequence, using the fact that $\gamma_2 := \gamma-\gamma_1$ and $\widetilde{X}(t_k) = X(t_k)1_{Y_j\notin[S_-^2,S_+^2]} - \overline{X}(t_k)$, we have

$$\frac{1}{\sqrt{n\mathcal{A}_2}}\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\ \widetilde{X}_j(t_k) \longrightarrow \sum_{s=1}^{m}\frac{a_s\rho(t_k,t_s^\star)\gamma_2}{\sqrt{\mathcal{A}_2}}\ . \tag{34}$$

Besides, according to formula (2.10) of Supplement A of [53],

$$\sum_{j=1}^{n}\frac{\sigma\varepsilon_j\overline{X}_j(t_k)}{\sqrt{n\mathcal{A}_1}} \longrightarrow \mathcal{N}\left(\frac{\sum_{s=1}^{m}\rho(t_s^\star,t_k)a_s}{\sqrt{\mathcal{A}_1}}\left\{z_{\gamma_1^+}\varphi(z_{\gamma_1^+}) - z_{1-\gamma_1^-}\varphi(z_{1-\gamma_1^-})\right\},1\right)\ , \tag{35}$$

$$\sum_{j=1}^{n}\frac{\sigma\varepsilon_j X_j(t_k)1_{Y_j\notin[S_-^2,S_+^2]}}{\sqrt{n\mathcal{B}}} \longrightarrow \mathcal{N}\left(\frac{\sum_{s=1}^{m}\rho(t_s^\star,t_k)a_s}{\sqrt{\mathcal{B}}}\left\{z_{\gamma^+}\varphi(z_{\gamma^+}) - z_{1-\gamma^-}\varphi(z_{1-\gamma^-})\right\},1\right)\ .$$

We have, using a technical proof present below formula (4.3) in Section 4 of Supplement A of [53],

$$\begin{aligned}
&\mathrm{Cov}\left(\sigma\varepsilon_j X_j(t_k)1_{Y_j\notin[S_-^2,S_+^2]},\ \sigma\varepsilon_j\overline{X}_j(t_k)\right)\\
&=\mathrm{E}\left(\sigma^2\varepsilon_j^2 1_{Y_j\notin[S_-^1,S_+^1]}\right)-\mathrm{E}\left(\sigma\varepsilon_j X_j(t_k)1_{Y_j\notin[S_-^2,S_+^2]}\right)\mathrm{E}\left(\sigma\varepsilon_j\overline{X}_j(t_k)\right)\\
&=\mathrm{E}\left(\sigma^2\varepsilon_j^2 1_{Y_j\notin[S_-^1,S_+^1]}\right)-\left[\left\{z_{\gamma^+}\varphi(z_{\gamma^+})-z_{1-\gamma^-}\varphi(z_{1-\gamma^-})\right\}\sum_{s=1}^m\rho(t_s^\star,t_k)q_s\right.\\
&\left.\times\left\{z_{\gamma_1^+}\varphi(z_{\gamma_1^+})-z_{1-\gamma_1^-}\varphi(z_{1-\gamma_1^-})\right\}\sum_{s=1}^m\rho(t_s^\star,t_k)q_s\right]+o(\max_{1\le s\le m}|q_s|^2)\\
&\longrightarrow\mathcal{A}_1\ .
\end{aligned}$$

As a consequence, we have

$$\begin{aligned}
&\sum_{j=1}^n\frac{\sigma\varepsilon_j X_j(t_k)1_{Y_j\notin[S_-^2,S_+^2]}}{\sqrt{n}}-\sum_{j=1}^n\frac{\sigma\varepsilon_j\overline{X}_j(t_k)}{\sqrt{n}}\longrightarrow\mathcal{N}\left(\sum_{s=1}^m\rho(t_s^\star,t_k)a_s\left\{z_{\gamma^+}\varphi(z_{\gamma^+})\right.\right.\\
&\left.\left.-z_{1-\gamma^-}\varphi(z_{1-\gamma^-})-z_{\gamma_1^+}\varphi(z_{\gamma_1^+})+z_{1-\gamma_1^-}\varphi(z_{1-\gamma_1^-})\right\},\mathcal{B}-\mathcal{A}_1\right)
\end{aligned}$$

Then, since by definition $\mathcal{A}_2=\mathcal{B}-\mathcal{A}_1$, we have :

$$\begin{aligned}
&\sum_{j=1}^n\frac{\sigma\varepsilon_j\widetilde{X}_j(t_k)}{\sqrt{n\mathcal{A}_2}}\longrightarrow\mathcal{N}\left(\frac{\sum_{s=1}^m\rho(t_s^\star,t_k)a_s}{\sqrt{\mathcal{A}_2}}\left\{z_{\gamma^+}\varphi(z_{\gamma^+})-z_{1-\gamma^-}\varphi(z_{1-\gamma^-})\right.\right.\\
&\left.\left.-z_{\gamma_1^+}\varphi(z_{\gamma_1^+})+z_{1-\gamma_1^-}\varphi(z_{1-\gamma_1^-})\right\},1\right)\ . \qquad (36)
\end{aligned}$$

Finally, we obtain using formulae (33), (34), (35) and (36)

$$\begin{aligned}
&\forall t_k\in\mathrm{T}_K^1\ V_{1,n}(t_k)\longrightarrow\mathcal{N}\left(\frac{\sqrt{\mathcal{A}_1}}{\sigma^2}\sum_{s=1}^m\rho(t_s^\star,t_k)a_s,1\right)\quad\text{and}\\
&\forall t_k\in\mathrm{T}_K^2\ V_{2,n}(t_k)\longrightarrow\mathcal{N}\left(\frac{\sqrt{\mathcal{A}_2}}{\sigma^2}\sum_{s=1}^m\rho(t_s^\star,t_k)a_s,1\right)\ .
\end{aligned}$$

As a consequence, using the interpolations :

$$S_n(t)\longrightarrow\mathcal{N}(\Omega,1) \qquad (37)$$

where

$$\begin{aligned}
\Omega=&\frac{\mathcal{A}_1\left\{\alpha_1(t)\sum_{s=1}^m\rho(t_s^\star,t^{\ell,1})a_s+\beta_1(t)\sum_{s=1}^m\rho(t_s^\star,t^{r,1})a_s\right\}}{\sigma^2\sqrt{\mathcal{A}_1\ \xi_1^2(t)+\mathcal{A}_2\ \xi_2^2(t)}}\\
&+\frac{\mathcal{A}_2\left\{\alpha_2(t)\sum_{s=1}^m\rho(t_s^\star,t^{\ell,2})a_s+\beta_2(t)\sum_{s=1}^m\rho(t_s^\star,t^{r,2})a_s\right\}}{\sigma^2\sqrt{\mathcal{A}_1\ \xi_1^2(t)+\mathcal{A}_2\ \xi_2^2(t)}}\ .
\end{aligned}$$

Study of the LRT process

Since the model with t fixed is regular, it is easy to prove that for fixed t

$$\Lambda_n(t) = S_n^2(t) + o_P(1) \tag{38}$$

under the null hypothesis.

Our goal is now to prove that the remainder is uniform in t.

Let us consider now t as an extra parameter. Let $t_1^\star$, $\theta_1^\star$ be the true parameter that will be assumed to belong to H_0. Note that $t_1^\star$ makes no sense for θ_1 belonging to H_0. It is easy to check that at H_0 the Fisher information relative to t is zero so that the model is not regular.

It can be proved that assumptions 1, 2 and 3 of [3] hold. So, we can apply Theorem 1 of [3] and we have

$$\sup_{(t,\theta)} l_t^n(\theta) - l_{t_1^\star}^n(\theta_1^\star) = \sup_{d\in\mathcal{D}} \left(\left(\frac{1}{\sqrt{n}} \sum_{j=1}^n d(X_j) \right)^2 1_{\sum_{j=1}^n d(X_j)\geq 0} \right) + o_P(1) \tag{39}$$

where the observation X_j stands for $\left(Y_j,\ \overline{X}_j(t^{\ell,1}),\ \widetilde{X}_j(t^{\ell,2}),\ \overline{X}_j(t^{r,1}),\ \widetilde{X}_j(t^{r,2})\right)$ and where $\mathcal{D}$ is the set of scores defined in [3], see also [21] and [4]. A similar result is true under H_0 with a set $\mathcal{D}_0$. Let us precise the sets of scores $\mathcal{D}$ and $\mathcal{D}_0$. These sets are defined at the sets of scores of one parameter families that converge to the true model $p_{t_1^\star,\theta_1^\star}$ and that are differentiable in quadratic mean.

It is easy to see that

$$\mathcal{D} = \left\{ \frac{\langle W, l_t'(\theta_1^\star)\rangle}{\sqrt{\mathrm{Var}_{H_0}\left(\langle W, l_t'(\theta_1^\star)\rangle\right)}}, W \in \mathrm{R}^3, t \in [t^{\ell,2}, t^{r,2}] \right\}$$

where l' is the gradient with respect to θ_1. In the same manner

$$\mathcal{D}_0 = \left\{ \frac{\langle W, l_t'(\theta_1^\star)\rangle}{\sqrt{\mathrm{Var}_{H_0}\left(\langle W, l_t'(\theta_1^\star)\rangle\right)}}, W \in \mathrm{R}^2 \right\},$$

where now the gradient is taken with respect to μ and σ only. Of course this gradient does not depend on t.

Using the transform $W \to -W$ in the expressions of the sets of score, we see that the indicator function can be removed in formula (39). Then, since the Fisher information matrix is diagonal (see formula 24) , it is easy to see that

$$\sup_{d\in\mathcal{D}} \left(\left(\frac{1}{\sqrt{n}} \sum_{j=1}^n d(X_j) \right)^2 \right) - \sup_{d\in\mathcal{D}_0} \left(\left(\frac{1}{\sqrt{n}} \sum_{j=1}^n d(X_j) \right)^2 \right)$$
$$= \sup_{t\in[t^{\ell,2},t^{r,2}]} \left(\left(\frac{1}{\sqrt{n}} \sum_{j=1}^n \frac{\frac{\partial l_t}{\partial q_1}(X_j) \mid_{\theta_0^1}}{\sqrt{\mathrm{Var}_{H_0}\left(\frac{\partial l_t}{\partial q_1}(X_j) \mid_{\theta_0^1}\right)}} \right)^2 \right).$$

This is exactly the desired result.

In other words, we have proved that under H_0:

$$\sup \Lambda_n(.) = \sup S_n^2(.) + o_P(1) \ . \tag{40}$$

Our goal is now to prove that it is also true under the alternative $\mathcal{H}_{a\vec{t}^\star}$.

Recall that K genetic markers are located at $0 = t_1 < t_2 < \ldots < t_K = T$ (i.e. on the map T_K^1). Besides, m QTLs lie on $[0,T]$ at locations $t_1^\star$, $t_2^\star$, ..., $t_m^\star$, that are distinct of marker locations. By definition $t_1^\star < t_2^\star < ... < t_m^\star$.

All the information is contained in the flanking markers of the QTLs locations, because of the Poisson process. As a consequence, let us compute the probability distribution of $\Big(Y, \overline{X}(t_1^{\star\ell,1}), \overline{X}(t_1^{\star r,1}), \ldots, \overline{X}(t_m^{\star\ell,1}), \overline{X}(t_m^{\star r,1}), \widetilde{X}(t_1^{\star\ell,2}), \widetilde{X}(t_1^{\star\ell,2}), \ldots, \widetilde{X}(t_m^{\star\ell,2}), \widetilde{X}(t_m^{\star r,2})\Big)$.

We have

$$\begin{aligned}
&\mathrm{P}(Y \in [y\ ,\ y+dy]\ ,\ Y \notin [S_-^1, S_+^1]\ ,\ \overline{X}(t_1^{\star\ell,1}), \overline{X}(t_1^{\star r,1}), \ldots, \overline{X}(t_m^{\star\ell,1}), \overline{X}(t_m^{\star r,1})) \\
&= \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \mathrm{P}(Y \in [y\ ,\ y+dy] \mid \overline{X}(t_1^\star) = u_1, \overline{X}(t_2^\star) = u_2, \ldots, \overline{X}(t_m^\star) = u_m) \\
&\times \mathrm{P}(\overline{X}(t_1^\star) = u_1, \overline{X}(t_2^\star) = u_2, \ldots, \overline{X}(t_m^\star) = u_m, \overline{X}(t_1^{\star\ell,1}), \overline{X}(t_1^{\star r,1}), \ldots, \overline{X}(t_m^{\star\ell,1}), \overline{X}(t_m^{\star r,1})) \ .
\end{aligned}$$

In the same way,

$$\begin{aligned}
&\mathrm{P}(Y \in [y\ ,\ y+dy]\ ,\ Y \in [S_-^1, S_-^2] \cup [S_+^2, S_+^1]\ ,\ \widetilde{X}(t_1^{\star\ell,2}), \widetilde{X}(t_1^{\star r,2}), \ldots, \widetilde{X}(t_m^{\star\ell,2}), \widetilde{X}(t_m^{\star r,2})) \\
&= \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \mathrm{P}(Y \in [y\ ,\ y+dy] \mid \widetilde{X}(t_1^\star) = u_1, \widetilde{X}(t_2^\star) = u_2, \ldots, \widetilde{X}(t_m^\star) = u_m) \\
&\times \mathrm{P}(\widetilde{X}(t_1^\star) = u_1, \widetilde{X}(t_2^\star) = u_2, \ldots, \widetilde{X}(t_m^\star) = u_m, \widetilde{X}(t_1^{\star\ell,2}), \widetilde{X}(t_1^{\star r,2}), \ldots, \widetilde{X}(t_m^{\star\ell,2}), \widetilde{X}(t_m^{\star r,2})) \ .
\end{aligned}$$

Besides,

$$\begin{aligned}
&\mathrm{P}(Y \in [y\ ,\ y+dy] \mid \overline{X}(t_1^\star) = u_1, \overline{X}(t_2^\star) = u_2, \ldots, \overline{X}(t_m^\star) = u_m) \\
&= \frac{\mathrm{P}(Y \in [y\ ,\ y+dy], Y \notin [S_-^1, S_+^1] \mid X(t_1^\star) = u_1, X(t_2^\star) = u_2, \ldots, X(t_m^\star) = u_m)}{\mathrm{P}(Y \notin [S_-^1, S_+^1] \mid X(t_1^\star) = u_1, X(t_2^\star) = u_2, \ldots, X(t_m^\star) = u_m)} \\
&= \frac{f_{(\mu+u_1q_1+u_2q_2+\ldots+u_mq_m,\sigma)}(y)\ 1_{y\notin[S_-^1,S_+^1]}}{\mathrm{P}(Y \notin [S_-^1, S_+^1] \mid X(t_1^\star) = u_1, X(t_2^\star) = u_2, \ldots, X(t_m^\star) = u_m)} \ .
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathrm{P}(\overline{X}(t_1^\star) = u_1, \overline{X}(t_2^\star) = u_2, \ldots, \overline{X}(t_m^\star) = u_m, \overline{X}(t_1^{\star\ell,1}), \overline{X}(t_1^{\star r,1}), \ldots, \overline{X}(t_m^{\star\ell,1}), \overline{X}(t_m^{\star r,1})) \\
&= \mathrm{P}(Y \notin [S_-^1, S_+^1] \mid X(t_1^\star) = u_1, X(t_2^\star) = u_2, \ldots, X(t_m^\star) = u_m) \\
&\mathrm{P}(X(t_1^\star) = u_1, X(t_2^\star) = u_2, \ldots, X(t_m^\star) = u_m, X(t_1^{\star\ell,1}), X(t_1^{\star r,1}), \ldots, X(t_m^{\star\ell,1}), X(t_m^{\star r,1})) \ .
\end{aligned}$$

As a result,

$$\begin{aligned}
&\mathrm{P}(Y \in [y\ ,\ y+dy]\ ,\ Y \notin [S_-^1, S_+^1]\ ,\ \overline{X}(t_1^{\star\ell,1}), \overline{X}(t_1^{\star r,1}), \ldots, \overline{X}(t_m^{\star\ell,1}), \overline{X}(t_m^{\star r,1})) \\
&= \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} f_{(\mu+u_1q_1+u_2q_2+u_mq_m,\sigma)}(y)\ 1_{y\notin[S_-^1,S_+^1]} \\
&\times \mathrm{P}(X(t_1^\star)=u_1, X(t_2^\star)=u_2, \ldots, X(t_m^\star)=u_m, X(t_1^{\star\ell,1}), X(t_1^{\star r,1}), \ldots, X(t_m^{\star\ell,1}), X(t_m^{\star r,1}))\ .
\end{aligned}$$

In the same way, we have:

$$\begin{aligned}
&\mathrm{P}(Y \in [y\ ,\ y+dy]\ ,\ Y \in [S_-^1, S_-^2] \cup [S_+^2, S_+^1]\ ,\ \widetilde{X}(t_1^{\star\ell,2}), \widetilde{X}(t_1^{\star r,2}), \ldots, \widetilde{X}(t_m^{\star\ell,2}), \widetilde{X}(t_m^{\star r,2})) \\
&= \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} f_{(\mu+u_1q_1+u_2q_2+u_mq_m,\sigma)}(y)\ 1_{y\in[S_-^1,S_-^2]\cup[S_+^2,S_+^1]} \\
&\times \mathrm{P}(X(t_1^\star)=u_1, X(t_2^\star)=u_2, \ldots, X(t_m^\star)=u_m, X(t_1^{\star\ell,2}), X(t_1^{\star r,2}), \ldots, X(t_m^{\star\ell,2}), X(t_m^{\star r,2}))\ .
\end{aligned}$$

Moreover, when the genome information is missing at marker locations (i.e. the phenotype is not extreme), we find

$$\begin{aligned}
&\mathrm{P}\Big(Y \in [y\ ,\ y+dy], \overline{X}(t_1^{\star\ell,1})=0, \overline{X}(t_1^{\star r,1})=0, \ldots, \overline{X}(t_m^{\star\ell,1})=0, \overline{X}(t_m^{\star r,1})=0, \\
&\widetilde{X}(t_1^{\star\ell,2})=0, \widetilde{X}(t_1^{\star r,2})=0, \ldots, \widetilde{X}(t_m^{\star\ell,2})=0, \widetilde{X}(t_m^{\star r,2})=0\Big) \qquad (41) \\
&= \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \\
&\mathrm{P}(Y \in [y\ ,\ y+dy], Y \in [S_-^2, S_+^2], X(t_1^\star)=u_1, X(t_2^\star)=u_2, \ldots, X(t_m^\star)=u_m) \\
&= \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \\
&f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(y)\ 1_{y\in[S_-^2,S_+^2]}\ \mathrm{P}(X(t_1^\star)=u_1, X(t_2^\star)=u_2, \ldots, X(t_m^\star)=u_m)\ .
\end{aligned}$$

Let $\theta^m = (q_1, ..., q_m, \mu, \sigma)$ denote the new parameter. Then, the probability distribution of $\Big(Y, \overline{X}(t_1^{\star\ell,1}), \overline{X}(t_1^{\star r,1}), \widetilde{X}(t_1^{\star\ell,2}), \widetilde{X}(t_1^{\star r,2}), \ldots, \overline{X}(t_m^{\star\ell,1}), \overline{X}(t_m^{\star r,1}), \widetilde{X}(t_m^{\star\ell,2}), \widetilde{X}(t_m^{\star r,2})\Big)$, with respect to the measure $\lambda \otimes N \otimes \ldots \otimes N$, is

$$\begin{aligned}
L_{\vec{t}^\star}^m(\theta^m) = \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} &\Big[w_{\vec{t}^\star}^1(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)\ 1_{Y\notin[S_-^1,S_+^1]} \\
&+ w_{\vec{t}^\star}^2(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)\ 1_{Y\in[S_-^1,S_-^2]\cup[S_+^2,S_+^1]} \\
&+ v_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)\ 1_{Y\in[S_-^2,S_+^2]}\Big]\ g^m(t_1^\star, \ldots, t_m^\star) \qquad (42)
\end{aligned}$$

with

$$
\begin{aligned}
&w^1_{\vec{t^\star}}(u_1,\ ...,\ u_m)\\
&= \mathrm{P}(X(t_1^\star)=u_1, X(t_2^\star)=u_2,\ldots,X(t_m^\star)=u_m \mid X(t_1^{\star\ell,1}), X(t_1^{\star r,1}),\ldots,X(t_m^{\star\ell,1}), X(t_m^{\star r,1}))\ ,\\
&w^2_{\vec{t^\star}}(u_1,\ ...,\ u_m)\\
&= \mathrm{P}(X(t_1^\star)=u_1, X(t_2^\star)=u_2,\ldots,X(t_m^\star)=u_m \mid X(t_1^{\star\ell,2}), X(t_1^{\star r,2}),\ldots,X(t_m^{\star\ell,2}), X(t_m^{\star r,2}))\ ,\\
&v_{\vec{t^\star}}(u_1,\ ...,\ u_m) = \mathrm{P}(X(t_1^\star)=u_1, X(t_2^\star)=u_2,\ldots,X(t_m^\star)=u_m)
\end{aligned}
$$

and

$$
\begin{aligned}
g^m(t_1^\star,\ldots,t_m^\star) = \mathrm{P}(X(t_1^{\star\ell,1}), X(t_1^{\star r,1}),\ldots,X(t_m^{\star\ell,1}), X(t_m^{\star r,1}))\ 1_{Y\notin[S^1_-,S^1_+]}\ +\ 1_{Y\in[S^2_-,S^2_+]}\\
+\ \mathrm{P}(X(t_2^{\star\ell,2}), X(t_1^{\star r,2}),\ldots,X(t_m^{\star\ell,2}), X(t_m^{\star r,2}))\ 1_{Y\in[S^1_-,S^2_-]\cup[S^2_+,S^1_+]}\ .
\end{aligned}
$$

Let us define the parameter θ_0^m in the following way : $\theta_0^m = (0,...,0,\mu,\sigma)$.

The likelihood $L^{m,n}_{\vec{t^\star}}(\theta^m)$ for n observations is obtained by the product of n terms as in formula (42) above. Let Q_n and P_n be two sequences of probability measures defined on the same space $(\Omega_n,\ \mathcal{A}_n)$. Q_n (respectively P_n) is the probability distribution with density $L^{m,n}_{\vec{t^\star}}(\theta^m)$ (respectively $L^{m,n}_{\vec{t^\star}}(\theta_0^m)$).

In what follows, $\log\frac{dQ_n}{dP_n}$ will denote the log likelihood ratio. By definition, we have the relationship,

$$
\log\frac{dQ_n}{dP_n} = \log\left\{\frac{L^{m,n}_{\vec{t^\star}}(\theta^m)}{L^{m,n}_{\vec{t^\star}}(\theta_0^m)}\right\}\ . \tag{43}
$$

Since the model is differentiable in quadratic mean at θ^m and according to the central limit theorem :

$$
\log\left(\frac{dQ_n}{dP_n}\right) \overset{\mathcal{H}_0}{\to} \mathcal{N}(-\frac{1}{2}\vartheta^2,\ \vartheta^2)\ \text{ with } \vartheta^2\in \mathrm{R}^{+\star}\ .
$$

As a result, according to iii) of Le Cam's first lemma, we have $Q_n \triangleleft P_n$, that is to say the sequence Q_n is contiguous with respect to the sequence P_n. Then, formula (40) is also true under the alternative $\mathcal{H}_{a\vec{t^\star}}$.

It concludes the proof of Theorem 4.1. ■

11. Proof of the skeleton of the covariance function of $Z(.)$

Using formulae (11) and (10), we obtain easily the following relationships:

$$
\forall (t_k,t_{k'})\in \mathrm{T}_K^2\times\mathrm{T}_K^2 \quad \mathrm{Cov}\,(Z(t_k),Z(t_{k'})) = \rho(t_k,t_{k'})\ ,
$$

$$\forall (t_k, t_{k'}) \in \mathrm{T}_K^1 \backslash \mathrm{T}_K^2 \times \mathrm{T}_K^1 \backslash \mathrm{T}_K^2$$
$$\mathrm{Cov}\left(Z(t_k), Z(t_{k'})\right) = \Big[\mathcal{A}_1 \rho(t_k, t_{k'}) + \mathcal{A}_2 \Big\{ \alpha_2(t_k)\alpha_2(t_{k'})\rho(t_k^{\ell,2}, t_{k'}^{\ell,2}) + \alpha_2(t_k)\beta_2(t_{k'})\rho(t_k^{\ell,2}, t_{k'}^{r,2}) + \beta_2(t_k)\alpha_2(t_{k'})\rho(t_k^{r,2}, t_{k'}^{\ell,2}) + \beta_2(t_k)\beta_2(t_{k'})\rho(t_k^{r,2}, t_{k'}^{r,2}) \Big\} \Big] / \sqrt{\left\{\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_k)\right\}\left\{\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_{k'})\right\}} \ .$$

Besides, since

$$\alpha_2(t_{k'})\rho(t_k^{\ell,2}, t_{k'}^{\ell,2}) + \beta_2(t_{k'})\rho(t_k^{\ell,2}, t_{k'}^{r,2}) = \rho(t_k^{\ell,2}, t_{k'})$$
$$\alpha_2(t_{k'})\rho(t_k^{r,2}, t_{k'}^{\ell,2}) + \beta_2(t_{k'})\rho(t_k^{r,2}, t_{k'}^{r,2}) = \rho(t_k^{r,2}, t_{k'}),$$

then,

$$\mathrm{Cov}\left(Z(t_k), Z(t_{k'})\right) = \frac{\mathcal{A}_1 \rho(t_k, t_{k'}) + \mathcal{A}_2 \left\{ \alpha_2(t_k)\rho(t_k^{\ell,2}, t_{k'}) + \beta_2(t_k)\rho(t_k^{r,2}, t_{k'}) \right\}}{\sqrt{\left\{\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_k)\right\}\left\{\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_{k'})\right\}}} \ .$$

Last, we have $\forall (t_k, t_{k'}) \in \mathrm{T}_K^2 \times \mathrm{T}_K^1 \backslash \mathrm{T}_K^2$

$$\mathrm{Cov}\left(Z(t_k), Z(t_{k'})\right) = \frac{\mathcal{A}_1 \rho(t_k, t_{k'}) + \mathcal{A}_2 \left\{ \alpha_2(t_{k'})\rho(t_k, t_{k'}^{\ell,2}) + \beta_2(t_{k'})\rho(t_k, t_{k'}^{r,2}) \right\}}{\sqrt{\mathcal{B}(\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_{k'}))}}$$
$$= \frac{\sqrt{\mathcal{B}}\rho(t_k, t_{k'})}{\sqrt{\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_{k'})}} \ .$$

12. Proof of Theorem 4.2

Let us consider $n^\star$ individuals for an experiment under a selective genotyping that varies along the genome. Recall that n is the number of individuals under the complete data situation ([2]), and also that $q_1 = a/\sqrt{n}$, ..., $q_m = a_m/\sqrt{n}$. In this context, let ζ be the quantity such as $\zeta = \frac{n^\star}{n}$. Then, using formula (37), we obtain easily that when $t \notin \mathrm{T}_K^1$,

$$S_{n^\star}(t) \longrightarrow \mathcal{N}\left(\sqrt{\zeta}\ \Omega, 1\right) \ .$$

where Ω is given in formula (37).

Under the complete data situation ([2]), we have $S_-^1 = S_-^2 = S_+^2 = S_+^1$, so that $\mathcal{A}_2 = 0$ and $\mathcal{A}_1 = \mathcal{B} = \sigma^2$. As a result,

$$S_n(t) \longrightarrow \mathcal{N}\left(\frac{\alpha_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{\ell,1})a_s + \beta_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{r,1})a_s}{\sigma\sqrt{\xi_1^2(t)}}, 1\right) \ .$$

As a consequence, if we suppose $\forall s\ a_s > 0$ and consider a one sided test, the statistical test for the selective genotyping that varies along the genome is more

powerful than the one regarding the complete data situation, as soon as

$$z_\alpha \; - \; \sqrt{\zeta}\,\Omega < \; z_\alpha \; - \frac{\left\{\alpha_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{\ell,1})a_s + \beta_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{r,1})a_s\right\}}{\sigma\sqrt{\xi_1^2(t)}}$$

$$\Leftrightarrow \zeta > \frac{\left\{\alpha_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{\ell,1})a_s + \beta_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{r,1})a_s\right\}^2}{\sigma^2\;\Omega^2\;\xi_1^2(t)} \; .$$

As a result, the efficiency κ is equal to $\frac{\sigma^2\;\Omega^2\;\xi_1^2(t)}{\left\{\alpha_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{\ell,1})a_s + \beta_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{r,1})a_s\right\}^2}$. It proves i). The cases ii) (i.e. $t_k \in \mathrm{T}_K^1 \backslash \mathrm{T}_K^2$) and iii) ($t_k \in \mathrm{T}_K^2$) can easily be obtained by continuity.

Proof of Remark 2 of Section 4.4: In order to make the results general, let us consider the case $t \notin \mathrm{T}_K^1$. To begin with, let us replace the term $\mathcal{A}_2$ by $\mathcal{B} - \mathcal{A}_1$ in the expression of the efficiency κ (see above). We have

$$\begin{aligned}
\Omega^2 = \Bigg[& \frac{\mathcal{A}_1^2\left\{\alpha_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{\ell,1})a_s + \beta_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{r,1})a_s\right\}^2}{\sigma^4\left\{\mathcal{A}_1\;\xi_1^2(t) + (\mathcal{B}-\mathcal{A}_1)\;\xi_2^2(t)\right\}} \\
& + \frac{(\mathcal{B}-\mathcal{A}_1)^2\left\{\alpha_2(t)\sum_{s=1}^m \rho(t_s^\star, t^{\ell,2})a_s + \beta_2(t)\sum_{s=1}^m \rho(t_s^\star, t^{r,2})a_s\right\}^2}{\sigma^4\left\{\mathcal{A}_1\;\xi_1^2(t) + (\mathcal{B}-\mathcal{A}_1)\;\xi_2^2(t)\right\}} \\
& + 2\frac{\mathcal{A}_1(\mathcal{B}-\mathcal{A}_1)\left\{\alpha_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{\ell,1})a_s + \beta_1(t)\sum_{s=1}^m \rho(t_s^\star, t^{r,1})a_s\right\}}{\sigma^4\left\{\mathcal{A}_1\;\xi_1^2(t) + (\mathcal{B}-\mathcal{A}_1)\;\xi_2^2(t)\right\}} \\
& \times \left\{\alpha_2(t)\sum_{s=1}^m \rho(t_s^\star, t^{\ell,2})a_s + \beta_2(t)\sum_{s=1}^m \rho(t_s^\star, t^{r,2})a_s\right\}\Bigg] \; .
\end{aligned}$$

We have to answer the following question : how must we choose γ_1^+, γ_1^-, γ^+ and γ^- to maximize the efficiency ? Recall that by definition, $\gamma_1^+ + \gamma_1^- = \gamma_1$, $\gamma^+ + \gamma^- = \gamma$ and $\gamma_1 \leq \gamma$, $\gamma_1^+ \leq \gamma^+$, $\gamma_1^- \leq \gamma^-$. Recall also that $\varphi(.)$ denote the density of the standard normal distribution. Moreover, let $\Phi(.)$ denote the cumulative distribution of the standard normal distribution, and let $u_1(.)$ be the function such as: $u_1(z_{\gamma_1^+}) = \Phi^{-1}\left\{\gamma_1 - 1 + \Phi(z_{\gamma_1^+})\right\}$. Then, $z_{1-\gamma_1^-} = u_1(z_{\gamma_1^+})$. In the same way, let $u(.)$ be the function such as : $u(z_{\gamma_+}) = \Phi^{-1}\left\{\gamma - 1 + \Phi(z_{\gamma_+})\right\}$. Then, $z_{1-\gamma_-} = u(z_{\gamma_+})$.

Let $k_1(.)$ be the following function : $k_1(z_{\gamma_1^+}) = z_{\gamma_1^+}\varphi(z_{\gamma_1^+}) - u(z_{\gamma_1^+})\;\varphi\left\{u(z_{\gamma_1^+})\right\}$.
We have $\mathcal{A}_1 = \sigma^2\left\{\gamma_1 + k_1(z_{\gamma_1^+})\right\}$ and we have

$$k_1'(z_{\gamma_1^+}) = \varphi(\gamma_1^+) \; + \; z_{\gamma_1^+}\varphi'(z_{\gamma_1^+}) \; - \; u_1'(z_{\gamma_1^+})\;\varphi\left\{u_1(z_{\gamma_1^+})\right\} \; - \; u_1(z_{\gamma_1^+})\;u_1'(z_{\gamma_1^+})\;\varphi'\left\{u_1(z_{\gamma_1^+})\right\} \; ,$$

$$u_1'(z_{\gamma_1^+}) = \frac{\varphi(z_{\gamma_1^+})}{\varphi(z_{1-\gamma_1^-})} \; .$$

Then, we have

$$k_1'\left(z_{\gamma_+}\right) = \varphi\left(z_{\gamma_+}\right)\left(z_{1-\gamma_1^-}^2 - z_{\gamma_1^+}^2\right) .$$

As a result, when $\gamma_1^+ = \gamma_1/2$, we have $k_1'(z_{\gamma_1/2}) = 0$. Besides, when $\gamma_1^+ = 0$, we have $z_{\gamma_1^+} = +\infty$ and $k_1'(z_{\gamma_1^+}) = 0$.

In the same way, let $k(.)$ be the following function : $k(z_{\gamma^+}) = z_{\gamma^+}\varphi(z_{\gamma^+}) - u(z_{\gamma^+})\,\varphi\left\{u(z_{\gamma^+})\right\}$. We have $\mathcal{B} = \sigma^2\left\{\gamma + k(z_{\gamma^+})\right\}$ and as before, $k'(z_{\gamma/2}) = 0$, and $k'(z_{\gamma^+}) = 0$ when $\gamma^+ = 0$.

Let us rewrite Ω^2 as the function $\Omega^2(z_{\gamma^+}, z_{\gamma_1^+})$. Next, after straightforward calculations, we obtain:

$$\frac{\partial\Omega^2}{\partial z_{\gamma_1^+}}\,|_{(z,z_{\gamma_1/2})} = 0\ ,\ \frac{\partial\Omega^2}{\partial z_{\gamma^+}}\,|_{(z_{\gamma/2},z)} = 0\ ,\ \frac{\partial\Omega^2}{\partial z_{\gamma_1^+}}\,|_{(z,+\infty)} = 0\ ,\ \frac{\partial\Omega^2}{\partial z_{\gamma^+}}\,|_{(+\infty,z)} = 0\ .$$

As a result, the setting $\gamma_+/\gamma = \frac{1}{2}$ and $\gamma_1^+/\gamma_1 = \frac{1}{2}$, and the setting $\gamma^+/\gamma = 1$ and $\gamma_1^+/\gamma_1 = 1$ are optimums of the function.

13. Comparison between selective genotyping that varies along the genome, the classical selective genotyping, and the complete data situation

Recall that n and $n^\star$ denote respectively the number of individuals under the complete data situation ([2]) and under the selective genotyping that varies along the genome. Recall also that κ refers to the efficiency for the selective genotyping that varies along the genome (see in Theorem 4.2). Then, assuming that phenotyping is free, the selective genotyping that varies along the genome is more interesting than the complete data situation, as soon as we have:

$$\begin{aligned} & n^\star\gamma_1 K + n^\star(\gamma - \gamma_1)\#\mathrm{T}_K^2 < nK \\ \Leftrightarrow\ & \kappa > \gamma_1 + (\gamma - \gamma_1)\frac{\#\mathrm{T}_K^2}{K}\ . \end{aligned}$$

In the same way, let $\widetilde{n}$ be the number of individuals required under the classical selective genotyping situation, in order to reach the same power as under the complete data situation. $\kappa_{\mathrm{clSgeno}}$ will denote the associated efficiency. Then, the selective genotyping that varies along the genome is more interesting than the classical selective genotyping as soon as we have

$$\begin{aligned} & n^\star\gamma_1 K + n^\star(\gamma - \gamma_1)\#\mathrm{T}_K^2 < \widetilde{n}\gamma_1 K \\ \Leftrightarrow\ & \kappa > \kappa_{\mathrm{clSgeno}}\left\{1 + \frac{(\gamma - \gamma_1)\#\mathrm{T}_K^2}{\gamma_1 K}\right\} \\ \Leftrightarrow\ & \kappa > \frac{\kappa_{\mathrm{clSgeno}}}{\gamma_1}\left\{\gamma_1 + \frac{(\gamma - \gamma_1)\#\mathrm{T}_K^2}{K}\right\} \\ \Leftrightarrow\ & \kappa > \left\{1 + \frac{z_{\gamma_1^+}\varphi(z_{\gamma_1^+}) - z_{1-\gamma_1^-}\varphi(z_{1-\gamma_1^-})}{\gamma_1}\right\}\left\{\gamma_1 + \frac{(\gamma - \gamma_1)\#\mathrm{T}_K^2}{K}\right\}\ . \end{aligned}$$

The last inequality is obtained by replacing $\kappa_{\mathrm{clSgeno}}$ by its expression, $\gamma_1 + z_{\gamma_1^+}\varphi(z_{\gamma_1^+}) - z_{1-\gamma_1^-}\varphi(z_{1-\gamma_1^-})$, given in Rabier [52].

References

[1] Azaïs, J.M. and Cierco-Ayrolles, C., 2002. An asymptotic test for quantitative gene detection. *Ann. Inst. Henri Poincaré (B)*, **38(6)** 1087-1092.

[2] Azaïs, J.M., Delmas, C., and Rabier, C.E. (2012). Likelihood ratio test process for Quantitative Trait Locus detection. *Statistics*, **48(4)** 787-801.

[3] Azaïs, J.M., Gassiat, E., and Mercadier, C. (2006). Asymptotic distribution and local power of the likelihood ratio test for mixtures. *Bernoulli*, **12(5)** 775-799.

[4] Azaïs, J.M., Gassiat, E., and Mercadier, C. (2009). The likelihood ratio test for general mixture models with possibly structural parameter. *ESAIM*, **13** 301-327.

[5] Azaïs, J.M. and Wschebor, M. (2009). *Level sets and extrema of random processes and fields*. Wiley, New-York.

[6] Arias-Castro, E., Candes, E.J., and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, **39(5)** 2533-2556.

[7] Auinger, H. J., Schonleben, M., Lehermeier, C., Schmidt, M., Korzun, V., Geiger, H. H., ... Schön, C.C. (2016). Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (Secale cereale L.). *Theor. Appl. Genet.*, **129(11)**, 2043-2053.

[8] Bastide, H., Betancourt, A., Nolte, V., Tobler, R., Stöbe, P., Futschik, A., Schlötterer, C. (2013). A genome-wide, fine-scale map of natural pigmentation variation in Drosophila melanogaster. *PLoS genetics,* **9(6)**, e1003534.

[9] Begum, H., Spindel, J.E., Lalusin, A., Borromeo, T., Gregorio, G., Hernandez, J., ..., and McCouch, S.R. (2015). Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa). *PloS one*, **10(3)** e0119873.

[10] Boligon, A.A., Long, N., Albuquerque, L.G.D., Weigel, K.A., Gianola, D., and Rosa, G.J.M. (2012). Comparison of selective genotyping strategies for prediction of breeding values in a population undergoing selection. *Journal of animal science*, **90(13)** 4716-4722.

[11] Brandariz, S. P., Bernardo, R. (2018). Maintaining the Accuracy of Genomewide Predictions when Selection Has Occurred in the Training Population. *Crop Science*, **58**, (3), 1226-1231.

[12] Broman, K. and Speed T. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64(4)** 641-656.

[13] Bühlmann, P. and Van de Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications, Springer Science.

[14] Chang, M.N., Wu, R., Wu, S.S., and Casella, G. (2009). Score statistics for mapping quantitative trait loci. *Stat. Appl. Genet. Mol. Biol.*, **8**, (1), 16.

[15] Chen, Z., and Chen, H. (2005). On some statistical aspects of the interval mapping for QTL detection. *Statistica Sinica*, **15** 909-925.

[16] Cierco, C. (1998). Asymptotic distribution of the maximum likelihood ratio test for gene detection. *Statistics*, **31** 261-285.

[17] Cordoba, S., Balcells, I., Castello, A., Ovilo, C., Noguera, J. L., Timoneda, O., Sanchez, A. (2015). Endometrial gene expression profile of pregnant sows with extreme phenotypes for reproductive efficiency. *Scientific reports*, **5**, 14416.

[18] Darvasi D. and Soller M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theor. Appl. Genet.*, **85** 353-359.

[19] Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space *Journal of the Royal Statistical Society: Series B*, **70(5)** 849-911.

[20] Ferrao, L. F. V., Ferrao, R. G., Ferrao, M. A. G., Fonseca, A., Carbonetto, P., Stephens, M., Garcia, A. A. F. (2018). Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. *Heredity*.

[21] Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. Henri Poincaré (B)*, **6** 897-906.

[22] Gezan, S. A., Osorio, L. F., Verma, S., Whitaker, V. M. (2017). An experimental validation of genomic selection in octoploid strawberry. *Horticulture research*. **4**, 16070.

[23] Gutierrez, A., Hoy, J., Kimbeng, C., Baisakh, N. (2018). Identification of genomic regions controlling leaf scald resistance in sugarcane using a bi-parental mapping population and selective genotyping by sequencing. *Frontiers in plant science*, **9**, 877.

[24] Haldane, J.B.S. (1919). The combination of linkage values and the calculation of distance between the loci of linked factors. *Journal of Genetics*, **8** 299-309.

[25] Hayes, B., Bowman, P., Chamberlain, A. & Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science*. **92**, (2), 433-443.

[26] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning theory*. Springer, New York.

[27] Hayes, B (2007). QTL Mapping, MAS, and Genomic Selection. *Short course organized by Iowa State University*.

[28] Fan, J., Li, Q., and Wang, Y. (2017). 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)*, **79(1)** 247-265.

[29] Fernandes, G. R., Massironi, S. M., Pereira, L. V. (2016). Identification of Loci Modulating the Cardiovascular and Skeletal Phenotypes of Marfan Syndrome in Mice. *Scientific reports*, **6**, 22426.

[30] Kurz, J. P., Yang, Z., Weiss, R. B., Wilson, D. J., Rood, K. A., Liu, G. E., Wang, Z. (2019). A genome-wide association study for mastitis resistance in phenotypically well-characterized Holstein dairy cattle using a selective genotyping approach. *Immunogenetics*, **71**, (1), 35-47.

[31] Lander, E.S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **138** 235-240.

[32] Lebowitz, R.J., Soller, M., and Beckmann, J.S. (1987). Trait-based analyses for the detection of

linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theor. Appl. Genet.*, **73** 556-562.

[33] Lynch, M., and Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sinauer Sunderland, MA.

[34] Manichaikul, A., Palmer, A., Sen, S., and Broman, K. (2007). Significance thresholds for Quantitative Trait Locus mapping under selective genotyping. *Genetics*, **177** 1963-1966.

[35] Meuwissen, T.H., Hayes, B. & Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. **157**, (4), 1819-1829.

[36] Minamikawa, M. F., Takada, N., Terakami, S., Saito, T., Onogi, A., Kajiya-Kanegae, H., ... Iwata, H. (2018). Genome-wide association study and genomic prediction using parental and breeding populations of Japanese pear (Pyrus pyrifolia Nakai). *Scientific reports*. **8(1)**, 11994.

[37] Momen, M., Mehrgardi, A. A., Sheikhi, A., Kranis, A., Tusell, L., Morota, G., ... Gianola, D. (2018). Predictive ability of genome-assisted statistical models under various forms of gene action. *Scientific reports*. **8**.

[38] Muranty, H. and Goffinet, B. (1997). Selective genotyping for location and estimation of the effect of the effect of a quantitative trait locus. *Biometrics*, **53** 629-643.

[39] Muranty, H., Troggio, M., Sadok, I. B., ... Kumar, S. (2015). Accuracy and responses of genomic selection on key traits in apple breeding. *Horticulture research*. **2**, 15060.

[40] Neyhart, J. L., Tiede, T., Lorenz, A. J., Smith, K. P. (2017). Evaluating methods of updating training data in long-term genomewide selection. *G3: Genes, Genomes, Genetics*, **7**, (5), 1499-1510.

[41] Nyine, M., Uwimana, B., Blavet, N., ... Dolezel, J. (2018). Genomic prediction in a multiploid crop: genotype by environment interaction and allele dosage effects on predictive ability in banana. *The Plant Genome*. **11(2)**, 170090.

[42] Ohlson, E. W., Ashrafi, H., Foolad, M. R. (2018). Identification and Mapping of Late Blight Resistance Quantitative Trait Loci in Tomato Accession PI 163245. *The plant genome*.

[43] Park, T., Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103**, (482), 681-686.

[44] Phansak, P., Soonsuwon, W., Hyten, D.L., ..., and Specht, J.E. (2016). Multi-population selective genotyping to identify soybean (Glycine max (L.) Merr.) seed protein and oil QTLs. *G3: Genes, Genomes, Genetics*, **6** 1635.

[45] Pszczola, M., Calus, M. P. L. (2016). Updating the reference population to achieve constant genomic prediction reliability across generations. *animal*, **10**, (6), 1018-1024.

[46] Rabbee, N., Speca, D., Armstrong, N., and Speed, T. (2004). Power calculations for selective genotyping in QTL mapping in backcross mice. *Genet. Res. Camb.*, **84** 103-108.

[47] Rabier, C.E., Barre, P., Asp, T., Charmet, G. & Mangin, B. (2016). On the Accuracy of Genomic Selection. *PloS One*. **11**, (6), e0156086. doi:10.1371/ journal.pone.0156086.

[48] Rabier, C.E., Mangin, B. & Grusea, S. (2019). On the accuracy in high dimensional linear models and its application to genomic selection. *Scandinavian Journal of Statistics*. **46**, (1), 289-313.

[49] Rabier, C.E. (2014a). On statistical inference for selective genotyping. *J. Stat. Plan. Infer.*, **147** 24-52.

[50] Rabier, C.E. (2014b). An asymptotic test for Quantitative Trait Locus detection in presence of missing genotypes. *Annales de la Faculté des Sciences de Toulouse Mathématiques*, **6(23)** 755-778.

[51] Rabier, C.E. (2014c). On empirical processes for Quantitative Trait Locus mapping under the presence of a selective genotyping and an interference phenomenon. *J. Stat. Plan. Infer.*, **153** 42-55.

[52] Rabier, C.E. (2015). On stochastic processes for Quantitative Trait Locus mapping under selective genotyping. *Statistics*, **49(1)** 19-34.

[53] 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.

[54] Rebaï, A., Goffinet, B., and Mangin, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, **138** 235-240.

[55] Rebaï, A., Goffinet, B., and Mangin, B. (1995). Comparing power of different methods for QTL detection. *Biometrics*, **51** 87-99.

[56] Siegmund, D. and Yakir, B. (2007). *The statistics of gene mapping*. Springer, New York.

[57] Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., ..., and McCouch, S.R. (2015). Genomic Selection and Association Mapping in rice (Oryza sativa): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genetics*. **11(2)**, e1004982.

[58] Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B., Ingvarsson, P. K. (2017). Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC plant biology*. **17**, (1),110.

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

[60] Upadhyaya, H. D., Bajaj, D., Narnoliya, L., Das, S., Kumar, V., Gowda, C. L. L., ... Parida, S. K. (2016). Genome-wide scans for delineation of candidate genes regulating seed-protein content in chickpea. *Frontiers in Plant Science*, **7**, 302.

[61] Van der Vaart, A.W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics.

[62] Visscher, P.M., Yang, J. & Goddard, M.E. (2010). A commentary on "common SNPs explain a large proportion of the heritability for human height" by Yang et al.(2010). *Twin Research and Human Genetics*. **13**, (06), 517-524.

[63] Vuong, T. D., Walker, D. R., Nguyen, B. T., Nguyen, T. T., Dinh, H. X., Hyten, D. L., ..., and Nguyen, H. T. (2016). Molecular Characterization of Resistance to Soybean Rust (Phakopsora pachyrhizi Syd. Syd.) in Soybean Cultivar DT 2000 (PI 635999). *PloS one*, **11**, (12), e0164493.

[64] Wolc, A., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., Preisinger, R., ... Dekkers, J. C. (2011). Persistence of accuracy of genomic estimated breeding values over generations in layer chickens.

Genetics Selection Evolution, **43**, (1), 23.

[65] Wu, R., Ma, C.X., and Casella, G. (2007). *Statistical Genetics of Quantitative Traits*. Springer, New York.

[66] Yan, L., Hofmann, N., Li, S., Ferreira, M. E., Song, B., Jiang, G., ... Song, Q. (2017). Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. *BMC genomics*, **18**, (1), 529.

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

[68] Zabaneh, D., Krapohl, E., Gaspar, H. A., Curtis, C., Lee, S. H., Patel, H., ... Lubinski, D. (2018). A genome-wide association study for extremely high intelligence. *Molecular psychiatry*, **23**, (5), 1226.

[69] Zhang, F., Guo, X., Zhang, Y., Wen, Y., Wang, W., Wang, S., ... Tan, L. (2014). Genome-wide copy number variation study and gene expression analysis identify ABI3BP as a susceptibility gene for KashinâBeck disease. *Human genetics*, **133**, (6), 793-799.

[70] Zhao, Y., Gowda, M., Longin, F. H., WÃ¼rschum, T., Ranc, N., Reif, J. C. (2012). Impact of selective genotyping in the training population on accuracy and bias of genomic selection. *Theoretical and Applied Genetics*, **125**, (4), 707-713.

[71] Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. **67**, (2), 301-320.

[72] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*. **101**, (476), 1418-1429.

[73] Zou, C., Wang, P., Xu, Y. (2016). Bulked sample analysis in genetics, genomics and crop improvement. *Plant biotechnology journal*. **14**, (10), 301-320.

Charles-Elie Rabier (charles-elie.rabier@umontpellier.fr)
IMAG, Université de Montpellier, CNRS, France.