The Adaptive SgLasso : when selective genotyping and extreme sampling vary along the genome

C.E. Rabier[a,b]

[a]*ISEM, Université de Montpellier, CNRS, France*
[b]*IMAG, Université de Montpellier, CNRS, France*

Abstract We introduce here the AdaptSgLasso, a new method for gene mapping. It relies on the new concept of a selective genotyping that varies along the genome. The selective genotyping, in its original version, 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. By considering a selective genotyping that varies along the genome, we allow geneticists to impose more weights on some loci of interest, known to be responsible for variation of the quantitative trait. The resulting signal is now dedicated to each locus. We show on simulated data the superiority of this new approach over the classical selective genotyping approach. It allows us to propose a new max test for Interval Mapping, and also a new penalized adaptive method to locate multiple genes along the genome.

1. Motivation

Today, more and more genomic data are available thanks to advances in molecular biology and to technology. This makes statistical science very exciting for geneticists, statisticians and mathematicians always eager to propose new methods (e.g. Momen et al. (2018)).

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 propose to introduce here an adaptive variable selection method, relying on an old concept, called selective genotyping, and that meets big data needs. Although genotyping costs have largely dropped recently, selective genotyping or extreme sampling, is still a relevant concept in the modern genomic era. It was first introduced by Lebowitz et al. (1987) who noticed that most of the information about Quantitative Trait Loci, so-called QTL (genes influencing a quantitative trait which is able to be measured) is present in the extreme phenotypes (i.e. extreme traits). Later, Lander and Bostein (1989) formalized this approach and called it selective genotyping. Today, application fields of selective genotyping lie in Genome Wide Association Study (GWAS), and in Genomic Selection (GS). We can also find applications in biotechnology Zou et al. (2016).

The aim of GWAS is to find associations between locations (i.e. loci) of the genome and a trait of interest. We denote some recent association studies using selective genotyping in plants (e.g. sugarcane Gutierrez et al. (2018), soybean Phansak et al. (2016); Yan et al. (2017); Vuong et al. (2016), chickpea Upadhyaya et al. (2016), tomatoes Ohlson et al. (2018)), in animals (e.g. dairy cattle Kurz et al. (2019), drosophila Bastide et al. (2013), sow Cordoba et al. (2015), mouse Fernandes et al. (2016)), and in humans (e.g. on Kashin-Beck disease Zhang et al. (2014) and on intelligence Zabaneh et al. (2018)). Selective genotyping is particularly rewarding for finding QTLs: by considering the extremes, the signal is significantly increased.

The second application field is Genomic Selection (GS) (Hayes et al. (2001)), which is nowadays a very popular topic in genomics (e.g. strawberry Genzan et al. (2017), banana Nyine et al. (2018)). GS consists in selecting individuals on the basis of genomic predictions (see for instance Rabier et al. (2016)). The goal is to predict the future phenotype of young candidates as soon as their DNA has been collected. As a consequence, many generations can be considered in GS without having to wait for observing the phenotype of the candidates at adult age. Some studies have shown that it is essential to update the model over time, otherwise predictions are not reliable anymore (see Goddard et al. (2019)). In this case, when the model is updated, extreme individuals are used to train the model since these individuals were selected at the previous generation because of the superiority of their genomic predictions.

2. Model

We study a backcross population: $A \times (A \times B)$, where A and B are purely homozygous lines and we address the problem of detecting a Quantitative Trait Locus, so-called QTL (a gene influencing a quantitative trait which is able to be measured) on a given chromosome. The trait is observed on n individuals (progenies) and we denote by $Y_j,\ j = 1, ..., n$, these observations.

The chromosome will be represented by the segment $[0, T]$. The distance on $[0, T]$ is called the genetic distance, it is measured in Morgans (see for instance Wu et al. (2007) or Siegmund and Yakir (2007)). 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. The Haldane modeling, which assumes no crossover interference, can be represented as follows: $X(0)$ is a random sign and $X(t) = X(0)(-1)^{N(t)}$ where $N(.)$ is a standard Poisson process on $[0, T]$. Calculations on the Poisson distribution show that

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

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 QTL 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 (e.g. in aquaculture and livestock, see Hayes (2007)). Besides, we will consider $0 < t_1^\star < ... < t_m^\star < T$. We will study the concept of QTL mapping: we will look for associations between allele variations at the QTLs and variation in the quantitative trait of interest.

Usually, in the classical problem of QTL mapping, the "genome information" is available only at fixed locations $t_1 = 0 < t_2 < ... < t_K = T$, called genetic markers. So, usually an observation is

$$(Y,\ X(t_1),\ ...,\ X(t_K))$$

and the challenge is that the number of QTLs m and their locations $t_1^\star$, ..., $t_m^\star$ are unknown. In our present study, we consider the classical problem, but in order to reduce the costs of genotyping, a selective genotyping that varies along the genome has been performed. The originality of this paper lies in the fact that we consider that the selective genotyping varies along the genome. Let us first describe the selective genotyping concept in its original version (Lebowitz et al. (1987); Darvasi and Soller (1992)), the one that does not vary along the genome. We consider two real thresholds S_-^1 and S_+^1, with $S_-^1 \leq S_+^1$ and we collect the genome information at all the marker locations (i.e. we genotype) if and only if the phenotype Y is extreme, that is to say $Y \leq S_-^1$ or $Y \geq S_+^1$.

In order to introduce the selective genotyping that varies along the genome, let us consider four thresholds S_-^1, S_-^2, S_-^2, S_+^1 that belong to $\mathbb{R}$ such as $S_-^1 \leq S_-^2 \leq S_+^2 \leq S_+^1$. As in the classical selective genotyping, we collect the genome information 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 dense map, and we collect the genome information at these marker locations, if and only if $Y \in \left[S_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ S_+^1\right]$. Intuitively, it enables to put more weights on some markers that are well known by geneticists.

In order to describe this concept more precisely, let $\mathbb{T}_K^1$ denote the set $\{t_1, \ldots, t_K\}$ of marker locations, and let $\mathbb{T}_K^2$ be a subset of $\mathbb{T}_K^1$ (i.e. $\mathbb{T}_K^2 \subseteq \mathbb{T}_K^1$). As previously said, $\mathbb{T}_K^2$ can be viewed as a sparser genetic map than $\mathbb{T}_K^1$. In order to fully describe $\mathbb{T}_K^2$, let $\#\mathbb{T}_K^2$ denote the cardinality of $\mathbb{T}_K^2$, and let $\sigma(.)$ denote a one-to-one map $\sigma : \left\{1, \ldots, \#\mathbb{T}_K^2\right\} \to \{1, \ldots, K\}$. Besides, we impose that $\sigma(k) < \sigma(k')$ for $k < k'$. Then, we define $\mathbb{T}_K^2$ the set such as $\mathbb{T}_K^2 = \left\{t_{\sigma(1)}, t_{\sigma(2)}, \ldots, t_{\sigma(\#\mathbb{T}_K^2)}\right\}$. Besides, we will assume $\sigma(1) = 1$ and $\sigma(\#\mathbb{T}_K^2) = K$, so that the markers located at 0 and at T are also located on the sparsest map. If we call $\overline{X}(t)$ and $\widetilde{X}(t)$ the random variables such as

$$\overline{X}(t) = \begin{cases} X(t) & \text{if } \ Y \notin \left[S_-^1\ ,\ S_+^1\right] \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\widetilde{X}(t) = \begin{cases} X(t) & \text{if } \ Y \in \left[S_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ S_+^1\right] \\ 0 & \text{otherwise,} \end{cases}$$

then, in our problem, one observation is

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

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

3. Roadmap

In Section 4, we present Theorem 1 that gives the asymptotic distribution of the score process and 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 $V(.)$ and $W(.)$. The processes $V(.)$ and $W(.)$ are devoted to the dense map and to the sparse map, respectively.

Corollary 1 states results regarding the reverse configuration of selective genotyping: only the non extreme individuals are genotyped.

Section 6 introduces AdaptSgLasso, our new adaptive and penalized likelihood method relying on results of Theorem 1. AdaptSgLasso allows to estimate the QTLs location, their effects and their number. Note that its ElasticNet cousin, AdaptSgEN is also described. The link with the non adaptive versions, SgLasso and SgEN (Rabier and Delmas (2019)) is also established. Last, Section 7 investigates the asymptotic theory for AdaptSgLasso under complete Linkage Disequilibrium.

At the end of the manuscript, Section 8.1 proposes a simulation study regarding 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.

4. Some theoretical results

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

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

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

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 $t \in [t_1, t_K]\backslash\mathbb{T}_K^1$, the likelihood of the couple $\Big(Y,\ \overline{X}(t^{\ell,1}),\ \widetilde{X}(t^{\ell,2}),\ \overline{X}(t^{r,1}),\ \widetilde{X}(t^{r,2})\Big)$ with respect to the measure $\lambda\otimes N\otimes N\otimes N\otimes N$, λ being the Lebesgue measure, N the counting measure on $\mathbb{N}$, 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 $\mathbb{P}(X(t) =\mid X(t^{\ell,1}), X(t^{r,1}))$ and $\mathbb{P}(X(t) =\mid X(t^{\ell,2}), X(t^{r,2}))$,

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

and

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

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

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

with

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

and

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

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

$$S_n(t) = \frac{\frac{\partial l^n_t}{\partial q_1}\ |_{\theta^1_0}}{\sqrt{\mathrm{Var}\left(\frac{\partial l^n_t}{\partial q_1}\ |_{\theta^1_0}\right)}}\ , \tag{3}$$

where $\bar{l}^n_t$ denotes the log likelihood at t, associated to n observations, and $\theta^1_0 = (0,\mu,\sigma)$ refers to the parameter θ_1 under $\mathcal{H}_0$.

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

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

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

$$\gamma_1 := \mathbb{P}_{\mathcal{H}_0}\left(Y \notin \left[S_-^1,\ S_+^1\right]\right)\ ,\ \gamma_1^+ := \mathbb{P}_{\mathcal{H}_0}\left(Y > S_+^1\right)\ ,\ \gamma_1^- := \mathbb{P}_{\mathcal{H}_0}\left(Y < S_-^1\right)\ , \tag{4}$$
$$\gamma := \mathbb{P}_{\mathcal{H}_0}\left(Y \notin \left[S_-^2,\ S_+^2\right]\right)\ ,\ \gamma^+ := \mathbb{P}_{\mathcal{H}_0}\left(Y > S_+^2\right)\ ,\ \gamma^- := \mathbb{P}_{\mathcal{H}_0}\left(Y < S_-^2\right)\ , \tag{5}$$
$$\mathcal{A} := \sigma^2\left\{\gamma_1 + z_{\gamma_1^+}\varphi(z_{\gamma_1^+}) - z_{1-\gamma_1^-}\varphi(z_{1-\gamma_1^-})\right\}\ , \tag{6}$$
$$\mathcal{B} := \sigma^2\left\{\gamma + z_{\gamma^+}\varphi(z_{\gamma^+}) - z_{1-\gamma^-}\varphi(z_{1-\gamma^-})\right\}\ , \tag{7}$$
$$\mathcal{C} := \mathcal{B} - \mathcal{A}, \tag{8}$$

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.

Our main result is the following:

Theorem 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,*

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

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

$$Z(t) = \frac{\sqrt{\mathcal{A}}\ \xi_1(t)V(t)\ +\ \sqrt{\mathcal{C}}\ \xi_2(t)W(t)}{\sqrt{\mathcal{A}\ \xi_1^2(t) + \mathcal{C}\ \xi_2^2(t)}}\ .$$

$V(.)$ and $W(.)$ are independent Gaussian processes with unit variance such as

$$V(t) = \left\{\alpha_1(t)\ V(t^{\ell,1})\ +\ \beta_1(t)\ V(t^{r,1})\right\}/\xi_1(t)$$
$$W(t) = \left\{\alpha_2(t)\ W(t^{\ell,2})\ +\ \beta_2(t)\ W(t^{r,2})\right\}/\xi_2(t)$$
$$\forall (t_k,t_{k'}) \in \mathbb{T}_K^1 \times \mathbb{T}_K^1\ \ Cov(V(t_k),V(t_{k'})) = \rho(t_k,t_{k'})$$
$$\forall (t_k,t_{k'}) \in \mathbb{T}_K^2 \times \mathbb{T}_K^2\ \ Cov(W(t_k),W(t_{k'})) = \rho(t_k,t_{k'})\ .$$

The mean function of $Z(.)$ *is*

$$m_{Z,\vec{t}^\star}(t) = \frac{\sqrt{\mathcal{A}}\ \xi_1(t) m_{V,\vec{t}^\star}(t)\ +\ \sqrt{\mathcal{C}}\ \xi_2(t) m_{W,\vec{t}^\star}(t)}{\sqrt{\mathcal{A}\ \xi_1^2(t) + \mathcal{C}\ \xi_2^2(t)}}$$

with

$$\begin{aligned}
m_{V,\vec{t}^\star}(t) &= \Big\{\alpha_1(t)\ m_{V,\vec{t}^\star}(t^{\ell,1}) + \beta_1(t)\ m_{V,\vec{t}^\star}(t^{r,1})\Big\}\ /\xi_1(t) \\
m_{W,\vec{t}^\star}(t) &= \Big\{\alpha_2(t)\ m_{W,\vec{t}^\star}(t^{\ell,2}) + \beta_2(t)\ m_{W,\vec{t}^\star}(t^{r,2})\Big\}\ /\xi_2(t) \\
\forall t_k \in \mathbb{T}_K^1\ \ m_{V,\vec{t}^\star}(t_k) &= \frac{\sqrt{\mathcal{A}}}{\sigma^2}\sum_{s=1}^{m} \rho(t_s^\star, t_k) a_s \\
\forall t_k \in \mathbb{T}_K^2\ \ m_{W,\vec{t}^\star}(t_k) &= \frac{\sqrt{\mathcal{C}}}{\sigma^2}\sum_{s=1}^{m} \rho(t_s^\star, t_k) a_s\ .
\end{aligned}$$

The proof is given in Section 9.

Note that when $S_-^1 = S_-^2$ and $S_+^1 = S_+^2$, we have $\mathcal{C} = 0$ and the process $Z(.)$ matches the process $V(.)$ of Rabier (2015). In the same way, when $S_-^1 = -\infty$ and $S_+^1 = +\infty$, we have $\mathcal{A} = 0$ and the process $Z(.)$ matches the process $W(.)$.

By continuity, it is easy to see that when t_k belongs to $\mathbb{T}_K^2$:

$$\begin{aligned}
Z(t_k) &= \frac{\sqrt{\mathcal{A}}\ V(t_k)\ +\ \sqrt{\mathcal{C}}\ W(t_k)}{\sqrt{\mathcal{A}\ + \mathcal{C}}}\ , \qquad (10) \\
m_{Z,\vec{t}^\star}(t_k) &= \frac{\sqrt{\mathcal{B}}}{\sigma^2}\sum_{s=1}^{m} \rho(t_s^\star, t_k) a_s\ .
\end{aligned}$$

However, at a location t_k that belongs to $\mathbb{T}_K^1 \backslash \mathbb{T}_K^2$, the signal is weaker:

$$\begin{aligned}
Z(t_k) &= \frac{\sqrt{\mathcal{A}}\ V(t_k)\ +\ \sqrt{\mathcal{C}}\ \Big\{\alpha_2(t_k) W(t_k^{\ell,2}) + \beta_2(t_k) W(t_k^{r,2})\Big\}}{\sqrt{\mathcal{A}\ + \mathcal{C}\ \xi_2^2(t_k)}}\ , \qquad (11) \\
m_{Z,\vec{t}^\star}(t_k) &= \frac{\frac{\mathcal{A}}{\sigma^2}\sum_{s=1}^{m} \rho(t_s^\star, t_k) a_s\ +\ \frac{\mathcal{C}}{\sigma^2}\Big\{\alpha_2(t_k)\sum_{s=1}^{m} \rho(t_s^\star, t_k^{\ell,2}) a_s + \beta_2(t_k)\sum_{s=1}^{m} \rho(t_s^\star, t_k^{r,2}) a_s\Big\}}{\sqrt{\mathcal{A} + \mathcal{C}\ \xi_2^2(t_k)}}
\end{aligned}$$

where $t_k^{\ell,2}$ and $t_k^{r,2}$ are defined according to formula (2), using a small abuse of notation. The decrease regarding the signal is due to the fact that (by continuity) $\xi_2^2(t_k) \neq 1$, $\alpha_2(t_k) \neq 1$, $\beta_2(t_k) \neq 0$.

Using formulae (11) and (10), we can easily compute the squeleton of the covariance function of $Z(.)$:

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

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

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

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

The proof is given in Section 10.

Let us now 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 all the genotypes are known. In other words, under the complete data situation, we have $S_-^1 = S_-^2 = S_+^2 = S_+^1$, so that $\mathcal{C} = 0$ and $\mathcal{A} = \mathcal{B} = \sigma^2$. Note also that the complete data situation is the one studied in Azaïs et al. (2012).

Theorem 2. *Let κ denote the ARE, then we have*

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

where

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

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

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

iii) if $t_k \in \mathbb{T}_K^2$, $\quad \kappa = \mathcal{B}/\sigma^2$.

The proof is given in Section 11. Note that ii) and iii) can be obtained from i) by continuity.

According to Theorem 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 $\mathcal{A}$ (i.e. $\mathcal{B}$ since $\mathcal{B} = \mathcal{A}$) linked to the selection intensity (see Theorem 4.2 of Rabier (2015)).

The situation iii), i.e. $t_k \in \mathbb{T}_K^2$, can be viewed as a “classical” selective genotyping situation at one marker, 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 match exactly the ARE presented in Theorem 1 of Rabier (2014a).

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

Figures 1 and 2 illustrate the efficiency κ, given in expression i) of Theorem 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 maps 1 and 2, 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 $t_1^\star = 0.85$, and the test is performed exactly at the QTL location ($t = t_1^\star$). The constant a linked to the QTL effect is set to the value 2. The main genetic map is such as $t^{\ell,1} = 0.80$ and $t^{r,1} = 0.90$, and two scenarios are investigated for the extra map that targets a few loci: either a) $t^{\ell,2} = 0.20$ and $t^{r,2} = 1.50$, or 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 map a) as compared to map b), when γ_1 increases. It was expected since on map a), markers and the QTL are far apart. When γ_1 increases, more and more individuals are genotyped at markers of the main map, and since these markers are closer to the QTL location, it helps for for the statistical test. In contrast, on map b), markers are already close to the QTL location and the main 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 map b) than on map a). This result was expected in view of the previous experiment.

Remark : According to the figures, the efficiencies reached their maximum for $\gamma_1^+/\gamma_1 = 1/2$ and $\gamma^+/\gamma = 1/2$. In Appendix, we show 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 Section 8.1).

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 map 1, and every 25cM on map 2 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	n	1(+)	2(++)	2(+-)	3(+-+)
1/2	Sgeno Varies	$+\infty$	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	Sgeno Classical	$+\infty$	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	Sgeno Varies	$+\infty$	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	Sgeno Classical	$+\infty$	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	Sgeno Varies	$+\infty$	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	Sgeno Classical	$+\infty$	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%

5. Extra results

5.1. The reverse configuration

As in Rabier (2014b), let us consider the reverse configuration: it consists in genotyping only individuals for which $Y \in [S_-^2, S_+^2]$ or $Y \in \left[S_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ S_+^1\right]$. More precisely, $\overline{X}(t)$ is now defined in the following way:

$$\overline{X}(t) = \begin{cases} X(t) & \text{if } Y \in \left[S_-^2\ ,\ S_+^2\right] \\ 0 & \text{otherwise.} \end{cases}$$

In constrast, we still have

$$\widetilde{X}(t) = \begin{cases} X(t) & \text{if } Y \in \left[S_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ S_+^1\right] \\ 0 & \text{otherwise} \end{cases}$$

and one observation is still

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

In this context, let us define the quantity $\mathcal{F}$ in the following way:

$$\mathcal{F} := \sigma^2 \left\{1 - \gamma - z_{\gamma^+}\varphi(z_{\gamma^+}) + z_{1-\gamma^-}\varphi(z_{1-\gamma^-})\right\}\ .$$

Corollary 1. *Under the reverse configuration, that is to say if* $\overline{X}(t_k) = X(t_k)\ 1_{Y\in\left[S_-^2\ ,\ S_+^2\right]}$. *and* $\widetilde{X}(t) = X(t)1_{Y\in\left[S_-^1\ ,\ S_-^2\right]\cup\left[S_+^2\ ,\ S_+^1\right]}$, *then we have the same result as in Theorem 1 provided that we replace the quantity* $\mathcal{A}$ *by the quantity* $\mathcal{F}$.

The proof is easily deduced from the proof of Theorem 1 and from Rabier (2014b).

6. An adaptive gene mapping method

In this section, the goal is to propose an adaptive method to estimate the number of QTLs, their effects and their positions combining results of Theorem 1 and a penalized likelihood method. Since this method is an adaptive version of the SgLasso (Rabier and Delmas (2019)), we will call it the AdaptSgLasso. Contrary to its ancestor SgLasso, AdaptSgLasso allows to put some weights on some loci along the genome. We will also introduce AdaptSgEN which is the Elastic Net version of our adaptive method (see formulas (18) below)

Notations 1. $\mathcal{G}_k$ *denotes either* $\sqrt{\mathcal{A}}/\sigma$ *or* $\sqrt{\mathcal{B}}/\sigma$ *depending if* t_k *belongs to* $\mathbb{T}_K^1\backslash\mathbb{T}_K^2$ *or* $\mathbb{T}_K^2$, *respectively.*

According to Theorem 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 $\vec{S}_n = (S_n(t_1)\ ,\ S_n(t_2)\ ,\ ...\ ,\ S_n(t_K))'$, $\vec{m}_{\vec{t}^\star} = (m_{\vec{t}^\star}(t_1)\ ,\ m_{\vec{t}^\star}(t_2)\ ,\ ..., m_{\vec{t}^\star}(t_K))'$ 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.

Using the the Cholesky decomposition $\Sigma = AA'$, we have

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

where

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

and B is a matrix of size $K \times K$ such as

$$\forall (t_k, t_{k'}) \in \mathbb{T}_K^2 \times \mathbb{T}_K^2 \ \ B_{k,k'} = \rho(t_k, t_{k'})$$

$$\forall (t_k, t_{k'}) \in \mathbb{T}_K^2 \times \mathbb{T}_K^1 \backslash \mathbb{T}_K^2 \ \ B_{k,k'} = \frac{\rho(t_k, t_{k'})\sqrt{\mathcal{B}}}{\sqrt{\mathcal{A}}}$$

$$\forall (t_k, t_{k'}) \in \mathbb{T}_K^1 \backslash \mathbb{T}_K^2 \times \mathbb{T}_K^2 \ \ B_{k,k'} = \rho(t_k, t_{k'})\sqrt{\mathcal{B}} / \sqrt{\mathcal{A} + \mathcal{C}\xi_2^2(t_k)}$$

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

$$B_{k,k'} = \left(\frac{\mathcal{A}\rho(t_k, t_{k'})}{\sqrt{\mathcal{A}}} + \frac{\mathcal{C}}{\sqrt{\mathcal{A}}} \left\{ \alpha_2(t_k)\rho(t_k^{\ell,2}, t_{k'}) + \beta_2(t_k)\rho(t_k^{r,2}, t_{k'}) \right\} \right) / \sqrt{\mathcal{A} + \mathcal{C}\xi_2^2(t_k)}\ .$$

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:

$$\hat{\Delta}_{\text{AdaptSg}}(\lambda, \alpha) = \arg\min_{\Delta} \left(\left\| A^{-1}\vec{S}_n - A^{-1}B\Delta \right\|_2^2 + \lambda \ \mathrm{pen}(\alpha) \right) \tag{16}$$

where:

$$\mathrm{pen}(\alpha) = \frac{1-\alpha}{2} \left\|\Delta\right\|_2^2 + \alpha \left\|\Delta\right\|_1 \tag{17}$$

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 AdaptSgLasso and the AdaptSgEN in the following way:

$$\begin{aligned}\hat{\Delta}_{\text{AdaptSgLasso}}(\lambda) &= \hat{\Delta}_{\text{AdaptSg}}(\lambda, 1)\\ \hat{\Delta}_{\text{AdaptSgEN}}(\lambda, \alpha) &= \hat{\Delta}_{\text{AdaptSg}}(\lambda, \alpha).\end{aligned} \tag{18}$$

Note that for $S_-^1 = S_-^2$ and $S_+^1 = S_+^2$ (classical selective genotyping), since $\mathcal{C} = 0$ and $\mathcal{B} = \mathcal{A}$, each entry of the matrices B and Σ is equal to $\rho(t_k, t_{k'})$. As expected, in this case, formula (16) is identical to formula (14) of Rabier and Delmas (2019), and the AdaptSgLasso (resp. AdaptSgEN) matches the SgLasso (resp. SgEN) under complete Linkage Disequilibrium.

7. Asymptotic theory for AdaptSgLasso under complete Linkage Disequilibrium

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 the relationships $B = \Sigma\ \Upsilon$, where

$$\forall t_k \in \mathbb{T}_K^2 \ \ \Upsilon_{k,k} = 1\ ,\ \forall t_k \in \mathbb{T}_K^1 \backslash \mathbb{T}_K^2 \ \ \Upsilon_{k,k} = \frac{\sqrt{\mathcal{A} + \mathcal{C}\xi_2^2(t_k)}}{\sqrt{\mathcal{A}}}\ ,\ \forall t_k \neq t_{k'} \ \ \Upsilon_{k,k'} = 0\ .$$

As a result, $A^{-1}B = A'\Upsilon$ and we have:

$$\hat{\Delta}_{\text{AdaptSgLasso}}(\lambda, \alpha) = \arg\min_{\Delta} \left(\left\| A^{-1}\vec{S}_n - A'\Upsilon\Delta \right\|_2^2 + \lambda \left\|\Delta\right\|_1 \right)\ . \tag{19}$$

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. Bühlmann and Van de Geer (2011) page 108). Since $\hat{\sigma}_k^2 := \frac{1}{K}(\Upsilon' A A' \Upsilon)_{kk} = \frac{\Upsilon_{kk}^2}{K}$, let us set $(A'\Upsilon)_{\text{scal}} := \sqrt{K} A'\Upsilon\Xi$ where $\Xi = \begin{pmatrix} \frac{1}{\Upsilon_{1,1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\Upsilon_{2,2}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{\Upsilon_{K,K}} \end{pmatrix}$. Then, let us define

$$\begin{aligned}\hat{\Delta}_{\text{AdaptSgLasso}_{\text{scal}}}(\lambda) := \arg\min_{\Delta} \Bigg(& \frac{\left\| A^{-1}\vec{S}_n - (A'\Upsilon)_{\text{scal}}\, \Xi^{-1}\Delta/\sqrt{K} \right\|_2^2}{K} \\ & + \lambda \frac{\Upsilon_{11}}{\sqrt{K}} \left|\Delta_1\right| + \lambda \frac{\Upsilon_{22}}{\sqrt{K}} \left|\Delta_2\right| + \ldots + \lambda \frac{\Upsilon_{KK}}{\sqrt{K}} \left|\Delta_K\right| \Bigg)\ .\end{aligned}$$

As soon as we set $\tilde{\Delta} := \Xi^{-1}\Delta/\sqrt{K}$, this problem can be rewritten in the following way:

$$\hat{\hat{\Delta}}_{\text{AdaptSgLasso}_{\text{scal}}}(\lambda) := \arg\min_{\tilde{\Delta}} \left(\frac{\left\| A^{-1}\vec{S}_n - (A'\Upsilon)_{\text{scal}}\,\tilde{\Delta} \right\|_2^2}{K} + \lambda \left\| \tilde{\Delta} \right\|_1 \right) . \tag{20}$$

We can apply Corollary 6.1 of Bühlmann and Van de Geer (2011) with $\hat{\sigma} = 1$ (cf. our linear model in formula (15)), that establishes the slow rate of convergence

$$\frac{\left\| (A'\Upsilon)_{\text{scal}}\,(\hat{\hat{\Delta}}_{\text{AdaptSgLasso}_{\text{scal}}} - \Delta') \right\|_2^2}{K} = O_P \left(\frac{\sqrt{\log(K)}}{K} \left\{ \sum_{s|t_s^\star \in \mathbb{T}_K^2} \frac{|a_s|\,\sqrt{\mathcal{B}}}{\sigma^2} + \sum_{s|t_s^\star \in \mathbb{T}_K^1 \setminus \mathbb{T}_K^2} \frac{|a_s|\,\sqrt{\mathcal{A} + \mathcal{C}\xi_2^2(t_s^\star)}}{\sigma^2} \right\} \right) \tag{21}$$

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 Bühlmann and Van de Geer (2011) applies and we obtain the fast rate of convergence.

Let us state the classical Lasso conditions in the "AdaptSgLasso" context.

The β-min condition:

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

where m is the number of QTLs (factor linked to the sparsity), and $\mathit{\Phi}^2$ is a restricted eigen value of the design matrix $(A'\Upsilon)_{\text{scal}}$.

The irrepresentable condition:

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

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

The β-min condition and the irrepresentable condition, ensure consistent variable selection for AdaptiveSgLasso under selective genotyping.

Note that under a classical selective genotyping, the β-min condition would have been

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

where $\tilde{\mathit{\Phi}}^2$ is a restricted eigen value of the design matrix (see Rabier and Delmas (2019)).

8. Simulation study

8.1. About 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 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 was considered, γ 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 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 symetrically, the asymptotic power associated to our approach was found equal to 58.55% for $m = 1$ and 46.69% for $m = 3$, whereas the one associated to the classical approach was only estimated to 48.09% for $m = 1$ and 40.83% for $m = 3$.

Surprisingly, the classical approach was the best method when the selective genotyping was unidirectional ($\gamma_+/\gamma = 1$). However, overall, it is clear in view of our simulation study that we should adopt our new method and perform a symetrical selective genotyping.

8.2. Association study

Note that in what follows, we will consider as a false discovery, a selected regressor which is not located in a neighbourhood of 5cM of a true QTL locations (e.g. Broman and Speed (2002)). In that sense, the definition of a false discovery differs slightly from the one of a false positive. The FDR will be the percentage of such false discoveries among all the discoveries.

8.3. Genomic selection

9. Proof of Theorem 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 $H_{at^\star}$
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

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

because of the key Lemma (Lemma 2.6 of Rabier (2015)), 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

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

and

$$\mathbb{E}\left[\left(\frac{\partial l_t}{\partial q_1}\ |_{\theta_0^1}\right)^2\right] = \frac{\mathcal{A}}{\sigma^4}\xi_1^2(t) + \frac{\mathcal{C}}{\sigma^4}\xi_2^2(t)\ .$$

Indeed, by definition, according to Rabier (2014a), we have $\mathcal{A} = \mathbb{E}_{\mathcal{H}_0}\left[(Y-\mu)^2 1_{Y\notin[S_-^1,S_+^1]}\right]$. In the same way, $\mathcal{C} = \mathbb{E}_{\mathcal{H}_0}\left[(Y-\mu)^2 1_{Y\in[S_-^1,S_-^2]\cup[S_+^2,S_+^1]}\right]$.

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

$$I_{\theta_0} = Diag\left[\frac{\mathcal{A}}{\sigma^4}\xi_1^2(t) + \frac{\mathcal{C}}{\sigma^4}\xi_2^2(t),\ \frac{1}{\sigma^2}\ ,\ \frac{2}{\sigma^2}\right]\ . \tag{22}$$

9.1. Study under $\mathcal{H}_0$

In what follows, we define the processes $V_n(.)$ and $W_n(.)$ in the following way:

$$\forall t_k \in \mathbb{T}_K^1\ \ V_n(t_k) := \frac{1}{\sqrt{n\mathcal{A}}}\sum_{j=1}^n (Y_j-\mu)\ \overline{X}_j(t_k)\ ,\ \forall t_k \in \mathbb{T}_K^2\ \ W_n(t_k) := \frac{1}{\sqrt{n\mathcal{C}}}\sum_{j=1}^n (Y_j-\mu)\ \widetilde{X}_j(t_k)\ ,$$
$$V_n(t) := \left\{\alpha_1(t)V_n(t^{\ell,1}) + \beta_1(t)V_n(t^{r,1})\right\}/\xi_1(t)\ ,$$
$$W_n(t) := \left\{\alpha_2(t)W_n(t^{\ell,2}) + \beta_2(t)W_n(t^{r,2})\right\}/\xi_2(t)\ .$$

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

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial l_t^n}{\partial q_1}\,|_{\theta_0^1} &= \frac{\alpha_1(t)}{\sigma\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j\,\overline{X}_j(t^{\ell,1}) \;+\; \frac{\beta_1(t)}{\sigma\sqrt{n}}\;\sum_{j=1}^{n}\varepsilon_j\,\overline{X}_j(t^{r,1}) + \frac{\alpha_2(t)}{\sigma\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j\,\widetilde{X}_j(t^{\ell,2}) \;+\; \frac{\beta_2(t)}{\sigma\sqrt{n}}\;\sum_{j=1}^{n}\varepsilon_j\,\widetilde{X}(t^{r,2})\\ &= \frac{\alpha_1(t)\sqrt{\mathcal{A}}}{\sigma^2\sqrt{n\mathcal{A}}}\sum_{j=1}^{n}\sigma\varepsilon_j\,\overline{X}(t^{\ell,1}) \;+\; \frac{\beta_1(t)\sqrt{\mathcal{A}}}{\sigma^2\sqrt{n\mathcal{A}}}\;\sum_{j=1}^{n}\sigma\varepsilon_j\,\overline{X}(t^{r,1}) + \frac{\alpha_2(t)\sqrt{\mathcal{C}}}{\sigma\sqrt{n\mathcal{C}}}\sum_{j=1}^{n}\sigma\varepsilon_j\,\widetilde{X}(t^{\ell,2})\\ &+\; \frac{\beta_2(t)\sqrt{\mathcal{C}}}{\sigma\sqrt{n\mathcal{C}}}\;\sum_{j=1}^{n}\sigma\varepsilon_j\,\widetilde{X}(t^{r,2})\\ &= \frac{\alpha_1(t)\sqrt{\mathcal{A}}}{\sigma^2}\,V_n(t^{\ell,1}) \;+\; \frac{\beta_1(t)\sqrt{\mathcal{A}}}{\sigma^2}\,V_n(t^{r,1}) \;+\; \frac{\alpha_2(t)\sqrt{\mathcal{C}}}{\sigma^2}\,W_n(t^{\ell,2}) \;+\; \frac{\beta_2(t)\sqrt{\mathcal{C}}}{\sigma^2}\,W_n(t^{r,2})\;. \end{aligned} \tag{23}$$

According to formulae (3), (22) and (23), we obtain easily that

$$S_n(t) = \frac{\sqrt{\mathcal{A}}\;\xi_1(t)V_n(t) \;+\; \sqrt{\mathcal{C}}\;\xi_2(t)W_n(t)}{\sqrt{\mathcal{A}\;\xi_1^2(t) + \mathcal{C}\;\xi_2^2(t)}}\;,$$

According to the proof of Theorem 2.5 of Rabier (2015), we have:

$$\forall t_k \in \mathbb{T}_K^1\;\; V_n(t_k) \longrightarrow \mathcal{N}(0,1) \quad .$$

In the same way, we obtain easily that:

$$\forall t_k \in \mathbb{T}_K^2\;\; W_n(t_k) \longrightarrow \mathcal{N}(0,1)\;.$$

Furthermore, according to the proof of Theorem 2.5 of Rabier (2015), we have:

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

In the same way, we obtain easily that:

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

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

$$V_n(t) \longrightarrow \mathcal{N}(0,1) \;\text{ and }\; W_n(t) \longrightarrow \mathcal{N}(0,1)\;.$$

As a consequence, according to the continuous mapping theorem

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

which proves the convergence of 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 Azaïs and

Wschebor (2009)). 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 $t_2 \notin \mathbb{T}_K^2$ and $t_3 \notin \mathbb{T}_K^2$. Besides, let us impose that $\{t_1, t_4\} \subset \mathbb{T}_K^2$. 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 $[t_2, t_3]$:

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

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 $\mathbb{P}\left(w_{S_n}(\delta) \geq \eta\right) \leq \varepsilon \quad \forall n \geq n_0$.

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

$$\begin{aligned} \forall i = 1,2 \quad \tilde{\alpha_i}(t) &= \alpha_i(t) / \sqrt{\mathcal{A}\xi_1^2(t) + \mathcal{C}\xi_2^2(t)}, \\ \tilde{\beta_i}(t) &= \beta_i(t) / \sqrt{\mathcal{A}\xi_1^2(t) + \mathcal{C}\xi_2^2(t)}. \end{aligned}$$

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

$$\begin{aligned} w_{S_n}(\delta) &= \sup_{|t'-t|<\delta} |S_n(t') - S_n(t)| \\ &= \sup_{|t'-t|<\delta} \Big| \sqrt{\mathcal{A}} \left\{ \tilde{\alpha_1}(t') V_n(t'^{\ell,1}) + \tilde{\beta_1}(t') V_n(t'^{r,1}) \right\} + \sqrt{\mathcal{C}} \left\{ \tilde{\alpha_2}(t') W_n(t'^{\ell,2}) + \tilde{\beta_2}(t') W_n(t'^{r,2}) \right\} \\ &\quad - \sqrt{\mathcal{A}} \left\{ \tilde{\alpha_1}(t) V_n(t^{\ell,1}) + \tilde{\beta_1}(t) V_n(t^{r,1}) \right\} - \sqrt{\mathcal{C}} \left\{ \tilde{\alpha_2}(t) W_n(t^{\ell,2}) + \tilde{\beta_2}(t) W_n(t^{r,2}) \right\} \Big|. \end{aligned} \tag{24}$$

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

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

Since the events are independent,

$$\begin{aligned}
&\mathbb{P} \Big(\max \Big\{ 2\sqrt{\mathcal{A}} \Big\{ w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta}_1}(\delta) \Big\} \max \left(\left| V_n(t'^{\ell,1}) \right|, \left| V_n(t'^{r,1}) \right| \right), \\
&2\sqrt{\mathcal{C}} \Big\{ w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta}_2}(\delta) \Big\} \max \left(\left| W_n(t'^{\ell,2}) \right|, \left| W_n(t'^{r,2}) \right| \right) \Big\} \geq \eta \Big) \\
&= 1 - \mathbb{P} \Big(2\sqrt{\mathcal{A}} \Big\{ w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta}_1}(\delta) \Big\} \max \left(\left| V_n(t'^{\ell,1}) \right|, \left| V_n(t'^{r,1}) \right| \right) \leq \eta \Big) \\
&\times \mathbb{P} \Big(2\sqrt{\mathcal{C}} \Big\{ w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta}_2}(\delta) \Big\} \max \left(\left| W_n(t'^{\ell,2}) \right|, \left| W_n(t'^{r,2}) \right| \right) \leq \eta \Big)
\end{aligned}$$

Let us consider $0 < \varepsilon_1 < 1$ and $\eta > 0$. Since the sequence $\max \left(\left| V_n(t'^{\ell,1}) \right|, \left| V_n(t'^{r,1}) \right| \right)$ is uniformly tight,

$$\exists M_1 > 0 \;\; \forall n \geq 1 \;\; \mathbb{P} \left(\max \left(\left| V_n(t'^{\ell,1}) \right|, \left| V_n(t'^{r,1}) \right| \right) \geq M_1 \right) \leq \varepsilon_1. \tag{25}$$

In other words,

$$\exists M_1 > 0 \;\; \forall n \geq 1 \;\; \mathbb{P} \left(\max \left(\left| V_n(t'^{\ell,1}) \right|, \left| V_n(t'^{r,1}) \right| \right) \leq M_1 \right) \geq 1 - \varepsilon_1. \tag{26}$$

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

$$\exists M_2 > 0 \;\; \forall n \geq 1 \;\; \mathbb{P} \left(\max \left(\left| W_n(t'^{\ell,2}) \right|, \left| W_n(t'^{r,2}) \right| \right) \geq M_2 \right) \leq \varepsilon_1. \tag{27}$$

In other words,

$$\exists M_2 > 0 \;\; \forall n \geq 1 \;\; \mathbb{P} \left(\max \left(\left| W_n(t'^{\ell,2}) \right|, \left| W_n(t'^{r,2}) \right| \right) \leq M_2 \right) \geq 1 - \varepsilon_1. \tag{28}$$

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,

$$\exists \delta \;\text{ such as }\; 0 < \delta < t_3 - t_2,\; w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta_1}}(\delta) < \frac{\eta}{2M_1\sqrt{\mathcal{A}}} \tag{29}$$

$$w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta_2}}(\delta) < \frac{\eta}{2M_2\sqrt{\mathcal{C}}}. \tag{30}$$

As a consequence, we have:

$$\mathbb{P}\left(2\sqrt{\mathcal{A}}\left\{w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta_1}}(\delta)\right\} \max\left(\left|V_n(t'^{\ell,1})\right|, \left|V_n(t'^{r,1})\right|\right) \leq \eta\right) \geq 1 - \varepsilon_1.$$

$$\mathbb{P}\left(2\sqrt{\mathcal{C}}\left\{w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta_2}}(\delta)\right\} \max\left(\left|W_n(t'^{\ell,2})\right|, \left|W_n(t'^{r,2})\right|\right) \leq \eta\right) \geq 1 - \varepsilon_1.$$

Then,

$$\begin{aligned}&\mathbb{P}\left(2\sqrt{\mathcal{A}}\left\{w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta_1}}(\delta)\right\} \max\left(\left|V_n(t'^{\ell,1})\right|, \left|V_n(t'^{r,1})\right|\right) \leq \eta\right) \\ &\times \mathbb{P}\left(2\sqrt{\mathcal{C}}\left\{w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta_2}}(\delta)\right\} \max\left(\left|W_n(t'^{\ell,2})\right|, \left|W_n(t'^{r,2})\right|\right) \leq \eta\right) \geq (1 - \varepsilon_1)^2.\end{aligned}$$

As a result,

$$\begin{aligned}&1 - \mathbb{P}\left(2\sqrt{\mathcal{A}}\left\{w_{\tilde{\alpha_1}}(\delta) + w_{\tilde{\beta_1}}(\delta)\right\} \max\left(\left|V_n(t'^{\ell,1})\right|, \left|V_n(t'^{r,1})\right|\right) \leq \eta\right) \\ &\times \mathbb{P}\left(2\sqrt{\mathcal{C}}\left\{w_{\tilde{\alpha_2}}(\delta) + w_{\tilde{\beta_2}}(\delta)\right\} \max\left(\left|W_n(t'^{\ell,2})\right|, \left|W_n(t'^{r,2})\right|\right) \leq \eta\right) \leq 1 - (1 - \varepsilon_1)^2.\end{aligned}$$

Last,

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

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.

9.2. 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}\ |_{\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)}{\sigma^2\sqrt{n}}\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\overline{X}_j(t^{\ell,1})\ +\ \frac{\beta_1(t)}{\sigma^2\sqrt{n}}\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\overline{X}_j(t^{r,1}) \\
& + \frac{\alpha_2(t)}{\sigma^2\sqrt{n}}\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\ \widetilde{X}_j(t^{\ell,2}) + \ \frac{\beta_2(t)}{\sigma^2\sqrt{n}}\ \sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\ \widetilde{X}(t^{r,2}) \\
& + \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}}}{\sigma^2\sqrt{n\mathcal{A}}}\ \left\{\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\overline{X}_j(t^{\ell,1})\ +\ \sum_{j=1}^{n}\sigma\varepsilon_j\ \overline{X}_j(t^{\ell,1})\right\} \\
& + \frac{\beta_1(t)\sqrt{\mathcal{A}}}{\sigma^2\sqrt{n\mathcal{A}}}\ \left\{\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\overline{X}_j(t^{r,1})\ +\ \sum_{j=1}^{n}\sigma\varepsilon_j\ \overline{X}_j(t^{r,1})\right\} \\
& + \frac{\alpha_2(t)\sqrt{\mathcal{C}}}{\sigma^2\sqrt{n\mathcal{C}}}\ \left\{\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\widetilde{X}_j(t^{\ell,1})\ +\ \sum_{j=1}^{n}\sigma\varepsilon_j\ \widetilde{X}_j(t^{\ell,1})\right\} \\
& + \frac{\beta_2(t)\sqrt{\mathcal{C}}}{\sigma^2\sqrt{n\mathcal{C}}}\ \left\{\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\widetilde{X}_j(t^{r,1})\ +\ \sum_{j=1}^{n}\sigma\varepsilon_j\ \widetilde{X}_j(t^{r,1})\right\}\ .
\end{aligned}
$$

In other words, under $\mathcal{H}_{a\vec{t^\star}}$, we have the relationship:

$$
\frac{1}{\sqrt{n}}\frac{\partial l_t^n}{\partial q_1}\ |_{\theta_0^1} = \frac{\alpha_1(t)\sqrt{\mathcal{A}}}{\sigma^2}\ V_n(t^{\ell,1})\ +\ \frac{\beta_1(t)\sqrt{\mathcal{A}}}{\sigma^2}\ V_n(t^{r,1})\ +\ \frac{\alpha_2(t)\sqrt{\mathcal{C}}}{\sigma^2}\ W_n(t^{\ell,2})\ +\ \frac{\beta_2(t)\sqrt{\mathcal{C}}}{\sigma^2}\ W_n(t^{r,2})
$$

where

$$
\begin{aligned}
\forall t_k \in \mathbb{T}_K^1\ V_n(t_k) &= \frac{1}{\sqrt{n\mathcal{A}}}\ \left\{\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\overline{X}_j(t_k)\ +\ \sum_{j=1}^{n}\sigma\varepsilon_j\ \overline{X}_j(t_k)\right\}\ , \\
\forall t_k \in \mathbb{T}_K^2\ W_n(t_k) &= \frac{1}{\sqrt{n\mathcal{C}}}\ \left\{\sum_{j=1}^{n}\sum_{s=1}^{m} q_s X_j(t_s^\star)\widetilde{X}_j(t_k)\ +\ \sum_{j=1}^{n}\sigma\varepsilon_j\ \widetilde{X}_j(t_k)\right\}\ .
\end{aligned}
$$

By definition, we have the relationship $\mathcal{B} = \mathbb{E}_{\mathcal{H}_0}\left[(Y-\mu)^2 1_{Y\notin[S_-^2,S_+^2]}\right]$.

According to formula (2.9) of Supplement A of Rabier and Delmas (2019),

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

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

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

Besides, according to formula (2.10) of Supplement A of Rabier and Delmas (2019),

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

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

We have, using a technical proof present in Section 4 of Supplement A of Rabier and Delmas (2019),

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

As a consequence, we have

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

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

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

Finally, we obtain

$$\forall t_k \in \mathbb{T}_K^1 \; V_n(t_k) \longrightarrow \mathcal{N}\left(\frac{\sqrt{\mathcal{A}}}{\sigma^2} \sum_{s=1}^{m} \rho(t_s^\star, t_k) a_s, 1 \right) \quad \text{and} \quad \forall t_k \in \mathbb{T}_K^2 \; W_n(t_k) \longrightarrow \mathcal{N}\left(\frac{\sqrt{\mathcal{C}}}{\sigma^2} \sum_{s=1}^{m} \rho(t_s^\star, t_k) a_s, 1 \right) \, .$$

As a consequence, using the interpolations :

$$S_n(t) \longrightarrow \mathcal{N}(\Omega, 1) \tag{31}$$

where

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

Study of the LRT process

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

$$\Lambda_n(t) = S_n^2(t) + o_P(1) \tag{32}$$

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 $t^\star$, $\theta^\star$ be the true parameter that will be assumed to belong to H_0. Note that $t^\star$ makes no sense for θ 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 Azaïs et al. (2006) hold. So, we can apply Theorem 1 of Azaïs et al. (2006) and we have

$$\sup_{(t,\theta)} l_t(\theta) - l_{t^\star}(\theta^\star) = \sup_{d \in \mathcal{D}} \left(\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} d(X_j) \right)^2 1_{\sum_{j=1}^{n} d(X_j) \geq 0} \right) + o_P(1) \tag{33}$$

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 Azaïs et al. (2006), see also Gassiat (2002) and Azaïs et al. (2009). 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 $p_{t^\star,\theta^\star}$ and that are differentiable in quadratic mean.

It is easy to see that

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

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

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

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 (33). Then, since the Fisher information matrix is diagonal (see formula (22)) , it is easy to see that

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

This is exactly the desired result.

In other words, we have proved that under H_0:

$$\sup \Lambda_n(.) = \sup S_n^2(.) + o_P(1)\ . \tag{34}$$

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 $\mathbb{T}_K^1$). 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

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

In the same way,

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

Besides,

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

On the other hand,

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

As a result,

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

In the same way, we have:

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

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

$$\begin{aligned}
&\mathbb{P}\Big(Y \in [y\ ,\ y+dy]\,, \overline{X}(t_1^{\star\ell,1}) = 0, \overline{X}(t_1^{\star r,1}) = 0, \ldots, \overline{X}(t_m^{\star\ell,1}) = 0, \overline{X}(t_m^{\star r,1}) = 0,\\
&\widetilde{X}(t_1^{\star\ell,2}) = 0, \widetilde{X}(t_1^{\star r,2}) = 0, \ldots, \widetilde{X}(t_m^{\star\ell,2}) = 0, \widetilde{X}(t_m^{\star r,2}) = 0\Big) \qquad (35)\\
&= \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \mathbb{P}(Y \in [y\ ,\ y+dy]\,, Y \in [S^2_-, S^2_+], X(t_1^\star) = u_1, X(t_2^\star) = u_2, \ldots, X(t_m^\star) = u_m)\\
&= \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(y)\ 1_{y\in[S^2_-,S^2_+]}\ \mathbb{P}(X(t_1^\star) = u_1, X(t_2^\star) = u_2, \ldots, X(t_m^\star) = u_m)\ .
\end{aligned}$$

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

$$\begin{aligned} L^m_{\vec{t}^\star}(\theta^m) = \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} & \Big[w^1_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)\ 1_{Y\notin[S^1_-,S^1_+]} \\ & + w^2_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)\ 1_{Y\in[S^1_-,S^2_-]\cup[S^2_+,S^1_+]} \\ & + v_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)\ 1_{Y\in[S^2_-,S^2_+]}\Big]\ g^m(t_1^\star, \ldots, t_m^\star) \end{aligned} \tag{36}$$

with

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

and

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

Let us define the parameter θ^m_0 in the following way : $\theta^m_0 = (0, ..., 0, \mu, \sigma)$.

The likelihood $L^{m,n}_{\vec{t}^\star}(\theta^m)$ for n observations is obtained by the product of n terms as in formula (36) 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 $L^{m,n}_{\vec{t}^\star}(\theta^m_0)$).

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

$$\log \frac{dQ_n}{dP_n} = \log \left\{ \frac{L^{m,n}_{\vec{t}^\star}(\theta^m)}{L^{m,n}_{\vec{t}^\star}(\theta^m_0)} \right\}\ . \tag{37}$$

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

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

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 (34) is also true under the alternative $\mathcal{H}_{a\vec{t}^\star}$.

It concludes the proof of Theorem 1. ■

10. Proof of the squeleton of the covariance function of $Z(.)$

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

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

$$\begin{aligned}
&\forall (t_k, t_{k'}) \in \mathbb{T}_K^1 \backslash \mathbb{T}_K^2 \times \mathbb{T}_K^1 \backslash \mathbb{T}_K^2 \\
&\mathrm{Cov}\left(Z(t_k), Z(t_{k'})\right) = \Big\{ \mathcal{A}\rho(t_k, t_{k'}) + \mathcal{C}\Big\{ \alpha_2(t_k)\alpha_2(t_{k'})\rho(t_k^{\ell,2}, t_{k'}^{\ell,2}) + \alpha_2(t_k)\beta_2(t_{k'})\rho(t_k^{\ell,2}, t_{k'}^{r,2}) \\
&+\beta_2(t_k)\alpha_2(t_{k'})\rho(t_k^{r,2}, t_{k'}^{\ell,2}) + \beta_2(t_k)\beta_2(t_{k'})\rho(t_k^{r,2}, t_{k'}^{r,2}) \Big\}\Big\} / \sqrt{\{\mathcal{A} + \mathcal{C}\xi_2^2(t_k)\}\{\mathcal{A} + \mathcal{C}\xi_2^2(t_{k'})\}} \ .
\end{aligned}$$

Besides, since

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

then,

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

Last, we have

$$\begin{aligned}
\forall (t_k, t_{k'}) \in \mathbb{T}_K^2 \times \mathbb{T}_K^1 \backslash \mathbb{T}_K^2 \quad \mathrm{Cov}\left(Z(t_k), Z(t_{k'})\right) &= \frac{\mathcal{A}\rho(t_k, t_{k'}) + \mathcal{C}\Big\{\alpha_2(t_{k'})\rho(t_k, t_{k'}^{\ell,2}) + \beta_2(t_{k'})\rho(t_k, t_{k'}^{r,2})\Big\}}{\sqrt{(\mathcal{A} + \mathcal{C})(\mathcal{A} + \mathcal{C}\xi_2^2(t_{k'}))}} \\
&= \frac{\sqrt{\mathcal{B}}\rho(t_k, t_{k'})}{\sqrt{\mathcal{A} + \mathcal{C}\xi_2^2(t_{k'})}} \ .
\end{aligned}$$

11. Proof of Theorem 2

Let us consider $n^\star$ individuals for an experiment under selective genotyping. Recall that n is the number of individuals under the complete data situation, 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 (31), we obtain easily that when $t \notin \mathbb{T}_K^1$,

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

where

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

Under the complete data situation (Azaïs et al. (2012)), we have $S_-^1 = S_-^2 = S_+^2 = S_+^1$, so that $\mathcal{C} = 0$ and $\mathcal{A} = \mathcal{B} = \sigma^2$. As a result,

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

As a consequence, if we suppose $\forall s \; a_s > 0$ and consider a one sided test, the statistical test in selective genotyping is more powerful than the one regarding the complete data situation, as soon as

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

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. $t_k \in \mathbb{T}_K^1 \backslash \mathbb{T}_K^2$) and iii) ($t_k \in \mathbb{T}_K^2$) can easily be obtained by continuity.

Let us proove iv). In order to make the results general, we will consider the case $t \notin \mathbb{T}_K^1$. To begin with, let us replace the term $\mathcal{C}$ by $\mathcal{B} - \mathcal{A}$ in the expression of the efficiency κ (see above). We have

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

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 $\mathcal{A} = \sigma^2\left\{\gamma_1 + k_1(z_{\gamma_1^+})\right\}$ 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

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

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.

References

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

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

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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

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.

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

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

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.

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.

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.

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.*

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

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.

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.

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.

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.

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

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

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.

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.

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.

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

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.

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

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

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.

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.

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**.

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.

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.

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.

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.

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.*

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

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.

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

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.

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.

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.

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

Rabier, C.E. (2014b). An asymptotic test for Quantitative Trait Locus detection in presence of missing genotypes. *Annales de la facultÃ© des sciences de Toulouse*, **6(23)** 755-778.

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.

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

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

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

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

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.

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.

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

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.

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

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.

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.

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.

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

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.

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.

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.

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.

Zhao, Y., Gowda, M., Longin, F. H., WÃ$\frac{1}{4}$rschum, 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.

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.

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

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

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

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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

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.

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

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

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.

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.

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.

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.*

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

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.

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.

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.

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.

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

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

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.

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.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, **136**, (2), 245-257.

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.

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

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.

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

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

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.

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.

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**.

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.

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.

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.

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.

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.*

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

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.

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

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.

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.

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.

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

Rabier, C.E. (2014b). An asymptotic test for Quantitative Trait Locus detection in presence of missing genotypes. *Annales de la facultÃ© des sciences de Toulouse*, **6(23)** 755-778.

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.

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

Rabier, C.E, Delmas, C. (2019). The SgLasso and its cousins for selective genotyping and extreme sampling: application to association studies and genomic selection. *submitted*, hal-02123295.

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

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

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

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.

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.

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

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.

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

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.

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.

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.

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

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.

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.

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.

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.

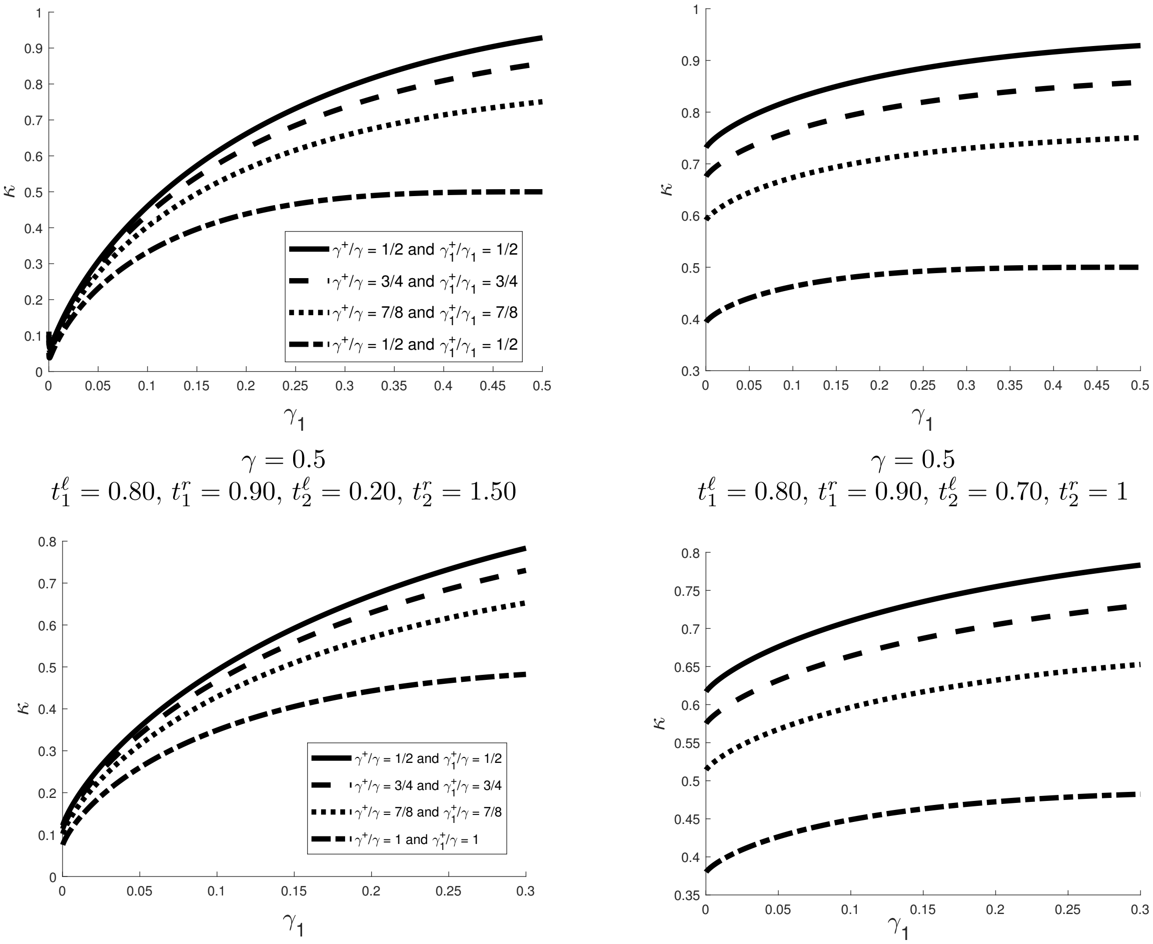
Zhao, Y., Gowda, M., Longin, F. H., WÃ¼rschum, T., Ranc, N., Reif, J. C. (2012). Impact of selective genotyping in the training population on accuracy and bias of genomic selection. *Theoretical and Applied Genetics*, **125**, (4), 707-713.

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.

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 (ce.rabier@gmail.com)
ISEM, Université de Montpellier, CNRS, France.

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. Two different genetic maps are considered, only one QTL is considered ($m = 1$, $a = 2$, $\sigma = 1$) and the test is performed exactly at the QTL location ($t = t_1^\star = 0.85$).



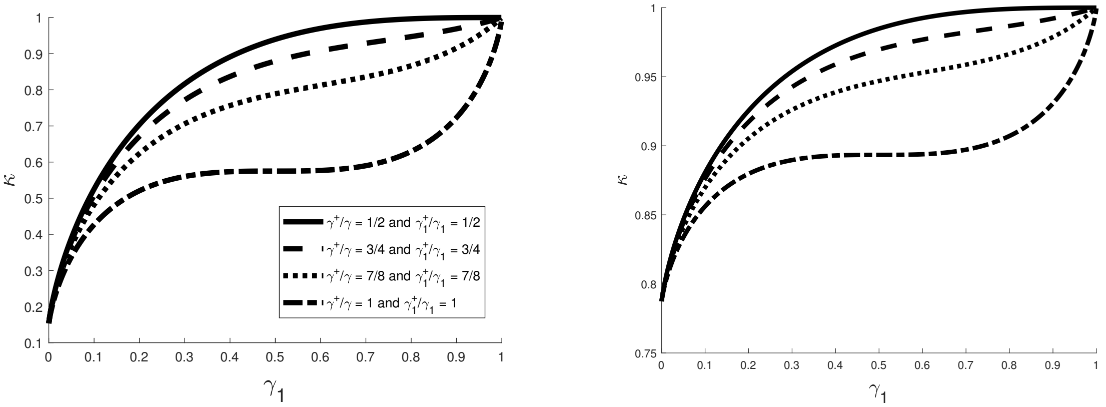
$\gamma = 0.5$
$t_1^\ell = 0.80$, $t_1^r = 0.90$, $t_2^\ell = 0.20$, $t_2^r = 1.50$

$\gamma = 0.5$
$t_1^\ell = 0.80$, $t_1^r = 0.90$, $t_2^\ell = 0.70$, $t_2^r = 1$

$\gamma = 0.3$
$t_1^\ell = 0.80$, $t_1^r = 0.90$, $t_2^\ell = 0.20$, $t_2^r = 1.50$

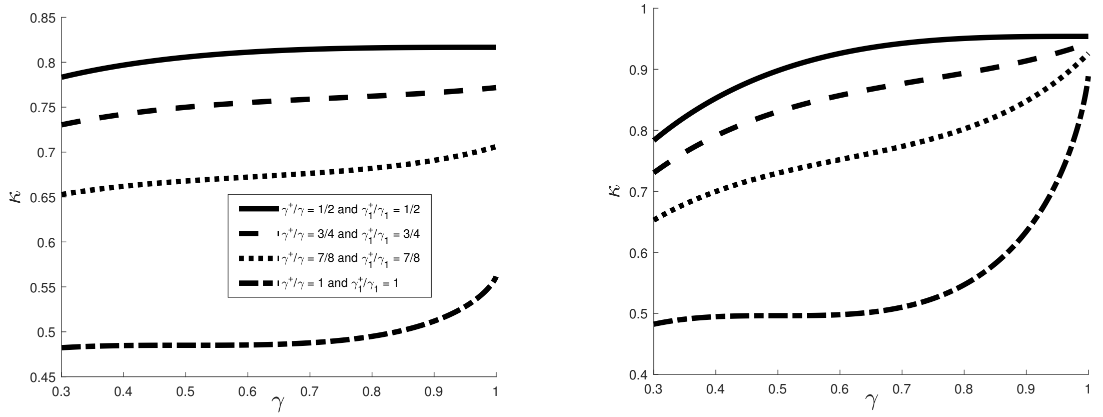
$\gamma = 0.3$
$t_1^\ell = 0.80$, $t_1^r = 0.90$, $t_2^\ell = 0.70$, $t_2^r = 1$

Figure 2: Efficiency κ as a function of γ_1, and as a function of the ratios γ_1^+/γ_1 and γ^+/γ. The two graphs correspond to two different genetic maps. 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 ($t = t_1^\star = 0.85$).



$t_1^\ell = 0.80$, $t_1^r = 0.90$, $t_2^\ell = 0.20$, $t_2^r = 1.50$ $t_1^\ell = 0.80$, $t_1^r = 0.90$, $t_2^\ell = 0.70$, $t_2^r = 1$

Figure 3: Efficiency κ as a function of γ, and as a function of the ratios γ_1^+/γ_1 and γ^+/γ. The two graphs correspond to two different genetic maps. 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 ($t = t_1^\star = 0.85$).



$t_1^\ell = 0.80$, $t_1^r = 0.90$, $t_2^\ell = 0.20$, $t_2^r = 1.50$ $t_1^\ell = 0.80$, $t_1^r = 0.90$, $t_2^\ell = 0.70$, $t_2^r = 1$