On gene mapping with the mixture model and the extremes

Charles-Elie Rabier, Céline Delmas

arXiv: arXiv:0000.0000

On gene mapping with the mixture model and the extremes

Charles-Elie Rabier, Céline Delmas

CHARLES-ELIE RABIER
INRA, UR 875 MIAT
BP 52627, Castanet-Tolosan
31326 Cedex, France
e-mail: ce.rabier@gmail.com

CELINE DELMAS
INRA, UR 875 MIAT
BP 52627, Castanet-Tolosan
31326 Cedex, France
e-mail: celine.delmas@toulouse.inra.fr

Abstract: We introduce a new variable selection method, suitable when the correlation between regressors is known. It is appropriate in genomics since once the genetic map has been built, the correlation is perfectly known. Our method, based on the LASSO , is original since the number of selected variables is bounded by the number of predictors, instead of being bounded by the number of observations as in the classical LASSO. It is made possible by the construction of a specific statistical test, a transformation of the data and by the knowledge of the correlation between regressors. We prove that the signal to noise ratio is largely increased by considering the extremes. This new technique is inspired by stochastic processes arising from statistical genetics. It is described in a statistical genetics context, considering a large panel of models present in the literature. Our method is insensitive to interactions between regressors. An illustration on simulated data is given.

AMS 2000 subject classifications: Primary 60G15, 62F03, 62F05.
Keywords and phrases: Gaussian process, Mixture model, Hypothesis testing, Extreme values, Selective genotyping, Quantitative Trait Locus detection.

1. Introduction and background

1.1. Preliminaries

There are many issues related to high dimensional data. As mentioned in Fan and Lv [15], one of the challenge is that "important predictors can be highly correlated with some unimportant ones". In genomics, correlation between predictors is highly linked to recombination between genetic markers. Then, once the genetic map is built (see Wu et al. [42] for instance), the correlation between predictors is perfectly known and we do not have to estimate these correlations.

In this context, we propose in this study, to exploit this extra information and to introduce a new variable selection method.

The number of selected variables by our method, is bounded by the number of predictors, instead of being bounded by the number of observations as in the classical LASSO (Tibshirani [39]). It is made possible by the construction of a specific statistical test, a transformation of the data and by this knowledge of the correlation between regressors. In other words, we transform the original problem of n observations and K predictors ($K \geq n$), to a situation where K is now the number of observations and L the number of predictors ($L \geq K$). The use of a mixture model allows us to look for variables that can be viewed as unobserved predictors. The quantity L denotes the total number of predictors (observed and unobserved).

Moreover, in high dimensional problem, it is well known that the sparse coefficient should be large enough (see for instance Bühlmann and Van de Geer [8]) in order to recover the true model (β-min condition). We prove that the performances of our method can be largely increased by considering the extremes. Indeed, the signal to noise ratio can be largely improved with the help of the selective genotyping concept proposed by Lebowitz and al. [22], and heavily used in agronomy.

Our study, inspired by stochastic processes arising from biology, focuses on the backcross design (see below): the mathematical theory behind this concept has been largely studied for many years (e.g. [12]). Note that we could have focused on an evolutionary process such as the Wright Fisher model. It could be investigated in future research.

1.2. A statistical genetic context

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 Quantitative Trait Loci, so-called QTL (genes 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 $Y_j,\ j = 1, ..., n$, the observations, which we will assume to be independent and identically distributed (i.i.d.). The mechanism of genetics, or more precisely of meiosis, implies that among the two chromosomes of each individual, one is purely inherited from A while the other (the "recombined" one), consists of parts originated from A and parts originated from B, due to crossing-overs. 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. The genome $X(t)$ of one individual takes the value $+1$ if, for example, the "recombined chromosome" is originated from A at location t and takes the value -1 if it is originated from B . The admitted model for the stochastic structure of $X(.)$ is due to Haldane [18] which states that:

$$X(0) \sim \frac{1}{2}(\delta_{+1} + \delta_{-1}),\quad X(t) = X(0)(-1)^{N(t)}$$

where for any $b \in \mathbb{R}$, δ_b denotes the point mass at b and $N(.)$ is a standard Poisson process on $[0,T]$. In a more practical point of view, the Haldane [18] model assumes no crossover interference and the Poisson process represents the number of crossovers on $[0,T]$ which happen during meiosis.

1.3. Analysis of variance model

The quantitative trait Y is affected by m additive QTLs located on the chromosome. 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 [20]). Let q_s and $t_s^\star$ denote respectively the QTL effect and the location of the sth QTL. Besides, we will consider $0 < t_1^\star < ... < t_m^\star < T$. We assume an "analysis of variance model" for the quantitative trait:

$$Y = \mu + \sum_{s=1}^{m} X(t_s^\star)\, q_s + \sigma\varepsilon \tag{1.1}$$

where ε is a Gaussian white noise.

1.4. Mixture model

In fact the "genome information" will be available only at marker locations, that is to say at certain fixed locations $t_1 = 0 < t_2 < ... < t_K = T$, and the observation will be

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

So, we observe n observations $(Y_j,\ X_j(t_1),\ ...,\ X_j(t_K))$ i.i.d.
The aim of this study is to estimate the number m of QTLs, their locations $t_1^\star$, ..., $t_m^\star$ and their effects q_1, ..., q_m. If the QTLs were located exactly on marker locations, a classical way to solve this problem would be to perform the "least absolute shrinkage and selection operator", so called LASSO (Tibshirani [39]) using as regressors the genome information at marker locations. However, since QTLs lie on the chromosome at unknown locations, we should not look for QTLs only at marker locations but rather focus on the whole chromosome. As a result, our problem can not be solved using classical variable selection tools: QTL mapping requires the use of mixture models in a way we will explain below.

In what follows, $r(t,t')$ will denote the probability of recombination between two loci (i.e. positions) located at t and t'. Calculation on the Poisson distribution show that

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

we set in addition

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

When there is only one QTL (i.e. $m = 1$), conditionally to $X(t_1), \ldots, X(t_K)$, Y obeys to a mixture model with known weights

$$p(t_1^\star) f_{(\mu+q_1,\sigma)}(.) + \{1 - p(t_1^\star)\} f_{(\mu-q_1,\sigma)}(.), \tag{1.2}$$

where $f_{(\mu,\sigma)}$ is the Gaussian density with parameters (μ, σ) and where $p(t_1^\star)$ is the probability $\mathbb{P}(X(t_1^\star) = 1)$ conditionally to the observations of the markers. It can be expressed from the functions r and $\overline{r}$ (see formula 4.1 for $K = 2$, and formula 5.1 for $K > 2$).

1.5. The "Interval Mapping" of Lander and Botstein [21]

In a famous article, Lander and Botstein [21] proposed to test the presence of the QTL (i.e. $m = 0$ vs $m = 1$), performing a likelihood ratio test (LRT) of the null hypothesis "$q_1 = 0$" in equation (1.2). Since t_1^* is unknown, the authors suggested to scan the chromosome and to perform a LRT, $\Lambda_n(t)$, at each location t of the interval $[0, T]$. It leads to a "LRT process", $\Lambda_n(.)$, and considering the supremum of $\Lambda_n(.)$ gives the LRT of "$q_1 = 0$" on the whole chromosome. Note that when the null hypothesis of the absence of QTL on $[0, T]$ is rejected, $\arg\sup \Lambda_n(t)$ is a natural estimator of the QTL location. This method, very popular in genetics, is called the "Interval Mapping". The distribution of the LRT statistic, $\sup \Lambda_n(.)$, has been given using some approximations by Cierco [12], Azaïs and Cierco-Ayrolles [1], Azaïs and Wschebor [5]. In Rebaï et al. [36], Rebaï et al. [35] and Chang et al. [9], the authors focus only on the null hypothesis and are still using some approximations. Theoretical results are also present in Chen and Chen [10] under non contiguous hypotheses. However, geneticists are usually interested in detecting QTLs with small effects (see Hayes [20]). Then, in Azaïs et al. [2], we have recently given the exact asymptotic distribution of the LRT statistic under the null and contiguous hypotheses. We showed that the LRT process, $\Lambda_n(.)$, is asymptotically the square of a "non linear interpolated process" centered under H_0 (i.e. no QTL on the chromosome) and uncentered of a mean function under the alternative which depends on the QTL effect q_1 and its location $t_1^\star$. Then, we presented a formula (due to the interpolation) to compute the supremum of $\Lambda_n(.)$.

1.6. First contribution: Asymptotic results on max test and LRT process, and a new gene mapping method

The problem is that the use of the test statistic $\sup \Lambda_n(.)$ is appropriate for testing and localizing one QTL on $[0, T]$, but it is not so rewarding when more than one QTL (i.e. $m > 1$) lie on $[0, T]$. When there are m QTLs, conditionally on $X(t_1), \ldots, X(t_K)$, Y obeys to a mixture of 2^m components

$$\sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} w_{\vec{t}^\star}(u_1, \ldots, u_m)\, f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)$$

where $w_{\vec{t}^\star}(u_1,\ ...,\ u_m)$ is the probability $\mathbb{P}\{X(t_1^\star)=u_1,...,X(t_m^\star)=u_m\}$ conditionally on the observations of the markers (see Sections 4 and 5).

In this paper, we propose to generalize the results of Azaïs et al. [2] to the general alternative that there exist m QTLs on $[0,T]$ at $t_1^\star,\cdots,t_m^\star$ with additive effects $q_1,\cdots,q_m$. We will show that under the general alternative, the LRT process is still asymptotically the square of a "non linear interpolated process". However, the mean function depends this time on the number of QTLs, their positions and their effects. This theoretical result allows us to propose a new method to estimate the number of QTLs, their positions and their effects using the LASSO.

Note that we will also give the asymptotic distribution of the statistic $\sup\Lambda_n(.)$ when $m>1$, since this test can be viewed as a global test or max test (see for instance [6]). In this context, $\sup\Lambda_n(.)$ matches the test statistic corresponding to the statistical test with the smallest pvalue in a multiple testing framework. It could be used before performing our new gene mapping method, in order to look for "some signal" on the chromosome.
Section 10.1 illustrates on simulated data, our theoretical result regarding the max test. We will show that the empirical power matches the theoretical power for moderates values of n. Besides, we will illustrate in Section 10.3, that our new method is able to recover the genes (lying between markers) in the ideal noiseless situation (see Donoho [14]) provided that the correlation between regressors is not too high. Recall that the correlation depends on the intensity of the Poisson process that models recombinations along the genome. Note that as expected, performances will deteriorate in the noisy setting, with false positives appearing. A way to increase the signal to noise ratio is to focus on the extremes (see below).

1.7. *Some background about the use of the extremes in genomics: the selective genotyping*

In the past, collecting the genome information at one marker for all the individuals was very expensive. In such a context, Lebowitz and al. [22] proposed to genotype only the individuals who present an extreme phenotype (i.e. the smallest and the largest Y), since they noticed that most of the information about the QTL is present in the extreme phenotypes. This way, at a given power, a large increase of the number of individuals leads to a decrease of the number of individuals genotyped. Later, Lander and Botstein [21] formalized this approach and called it "selective genotyping". This design has been studied theoretically by many authors considering only one fixed location of the genome (e.g. Lebowitz and al. [22], Lander and Botstein [21], Darvasi and Soller [13], Muranty and Goffinet [28], Rabier [30]). More recently, in Rabier [32], we investigated the asymptotic properties of the LRT statistic on the chromosome: it can be viewed as an answer to the simulation study presented by Rabbee et al. [29].

The model corresponding to selective genotyping is the following: we consider two real thresholds S_- and S_+, with $S_-\leq S_+$ and we genotype if and

only if the phenotype Y is extreme, that is to say $Y \leq S_-$ or $Y \geq S_+$. Note that in practice, the cutoffs for genotyping are based on quantiles. However, in most of the theoretical studies about selective genotyping, authors consider fixed thresholds. This approximation is reasonable when we deal with a large number of observations.

If we call $\overline{X}(t)$ the random variable such as

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

then, in our problem, one observation will be now

$$\left(Y, \ \overline{X}(t_1), \ ..., \ \overline{X}(t_K)\right) .$$

Note that with our notations :

- when $Y \notin [S_- \ , \ S_+]$, we have $\overline{X}(t_1) = X(t_1), \ ..., \overline{X}(t_K) = X(t_K)$.
- when $Y \in [S_- \ , \ S_+]$, we have $\overline{X}(t_1) = 0, \ ..., \overline{X}(t_K) = 0$, which means that the genome information is missing at the marker locations.

When there is only one QTL (i.e. $m = 1$), we have proved (see Rabier [32]) that the probability distribution of $\left(Y, \ \overline{X}(t_1), \ ..., \ \overline{X}(t_K)\right)$ is proportional to the mixture

$$\begin{aligned} & p(t_1^\star) \ f_{(\mu+q_1,\sigma)}(Y) \ 1_{Y \notin [S_-,S_+]} \ + \ \{1 - p(t_1^\star)\} \ f_{(\mu-q_1,\sigma)}(Y) \ 1_{Y \notin [S_-,S_+]} \\ & + \ \frac{1}{2} \ f_{(\mu+q_1,\sigma)}(Y) \ 1_{Y \in [S_-,S_+]} \ + \ \frac{1}{2} \ f_{(\mu-q_1,\sigma)}(Y) \ \ 1_{Y \in [S_-,S_+]} \ . \end{aligned} \tag{1.3}$$

Recall that the function $p(t_1^\star)$ is the probability $\mathbb{P}(X(t_1^\star) = 1 \mid X(t_1), \ldots, X(t_K))$. Note that although $p(t_1^\star)$ is a function of $X(t_1), \ldots, \ X(t_K)$, the quantity $p(t_1^\star) 1_{Y \notin [S_-,S_+]}$ present in (1.3) is a function of $\overline{X}(t_1), \ ..., \ \overline{X}(t_K)$. In this context, we have proved (Rabier [32]) that the LRT process, $\Lambda_n(.)$, converges to the square of a non linear interpolated process. This limiting process is the same as the one of the complete data situation (as above) except that the mean functions are proportional of a factor linked to the selective genotyping.

1.7.1. Second contribution: Asymptotic results regarding the extremes (selective genotyping)

As explained before, we propose to tackle in this study, the problem of recovering several genes lying on the genome. So, under selective genotyping, when there are m QTLs, the probability distribution of $\left(Y, \ \overline{X}(t_1), \ ..., \ \overline{X}(t_K)\right)$ is proportional to the mixture of 2^m components

$$\begin{aligned} & \sum_{(u_1,\ldots,u_m) \in \{-1,1\}^m} w_{\vec{t^\star}}(u_1, \ ..., \ u_m) \ f_{(\mu+u_1 q_1+\ldots+u_m q_m,\sigma)}(Y) \ 1_{Y \notin [S_-,S_+]} \\ & + v_{\vec{t^\star}}(u_1, \ ..., \ u_m) \ f_{(\mu+u_1 q_1+\ldots+u_m q_m,\sigma)}(Y) \ 1_{Y \in [S_-,S_+]} \end{aligned}$$

where $v_{\vec{t}^\star}(u_1,\ ...,\ u_m)$ is the probability $\mathbb{P}(X(t_1^\star)=u_1, X(t_2^\star)=u_2,\ldots,X(t_m^\star)=u_m)$ and where $w_{\vec{t}^\star}(u_1,\ ...,\ u_m)$ is the same quantity as previously. A proof about this mixture model is given in Section 11.3.1.

We will show that the mean function of the LRT process is still proportional to the one of the complete data situation. Note that the proof is somewhat technical and some parts are presented in supplementary material. Besides, we will compare theoretically the case where the n genotyped individuals are extreme or not. Last, we will illustrate in Section 10.2, our new gene mapping method under selective genotyping. As expected, the signal to noise ratio is largely increased by considering extreme individuals. Recall that our proposed method takes into account explicitly the fact that the individuals are extreme, since it relies on the mean function of the LRT process under selective genotyping.

2. Extra models studied in this paper

In this article, we will also investigate the asymptotic properties of the LRT process regarding other models present in the statistical genetics litterature. Then, our new gene mapping method, will be suitable in a general framework.

2.1. Epistatic model

It is well known that interactions between QTLs (so-called epistasis phenomenon) can be responsible for a non-negligible part of the genetic variability of a quantitative trait (see for instance Wu et al. [42]). Then, we propose to include interactions into the model. We will assume that only loci with additive effects on the trait, are involved in the interactions. The "analysis of variance model" of formula (1.1) for the quantitative trait becomes

$$Y=\mu\ +\ \sum_{s=1}^{m} X(t_s^\star)\ q_s\ +\ \sum_{s=1}^{m-1}\sum_{\tilde{s}=s+1}^{m} X(t_s^\star)X(t_{\tilde{s}}^\star)\ q_{s,\tilde{s}}\ +\ \sigma\varepsilon \tag{2.1}$$

where ε is a Gaussian white noise, and $q_{s,\tilde{s}}$ is the interaction effect between loci $t_s^\star$ and $t_{\tilde{s}}^\star$.

Conditionally on $X(t_1),\ldots,X(t_K)$, Y obeys now to the following mixture of 2^m components

$$\sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} w_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{\left(\mu+\sum_{s=1}^{m} u_s q_s+\sum_{s=1}^{m-1}\sum_{\tilde{s}=s+1}^{m} u_s u_{\tilde{s}} q_{s,\tilde{s}}\ ,\ \sigma\right)}(Y)\ . \tag{2.2}$$

Recall that $w_{\vec{t}^\star}(u_1,\ ...,\ u_m)$ is the probability $\mathbb{P}\left\{X(t_1^\star)=u_1,...,X(t_m^\star)=u_m\right\}$ conditionally on the observations of markers.

In this context, we will prove that the interaction effects are not included in the mean function of the LRT process. In other words, those effects are unidentifiable when the classical LRT is used. As a consequence, our gene mapping method is insensitive to epistatic loci in the genome. The method enables exclusively the detection of additive effects.

2.2. Interference model

Recall that Haldane modeling assumes that crossovers occur independently along the genome. In fact, biologists have observed that a recombination event can inhibit the formation of another recombination event nearby (e.g. Sturtevant [38], Muller [27], McPeek and Speed [26], Lobo and Shaw [23]). This phenomenon is called interference phenomenon.

We propose to focus here on the model introduced by Rebaï et al. [36] (see in particular their Section 2) in which double recombination between the QTL and its flanking markers is not allowed. Note that in Rebaï et al. [35], the authors extended their previous model to several markers, keeping Haldane [18] modeling for the genetic information at marker locations. As a result, the probability distribution of $(X(t_1), \ldots, X(t_K))$ is unchanged. In order to extend the interference model to m QTLs, we will impose that the QTLs do not belong to the same marker intervals. Obviously, double recombination between each QTL and the corresponding flanking markers is not allowed. Then, under the interference model, the "analysis of variance model" for the quantitative trait is the following:

$$Y = \mu \ + \ \sum_{s=1}^{m} U(t_s^\star)\ q_s \ + \ \sigma\varepsilon \tag{2.3}$$

where ε is a Gaussian white noise, and $U(t_s^\star)$ is the analogue of $X(t_s^\star)$ under interference (more details in Section 8).

When there are m QTLs, conditionally to $X(t_1), \ldots, X(t_K)$, we will show that Y obeys to a mixture of 2^m components

$$\sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \widetilde{w}_{\vec{t}^\star}(u_1,\ \ldots,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)$$

where $\widetilde{w}_{\vec{t}^\star}(u_1,\ \ldots,\ u_m)$ is the probability $\mathbb{P}\left\{U(t_1^\star) = u_1, \ldots, U(t_m^\star) = u_m\right\}$ conditionally to the observations of the markers. Note that the weights $\widetilde{w}_{\vec{t}^\star}(u_1,\ \ldots,\ u_m)$ are different from the weights $w_{\vec{t}^\star}(u_1,\ \ldots,\ u_m)$ obtained under Haldane (cf. formula 2.2). In this study, we will give asymptotic results about the LRT process under interference.

3. Roadmap

In this article, we will present the following theorems and lemmas corresponding to the different models studied:

- Theorem 4.1 : Two genetic markers
- Theorem 5.1: Several markers
- Theorem 6.1: Epistasis
- Theorem 7.1 : Selective genotyping
- Lemma 7.1 : Asymptotic Relative Efficiency for selective genotyping

- Theorem 7.2 : Selective genotyping + epistasis
- Lemma 7.2: Reverse configuration of selective genotyping
- Theorem 8.1: Interference
- Theorem 8.2: Interference + epistasis

Finally, in Section 9, we will introduce our new gene mapping method.

4. Two genetic markers

To begin, we suppose that there are only two markers ($K = 2$) located at 0 and $T : 0 = t_1 < t_2 = T$. Besides, let us consider the case $m = 1$ (i.e. one QTL located at $t_1^\star$). For $t \in [t_1, t_2]$ we define

$$p(t) = \mathbb{P}\left\{X(t) = 1 \middle| X(t_1), X(t_2)\right\}$$

and

$$x(t) = \mathbb{E}\left\{X(t) \middle| X(t_1), X(t_2)\right\} = 2p(t) - 1.$$

It is clear that $p(t_1^\star)$ is effectively the probability appearing in (1.2). An application of the rule of total probabilities leads to

$$\begin{aligned} p(t) = {} & Q_t^{1,1}\ 1_{X(t_1)=1}1_{X(t_2)=1} \ + \ Q_t^{1,-1}\ 1_{X(t_1)=1}1_{X(t_2)=-1} \\ & + Q_t^{-1,1}\ 1_{X(t_1)=-1}1_{X(t_2)=1} \ + \ Q_t^{-1,-1}\ 1_{X(t_1)=-1}1_{X(t_2)=-1} \end{aligned} \tag{4.1}$$

where

$$\begin{aligned} Q_t^{1,1} &= \frac{\bar{r}(t_1,t)\ \bar{r}(t,t_2)}{\bar{r}(t_1,t_2)} \quad , \quad Q_t^{1,-1} = \frac{\bar{r}(t_1,t)\ r(t,t_2)}{r(t_1,t_2)} \\ Q_t^{-1,1} &= \frac{r(t_1,t)\ \bar{r}(t,t_2)}{r(t_1,t_2)} \quad , \quad Q_t^{-1,-1} = \frac{r(t_1,t)\ r(t,t_2)}{\bar{r}(t_1,t_2)}. \end{aligned}$$

We can notice that we have

$$Q_t^{-1,-1} = 1 - Q_t^{1,1} \quad \text{and} \quad Q_t^{-1,1} = 1 - Q_t^{1,-1}.$$

Let $\theta^1 = (q_1,\ \mu,\ \sigma)$ be the parameter of the model at t fixed. The likelihood of the triplet $(Y,\ X(t_1),\ X(t_2))$ with respect to the measure $\lambda \otimes N \otimes N$, λ being the Lebesgue measure, N the counting measure on $\mathbb{N}$, is $\forall t \in [t_1, t_2]$:

$$L_t(\theta^1) = \left[p(t) f_{(\mu+q_1,\sigma)}(Y) + \{1 - p(t)\}\, f_{(\mu-q_1,\sigma)}(Y)\right]\ g(t) \tag{4.2}$$

where the function

$$\begin{aligned} g(t) = {} & \frac{1}{2}\left\{\bar{r}(t_1,t_2)\ 1_{X(t_1)=1}1_{X(t_2)=1} \ + \ r(t_1,t_2)\ 1_{X(t_1)=1}1_{X(t_2)=-1}\right\} \\ & + \frac{1}{2}\left\{r(t_1,t_2)\ 1_{X(t_1)=-1}1_{X(t_2)=1} \ + \ \bar{r}(t_1,t_2)\ 1_{X(t_1)=-1}1_{X(t_2)=-1}\right\} \end{aligned} \tag{4.3}$$

can be removed because it does not depend on the parameters. By a small abuse of notation we still denote $L_t(\theta^1)$ for the likelihood without this function. Thus we set

$$L_t(\theta^1) = \left[p(t) f_{(\mu+q_1,\sigma)}(Y) + \{1-p(t)\} f_{(\mu-q_1,\sigma)}(Y)\right]$$

and $l_t(\theta^1)$ will be the loglikelihood.

Before defining the score statistic and the LRT statistic at t, let us introduce the notation $\theta_0^1 = (0,\ \mu,\ \sigma)$ which will refer to the parameter θ^1 under H_0. Since the Fisher Information matrix is diagonal (cf. Section 11.1), the score statistic of the hypothesis "$q_1 = 0$" at t, for n independent observations, will be defined as

$$S_n(t) = \frac{\frac{\partial l_t^n}{\partial q_1} \mid_{\theta_0^1}}{\sqrt{\mathbb{V}_{H_0}\left(\frac{\partial l_t^n}{\partial q_1} \mid_{\theta_0^1}\right)}}, \tag{4.4}$$

where $l_t^n(\theta^1)$ denotes the log-likelihood at t, associated to n observations.

The LRT at t, for n independent observations, will be defined as

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

where $\widehat{\theta^1}$ is the maximum likelihood estimator (MLE), and $\widehat{\theta^1}_{|H_0}$ the MLE under H_0. As previously said, $\sup \Lambda_n(.)$ is the LRT statistic of $q_1 = 0$ on the whole interval $[0,T]$. We refer to Azaïs et al. [2] for the asymptotic distributions of the LRT statistic.

Let us now suppose that more than one QTL (i.e. $m > 1$) lie on $[0,T]$. In what follows, $t_1^\star$, ..., $t_m^\star$ will denote the QTL locations, and we define the parameter θ^m and θ_0 in the following way : $\theta^m = (q_1, ..., q_m, \mu, \sigma)$ and $\theta_0^m = (0, ..., 0, \mu, \sigma)$. Then, the full likelihood of the triplet $(Y,\ X(t_1),\ X(t_2))$, with respect to the measure $\lambda \otimes N \otimes N$, is

$$L_{\vec{t}^\star}^m(\theta^m) = \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} w_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)\ g^m(t_1^\star,\ldots,t_m^\star) \tag{4.5}$$

where the function $g^m(.)$ is equal to the function $g(.)$ given in formula (4.3) and where

$$w_{\vec{t}^\star}(u_1,\ ...,\ u_m) = \mathbb{P}\left\{X(t_1^\star) = u_1, ..., X(t_m^\star) = u_m \mid X(t_1), X(t_2)\right\}.$$

Note that at this time, it is clear that $g^m(.)$ does not depend on $t_1^\star$, $\ldots$,$t_m^\star$. However, these parameters will be useful in the generalization in the next Section. Note that calculations on the Poisson process lead to (proof included in the proof of Theorem 4.1, see Section 11.1)

$$\begin{aligned}
&w_{\vec{t}^\star}(u_1,\ ...,\ u_m)\\
&= \left\{r(t_1,t_1^\star)\ 1_{X(t_1)u_1=-1} + \bar{r}(t_1,t_1^\star)\ 1_{X(t_1)u_1=1}\right\}\left\{r(t_1^\star,t_2^\star)\ 1_{u_1u_2=-1} + \bar{r}(t_1^\star,t_2^\star)\ 1_{u_1u_2=1}\right\}\\
&\ldots\left\{r(t_{m-1}^\star,t_m^\star)\ 1_{u_{m-1}u_m=-1} + \bar{r}(t_{m-1}^\star,t_m^\star)\ 1_{u_{m-1}u_m=1}\right\}\\
&\left\{r(t_m^\star,t_2)\ 1_{u_mX(t_2)=-1} + \bar{r}(t_m^\star,t_2)1_{u_mX(t_2)=1}\right\} / \left\{r(t_1,t_2)\ 1_{X(t_1)X(t_2)=-1} + \bar{r}(t_1,t_2)\ 1_{X(t_1)X(t_2)=1}\right\}.
\end{aligned}$$

As previously, we set $L^m_{\vec{t}^\star}(\theta^m)$ the likelihood without the function $g(.)$. Note that since

$$\sum_{(u_2,\ldots,u_m)\in\{-1,1\}^{m-1}} \mathbb{P}\left\{X(t_1^\star)=1, X(t_2^\star)=u_2,\ldots,X(t_m^\star)=u_m \;\middle|\; X(t_1),X(t_2)\right\} = p(t^\star),$$

we have the relationship

$$L^m_{\vec{t}^\star}\left(\theta^m_{10\ldots0}\right) = L_{t_1^\star}(\theta^1) \;\text{ where }\; \theta^m_{10\ldots0} = (q_1,0,\ldots,0,\mu,\sigma)\,.$$

In the same way, under H_0,

$$L^m_{\vec{t}^\star}\left(\theta^m_0\right) = L_{t_1^\star}(\theta^1_0).$$

Notations 4.1. *$\Rightarrow$ is the weak convergence, $\overset{F.d.}{\to}$ is the convergence of finite-dimensional distributions and $\overset{\mathcal{L}}{\longrightarrow}$ is the convergence in distribution.*

Our main result is the following:

Theorem 4.1. *Suppose that the parameters $(q_1,\ldots,q_m,\mu,\sigma^2)$ vary in a compact and that σ^2 is bounded away from zero, and also that m is finite. Let H_0 be the null hypothesis of no QTL on $[0,T]$, and let define the following local alternatives $H_{a\vec{t}^\star}$: "there are m QTLs located respectively at $t_1^\star$, ..., $t_m^\star$ with effect $q_1 = a_1/\sqrt{n},\ldots,q_m = a_m/\sqrt{n}$ where $a_1\neq 0,\ldots,a_m\neq 0$" . Then,*

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

as n tends to infinity, under H_0 and $H_{a\vec{t}^\star}$ where $Z(.)$ is the Gaussian process with unit variance such as

$$Z(t) = \frac{\alpha(t)\;Z(t_1)\;+\;\beta(t)\;Z(t_2)}{\sqrt{\alpha^2(t)+\beta^2(t)\;+\;2\alpha(t)\beta(t)\rho(t_1,t_2)}}\;,$$
$$Cov\left\{Z(t_1),Z(t_2)\right\} = \rho(t_1,t_2) = e^{-2|t_1-t_2|}$$

where $\alpha(t) = Q_t^{1,1} - Q_t^{-1,1}$, $\beta(t) = Q_t^{1,1} - Q_t^{1,-1}$ and with mean function

- *under H_0, $m(t)=0$*
- *under $H_{a\vec{t}^\star}$,*

$$m_{\vec{t}^\star}(t) = \frac{\alpha(t)\;m_{\vec{t}^\star}(t_1)+\beta(t)\;m_{\vec{t}^\star}(t_2)}{\sqrt{\alpha^2(t)+\beta^2(t)\;+\;2\alpha(t)\beta(t)\rho(t_1,t_2)}}$$

where

$$m_{\vec{t}^\star}(t_1) = \sum_{s=1}^m a_s\;\rho(t_1,t_s^\star)\;/\;\sigma \;\;,\;\; m_{\vec{t}^\star}(t_2) = \sum_{s=1}^m a_s\;\rho(t_2,t_s^\star)\;/\;\sigma.$$

A proof is given in Section 11.1. This theorem is a generalization of Theorem 2.1 of Azaïs et al. [2], that considers only one QTL on the chromosome. According to Theorem 4.1, under the general alternative, the LRT process is still asymptotically the square of a non linear interpolated process. However, the mean function depends this time on the number of QTLs, their positions and their effects.

5. Several markers

In that case suppose that there are K markers $0 = t_1 < t_2 < ... < t_K = T$. We consider values t, t' or $t^\star$ of the parameters that are distinct of the markers positions, and the result will be prolonged by continuity at the markers positions. For $t \in [t_1, t_K] \backslash \mathbb{T}_K$ where $\mathbb{T}_K = \{t_1, ..., t_K\}$, we define t^ℓ and t^r as :

$$t^\ell = \sup \{t_k \in \mathbb{T}_K : t_k < t\} \quad , \quad t^r = \inf \{t_k \in \mathbb{T}_K : t < t_k\} .$$

In other words, t belongs to the "Marker interval" (t^ℓ, t^r).

Let us briefly describe the changes with the previous section. Due to the independent increments of the Poisson process,

$$\mathbb{P}\{X(t) \mid X(t_1), \ldots, X(t_K)\} = \mathbb{P}\left\{X(t) \mid X(t^\ell), X(t^r)\right\} .$$

As a consequence, the likelihood ratio test $\Lambda_n(t)$ is now built on the likelihood of the triplet $\left(Y,\ X(t^\ell),\ X(t^r)\right)$ and the quantities $p(t)$ and $g(t)$, introduced in formulae (4.1) and (4.3), become

$$p(t) = \mathbb{P}\left\{X(t) = 1 \mid X(t^\ell), X(t^r)\right\} \quad , \quad g(t) = \mathbb{P}\left\{X(t^\ell), X(t^r)\right\} . \tag{5.1}$$

Recall that our test statistic $\Lambda_n(t)$, is the LRT corresponding to the test of the presence of only one QTL at t. Let us now consider the true probability distribution. Since

$$\mathbb{P}\{X(t_1^\star), \ldots, X(t_m^\star) \mid X(t_1), \ldots, X(t_K)\} = \mathbb{P}\left\{X(t_1^\star), \ldots, X(t_m^\star) \mid X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r})\right\}$$

all the information is contained in the markers flanking the QTL locations. As a result, the focus is on the probability distribution of $(Y, X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r}))$ and the quantities $w_{\vec{t}^\star}(u_1,\ ...,\ u_m)$ and $g^m(.)$ present in formula (4.5), verify now

$$w_{\vec{t}^\star}(u_1,\ ...,\ u_m) = \mathbb{P}\left\{X(t_1^\star) = u_1, ..., X(t_m^\star) = u_m \;\middle|\; X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r})\right\} \tag{5.2}$$

and

$$g^m(t_1^\star, \ldots, t_m^\star) = \mathbb{P}\left\{X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r})\right\} . \tag{5.3}$$

Let us now introduce Theorem 5.1.

Theorem 5.1. *We have the same result as in Theorem 4.1, provided that we make some adjustments and that we redefine $Z(.)$ in the following way :*

- *in the definition of $\alpha(t)$ and $\beta(t)$, t_1 becomes t^ℓ and t_2 becomes t^r*
- *under the null hypothesis, the process $Z(.)$ considered at marker positions is the "squeleton" of an Ornstein-Uhlenbeck process: the stationary Gaussian process with covariance $\rho(t_k, t_{k'}) = \exp(-2|t_k - t_{k'}|)$*
- *at the other positions, $Z(.)$ is obtained from $Z(t^\ell)$ and $Z(t^r)$ by interpolation and normalization using the functions $\alpha(t)$ and $\beta(t)$*

- *at the marker positions, the expectation is such as* $m_{\vec{t}^\star}(t_k) = \sum_{s=1}^{m} a_s \rho(t_k, t_s^\star)/\sigma$
- *at other positions, the expection is obtained from* $m_{\vec{t}^\star}(t^\ell)$ *and* $m_{\vec{t}^\star}(t^r)$ *by interpolation and normalization using the functions* $\alpha(t)$ *and* $\beta(t)$.

The proof of the theorem is the same as the proof of Theorem 4.1 since we can limit our attention to the interval (t^ℓ, t^r) when considering a unique instant t.

6. Epistasis

We propose now to include interactions between QTLs (so-called espistasis phenomenon) into our model (see for instance Wu et al. [42]). We will assume that only loci with additive effects on the trait, are involved in interactions. The "analysis of variance model" of formula (1.1) for the quantitative trait becomes

$$Y = \mu \ + \ \sum_{s=1}^{m} X(t_s^\star)\, q_s \ + \ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^{m} X(t_s^\star) X(t_{\tilde{s}}^\star)\, q_{s,\tilde{s}} \ + \ \sigma\varepsilon$$

where ε is a Gaussian white noise, and $q_{s,\tilde{s}}$ is the interaction effect between loci $t_s^\star$ and $t_{\tilde{s}}^\star$.

The probability distribution of $(Y, X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r}))$ is

$$\sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} w_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{\left(\mu+\sum_{s=1}^{m} u_s q_s + \sum_{s=1}^{m-1}\sum_{\tilde{s}=s+1}^{m} u_s u_{\tilde{s}} q_{s,\tilde{s}}\ ,\ \sigma\right)}(Y)\ g^m(t_1^\star, \ldots, t_m^\star) \tag{6.1}$$

where $w_{\vec{t}^\star}(u_1,\ ...,\ u_m)$ and $g^m(.)$ are given in formulae (5.2) and (5.3).

Theorem 6.1. *Suppose that the parameters* $(q_1, ..., q_m, q_{12}, ..., q_{m-1m}, \mu, \sigma^2)$ *vary in a compact and that* σ^2 *is bounded away from zero, and also that* m *is finite. Let define the local alternative*

- $H_{a\vec{t}^\star, b\vec{t}^\star}$ *:"There are* m *additive QTLs located respectively at* $t_1^\star$, ..., $t_m^\star$ *with effects respectively* $q_1 = a_1/\sqrt{n}$, ..., $q_m = a_m/\sqrt{n}$ *where* $a_1 \neq 0$, ..., $a_m \neq 0$ *. Besides, all these QTLs interact with each other : the interaction effects are respectively* $q_{1,2} = b_{1,2}/\sqrt{n}$ *for loci* $t_1^\star$ *and* $t_2^\star$, ..., $q_{m-1,m} = b_{m-1,m}/\sqrt{n}$ *for loci* $t_{m-1}^\star$ *and* $t_m^\star$ *where* $b_{1,2} \neq 0$, ..., $b_{m-1,m} \neq 0$*".*

then, with the previous notations, under $H_{a\vec{t}^\star, b\vec{t}^\star}$ *,*

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

where $Z(.)$ *is the Gaussian process of Theorem 5.1 uncentered with mean function* $m_{\vec{t}^\star}(.)$ *defined in Theorem 5.1.*

A proof is given in Section 11.2. Note that the interaction effects are not included in the mean function. In other words, those effects are unidentifiable when the classical LRT is used. It is due to independent increments of the Poisson process (cf. proof in Section 11.2).

7. Selective genotyping

We propose to consider here the classical problem (as in Sections 4 and 5), but incorporating now a selective genotyping in order to reduce the costs of genotyping. To begin with, in order to make the reading easier, we won't consider interactions in our model. However, the epistasis will be investigated later in this section. As mentioned in Section 1.7, the selective genotyping model is the following: we consider two real thresholds S_- and S_+, with $S_- \leq S_+$ and we genotype if and only if the phenotype Y is extreme, that is to say $Y \leq S_-$ or $Y \geq S_+$.

If we call $\overline{X}(t)$ the random variable such as

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

then, in our problem, one observation will be now

$$\left(Y, \ \overline{X}(t_1), \ ..., \ \overline{X}(t_K)\right) .$$

Note that with our notations :

- when $Y \notin [S_- \ , \ S_+]$, we have $\overline{X}(t_1) = X(t_1), \ ..., \overline{X}(t_K) = X(t_K)$.
- when $Y \in [S_- \ , \ S_+]$, we have $\overline{X}(t_1) = 0, \ ..., \overline{X}(t_K) = 0$, which means that the genome information is missing at the marker locations.

To begin with, let us consider the case $m = 1$. According to Rabier [32], the likelihood of the triplet $\left(Y, \ \overline{X}(t^\ell), \ \overline{X}(t^r)\right)$ with respect to the measure $\lambda \otimes N \otimes N$, λ is $\forall t \in [t_1, t_K] \backslash \mathbb{T}_K$:

$$\begin{aligned} \overline{L}_t(\theta^1) = \Big[\ p(t) \ f_{(\mu+q_1,\sigma)}(Y) 1_{Y \notin [S_-,S_+]} + \{1 - p(t)\} \ f_{(\mu-q_1,\sigma)}(Y) 1_{Y \notin [S_-,S_+]} \\ + \ \frac{1}{2} \ f_{(\mu+q_1,\sigma)}(Y) 1_{Y \in [S_-,S_+]} \ + \ \frac{1}{2} \ f_{(\mu-q_1,\sigma)}(Y) 1_{Y \in [S_-,S_+]} \Big] \ \overline{g}(t) \end{aligned} \tag{7.1}$$

with

$$\overline{g}(t) = \mathbb{P}\left\{X(t^\ell), X(t^r)\right\} \ 1_{Y \notin [S_-,S_+]} \ + \ 1_{Y \in [S_-,S_+]} \ . \tag{7.2}$$

Note that we use the same notations $p(t)$ for the weights since they are exactly the same as in Section 5. Recall that

$$\begin{aligned} p(t) &= \mathbb{P}\left\{X(t) = 1 \mid X(t^\ell), X(t^r)\right\} \\ &= Q_t^{1,1} \ 1_{X(t^\ell)=1} 1_{X(t^r)=1} \ + \ Q_t^{1,-1} \ 1_{X(t^\ell)=1} 1_{X(t^r)=-1} \\ &+ Q_t^{-1,1} \ 1_{X(t^\ell)=-1} 1_{X(t^r)=1} \ + \ Q_t^{-1,-1} \ 1_{X(t^\ell)=-1} 1_{X(t^r)=-1} \ . \end{aligned}$$

Although $p(t)$ is not a function of $X(t^\ell)$ and $X(t^r)$, $p(t) 1_{Y \notin [S_-,S_+]}$ is the following function of $\overline{X}(t^\ell)$ and $\overline{X}(t^r)$:

$$\begin{aligned} p(t) 1_{Y \notin [S_-,S_+]} = Q_t^{1,1} \ 1_{\overline{X}(t^\ell)=1} 1_{\overline{X}(t^r)=1} \ + \ Q_t^{1,-1} \ 1_{\overline{X}(t^\ell)=1} 1_{\overline{X}(t^r)=-1} \\ + Q_t^{-1,1} \ 1_{\overline{X}(t^\ell)=-1} 1_{\overline{X}(t^r)=1} \ + \ Q_t^{-1,-1} \ 1_{\overline{X}(t^\ell)=-1} 1_{\overline{X}(t^r)=-1} \ . \end{aligned}$$

In the same way, the quantity $\mathbb{P}\left\{X(t^\ell), X(t^r)\right\}\ 1_{Y\notin[S_-,S_+]}$ present in the definition of $\overline{g}(t)$ verifies

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

As a result, as expected, the likelihood is a function of Y, $\overline{X}(t^\ell)$, $\overline{X}(t^r)$, which was not obvious at first reading. However, the expression given in formula (7.1) will be very convenient for the generalization to several QTLs.

The score statistic of the hypothesis "$q_1 = 0$" at t, for n independent observations, will be defined as

$$\overline{S}_n(t) = \frac{\frac{\partial \overline{l}_t^n}{\partial q_1}\ |_{\theta_0^1}}{\sqrt{\mathbb{V}\left(\frac{\partial \overline{l}_t^n}{\partial q_1}\ |_{\theta_0^1}\right)}}\ ,$$

where $\overline{l}_t^n(\theta^1)$ denotes the log likelihood at t, associated to n observations.

In the same way, the LRT at t, for n independent observations, will be defined as

$$\overline{\Lambda}_n(t) = 2\left\{\overline{l}_t^n(\widehat{\theta^1}) - \overline{l}_t^n(\widehat{\theta^1}_{|H_0})\right\}\ ,$$

where $\widehat{\theta^1}$ is the maximum likelihood estimator (MLE), and $\widehat{\theta^1}_{|H_0}$ the MLE under H_0.

Let us now suppose that more than one QTL (i.e. $m > 1$) lie on $[0,T]$. Using the same notations as in Sections 4 and 5, $t_1^\star$, ..., $t_m^\star$ denote the QTL locations, and the parameter θ^m and θ_0 are defined in the following way : $\theta^m = (q_1, ..., q_m, \mu, \sigma)$ and $\theta_0^m = (0, ..., 0, \mu, \sigma)$. Besides, recall that all the information is contained in the markers flanking the QTL locations. Then, the probability distribution of $\left(Y, \overline{X}(t_1^{\star\ell}), \overline{X}(t_1^{\star r}), \ldots, \overline{X}(t_m^{\star\ell}), \overline{X}(t_m^{\star r})\right)$, with respect to the measure $\lambda \otimes N \otimes \cdots \otimes N$, is

$$\begin{aligned}\overline{L}_{\vec{t}^\star}^m(\theta^m) = \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} &\left[w_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)\ 1_{Y\notin[S_-,S_+]}\right.\\ &\left.+v_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)\ 1_{Y\in[S_-,S_+]}\right]\ \overline{g}^m(t_1^\star,\ldots,t_m^\star)\end{aligned} \tag{7.3}$$

with

$$v_{\vec{t}^\star}(u_1,\ ...,\ u_m) = \mathbb{P}\left\{X(t_1^\star) = u_1, X(t_2^\star) = u_2, \ldots, X(t_m^\star) = u_m\right\}\ ,$$

$$\overline{g}^m(t_1^\star,\ldots,t_m^\star) = \mathbb{P}\left\{X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r})\right\}\ 1_{Y\notin[S_-,S_+]}\ +\ 1_{Y\in[S_-,S_+]}\,.$$

Recall also the definition of $w_{\vec{t}^\star}(u_1,\ ...,\ u_m)$ given in formula (5.2):

$$w_{\vec{t}^\star}(u_1,\ ...,\ u_m) = \mathbb{P}\left\{X(t_1^\star) = u_1, X(t_2^\star) = u_2, \ldots, X(t_m^\star) = u_m \mid X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r})\right\}\ .$$

As before, formula (7.3) is a function of Y, $\overline{X}(t_1^{\star\ell})$, , ..., $\overline{X}(t_m^{\star r})$. Note that the proof of formula (7.3) is included in the proof of the following Theorem 7.1.

Notations 7.1. *γ, γ_+ and γ_- are respectively the quantities $\mathbb{P}_{H_0}(Y \notin [S_-,\ S_+])$, $\mathbb{P}_{H_0}(Y > S_+)$ and $\mathbb{P}_{H_0}(Y < S_-)$.*

Notations 7.2. *$\mathcal{A}$ is the quantity such as $\mathcal{A} = \sigma^2 \left\{\gamma + z_{\gamma_+}\,\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\,\varphi(z_{1-\gamma_-})\right\}$, 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.*

Theorem 7.1. *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. Then,*

$$\overline{S}_n(.) \Rightarrow V(.) \quad , \quad \overline{\Lambda}_n(.) \stackrel{F.d.}{\to} V^2(.) \quad , \quad \sup \overline{\Lambda}_n(.) \stackrel{\mathcal{L}}{\longrightarrow} \sup V^2(.)$$

as n tends to infinity, under H_0 and $H_{a\vec{t}^\star}$ where $V(.)$ is the Gaussian process with unit variance such as

$$V(t) = \frac{\alpha(t)\ V(t^\ell) + \beta(t)\ V(t^r)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t^\ell, t^r)}} \quad ,$$
$$Cov\left\{V(t^\ell), V(t^r)\right\} = \rho(t^\ell, t^r) = e^{-2|t^\ell - t^r|}$$

where the functions $\alpha(.)$ and $\beta(.)$ are given in Theorem 4.1, and with mean function

- *under H_0, $m(t) = 0$*
- *under $H_{a\vec{t}^\star}$,*

$$\overline{m}_{\vec{t}^\star}(t) = \frac{\alpha(t)\ \overline{m}_{\vec{t}^\star}(t^\ell) + \beta(t)\ \overline{m}_{\vec{t}^\star}(t^r)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t^\ell, t^r)}}$$

where

$$\overline{m}_{\vec{t}^\star}(t^\ell) = \sum_{s=1}^{m} a_s \sqrt{\mathcal{A}}\ \rho(t^\ell, t_s^\star) \ / \ \sigma^2 \quad , \quad \overline{m}_{\vec{t}^\star}(t^r) = \sum_{s=1}^{m} a_s \sqrt{\mathcal{A}}\ \rho(t^r, t_s^\star) \ / \ \sigma^2 \quad .$$

The proof is given in Section 11.3. Under the null hypothesis, despite the selective genotyping, $V(.)$ is exactly the same process as the process $Z(.)$ of Theorem 5.1 obtained for the complete data situation. However, under the general alternative, the mean functions of the two processes are not the same anymore : the mean functions are proportional of a factor $\sqrt{\mathcal{A}}/\sigma$.

Before introducing our Lemma 7.1, let us recall that the Asymptotic Relative Efficiency (ARE) determines the relative sample size required to obtain the same local asymptotic power as the one of the test under the complete data situation where the genome information at markers is known for all the individuals.

Lemma 7.1. *Let κ denote the ARE, then we have*

i) $\kappa = \gamma + z_{\gamma_+}\,\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\,\varphi(z_{1-\gamma_-})$
ii) κ reaches its maximum for $\gamma_+ = \gamma_- = \gamma/2$.

This lemma is a generalization of Theorem 4.2 of Rabier [32] where the focus was only on the case $m = 1$. To prove Lemma 7.1, just use the same proof as the one of Theorem 4.2 of Rabier [32].

According to i) of Lemma 7.1, the ARE with respect to the complete data situation, does not depend on the number of QTLs m, the constants a_1, ..., a_m linked to the QTL effects, and the QTLs locations $t_1^\star$, ..., $t_m^\star$. Indeed, since the mean functions (complete data situation and selective genotyping) are proportional of a factor $\sqrt{\mathcal{A}}/\sigma$, it is obvious that the ARE does not depend on those parameters. On the other hand, according to ii) of Lemma 7.1, if we want to genotype only a percentage γ of the population, we should genotype the $\gamma/2\%$ individuals with the largest phenotypes and $\gamma/2\%$ individuals with the smallest phenotypes.

Let us consider now $n^\star$ individuals for a selective genotyping experiment, and let us assume that we have the relationship $n = n^\star\gamma$. In other words, we focus on the case where, for economical reasons, we are allowed to genotype only n individuals. By considering $n = n^\star\gamma$, we are allowed to genotype n extreme individuals, provided that the overall population size has been increased to $n^\star$. In this context, we have

$$\overline{S}_{n^\star}(t_k) \overset{H_0}{\to} N\left(\sqrt{\frac{\mathcal{A}}{\gamma\,\sigma^4}}\ \sum_{s=1}^{m} a_s\ \rho(t_k, t_s^\star),\ 1\right)$$

and the mean function of the process is still interpolated. As a result, the ratio between the signal corresponding to selective genotyping and the one matching the complete data situation is equal to $\sqrt{\frac{\mathcal{A}}{\gamma\,\sigma^2}}$. This quantity verifies the following relationship

$$\sqrt{\frac{\mathcal{A}}{\gamma\,\sigma^2}} = \sqrt{z_{\gamma_+}\ \varphi(z_{\gamma_+})/\gamma\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})/\gamma\ +\ 1}$$

and if we are willing to genotype symmetrically (i.e. $\gamma_+ = \gamma_-$), it becomes

$$\sqrt{\frac{\mathcal{A}}{\gamma\,\sigma^2}} = \sqrt{2z_{\gamma/2}\varphi(z_{\gamma/2})/\gamma\ +\ 1}\ .$$

In other words, provided that the phenotyping is free, the signal can be largely increased, by genotyping extreme individuals (i.e. selective genotyping) instead of genotyping random individuals (i.e. complete data situation). According to Figure 1, when the selective genotyping is performed symmetrically, the signal corresponding respectively to the cases $\gamma = 0.1$, $\gamma = 0.2$ and $\gamma = 0.3$, is respectively 2.09, 1.80 and 1.61 times larger under selective genotyping than under random genotyping. The worst case is obtained when genotyping only the largest phenotypes (see $\gamma_+/\gamma = 1$) or genotyping only the smallest phenotypes (same curve as the one for $\gamma_+/\gamma = 1$). In that case, the selective genotyping can be less rewarding than the random genotyping (cf. $\gamma_+ = \gamma = 0.8$).

Obviously, when all the individuals are genotyped ($\gamma = 1$), all the efficiencies are equal to one.

Let us now move on to the case where interactions are present into our model (see formula 2.1). Then, under selective genotyping, we have the following result:

Theorem 7.2. *Suppose that the parameters* $(q_1, ..., q_m, q_{1,2}, ..., q_{m-1,m}, \mu, \sigma^2)$ *vary in a compact and that* σ^2 *is bounded away from zero, and also that* m *is finite. Then, with the previous notations, under* $H_{a\vec{t}^\star, b\vec{t}^\star}$ *,*

$$\overline{S}_n(.) \Rightarrow V(.) \quad , \quad \overline{\Lambda}_n(.) \stackrel{F.d.}{\rightarrow} V^2(.) \quad , \quad \sup \overline{\Lambda}_n(.) \stackrel{\mathcal{L}}{\longrightarrow} \sup V^2(.)$$

where $V(.)$ *is the Gaussian process of Theorem 7.1 uncentered with mean function* $m_{t^\star}(.)$ *defined in Theorem 7.1.*

The proof is given in Section 11.4. As under the complete data situation (Theorem 6.1), the interaction effects are not included in the mean function.

Sometimes, for some biological reasons, we are only able to genotype the non extreme individuals (i.e. the individuals for which $Y \in [S_-, S_+]$). In this context, we present the following lemma.

Lemma 7.2. *Under the reverse configuration, that is to say if* $\overline{X}(t_k) = X(t_k)\ 1_{Y \in [S_-\ ,\ S_+]}$*, then we have the same results as in Theorem 7.1 and Theorem 7.2 provided that we replace the quantity* $\mathcal{A}$ *by the quantity* $\mathcal{B}$ *defined in the following way*

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

The proof is largely inspired of the proof of Theorems 7.1 and 7.2, and also from Rabier [33] where this configuration is studied under the local alternative of one QTL at $t^\star$ on $[0, T]$.

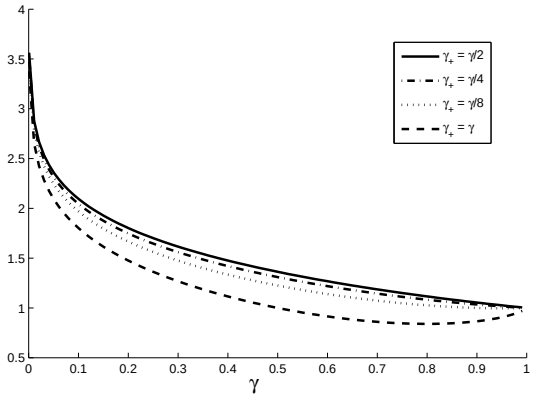


FIG 1. *Function* $\sqrt{z_{\gamma_+}\ \varphi(z_{\gamma_+})/\gamma \ - \ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})/\gamma \ + \ 1}$ *as a function of the percentage* γ *of individuals genotyped and as a function of the ratio* γ_+/γ.

8. Interference phenomenon

In order to handle the interference phenomenon, we will focus on the model introduced by Rebaï et al. [36] (see in particular their Section 2) in which double recombination between the QTL and its flanking markers is not allowed. Let us consider the case $m = 1$. Then, under the model considered by Rebaï et al. [36], if the QTL is lying for instance between the first two markers (i.e. $t_1^\star \in]t_1, t_2[$), we can not have the scenario $X(t_1) = 1, U(t_1^\star) = -1$ and $X(t_2) = 1$. Indeed, this would have supposed that there had been a recombination between the first marker and the QTL, and also a recombination between the second marker and the QTL. In particular, the model considers that if we have a recombination between the QTL and one of its flanking marker, we could not have a recombination between the QTL and the other flanking marker. In other words, if $X(t_1) = 1$ and $U(t_1^\star) = -1$, then we have automatically $X(t_2) = -1$. In the same way, if $X(t_2) = 1$ and $U(t_1^\star) = -1$, then we have automatically $X(t_1) = -1$. Note that in Rebaï et al. [35], the authors extended their previous model to several markers, keeping Haldane [18] modeling for the genetic information at marker locations. In other words, as previously, $X(0)$ is a random sign and $X(t_k) = X(0)(-1)^{N(t_k)}$ where $N(.)$ is a standard Poisson process on $[0, T]$.

In this section, we will first study the classical model (see Sections 4 and 5) under interference, and later we will consider the epistatic model. Note that the results can easily be generalized to selective genotyping experiments.

To begin with, let us consider the case $m = 1$ (i.e. one QTL on $[0, T]$). According to Rabier [31], the likelihood of the triplet $\left(Y,\ X(t^\ell),\ X(t^r)\right)$ with respect to the measure $\lambda \otimes N \otimes N$, λ being the Lebesgue measure, N the counting measure on $\mathbb{N}$, is $\forall t \in]t^\ell, t^r[$:

$$\widetilde{L}_t(\theta^1) = \left[\widetilde{p}(t) f_{(\mu+q_1,\sigma)}(Y) + \{1 - \widetilde{p}(t)\} f_{(\mu-q_1,\sigma)}(Y)\right]\ g(t) \tag{8.1}$$

where the function

$$\widetilde{p}(t) = \mathbb{P}\left\{U(t) = 1 \mid X(t^\ell), X(t^r)\right\}\ \ ,\ \ g(t) = \mathbb{P}\left\{X(t^\ell), X(t^r)\right\}\ .$$

In particular,

$$\widetilde{p}(t) = 1_{X(t^\ell)=1} 1_{X(t^r)=1}\ +\ \frac{t^r - t}{t^r - t^\ell}\ 1_{X(t^\ell)=1} 1_{X(t^r)=-1}\ +\ \frac{t - t^\ell}{t^r - t^\ell}\ 1_{X(t^\ell)=-1} 1_{X(t^r)=1}\ . \tag{8.2}$$

As previously, the score statistic of the hypothesis "$q_1 = 0$" at t, for n independent observations, will be defined as

$$\widetilde{S}_n(t) = \frac{\frac{\partial \widetilde{l}_t^n}{\partial q_1} \mid_{\theta_0^1}}{\sqrt{\mathbb{V}\left(\frac{\partial \widetilde{l}_t^n}{\partial q_1} \mid_{\theta_0^1}\right)}}\ ,$$

where $\widetilde{l}_t^n(\theta^1)$ denotes the log likelihood at t, associated to n observations.

In the same way, the LRT at t, for n independent observations, will be defined as

$$\widetilde{\Lambda}_n(t) = 2\left\{\widetilde{l}_t^n(\widehat{\theta^1}) - \widetilde{l}_t^n(\widehat{\theta^1}_{|H_0})\right\} .$$

Let us now move on to the case $m > 1$.

Recall that we impose that the QTLs do not belong to the same marker intervals and that we consider Haldane modeling for the genome information at genetic markers. As a result, since all the information is contained in the markers flanking the QTL locations, we have the relationship

$$\mathbb{P}\left\{U(t_1^\star), \ldots, U(t_m^\star) \mid X(t_1), \ldots, X(t_K)\right\} = \mathbb{P}\left\{U(t_1^\star), \ldots, U(t_m^\star) \mid X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r})\right\} .$$

The full likelihood of $(Y, X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r}))$ is

$$\widetilde{L}_{\vec{t}^\star}^m(\theta^m) = \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \widetilde{w}_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)\ g^m(t_1^\star, \ldots, t_m^\star) \tag{8.3}$$

where

$$\widetilde{w}_{\vec{t}^\star}(u_1,\ ...,\ u_m) = \mathbb{P}\left\{U(t_1^\star) = u_1, ..., U(t_m^\star) = u_m \;\middle|\; X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r})\right\},$$
$$g^m(t_1^\star, \ldots, t_m^\star) = \mathbb{P}\left\{X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r})\right\} .$$

Theorem 8.1. *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. Then,*

$$\widetilde{S}_n(.) \Rightarrow W(.) \ \ , \ \ \widetilde{\Lambda}_n(.) \overset{F.d.}{\to} W^2(.) \quad , \quad \sup\widetilde{\Lambda}_n(.) \xrightarrow{\mathcal{L}} \sup W^2(.)$$

as n tends to infinity, under H_0 and $H_{a\vec{t}^\star}$ where $W(.)$ is the Gaussian process with unit variance such as

$$W(t) = \frac{\widetilde{\alpha}(t)\ W(t^\ell)\ +\ \widetilde{\beta}(t)\ W(t^r)}{\sqrt{\widetilde{\alpha}^2(t) + \widetilde{\beta}^2(t)\ +\ 2\widetilde{\alpha}(t)\widetilde{\beta}(t)\rho(t^\ell, t^r)}} \ ,$$
$$Cov\left\{W(t^\ell), W(t^r)\right\} = \rho(t^\ell, t^r) = e^{-2|t^\ell - t^r|}$$

where $\widetilde{\alpha}(t) = \frac{t^r - t}{t^r - t^\ell}$, $\widetilde{\beta}(t) = \frac{t - t^\ell}{t^r - t^\ell}$ and with mean function

- *under H_0, $m(t) = 0$*
- *under $H_{a\vec{t}^\star}$,*

$$\widetilde{m}_{\vec{t}^\star}(t) = \frac{\widetilde{\alpha}(t)\ \widetilde{m}_{\vec{t}^\star}(t^\ell) + \widetilde{\beta}(t)\ \widetilde{m}_{\vec{t}^\star}(t^r)}{\sqrt{\widetilde{\alpha}^2(t) + \widetilde{\beta}^2(t)\ +\ 2\widetilde{\alpha}(t)\widetilde{\beta}(t)\rho(t^\ell, t^r)}}$$

where

$$\widetilde{m}_{\vec{t}^\star}(t^\ell) = \sum_{s=1}^{m} a_s\ h(t^\ell, t_s^\star)\ /\ \sigma \quad , \quad \widetilde{m}_{\vec{t}^\star}(t^r) = \sum_{s=1}^{m} a_s\ h(t^r, t_s^\star)\ /\ \sigma \quad ,$$

$$\forall k \quad h(t_k, t_s^\star) = \rho(t_k, t_s^{\star\ell}) \left\{ \widetilde{\alpha}(t_s^\star) + \widetilde{\beta}(t_s^\star) \rho(t_s^{\star\ell}, t_s^{\star r}) \right\} 1_{t_s^\star > t_k}$$
$$+ \rho(t_k, t_s^{\star r}) \left\{ \widetilde{\alpha}(t_s^\star) \rho(t_s^{\star\ell}, t_s^{\star r}) + \widetilde{\beta}(t_s^\star) \right\} 1_{t_s^\star < t_k} \ .$$

The proof is given in Section 11.5. Note that the functions $\widetilde{\alpha}(t)$ and $\widetilde{\beta}(t)$ are different from the ones of Theorems 4.1 and 5.1. As in Rabier [31], the limiting process is the square of a linear interpolated process. As expected, the mean function depends now on the number of QTLs, their positions and their effects.

Theorem 8.2. *Suppose that the parameters* $(q_1, ..., q_m, q_{1,2}, ..., q_{m-1,m}, \mu, \sigma^2)$ *vary in a compact and that* σ^2 *is bounded away from zero, and also that* m *is finite. Then, with the previous notations, under* $H_{a\vec{t}^\star, b\vec{t}^\star}$ *,*

$$\widetilde{S}_n(.) \Rightarrow W(.) \ , \ \widetilde{\Lambda}_n(.) \overset{F.d.}{\to} W^2(.) \quad , \quad \sup \widetilde{\Lambda}_n(.) \xrightarrow{\mathcal{L}} \sup W^2(.)$$

where $W(.)$ *is the Gaussian process of Theorem 8.1 uncentered with mean function* $\widetilde{m}_{\vec{t}^\star}(.)$ *defined in Theorem 8.1.*

In the same way as before, the interaction effects are not present in the mean function.

9. A new method for gene mapping

In this section, the goal is to propose a method to estimate the number of QTLs, their effects and their positions combining results of the theorem and a penalized likelihood method. Note that in order to make the reading easier, we will introduce the method in the context of Sections 4 and 5 (Haldane mapping, no selective genotyping and no epistasis). However, it can easily be adapted to the different models studied in this paper.

9.1. Introducing our method

According to Theorem 5.1, 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 $\Sigma_{kk'} = \rho(t_k, t_{k'})$.
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. Recall that $S_n(.)$ is an "interpolated process". Let us consider the Cholesky decomposition $\Sigma = AA'$. We have

$$A^{-1}\vec{S}_n = A^{-1}B\left(\frac{a_1}{\sigma}\ ,\ \ldots\ ,\ \frac{a_m}{\sigma}\right)'\ +\ A^{-1}\vec{\varepsilon}\ +\ o_P(1)$$

where B is a matrix of size $K \times m$ such as $B_{ks} = e^{-2|t_k - t_s^\star|}$.
Since the number m of QTLs and their positions $t_1^\star$,...,$t_m^\star$ are unknown, we propose to focus on a new discretization of $[0, T]$ corresponding to all the putative QTL locations: $0 \leq t'_1 < t'_2 < \ldots < t'_L \leq T$. $\Delta_1, \ldots, \Delta_L$ will be the corresponding effects divided by σ. Note that although we focus only on the discretized process at markers locations, we look for QTL not only on markers. The model can be rewritten in the following way:

$$A^{-1}\vec{S}_n = A^{-1}C\,(\Delta_1\ ,\ \ldots\ ,\ \Delta_L)'\ +\ A^{-1}\vec{\varepsilon}\ +\ o_P(1) \tag{9.1}$$

where C is a matrix of size $K \times L$ such as $C_{kl} = e^{-2|t_k - t'_l|}$.
Last, in order to find the non zero Δ_l, a natural approach is to use the L1 penalized regression, so-called LASSO (Tibshirani [39]):

$$\arg \min_{(\Delta_1,\ldots,\Delta_L)} \left\| A^{-1}\vec{S}_n - A^{-1}C\Delta \right\|_2^2 \quad + \zeta \left\|\Delta\right\|_1$$

where $\|\ \|_2$ is the L2 norm, $\|\ \|_1$ is the L1 norm, $\Delta = (\Delta_1, \ldots, \Delta_L)'$ and ζ denotes the tuning parameter. ζ will be estimated using cross validation as described in Chapter 7 of [19].

10. Illustrations

10.1. About the max test

To begin with, let us briefly illustrate our theoretical results regarding the max test. Recall that it relies on the test statistic, $\sup \Lambda_n(.)$. The focus is on a sparse map: a chromosome of length 1M ($T = 1$), with 21 markers ($K = 21$) equally spaced every 5cM. In this context, Table 1 compares the theoretical power and the empirical power, under different configurations : either 1 QTL ($m = 1$) at 3cM, either 2 QTLs ($m = 2$) at 3cM and 28cM, or 3 QTLs ($m = 3$) at 3cM, 28cM and 72cM. 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 theoretical power was obtained by generating 10,000 paths of the asymptotic process, whereas 1,000 samples of size n equal to 1,000 , 200 or 100 were considered for the empirical power. The threshold (i.e. critical value) at the 5% level was set to 7.84 using the Monte-Carlo Quasi Monte-Carlo method, proposed by Azaïs et al. [2] and based on Genz [17]. In order to compute the maximum of the process, simulated data were analyzed using Lemma 1 of Azaïs et al. [2], that is to say performing LRT on markers

γ	n \ m	1 (+)	2 (++)	2 (+-)	3 (+-+)
1	$+\infty$	60.20%	99.35%	15.27%	49.74%
	1,000	59.7%	98.90%	15.70%	49.00%
	200	60.00%	98.80%	15.50%	47.30%
	100	53.90%	98.50%	13.70%	45.80%
0.3	$+\infty$	48.21%	97.47%	12.71%	39.36%
	1,000	47.90%	97.10%	12.20%	39.50%
	200	47.70%	96.80%	10.50%	37.50%
	100	46.10%	96.50%	9.40%	32.80%

TABLE 1

Theoretical power and empirical power associated to the test statistic $\sup \Lambda_n(.)$*, and as a function of the number* m *of QTLs and the percentage* γ *of genotyped individuals (* $T = 1$*,* $K = 21$*,* $t_k = 0.05(k-1)$*, (*$m = 1$*,* $t_1^\star = 0.03$*), (*$m = 2$*,* $t_1^\star = 0.03$*,* $t_2^\star = 0.80$*), (*$m = 3$*,* $t_1^\star = 0.03$*,* $t_2^\star = 0.28$*,* $t_3^\star = 0.72$*), all* $|a_s| = 2.828$*, + for positive effect, − for negative effect, 10,000 paths for the theoretical power, 1,000 samples of size* n *for the empirical power,* $\gamma_+/\gamma = 1/2$*).*

and performing only one test in each marker interval if the ratio of the score statistics on markers fulfills the given condition.

According to Table 1, we can notice a good agreement between the empirical power and the theoretical power for $n = 200$. However, the asymptotic seems to be really reached for n =1,000. We also investigated the behavior of the test under a selective genotyping performed symmetrically (i.e. $\gamma_+ = \gamma/2$). Recall that the threshold remains the same under selective genotyping (cf. Theorem 7.1). We can observe that when $\gamma = 0.3$, the empirical power still matches the theoretical power for n =1,000. This validates our theoretical results presented in Theorems 5.1 and 7.1.

Last, the power of the test is reported as a function of the QTL effect signs. We can see that when the two QTLs at 3cM and 28cM have the same signs, the power is almost equal to 1 whereas it largely decreases ($\approx 15\%$ for $\gamma = 1$) when the signs are opposite. In this case, the max test is clearly not the most appropriate test to perform. We refer to the recent study of [6] where the authors compared performances of the max test and the ANOVA in another context.

10.2. Selective genotyping improves the detection process

We propose to investigate here the performances of our gene mapping method (see Section 9). Figure 2, based on one simulated data set, illustrates the performances of the method under selective genotyping. The considered genome is of length 10M ($T = 10$), with 201 markers ($K = 201$) equally spaced every 5cM. 16 QTLs ($m = 16$) lie on the interval [0M,4M] whereas no QTL are present on the rest of the genome (i.e. [6M,10M]). The QTL effects are equal to either +0.2 or −0.2, each QTL having its own random sign. The presence of QTL is tracked every 2.5cM. As a consequence, 401 regressors ($L = 401$) are present in the linear model (formula 9.1). In other words, we use the discretization $t'_l = 0.025(l-1)$, $l = 1, \ldots, 401$. Recall that this grid is different from the one corresponding to

marker locations: $t_k = 0.05(k-1)$, $k = 1, \ldots, 201$. Figure 2A refers to the case $n = 200$ whereas Figure 2B focuses on $n = 100$.

Assuming that, for economical reasons, the geneticist is allowed to genotype only n individuals, we compare here the case where those n individuals are extreme or not. In particular, when a selective genotyping was performed, the total number of individuals was increased to $n^\star$, with the relationship $n^\star = n/\gamma$. This way, on average, n individuals are genotyped under selective genotyping (cf. Section 7). Then, when the sample size n was equal to 100 under complete genotyping ($\gamma = 1$), we considered respectively 1000, 500 and 333 individuals to handle the cases $\gamma = 0.1$, 0.2, and 0.3 respectively. In the same way, when $n = 200$, we considered either 2000, either 1000, or 666 individuals. According to Figure 2A, the largest estimated effects are the ones corresponding to the case $\gamma = 0.1$: a few QTL effects are estimated at approximately 5 (see around 1M and 4M), and at -6 around 2M. It was expected since under such selective genotyping (i.e with $n^\star = n/\gamma$), the quantities Δ_l, present in formula (9.1), are increased by a factor $\sqrt{\mathcal{A}}/\sqrt{\gamma}$ at each gene location. Then, under the configuration studied, the quantities $|a|\sqrt{\mathcal{A}}/\sqrt{\gamma}$ are equal respectively to 5.92, 4.56 and 2.50 when γ takes respectively the values 0.1, 0.3, and 1. Note that the number of selected regressors was between 15 and 17 in all studied cases.

In what follows, the L1 ratio will denote the ratio L1 norm of estimated effects on [0M,4M] to L1 norm of estimated effects on [0M,10M]. This L1 ratio is an indicator of whether or not the detected QTLs belong to the "signal area". Recall that on this example, all the simulated QTLs belong to the interval [0M,4M]. Then, the L1 ratio is the quantity $\sum_{i=1}^{161} |\hat{\Delta}_i| / \sum_{i=1}^{401} |\hat{\Delta}_i|$. The one associated to the case $n = 200$ took the values 98.47% for $\gamma = 0.1$, 90.79% for $\gamma = 0.2$, and 76.44% for $\gamma = 1$. On the other hand, the L1 ratio corresponding to $n = 100$ was found equal to 99.55% for $\gamma = 0.1$, to 95.25% for $\gamma = 0.2$, and to 61.10% for $\gamma = 1$. In other words, by considering extreme individuals, we largely improve the detection process.

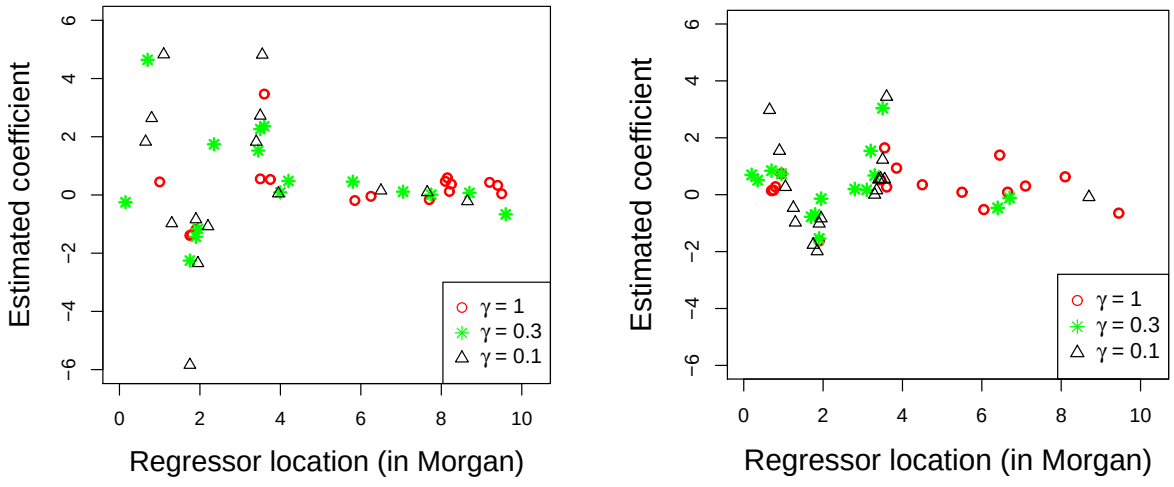
To confirm this finding based on one single data set, Tables 2 and 3 report in a more general framework, the mean L1 ratio over 100 samples of size $n = 200$ and $n = 100$ respectively. Different QTL effects are taken into consideration : $|q_s|$ is either equal to 0.2, 0.1, or 0.05. We can notice that whatever the parameter values, the more extremes the genotyped individuals are, the larger the L1 ratio is. Another interesting aspect (not shown here) is in the choice of the tuning parameter value. The MSE curve obtained by cross validation is flat under complete genotyping ($\gamma = 1$), suggesting the absence of signal. In contrast, under selective genotyping, we can clearly distinguished the minimum of the curve, due to to the increase of signal.

Last, Table 4 focuses on different ways of performing the selective genotyping: different ratios γ_+/γ are investigated. As expected, when only the largest (or the smallest) individuals are genotyped ($\gamma_+/\gamma = 1$), the L1 ratio is the smallest. For instance, we can see that for $\gamma = 0.1$, the L1 ratio is equal to 68.87% when the selective genotyping is performed unilaterally, whereas it increases to 82.86% when the selective genotyping is performed symmetrically. It confirms

our theoretical results presented in Section 7 and illustrated in Figure 1.

To conclude, selective genotyping is largely more rewarding for localizing genes.

A) B)



$n = 200$ $n = 100$

FIG 2. *Estimated coefficients according to our method as a function of the percentage γ of genotyped individuals (1 sample, $m = 16$, $T = 10$, $|q_1| = \ldots = |q_{16}| = 0.2$, QTLs randomly located only on [0M,4M], $K = 201$, $t_k = 0.05(k-1)$, $L = 401$, $t'_l = 0.025(k-1)$, $\gamma_+/\gamma = 1/2$, on average n individuals genotyped).*

10.3. The mixture model allows to detect genes lying between markers

Let us illustrate here that our method allows to detect genes lying between markers. In order to present the performances of the method in a general framework, the focus is on a configuration without selective genotyping. The genome is of length 10M ($T = 10$), with 101 markers ($K = 101$) equally spaced every 10cM. The presence of QTL is tracked every 2.5cM, so 401 regressors ($L = 401$) are present in our linear model. This genetic map is slightly sparser than the previous one in order to deal with more regressors located between markers.

10 QTLs ($m = 10$) are equally spaced every 50cM on the interval [5cM, 455cM], and we consider random QTL effects signs accross simulations. Note that these QTL locations have been chosen between markers. In this context, Table 5 compares the performances of the method as a function of the intensity λ of the Poisson process, the absolute value $|a|$ of the constant linked to the

all $\lvert q_s\rvert$	γ	$\sum_{i=1}^{161}\lvert\hat{\Delta}_i\rvert/\sum_{i=1}^{401}\lvert\hat{\Delta}_i\rvert$	$\sum_{i=162}^{401}\lvert\hat{\Delta}_i\rvert/\sum_{i=1}^{401}\lvert\hat{\Delta}_i\rvert$	$\hat{m}$
0.2	0.1	99.61%	0.39%	15.54
	0.2	97.99%	2.01%	15.3
	0.3	95.84%	4.16%	17.22
	1	82.57%	17.43%	16.94
0.1	0.1	89.79%	10.21%	16.11
	0.2	84.00%	16.00%	18.47
	0.3	83.10%	16.90%	16.83
	1	62.35%	37.65%	17.62
0.05	0.1	68.66%	31.34%	15.17
	0.2	63.70%	36.30%	15.86
	0.3	68.24%	31.76%	16.5
	1	46.49%	53.51%	18.07

TABLE 2

Performances of the new method as a function of the percentage γ of genotyped individuals and as a function of the QTL effects (Mean over 100 samples, $m = 16$, $T = 10$, QTLs randomly located only on [0M,4M], $K = 201$, $t_k = 0.05(k-1)$, $L = 401$, $t'_l = 0.025(k-1)$, $\gamma_+/\gamma = 1/2$, on average $n = 200$ individuals genotyped). The L1 ratio corresponds to the quantity $\sum_{i=1}^{161}\lvert\hat{\Delta}_i\rvert/\sum_{i=1}^{401}\lvert\hat{\Delta}_i\rvert$, and $\hat{m}$ denotes the estimated QTL number.

all $\lvert q_s\rvert$	γ	$\sum_{i=1}^{161}\lvert\hat{\Delta}_i\rvert/\sum_{i=1}^{401}\lvert\hat{\Delta}_i\rvert$	$\sum_{i=162}^{401}\lvert\hat{\Delta}_i\rvert/\sum_{i=1}^{401}\lvert\hat{\Delta}_i\rvert$	$\hat{m}$
0.2	0.1	96.83%	3.17%	14.75
	0.2	90.32%	9.68%	18.17
	0.3	88.03%	11.97%	17.45
	1	70.91%	29.09%	18.47
0.1	0.1	82.26%	17.74%	14.74
	0.2	73.43%	26.57%	15.64
	0.3	70.95%	29.05%	16.59
	1	55.41%	44.59%	18.57
0.05	0.1	61.00%	39.00%	15.06
	0.2	52.73%	47.27%	15.07
	0.3	52.27%	47.73%	15.38
	1	45.34%	54.66%	15.64

TABLE 3

Performances of the new method as a function of the percentage γ of genotyped individuals and as a function of the QTL effects (Mean over 100 samples, $m = 16$, $T = 10$, QTLs randomly located only on [0M,4M], $K = 201$, $t_k = 0.05(k-1)$, $L = 401$, $t'_l = 0.025(k-1)$, $\gamma_+/\gamma = 1/2$, on average $n = 100$ individuals genotyped). The L1 ratio corresponds to the quantity $\sum_{i=1}^{161}\lvert\hat{\Delta}_i\rvert/\sum_{i=1}^{401}\lvert\hat{\Delta}_i\rvert$, and $\hat{m}$ denotes the estimated QTL number.

γ	γ_+/γ	$\sum_{i=1}^{161} \lvert\hat{\Delta}_i\rvert / \sum_{i=1}^{401} \lvert\hat{\Delta}_i\rvert$	$\sum_{i=162}^{401} \lvert\hat{\Delta}_i\rvert / \sum_{i=1}^{401} \lvert\hat{\Delta}_i\rvert$	$\hat{m}$
0.1	1/2	82.86%	17.74%	14.74
	3/4	79.17%	20.83%	15.35
	7/8	74.61%	25.39%	15.89
	1	68.87%	31.13%	16.26
0.2	1/2	73.43%	26.57%	15.64
	3/4	71.27%	28.73%	16.36
	7/8	68.19%	31.81%	17.15
	1	63.80%	36.20%	16.95
0.3	1/2	70.95%	29.05%	16.59
	3/4	68.84%	31.16%	15.39
	7/8	65.36%	34.63%	15.75
	1	61.76%	36.24%	16.63

TABLE 4

Performances of the new method as a function of the ratio γ_+/γ (Mean over 100 samples, $m = 16$, $\lvert q_1\rvert = \ldots = \lvert q_{16}\rvert = 0.1$, $T = 10$, QTLs randomly located only on [0M,4M], $K = 201$, $t_k = 0.05(k-1)$, $L = 401$, $t'_l = 0.025(k-1)$, $\gamma_+/\gamma = 1/2$, on average $n = 100$ individuals genotyped). The L1 ratio corresponds to the quantity $\sum_{i=1}^{161} \lvert\hat{\Delta}_i\rvert / \sum_{i=1}^{401} \lvert\hat{\Delta}_i\rvert$, and $\hat{m}$ denotes the estimated QTL number.

QTL effects, and whether or not some noise is present in the model. Recall that under Haldane mapping, recombination is modeled according to a standard Poisson process on $[0, T]$ (i.e. $\lambda = 1$). However, according to many biological studies (e.g [40]), a higher recombination rate is observed in some areas, so called hotspots (mostly located at the ends of the chromosomes). As a result, we studied different recombination rates accross the genome. So, let us assume now that $N(.)$ is a Poisson process of intensity λ on $[0, T]$. Recall that $r(t, t')$ denotes the probability of recombination between two loci located at t and t'. Then, we have the relationships

$$r(t,t') = \mathbb{P}(X(t)X(t') = -1) = \mathbb{P}(|N(t) - N(t')| \text{ odd}) \quad = \sum_{k=0}^{\infty} \frac{(\lambda\,|t-t'|)^{2k+1}}{(2k+1)!}$$

$$= e^{-\lambda|t-t'|} sh(\lambda\,|t-t'|) = \frac{1}{2}\left(1 - e^{-2\lambda|t-t'|}\right) \ .$$

Since $1 - 2r(t, t') = e^{-2\lambda|t-t'|}$, all the quantities $\rho(t, t')$ present in the covariance and in the mean functions of our different theorems, become equal to $e^{-2\lambda|t-t'|}$. In other words, the parameter λ acts as a multiplying factor for the distance between loci.

To begin with, in order to check if our method is able to recover the genes, let us investigate the L1 penalization in the ideal noiseless situation (see Donoho [14]). Recall that, even with noiseless data, we have to deal with K equations and L unknown parameters ($L > K$), which makes difficult the resolution of such system of equations. However, Donoho [14] has shown that the solution given by L1 penalization is a good approximation for recovering the sparse unknown vector.

According to Table 5, when λ takes respectively the values 5 and 10, the number of truly detected genes is equal to 9.93 and 9.84 respectively, without

λ	Method	$\lvert a\rvert$	True positives	False positives	FDR
1	Noiseless	-	8.36	6.94	9.83%
	Noisy	20	8.40	12.55	35.44%
		12	7.55	16.34	46.07%
		8	5.74	18.35	55.39%
		4	1.93	11.62	84.61%
5	Noiseless	-	9.93	0	0%
	Noisy	20	9.87	3.68	2.64%
		12	9.52	6.42	10.37%
		8	8.17	7.69	17.01%
		4	0.95	1.71	34.74%
10	Noiseless	-	9.84	0	0%
	Noisy	20	8.07	1.33	0.81%
		12	4.68	1.78	4.03%
		8	1.41	0.99	14.25%
		4	0.03	0.03	48.45%

TABLE 5

Performances of the method as a function of the intensity λ of the Poisson process, the absolute value $|a|$ of the constant linked to the QTL effects, and whether or not some noise is present in the model ($T = 10$, $m = 10$, $t_s^\star = 0.5(s-1) + 0.05$, random QTL effect signs, $K = 101$, $t_k = 0.10(k-1)$, $L = 401$, $t'_l = 0.025(k-1)$, 1,000 paths of the asymptotic processes, FDR=False Discovery Rate within 5cM)

any false positives in both cases. In other words, our method is able to find the genes perfectly in this noiseless setting. Note that when the recombination rate decreases ($\lambda = 1$), the number of true positives drops to 8.36 and some false positives are even detected (6.94). 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 [7]). 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. Coming back to our example ($\lambda = 1$), we can notice that the FDR is fair since it is maintained slightly below 10%.

On the other hand, in the noisy setting, we can see that the best configuration seems to match the case λ equal to 5, corresponding to a moderate recombination rate between regressors. Indeed, the high correlation ($\lambda = 1$) makes the problem more difficult, whereas when the correlation between regressor is too low ($\lambda = 10$), the signal captured by the test statistics on markers is too small. Recall that signal depends on the correlation between markers and QTLs. Last, Table 6 shows that, as expected, for a given value of the absolute value $|q|$ of the QTL effects, the detection power increases with the number of individuals n.

11. Proofs

11.1. Proof of Theorem 4.1

The proof is divided into four parts:

- Preliminaries (i.e. computation of the Fisher Information Matrix)

λ	Method	n	$\|q\|$	True positives	False positives	FDR
1	Noiseless	-	-	8.36	6.94	9.83%
	Noisy	5,000	0.4	8.32	11.95	34.21%
			0.2	7.36	15.52	45.13%
		1,000	0.4	5.66	17.54	55.15%
			0.2	3.57	16.74	64.09%
5	Noiseless	-	-	9.93	0	0%
	Noisy	5,000	0.4	9.9	4.37	3.31%
			0.2	9.42	6.2	9.26%
		1,000	0.4	7.91	8.39	20.62%
			0.2	3.4	4.97	25.77%
10	Noiseless	-		9.84	0	0%
	Noisy	5,000	0.4	8.22	1.62	2.05%
			0.2	4.91	1.44	2.86%
		1,000	0.4	1.93	1.24	16.25%
			0.2	0.2	0.21	20%

TABLE 6

Performances of the method as a function of the intensity λ of the Poisson process, the absolute value $|q|$ of the QTL effects, the number of individuals n, and whether or not some noise is present in the model ($T = 10$, $m = 10$, $t_s^\star = 0.5(s-1) + 0.05$, random QTL effect signs, $K = 101$, $t_k = 0.10(k-1)$, $L = 401$, $t'_l = 0.025(k-1)$, 1,000 paths of the asymptotic processes, FDR=False Discovery Rate within 5cM)

- Weak convergence of the score process under H_0
- Study of the score process under the local alternative $H_{a\vec{t}^\star}$
- Study of the supremum of the LRT process.

Note that under H_0, the proof has already been given in Azaïs et al. [2]. However, the weak convergence of the score process has not been proved in details. Indeed, the authors only mentioned the continuous mapping theorem, after having proved the convergence of finite-dimensional. As a consequence, we propose to give here a more rigorous proof by showing the tightness of the score process. 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 [5]).

In what follows, we will consider values t, $t_1^\star$, ..., $t_m^\star$ of the parameters that are distinct of the markers positions (i.e. t_1 and t_2), and the result will be extended by continuity at the markers positions.

11.1.1. Preliminaries

The proof starts with the computation of the Fisher Information Matrix. As a result, calculations are exactly the same as in Azaïs et al. [2], see Section "Study of the score process under the null hypothesis" of the proof of their Theorem 2.1. We propose to recall here the key elements of the proof.

First, the authors compute the score function at a point $\theta_0^1 = (0,\ \mu,\ \sigma)$ that belongs to H_0:

$$\frac{\partial l_t}{\partial q_1}\,|_{\theta_0^1} = \frac{Y-\mu}{\sigma^2}\,x(t) \tag{11.1}$$

$$\frac{\partial l_t}{\partial \mu} \mid_{\theta_0^1} = \frac{Y-\mu}{\sigma^2} \quad , \quad \frac{\partial l_t}{\partial \sigma} \mid_{\theta_0^1} = -\frac{1}{\sigma} + \frac{(Y-\mu)^2}{\sigma^3} .$$

Then, they introduce their key Lemma (Lemma 2.3 of Azaïs et al. [2]), which states that

$$x(t) = \alpha(t)X(t_1) + \beta(t)X(t_2) \tag{11.2}$$

where $\alpha(t) = Q_t^{1,1} - Q_t^{-1,1}$ and $\beta(t) = Q_t^{1,1} - Q_t^{1,-1}$.

As a result, the Fisher information at θ_0^1, denoted $I_{\theta_0^1}$, verifies

$$I_{\theta_0^1} = Diag\left\{ \frac{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t_1,t_2)}{\sigma^2} , \frac{1}{\sigma^2} , \frac{2}{\sigma^2} \right\} . \tag{11.3}$$

11.1.2. Weak convergence of the score process under H_0

Convergence of finite-dimensional

We have $\forall k = 1, 2$:

$$S_n(t_k) = \frac{\frac{\partial l_{t_k}^n}{\partial q_1} \mid_{\theta_0^1}}{\sqrt{\mathbb{V}_{H_0}\left(\frac{\partial l_{t_k}^n}{\partial q_1} \mid_{\theta_0^1}\right)}} = \sum_{j=1}^{n} \frac{\varepsilon_j \; X_j(t_k)}{\sqrt{n}}.$$

Since $\frac{\partial l_{t_k}^n}{\partial q_1} \mid_{\theta_0^1}$ is centered under H_0, a direct application of the central limit theorem implies that

$$S_n(t_k) \xrightarrow{\mathcal{L}} N(0,\ 1) .$$

Then, since we have the relationship (cf. formula (11.2))

$$S_n(t) = \frac{\alpha(t)S_n(t_1) + \beta(t)S_n(t_2)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t_1,t_2)}},$$

the continous mapping theorem implies that

$$S_n(t) \xrightarrow{\mathcal{L}} Z(t) .$$

It proves the convergence of finite-dimensional.

Tightness

Since we have already proved the convergence of finite-dimensional, let us focus on the tightness of the score process. Since $p(t)$ and $\alpha^2(t)+\beta^2(t)+2\alpha(t)\beta(t)\rho(t_1,t_2)$

are continuous functions, each path of the process $S_n(.)$ is a continuous function on $[t_1, t_2]$. Recall the modulus of continuity of a continous function $h(t)$ on $[t_1, t_2]$:

$$\varpi_h(\delta) = \sup_{|t'-t|<\delta} |h(t') - h(t)| \quad \text{where} \quad t_1 < \delta \leq t_2.$$

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 $S_n(t_1)$ is tight.
2. For each positive ε and η, there exists a δ, with $t_1 < \delta < t_2$, and an integer n_0 such that $\mathbb{P}(\varpi_{S_n}(\delta) \geq \eta) \leq \varepsilon \quad \forall n \geq n_0$.

According to Prohorov, the sequence $S_n(t_1)$ is tight. Then, Condition 1 is verified. Let us define the functions $\alpha'(t)$ and $\beta'(t)$ in the following way:

$$\alpha'(t) = \alpha(t)/\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t_1, t_2)},$$
$$\beta'(t) = \beta(t)/\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t_1, t_2)}.$$

First, we can notice that $\forall \delta$ such as $t_1 < \delta \leq t_2$,

$$\begin{aligned}
\varpi_{S_n}(\delta) &= \sup_{|t'-t|<\delta} |S_n(t') - S_n(t)| \\
&= \sup_{|t'-t|<\delta} |(\alpha'(t') - \alpha'(t)) S_n(t_1) + (\beta'(t') - \beta'(t)) S_n(t_2)| \\
&\leq \max(|S_n(t_1)|, |S_n(t_2)|) \quad (\varpi_{\alpha'}(\delta) + \varpi_{\beta'}(\delta)). \qquad (11.4)
\end{aligned}$$

Furthermore, the sequence $\max(|S_n(t_1)|, |S_n(t_2)|)$ is uniformly tight. This way,

$$\forall \varepsilon > 0 \ \ \exists M > 0 \ \ \forall n \geq 1 \ \ \mathbb{P}(\max(|S_n(t_1)|, |S_n(t_2)|) \geq M) \leq \varepsilon. \qquad (11.5)$$

According to Heine's theorem, since $\alpha'(t)$ and $\beta'(t)$ are continuous on the compact $[t_1, t_2]$, these functions are uniformly continuous. So,

$$\forall \upsilon > 0 \ \ \exists \delta \ \text{ such as } \ t_1 < \delta < t_2, \ \ \varpi_{\alpha'}(\delta) + \varpi_{\beta'}(\delta) < \upsilon. \qquad (11.6)$$

Let η be a positive quantity. Using formulae (11.5) and (11.6) and imposing $\upsilon = \eta/M$, we have

$$\mathbb{P}(\max(|S_n(t_1)|, |S_n(t_2)|)(\varpi_{\alpha'}(\delta) + \varpi_{\beta'}(\delta)) \geq \eta) \leq \varepsilon.$$

As a consequence, according to formula (11.4), we have

$$\forall n \geq 1 \quad \mathbb{P}(\varpi_{S_n}(\delta) \geq \eta) \leq \varepsilon.$$

It proves Condition 2 of Theorem 8.2 of Billingsley (1999). As a result, the tightness of the score process is proved. To conclude, the tightness and the convergence of finite-dimensional imply the weak convergence of the score process.

11.1.3. Study of the score process under the local alternative $H_{a\vec{t}^\star}$

There are m QTLs located on $[0,T]$ and the model for the quantitative trait is the following:

$$Y = \mu \ + \ \sum_{s=1}^{m} X(t_s^\star)\ q_s \ + \ \sigma\varepsilon \tag{11.7}$$

where ε is a Gaussian white noise.

Since the score test statistic at t can be obtained using the following non linear interpolation

$$S_n(t) = \frac{\alpha(t)\ S_n(t_1) \ + \ \beta(t)\ S_n(t_2)}{\sqrt{\alpha^2(t) + \beta^2(t) \ + \ 2\alpha(t)\beta(t)\rho(t_1,t_2)}} \ ,$$

, the mean function will be also a non linear interpolation

$$m_{\vec{t}^\star}(t) = \frac{\alpha(t)\ m_{\vec{t}^\star}(t_1) \ + \ \beta(t)\ m_{\vec{t}^\star}(t_2)}{\sqrt{\alpha^2(t) + \beta^2(t) \ + \ 2\alpha(t)\beta(t)\rho(t_1,t_2)}} \ .$$

Let us compute the quantities $m_{\vec{t}^\star}(t_1)$ and $m_{\vec{t}^\star}(t_2)$.

Without loss of generality, let's consider location t_k which refers to the location of marker k. According to formulae (11.1) and (11.12), we have

$$\begin{aligned} S_n(t_k) &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j\ X_j(t_k) \ + \ \frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} X_j(t_s^\star)\ a_s \right\} X_j(t_k) \\ &= S_n^0(t_k) \ + \ \frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} X_j(t_s^\star)\ a_s \right\} X_j(t_k) \end{aligned} \tag{11.8}$$

where $S_n^0(t_k)$ is the score obtained under H_0 at location t_k.
By the law of large number :

$$\frac{1}{n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} X_j(t_s^\star)\ a_s \right\} \ X_j(t_k) \to \mathbb{E}\left[\left\{ \sum_{s=1}^{m} X(t_s^\star)\ a_s \right\} X(t_k) \right] \ .$$

We have

$$\mathbb{E}\left[\left\{ \sum_{s=1}^{m} X(t_s^\star)\ a_s \right\} X(t_k) \right] = \sum_{s=1}^{m} a_s\ e^{-2|t_s^\star - t_k|} = \sum_{s=1}^{m} a_s\ \rho(t_s^\star, t_k) \ .$$

Then,

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

As a consequence, if we consider $t_k = t_1$ or $t_k = t_2$, we have

$$m_{\vec{t}^\star}(t_1) = \frac{1}{\sigma} \sum_{s=1}^{m} a_s\ \rho(t_s^\star, t_1) \quad , \quad m_{\vec{t}^\star}(t_2) = \frac{1}{\sigma} \sum_{s=1}^{m} a_s\ \rho(t_s^\star, t_2) \ .$$

11.1.4. Study of the supremum of the LRT process

At fixed t, the model is regular and it is well known that we have the following relationship under H_0 (i.e. no QTL on the whole interval studied)

$$\Lambda_n(t) = S_n^2(t) \ + \ o_P(1)$$

and where $o_P(1)$ is short for a sequence of random vectors that converges to zeros in probability. Let us consider now t as an extra parameter. It easy to check that at H_0 the Fisher Information relative to t is zero so that the model is not regular. As a consequence, Azaïs et al. [2] studied this irregular model and proved that

$$\sup \Lambda_n(t) = \sup S_n^2(t) + o_P(1) \ . \tag{11.9}$$

Note that the proof is based on results of Azaïs et al. [4], Azaïs et al. [3] and Gassiat [16] on empirical process theory. This result has been obtained under H_0 and under the local alternative of only one QTL (i.e. $m = 1$), located at $t_1^\star$ on $[0, T]$. This way, our goal is now to show that the remainder converges also to zero under $H_{a\vec{t}^\star}$.

Recall that the parameters θ^m and θ_0^m are defined in the following way : $\theta^m = (q_1, ..., q_m, \mu, \sigma)$ and $\theta_0^m = (0, ..., 0, \mu, \sigma)$. Recall also that the full likelihood of the triplet $(Y,\ X(t_1),\ X(t_2))$, with respect to the measure $\lambda \otimes N \otimes N$, is given in formula (4.5). According to Bayes rules,

$$w_{\vec{t}^\star}(u_1,\ ...,\ u_m) = \frac{\mathbb{P}\left\{X(t_1^\star) = u_1, ..., X(t_m^\star) = u_m, X(t_1), X(t_2\right\}}{\mathbb{P}\left\{X(t_1), X(t_2)\right\}} \ .$$

Besides, we have the relationships

$$\begin{aligned}
&\mathbb{P}\left\{X(t_1^\star) = u_1, ..., X(t_m^\star) = u_m, X(t_1), X(t_2)\right\} \\
&= \mathbb{P}\left\{X(t_1)\right\} \mathbb{P}\left\{X(t_1^\star) = u_1 \mid X(t_1)\right\} \\
&\times \mathbb{P}\left\{X(t_2^\star) = u_2 \mid X(t_1^\star) = u_1\right\} \cdots \mathbb{P}\left\{X(t_m^\star) = u_m \mid X(t_{m-1}^\star) = u_{m-1}\right\} \\
&\times \mathbb{P}\left\{X(t_2) \mid X(t_m^\star) = u_m\right\} \\
&= \frac{1}{2}\left\{r(t_1, t_1^\star)\ 1_{X(t_1)u_1=-1} + \bar{r}(t_1, t_1^\star)\ 1_{X(t_1)u_1=1}\right\} \\
&\times \left\{r(t_1^\star, t_2^\star)\ 1_{u_1u_2=-1} + \bar{r}(t_1^\star, t_2^\star)\ 1_{u_1u_2=1}\right\} \\
&\times \ldots \times \left\{r(t_{m-1}^\star, t_m^\star)\ 1_{u_{m-1}u_m=-1} + \bar{r}(t_{m-1}^\star, t_m^\star)\ 1_{u_{m-1}u_m=1}\right\} \\
&\times \left\{r(t_m^\star, t_2)\ 1_{u_mX(t_2)=-1} + \bar{r}(t_m^\star, t_2) 1_{u_mX(t_2)=1}\right\}
\end{aligned}$$

and in the same way,

$$\mathbb{P}\left\{X(t_1), X(t_2)\right\} = \frac{1}{2}\left\{r(t_1, t_2)\ 1_{X(t_1)X(t_2)=-1} + \bar{r}(t_1, t_2)\ 1_{X(t_1)X(t_2)=1}\right\} \ .$$

As a result,

$$
\begin{aligned}
&w_{\vec{t}^\star}(u_1,\ ...,\ u_m)\\
&= \left\{r(t_1,t_1^\star)\ 1_{X(t_1)u_1=-1} + \bar{r}(t_1,t_1^\star)\ 1_{X(t_1)u_1=1}\right\}\left\{r(t_1^\star,t_2^\star)\ 1_{u_1u_2=-1} + \bar{r}(t_1^\star,t_2^\star)\ 1_{u_1u_2=1}\right\}\\
&\ldots\left\{r(t_{m-1}^\star,t_m^\star)\ 1_{u_{m-1}u_m=-1} + \bar{r}(t_{m-1}^\star,t_m^\star)\ 1_{u_{m-1}u_m=1}\right\}\\
&\left\{r(t_m^\star,t_2)\ 1_{u_mX(t_2)=-1} + \bar{r}(t_m^\star,t_2)1_{u_mX(t_2)=1}\right\} / \left\{r(t_1,t_2)\ 1_{X(t_1)X(t_2)=-1} + \bar{r}(t_1,t_2)\ 1_{X(t_1)X(t_2)=1}\right\}.
\end{aligned}
$$

The likelihood $L_{\vec{t}^\star}^{m,n}(\theta^m)$ for n observations is obtained by the product of n terms as in formula (4.5). 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 $L_{\vec{t}^\star}^{m,n}(\theta^m)$ (respectively $L_{\vec{t}^\star}^{m,n}(\theta_0^m)$).

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

$$
\log\frac{dQ_n}{dP_n} = \log\left\{\frac{L_{\vec{t}^\star}^{m,n}(\theta^m)}{L_{\vec{t}^\star}^{m,n}(\theta_0^m)}\right\}\ . \tag{11.10}
$$

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

$$
\log\left(\frac{dQ_n}{dP_n}\right) \overset{H_0}{\to} N(-\frac{1}{2}\vartheta^2,\ \vartheta^2)\ \text{ with } \vartheta^2 \in \mathbb{R}^{+\star}\ .
$$

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 (11.9) is also true under the alternative $H_{a\vec{t}^\star}$.

11.2. Proof of Theorem 6.1

Since the process $S_n(.)$ is an interpolated process, we can focus, without loss of generality, only on location t_k (i.e. the location of marker k). According to formulae (2.1) and (11.1), we have

$$
\begin{aligned}
S_n(t_k) &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j\ X_j(t_k)\ +\ \frac{1}{\sigma n}\sum_{j=1}^{n}\left\{\sum_{s=1}^{m}X_j(t_s^\star)\ a_s\right\}X_j(t_k)\\
&\quad+\frac{1}{\sigma n}\sum_{j=1}^{n}\left\{\sum_{s=1}^{m-1}\sum_{\tilde{s}=s+1}^{m}X_j(t_s^\star)X_j(t_{\tilde{s}}^\star)\ b_{s,\tilde{s}}\right\}X_j(t_k)\\
&= S_n^0(t_k)\ +\ \frac{1}{\sigma n}\sum_{j=1}^{n}\left\{\sum_{s=1}^{m}X_j(t_s^\star)\ a_s\right\}X_j(t_k)\\
&\quad+\frac{1}{\sigma n}\sum_{j=1}^{n}\left\{\sum_{s=1}^{m-1}\sum_{\tilde{s}=s+1}^{m}X_j(t_s^\star)X_j(t_{\tilde{s}}^\star)\ b_{s,\tilde{s}}\right\}X_j(t_k)
\end{aligned}
$$

where $S_n^0(t_k)$ is the score obtained under H_0 at location t_k.

As in the previous proofs, by the law of large number

$$\frac{1}{n}\sum_{j=1}^{n}\left\{\sum_{s=1}^{m} X_j(t_s^\star)\ a_s\right\}\ X_j(t_k) \to \sum_{s=1}^{m} a_s\ \rho(t_s^\star, t_k)\ \ .$$

In the same way,

$$\frac{1}{\sigma n}\sum_{j=1}^{n}\left\{\sum_{s=1}^{m-1}\sum_{\tilde{s}=s+1}^{m} X_j(t_s^\star)X_j(t_{\tilde{s}}^\star)\ b_{s,\tilde{s}}\right\} X_j(t_k) \to \mathbb{E}\left[\left\{\sum_{s=1}^{m-1}\sum_{\tilde{s}=s+1}^{m} X(t_s^\star)X(t_{\tilde{s}}^\star)\ b_{s,\tilde{s}}\right\} X(t_k)\right]$$

We have the relationship

$$\begin{aligned}\mathbb{E}\left\{X(t_s^\star)X(t_{\tilde{s}}^\star)X(t_k)\right\} &= \mathbb{E}\left[X(t_s^\star)X(t_{\tilde{s}}^\star)\left\{21_{X(t_k)=1}-1\right\}\right]\\ &= 2\,\mathbb{E}\left[X(t_s^\star)X(t_{\tilde{s}}^\star)1_{X(t_k)=1}\right] - \rho(t_s^\star, t_{\tilde{s}}^\star)\end{aligned}$$

Then, if $t_k < \min(t_s^\star, t_{\tilde{s}}^\star)$ or $t_k > \max(t_s^\star, t_{\tilde{s}}^\star)$, we have

$$\mathbb{E}\left[X(t_s^\star)X(t_{\tilde{s}}^\star)1_{X(t_k)=1}\right] = \mathbb{E}\left[X(t_s^\star)X(t_{\tilde{s}}^\star)\mid X(t_k)=1\right]/2 = \mathbb{E}\left[X(t_s^\star)X(t_{\tilde{s}}^\star)\right]/2 = \rho(t_s^\star, t_{\tilde{s}}^\star)/2\ .$$

Besides, if $\min(t_s^\star, t_{\tilde{s}}^\star) < t_k < \max(t_s^\star, t_{\tilde{s}}^\star)$,

$$\begin{aligned}\mathbb{E}\left[X(t_s^\star)X(t_{\tilde{s}}^\star)1_{X(t_k)=1}\right] &= \mathbb{E}\left[X(t_s^\star)X(t_{\tilde{s}}^\star)\mid X(t_k)=1\right]/2\\ &= \mathbb{E}\left[X(t_s^\star)\mid X(t_k)=1\right]\mathbb{E}\left[X(t_{\tilde{s}}^\star)\mid X(t_k)=1\right]/2\\ &= \rho(t_s^\star, t_k)\rho(t_k, t_{\tilde{s}}^\star)/2 = \rho(t_s^\star, t_{\tilde{s}}^\star)/2\ .\end{aligned}$$

As a result, we always have

$$\mathbb{E}\left\{X(t_s^\star)X(t_{\tilde{s}}^\star)X(t_k)\right\} = 0\ .$$

To conclude,

$$\begin{aligned}&\frac{1}{n}\sum_{j=1}^{n}\left\{\sum_{s=1}^{m} X_j(t_s^\star)\ a_s\right\}\ X_j(t_k)\ +\ \frac{1}{\sigma n}\sum_{j=1}^{n}\left\{\sum_{s=1}^{m-1}\sum_{\tilde{s}=s+1}^{m} X_j(t_s^\star)X_j(t_{\tilde{s}}^\star)\ b_{s,\tilde{s}}\right\} X_j(t_k) \qquad (11.11)\\ &\to \sum_{s=1}^{m} a_s\ \rho(t_s^\star, t_k)\ \ .\end{aligned}$$

We can notice that the interaction effects have disappeared and that we have exactly the same mean function as in Theorem 5.1: $m_{\vec{t}^\star}(t_k) = \sum_{s=1}^{m} a_s\rho(t_k, t_s^\star)/\sigma$. Since the model (based on formula 6.1) is differentiable in quadratic mean at $(0, ..., 0, \mu, \sigma^2)$ (i.e. no additive and no epistatic effect), this result is also suitable for the LRT process.

11.3. Proof of Theorem 7.1

11.3.1. Proof of formula (7.3)

Recall that K genetic markers are located at $0 = t_1 < t_2 < \ldots < t_K = T$. 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 $t_1^\star < t_2^\star < \ldots < t_m^\star$. Let us compute the probability distribution of $\big(Y, \overline{X}(t_1^{\star\ell}), \overline{X}(t_1^{\star r}), \ldots, \overline{X}(t_m^{\star\ell}), \overline{X}(t_m^{\star r})\big)$.

We have

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

Besides,

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

On the other hand,

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

As a result,

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

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

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

Then, the probability distribution of $\big(Y, \overline{X}(t_1^{\star\ell}), \overline{X}(t_1^{\star r}), \ldots, \overline{X}(t_m^{\star\ell}), \overline{X}(t_m^{\star r})\big)$, with respect to the measure $\lambda \otimes N \otimes \ldots \otimes N$, is

$$\overline{L}^m_{\vec{t}^\star}(\theta^m) = \sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m} \big[w_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)\ 1_{Y\notin[S_-,S_+]} + v_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+\ldots+u_mq_m,\sigma)}(Y)\ 1_{Y\in[S_-,S_+]}\big]\ \overline{g}^m(t_1^\star,\ldots,t_m^\star)$$

with

$$w_{\vec{t}^\star}(u_1,\ ...,\ u_m) = \mathbb{P}(X(t_1^\star)=u_1, X(t_2^\star)=u_2, \ldots, X(t_m^\star)=u_m \mid X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r}))\ ,$$

$$v_{\vec{t}^\star}(u_1,\ ...,\ u_m) = \mathbb{P}(X(t_1^\star)=u_1, X(t_2^\star)=u_2, \ldots, X(t_m^\star)=u_m)$$

and

$$\overline{g}^m(t_1^\star,\ldots,t_m^\star) = \mathbb{P}(X(t_1^{\star\ell}), X(t_1^{\star r}), \ldots, X(t_m^{\star\ell}), X(t_m^{\star r}))\ 1_{Y\notin[S_-,S_+]}\ +\ 1_{Y\in[S_-,S_+]}\ .$$

11.3.2. Study of the score process under H_0

The score process $\overline{S}_n(.)$ has already been studied in Rabier [32]. Let us recall the key elements of the proof. By definition, the score statistic at t is the following

$$\overline{S}_n(t) = \frac{\frac{\partial \overline{l}^n_t}{\partial q_1}\mid_{\theta_0^1}}{\sqrt{\mathbb{V}\left(\frac{\partial \overline{l}^n_t}{\partial q_1}\mid_{\theta_0^1}\right)}} \text{ where } \theta_0^1 = (0,\ \mu,\ \sigma)\ .$$

The score function verifies

$$\begin{aligned}\frac{\partial \overline{l}^n_t}{\partial q_1}\mid_{\theta_0^1} &= \sum_{j=1}^n \frac{Y_j-\mu}{\sigma^2}\ \{2p_j(t)-1\}\ 1_{Y_j\notin[S_-,S_+]}\\ &= \frac{\alpha(t)}{\sigma}\sum_{j=1}^n \varepsilon_j\ \overline{X}_j(t^\ell)\ +\ \frac{\beta(t)}{\sigma}\sum_{j=1}^n \varepsilon_j\ \overline{X}_j(t^r)\ .\end{aligned}$$

As a result, the limiting process is a non linear interpolated process.

On the other hand, at location t_k:

$$\overline{S}_n(t_k) = \frac{\frac{\partial \overline{l}^n_{t_k}}{\partial q}\mid_{\theta_0}}{\sqrt{\mathbb{V}\left(\frac{\partial \overline{l}^n_{t_k}}{\partial q}\mid_{\theta_0}\right)}} = \sum_{j=1}^n \frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}}\ .$$

According to the Central Limit Theorem,

$$\overline{S}_n(t_k) \xrightarrow{\mathcal{L}} N(0,\ 1)\ .$$

Besides, we have the relationship

$$\mathrm{Cov}_{H_0}\left\{\overline{S}_n(t_k), \overline{S}_n(t_{k'})\right\} = \rho(t_k, t_{k'})\ .$$

11.3.3. Study of the score process under the local alternative $H_{a\vec{t}^\star}$

There are m QTLs located on $[0,T]$ and that the model for the quantitative trait is the following:

$$Y = \mu \ + \ \sum_{s=1}^{m} X(t_s^\star)\ q_s \ + \ \sigma\varepsilon \tag{11.12}$$

where ε is a Gaussian white noise.

Since the score test statistic at t can be obtained using the following non linear interpolation

$$\overline{S}_n(t) = \frac{\alpha(t)\ \overline{S}_n(t^\ell) \ + \ \beta(t)\ \overline{S}_n(t^r)}{\sqrt{\alpha^2(t) + \beta^2(t) \ + \ 2\alpha(t)\beta(t)\rho(t^\ell, t^r)}} \ ,$$

, the mean function will be also a non linear interpolation

$$\overline{m}_{\vec{t}^\star}(t) = \frac{\alpha(t)\ \overline{m}_{\vec{t}^\star}(t^\ell) \ + \ \beta(t)\ \overline{m}_{\vec{t}^\star}(t^r)}{\sqrt{\alpha^2(t) + \beta^2(t) \ + \ 2\alpha(t)\beta(t)\rho(t^\ell, t^r)}} \ .$$

Let us compute the quantities $\overline{m}_{\vec{t}^\star}(t^\ell)$ and $\overline{m}_{\vec{t}^\star}(t^r)$.

Without loss of generality, let's consider location t_k which refers to the location of marker k.

$$\begin{aligned}\overline{S}_n(t_k) &= \sum_{j=1}^{n} \frac{(Y_j - \mu)\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}} \\ &= \sum_{j=1}^{n}\sum_{s=1}^{m} \frac{q_s\ \overline{X}_j(t_s^\star)\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}} \ + \ \sum_{j=1}^{n} \frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}} \ .\end{aligned} \tag{11.13}$$

We will see, that we can apply the Law of Large Numbers for the first term and the Central Limit Theorem for the second term. To begin, let's focus on the first term. We have

$$\begin{aligned}&\mathbb{E}\left\{\overline{X}(t_s^\star)\ \overline{X}(t_k)\right\} = \\ &\mathbb{E}\left[1_{Y\notin[S_-,S_+]}\left\{1_{X(t_s^\star)=1}1_{X(t_k)=1} + 1_{X(t_s^\star)=-1}1_{X(t_k)=-1}\right\}\right] \\ &- \mathbb{E}\left[1_{Y\notin[S_-,S_+]}\left\{1_{X(t_s^\star)=-1}1_{X(t_k)=1} + 1_{X(t_s^\star)=1}1_{X(t_k)=-1}\right\}\right] .\end{aligned}$$

According to calculations present in the joint supplementary material,

$$\begin{aligned}&\mathbb{E}\left[1_{Y\notin[S_-,S_+]}\left\{1_{X(t_s^\star)=1}1_{X(t_k)=1} + 1_{X(t_s^\star)=-1}1_{X(t_k)=-1}\right\}\right] \\ &= \overline{r}(t_k, t_s^\star)\ \left\{1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) + \Phi\left(\frac{S_- - \mu}{\sigma}\right)\right\} \ + \ o(1) \ ,\end{aligned}$$

where Φ is the cumulative distribution of a standard normal distribution. In the same way,

$$\begin{aligned}&\mathbb{E}\left[1_{Y\notin[S_-,S_+]}\left\{1_{X(t_s^\star)=-1}1_{X(t_k)=1} + 1_{X(t_s^\star)=1}1_{X(t_k)=-1}\right\}\right] \\ &= r(t_k, t_s^\star)\ \left\{1 - \Phi\left(\frac{S_+ - \mu}{\sigma}\right) + \Phi\left(\frac{S_- - \mu}{\sigma}\right)\right\} \ + \ o(1) \ .\end{aligned}$$

Since we have the relationships

$$1-\Phi\left(\frac{S_+-\mu}{\sigma}\right)+\Phi\left(\frac{S_--\mu}{\sigma}\right)=\gamma \ \text{ and } \ \overline{r}(t_k,t_s^\star)-r(t_k,t_s^\star)=\rho(t_k,t_s^\star),$$

then we have

$$\mathbb{E}\left\{\overline{X}(t_s^\star)\ \overline{X}(t_k)\right\}=\rho(t_k,t_s^\star)\ \gamma\ +\ o(1)\ .$$

As a consequence, according to the Law of Large Numbers,

$$\sum_{j=1}^{n}\sum_{s=1}^{m}\frac{q_s\ \overline{X}_j(t_s^\star)\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}}\to\sum_{s=1}^{m}\frac{a_s\ \rho(t_k,t_s^\star)\ \gamma}{\sqrt{\mathcal{A}}}\ . \tag{11.14}$$

Let us now focus on the second term of formula (11.13). According to a technical proof present in the supplementary material, we have

$$\mathbb{E}\left\{\sigma\varepsilon\ \overline{X}(t_k)\right\}=\left\{z_{\gamma^+}\ \varphi(z_{\gamma^+})-z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\right\}\ \sum_{s=1}^{m}\rho(t_s^\star,t_k)\ q_s+\ o(\max_{1\le s\le m}|q_s|)\ .$$

Besides, according to iii) of Lemma 5 of Rabier [30],

$$\begin{aligned}
&\mathbb{E}\left[\left\{\sigma\varepsilon\ \overline{X}(t_k)\right\}^2\right]=\mathbb{E}\left(\sigma^2\ \varepsilon^2\ 1_{Y\notin[S_-,S_+]}\right)\\
&=\sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m}\mathbb{E}\left\{\sigma^2\ \varepsilon^2\ 1_{Y\notin[S_-,S_+]}\mid X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\}\\
&\times\mathbb{P}\left\{X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\}\\
&\to\sum_{(u_1,\ldots,u_m)\in\{-1,1\}^m}\mathcal{A}\,\mathbb{P}\left\{X(t_1^\star)=u_1,\ldots,X(t_m^\star)=u_m\right\}\ \to\mathcal{A}\ .
\end{aligned}$$

As a result,

$$\mathbb{E}\left[\left\{\sigma\varepsilon\ \overline{X}(t_k)\right\}^2\right]\to\mathcal{A}\ \text{ and }\ \mathbb{V}\left\{\sum_{j=1}^{n}\frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}}\right\}\to 1\ .$$

Then, according to the Central Limit Theorem,

$$\sum_{j=1}^{n}\frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}}\xrightarrow{\mathcal{L}}N\left[\frac{\sum_{s=1}^{m}\rho(t_s^\star,t_k)\ a_s}{\sqrt{\mathcal{A}}}\ \left\{z_{\gamma^+}\ \varphi(z_{\gamma^+})-z_{1-\gamma^-}\ \varphi(z_{1-\gamma^-})\right\},\ 1\right]\ . \tag{11.15}$$

Finally, according to formulae (11.14) and (11.15),

$$\overline{S}_n(t_k)\xrightarrow{\mathcal{L}}N\left[\sum_{s=1}^{m}\rho(t_k,t_s^\star)\ a_s\sqrt{\mathcal{A}}/\sigma^2,\ 1\right]\ .$$

11.3.4. Study of the supremum of the LRT process

At fixed t, the model is regular and it is well known that we have the following relationship under H_0 (i.e. no QTL on the whole interval studied)

$$\overline{\Lambda}_n(t) = \overline{S}_n^{\,2}(t) \; + \; o_P(1)$$

and where $o_P(1)$ is short for a sequence of random vectors that converges to zeros in probability. The problem is that, when t is not fixed, the Fisher Information relative to t at H_0 is zero so that the model is not regular. As a result, let us consider now t as an extra parameter. Let $t_1^\star$ and $\theta^{1\star}$, be the true parameters that will be assumed to belong to H_0. Note that $t_1^\star$ makes no sense for θ belonging to H_0.

Without loss of generality, let us consider that μ and σ are known ($\mu = 0$ and $\sigma = 1$) and as previously, let us consider values of t distinct of the markers positions. Besides, let us consider only two genetic markers located at $t_1 = 0$ and $t_2 = T$. Note that in order to make the reading easier, we will use the notation $f_q(.)$ instead of $f_{(q,1)}(.)$ to denote a Gaussian density with mean q and unit variance.

For computing scores at each t separately, the likelihood of $(Y,\ X(t_1),\ X(t_2))$ can be considered in the following way $\forall t \in]t_1, t_2[$:

$$\begin{aligned}\overline{L}(\psi\ q_1, t(q_1)) = \big[\ p\left\{t(q_1)\right\}\ f_{\psi q_1}(Y) 1_{Y\notin[S_-,S_+]} + \left[1 - p\left\{t(q_1)\right\}\right]\ f_{-\psi q_1}(Y) 1_{Y\notin[S_-,S_+]} \\ + \frac{1}{2}\, f_{\psi q_1}(Y) 1_{Y\in[S_-,S_+]} \; + \; \frac{1}{2}\, f_{-\psi q_1}(Y) 1_{Y\in[S_-,S_+]}\Big]\ \overline{g}(t)\end{aligned} \tag{11.16}$$

where $t(q_1)$ is a continuous function on $[0,1]$ into $]t_1, t_2[$, $p(t)$ is the classical weight (cf. Sections 4 and 5) and $\overline{g}(t)$ is given in formula (7.2). Then, at each value of q_1 corresponds a value of t, denoted $t(q_1)$.

Let us compute the score function corresponding to $\tilde{L}$ at $q_1 = 0$. Recall that the null hypothesis is reached if and only if the QTL effect is null.

$$\begin{aligned}\frac{\partial \log \overline{L}}{\partial q_1}\,|_{q_1=0} \; &= \psi\left\{2Q^{1,1}_{t(0)} - 1\right\}\ Y\ 1_{\overline{X}(t_1)=1} 1_{\overline{X}(t_2)=1} \; + \; \psi\left\{2Q^{1,-1}_{t(0)} - 1\right\}\ Y\ 1_{\overline{X}(t_1)=1} 1_{\overline{X}(t_2)=-1} \\ &+ \psi\left\{2Q^{-1,1}_{t(0)} - 1\right\}\ Y\ 1_{\overline{X}(t_1)=-1} 1_{\overline{X}(t_2)=1} \; + \; \psi\left\{2Q^{-1,-1}_{t(0)} - 1\right\}\ Y\ 1_{\overline{X}(t_1)=-1} 1_{\overline{X}(t_2)=-1} \\ &= \psi\left[\alpha\left\{t(0)\right\} + \beta\left\{t(0)\right\}\right]\ Y\ 1_{\overline{X}(t_1)=1} 1_{\overline{X}(t_2)=1} \\ &+ \psi\left[\alpha\left\{t(0)\right\} - \beta\left\{t(0)\right\}\right]\ Y\ 1_{\overline{X}(t_1)=1} 1_{\overline{X}(t_2)=-1} \\ &+ \psi\left[\beta\left\{t(0)\right\} - \alpha\left\{t(0)\right\}\right]\ Y\ 1_{\overline{X}(t_1)=-1} 1_{\overline{X}(t_2)=1} \\ &- \ \psi\left[\beta\left\{t(0)\right\} + \alpha\left\{t(0)\right\}\right]\ Y\ 1_{\overline{X}(t_1)=-1} 1_{\overline{X}(t_2)=-1} \\ &= \psi\ \alpha\left\{t(0)\right\}\ Y\ \overline{X}(t_1) \; + \; \psi\ \beta\left\{t(0)\right\}\ Y\ \overline{X}(t_2)\end{aligned}$$

where $\alpha(.)$ and $\beta(.)$ are the classical quantities introduced in Theorems 4.1 and

5.1. We have

$$\mathbb{E}_{H_0}\left\{\left(\frac{\partial \log \overline{L}}{\partial q_1}\mid_{q_1=0}\right)^2\right\} = \frac{\mathcal{A}}{\sigma^4}\psi^2\ \left[\alpha^2\left\{t(0)\right\}\ +\ \beta^2\left\{t(0)\right\}\ +\ 2\alpha\left\{t(0)\right\}\beta\left\{t(0)\right\}\ \rho(t_1,t_2)\right]\ .$$

This quantity is always different from 0. As previously, for any function $t(q_1)$, each sub-model $\tilde{L}(q_1, t(q_1))_{q_1\in\mathbb{R}}$ is differentiable in quadratic mean and Assumption 2 of Azaïs et al. [4] is verified. Besides, the set of log likelihood is Glivenko-Cantelli (cf. example 19.7 with $r = 1$ of Van der Vaart [41]), so Assumption 1 of Azaïs et al. [4] holds.

As in Azaïs et al. [4], let us define the set of scores renormalized $\mathcal{D}$. In our case :

$$\mathcal{D} = \left\{\ sign(\psi)\frac{\alpha\left\{t(0)\right\}\ Y\ \overline{X}(t_1)\ \ +\ \ \beta\left\{t(0)\right\}\ Y\ \overline{X}(t_2)}{\sqrt{\alpha^2\left\{t(0)\right\}\ +\ \beta^2\left\{t(0)\right\}\ +\ 2\alpha\left\{t(0)\right\}\beta\left\{t(0)\right\}\ \rho(t_1,t_2)}}\ \ ;\ \ t(0)\in]t_1,t_2[\ \ ;\ \ \psi\in\mathbb{R}\right\}$$

which can be rewritten

$$\mathcal{D} = \left\{\ \psi'\frac{\alpha(t)\ Y\ \overline{X}(t_1)\ \ +\ \ \beta(t)\ Y\ \overline{X}(t_2)}{\sqrt{\alpha^2(t)\ +\ \beta^2(t)\ +\ 2\alpha(t)\beta(t)\ \rho(t_1,t_2)}}\ \ ;\ \ t\in]t_1,t_2[\ \ ;\ \ \psi'\in\{-1,1\}\ \right\}\ .$$

Since we have already shown the tightness of the score process $S_n(.)$, $\tilde{\mathcal{D}}$ is Donsker (cf. [41]). In particular it proves that Theorem 1 of Azaïs et al. [3] applies in the sense that

$$\sup_{(t,\theta)}\overline{l}_t^n(\theta) - \overline{l}_{t_1^\star}^n(\theta^{1\star}) = \sup_{d\in\mathcal{D}}\left[\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^n d(X_j)\right\}^2 1_{\sum_{j=1}^n d(X_j)\geq 0}\right] + o_P(1) \quad (11.17)$$

where the observation X_j stands for $(Y_j, X_j(t_1), X_j(t_2))$. Note that since $\psi' \in \{-1, 1\}$, the indicator function can be removed in formula (11.17).

Since the model (based on formula 7.3) is differentiable in quadratic mean at $\theta_0^m = (0, ..., 0, \mu, \sigma)$, we can apply Le Cam first lemma and formula (11.17) is also suitable under the contiguous alternative $H_{a\vec{t}^\star}$.

11.4. *Proof of Theorem 7.2*

Since the process $\overline{S}_n(.)$ is an interpolated process, we can focus, without loss of generality, only on location t_k (i.e. the location of marker k). According to formulae (2.1) and (11.13), we have

$$\begin{aligned}\overline{S}_n(t_k) = &\sum_{j=1}^n\sum_{s=1}^m \frac{a_s\ \overline{X}_j(t_s^\star)\ \overline{X}_j(t_k)}{n\ \sqrt{\mathcal{A}}}\ +\ \sum_{j=1}^n\frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)}{\sqrt{n\ \mathcal{A}}} \\ &+\frac{1}{n\ \sqrt{\mathcal{A}}}\sum_{j=1}^n\left\{\sum_{s=1}^{m-1}\sum_{\tilde{s}=s+1}^m \overline{X}_j(t_s^\star)\overline{X}_j(t_{\tilde{s}}^\star)\ b_{s,\tilde{s}}\right\}\overline{X}_j(t_k)\ .\end{aligned} \quad (11.18)$$

According to calculations present in the supplementary material, when $1 \leq s \leq m-1$ and $s+1 \leq \tilde{s} \leq m$,

$$\mathbb{E}\left\{\overline{X}(t_s^\star)\overline{X}(t_{\tilde{s}}^\star)\overline{X}(t_k)\right\} = o(1) \ .$$

Then, according to the law of large numbers,

$$\overline{S}_n(t_k) = \sum_{j=1}^{n}\sum_{s=1}^{m} \frac{a_s \ \overline{X}_j(t_s^\star) \ \overline{X}_j(t_k)}{n\sqrt{\mathcal{A}}} \ + \ \sum_{j=1}^{n} \frac{\sigma\varepsilon_j \ \overline{X}_j(t_k)}{\sqrt{n\,\mathcal{A}}} + o_P(1) \ .$$

As a result, using formulae (11.14) and (11.15),

$$\overline{S}_n(t_k) \xrightarrow{\mathcal{L}} N\left[\sum_{s=1}^{m} \rho(t_k, t_s^\star) \ a_s\sqrt{\mathcal{A}}/\sigma^2, \ 1\right] \ .$$

We can notice that we have exactly the same function as in Theorem 7.2.

11.5. Proof of Theorem 8.1

11.5.1. Study of the score process under H_0

The score process $\widetilde{S}_n(.)$ has already been studied in details in Rabier [31], under H_0. We propose to recall here the main elements of the proof. By definition, the score statistic at t is the following

$$\widetilde{S}_n(t) = \frac{\frac{\partial \widetilde{l}_t^n}{\partial q_1} \mid_{\theta_0^1}}{\sqrt{\mathbb{V}\left(\frac{\partial \widetilde{l}_t^n}{\partial q_1} \mid_{\theta_0^1}\right)}} \text{ where } \theta_0^1 = (0, \ \mu, \ \sigma) \ .$$

The score function verifies

$$\begin{aligned}\frac{\partial \widetilde{l}_t^n}{\partial q_1} \mid_{\theta_0^1} &= \sum_{j=1}^{n} \frac{Y_j - \mu}{\sigma^2} \ \{2\widetilde{p}_j(t) - 1\} \\ &= \frac{\widetilde{\alpha}(t)}{\sigma}\sum_{j=1}^{n} \varepsilon_j \ X_j(t^\ell) \ + \ \frac{\widetilde{\beta}(t)}{\sigma}\sum_{j=1}^{n} \ \varepsilon_j \ X_j(t^r) \ . \end{aligned} \tag{11.19}$$

Recall also that the score statistic at t_k which refers to the location of marker k, verifies:

$$\widetilde{S}_n(t_k) = \sum_{j=1}^{n} \frac{(Y_j - \mu) \ X_j(t_k)}{\sqrt{n}} = \sum_{j=1}^{n} \frac{\sigma\varepsilon_j \ X_j(t_k)}{\sqrt{n}} \ .$$

Finally, according to formula (11.19) and the computation of the Fisher information matrix, the score statistic at t can be obtained using the following linear interpolation

$$\widetilde{S}_n(t) = \frac{\widetilde{\alpha}(t) \ \widetilde{S}_n(t^\ell) \ + \ \widetilde{\beta}(t) \ \widetilde{S}_n(t^r)}{\sqrt{\widetilde{\alpha}^2(t) + \widetilde{\beta}^2(t) \ + \ 2\widetilde{\alpha}(t)\widetilde{\beta}(t)\rho(t^\ell, t^r)}} \ .$$

11.5.2. Study of the score process under the local alternative $H_{a\vec{t}^\star}$

Since the mean function is the following linear interpolation

$$\widetilde{m}_{\vec{t}^\star}(t) = \frac{\widetilde{\alpha}(t)\ \widetilde{m}_{\vec{t}^\star}(t^\ell)\ +\ \widetilde{\beta}(t)\ \widetilde{m}_{\vec{t}^\star}(t^r)}{\sqrt{\widetilde{\alpha}^2(t) + \widetilde{\beta}^2(t)\ +\ 2\widetilde{\alpha}(t)\widetilde{\beta}(t)\rho(t^\ell, t^r)}}\ ,$$

we only need to compute the quantities $\widetilde{m}_{\vec{t}^\star}(t^\ell)$ and $\widetilde{m}_{\vec{t}^\star}(t^r)$.

Without loss of generality, let us consider location t_k. According to formula (2.3),

$$\begin{aligned}\widetilde{S}_n(t_k) &= \sum_{j=1}^{n} \frac{(Y_j - \mu)\ X_j(t_k)}{\sqrt{n}} \\ &= \frac{1}{\sigma n}\sum_{j=1}^{n}\sum_{s=1}^{m} a_s\ U_j(t_s^\star)\ X_j(t_k)\ +\ \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j\ X_j(t_k)\ .\end{aligned}$$

As in the previous proofs, we can apply the law of large number to the first term. Then, we have

$$\frac{1}{\sigma n}\sum_{j=1}^{n}\left\{\sum_{s=1}^{m} U_j(t_s^\star)\ a_s\right\}\ X_j(t_k) \to \sum_{s=1}^{m} a_s\ h(t_k, t_s^\star)/\sigma$$

where

$$\begin{aligned}h(t_k, t_s^\star) = \rho(t_k, t_s^{\star\ell})&\left\{\widetilde{\alpha}(t_s^\star) + \widetilde{\beta}(t_s^\star)\rho(t_s^{\star\ell}, t_s^{\star r})\right\} 1_{t_s^\star > t_k} \\ &+ \rho(t_k, t_s^{\star r})\left\{\widetilde{\alpha}(t_s^\star)\rho(t_s^{\star\ell}, t_s^{\star r}) + \widetilde{\beta}(t_s^\star)\right\} 1_{t_s^\star < t_k}\ .\end{aligned}$$

Indeed, the quantity $h(t_k, t_s^\star)$ is equal to $\mathbb{E}\left\{U_j(t_s^\star)X_j(t_k)\right\}$, which was computed in Theorem 2 of Rabier [31].

11.6. Proof of Theorem 8.2

Let us first introduce the model for the quantitative trait in presence of interference and epistasis:

$$Y = \mu\ +\ \sum_{s=1}^{m} U(t_s^\star)\ q_s\ +\ \sum_{s=1}^{m-1}\sum_{\tilde{s}=s+1}^{m} U(t_s^\star)U(t_{\tilde{s}}^\star)\ q_{s,\tilde{s}}\ +\ \sigma\varepsilon \tag{11.20}$$

where ε is a Gaussian white noise, and $q_{s,\tilde{s}}$ is the interaction effect between loci $t_s^\star$ and $t_{\tilde{s}}^\star$. Recall that we impose that the QTLs do not belong to the same marker intervals.

Since the process $\widetilde{S}_n(.)$ is an interpolated process, we can focus, without loss of generality, only on location t_k (i.e. the location of marker k). According to formulae (11.20), we have

$$\widetilde{S}_n(t_k) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j \; X_j(t_k) \; + \; \frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} U_j(t_s^\star) \; a_s \right\} X_j(t_k)$$
$$+ \frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^{m} U_j(t_s^\star) U_j(t_{\tilde{s}}^\star) \; b_{s,\tilde{s}} \right\} X_j(t_k) \; .$$

According to the previous section, by the law of large number, we have

$$\frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} U_j(t_s^\star) \; a_s \right\} \; X_j(t_k) \to \sum_{s=1}^{m} a_s \; h(t_s^\star, t_k)/\sigma \quad .$$

In the same way,

$$\frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^{m} U_j(t_s^\star) U_j(t_{\tilde{s}}^\star) \; b_{s,\tilde{s}} \right\} X_j(t_k) \to \mathbb{E} \left[\left\{ \sum_{s=1}^{m-1} \sum_{\tilde{s}=s+1}^{m} U(t_s^\star) U(t_{\tilde{s}}^\star) \; b_{s,\tilde{s}} \right\} X(t_k) \right]$$

If $t_k < \min(t_s^\star, t_{\tilde{s}}^\star)$ or $t_k > \max(t_s^\star, t_{\tilde{s}}^\star)$, since under the interference model $t_s^\star$ and $t_{\tilde{s}}^\star$ do not belong to the same marker interval, we have

$$\mathbb{E}\left[U(t_s^\star)U(t_{\tilde{s}}^\star)1_{X(t_k)=1}\right] = \mathbb{E}\left[U(t_s^\star)U(t_{\tilde{s}}^\star) \mid X(t_k)=1\right]/2 = \mathbb{E}\left[U(t_s^\star)U(t_{\tilde{s}}^\star)\right]/2 \; .$$

As a consequence,

$$\mathbb{E}\left\{U(t_s^\star)U(t_{\tilde{s}}^\star)X(t_k)\right\} = \mathbb{E}\left[U(t_s^\star)U(t_{\tilde{s}}^\star)\left\{21_{X(t_k)=1}-1\right\}\right] = 0 \; .$$

Let us consider now the case $\min(t_s^\star, t_{\tilde{s}}^\star) < t_k < \max(t_s^\star, t_{\tilde{s}}^\star)$. By definition, conditionnally to $X(t_k)$, $U(t_s^\star)$ and $U(t_{\tilde{s}}^\star)$ are independent. As a consequence, we have the relationship

$$\begin{aligned}
&\mathbb{E}\left\{U(t_s^\star)U(t_{\tilde{s}}^\star)X(t_k)\right\} \\
&= \mathbb{E}\left\{U(t_s^\star)U(t_{\tilde{s}}^\star) \mid X(t_k)=1\right\}\mathbb{P}\left\{X(t_k)=1\right\} \; - \; \mathbb{E}\left\{U(t_s^\star)U(t_{\tilde{s}}^\star) \mid X(t_k)=-1\right\}\mathbb{P}\left\{X(t_k)=-1\right\} \\
&= \mathbb{E}\left\{U(t_s^\star) \mid X(t_k)=1\right\}\mathbb{E}\left\{U(t_{\tilde{s}}^\star) \mid X(t_k)=1\right\}/2 \\
&- \mathbb{E}\left\{U(t_s^\star) \mid X(t_k)=-1\right\}\mathbb{E}\left\{U(t_{\tilde{s}}^\star) \mid X(t_k)=-1\right\}/2.
\end{aligned}$$

Besides, we have

$$\begin{aligned}
\mathbb{E}\left\{U(t_s^\star)\right\} &= \frac{1}{2}\,\mathbb{E}\left\{U(t_s^\star) \mid X(t_k)=1\right\} + \frac{1}{2}\,\mathbb{E}\left\{U(t_s^\star) \mid X(t_k)=-1\right\} \\
\mathbb{E}\left\{U(t_{\tilde{s}}^\star)\right\} &= \frac{1}{2}\,\mathbb{E}\left\{U(t_{\tilde{s}}^\star) \mid X(t_k)=1\right\} + \frac{1}{2}\,\mathbb{E}\left\{U(t_{\tilde{s}}^\star) \mid X(t_k)=-1\right\} \; .
\end{aligned}$$

It is easy to check that $U(t_s^\star)$ and $U(t_{\tilde{s}}^\star)$ take value $+1$ and -1 with equal probability (cf. formula 8.2). Then,

$$\begin{aligned}
\mathbb{E}\left\{U(t_s^\star) \mid X(t_k)=1\right\} &= -\mathbb{E}\left\{U(t_s^\star) \mid X(t_k)=-1\right\} \\
\mathbb{E}\left\{U(t_{\tilde{s}}^\star) \mid X(t_k)=1\right\} &= -\mathbb{E}\left\{U(t_{\tilde{s}}^\star) \mid X(t_k)=-1\right\} \quad .
\end{aligned}$$

As a result,

$$\mathbb{E}\left\{U(t_s^\star)U(t_{\tilde{s}}^\star)X(t_k)\right\} = 0 \ .$$

This gives the result.

Acknowledgements

We thank Professor Jean-Marc Azaïs from university Paul Sabatier Toulouse (FR) for fruitful discussions.

References

[1] AZAÏS, J.M. AND CIERCO-AYROLLES, C. (2002). An asymptotic test for quantitative gene detection. *Ann. Inst. Henri Poincaré (B)*, **38(6)** 1087-1092.

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

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

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

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

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

[7] 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.

[8] BUHLMANN, P. AND VAN DE GEER, S. (2011). Statistics for high-dimensional data: methods, theory and applications, Springer Science.

[9] CHANG, M.N., WU, R., WU, S.S., CASELLA, G. (2009). Score statistics for mapping quantitative trait loci. *Stat. Appl. Genet. Mol. Biol.*, **8(1)** 16.

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

[11] CHURCHILL, G.A. AND DOERGE, R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, **138** 963-971.

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

[13] 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.

[14] DONOHO D. (2006). For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution *Comm. Pure Appl. Math.*, **59(6)** 797-829.
[15] FAN, J., LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space *Journal of the Royal Statistical Society: Series B*, **70(5)** 849-911.
[16] GASSIAT, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. Henri Poincaré (B)*, **6** 897-906.
[17] GENZ, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comp. Graph. Stat.*, **1** 141-149.
[18] 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.
[19] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. (2001) *The elements of statistical learning theory*. Springer, New York.
[20] HAYES, B (2007). QTL Mapping, MAS, and Genomic Selection. *Short course organized by Iowa State University*.
[21] LANDER, E.S. and BOTSTEIN, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **138** 235-240.
[22] LEBOWITZ, R.J., SOLLER, M., BECKMANN, J.S. (1987). Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theor. Appl. Genet.*, **73** 556-562.
[23] LOBO, I., Shaw, K. (2008). Thomas Hunt Morgan, genetic recombination, and gene mapping. Nat. Educ., 1:1.
[24] MANICHAIKUL, A., PALMER, A., SEN, S., BROMAN, K. (2007). Significance thresholds for Quantitative Trait Locus mapping under selective genotyping. *Genetics*, **177** 1963-1966.
[25] MARTINI, E., DIAZ, L.D., HUNTER, N., KEENEY, S. (2006). Crossover homeostasis in yeast meiosis. *Cell*, **126** 285-295.
[26] MCPEEK, M.S AND SPEED, T. P. (1995). Modeling interference in genetic recombination. *Genetics*, **139** 1031-1044.
[27] MULLER, H.J. (1916). The mechanism of crossing-over. *Am. Nat.*, **50** 193-221, 284-305, 350-366, 421-434.
[28] 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.
[29] RABBEE, N., SPECA, D., ARMSTRONG, N., SPEED, T. (2004). Power calculations for selective genotyping in QTL mapping in backcross mice. *Genet. Res. Camb.*, **84** 103-108.
[30] RABIER, C.E. (2014). On statistical inference for selective genotyping. *J. Stat. Plan. Infer.*, **147** 24-52.
[31] RABIER, C.E. (2014). On Quantitative Trait Locus mapping with an interference phenomenon. *TEST*, **23(2)** 311-329.
[32] RABIER, C.E. (2013). On stochastic processes for Quantitative Trait Locus mapping under selective genotyping. *Statistics*, DOI

:10.1080/02331888.2013.858720.

[33] RABIER, C.E. (2014). An asymptotic test for Quantitative Trait Locus detection in presence of missing genotypes. *Annales de la faculté des sciences de Toulouse*, **6(23)** 755-778.

[34] RABIER, C.E. (2014). 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.

[35] REBAÏ, A., GOFFINET, B., MANGIN, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, **138** 235-240.

[36] REBAÏ, A., GOFFINET, B., MANGIN, B. (1995). Comparing power of different methods for QTL detection. *Biometrics*, **51** 87-99.

[37] SIEGMUND, D. AND YAKIR, B. (2007). *The statistics of gene mapping*. Springer, New York.

[38] STURTEVANT, A.H. (1915). The behavior of the chromosomes as studied through linkage. *Z. Indukt. Abstammungs. Vererbungsl.*, **13** 234-287.

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

[40] TORTEREAU, F., SERVIN, B., et al. (2012). A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC genomics*, **13**, 586.

[41] VAN DER VAART, A.W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics.

[42] WU, R., MA, C.X., CASELLA, G. (2007). *Statistical Genetics of Quantitative Traits*. Springer, New York.