On stochastic processes for QTL mapping under selective genotyping

Charles-Elie Rabier

On stochastic processes for Quantitative Trait Locus mapping under selective genotyping

Charles-Elie Rabier

Université de Toulouse, Institut de Mathématiques de Toulouse, U.P.S., Toulouse, France.

Summary.
We consider the likelihood ratio test (LRT) process related to the test of the absence of QTL (a QTL denotes a quantitative trait locus, i.e. a gene with quantitative effect on a trait) on the interval $[0,T]$ representing a chromosome. The originality is in the fact that we are under selective genotyping : only the individuals with extreme phenotypes are genotyped. We give the asymptotic distribution of this LRT process under the null hypothesis that there is no QTL on $[0,T]$ and under local alternatives with a QTL at $t^\star$ on $[0,T]$. We show that the LRT is asymptotically the square of " non-linear interpolated and normalized Gaussian process ". We have an easy formula in order to compute the supremum of the square of this interpolated process. We prove that the efficiency with respect to the oracle situation where all the individuals are genotyped, is exactly the same as when we focus only on one locus. Besides, we proove that we have to genotype symetrically and that the non extreme phenotypes (ie. the phenotypes for which the genotypes are missing) don't bring any information for statistical inference. Finally, we show that the threshold is exactly the same as in the oracle situation.

Keywords: Gaussian process, Likelihood Ratio Test, Mixture models, Nuisance parameters present only under the alternative, QTL detection, selective genotyping.

1. Introduction

We study a backcross population: $A \times (A \times B)$, where A and B are purely homozygous lines and we address the problem of detecting a Quantitative Trait Locus, so-called QTL (a gene influencing a quantitative trait which is able to be measured) on a given chromosome. The trait is observed on n individuals (progenies) and we denote by $Y_j,\ j = 1, ..., n$, the observations, which we will assume to be Gaussian, 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 . We use the Haldane modeling that can be represented as follows: $X(0)$ is a random sign and $X(t) = X(0)(-1)^{N(t)}$ where $N(.)$ is a standard Poisson process on $[0,T]$. Calculation on the Poisson distribution show that

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

we set in addition

$$\bar{r}(t,t') = 1 - r(t,t').$$

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

$$Y = \mu \ + \ X(t^\star)\, q \ + \ \sigma\varepsilon \tag{1}$$

where ε is a Gaussian white noise and t^* is the true location of the QTL.

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

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

and the challenge is that the location $t^\star$ of the QTL is unknown.

The originality of this paper is that we consider the classical problem, but this time, in order to reduce the costs of genotyping, a selective genotyping has been performed : we consider two real thresholds S_- and S_+, with $S_- \leqslant S_+$ and we genotype if and only if the phenotype Y is extreme, that is to say $Y \leqslant S_-$ or $Y \geqslant S_+$. 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.

We will observe n observations $\left(Y_j,\ \overline{X}_j(t_1),\ ...,\ \overline{X}_j(t_K)\right)$ i.i.d.

It can be proved that $\left(Y,\ \overline{X}(t_1),\ ...,\ \overline{X}(t_K)\right)$ obeys to a mixture model with known weights, times a function g(.) which does not depend of the parameters μ, q and σ :

$$\begin{aligned} \Big[& p(t^*)\ f_{(\mu+q,\sigma)}(y)\ 1_{y\notin[S_-,S_+]} \ + \ \{1-p(t^*)\}\ f_{(\mu-q,\sigma)}(y)\ 1_{y\notin[S_-,S_+]} \\ & + \ \frac{1}{2}\ f_{(\mu+q,\sigma)}(y)\ 1_{y\in[S_-,S_+]} \ + \ \frac{1}{2}\ f_{(\mu-q,\sigma)}(y)\ \ 1_{y\in[S_-,S_+]} \Big]\ g(.) \end{aligned} \tag{2}$$

where $f_{(m,\sigma)}$ is the Gaussian density with parameters (m, σ) and where the function $p(t)$ is fully given in Section 2.

As said before, the challenge is that t^* is unknown. So, at every location $t \in [0, T]$, we perform a Likelihood Ratio Test (LRT), $\Lambda_n(t)$, of the hypothesis "$q = 0$". It leads to a LRT process $\Lambda_n(.)$ and taking as test statistic the maximum of this process comes down to perform a LRT in a model when the localisation of the QTL is an extra parameter.

In the classical problem of detecting a QTL on a chromosome, that is to say in the oracle situation where all the individuals are genotyped (ie. without selective genotyping), the asymptotic distribution of the LRT statistic has been given under some approximations by Rebaï et al. [1995], Rebaï et al. [1994], Cierco [1998], Azaïs and Cierco-Ayrolles [2002], Azaïs and Wschebor [2009], Chang et al. [2009]. Recently, Azaïs et al. [2011] have shown that the distribution of the LRT statistic is asymptotically that of the maximum of the square of a "non linear normalized interpolated process".

The originality of this paper is in the fact that we study a problem which has never been studied before : the detection of a QTL on a chromosome with a selective genotyping. Selective genotyping has been studied by many authors : for instance Lebowitz and al. [1987], Lander and Botstein [1989], Darvasi and Soller [1992], Muranty and Goffinet [1997], Rabier [2011]... However, in all these articles, the focus is only on one fixed location of the genome. This way, our study which focus on the whole chromosome is totally new, with a real impact for geneticists.

The main result of the paper (Theorem 1 and 2) is that the distribution of the LRT statistic is asymptotically that of the maximum of the square of a "non linear normalized interpolated process". This is a generalization of the results obtained by Azaïs et al. [2011] only for the oracle situation. Under the null hypothesis, despite the selective genotyping, our process is exactly the same as the one obtained by Azaïs et al. [2011]. However, under the alternative, we show that the mean functions of the two processes are not the same anymore.
Another important result is introduced in Theorem 3 : we have exactly the same Asymptotic Relative Efficiency (ARE) with respect to the oracle situation, if we look for a QTL on a whole chromosome and if we look for a QTL only on one given genetic marker. Theorem 3 also says that if we want to genotype only a percentage γ of the population, we should genotype symetrically : the $\gamma/2\%$ individuals with the largest phenotypes and $\gamma/2\%$ individuals with the smallest phenotypes. This is a generalization of Rabier [2011], where it is prooved that we have to genotype symetrically, when we focus only on one genetic marker.
Furthermore, we have an easy formula (see Lemma 2 and formula 20) to compute the maximum of the square of the non linear interpolated process. This formula is original. Usually when we look for a QTL on a chromosome with a selective genotyping, we have to compute an EM algorithm at each location, so it is quite challenging. With our formula, we don't need to perform any EM algorithm and we only have to focus on given locations on the chromosome. Note that in this paper, we also prove that the non extreme phenotypes (for which the genotypes are missing) don't bring any extra information for statistical inference (same result as in Rabier [2011] but for the whole chromosome).

To conclude, we will illustrate our theoretical results with the help of simulated data. Note that, according to Theorem 1 and 2, the threshold (ie. critical value) in selective genotyping, is exactly the same as the classical threshold used in the oracle situation. So, in order to obtain our threshold, the Monte Carlo Quasi Monte-Carlo methods of Azaïs et al. [2011], based on Genz [1992] is still suitable here. We refer to the book of Van der Vaart [1998] for elements of asymptotic statistics used in proofs.

2. Main results : two genetic markers

To begin, we suppose that there are only two markers ($K = 2$) located at 0 and T : $0 = t_1 < t_2 = T$. We look for a QTL located at $t^\star \in [t_1, t_2]$. As said before, since $t^\star$ is unknown, we have to consider every locations $t \in [t_1, t_2]$. So, let's consider a location $t \in [t_1, t_2]$, and let's suppose $t = t^\star$.

For $(i, i') \in \{-1, 1\}^2$ we define the quantity $Q_t^{i,i'}$ such as

$$Q_t^{i,i'} = \mathbb{P}\left\{X(t) = 1 \middle| X(t_1) = i, X(t_2) = i'\right\} .$$

Using Bayes rules, we have

$$Q_t^{1,1} = \frac{\bar{r}(t_1,t)\,\bar{r}(t,t_2)}{\bar{r}(t_1,t_2)} \quad , \quad Q_t^{1,-1} = \frac{\bar{r}(t_1,t)\,r(t,t_2)}{r(t_1,t_2)} \tag{3}$$
$$Q_t^{-1,1} = \frac{r(t_1,t)\,\bar{r}(t,t_2)}{r(t_1,t_2)} \quad , \quad Q_t^{-1,-1} = \frac{r(t_1,t)\,r(t,t_2)}{\bar{r}(t_1,t_2)} \ .$$

We can remark 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 $\mathbb{P}\{k \mid i\}$ be the quantity such as $\forall\ i \in \{-1,1\}$ and $\forall\ l \in \{-1,0,1\}$

$$\mathbb{P}\{l \mid i\} = \mathbb{P}(\overline{X}(t) = l \mid X(t) = i) \ .$$

In order to compute the likelihood, we have to study the different probability laws. To begin, let's compute $\mathbb{P}(Y \in [y\ ,\ y+dy] \ \cap\ \overline{X}(t_1) = 1 \ \cap\ \overline{X}(t_2) = 1)$ for instance. We have, according to Bayes rules,

$$\begin{aligned}
&\mathbb{P}(Y \in [y\ ,\ y+dy] \ \cap\ \overline{X}(t_1) = 1 \ \cap\ \overline{X}(t_2) = 1) \\
&= \sum_{i \in \{-1,1\}} \mathbb{P}(Y \in [y\ ,\ y+dy] \mid \overline{X}(t) = i)\ \ \mathbb{P}(\overline{X}(t) = k \ \cap\ \overline{X}(t_1) = 1 \ \cap\ \overline{X}(t_2) = 1)
\end{aligned}$$

Besides,

$$\begin{aligned}
\mathbb{P}(Y \in [y\ ,\ y+dy] \mid \overline{X} = i) &= \frac{\mathbb{P}(Y \in [y\ ,\ y+dy] \cap \overline{X} \neq 0 \mid X(t) = i)}{\mathbb{P}(\overline{X}(t) \neq 0 \mid X(t) = i)} \\
&= \frac{f_{(\mu+q,\sigma)}(y)\ 1_{y \notin [S_-,S_+]}}{\mathbb{P}\{i \mid i\}}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}(\overline{X}(t) = i \ \cap\ \overline{X}(t_1) = 1 \ \cap\ \overline{X}(t_2) = 1) \\
&= \mathbb{P}(\overline{X}(t) \neq 0 \ \cap\ X(t) = i \ \cap\ X(t_1) = 1 \ \cap\ X(t_2) = 1) \\
&= \mathbb{P}\{i \mid i\}\ \mathbb{P}(X(t) = i \ \cap\ X(t_1) = 1 \ \cap\ X(t_2) = 1) \\
&= \frac{1}{2}\,\mathbb{P}\{1 \mid 1\}\ \bar{r}(t_1,t)\,\bar{r}(t,t_2) 1_{i=1} \ + \frac{1}{2}\,\mathbb{P}\{-1 \mid -1\}\ r(t_1,t)\,r(t,t_2) 1_{i=-1} \ .
\end{aligned}$$

It comes, using formula (3),

$$\begin{aligned}
&\mathbb{P}(Y \in [y\ ,\ y+dy] \ \cap\ \overline{X}(t_1) = 1 \ \cap\ \overline{X}(t_2) = 1) \\
&= \frac{1}{2}\ f_{(\mu+q,\sigma)}(y)\ 1_{y \notin [S_-,S_+]}\ \bar{r}(t_1,t_2)\ Q_t^{1,1} \ + \ \frac{1}{2}\ f_{(\mu-q,\sigma)}(y)\ 1_{y \notin [S_-,S_+]}\ \bar{r}(t_1,t_2)\ Q_t^{-1,-1} \ .
\end{aligned}$$

In the same way, after some calculations, we find

$$\begin{aligned}
&\mathbb{P}(Y \in [y\ ,\ y+dy] \ \cap\ \overline{X}(t_1) = 1 \ \cap\ \overline{X}(t_2) = -1) \\
&= \frac{1}{2}\ f_{(\mu+q,\sigma)}(y)\ 1_{y \notin [S_-,S_+]}\ r(t_1,t_2)\ Q_t^{1,-1} \ + \ \frac{1}{2}\ f_{(\mu-q,\sigma)}(y)\ 1_{y \notin [S_-,S_+]}\ r(t_1,t_2)\ Q_t^{-1,1} \ ,
\end{aligned}$$

$$\begin{aligned}
&\mathbb{P}(Y \in [y\ ,\ y+dy] \ \cap\ \overline{X}(t_1) = -1 \ \cap\ \overline{X}(t_2) = -1) \\
&= \frac{1}{2}\ f_{(\mu+q,\sigma)}(y)\ 1_{y \notin [S_-,S_+]}\ r(t_1,t_2)\ Q_t^{-1,1} \ + \ \frac{1}{2}\ f_{(\mu-q,\sigma)}(y)\ 1_{y \notin [S_-,S_+]}\ r(t_1,t_2)\ Q_t^{1,-1} \ ,
\end{aligned}$$

$$\mathbb{P}(Y \in [y\ ,\ y+dy]\ \cap\ \overline{X}(t_1) = -1\ \cap\ \overline{X}(t_2) = -1)$$
$$= \frac{1}{2}\ f_{(\mu+q,\sigma)}(y)\ 1_{y\notin[S_-,S_+]}\ \overline{r}(t_1,t_2)\ Q_t^{-1,-1}\ +\ \frac{1}{2}\ f_{(\mu-q,\sigma)}(y)\ 1_{y\notin[S_-,S_+]}\ \overline{r}(t_1,t_2)\ Q_t^{1,1}\ .$$

Finally, when the genome information is missing at marker locations (ie the phenotype is not extreme), we find

$$\mathbb{P}(Y \in [y\ ,\ y+dy]\ \cap\ \overline{X}(t_1) = 0\ \cap\ \overline{X}(t_2) = 0)$$
$$= \frac{1}{2}\ f_{(\mu+q,\sigma)}(y)\ 1_{y\notin[S_-,S_+]}\ +\ \frac{1}{2}\ f_{(\mu-q,\sigma)}(y)\ 1_{y\notin[S_-,S_+]}\ .$$

Let's define the quantity $p(t)$ such as

$$p(t) = Q_t^{1,1}\ 1_{\overline{X}(t_1)=1}1_{\overline{X}(t_2)=1}\ +\ Q_t^{1,-1}\ 1_{\overline{X}(t_1)=1}1_{\overline{X}(t_2)=-1}$$
$$+\ Q_t^{-1,1}\ 1_{\overline{X}(t_1)=-1}1_{\overline{X}(t_2)=1}\ +\ Q_t^{-1,-1}\ 1_{\overline{X}(t_1)=-1}1_{\overline{X}(t_2)=-1} \qquad (4)$$

and let $\theta = (q,\ \mu,\ \sigma)$ be the parameter of the model at t fixed. It comes, the likelihood of the triplet $\big(Y,\ \overline{X}(t_1),\ \overline{X}(t_2)\big)$ with respect to the measure $\lambda\otimes N\otimes N$, λ being the Lebesgue measure, N the counting measure on $\mathbb{N}$, is $\forall t \in [t_1,t_2]$:

$$L_t(\theta) = \big[\ p(t)\ f_{(\mu+q,\sigma)}(y)1_{y\notin[S_-,S_+]} + \{1-p(t)\}\ f_{(\mu-q,\sigma)}(y)1_{y\notin[S_-,S_+]} \qquad (5)$$
$$+\ \frac{1}{2}\ f_{(\mu+q,\sigma)}(y)1_{y\in[S_-,S_+]}\ +\ \frac{1}{2}\ f_{(\mu-q,\sigma)}(y)1_{y\in[S_-,S_+]}\Big]\ g(t)$$

where the function

$$g(t) = \frac{1}{2}\Big\{\overline{r}(t_1,t_2)\ 1_{\overline{X}(t_1)=1}1_{\overline{X}(t_2)=1}\ +\ r(t_1,t_2)\ 1_{\overline{X}(t_1)=1}1_{\overline{X}(t_2)=-1}\Big\}$$
$$+\ \frac{1}{2}\Big\{r(t_1,t_2)\ 1_{\overline{X}(t_1)=-1}1_{\overline{X}(t_2)=1}\ +\ \overline{r}(t_1,t_2)\ 1_{\overline{X}(t_1)=-1}1_{\overline{X}(t_2)=-1}\Big\}$$
$$+\ 1_{\overline{X}(t_1)=0}1_{\overline{X}(t_2)=0}$$

can be removed because it does not depend on the parameters. Note that for $t = t^\star$, we find our formula (2) of the introduction where $p(t^\star)$ is described in formula (4).

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

Notations : $\mathcal{A}$ is the quantity such as $\mathcal{A} = \sigma^2\ \big\{\gamma\ +\ z_{\gamma_+}\ \varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})\big\}$, where $\varphi(x)$ and z_α denote respectively the density of a standard normal distribution taken at the point x, and the quantile of order $1-\alpha$ of a standard normal distribution.

Our main result is the following

THEOREM 1. *Suppose that the parameters (q,μ,σ^2) vary in a compact and that σ^2 is bounded away from zero. Let H_0 be the null hypothesis $q = 0$ and define the following local alternative*

$H_{at^\star}$: *"the QTL is located at the position $t^\star$ with effect $q = a/\sqrt{n}$ where $a \neq 0$ ".*

With the previous defined notations,

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

as n tends to infinity, under H_0 and $H_{at^\star}$ where :

- $\Rightarrow$ *is the weak convergence,* $\xrightarrow{F.d.}$ *is the convergence of finite-dimensional distributions and* $\xrightarrow{\mathcal{L}}$ *is the convergence in distribution*
- $V(.)$ *is the Gaussian process with unit variance such as :*

$$V(t) = \frac{\alpha(t)V(t_1) + \beta(t)V(t_2)}{\sqrt{\mathbb{V}\left\{\alpha(t)V(t_1) + \beta(t)V(t_2)\right\}}} \tag{6}$$

where

$$\mathrm{Cov}\left\{\mathrm{V}(\mathrm{t}_1), \mathrm{V}(\mathrm{t}_2)\right\} = \rho(\mathrm{t}_1, \mathrm{t}_2) = \exp(-2|\mathrm{t}_1 - \mathrm{t}_2|)$$
$$\alpha(t) = Q_t^{1,1} - Q_t^{-1,1} \quad , \quad \beta(t) = Q_t^{1,1} - Q_t^{1,-1}$$

and with expectation :

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

$$m_{t^\star}(t) = \frac{\alpha(t)\ m_{t^\star}(t_1) + \beta(t)\ m_{t^\star}(t_2)}{\sqrt{\mathbb{V}\left\{\alpha(t)V(t_1) + \beta(t)V(t_2)\right\}}}$$

where

$$m_{t^\star}(t_1) = \frac{a\ \sqrt{\mathcal{A}}\ \rho(t_1, t^\star)}{\sigma^2} \quad , \quad m_{t^\star}(t_2) = \frac{a\ \sqrt{\mathcal{A}}\ \rho(t^\star, t_2)}{\sigma^2} \ .$$

In the sense of this equation, $V(.)$ will be called a "non linear normalized interpolated process". We can see that under the null hypothesis, despite the selective genotyping, $V(.)$ is exactly the same process as the process $Z(.)$ of Theorem 1 of Azaïs et al. [2011] obtained for the oracle situation. However, under the alternative, the mean functions of the two processes are not the same anymore : the mean functions are proportional of a factor $\sqrt{\mathcal{A}}/\sigma$. Note also that $V(.)$ is the generalization of $Z(.)$. Indeed, if we choose $S_- = S_+$, that is to say we genotype all the individuals, the factor $\sqrt{\mathcal{A}}/\sigma$ is equal to 1, and $V(.)$ is the same process as $Z(.)$.

Proof of Theorem 1 :

Fisher Information Matrix
Let $l_t(\theta)$ be the loglikelihood. We first compute the Fisher information at a point θ_0 that belongs to H_0. We have

$$\frac{\partial l_t}{\partial q}\ |_{\theta_0} = \frac{y - \mu}{\sigma^2}\ \{2p(t) - 1\}\ 1_{y \notin [S_-, S_+]} \tag{7}$$

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

Then,

$$\mathbb{E}_{H_0}\left\{\left(\frac{\partial l_t}{\partial q}\mid_{\theta_0}\right)^2\right\} = \mathbb{E}_{H_0}\left\{\left(\frac{y-\mu}{\sigma^2}\right)^2 \{2p(t)-1\}^2 1_{y\notin[S_-,S_+]}\right\} \tag{8}$$

Let's introduce a key lemma : a version of Lemma 2 of Azaïs et al. [2011] adapted for selective genotyping.

LEMMA 1. *We have the following relationship :*

$$\{2p(t)-1\}\, 1_{y\notin[S_-,S_+]} = \left\{\alpha(t)\overline{X}(t_1) + \beta(t)\overline{X}(t_2)\right\} 1_{y\notin[S_-,S_+]}$$

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

To prove the lemma, use formula (4) and check that both sides coincide when $y \notin [S_-, S_+]$.

It comes

$$\begin{aligned}
&\mathbb{E}_{H_0}\left\{\left(\frac{\partial l_t}{\partial q}\mid_{\theta_0}\right)^2\right\} \qquad (9)\\
&= \mathbb{E}_{H_0}\left[\left(\frac{y-\mu}{\sigma^2}\right)^2 \left\{\alpha(t)\overline{X}(t_1) + \beta(t)\overline{X}(t_2)\right\}^2 1_{y\notin[S_-,S_+]}\right]\\
&= \mathbb{E}_{H_0}\left\{\left(\frac{y-\mu}{\sigma^2}\right)^2 1_{y\notin[S_-,S_+]}\right\}\ \mathbb{E}_{H_0}\left[\left\{\alpha(t)X(t_1) + \beta(t)X(t_2)\right\}^2\right]\\
&= \mathcal{A}\left\{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)e^{-2(t_2-t_1)}\right\}/\sigma^4 \, .
\end{aligned}$$

To conclude, after some calculations, we find

$$I_{\theta_0} = Diag\left[\mathcal{A}\left\{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)e^{-2(t_2-t_1)}\right\}/\sigma^4 \, , \, \frac{1}{\sigma^2} \, , \, \frac{2}{\sigma^2}\right] \quad . \tag{10}$$

Only the computation of $\mathbb{E}_{H_0}\left\{-\frac{\partial l_t}{\partial q \partial \mu}\mid_{\theta_0}\right\}$ and $\mathbb{E}_{H_0}\left\{-\frac{\partial l_t}{\partial q \partial \sigma}\mid_{\theta_0}\right\}$, were not easy. Let's prove now why these two terms are equal to zero. We have

$$\frac{\partial l_t}{\partial q \partial \mu}\mid_{\theta_0} = -\frac{2p(t)-1}{\sigma^2} 1_{y\notin[S_-,S_+]} \quad .$$

It comes, using Lemma 1,

$$\begin{aligned}
\mathbb{E}_{H_0}\left\{-\frac{\partial l_t}{\partial q \partial \mu}\mid_{\theta_0}\right\} &= \frac{1}{\sigma^2}\, \mathbb{E}_{H_0}\left[\left\{\alpha(t)\overline{X}(t_1) + \beta(t)\overline{X}(t_2)\right\} \mid y \notin [S_-, S_+]\ \right]\ \mathbb{P}_{H_0}(y \notin [S_-, S_+])\\
&= \frac{1}{\sigma^2}\, \mathbb{E}_{H_0}\left\{\alpha(t)X(t_1) + \beta(t)X(t_2)\right\}\ \mathbb{P}_{H_0}(y \notin [S_-, S_+]) = 0 \, .
\end{aligned}$$

Besides,

$$\frac{\partial l_t}{\partial q \partial \sigma} \mid_{\theta_0} = -\frac{2}{\sigma^3}(y-\mu)\left\{2p(t)-1\right\} 1_{y \notin [S_-,S_+]} \ .$$

It comes

$$\begin{aligned} &\mathbb{E}_{H_0}\left(\frac{\partial l_t}{\partial q \partial \sigma} \mid_{\theta_0}\right) \\ &= -\frac{2}{\sigma^3}\, \mathbb{E}_{H_0}\left\{(y-\mu) 1_{y \notin [S_-,S_+]}\right\}\, \mathbb{E}_{H_0}\left\{\alpha(t)X(t_1)+\beta(t)X(t_2)\right\} = 0 \ . \end{aligned} \tag{11}$$

It concludes the proof for the Fisher Information matrix.

Study of the score process under H_0
Since the Fisher Information matrix is diagonal, the score statistic of the hypothesis "$q = 0$" will be defined as

$$S_n(t) = \frac{\frac{\partial l_t^n}{\partial q} \mid_{\theta_0}}{\sqrt{\mathbb{V}\left(\frac{\partial l_t^n}{\partial q} \mid_{\theta_0}\right)}} \ .$$

Now using (7) and using Lemma 1, it is clear that

$$\begin{aligned} \frac{\partial l_t^n}{\partial q} \mid_{\theta_0} &= \sum_{j=1}^{n} \frac{y_j-\mu}{\sigma^2}\ \left\{2p_j(t)-1\right\}\ 1_{y_j \notin [S_-,S_+]} \\ &= \frac{\alpha(t)}{\sigma}\sum_{j=1}^{n} \varepsilon_j\ \overline{X}_j(t_1)\ 1_{y_j \notin [S_-,S_+]}\ +\ \frac{\beta(t)}{\sigma}\sum_{j=1}^{n}\ \varepsilon_j\ \overline{X}_j(t_2)\ 1_{y_j \notin [S_-,S_+]} \end{aligned} \tag{12}$$

this proves that $V(.)$ is a non linear interpolated process.

On the other hand, according to formula (4) (with $p = 1/2$) of Section 8.3.1 of Rabier [2011], $\forall k = 1, 2$:

$$\begin{aligned} S_n(t_k) &= \sum_{j=1}^{n} \frac{\sigma \varepsilon_j\ \overline{X}_j(t_k)\ 1_{y_j \notin [S_-,S_+]}}{\sqrt{n\ \mathcal{A}}} \\ S_n(t_k) &\xrightarrow{\mathcal{L}} N(0,\ 1) \end{aligned}$$

Let's compute the covariance of the score statistics on markers, ie. $\mathrm{Cov}\left\{\mathrm{S_n(t_1), S_n(t_2)}\right\}$. Since $\mathbb{E}\left\{(y-\mu)^2 1_{y \notin [S_-,S_+]}\right\} = \mathcal{A}$ (see our lemma), we have :

$$\begin{aligned} \mathbb{E}\left\{S_n(t_1)S_n(t_2)\right\} &= \frac{1}{\mathcal{A}}\, \mathbb{E}\left\{(y-\mu)^2\ \overline{X}(t_1)\ \overline{X}(t_2)\ 1_{y \notin [S_-,S_+]}\right\} \\ &= \frac{1}{\mathcal{A}}\, \mathbb{E}\left\{(y-\mu)^2\ X(t_1)\ X(t_2) \mid y \notin [S_-,S_+]\right\} \mathbb{P}(y \notin [S_-,S_+]) \\ &= \frac{1}{\mathcal{A}}\, \mathbb{E}\left\{(y-\mu)^2 1_{y \notin [S_-,S_+]}\right\} \mathbb{E}\left\{X(t_1)X(t_2)\right\} = e^{-2(t_2-t_1)} \end{aligned}$$

As a consequence, $\mathrm{Cov}\left\{\mathrm{S_n(t_1), S_n(t_2)}\right\} = \mathrm{e}^{-2(\mathrm{t_2}-\mathrm{t_1})}$. The weak convergence of the score process, $S_n(.)$, is then a direct consequence of (12), the convergence of $(S_n(t_1), S_n(t_2))$ and the Continuous Mapping Theorem.

Study under the local alternative

Let us consider a local alternative defined by t^* and $q = a/\sqrt{n}$.

It remains to compute the asymptotic distribution of $S_n(t)$ under this alternative. Indeed, under the alternative

$$\begin{aligned} S_n(t_k) &= \sum_{j=1}^{n} \frac{(y_j - \mu)\ \overline{X}_j(t_k)\ 1_{y_j \notin [S_-, S_+]}}{\sqrt{n\ \mathcal{A}}} \\ &= \sum_{j=1}^{n} \frac{q\overline{X}_j(t^\star)\ \overline{X}_j(t_k)\ 1_{y_j \notin [S_-, S_+]}}{\sqrt{n\ \mathcal{A}}} \ + \ \sum_{j=1}^{n} \frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)\ 1_{y_j \notin [S_-, S_+]}}{\sqrt{n\ \mathcal{A}}} \quad . \end{aligned}$$

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

$$\begin{aligned} &\mathbb{E}\left\{\sigma\varepsilon\ \overline{X}(t_k)\ 1_{y \notin [S_-, S_+]}\right\} \qquad (13) \\ &= \frac{1}{2}\ \mathbb{E}\left\{\sigma\varepsilon\ \overline{X}(t_k)\ 1_{y \notin [S_-, S_+]} \mid X(t^\star) = 1\right\} \ + \ \frac{1}{2}\ \mathbb{E}\left\{\sigma\varepsilon\ \overline{X}(t_k)\ 1_{y \notin [S_-, S_+]} \mid X(t^\star) = -1\right\} \quad . \end{aligned}$$

We have

$$\begin{aligned} &\mathbb{E}\left\{\sigma\varepsilon\ \overline{X}(t_k)\ 1_{y \notin [S_-, S_+]} \mid X(t^\star) = 1\right\} \\ &= \mathbb{E}\left\{\sigma\varepsilon\ 1_{y \notin [S_-, S_+]} \mid X(t^\star) = 1\right\} \bar{r}(t_k, t^\star) \ - \ \mathbb{E}\left\{\sigma\varepsilon\ 1_{y \notin [S_-, S_+]} \mid X(t^\star) = 1\right\} r(t_k, t^\star) \\ &= \mathbb{E}\left\{\sigma\varepsilon\ 1_{y \notin [S_-, S_+]} \mid X(t^\star) = 1\right\}\ e^{-2|t_k - t^\star|} \quad . \end{aligned}$$

Besides, according to iv) of Lemma 3 introduced in Section 8.2.2 of Rabier [2011],

$$\mathbb{E}\left\{\sigma\varepsilon\ 1_{y \notin [S_-, S_+]} \mid X(t^\star) = 1\right\} = \sigma\varphi\left(\frac{S_+ - \mu - q}{\sigma}\right) - \sigma\varphi\left(\frac{S_- - \mu - q}{\sigma}\right) \quad .$$

It comes

$$\mathbb{E}\left\{\sigma\varepsilon\ \overline{X}(t_k)\ 1_{y \notin [S_-, S_+]} \mid X(t^\star) = 1\right\} = e^{-2|t_k - t^\star|}\ \sigma\ \left\{\varphi\left(\frac{S_+ - \mu - q}{\sigma}\right) - \varphi\left(\frac{S_- - \mu - q}{\sigma}\right)\right\} \quad .$$

In the same way, after some calculations, we obtain

$$\mathbb{E}\left\{\sigma\varepsilon\ \overline{X}(t_k)\ 1_{y \notin [S_-, S_+]} \mid X(t^\star) = 1\right\} = -e^{-2|t_k - t^\star|}\ \sigma\ \left\{\varphi\left(\frac{S_+ - \mu + q}{\sigma}\right) - \varphi\left(\frac{S_- - \mu + q}{\sigma}\right)\right\} \quad .$$

Finally, since we consider q small, using Taylor linearization as at the top of page 21 of Rabier [2011], we obtain

$$\mathbb{E}\left\{\sigma\varepsilon\ \overline{X}(t_k)\ 1_{y \notin [S_-, S_+]}\right\} = e^{-2|t_k - t^\star|}\ q\ \left\{z_{\gamma^+}\varphi(z_{\gamma^+}) - z_{1-\gamma^-}\varphi(z_{1-\gamma^-})\right\} \ + \ o(q) \quad .$$

It comes

$$\mathbb{E}\left\{\sum_{j=1}^{n} \frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)\ 1_{y_j \notin [S_-, S_+]}}{\sqrt{n\ \mathcal{A}}}\right\} \to \frac{e^{-2|t_k - t^\star|}}{\sqrt{\mathcal{A}}}\ a\ \left\{z_{\gamma^+}\varphi(z_{\gamma^+}) - z_{1-\gamma^-}\varphi(z_{1-\gamma^-})\right\} \quad . \qquad (14)$$

We have now just to remark that

$$\begin{aligned}
&\mathbb{E}\left(\left\{\sigma\varepsilon\ \overline{X}(t_k)\ 1_{y\notin[S_-,S_+]}\right\}^2\right) = \mathbb{E}\left\{\sigma^2\ \varepsilon^2\ 1_{y\notin[S_-,S_+]}\right\} \\
&= \mathbb{E}\left\{\sigma^2\ \varepsilon^2\ 1_{y\notin[S_-,S_+]} \mid X(t^\star)=1\right\}/2 + \mathbb{E}\left\{\sigma^2\ \varepsilon^2\ 1_{y\notin[S_-,S_+]} \mid X(t^\star)=-1\right\}/2 \\
&= \mathcal{A}/2\ +\ \mathcal{A}/2 = \mathcal{A}\ .
\end{aligned}$$

It comes

$$\mathbb{V}\left\{\sum_{j=1}^{n} \frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)\ 1_{y_j\notin[S_-,S_+]}}{\sqrt{n\ \mathcal{A}}}\right\} \to \mathcal{A}\ ,$$

and according to the Central Limit Theorem

$$\sum_{j=1}^{n} \frac{\sigma\varepsilon_j\ \overline{X}_j(t_k)\ 1_{y_j\notin[S_-,S_+]}}{\sqrt{n\ \mathcal{A}}} \xrightarrow{\mathcal{L}} N\left[\frac{e^{-2|t_k-t^\star|}}{\sqrt{\mathcal{A}}}\ a\ \left\{z_{\gamma^+}\varphi(z_{\gamma^+}) - z_{1-\gamma^-}\varphi(z_{1-\gamma^-})\right\},\ 1\right]\ . \tag{15}$$

Besides,

$$\begin{aligned}
&\mathbb{E}\left\{\overline{X}(t^\star)\ \overline{X}(t_k)\ 1_{y\notin[S_-,S_+]}\right\} \qquad (16)\\
&= \frac{1}{2}\ \mathbb{P}\left\{1 \mid 1\right\}\ \mathbb{E}\left\{X(t_k) \mid \overline{X}(t^\star)=1\right\}\ -\ \frac{1}{2}\ \mathbb{P}\left\{-1 \mid -1\right\}\ \mathbb{E}\left\{X(t_k) \mid \overline{X}(t^\star)=-1\right\} \\
&= \frac{1}{2}\ e^{-2|t_k-t^\star|}\left(\mathbb{P}\left\{1 \mid 1\right\}\ +\ \mathbb{P}\left\{-1 \mid -1\right\}\right)\ .
\end{aligned}$$

We have (see formula for $\mathbb{P}\{1 \mid 1\}$ at the top of page 21 of Rabier [2011])

$$\mathbb{P}\left\{1 \mid 1\right\}\ +\ \mathbb{P}\left\{-1 \mid -1\right\} = 2 + 2\ \Phi\left(\frac{S_- - \mu}{\sigma}\right) - 2\ \Phi\left(\frac{S_+ - \mu}{\sigma}\right)\ +\ o(q)\ .$$

It comes

$$\begin{aligned}
&\mathbb{E}\left\{\overline{X}(t^\star)\ \overline{X}(t_k)\ 1_{y\notin[S_-,S_+]}\right\} \quad = e^{-2|t_k-t^\star|}\left\{1 + \Phi\left(\frac{S_- - \mu}{\sigma}\right) + \Phi\left(\frac{S_+ - \mu}{\sigma}\right)\right\} + o(q) \\
&= e^{-2|t_k-t^\star|}\ \gamma\ +\ o(q)\ .
\end{aligned}$$

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

$$\sum_{j=1}^{n} \frac{q\overline{X}_j(t^\star)\ \overline{X}_j(t_k)\ 1_{y_j\notin[S_-,S_+]}}{\sqrt{n\ \mathcal{A}}} \to e^{-2|t_k-t^\star|}\ \gamma\ . \tag{17}$$

Finally, using formulae (15) and (17), we obtain

$$S_n(t_k) \xrightarrow{\mathcal{L}} N(\frac{a\sqrt{\mathcal{A}}}{\sigma^2}e^{-2|t_k-t^\star|},\ 1)\ , \tag{18}$$

which concludes the proof.

Study of the supremum of the LRT process

Let $l_t^n(\theta)$ be the log likelihood for n observations log likelihood. Let $l_t^n(\widehat{\theta})$ be the maximized log likelihood and let $l_t^n(\widehat{\theta}_{|H_0})$ be the maximized log likelihood under H_0, with $\widehat{\theta}_{|H_0} = \big(0, \overline{Y} = \sum Y_j/n, 1/n\sum(Y_j - \overline{Y})^2\big)$ (the genetic markers are useless under H_0). The likelihood ratio statistics will be defined as

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

on n independent observations.

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

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

under the null hypothesis. Our goal is now to prove that the rest above is uniform in t.

Let us consider now t as an extra parameter. Let t^*, θ^* be the true parameter that will be assumed to belong to H_0. Note that t^* makes no sense. It is easy to check that at H_0 the Fisher information relative to t is zero so that the model is not regular.

It can be proved that assumptions 1, 2 and 3 of Azaïs et al. [2009] holds. So, we can apply Theorem 1 of Azaïs et al. [2009] and we have

$$\sup_{(t,\theta)} l_t(\theta) - l_{t^*}(\theta^*) = \sup_{d\in\mathcal{D}} \left[\left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} d(X_j) \right\}^2 1_{d(X_j)\geqslant 0} \right] + o_P(1) \tag{19}$$

where the observation X_j stands for $Y_j, \overline{X}_j(t_1), \overline{X}_j(t_2)$ and where $\mathcal{D}$ is the set of scores defined in Azaïs et al. [2009], see also Gassiat [2002]. A similar result is true under H_0 with a set $\mathcal{D}_0$. Let us precise the sets of scores $\mathcal{D}$ and $\mathcal{D}_0$. This sets are defined at the sets of scores of one parameter families that converge to the true model p_{t^*,θ^*} and that are differentiable in quadratic mean.

It is easy to see that

$$\mathcal{D} = \left\{ \frac{\langle V, l_t'(\theta^*)\rangle}{\mathbb{V}(\langle V, l_t'(\theta^*)\rangle)}, V \in \mathbb{R}^3, t \in [t_1, t_2] \right\}$$

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

$$\mathcal{D}_0 = \left\{ \frac{\langle V, l_t'(\theta^*)\rangle}{\mathbb{V}(\langle V, l_t'(\theta^*)\rangle)}, V \in \mathbb{R}^2 \right\},$$

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

Using the transform $V \to -V$ in the expressions of the sets of score, we see that the indicator function can be removed in (19). Then, since the Fisher

information matrix is diagonal (see formula (10)) , it is easy to see that

$$\sup_{d\in\mathcal{D}}\left[\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}d(X_j)\right\}^2\right]-\sup_{d\in\mathcal{D}_0}\left[\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}d(X_j)\right\}^2\right]$$
$$=\sup_{t\in[t_1,t_2]}\left(\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\frac{\partial l_t}{\partial q}(X_j)\mid_{\theta_0}}{\sqrt{\mathbb{V}\left\{\frac{\partial l_t}{\partial q}(X_j)\mid_{\theta_0}\right\}}}\right]^2\right).$$

This is exactly the desired result. Note that the model with t^* fixed is differentiable in quadratic mean, this implies that the alternative defines a contiguous sequence of alternatives. By Le Cam's first Lemma, relation (19) remains true under the alternative.

Remark : According to the law of large numbers, under the null hypothesis H_0 and under the local alternative $H_{at^\star}$, $\frac{1}{n}\sum 1_{y_j\notin[S_+,S_-]}\to\gamma$. So, γ corresponds asymptotically to the percentage of individuals genotyped. In the same way, γ_+ (resp. γ_-) corresponds asymptotically to the percentage of individuals genotyped with the largest (resp. the smallest) phenotypes.

3. An easy way to perform the statistical test

Since $V(.)$ is a "non linear normalized interpolated process", we can use Lemma 1 of Azaïs et al. [2011] in order to compute easily the supremum of $V^2(.)$. Note that this Lemma is suitable here because we have exactly the same interpolation as in Theorem 1 of Azaïs et al. [2011]. It comes

$$\max_{t\in[t_1,t_2]}\frac{\{\alpha(t)V(t_1)+\beta(t)V(t_2)\}^2}{\alpha^2(t)+\beta^2(t)+2\rho(t_1,t_2)\alpha(t)\beta(t)}$$
$$=\max\left(V^2(t_1)\ ,\ V^2(t_2)\ ,\ \frac{V^2(t_1)+V^2(t_2)-2\rho(t_1,t_2)V(t_1)V(t_2)}{1-\rho^2(t_1,t_2)}1_{\frac{V(t_2)}{V(t_1)}\in\ ]\ \rho(t_1,t_2)\ ,\ \frac{1}{\rho(t_1,t_2)}\ [}\right). \tag{20}$$

Note that since under H_0, the process $V(.)$ is exactly the same process as $W(.)$, we will have exactly the same threshold if we are under selective genotyping or not. So, the Monte-Carlo Quasi Monte-Carlo method of Azaïs et al. [2006] and based on Genz [1992], is still suitable here.

Let's focus now on the data analysis. Which test statistic should we use in order to make the data analysis easy ? It is well known that under selective genotyping, when we focus only on one location of the genome which is a marker location, performing a LRT or a Wald test is time consuming : an EM algorithm is required to obtain the maximum likelihood estimators. In Rabier [2011], we propose a very easy test which is almost a comparison of means and which has the same asymptotic properties as LRT and Wald tests. So, the idea now is to adapt this comparison of means to our problem which focus on the whole chromosome.

As a consequence, $\forall k = 1, 2$, let define now the test statistic $T_n(t_k)$ such as

$$T_n(t_k) = \frac{\sum_{j=1}^n (Y_j - \overline{Y})\ \overline{X}_j(t_k)\ 1_{Y_j \notin [S_-, S_+]}}{\sqrt{\sum_{j=1}^n (Y_j - \overline{Y})^2\ 1_{Y_j \notin [S_-, S_+]}}} \ .$$

We introduce the following lemma.

LEMMA 2. *Let $T_n(.)$ be the process such as*

$$T_n(t) = \frac{\alpha(t) T_n(t_1) + \beta(t) T_n(t_2)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\rho(t_1, t_2)\alpha(t)\beta(t)}} \quad , \textit{then } T_n(.) \Rightarrow V(.) \quad \textit{and} \quad T_n^2(.) \Rightarrow V^2(.) \ .$$

Then, for the data analysis, we just have to consider as a test statistic $\sup T_n^2(.)$, which can be obtained easily using formula (20) and replacing $V(t_1)$ and $V(t_2)$ by respectively $T_n(t_1)$ and $T_n(t_2)$. Note that, according to Lemma 2, this test has the same asymptotic properties as the test based on the test statistic $\sup \Lambda_n(.)$, which corresponds to a LRT on the whole chromosome.

On other hand, a consequence of Lemma 2 is that the non extreme phenotypes (for which the genotypes are missing) don't bring any information for statistical inference. Indeed, our test statistics $T_n(t)$ are based only on the extreme phenotypes, as soon as we replace the empirical mean $\overline{Y}$ by $\hat{\mu}$, an estimator $\sqrt{n}$ consistent based only on the extreme phenotypes ($\hat{\mu}$ can be obtained by the method of moments for instance).

Proof of Lemma 2 :

For $k = 1, 2$, we define $\tilde{T}(t_k)$ such as

$$\tilde{T}_n(t_k) = \frac{\sum_{j=1}^n (Y_j - \overline{Y})\ \overline{X}_j(t_k)\ 1_{Y_j \notin [S_-, S_+]}}{\sqrt{n\ \mathcal{A}}} \ .$$

To begin, in order to make the proof easier, let's consider that we are under H_0. Since $\overline{Y} = \mu + O_P\left(1/\sqrt{n}\right)$, we have

$$\tilde{T}_n(t_k) = \frac{\sum_{j=1}^n (Y_j - \mu)\ \overline{X}_j(t_k)\ 1_{Y_j \notin [S_-, S_+]}}{\sqrt{n\ \mathcal{A}}} \ + \ O_P\left(\frac{1}{\sqrt{n}}\right) \frac{\sum_{j=1}^n \overline{X}_j(t_k)\ 1_{Y_j \notin [S_-, S_+]}}{\sqrt{n\ \mathcal{A}}} \ .$$

Let's focus on the second term. We have

$$\begin{aligned} \mathbb{E}\left[\overline{X}(t_k)\ 1_{Y \notin [S_-, S_+]}\right] &= \mathbb{E}\left[\overline{X}(t_k) \mid Y \notin [S_-, S_+]\right]\ \mathbb{P}(Y \notin [S_-, S_+]) \\ &= \mathbb{E}\left[X(t_k)\right]\ \gamma \ \ = \ \ 0 \ . \end{aligned}$$

By Prohorov, it comes $\sum_{j=1}^n \overline{X}_j(t_k)\ 1_{Y_j \notin [S_-, S_+]} = O_P\left(1/\sqrt{n}\right)$.
It comes $\tilde{T}_n(t_k) = S_n(t_k) + O_P\left(1/\sqrt{n}\right)$ and as a consequence $\tilde{T}_n(t_k) = S_n(t_k) + o_P(1)$. As said before, the model with t^* fixed is differentiable in quadratic mean, this implies that the alternative defines a contiguous sequence of alternatives. By Le Cam's first Lemma, the remainder converges also to 0 in probability under the alternative.
So, if we apply the Multivariate Central Limit Theorem, we have now $\left(\tilde{T}_n(t_1), \tilde{T}_n(t_2)\right) \xrightarrow{\mathcal{L}}$ $(V(t_1), V(t_2))$ whatever the hypothesis. We set in addition

$$\hat{\mathcal{A}} = \frac{1}{n} \sum_{j=1}^n (Y_j - \overline{Y})^2\ 1_{Y_j \notin [S_-, S_+]} \ .$$

We have the relationship $(T_n(t_1), T_n(t_2)) = \sqrt{\frac{\mathcal{A}}{\hat{\mathcal{A}}}}\left(\tilde{T}_n(t_1), \tilde{T}_n(t_2)\right)$. Since $\hat{\mathcal{A}} \xrightarrow{\mathcal{L}} \mathcal{A}$ whatever the hypothesis (cf. Rabier [2011]), according to Slutsly and then Continuous Mapping theorem, we have $\sqrt{\frac{\mathcal{A}}{\hat{\mathcal{A}}}} \xrightarrow{\mathcal{L}} 1$. Using Slutsky, it comes $(T_n(t_1), T_n(t_2)) \xrightarrow{\mathcal{L}} (V(t_1), V(t_2))$. To conclude the proof, we just have to use the Continuous Mapping theorem : $T_n(.) \Rightarrow V(.)$ and obviously $T_n^2(.) \Rightarrow V^2(.)$.

4. Several markers : the "Interval Mapping" of Lander and Botstein [1989] under selective genotyping

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).

THEOREM 2. *We have the same result as in Theorem 1, provided that we make some adjustments and that we redefine $V(.)$ 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 $V(.)$ 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, $V(.)$ is obtained from $V(t^\ell)$ and $V(t^r)$ by interpolation and normalization using the functions $\alpha(t)$ and $\beta(t)$*
- *at the marker positions, the expectation is such as $m_{t^\star}(t_k) = \frac{a\sqrt{\mathcal{A}}\,\rho(t_k, t^\star)}{\sigma^2}$*
- *at other positions, the expection is obtained from $m_{t^\star}(t^\ell)$ and $m_{t^\star}(t^r)$ by interpolation and normalization using the functions $\alpha(t)$ and $\beta(t)$.*

The proof of the theorem is the same the proof of Theorem 1 since for a position t, we can limit our attention to the interval (t^ℓ, t^r). Note that it is due to Haldane model with Poisson increments. Another key point for the proof, is that when $t^\star$ does not belong to the marker interval (t^ℓ, t^r), we can still use the section "Study under the alternative" of the proof of Theorem 1.

Another important point is that since for a position t we can limit our attention to the interval (t^ℓ, t^r), Lemma 2 and formula (20) are still true here. We just have to replace t_1 and t_2 by t^ℓ and t^r in order to have the good expressions. As a consequence, we can easily compute $\sup T_n^2(.)$.
We introduce now our Theorem 3.

THEOREM 3. *Let κ be the Asymptotic Relative Efficiency (ARE) with respect to the oracle situation where all the genotypes are known. Then, we have*

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

The proof is obvious since the mean functions of the selective genotyping and the oracle situation, are proportional of a factor $\sqrt{\mathcal{A}}/\sigma$. Besides, in order to proove that the maximum is reached for $\gamma_+ = \gamma_- = \gamma/2$, we just have to use Theorem 1 of Rabier [2011].

According to i) of Theorem 3, we have exactly the same ARE as the ARE obtained in Theorem 1 of Rabier [2011]. In other words, we have exactly the same ARE if we look for a QTL on a chromosome (as here) and if we look for a QTL only on one given genetic marker (as in Rabier [2011]). Besides, ii) of Theorem 3 says that 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. It confirms by the theory what geneticists do in practice.

5. Applications

In this Section, we propose to illustrate the theoretical results obtained in this paper. For all the following applications, we will consider statistical tests at the 5% level. If we call

$$h_n(t_k,t_{k+1}) = \frac{T_n^2(t_k)+T_n^2(t_{k+1})-2\rho(t_k,t_{k+1})T_n(t_k)T_n(t_{k+1})}{1-\rho^2(t_k,t_{k+1})} 1_{\frac{T_n(t_{k+1})}{T_n(t_k)} \in]\rho(t_k,t_{k+1}),\frac{1}{\rho(t_k,t_{k+1})}[}$$

as explained before, an easy way to perform our statistical test is to use the test statistic

$$M_n = \max\left\{T_n^2(t_1), T_n^2(t_2), h_n(t_1,t_2), ..., T_n^2(t_{K-1}), T_n^2(t_K), h_n(t_{K-1},t_K)\right\}.$$

Our first result is that the threshold (ie. critical value) is the same if we are under selective genotyping or in the oracle situation. So, the Monte-Carlo Quasi Monte-Carlo method, proposed by Azaïs et al. [2011] (based on Genz [1992]) for the oracle situation, is still suitable here to obtain our threshold. This way, in Figure 1, we propose to check this asymptotic results on simulated data. We consider a chromosome of length $T = 1$M, with two genetic markers located at each extremity. For such a configuration, the corresponding threshold is 5.40. We consider here $\gamma = 0.3$, and different ways of performing the selective genotyping : genotyping symetrically (ie $\gamma_+ = \gamma/2$), genotyping only the individuals with the largest phenotypes (ie $\gamma_+ = \gamma$) .... We can see that, whatever the value of γ_+, the Percentage of False Positives is close to the true level of the test (ie. 5%) even for small values of n (see $n = 50$).

In Figures 2 and 3, we focus on the alternative hypothesis. In Figure 2, we consider the same map and the same value of γ as previously. For the QTL effect q, we consider $a = 4$: we remind that $q = a/\sqrt{n}$. We focus on different locations $t^\star$ of the QTL and different values of γ_+. As expected (cf. Theorem 3), we can see that the Theoretical Power is maximum when we genotype symetrically (ie. $\gamma_+ = \gamma/2$). Note that, we also give in brackets the Empirical Power obtained for $n = 1000$, just to confirm our asymptotic results. Finally, in Figure 3, we focus on a more dense genetic map, and we change the value of γ : $\gamma = 0.6$ here. We obtain the same kind of conclusions as before.

γ_+ \ n	1000	200	50
γ	4.98%	4.96%	4.50%
$\gamma/2$	5.27%	4.89%	4.65%
$\gamma/4$	4.79%	4.91%	4.43%
$\gamma/8$	5.21%	4.99%	4.58%

Fig. 1. Percentage of False Positives as a function of n and the percentage γ_+ of individuals genotyped in the right tail. The chromosome is of length $T = 1$M and two markers are located at each extremity ($\gamma = 0.3$, $a = 0$, $\mu = 0$, $\sigma = 1$, 10000 samples of size n).

γ_+ \ $t^\star$	10cM	30cM	60cM	80cM
γ	53.72% (53.84%)	30.70% (30.02%)	25.59% (25.88%)	39.88% (39.10%)
$\gamma/2$	76.22% (75.98%)	46.64% (46.15%)	38.91% (38.30%)	59.82% (59.07%)
$\gamma/4$	72.71% (72.41%)	43.80% (43.51%)	36.42% (35.72%)	56.53% (56.02%)
$\gamma/8$	67.95% (67.56%)	40.15% (39.65%)	33.22% (33.77%)	52.16% (51.44%)

Fig. 2. Theoretical Power and Empirical Power (in brackets) as a function of the location of the QTL $t^\star$ and the percentage γ_+ of individuals genotyped in the right tail . The chromosome is of length $T = 1$M and the markers are equally spaced ($\gamma = 0.3$, $a = 4$, $\mu = 0$, $\sigma = 1$, 10000 samples of size $n = 1000$, 100000 paths for the Theoretical Power).

γ_+ \ $t^\star$	18cM	44cM	70cM	90cM
γ	64.16% (63.34%)	62.45% (62.48%)	59.43% (58.86%)	58.05% (57.67%)
$\gamma/2$	91.57% (91.71%)	90.45% (89.47%)	87.87% (88.05%)	87.42% (87.25%)
$\gamma/4$	89.22% (88.82%)	87.84% (88.17%)	85.06% (84.92%)	84.35% (83.77%)
$\gamma/8$	84.19% (84.84%)	82.55% (82.09%)	79.66% (79.51%)	78.43% (78.55%)

Fig. 3. Theoretical Power and Empirical Power (in brackets) as a function of the location of the QTL $t^\star$ and the percentage γ_+ of individuals genotyped in the right tail . The chromosome is of length $T = 1$M and 6 markers are equally spaced every 20cM ($\gamma = 0.6$, $a = 4$, $\mu = 0$, $\sigma = 1$, 10000 samples of size $n = 1000$, 100000 paths for the Theoretical Power).

6. Acknowledgements

I thank Jean-Marc Azaïs and Céline Delmas for fruitful discussions.

Charles-Elie Rabier (rabier@stat.wisc.edu)
Université de Toulouse, Institut de Mathématiques de Toulouse, U.P.S.,
F-31062 Toulouse Cedex 9, France.

References

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

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.

Azaïs, J. M., Gassiat, E., Mercadier, C. (2009). The likelihood ratio test for general mixture models with possibly structural parameter. *ESAIM*, To appear.

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

Azaïs, J. M., Delmas, C., Rabier, C-E (2011). *Likelihood Ratio Test process for Quantitative Trait Locus detection.* submitted to ESAIM.

Billingsley, P. (1999). *Convergence of probability measures.* Wiley, New-York.

Chang, M. N., Wu, R., Wu, S. S., Casella, G. (2009). Score statistics for mapping quantitative trait loci. *Statistical Application in Genetics and Molecular Biology*, **8(1)**, 16.

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

Darvasi, D., Soller, M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and Applied Genetics*, **85**, 353-359.

Davies, R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **64**, 247-254.

Davies, R.B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **74**, 33-43.

Feingold, E., Brown, P.O., Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of idendity by descent. *Am. J. Human. Genet.*, **53**, 234-251.

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

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comp. Graph. Stat.*, 141-149.

Ghosh, J.K., Sen, P.K. (1984). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. *Inst. Statistics Mimeo Series*, 1467.

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.

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

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. *Theoretical and Applied Genetics*, **73**, 556-562.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*, Springer.

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

Rabier, C-E. (2011). On statistical inference for selective genotyping, *submitted to Sankhya Series A*.

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

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

Siegmund, D. (1985). Sequential analysis : tests and confidence intervals. *Springer, New York*.

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

Wu, R., MA, C.X., Casella, G. (2007) *Statistical Genetics of Quantitative Traits*, Springer