On interpolations in QTL detection

Céline Delmas, Charles-Elie Rabier

On interpolations in QTL detection

Céline Delmas

INRA UR631, Station d'Amélioration Génétique des Animaux, Auzeville, France.

Charles-Elie Rabier

Université de Toulouse, Institut de Mathématiques de Toulouse, U.P.S., Toulouse, France.
INRA UR631, Station d'Amélioration Génétique des Animaux, Auzeville, 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. Recently, Azaïs et al. (2011) proved that the LRT process was the square of a non linear interpolated process. However, in their study of the same problem, Chang et al. (2009) introduced another interpolation. So, why do Azaïs et al. (2011) and Chang et al. (2009) find different interpolations ? We correct errors present in the interpolation of Chang et al. (2009) and establish the link between the two interpolations. We finally generalize the interpolation of Chang et al. (2009) to the alternative hypothesis of a QTL located at $t^\star \in [0, T]$.

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

1. Introduction

We focus on the famous "Interval Mapping" of Lander and Botstein (1989). That is to say, 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. We study a backcross population: $A \times (A \times B)$, where A and B are purely homozygous lines. The trait is observed on n individuals (progenies) and we denote by Y_j, $j = 1, ..., n$, 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 (see for instance Wu et al. (2007)).

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. 2 genetic markers are located at fixed locations $t_1 = 0 < t_2 = T$. 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 (1919) 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 $[0, T]$. 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.

In fact the "genome information" will be available only at the marker locations and the observation will be

$$(Y,\ X(t_1),\ X(t_2))\,.$$

So, we observe n observations $(Y_j,\ X_j(t_1),\ X_j(t_2))$ i.i.d. 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

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

The challenge is that the location of the QTL $t^\star$ is unknown, so we will perform a likelihood ratio test (LRT) in order to test the presence of a QTL (ie. $q = 0$) at every location $t \in [0,T]$. It leads to a process $\{\Lambda_n(t),\ t \in [0,T]\}$ called "LRT process", 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. A key point is that since the "genome information" is only available at marker locations, we have to deal with a mixture model, when we perform a test at a location t which does not correspond to a marker location. This mixture model has two Gaussian components : boths components have variance σ^2, but the first component has expected value $\mu + q$ whereas the second one has expected value $\mu - q$. So, at such a location t, in order to obtain the weights of our mixture model, for the first (resp. second) component, we have to compute the probability that $X(t) = 1$ (resp. $X(t) = -1$) given the "genome information" at markers. In particular, according to Azaïs et al. (2011), if we call $p(t) = \mathbb{P}\{X(t) = 1 \mid X(t_1), X(t_2)\}$ (ie. the weight for the first component), we have :

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

where :

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

This problem has been studied under some approximations by Rebaï et al. (1995), Rebaï et al. (1994), Cierco (1998), Azaïs and Cierco-Ayrolles (2002), Azaïs and Wschebor (2009). Recently, Azaïs et al. (2011) have shown that the LRT process was the square of a non linear interpolated process. The aim of this present article is to establish the link between this non linear interpolation and another interpolation introduced also recently by Chang et al. (2009). The interpolation of Azaïs et al. (2011) is a non linear interpolation between test statistics at marker locations : it means that all the test statistics inside a marker interval, can be deduced by interpolation from the test statistics at flanking markers. So, it explains why when people analyze data, they obtain a

likelihood profile which is smooth between markers. Besides, this interpolation leads to an easy formula for computing the supremum of the LRT process (see Lemma 1 of Azaïs et al. (2011)). The interpolation of Chang et al. (2009) is less interesting from a genetic point of view : it is very difficult to interpret it graphically. However, it is always interesting to understand why the two articles Azaïs et al. (2011) and Chang et al. (2009), which study the same problem, present different results. We will correct here technical errors present in Chang et al. (2009), and establish the link between the two interpolations. Finally, we will generalize the interpolation of Chang et al. (2009) to the alternative hypothesis of a QTL located at $t^\star \in [0, T]$, since contrary to Azaïs et al. (2011), Chang et al. (2009) focused only on the null hypothesis.

We refer to the book of Van der Vaart (1998) for elements of asymptotic statistics used in proofs.

2. Two different interpolations

Let H_0 be the null hypothesis $q = 0$. Since in Chang et al. (2009), the authors study only the null hypothesis, we will first focus only under the null hypothesis. Besides, it is well known that for a regular model, the score test is equivalent to the square of the LRT, so without loss of generality, we will limit our study to score tests as in Chang et al. (2009).

2.1. *Under the null hypothesis*

Let's consider a location t, distinct from the marker locations, that is to say $t \in]t_1, t_2[$, and the result will be prolonged by continuity at marker locations. $S_n(t)$ will be the score test statistic at location t.

According to Theorem 1 and formula (5) of Azaïs et al. (2011), we have :

$$S_n(t) = \frac{\left\{Q_t^{1,1} - Q_t^{-1,1}\right\} S_n(t_1) + \left\{Q_t^{1,1} - Q_t^{1,-1}\right\} S_n(t_2)}{\sqrt{\mathbb{E}\left[\{2p(t) - 1\}^2\right]}} \quad , \tag{2}$$

where

$$\forall k = 1, 2 \qquad S_n(t_k) = \sum_{j=1}^{n} \frac{(y_j - \mu)\ (21_{X(t_k)=1} - 1)}{\sigma\ \sqrt{n}} \quad ,$$

$$\mathbb{E}\left[\{2p(t) - 1\}^2\right] = \left\{Q_t^{1,1} - Q_t^{-1,1}\right\}^2 + \left\{Q_t^{1,1} - Q_t^{1,-1}\right\}^2 + 2\left\{Q_t^{1,1} - Q_t^{-1,1}\right\}\left\{Q_t^{1,1} - Q_t^{1,-1}\right\} e^{-2}$$

This is the non linear interpolation of Azaïs et al. (2011), between statistics on markers. The couple $(S_n(t_1), S_n(t_2))$ follow a standard bivariate normal distribution with covariance $e^{-2(t_2 - t_1)}$. Let's now establish the link beween this interpolation and the interpolation of Chang et al. (2009).

According to formula (5) of Azaïs et al. (2011) and using the fact that $Q_t^{1,1} =$

$1 - Q_t^{-1,-1}$ and $Q_t^{1,-1} = 1 - Q_t^{-1,1}$, we have

$$S_n(t) = (1 - 2Q_t^{-1,-1}) \sum_{j=1}^{n} \frac{(y_j - \mu)\left\{1_{X_j(t_1)=1} 1_{X_j(t_2)=1} - 1_{X_j(t_1)=-1} 1_{X_j(t_2)=-1}\right\}}{\sigma \sqrt{n} \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}}$$

$$+ (1 - 2Q_t^{-1,1}) \sum_{j=1}^{n} \frac{(y_j - \mu)\left\{1_{X_j(t_1)=1} 1_{X_j(t_2)=-1} - 1_{X_j(t_1)=-1} 1_{X_j(t_2)=1}\right\}}{\sigma \sqrt{n} \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}} \quad .$$

Let $G_n^1(t)$ and G_n^2 be the quantities such as :

$$G_n^1 = \sum_{j=1}^{n} \frac{(y_j - \mu)\left\{1_{X_j(t_1)=1} 1_{X_j(t_2)=1} - 1_{X_j(t_1)=-1} 1_{X_j(t_2)=-1}\right\}}{\sigma \sqrt{n\, \bar{r}(t_1,t_2)}} \quad ,$$

$$G_n^2 = \sum_{j=1}^{n} \frac{(y_j - \mu)\left\{1_{X_j(t_1)=1} 1_{X_j(t_2)=-1} - 1_{X_j(t_1)=-1} 1_{X_j(t_2)=1}\right\}}{\sigma \sqrt{n\, r(t_1,t_2)}} \quad .$$

G_n^1 and G_n^2 are asymptotically independent standard normal variables under H_0. Besides, it is clear that we have the following relationship between $S_n(t)$, G_n^1 and G_n^2 :

$$S_n(t) = \left\{\sqrt{\bar{r}(t_1,t_2)}\,(1 - 2Q_t^{-1,-1})\, G_n^1 \;+\; \sqrt{r(t_1,t_2)}\,(1 - 2Q_t^{-1,1})\, G_n^2\right\} \Big/ \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]} \quad . \tag{3}$$

We will see later that this interpolation is the interpolation of Chang et al. (2009) but rewritten without approximations. This interpolation is difficult to describe graphically because it is an interpolation between two test statistics G_n^1 and G_n^2, which boths include the genome information at the two markers. The main difference is that G_n^1 and G_n^2 are not points of the process $S_n(.)$, contrary to $S_n(t_1)$ and $S_n(t_2)$ for the interpolation of Azaïs et al. (2011).

Note that if we want to obtain the score test, we just have to replace μ by $\hat{\mu} := \frac{1}{n}\sum_{j=1}^{n} Y_j$ and σ by $\hat{\sigma} := \left\{\frac{1}{n-1}\sum (Y_j - \hat{\mu})^2\right\}^{1/2}$ in formulae (2) and (3). These new expressions (ie with $\hat{\sigma}$ and $\hat{\mu}$) of G_n^1, G_n^2, $S_n(t_1)$ and $S_n(t_2)$ are asymptotically equivalent to the previous ones. We will call respectively $\widehat{G}_n^1$ and $\widehat{G}_n^2$ the new expressions of G_n^1, G_n^2.

Let's now focus on the work of Chang et al. (2009). With our notations, the score test statistic of formula (8) of Chang et al. (2009) is :

$$U_n(t) = \left\{\sqrt{\bar{r}(t_1,t_2)}\,(1 - 2Q_t^{-1,-1})\, \widetilde{G}_n^1 \;+\; \sqrt{r(t_1,t_2)}\,(1 - 2Q_t^{-1,1})\, \widetilde{G}_n^2\right\} \Big/ \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]} \quad ,$$

where

$$\widetilde{G}_n^1 = \frac{\sum_{j=1}^{n} y_j\; 1_{X_j(t_1)=1} 1_{X_j(t_2)=1} \;-\; \sum_{j=1}^{n} y_j\; 1_{X_j(t_1)=-1} 1_{X_j(t_2)=-1}}{\hat{\sigma}\sqrt{n\,\bar{r}(t_1,t_2)}} \quad ,$$

$$\widetilde{G}_n^2 = \frac{\sum_{j=1}^{n} y_j\; 1_{X_j(t_1)=1} 1_{X_j(t_2)=-1} \;-\; \sum_{j=1}^{n} y_j\; 1_{X_j(t_1)=-1} 1_{X_j(t_2)=1}}{\hat{\sigma}\sqrt{n\,\bar{r}(t_1,t_2)}} \quad .$$

Let $o_{P_{\theta_0}}(1)$ be a sequence of random vectors that converges to zero in probability under H_0 (i.e. no QTL on the whole interval studied) and let $O_p(1)$ be a sequence bounded in probability. We have :

$$\widehat{G}_n^1 = \frac{\sum_{j=1}^n y_j \ \left\{1_{X_j(t_1)=1}1_{X_j(t_2)=1} \ - \ 1_{X_j(t_1)=-1}1_{X_j(t_2)=-1}\right\}}{\hat{\sigma}\ \sqrt{n\ \bar{r}(t_1,t_2)}} - \hat{\mu}\frac{\sum_{j=1}^n \left\{1_{X_j(t_1)=1}1_{X_j(t_2)=1} \ - \ 1\right.}{\hat{\sigma}\ \sqrt{n\ \bar{r}(t_1,t}}$$
$$= \widetilde{G}_n^1 - \left\{O_p(1) + O_p(1/\sqrt{n})\right\} O_p(1) = \widetilde{G}_n^1 + O_p(1) + o_{P_{\theta_0}}(1) \ .$$

In the same way :

$$\widehat{G}_n^2 = \widetilde{G}_n^2 + O_p(1) + o_{P_{\theta_0}}(1) \ .$$

As a consequence, since $\widehat{G}_n^2 = G_n^2 + o_{P_{\theta_0}}(1)$ and $\widehat{G}_n^1 = G_n^1 + o_{P_{\theta_0}}(1)$, we can remark that $S_n(t) \neq U_n(t) + o_{P_{\theta_0}}(1)$. So, the interpolation introduced by Chang et al. (2009) is only an approximation as said before. The interpolation of Chang et al. (2009) rewritten without approximations is presented in formula (3).

2.2. *Under the alternative hypothesis*

Let define the alternative hypothesis :

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

In Azaïs et al. (2011), the authors show that this alternative hypothesis is contiguous to the null hypothesis. So, it makes the algebra easy under $H_{at^\star}$. We will still have the same interpolations as under H_0. In particular, for the non linear interpolation, according to Azaïs et al. (2011), we still have $(S_n(t_1), S_n(t_2))$ which follows a bivariate normal distribution with covariance $e^{-2(t_2-t_1)}$. However, $S_n(t_1)$ and $S_n(t_2)$ are not centered anymore : $\mathbb{E}\left\{S_n(t_1)\right\} = ae^{-2(t^\star - t_1)}/\sigma$ and $\mathbb{E}\left\{S_n(t_2)\right\} = ae^{-2(t_2-t^\star)}/\sigma$.

Let's focus now on the interpolation of Chang et al. (2009). After some calculations and using the fact that $Q_t^{1,1} = 1 - Q_t^{-1,-1}$ and $Q_t^{1,-1} = 1 - Q_t^{-1,1}$, we obtain

$$\mathbb{E}\left\{G_n^1\right\} = a\ \sqrt{\bar{r}(t_1,t_2)}\ (2Q_{t^\star}^{1,1} - 1)/\sigma \ ,$$
$$\mathbb{E}\left\{G_n^2\right\} = a\ \sqrt{r(t_1,t_2)}\ (2Q_{t^\star}^{1,-1} - 1)/\sigma \ .$$

Besides,

$$\mathrm{Cov}\left\{\mathrm{G_n^1}, \mathrm{G_n^2}\right\} = \mathbb{E}\left\{G_n^1 G_n^2\right\} - \mathbb{E}\left\{G_n^1\right\}\mathbb{E}\left\{G_n^2\right\} = 0 - \mathbb{E}\left\{G_n^1\right\}\mathbb{E}\left\{G_n^2\right\}$$
$$= -a^2\ \sqrt{\bar{r}(t_1,t_2)}\sqrt{r(t_1,t_2)}\ (2Q_{t^\star}^{1,-1} - 1)\ (2Q_{t^\star}^{1,1} - 1)/\sigma^2 \ .$$

So, under the alternative, G_n^1 and G_n^2 will still be asymptotically normal with unit variance. However, G_n^1 and G_n^2 are not independent anymore (contrary to under the null hypothesis).

Note that here, we limited our study to only two genetic markers located on the chromosome, but it can easily be generalized to several markers.

3. Acknowledgements

The authors thank Jean-Marc Azaïs and Jean-Michel Elsen for fruitful discussions. This work has been supported by the Animal Genetic Department of the French National Institute for Agricultural Research, SABRE, and the National Center for Scientific Research.

Céline Delmas (celine.delmas@toulouse.inra.fr)
INRA UR631, Station d'Amélioration Génétique des Animaux,
BP 52627-31326 Castanet-Tolosan Cedex, France.
Charles-Elie Rabier (rabier@stat.wisc.edu)
Université de Toulouse, Institut de Mathématiques de Toulouse, U.P.S.,
F-31062 Toulouse Cedex 9, France.
INRA UR631, Station d'Amélioration Génétique des Animaux,
BP 52627-31326 Castanet-Tolosan Cedex, 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. 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.

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.

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.

Rabier, C-E. (2010). *PhD thesis*, Université Toulouse 3, Paul Sabatier.

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.

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