THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Mathématiques-Statistiques*

Présentée et soutenue par *RABIER Charles-Elie*
Le *16/06/2010*

Titre : *Techniques statistiques pour la détection de gènes à effets quantitatifs*

JURY

AZAÏS Jean-Marc (Université Paul Sabatier, directeur de thèse)
BORDES Laurent (Université de Pau et des Pays de l'Adour, président et rapporteur)
DELMAS Céline (INRA Auzeville, examinatrice)
ELSEN Jean-Michel (INRA Auzeville, co-directeur de thèse)
MOREAU Laurence (INRA Le Moulon, examinatrice)
PRUM Bernard (Université d'Evry, rapporteur)

Ecole doctorale : *Mathématiques, Informatique et Télécommunications de Toulouse*
Unités de recherche : *Institut de Mathématiques de Toulouse, Université Paul Sabatier, UMR 5219*
Station d'Amélioration Génétique des Animaux, INRA Auzeville, UR631
Directeur(s) de Thèse : *AZAÏS Jean-Marc / ELSEN Jean-Michel*
Rapporteurs : *BORDES Laurent / PRUM Bernard*

Techniques statistiques pour la détection de gènes à effets quantitatifs

Charles-Elie Rabier

Ce travail a été réalisé au sein de la Station d'Amélioration Génétique des Animaux (SAGA) de l'INRA de Toulouse, et du Laboratoire de Statistique et Probabilités (LSP), membre désormais de l'Institut de Mathématiques de Toulouse (IMT). Il a été sponsorisé par le CNRS et le département Genétique Animale de l'INRA.

Je remercie tout d'abord mes deux directeurs de thèse, Jean-Michel Elsen et Jean-Marc Azaïs.

Merci à Jean-Michel Elsen pour m'avoir formé à la génétique et de m'avoir permis de travailler sur les problématiques actuelles des généticiens.

Merci à Jean-Marc Azaïs. J'ai pu profiter de ses grandes compétences théoriques et de son enthousiasme lors de nos nombreux rendez-vous dans l'unique suite avec balcon du LSP.

Ce travail doit beaucoup à Céline Delmas avec qui j'ai eu le plaisir de travailler au quotidien, notamment à la transcription en langage mathématique des techniques utilisées par les généticiens. J'espère que notre approche multi-QTL sera en mesure de concurrencer le fameux Composite Interval Mapping qui règne au pays de la génétique !

Je remercie mes deux rapporteurs, Bernard Prum et Laurent Bordes, pour leurs remarques et l'intérêt qu'ils ont porté au manuscript. Merci à Laurence Moreau d'avoir participé à ce jury.

Merci au directeur de l'école doctorale MITT de m'avoir permis de réaliser ce travail à l'intersection de deux disciplines. J'espère qu'aucune des deux communautés ne se sentira frustrée.

Merci à tous ceux qui m'ont motivé à faire une thèse : Régis Sabbadin qui m'a donné goût à la recherche lors de mon stage en BIA (INRA de Toulouse) sur la conservation de la biodiversité au Costa Rica, Béatrice Laurent qui m'a encouragé à poursuivre en Master 2 Recherche, et mes amis Irlandais Caroline Brophy et Mickaël Hawkins from UCD, qui après quelques Guiness m'ont convaincu de faire une thèse.

Comment ne pas remercier Michèle, ma collègue de bureau à l'INRA ? Encore désolé d'avoir oublié d'arroser ses plantes ... Merci à Zoubida pour les nombreux fous rires, Edouardo pour nos nombreuses discussions culturelles ou scientifiques, Robert pour les bons moments passés en congrès, Seyed pour nos footing ponctués de corde à sauter sur le canal du midi, David, pour nos soirées "Bonbon rose", Kileh, le futur diplomate Djiboutien, Ana, avec un seul “n", Benjamin, pas que pour ses magrets, et Anne, pour avoir eu la folie de se lancer dans une thèse. Une petite pensée pour les informaticiens qui ont tenté en vain de m'apprendre les différents vainqueurs du top 14. Enfin, merci à Eddy et Thomas qui ont passé mon portable au service de réanimation.

Bien entendu, je remercie les membres de la “MAFIA" du LSP qui lors de nos séjours à Nissan et en Espagne, m'ont motivé durant ma thèse : merci Jean-François, Thierry,

Jean-Claude, Fabrice ... Je tiens également à saluer le département de Génie Mathématiques et Modélisation de l'INSA et par conséquent Philippe Besse qui m'a accueilli en tant qu'ATER. Je tiens à remercier Olivier Mazet de m'avoir concocté un programme d'enseignement varié et plaisant. Merci à Antonietta, Simona, Cathy, Frédéric, Aldéric, Fabien, Benoît .... Et surtout merci à ceux qui m'ont donné un grand bol d'énergie : mes étudiants de première année. J'espère avoir été aussi bon que Laurence Di Poï qui m'enseignait l'algèbre linéaire à Poitiers.

Merci au groupe ski de Paul Sabatier, avec qui j'ai cru mourir plusieurs fois en hors piste à Saint-Lary, mais bon je vous rassure cette thèse n'arrive pas à titre posthume. Merci au Nick Bollettieri du tennis de la fac, Dédé et son cigare, Christophe pour m'avoir enfin appris le revers lifté long de ligne, Julien, le peintre de Port-Sud, et surtout Christiano (pas celui du Real), la recrue Colombienne en quête d'american express gold.

Enfin, as usual, je remercie ma grande soeur et mes parents qui y sont pour beaucoup. Après l'hydrogéologie, les "glide" les "shuffle" et Doudou lapin nous ont menés au pays des QTL. Certes, après un dérapage en ski, je ne peux pas voir des "glyde" et des "shuffle", mais avec un peu de chance, au détour d'une bosse, un rider peut surgir et tenter un backcross.

Merci à tous.

Table des matières

Introduction générale **9**

Rappels de génétique **13**

Détection de QTL **19**

I Première Partie : Etude du Selective Genotyping **29**

1 Introduction **35**
1.1 Motivation 35
1.2 L'étude dans sa globalité 35
1.2.1 Selective genotyping en présence d'un caractère quantitatif 36
1.2.2 Selective genotyping en présence de deux caractères quantitatifs corrélés 36

2 Eléments théoriques nécessaires à l'étude **37**

3 Selective genotyping en présence d'un caractère quantitatif **41**
3.1 Introduction 41
3.2 Test statistique en l'absence de censure 42
3.2.1 Modèle en l'absence de censure 42
3.2.2 Test statistique oracle (μ, q, σ) 42
3.3 Etude des différentes stratégies en selective genotyping 45
3.3.1 Modèle correspondant au selective genotyping 45
3.3.2 Efficacités et puissances des tests correspondant aux différentes stratégies (résultats principaux) 46
3.3.3 Résultats secondaires 57
3.3.4 Résumé des différents résultats 61
3.3.5 Remarques générales sur la modélisation du selective genotyping . 63
3.4 Annexe 64
3.4.1 Test de Wald (μ, q, σ) (situation oracle) 64

3.4.2 Test statistique oracle (q) 65
3.4.3 Test statistique oracle (μ, q) 66
3.4.4 Algorithme EM pour la première stratégie 67
3.4.5 Preuve du corollaire 3.9 69
3.4.6 Convergence vers l'asymptotique 72

4 Selective genotyping en présence de deux caractères quantitatifs corrélés 77
4.1 Introduction 77
4.2 Test statistique en l'absence de censure 78
4.2.1 Modèle en l'absence de censure 78
4.2.2 Test statistique oracle (μ_Z, q_Z) 78
4.3 Etude des différentes stratégies en selective genotyping 79
4.3.1 Modèle 79
4.3.2 Efficacités et puissances des tests correspondant aux différentes stratégies (résultats principaux) 79
4.3.3 Résultats secondaires 87
4.3.4 Résumé des différents résultats 90
4.4 Annexe 92
4.4.1 Convergence vers l'asymptotique 92

II Deuxième Partie : Génome Scan 95

1 Introduction 101
1.1 Contexte 101
1.2 Feuille de route 102

2 Asymptotic process for QTL detection 105
2.1 Introduction 105
2.2 Model 105
2.3 Only two genetic markers 106
2.3.1 Likelihood Ratio Test 106
2.3.2 Process under H_0 108
2.3.3 Process under $H_{\lambda t^\star}$ 109
2.3.4 Addendum about the process 111
2.4 Several markers : the "Interval Mapping" of Lander and Botstein (1989) 114
2.4.1 Process under H_0 114
2.4.2 Process under $H_{\lambda t^\star}$ 114
2.4.3 Addendum about the process 116

2.5 Generalization to families of sires with their own informative markers . . 118
2.5.1 Process under H_0 120
2.5.2 Process under $H_{\lambda t^\star}$ 121
2.5.3 Addendum about the processes 122
2.5.4 Relaxing some hypotheses 122
2.6 Appendix 125
2.6.1 Formula for $\mathbb{E}\left[(2p_t - 1)^2\right]$ 125
2.6.2 Covariances of the process (only 2 markers) 125
2.6.3 Covariance of the process (Interval Mapping) 125
2.6.4 Covariance of the process (I families of sires) 126
2.6.5 Proof of "Relaxing some hypotheses" 126
2.7 Article submitted "LRT process for QTL detection" 127
2.8 Article submitted "Threshold and power for QTL detection" 128

3 About the supremum of the linear interpolated process 129
3.1 Introduction 129
3.2 Study of the maximum 129
3.2.1 Only two genetic markers on $[0, 1]$ 129
3.2.2 Only two genetic markers on $[0, T]$ 132
3.2.3 Several markers : the "Interval Mapping" of Lander and Botstein (1989) 132
3.2.4 Graphical illustrations under H_0 134
3.2.5 Thresholds 135

4 About the supremum of Chi-Square processes 139

Conclusion et perspectives 141

Introduction générale

La recherche de gènes s'avère un des grands défis actuels. Les enjeux sont très variés : médicaux, économiques, et plus encore la compréhension du domaine du vivant. Grâce aux progrès récents de la biologie moléculaire, de nombreux marqueurs génétiques sont maintenant disponibles et organisés en cartes génétiques dans un grand nombre d'espèces. Cela constitue une avancée considérable dans la détection et l'identification de loci (ie. emplacements physiques précis sur un chromosome) où la variation allélique est associée à la variation d'un caractère quantitatif. On nomme QTL (Quantitative Trait Loci) de tels loci.

Les principes de bases de la détection de QTL datent de 1920. Le sujet n'est devenu très répandu qu'à la fin des années 80 avec l'avènement des microsatellites (séquences constituées de répétitions en tandem de mono-, di- ou trinucléotides et se caractérisant par un polymorphisme important). Jusqu'à cette période les marqueurs génétiques disponibles étaient seulement les marqueurs morphologiques et les groupes sanguins. Autant dire une couverture impossible de la totalité du génome.

Aujourd'hui, avec l'arrivée des SNPs (Single Nucleotid Polymorphisms) et par conséquent l'afflux d'informations moléculaires, l'analyse statistique demeure un outil encore plus précieux pour l'analyse du génome. En 1989, Lander et Botstein publiaient un article méthodologique présentant une méthode statistique consistant à scanner le génome, nommée "Interval Mapping". Aujourd'hui, cette méthode s'avère incontournable en détection de QTL. Elle a permis la découverte de milliers de QTL dans le domaine des plantes, dans le domaine animal et chez les humains.

L'objectif de cette thèse est d'étudier la théorie sous-jacente aux techniques de détection de QTL. A l'issue de cette étude théorique, on pourra contribuer à l'optimisation du processus de détection et proposer par la même occasion de nouvelles méthodes de détection de QTL.

Après une introduction rapide à la génétique, ce document s'articule autour de deux parties : l'étude du génotypage sélectif (selective genotyping), l'étude de la technique qui consiste à scanner le génome (génome scan).

L'introduction à la génétique fournit le vocabulaire et les bases de génétique utilisés dans ce travail. Elle est largement inspirée des ouvrages Lynch and Walsh (1997),

Hayes (2005) et Wu and al. (2007). On y présente le crossing-over, les marqueurs et distances génétiques. Les grands principes de la détection de QTL sont rappelés, notamment au travers d'un schéma expérimental très utilisé dans le domaine végétal : le backcross. On évoquera aussi les populations outbred que l'on retrouve chez les humains et très souvent dans le domaine animal.

La première partie est consacrée au selective genotyping. Il a été proposé par Lebowitz and al. (1987). Le selective genotyping consiste à génotyper uniquement les individus dont la valeur du caractère quantitatif est extrême. Cela permet de réduire les coûts dus au génotypage tout en gardant une bonne puissance pour le test statistique, à condition que le nombre d'individus ait été augmenté.

Dans cette partie, on s'intéresse aux propriétés statistiques de ce dispositif expérimental uniquement en un point précis du génome. A travers cette étude, on cherche non seulement à proposer différents tests pour la détection de QTL mais également à quantifier l'apport des phénotypes non extrêmes dans l'analyse statistique.

On s'attarde également sur la question de l'optimisation du génotypage en selective genotyping. Enfin, on porte notre attention sur le selective genotyping en présence de deux caractères quantitatifs corrélés. En effet, il est possible de réduire les coûts dus à la fois au génotypage et au phénotypage en effectuant un selective genotyping sur le premier caractère et en ne mesurant le deuxième caractère que pour les individus génotypés. On teste alors s'il existe un QTL affectant le deuxième caractère.

La deuxième partie est consacrée au genome scan. On ne se place plus uniquement en un point précis du génome mais on scanne désormais le génome. De plus, on suppose que l'ensemble des individus sont génotypés. On étudie le problème de détection de QTL sur un chromosome, dans une population de descendants structurés en familles de pères. Une population backcross, $C \times (C \times D)$, où C et D sont de pures lignées homozygotes, est un cas particulier de la population étudiée.

La présence de QTL est testée à chaque position sur le chromosome à l'aide d'un test du rapport de vraisemblance (Likelihood Ratio Test : LRT). Le choix comme statistique de test du maximum de ce processus, revient à effectuer un LRT dans un modèle où la localisation du QTL est un paramètre supplémentaire. Utilisant la modélisation de Haldane, on présente des résultats asymptotiques sur la distribution du processus de LRT sous l'hypothèse nulle d'absence de QTL, sous l'alternative où il existe un seul QTL sur le chromosome, et sous l'alternative générale où il existe m QTL sur le chromosome. Dans ce dernier cas, une hypothèse d'additivité des effets des QTL a été faite. Cependant, ces résultats sont généralisables au cas d'interaction entre les QTL : l'épistasie. On énoncera également des résultats qui permettront d'approcher les populations animales, dites "outbred". D'autre part, on proposera plusieurs méthodes de calcul de seuil adaptées à la carte génétique. Pour finir, on s'attardera sur l'optimisation du processus de détection

et sur le choix de l'endroit où procéder aux tests sur le chromosome.

A noter que dans cette deuxième partie, l'hypothèse de normalité des phénotypes pourra être levée, à condition d'utiliser des processus de score et non pas de LRT.

Rappels de génétique

Vocabulaire

Un **gène** est défini comme une séquence d'acides désoxyribonucléiques (ADN), destiné à être trancrit en acide ribonucléique (ARN). Si c'est le cas, la "séquence" est dite codante. La molécule d'ARN ainsi produite peut soit être traduite en protéine (on l'appelle dans ce cas ARN messager), soit être directement fonctionnelle. A titre d'exemple, il y a environ 13000 gènes dans l'ADN des cellules d'une drosophile et 20000 gènes chez l'homme.

Les gènes apparaissent généralement en paires : on parle alors d'espèce **diploïde**. Chez un individu appartenant à une espèce **sexuée** (i.e. nécessitant le concours d'une cellule sexuelle mâle et d'une cellule sexuelle femelle) et diploïde, chaque gène existe en deux exemplaires, l'un hérité du père, et l'autre de la mère.

Chaque gène peut exister en un ou plusieurs variants, appelés **allèles**. Un gène codant la couleur des yeux peut ainsi exister en plusieurs variants codant plusieurs coloris. Un gène est dit **polymorphe** lorsque plusieurs allèles sont présents dans l'espèce. Un individu est dit **homozygote** pour un gène lorsqu'il est porteur de deux allèles similaires. Par opposition, on dira qu'il est **hétérozygote** s'il est porteur de deux allèles différents.

On distingue les **allèles** dominants (le caractère correspondant s'exprime toujours) des allèles **récessifs** (muets lorsqu'ils sont associés à l'allèle dominant).

Chaque gène occupe une position constante sur le chromosome : on parle de **locus**. La paire d'allèles présente sur des chromosomes **homologues** (i.e. les chromosomes d'une même paire) constitue le **génotype** de l'individu pour ce locus.

Un **phénotype** désigne l'état d'un caractère observable chez l'individu.

Méïose et crossing-over

Il existe deux types de divisions cellulaires dans le monde vivant : la mitose qui assure la naissance de cellules identiques à la cellule mère et la méïose qui aboutit à la production de cellules sexuelles ou gamètes pour la reproduction.

Ainsi, la méïose est la division particulière qui permet le passage de la phase diploïde (les cellules contiennent $2n$ chromosomes) à la phase haploïde (les cellules contiennent n chromosomes).

Le point de départ est une cellule dite “germinale" à la base de la reproduction ($2n$ chromosomes). Les étapes successives sont les suivantes :

1. duplication du matériel génétique
2. appariement des chromosomes homologues
3. phénomène de brassage intra chromosomique plus connu sous le nom de “crossing over"
4. migration des chromosomes homologues vers les pôles
5. deux divisions successives qui aboutissent à la production de quatre cellules haploïdes (n chromosomes).

La figure 1 illustre les différentes étapes de la méïose dans le cas $n = 1$. Le phénomène de crossing-over est détaillé en figure 2.

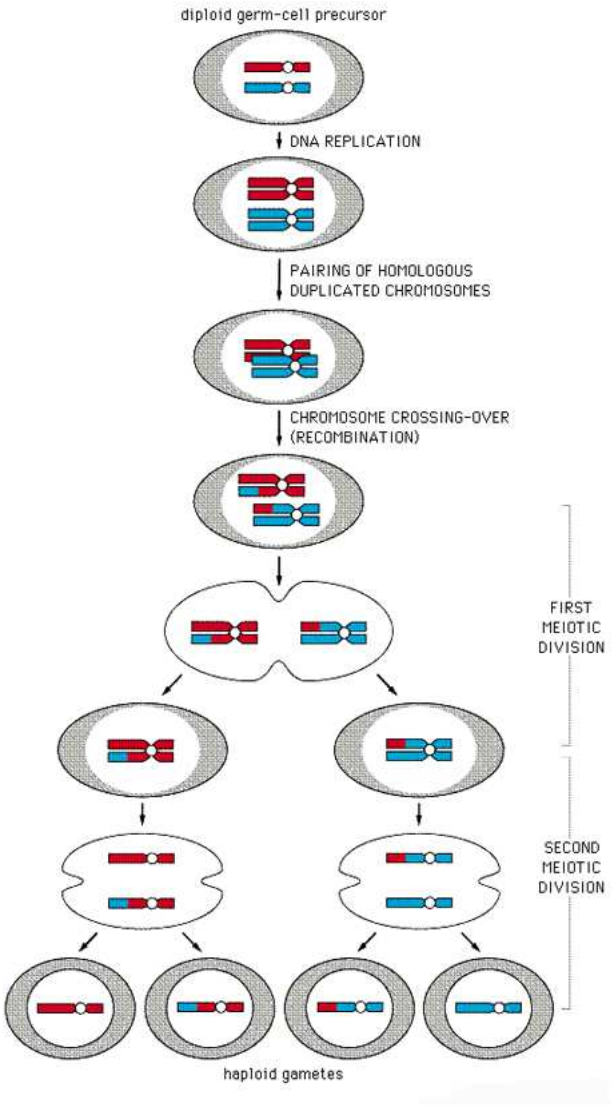

diploid germ-cell precursor
DNA REPLICATION
PAIRING OF HOMOLOGOUS
DUPLICATED CHROMOSOMES
CHROMOSOME CROSSING-OVER
(RECOMBINATION)
FIRST
MEIOTIC
DIVISION
SECOND
MEIOTIC
DIVISION
haploid gametes

FIG. 1 – Méïose

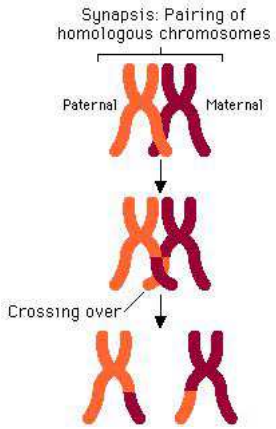


FIG. 2 – Le phénomène du crossing-over

Marqueurs génétiques

Un marqueur génétique est une séquence polymorphe d'ADN aisément détectable, utilisée en cartographie génétique afin de "baliser" le génome. On s'attarde ici sur les spécificités des microsatellites et des SNPs.

Un microsatellite est une séquence constituée en répétition en tandem de di-, tri- ou tetra nucléotides. Les microsatellites sont les marqueurs des années 90 et 2000. Ils sont très polymorphes en raison des répétitions très variables (en moyenne sept à huit allèles dans les espèces).

Un SNP (Single Nucleotide Polymorphism) désigne une variation d'un seul nucléotide. Les SNPs sont bien plus nombreux que les microsatellites. Il existe quelques millions de SNPs dans une espèce alors qu'il n'existe que mille à deux mille microsatellites disponibles. Les SNPs sont facilement détectables grâce aux programmes de séquençage à haut débit, c'est pourquoi ils constituent les marqueurs d'avenir. Ils présentent néanmoins un faible polymorphisme car ce sont des marqueurs bialléliques.

Distances génétiques

La distance génétique entre deux loci sur un chromosome est définie comme étant l'espérance du nombre de crossing-over entre ces deux loci durant la méïose. On mesure

cette distance en Morgans (M).

Grâce à la distance de Haldane (1919) ou celle de Kosambi (1944), il est possible d'exprimer la distance génétique entre deux loci en fonction de la probabilité de recombinaison entre ces deux loci.

Dans la modélisation de Haldane (1919), on suppose que les crossing-over sont indépendants et que leur nombre suit un processus de poisson d'intensité 1. Si l'on note r (resp. N) la probabilité de recombinaison (resp. le nombre de crossing-over) entre deux loci distants de d, alors :

$$r = \mathbb{P}(N \text{ impair}) = \sum_{k=0}^{+\infty} \mathbb{P}(N = 2k+1) = \sum_{k=0}^{+\infty} \frac{e^{-d}}{(2k+1)!}\, d^{2k+1} = e^{-d} sh(d) = \frac{1}{2}(1 - e^{-2d})$$

Par conséquent, la distance de Haldane d vérifie :

$$d = -\frac{1}{2}\log(1-2r) \quad \forall r \in [0, 1/2[$$

Cependant, certaines observations empiriques ont montré que la probabilité que deux crossing-over se produisent à proximité est toujours plus faible que celle prédite par la fonction de Haldane. En effet, l'occurence d'un crossing-over inhibe les chances d'arrivée d'un autre crossing-over à proximité (phénomène d'interférence).

La modélisation de Kosambi (1944) tient compte de ce phénomène d'interférence. La distance d est la suivante :

$$d = \frac{1}{4}\log\left(\frac{1+2r}{1-2r}\right) \quad \forall r \in [0, 1/2[$$

La figure 3 illustre les distances génétiques de Haldane et de Kosambi. Dans cette thèse, on considèrera principalement la distance de Haldane. Elle s'avère la plus fréquemment utilisée dans la communauté génétique.

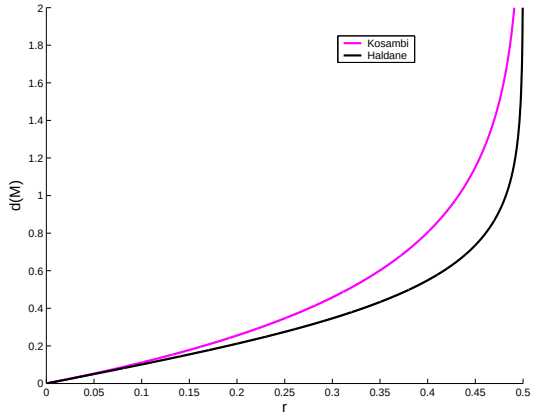


FIG. 3 – Distance génétique d en fonction de la probabilité de recombinaison r

Détection de QTL

Notion de QTL

Afin d'expliquer la variation génétique des caractères quantitatifs, deux modèles ont été proposés : le modèle infinitésimal et le modèle avec un nombre fini de loci.
Dans le modèle infinitésimal, on suppose que les caractères quantitatifs sont gouvernés par un nombre infini de loci additifs et indépendants, chacun ayant un faible effet (Fisher (1918)). Cependant, il y aurait environ 20000 gènes ou loci dans le génome. Par conséquent, il doit y avoir un nombre fini de loci à l'origine de la variation d'un caractère quantitatif.
La recherche de ces loci, nommés QTL (Quantitative Trait Loci), s'avère un grand défi à l'heure actuelle. A noter que peu de gènes ont un grand effet sur le caractère, alors que de nombreux gènes sont de faible effet (Hayes and Goddard (2001)).

Présentation d'un schéma expérimental : le backcross

Les croisements sont très fréquemment utilisés en recherche de QTL : ils permettent de disposer d'une population en ségrégation pour le QTL. Par la suite, à l'aide de marqueurs génétiques positionnés le long du génome et des valeurs phénotypiques, on cherchera à détecter et localiser les QTL.

On présente ici un schéma expérimental fondamental dans le domaine végétal : le backcross. Ce schéma repose sur le croisement de lignées pures homozygotes (on parlera de lignées "inbred").

La figure 4 détaille comment obtenir une population backcross. On y considère un intervalle défini par deux marqueurs A et B, qui possèdent chacun deux allèles (A_1, A_2 pour A et B_1, B_2 pour B). Un QTL Q présentant deux allèles Q_1 et Q_2, est supposé présent entre A et B à une position donnée.

Deux lignées inbred sont croisées afin de générer l'hétérozygote F1. Le F1 est alors rétrocroisé à un de ses parents afin d'obtenir une population en ségrégation pour le QTL.

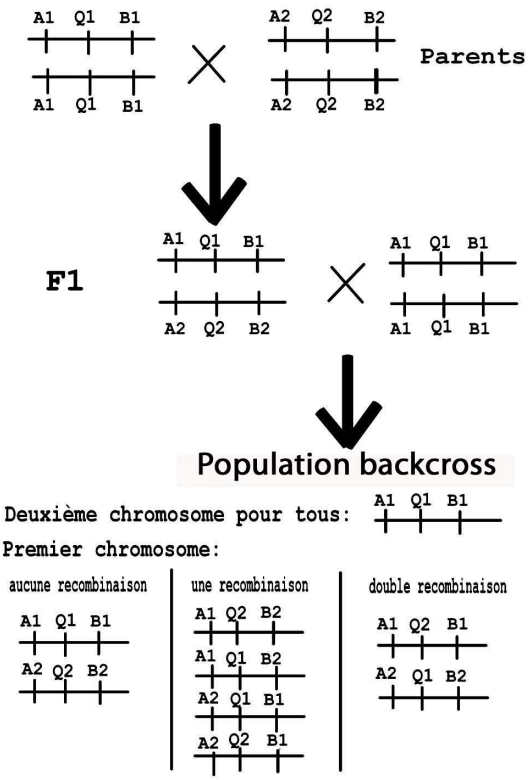


FIG. 4 – Une population backcross

Quelques techniques statistiques de détection de QTL

On détaille ici quelques techniques statistiques de détection de QTL dans le cadre du backcross présenté ci-dessus.

On considère un modèle additif (aucun effet de dominance entre les allèles). Le caractère d'intérêt Y vérifie :

$$Y = \begin{cases} \mu + q + \varepsilon & \text{si génotype } Q_1Q_1 \text{ au QTL} \\ \mu - q + \varepsilon & \text{si génotype } Q_1Q_2 \text{ au QTL} \end{cases}$$

où $\varepsilon \sim N(0, \sigma^2)$.

Première approche : le QTL est présent sur un marqueur

On considère que le QTL est positionné sur un marqueur donné. On a donc connaissance du génotype au QTL.
Afin de tester la nullité de l'effet du QTL, on confronte les deux hypothèses suivantes :

$$H_0 : q = 0 \quad \text{vs} \quad H_1 : q \neq 0$$

Une statistique exhaustive est une simple comparaison de moyenne entre les individus de génotype Q_1Q_1 et ceux Q_1Q_2 au QTL (cf. Cierco (1996)).

Deuxième approche : la position du QTL est inconnue

On supposait précédemment que le QTL était positionné sur un marqueur donné, A par exemple. On suppose désormais que le QTL est présent entre A et B à une position inconnue.
Nous allons parcourir l'intervalle délimité par A et B, et pour chaque position t, nous allons tester s'il y a présence d'un QTL en calculant la statistique du rapport de vraisemblance.

Interval Mapping

A chaque position t entre A et B, le génotype au QTL est inconnu. On va utiliser les probabilités de recombinaison entre les marqueurs et le QTL afin de reconstruire le génotype au QTL.

En notant $\theta = (q, \mu, \sigma)$, la vraisemblance à la position t, pour n observations j indépendantes et identiquement distribuées (iid), s'écrit :

$$L_n(\theta, t) = \prod_{j=1}^{n} p_t^j f_{(\mu+q,\sigma)}(y_j) \; + \; (1 - p_t^j) f_{(\mu-q,\sigma)}(y_j) \tag{1}$$

$f_{(\mu,\sigma)}(.)$ désigne une densité Gaussienne de moyenne μ et de variance σ^2. p_t^j désigne la probabilité que l'individu j soit de génotype Q_1Q_1 pour un QTL positionné en t, sachant son génotype aux marqueurs A et B. Ces poids s'obtiennent grâce à la formule de Bayes et à la formule de Haldane.

Afin de tester la présence d'un QTL en t, on confronte :

$$H_0 : q = 0 \quad \text{vs} \quad H_1 : q \neq 0$$

On note Λ_t la statistique du rapport de vraisemblance à la position t. Le processus $\Lambda_{(.)}$ est appelé processus de tests de rapport de vraisemblance. Cette technique se généralise

à plusieurs marqueurs localisés sur un chromosome en utilisant les marqueurs flanquant la position testée. Cette méthode a été proposée par Lander and Botstein (1989). On la nomme “Interval Mapping".
Les auteurs proposent d'utiliser comme statistique de test le supremum du processus $\Lambda_{(.)}$. En effet, le QTL aura tendance à se situer à la position t pour laquelle la statistique de test Λ_t sera la plus grande.

La figure 5 présente une trajectoire du processus $\Lambda_{(.)}$ lorsque deux marqueurs sont situés à 0cM et 20cM. Une telle trajectoire est nommée profil de vraisemblance. On constate que s'il existe un QTL, il se situe aux environs de 14cM.

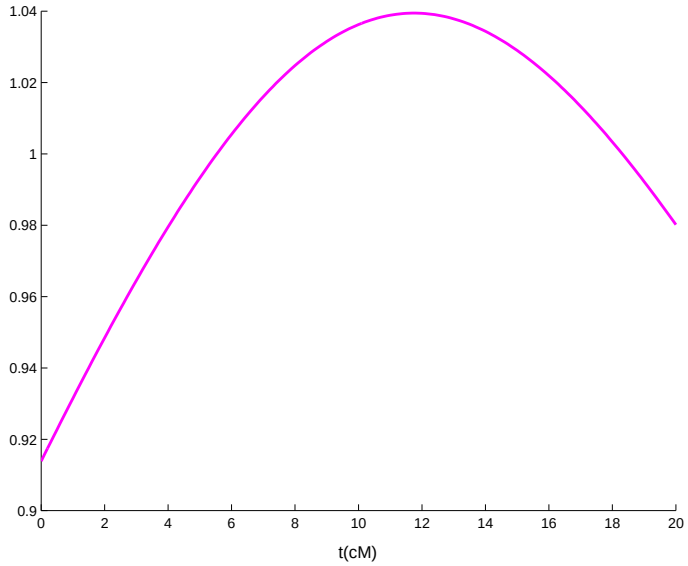


FIG. 5 – Une trajectoire du processus $\Lambda_{(.)}$ (2 marqueurs situés respectivement à 0cM et à 20cM)

Régression

L'“Interval Mapping" présente un inconvénient : les estimateurs du maximum de vraisemblance (EMV) ne peuvent être obtenus explicitement en raison du modèle de mélange. On peut par exemple utiliser l'algorithme EM afin de calculer les EMV.

Cependant, pour remédier à ces problèmes de maximisation numérique, certains auteurs (Knapp and al. (1990), Haley and Knott (1992), Haley and al. (1994)) proposent d'utiliser une méthode de régression. Cette méthode consiste à linéariser le mélange présent dans

la vraisemblance (cf. formule 1). La vraisemblance s'écrit dès lors :

$$L_n(\theta, t) = \prod_{j=1}^{n} f_{(\mu+q(2p_t^j-1),\sigma)}(y_j)$$

La vraisemblance s'avère très facilement maximisable et on peut proposer une statistique de Fisher pour tester la nullité de l'effet QTL q en un point t. Par la suite, on considère comme statistique de test le supremum du processus de Fisher.

Prise de décision

On considère ici le modèle correspondant à l'"Interval Mapping".

Afin de prendre une décision quant à l'existence d'un QTL sur le chromosome étudié, il nous faut connaître la région critique du test statistique basé sur le supremum du processus $\Lambda_{(.)}$.

On précise tout d'abord l'hypothèse nulle :

H_0 : "Il n'y a pas de QTL sur le chromosome étudié"

De nombreux travaux théoriques sur la valeur critique (ie. seuil) du supremum de $\Lambda_{(.)}$ ou sur une statistique de test très proche ont été effectués. On se contente ici d'évoquer ces travaux : on s'intéressera aux différents processus dans la deuxième partie ce cette thèse (partie génome scan).

Dans le cadre de cartes denses (ie. le nombre de marqueurs tend vers l'infini), on peut citer Delong (1981), Lander and Botstein (1989), Cierco (1996).

Dans le cadre de cartes non denses (ie. quelques marqueurs présents sur le chromosome), on peut citer Feingold and al. (1993), Rebaï and al. (1994), Piepho (2001).

Cependant, dans tous ces articles, on suppose que le nombre d'individus tend vers l'infini.

La méthode Churchill and Doerge (1994) dite des "permutations" ou "suffling" en anglais, présente l'avantage de ne pas être asymptotique. De plus, elle peut être utilisée dans le cadre de carte denses ou non denses. C'est pourquoi, elle s'avère très employée en détection de QTL à l'heure actuelle. Elle présente néanmoins un désavantage : elle est très lourde en terme de temps de calcul.

Afin de se placer sous l'hypothèse nulle d'absence de QTL sur le chromosome, Churchill and Doerge (1994) proposent de casser la liaison marqueur/phénotype en permutant l'ensemble des phénotypes entre eux. Ainsi, chaque phénotype se voit affecter l'information marqueur (ie. les génotypes aux différents marqueurs) d'un autre individu. La structure de corrélation entre les statistiques de test est par conséquent préservée.

Introduction aux populations “outbred"

Dans de nombreuses espèces, en particulier chez les humains et très souvent dans le domaine animal, on est incapable de générer des lignées “inbred" (comme pour le backcross) pour des raisons biologiques. On parle de population “outbred". On peut néanmoins à l'aide de croisements, générer une population en ségrégation pour le QTL.

Afin qu'un parent apporte une information de liaison, il doit être hétérozygote à la fois à un marqueur et à un QTL lié. Un QTL ne peut être détecté que dans cette situation. Quelques parents d'une population “outbred" sont de tels doubles hétérozygotes par opposition aux lignées “inbred" pour lesquelles les F1 sont hétérozygotes à tous les loci (cf. backcross figure 4 page 20).

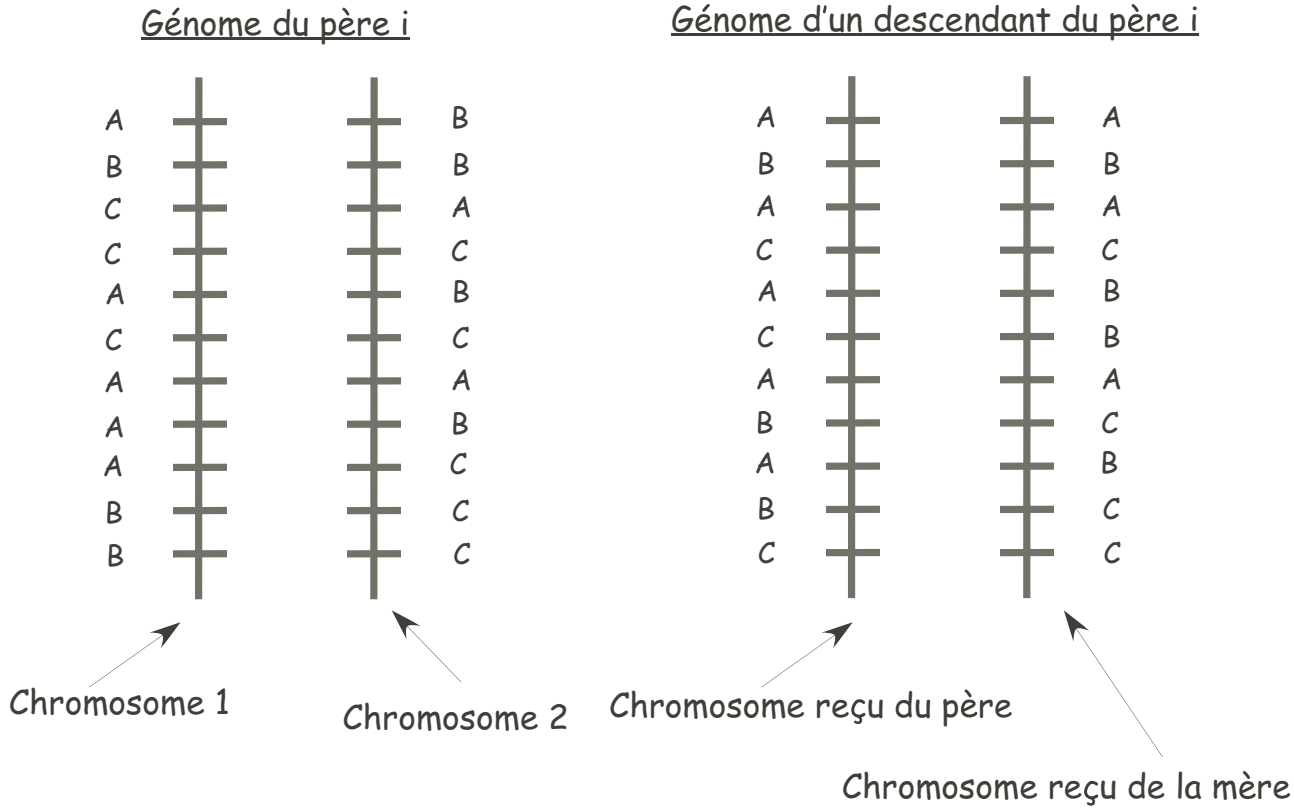
A noter que dans les croisements entre lignées “inbred", seulement deux allèles sont en ségrégation à chaque locus, alors que dans une population “outbred", n'importe quel nombre d'allèles est en ségrégation.

Dans cette thèse, on appellera famille de père, un père et ses descendants provenant d'accouplement avec de nombreuses mères. En présence de populations outbred, un dispositif largement employé consiste à chercher à détecter des QTL dans plusieurs familles de pères car seulement certains pères sont hétérozygotes au QTL. En effet, il s'avère impossible de détecter un QTL parmi les descendants d'un père homozygote au QTL. En considérant plusieurs familles de pères, on augmente les chances d'obtenir des pères hétérozygotes au QTL. Un problème se pose cependant. Afin de mener à bien l'analyse QTL, on a besoin de connaitre pour chaque descendant, la provenance des allèles aux différents marqueurs transmis par le père. Cela s'avère impossible pour certains marqueurs : ces marqueurs sont dits “non informatifs".

La figure 6 illustre le problème de l'informativité des marqueurs génétiques, propre aux populations outbred. On considère onze marqueurs trialléliques (allèles A,B ou C). Un exemple de génome pour un père est indiqué. Le génome d'un descendant de ce père est également représenté : on distingue clairement le chromosome reçu du père et celui reçu de la mère.

Cependant, en réalité, l'origine parentale est inconnue. On sait uniquement que ce descendant est AA au premier marqueur, BB au deuxième marqueur... Le tableau liste les marqueurs informatifs du descendant. Ce dernier est AA au premier marqueur, de plus, son père porte sur son chromosome 1, l'allèle A au premier marqueur, et sur son chromosome 2, l'allèle B. Le descendant aura donc forcément reçu l'allèle A de son père, autrement dit l'allèle porté par le chromosome 1 de son père. Le marqueur 1 est donc un marqueur informatif. Si l'on considère désormais le deuxième marqueur, le descendant et son père présentent tous les deux le génotype BB. Il s'avère impossible de déterminer l'origine de l'allèle reçu du père au marqueur 2. Le marqueur 2 sera donc non informatif.

On remarquera également que pour un descendant quelconque du père considéré, uniquement les marqueurs 1, 3, 5, 8, 9, 10, 11 pourront être informatifs car le père est hétérozygote à ces marqueurs. Parmi ces marqueurs susceptibles d'être informatifs, en raison du patrimoine génétique de la mère, seuls les marqueurs 1, 3, 8, 9 et 11 sont informatifs pour le descendant considéré en figure 6.



Numéro du marqueur	Génotype au marqueur	Origine de l'allèle du marqueur reçu du père
1	AA	Chromosome 1
2	BB	?
3	AA	Chromosome 2
4	CC	?
5	AB	?
6	CB	?
7	AA	?
8	BC	Chromosome 2
9	AB	Chromosome 1
10	BC	?
11	CC	Chromosome 2

Informativité des marqueurs du descendant du père i

FIG. 6 – Illustration du problème de l'informativité des marqueurs génétiques en population outbred

Première partie

Etude du Selective Genotyping

Cette partie présente une analyse théorique du “selective genotyping" dans le cas respectivement d’un seul caractère quantitatif, puis de deux caractères quantitatifs corrélés. Après une brève introduction, nous développons dans différents chapitres la méthodologie statistique sous-jacente au “selective genotyping".

Notations

n	nombre d'observations
j	indice de l'observation
φ	densité d'une loi normale centrée réduite
Φ	fonction de répartition de la loi normale centrée réduite
z_α	quantile d'ordre $1-\alpha$ d'une loi normale centrée réduite (ie. $\Phi(z_\alpha) = 1-\alpha$)
X	génotype au QTL
S_-, S_+	seuils fixes réels tels que $S_- \leqslant S_+$
T	statistique oracle de comparaison de moyenne
W_i	statistique de Wald pour la stratégie i

Chapitre 3

Y	phénotype (ie. caractère quantitatif)
q	effet du QTL
H_0	hypothèse nulle ($q = 0$)
H_a	hypothèse alternative locale ($q = \frac{a}{\sqrt{n}}$)
T_2	statistique de comparaison de moyenne pour la stratégie deux
θ	paramètre du modèle statistique
θ_0	paramètre du modèle statistique sous H_0
$\hat{\theta}$	estimateur du maximum de vraisemblance (EMV) de θ
I_{θ_0}	information de Fisher au point θ_0
$I_{ij}(\theta_0)$	élément ij de I_{θ_0}
$I_{\theta_0}^{-1}$	inverse de l'information de Fisher au point θ_0
$I_{ij}^{-1}(\theta_0)$	élément ij de $I_{\theta_0}^{-1}$
$n^\star$	nouveau nombre d'individus
ζ	ratio tel que $\zeta = \frac{n^\star}{n}$
ζ_{eff}	valeur de ζ pour laquelle la puissance d'un test 1 est égale à celle d'un test 2
κ_i	efficacité de la stratégie i (relativement au test oracle)
γ	quantité telle que $\gamma = \mathbb{P}_{H_0}(Y \notin [S_-, S_+])$
γ_+	quantité telle que $\gamma_+ = \mathbb{P}_{H_0}(Y > S_+)$
γ_-	quantité telle que $\gamma_- = \mathbb{P}_{H_0}(Y < S_-)$
$\mathcal{A}$	quantité telle que $\mathcal{A} = \sigma^2 \left\{\gamma + z_{\gamma_+}\, \varphi(z_{\gamma_+}) - z_{1-\gamma_-}\, \varphi(z_{1-\gamma_-})\right\}$

Chapitre 4

Y	phénotype sur lequel le selective genotyping est effectué
Z	deuxième phénotype (ie. deuxième caractère quantitatif)
r	coefficient de corrélation entre Y et Z au sein d'un génotype
q_Z	effet du QTL sur Z
q_Y	effet du QTL sur Y
H_{0Z}	hypothèse nulle sur Z ($q_Z = 0$)
H_{bZ}	hypothèse alternative locale sur Z ($q_Z = \frac{b}{\sqrt{n}}$)
H_{0Y}	hypothèse nulle sur Y ($q_Y = 0$)
H_{aY}	hypothèse alternative locale sur Y ($q_Y = \frac{a}{\sqrt{n}}$)
θ	paramétrisation classique du modèle statistique
$\theta^\star$	paramétrisation $\star$ du modèle statistique
L	vraisemblance en paramétrisation classique
$L^\star$	vraisemblance en paramétrisation $\star$
I_θ	information de Fisher au point θ (relative à L)
$I^\star_{\theta^\star}$	information de Fisher au point $\theta^\star$ (relative à $L^\star$)
$I_{ij}(\theta)$	élément ij de I_θ
$I^\star_{ij}(\theta^\star)$	élément ij de $I^\star_{\theta^\star}$
I_θ^{-1}	inverse de I_θ au point θ
$I^{\star\,-1}_{\theta^\star}$	inverse de $I^\star_{\theta^\star}$ au point $\theta^\star$
$I^{-1}_{ij}(\theta)$	élément ij de I_θ^{-1}
$I^{\star\,-1}_{ij}(\theta^\star)$	élément ij de $I^{\star\,-1}_{\theta^\star}$
$\tilde{\kappa}_i$	efficacité de la stratégie i (relativement au test oracle)
γ	quantité telle que $\gamma = \mathbb{P}_{H_{0Y}}(Y \notin [S_-, S_+])$
γ_+	quantité telle que $\gamma_+ = \mathbb{P}_{H_{0Y}}(Y > S_+)$
γ_-	quantité telle que $\gamma_- = \mathbb{P}_{H_{0Y}}(Y < S_-)$
$\mathcal{A}$	quantité telle que $\mathcal{A} = \sigma^2 \left\{\gamma + z_{\gamma_+}\,\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\,\varphi(z_{1-\gamma_-})\right\}$

Chapitre 1

Introduction

1.1 Motivation

Les nouvelles technologies en matière de génomique se révèlent être efficaces afin de percer les secrets de la variation génétique d'un caractère quantitatif. Ces technologies permettent la caractérisation moléculaire de marqueurs polymorphes (i.e. présentant plusieurs allèles) sur l'ensemble du génome. Ces derniers seront par la suite utilisés pour identifier et localiser les QTL. Néanmoins, les coûts dus au génotypage demeurent très élevés. C'est pourquoi l'optimisation du processus expérimental est primordiale. L'un de ces processus expérimentaux s'intitule selective genotyping. Il a été proposé par Lebowitz and al. (1987), et élaboré par Lander and Botstein (1989), Darvasi and Soller (1992), puis Muranty and Goffinet (1997). Le selective genotyping consiste à génotyper uniquement les individus dont la valeur du caractère quantitatif est extrême (plus grande ou plus petite qu'un seuil). Cela permet de réduire les coûts dus au génotypage tout en gardant une bonne puissance pour le test statistique, à condition que le nombre d'individus ait été augmenté.

Le selective genotyping peut également être employé en présence de deux caractères quantitatifs corrélés, notamment lorsque le deuxième caractère est difficile à mesurer pour des raisons biologiques. Il est ainsi possible de réduire les coûts dus à la fois au génotypage et au phénotypage : on effectue un selective genotyping sur le premier caractère, et on ne mesure le deuxième caractère que pour les individus génotypés. On teste alors s'il existe un QTL affectant le deuxième caractère.

1.2 L'étude dans sa globalité

L'étude porte successivement sur :

1. un selective genotyping en présence d'un caractère quantitatif

2. un selective genotyping en présence de deux caractères quantitatifs corrélés

Cette étude se situe dans le prolongement des travaux de Muranty and Goffinet (1997) où les auteurs s'intéressent au selective genotyping en présence d'un seul puis de deux caractères. Contrairement à Muranty and Goffinet (1997), nous ne nous attarderons pas sur l'estimation des effets QTL, mais sur la construction de tests statistiques.

1.2.1 Selective genotyping en présence d'un caractère quantitatif

Dans cette étude, on cherche non seulement à proposer des tests pour la détection de QTL, mais également à quantifier l'apport des phénotypes non extrêmes dans l'analyse statistique.

Par conséquent, différentes stratégies pour l'analyse en selective genotyping sont étudiées. La puissance des tests correspondant aux différentes stratégies est étudiée sous des alternatives contiguës. Tous ces tests sont comparés en terme d'efficacité au test oracle, celui où tous les génotypes sont connus.

1.2.2 Selective genotyping en présence de deux caractères quantitatifs corrélés

Par soucis de clareté, on note :
- Y le caractère sur lequel le selective genotyping est effectué
- Z le caractère difficile à mesurer pour des raisons biologiques

On cherche cette fois-ci à savoir s'il existe un QTL affectant le phénotype Z. Comme précédemment, on cherche non seulement à proposer différents tests mais également à quantifier l'apport des phénotypes Y non extrêmes dans l'analyse statistique.

Par conséquent, différentes stratégies sont étudiées. Tous les tests correspondant aux différentes stratégies sont comparés en terme d'efficacité au test oracle, celui où tous les génotypes ainsi que tous les phénotypes Z sont connus.

Chapitre 2

Eléments théoriques nécessaires à l'étude

On introduit ici les éléments théoriques nécessaires à l'étude du selective genotyping en utilisant les mêmes notations que celles employées dans Van der Vaart (1998). De plus, les théorèmes et lemmes énoncés par la suite, proviennent également de cet ouvrage. Uniquement le théorème 2.9 n'y est pas présent : il a été démontré dans le cadre de cette thèse.

Notation 2.1 *Soit $\Theta \in \mathbb{R}^d$. Soit X une variable aléatoire de loi P_θ $(\theta \in \Theta)$. Pour une fonction f donnée, on définit la notation suivante :*

$$P_\theta f = \mathbb{E}_\theta \left[f(X) \right]$$

De plus, on note p_θ une densité de P_θ par rapport à une mesure μ.

Notation 2.2 *Soit $\theta_0 \in \Theta$. La notation $o_{P_{\theta_0}}(1)$ désigne une suite de vecteurs aléatoires qui converge en probabilité vers 0 sous P_{θ_0} .*

Notation 2.3 *On notera I_{θ_0} la matrice d'information de Fisher au point θ_0. $I_{ij}(\theta_0)$ désigne l'élément ij de I_{θ_0}. De même, $I_{\theta_0}^{-1}$ sera l'inverse de I_{θ_0} et $I_{ij}^{-1}(\theta_0)$ désignera l'élément ij de $I_{\theta_0}^{-1}$.*

Théorème 2.4 *Soit $X_1, ..., X_n$ un échantillon iid provenant d'une distribution P_θ. Supposons que le modèle $(P_\theta : \theta \in \Theta)$ est différentiable en moyenne quadratique en un point intérieur θ_0 de $\Theta \subset \mathbb{R}^d$ et notons $\dot{\ell}_{\theta_0}$ le score au point θ_0.*
De plus, supposons qu'il existe une fonction mesurable $\dot{g}$ vérifiant $P_{\theta_0}\dot{g}^2 < \infty$ telle que, pour tout θ_1 et θ_2 au voisinage de θ_0,

$$|\log\, p_{\theta_1}(x) - \log\, p_{\theta_2}(x)| \leqslant \dot{g}(x)\, \|\theta_1 - \theta_2\|$$

Si la matrice d'information de Fisher I_{θ_0} n'est pas singulière et si l'estimateur du maximum de vraisemblance (EMV) $\hat{\theta}$ est consistant, alors :

$$\sqrt{n}(\hat{\theta} - \theta_0) = I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \dot{\ell}_{\theta_0}(X_j) \; + \; o_{P_{\theta_0}}(1)$$

En particulier, la séquence $\sqrt{n}(\hat{\theta} - \theta_0)$ est asymptotiquement normale de moyenne 0 et de covariance $I_{\theta_0}^{-1}$.

On se propose d'introduire la notion de contiguïté.

Définition 2.5 *Soient P_n et Q_n deux suites de mesures de probabilité sur des espaces $(\Omega_n, \mathcal{A}_n)$. On dit que la mesure de probabilité Q_n est contiguë par rapport à P_n si toute suite d'événements A_n vérifiant $\lim P_n(A_n) = 0$ vérifie aussi $\lim Q_n(A_n) = 0$.*
On note alors $Q_n \triangleleft P_n$. Si $Q_n \triangleright P_n$ et $P_n \triangleleft Q_n$, on dit alors que P_n et Q_n sont mutuellement contiguës et on note $P_n \triangleleft\triangleright Q_n$.

Le premier lemme de Le Cam énonce des conditions équivalentes à la contiguïté de Q_n par rapport à P_n, en terme du comportement asymptotique des rapports de vraisemblance dQ_n/dP_n et dP_n/dQ_n.

Lemme 2.6 (Premier lemme de Le Cam) *Soient P_n et Q_n deux suites de mesures de probabilité sur des espaces $(\Omega_n, \mathcal{A}_n)$. Alors il y a équivalence entre :*

1. $Q_n \triangleleft P_n$
2. *S'il existe une sous-suite de $\frac{dP_n}{dQ_n}$ qui converge en loi sous Q_n vers une variable aléatoire U (définie sur un espace de probabilité $(\Omega, \mathcal{F}, P)$), alors $P(U > 0) = 1$*
3. *S'il existe une sous-suite de $\frac{dQ_n}{dP_n}$ qui converge en loi sous P_n vers une variable aléatoire V (définie sur un espace de probabilité $(\Omega, \mathcal{F}, P)$), alors $\mathbb{E}(V) = 1$*
4. *Pour n'importe quelle statistique $T_n : \Omega_n \mapsto \mathbb{R}^d$: si $T_n \xrightarrow{P_n} 0$, alors $T_n \xrightarrow{Q_n} 0$*

Lemme 2.7 (Troisième lemme de Le Cam) *Soient P_n et Q_n deux suites de mesures de probabilité sur des espaces $(\Omega_n, \mathcal{A}_n)$ et $T_n : \Omega_n \mapsto \mathbb{R}^d$ une suite de variables aléatoires. Supposons que*

$$\left(T_n, \log\left(\frac{dQ_n}{dP_n}\right) \right)^t \xrightarrow{P_n} N_{d+1}\left(\begin{pmatrix} \xi \\ -\frac{1}{2}\nu^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^t & \nu^2 \end{pmatrix} \right)$$

alors

$$T_n \xrightarrow{Q_n} N_d(\xi + \tau, \Sigma)$$

(l'exposant t indiquant la transposée, et N_d une loi normale de dimension d).

Théorème 2.8 *Soient $X_1, ..., X_n$ un échantillon iid provenant d'une distribution P_θ. Supposons que Θ est un ouvert de $\mathbb{R}^d$ et que le modèle $(P_\theta : \theta \in \Theta)$ est différentiable en moyenne quadratique en $\theta_0 \in \Theta$. Posons $\ell_\theta(x) = \log p_\theta(x)$ où p_θ est une densité de P_θ par rapport à une mesure μ. Alors, $P_{\theta_0}\dot{\ell}_{\theta_0} = 0$ et la matrice d'information de Fisher $I_{\theta_0} = P_{\theta_0}\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^t$ existe. De plus, pour toute séquence convergente de type $h_n \to h$:*

$$\log \prod_{j=1}^{n} \frac{p_{\theta_0+h_n/\sqrt{n}}}{p_{\theta_0}}(X_j) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} h^t \dot{\ell}_{\theta_0}(X_j) \; - \; \frac{1}{2} h^t I_{\theta_0} h \; + \; o_{P_{\theta_0}}(1)$$

Le 3ème lemme de Le Cam indique que l'on passe de la loi asymptotique sous l'hypothèse nulle à celle sous une alternative contiguë par un phénomène de translation.
On énonce désormais un théorème qui nous permettra de calculer par la suite très facilement ces translations lorsque l'on considère des statistiques de Wald.

Théorème 2.9 *Soient $X_1, ..., X_n$ un échantillon iid provenant d'une distribution P_θ. Supposons que Θ est un ouvert de $\mathbb{R}^d$ et que le modèle $(P_\theta : \theta \in \Theta)$ est régulier. On note $\theta_0 \in \Theta$ et $\hat{\theta}$ l'EMV de θ, alors pour toute séquence convergente de type $h_n \to h$, on a :*

i) sous P_{θ_0} : $\sqrt{n}(\hat{\theta} - \theta_0) \to N(0,\, I^{-1}(\theta_0))$
ii) sous $P_{\theta_0+h_n/\sqrt{n}}$: $\sqrt{n}(\hat{\theta} - \theta_0) \to N(h,\, I^{-1}(\theta_0))$

Preuve.
On note :
- P_n la loi correspondant à $P_{\theta_0}^{\otimes n}$
- Q_n la loi correspondant à $P_{\theta_0+h_n/\sqrt{n}}^{\otimes n}$
- $\frac{dQ_n}{dP_n}$ le rapport de vraisemblance

Comme le modèle est régulier, on a immédiatement :

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{P_n}{\to} N(0,\, I^{-1}(\theta_0))$$

ce qui prouve i). Comme le modèle est régulier, on peut utiliser le théorème 2.8 qui permet d'obtenir une expression explicite de la log vraisemblance sous P_n. Grâce au théorème central limite, à la loi des grands nombres et aux propriétés de l'information de Fisher, on a :

$$\log\left(\frac{dQ_n}{dP_n}\right) \overset{P_n}{\to} N(-\frac{1}{2}\nu^2,\, \nu^2) \quad \text{avec } \nu^2 = h^t I_{\theta_0} h$$

Remarque 2.10 *Ce résultat implique $P_n \triangleleft\triangleright Q_n$. D'après le 3) du premier lemme de Le Cam, on a $Q_n \triangleleft P_n$. De plus, en échangeant les rôles de Q_n et P_n, en applicant le 2), on a $P_n \triangleleft Q_n$.*

Les conditions d'application du 3ème lemme de Le Cam sont bien remplies.
Comme le modèle est régulier, on utilise le théorème 2.4 :

$$\sqrt{n}(\hat{\theta} - \theta_0) = I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \dot{\ell}_{\theta_0}(X_j) \ + \ o_{P_{\theta_0}}(1)$$

D'après le théorème 2.8

$$\log\left(\frac{dQ_n}{dP_n}\right) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} h^t \dot{\ell}_{\theta_0}(X_j) \ - \ \frac{1}{2} h^t I_{\theta_0} h \ + \ o_{P_{\theta_0}}(1)$$

On note $h_{(i)}$ la ième composante de h.
A la ième ligne, on a :

$$\begin{aligned} \mathrm{Cov}\left(\log\left(\frac{dQ_n}{dP_n}\right), \ \sqrt{n}(\hat{\theta} - \theta_0)\right) &= \sum_{k=1}^{d} h_{(k)} \left\{ I_{i1}^{-1}(\theta_0) I_{1k}(\theta_0) + ... + I_{id}^{-1}(\theta_0) I_{dk}(\theta_0) \right\} \ + \ o_{P_{\theta_0}}(1) \\ &= h_{(i)} \ + \ o_{P_{\theta_0}}(1) \end{aligned}$$

Alors, par le 3ème lemme de Le Cam :

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{Q_n}{\to} N(h, \ I^{-1}(\theta_0))$$

Ce qui conclut la preuve. ■

Chapitre 3

Selective genotyping en présence d'un caractère quantitatif

3.1 Introduction

Dans ce chapitre, on s'intéresse à un seul caractère quantitatif (ie. un seul phénotype). Les coûts dus au génotypage étant relativement élevés, un selective genotyping est effectué. Ainsi, on génotype uniquement les individus présentant un phénotype extrême.

On s'intéresse aux propriétés statistiques de ce dispositif expérimental uniquement en un point précis du génome : on se place sur un marqueur génétique. Trois stratégies différentes pour l'analyse statistique en selective genotyping sont étudiées :

- la première consiste à conserver dans l'analyse statistique, tous les phénotypes, même les phénotypes qui ne sont pas considérés comme extrêmes et pour lesquels nous ne disposons pas du génotype
- la deuxième et la troisième stratégies sont basées sur la conservation uniquement des phénotypes extrêmes

Les tests correspondant aux différentes stratégies sont les suivants :

- tests de Wald (basés sur la normalité asymptotique des EMV) pour les stratégies une et trois
- comparaison de moyenne pour la stratégie deux

Ainsi, la deuxième stratégie se distingue par sa simplicité. Tous ces tests seront comparés en terme d'efficacité au test oracle, celui où tous les génotypes sont connus.

A travers cette étude théorique, on cherche non seulement à proposer différents tests pour la détection de QTL mais également à quantifier l'apport des phénotypes non extrêmes dans l'analyse statistique.

On s'intéresse par la suite à la question de l'optimisation du génotypage en selective genotyping : supposons que l'on souhaite génotyper uniquement un pourcentage γ de la

population, doit-on génotyper uniquement les individus présentant les plus grands phénotypes, ou au contraire ceux présentant les plus petits phénotypes, ou bien un mélange des deux ?

Les principaux résultats obtenus lors de cette étude du selective genotyping sont énoncés en théorème 3.3 page 46 et lemme 3.8 page 55.

3.2 Test statistique en l'absence de censure

3.2.1 Modèle en l'absence de censure

Soit X la variable aléatoire (v.a.) correspondant au génotype au QTL. On considérera 2 génotypes possibles au QTL :

$$X = \begin{cases} -1 & \text{avec probabilité } 1-p \\ 1 & \text{avec probabilité } p \end{cases}$$

On supposera $p \neq \{0, 1\}$.

Soit Y la variable aléatoire correspondant au phénotype. Le modèle pour la v.a. Y s'écrit :

$$Y = \mu + qX + \varepsilon$$

où ε est une v.a. de loi normale de moyenne 0 et de variance σ^2, et où q désigne l'effet du QTL.

On considèrera un échantillon de n observations (X_j, Y_j) indépendantes et équidistribuées (iid).

3.2.2 Test statistique oracle (μ, q, σ)

On suppose ici que l'on dispose d'un modèle statistique à trois paramètres (μ, q, σ). Afin de tester la présence d'un QTL, on confronte les 2 hypothèses suivantes :

$$H_0 : \ q = 0 \ \ vs \ \ H_1 : \ q \neq 0$$

Plus précisément, on utilisera comme hypothèse alternative, une alternative locale H_a : $q = \frac{a}{\sqrt{n}}$ où a est une constante.

Un estimateur naturel de l'effet QTL q est la comparaison de moyenne suivante :

$$\frac{1}{2}\left\{ \frac{\sum_{j=1}^{n} Y_j \, 1_{X_j=1}}{\sum_{j=1}^{n} 1_{X_j=1}} - \frac{\sum_{j=1}^{n} Y_j \, 1_{X_j=-1}}{\sum_{j=1}^{n} 1_{X_j=-1}} \right\}$$

Cependant cet estimateur n'est pas facilement exploitable en raison des dénominateurs aléatoires.
Par conséquent, on cherche à construire un estimateur bien plus simple.
Si on pose $\eta = qX + \varepsilon$, on remarque que sous l'alternative locale H_a :

$$\mathbb{E}_{H_a}\left\{\frac{1}{2n}\left(\sum_{j=1}^{n}\frac{\eta_j}{p}\,1_{X_j=1} - \frac{\eta_j}{1-p}\,1_{X_j=-1}\right)\right\} = q$$

Ce nouvel estimateur est donc un estimateur sans biais de q.
Sous l'hypothèse nulle H_0, on a :

$$\mathbb{E}_{H_0}\left(\frac{\eta}{p}\,1_{X=1} - \frac{\eta}{1-p}\,1_{X=-1}\right) = 0$$

Et :

$$\begin{aligned}\mathbb{E}_{H_0}\left\{\left(\frac{\eta}{p}\,1_{X=1} - \frac{\eta}{1-p}\,1_{X=-1}\right)^2\right\} &= \mathbb{E}_{H_0}\left(\frac{\eta^2}{p^2}\,1_{X=1} + \frac{\eta^2}{(1-p)^2}\,1_{X=-1}\right)\\ &= \frac{\sigma^2}{p(1-p)}\end{aligned}$$

Alors,

$$\mathrm{Var}_{H_0}\left(\frac{\eta}{p}\,1_{X=1} - \frac{\eta}{1-p}\,1_{X=-1}\right) = \frac{\sigma^2}{p(1-p)}$$

De plus, sous l'alternative locale H_a :

$$\mathbb{E}_{H_a}\left(\frac{\eta}{p}\,1_{X=1} - \frac{\eta}{1-p}\,1_{X=-1}\right) = 2q \tag{3.1}$$

$$\mathbb{E}_{H_a}\left\{\left(\frac{\eta}{p}\,1_{X=1} - \frac{\eta}{1-p}\,1_{X=-1}\right)^2\right\} = \frac{1}{p}(\sigma^2+q^2) + \frac{1}{1-p}(\sigma^2+q^2) \to \frac{\sigma^2}{p(1-p)}$$

$$\mathrm{Var}_{H_a}\left(\frac{\eta}{p}\,1_{X=1} - \frac{\eta}{1-p}\,1_{X=-1}\right) = \frac{1}{p}(\sigma^2+q^2) + \frac{1}{1-p}(\sigma^2+q^2) - 4q^2$$

On remarque que :

$$\mathrm{Var}_{H_a}\left(\frac{\eta}{p}\,1_{X=1} - \frac{\eta}{1-p}\,1_{X=-1}\right) \to \mathrm{Var}_{H_0}\left(\frac{\eta}{p}\,1_{X=1} - \frac{\eta}{1-p}\,1_{X=-1}\right)$$

On en déduit la statistique de test $\widetilde{T}$ suivante :

$$\widetilde{T} = \frac{\sum_{j=1}^{n} \frac{\eta_j}{p} 1_{X_j=1} - \frac{\eta_j}{1-p} 1_{X_j=-1}}{\sigma \sqrt{\frac{n}{p\,(1-p)}}}$$

Ses lois asymptotiques sont les suivantes :

$$\widetilde{T} \overset{H_0}{\to} N(0,\ 1) \quad \text{et} \quad \widetilde{T} \overset{H_a}{\to} N\left(\frac{2a\,\sqrt{p\,(1-p)}}{\sigma},\ 1\right)$$

Cependant, on observe non pas les v.a. η mais les phénotypes Y.
On note $\overline{Y}$ et $\overline{\eta}$ les moyennes empiriques : $\overline{Y} = \frac{1}{n}\sum Y_j$ et $\overline{\eta} = \frac{1}{n}\sum \eta_j$.
Alors, $\overline{Y} = \mu + \overline{\eta}$ et par conséquent $Y - \overline{Y} = \eta - \overline{\eta}$. On s'intéresse dès lors à la statistique de test T suivante :

$$T = \frac{\sum_{j=1}^{n} \frac{1}{p}(Y_j - \overline{Y}) 1_{X_j=1} - \frac{1}{1-p}(Y_j - \overline{Y}) 1_{X_j=-1}}{\sigma \sqrt{\frac{n}{p\,(1-p)}}} \tag{3.2}$$

On a :

$$T = \widetilde{T} + \overline{\eta}\, \frac{\sum_{j=1}^{n} \frac{1}{1-p} 1_{X_j=-1} - \frac{1}{p} 1_{X_j=1}}{\sigma \sqrt{\frac{n}{p\,(1-p)}}}$$

Notation 3.1 *On notera $o_P(1)$ une suite de vecteurs aléatoires qui converge vers 0 en probabilité, et $O_P(1)$ une suite bornée en probabilité.*

Par Prohorov, $\overline{\eta} = O_P(\frac{1}{\sqrt{n}})$ et $\sum_{j=1}^{n} \frac{1}{1-p} 1_{X_j=-1} - \frac{1}{p} 1_{X_j=1} = O_P(\sqrt{n})$.
Ainsi,

$$\overline{\eta}\ \frac{\sum_{j=1}^{n} \frac{1}{1-p} 1_{X_j=-1} - \frac{1}{p} 1_{X_j=1}}{\sigma \sqrt{\frac{n}{p\,(1-p)}}} \to 0$$

Il en découle que (on rappelle que l'on se situe sous H_0 ou sous H_a) :

$$T = \widetilde{T} + o_P(1)$$

T suit par conséquent les même lois asymptotiques que $\widetilde{T}$.
Il reste désormais à estimer la variance σ^2 qui demeure inconnue dans le modèle étudié. Au lieu d'utiliser la variance empirique, qui s'avère un estimateur consistant sous H_0 et

H_a par contiguïté, on propose un estimateur $\hat{\sigma}^2$ consistant sous ces deux hypothèses mais également sous une alternative non locale :

$$\hat{\sigma}^2 = \frac{1}{n}\left\{\sum_{j=1}^{n}(Y_j - \overline{Y}_1)^2\, 1_{X_j=1} \;+\; \sum_{j=1}^{n}(Y_j - \overline{Y}_{-1})^2\, 1_{X_j=-1}\right\}$$

où $\overline{Y}_1 = \frac{1}{\sum_{j=1}^{n} 1_{X_j=1}} \sum_{j=1}^{n} Y_j 1_{X_j=1}$ et $\overline{Y}_{-1} = \frac{1}{\sum_{j=1}^{n} 1_{X_j=-1}} \sum_{j=1}^{n} Y_j 1_{X_j=-1}$.

On constate que :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(Y_j)^2 1_{X_j=1} \;-\; \frac{1}{n}(\overline{Y}_1)^2 \sum_{j=1}^{n} 1_{X_j=1} \;+\; \frac{1}{n}\sum_{j=1}^{n}(Y_j)^2 1_{X_j=-1} \;-\; \frac{1}{n}(\overline{Y}_{-1})^2 \sum_{j=1}^{n} 1_{X_j=-1}$$

Par la loi des grands nombres et l'image continue :

$$\hat{\sigma}^2 \;\to\; \mathbb{E}\left(Y^2 1_{X=1}\right) \;-\; p\,\{\mathbb{E}\,(Y/X=1)\}^2 \;+\; \mathbb{E}\left(Y^2 1_{X=-1}\right) \;-\; (1-p)\,\{\mathbb{E}\,(Y/X=-1)\}^2$$

D'où, $\hat{\sigma}^2 \;\to\; \sigma^2$. Et ce pour un effet QTL q quelconque.

On conclut en effectuant un ajustement sur la statistique de test T afin d'estimer la variance :

$$T = \frac{\sum_{j=1}^{n} \frac{1}{p}(Y_j - \overline{Y})\, 1_{X_j=1} \;-\; \frac{1}{1-p}(Y_j - \overline{Y})\, 1_{X_j=-1}}{\hat{\sigma}\sqrt{\frac{n}{p\,(1-p)}}}$$

Alors :

$$T \overset{H_0}{\to} N\,(0,\,1) \quad \text{et} \quad T \overset{H_a}{\to} N\left(\frac{2a\sqrt{p\,(1-p)}}{\sigma},\,1\right)$$

Ce test présente les même lois asymptotiques qu'un test de Wald (cf. preuve en annexe 3.4.1). Les tests oracles (q) et (μ, q) sont également étudiés en annexe 3.4.2 et en annexe 3.4.3.

3.3 Etude des différentes stratégies en selective genotyping

3.3.1 Modèle correspondant au selective genotyping

On adopte le même modèle que celui de la situation oracle (cf. section 3.2.1). Seule différence, on n'observe plus la v.a. X mais la v.a. $\overline{X}$ définie de la manière suivante :

$$\overline{X} = \begin{cases} X & \text{si } Y \notin [S_-\,,\, S_+] \\ 0 & \text{sinon} \end{cases}$$

où S_- et S_+ sont deux réels tels que $S_- \leqslant S_+$.

3.3.2 Efficacités et puissances des tests correspondant aux différentes stratégies (résultats principaux)

On rappelle tout d'abord brièvement les différentes stratégies énoncées en introduction (cf. section 3.1) :

- la stratégie une est basée sur la conservation de l'ensemble des phénotypes
- les stratégies deux et trois sont basées sur la conservation uniquement des phénotypes extrêmes

Pour chacune des stratégies, afin de tester la présence de QTL, on confronte les deux hypothèses suivantes :

$$H_0 : \ q = 0 \ \ vs \ \ H_1 : \ q \neq 0$$

Plus précisément, on utilise comme hypothèse alternative, une alternative locale H_a : $q = \frac{a}{\sqrt{n}}$ où a est une constante.

Le théorème 3.3 et le lemme 3.8 énoncés dans cette section, résument les principaux résultats obtenus en terme d'efficacité et de puissance quant aux différentes stratégies.
En effet, il peut être intéressant de regarder comment évolue la puissance de chacun des tests lorsque l'on augmente le nombre d'individus, tout en gardant la même valeur de l'effet QTL q.

Soit $n^\star$ le nouveau nombre d'individus et ζ le ratio tel que $\zeta = \frac{n^\star}{n}$.

Définition 3.2 *On définit l'efficacité d'un test 2 relativement à un test 1,* $\kappa = \frac{1}{\zeta_{eff}}$ *où* ζ_{eff} *désigne la valeur de* ζ *pour laquelle la puissance du test 2 est égale à celle du test 1.*

Dans notre cas, le test oracle sert de test de référence. Tous les tests considérés sont des tests unilatéraux.

Théorème présentant les différentes efficacités

Théorème 3.3 *Soient* κ_1, κ_2 *et* κ_3 *les efficacités correspondant respectivement aux stratégies une, deux et trois, énoncées en section 3.1.*
Soient γ, γ_+ *et* γ_-, *les quantités respectives* $\mathbb{P}_{H_0}(Y \notin [S_-, \ S_+])$, $\mathbb{P}_{H_0}(Y > S_+)$ *et* $\mathbb{P}_{H_0}(Y < S_-)$.
Alors, si l'on considère un modèle statistique à trois paramètres (μ, q, σ), $\forall p$:

i) $\kappa_1 = \kappa_2 = \kappa_3 = \gamma \ + \ z_{\gamma_+}\varphi(z_{\gamma_+}) \ - \ z_{1-\gamma_-}\,\varphi(z_{1-\gamma_-})$

ii) κ_1, κ_2 *et* κ_3 *atteignent leur maximum,* M, *pour* $\gamma_+ = \gamma_- = \dfrac{\gamma}{2}$, *où*

$$M = \gamma \ + \ 2\, z_{\gamma/2}\, \varphi(z_{\gamma/2})$$

$\varphi(x)$ *et* z_α *désignant respectivement la densité d'une loi normale centrée réduite prise au point* x, *et le quantile d'ordre* $1 - \alpha$ *d'une loi normale centrée réduite.*

Ainsi, les trois stratégies présentent la même efficacité : il n'y a donc aucun gain de puissance à considérer les phénotypes non extrêmes dans l'analyse statistique et ce, quel que soit p. On rappelle que dans le modèle étudié, le cas $p = 1/2$ correspond à une population backcross.

Par la loi des grands nombres, à la fois sous l'hypothèse nulle H_0 et sous l'alternative locale H_a, $\frac{1}{n}\sum 1_{\overline{X}_j \neq 0} \to \gamma$. Ainsi, γ correspond asymptotiquement au pourcentage d'individus génotypés. De la même manière, γ_+ (resp. γ_-) correspond asymptotiquement au pourcentage d'individus génotypés à droite (resp. à gauche). D'après le théorème 3.3, les efficacités correspondant aux différentes stratégies sont maximum lorsque l'on génotype asymptotiquement le même pourcentage d'individus à droite qu'à gauche : $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.

Preuve du théorème

Première stratégie (test de Wald utilisant l'ensemble des phénotypes)

Ecriture de la vraisemblance

Le couple $(\overline{X}\ , Y)$ possède une densité par rapport à la mesure de Lebesgue $\times$ comptage.

Notation 3.4 *On note* $\forall\ i \in \{-1, 1\}$ *et* $\forall\ k \in \{-1, 0, 1\}$ *:*

$$\overline{\mathbb{P}}\{i \mid k\} = \mathbb{P}(X = i/\overline{X} = k) \quad et \quad \mathbb{P}\{k \mid i\} = \mathbb{P}(\overline{X} = k/X = i)$$

On a :

$$\mathbb{P}\{i \mid i\} = \Phi\left(\frac{S_- - \mu - iq}{\sigma}\right) + 1 - \Phi\left(\frac{S_+ - \mu - iq}{\sigma}\right)$$

où Φ désigne la fonction de répartition d'une normale centrée réduite.

On définit q_{-1} et q_1 de la manière suivante :

$$\begin{aligned} q_{-1} = \mathbb{P}(\overline{X} = -1) &= \mathbb{P}(\overline{X} = -1\ et\ X = -1) \\ &= \mathbb{P}(\overline{X} = -1/X = -1)\,(1-p) \\ &= \mathbb{P}\{-1 \mid -1\}\ (1-p) \end{aligned}$$

$$q_1 = \ \mathbb{P}(\overline{X} = 1) = \mathbb{P}\{1 \mid 1\}\ p$$

De la même façon :

$$q_0 \;=\; \mathbb{P}(\overline{X}=0) \;=\; (1-\mathbb{P}\{-1 \mid -1\})\,(1-p) \;+\; (1-\mathbb{P}\{1 \mid 1\})\,p$$

D'où :

$$\overline{\mathbb{P}}\{-1 \mid k\} = \frac{\mathbb{P}\{k \mid -1\}\;(1-p)}{q_k} \quad , \quad \overline{\mathbb{P}}\{1 \mid k\} = \frac{\mathbb{P}\{k \mid 1\}\;p}{q_k}$$

On a $\forall\, k \in \{-1,1\}$ et $\forall\, y \in \mathbb{R}$:

$$\begin{aligned}\mathbb{P}(Y \in [y\,,\; y+dy]\,/\overline{X}=k) &= \mathbb{P}(Y \in [y\,,\; y+dy]\,/X=k \; et\; \overline{X}\neq 0)\\ &= \frac{\mathbb{P}(Y \in [y\,,\; y+dy]\;\; et\; \overline{X}\neq 0\;/\;X=k)}{\mathbb{P}(\overline{X}\neq 0/X=k)}\\ &= \frac{\frac{1}{\sigma}\,\varphi(\frac{y-\mu-kq}{\sigma})1_{y\notin[S_-\,,\;S_+]}}{\mathbb{P}\{k \mid k\}}\,dy\end{aligned}$$

$$\begin{aligned}\mathbb{P}(Y \in [y\,,\; y+dy]\;\; et\; \overline{X}=k) &= \mathbb{P}(Y \in [y\,,\; y+dy]\,/\overline{X}=k)\,\mathbb{P}(\overline{X}=k)\\ &= \frac{\frac{1}{\sigma}\,\varphi(\frac{y-\mu-kq}{\sigma})1_{y\notin[S_-\,,\;S_+]}}{\mathbb{P}\{k \mid k\}}\,q_k\,dy\end{aligned}$$

D'où

$$\mathbb{P}(Y \in [y\,,\; y+dy]\;\; et\; \overline{X}=-1) = \frac{1-p}{\sigma}\,\varphi\left(\frac{y-\mu+q}{\sigma}\right)1_{y\notin[S_-\,,\;S_+]}\,dy$$

$$\mathbb{P}(Y \in [y\,,\; y+dy]\;\; et\; \overline{X}=1) = \frac{p}{\sigma}\,\varphi\left(\frac{y-\mu-q}{\sigma}\right)1_{y\notin[S_-\,,\;S_+]}\,dy$$

De plus :

$$\begin{aligned}\mathbb{P}(Y \in [y\,,\; y+dy]\,/\overline{X}=0) &= \sum_{i\in\{-1,1\}} \mathbb{P}(Y \in [y\,,\; y+dy]\;\; et\; X=i\;/\;\overline{X}=0)\\ &= \sum_{i\in\{-1,1\}} \mathbb{P}(Y \in [y\,,\; y+dy]\,/X=i\; et\; \overline{X}=0)\;\;\overline{\mathbb{P}}\{i \mid 0\}\\ &= \frac{\frac{p}{\sigma}\,\varphi(\frac{y-\mu-q}{\sigma})1_{y\in[S_-\,,\;S_+]}}{q_0}\,dy \;+\; \frac{\frac{1-p}{\sigma}\,\varphi(\frac{y-\mu+q}{\sigma})1_{y\in[S_-\,,\;S_+]}}{q_0}\,dy\end{aligned}$$

Alors,

$$\begin{aligned}&\mathbb{P}(Y \in [y\ ,\ y+dy]\ et\ \overline{X}=0) = \mathbb{P}(Y \in [y\ ,\ y+dy]\,/\overline{X}=0)\ \mathbb{P}(\overline{X}=0)\\ &= \frac{p}{\sigma}\,\varphi\left(\frac{y-\mu-q}{\sigma}\right)1_{y\in[S_-\ ,\ S_+]}\,dy\ +\ \frac{1-p}{\sigma}\,\varphi\left(\frac{y-\mu+q}{\sigma}\right)1_{y\in[S_-\ ,\ S_+]}\,dy\end{aligned}$$

La vraisemblance L pour une observation $\left(\overline{X}, Y\right)$ s'écrit :

$$\begin{aligned}L = &\frac{1-p}{\sigma}\,\varphi\left(\frac{y-\mu+q}{\sigma}\right)1_{\overline{X}=-1}\ +\ \frac{p}{\sigma}\,\varphi\left(\frac{y-\mu-q}{\sigma}\right)1_{\overline{X}=1}\\ &+\left\{\frac{1-p}{\sigma}\,\varphi\left(\frac{y-\mu+q}{\sigma}\right)\ +\ \frac{p}{\sigma}\,\varphi\left(\frac{y-\mu-q}{\sigma}\right)\right\}1_{\overline{X}=0}\end{aligned}$$

Test statistique (μ, q)

Afin de calculer la statistique de Wald, on introduit tout d'abord un lemme démontré dans le cadre de cette thèse. La démonstration étant très lourde en terme de calculs, elle n'est pas présentée ici.

Lemme 3.5 *Soit* $Y \sim N(\mu,\ \sigma^2)$, *alors :*

1.

$$\begin{aligned}\mathbb{E}\left(Y^2 1_{Y\notin[S_-,\ S_+]}\right) = &\left(\mu^2+\sigma^2\right)\mathbb{P}(Y \notin [S_-,\ S_+])\ +\ \sigma\,(S_+ + \mu)\,\varphi\left(\frac{S_+-\mu}{\sigma}\right)\\ &-\sigma\,(S_- + \mu)\,\varphi\left(\frac{S_--\mu}{\sigma}\right)\end{aligned}$$

2.

$$\mathbb{E}\left(Y 1_{Y\notin[S_-,\ S_+]}\right) =\ \mu\,\mathbb{P}(Y \notin [S_-,\ S_+])\ +\ \sigma\,\varphi\left(\frac{S_+-\mu}{\sigma}\right)\ -\ \sigma\,\varphi\left(\frac{S_--\mu}{\sigma}\right)$$

3.

$$\begin{aligned}\mathbb{E}\left\{(Y-\mu)^2 1_{Y\notin[S_-,\ S_+]}\right\} = &\sigma^2\,\mathbb{P}(Y \notin [S_-,\ S_+])\ +\ \sigma\,(S_+ - \mu)\,\varphi\left(\frac{S_+-\mu}{\sigma}\right)\\ &-\sigma\,(S_- - \mu)\,\varphi\left(\frac{S_--\mu}{\sigma}\right)\end{aligned}$$

4.

$$\mathbb{E}\left\{(Y-\mu) 1_{Y\notin[S_-,\ S_+]}\right\} = \sigma\,\varphi\left(\frac{S_+-\mu}{\sigma}\right)\ -\ \sigma\,\varphi\left(\frac{S_--\mu}{\sigma}\right)$$

5.

$$\mathbb{E}\left\{(Y-\mu)^2 1_{Y\in[S_-,\,S_+]}\right\} = \sigma^2 \;-\; \sigma^2\,\mathbb{P}(Y\notin[S_-,\,S_+]) \;-\; \sigma\,(S_+-\mu)\,\varphi\left(\frac{S_+-\mu}{\sigma}\right) + \sigma\,(S_--\mu)\,\varphi\left(\frac{S_--\mu}{\sigma}\right)$$

Notation 3.6 *On note $\mathcal{A}$ la quantité suivante :*

$$\mathcal{A} := \sigma^2\,\left\{\gamma \;+\; z_{\gamma_+}\,\varphi(z_{\gamma_+}) \;-\; z_{1-\gamma_-}\,\varphi(z_{1-\gamma_-})\right\}$$

D'après le lemme ci-dessus, on a $\mathcal{A} = \mathbb{E}_{H_0}\left\{(Y-\mu)^2 1_{Y\notin[S_-,\,S_+]}\right\}$.
On pose $\theta = (\mu,\, q)$ et $\theta_0 = (\mu,\, 0)$. Le calcul de la fonction score et de l'information de Fisher est le suivant :

$$\frac{\partial logL}{\partial q}\,|_{\theta_0} = -\left(\frac{y-\mu}{\sigma^2}\right) 1_{\overline{X}=-1} \;+\; \left(\frac{y-\mu}{\sigma^2}\right) 1_{\overline{X}=1} \;+\; \left(\frac{y-\mu}{\sigma^2}\right)(2p-1) 1_{\overline{X}=0}$$

$$\left(\frac{\partial logL}{\partial q}\,|_{\theta_0}\right)^2 = \quad \frac{(y-\mu)^2}{\sigma^4} 1_{\overline{X}=-1} \;+\; \frac{(y-\mu)^2}{\sigma^4} 1_{\overline{X}=1} \;+\; \frac{(y-\mu)^2}{\sigma^4}\,(2p-1)^2\,1_{\overline{X}=0}$$

D'où :

$$I_{22}(\theta_0) = \frac{\mathcal{A}}{\sigma^4} \;+\; \frac{(2p-1)^2}{\sigma^4}\,(\sigma^2-\mathcal{A})$$

De plus,

$$\frac{\partial logL}{\partial \mu}\,|_{\theta_0} = \frac{y-\mu}{\sigma^2} \quad \text{d'où} \quad I_{11}(\theta_0) \;=\; \frac{1}{\sigma^2}$$

$$\frac{\partial logL}{\partial q\,\partial \mu}\,|_{\theta_0} = \frac{1}{\sigma^2} 1_{\overline{X}=-1} \;-\; \frac{1}{\sigma^2} 1_{\overline{X}=1} \;-\; \frac{1}{\sigma^2}\,(2p-1) 1_{\overline{X}=0}$$

Comme on se place sous H_0, $P_{H_0}\{-1 \mid -1\} = P_{H_0}\{1 \mid 1\}$. D'où :

$$I_{12}(\theta_0) = \frac{1}{\sigma^2}(2p-1)$$

On en déduit :

$$I_{22}^{-1}(\theta_0) = \frac{\sigma^4}{4\,\mathcal{A}\,p(1-p)}$$

On notera $\hat{q}$ l'EMV de q. Il sera obtenu au moyen de l'algorithme EM présenté en annexe 3.4.4. Comme le modèle est régulier :

$$\sqrt{n}\,\hat{q} \overset{H_0}{\rightarrow} N(\,0\,,\, I_{22}^{-1}(\theta_0)\,)$$

D'où le test :

$$W_1 := \frac{2\sqrt{n}}{\sigma^2}\sqrt{\mathcal{A}\,p(1-p)}\;\hat{q} \overset{H_0}{\rightarrow} N(0,\,1)$$

En appliquant le théorème 2.9 avec $h_n = h = (0, a)$:

$$W_1 \overset{H_a}{\rightarrow} N\left(\frac{2a}{\sigma^2}\sqrt{\mathcal{A}\,p(1-p)},\,1\right) \tag{3.3}$$

Deuxième stratégie (comparaison de moyenne basée uniquement sur les phénotypes extrêmes)

Test statistique (μ, q, σ)

Soit $\hat{\delta}$ l'estimateur vérifiant :

$$\hat{\delta} = \frac{1}{p}(Y-\mu)1_{X=1} \;-\; \frac{1}{1-p}(Y-\mu)1_{X=-1}$$

D'après la formule (3.1) page 43, $\mathbb{E}_{H_a}(\hat{\delta}) = 2q$ lorsque l'on se situe dans la situation oracle. $\hat{\delta}$ est donc un estimateur sans biais de deux fois l'effet QTL. Si l'on se replace désormais dans le cadre du selective genotyping, on est tenté de définir $\hat{\delta}$ de la manière suivante :

$$\hat{\delta} = \frac{1}{p}(Y-\mu)1_{\overline{X}=1} \;-\; \frac{1}{1-p}(Y-\mu)1_{\overline{X}=-1}$$

D'après le lemme 3.5 page 49 :

$$\begin{aligned}\mathbb{E}\left(\hat{\delta}\right) &= \frac{1}{p}\,\mathbb{E}\left(Y-\mu/\overline{X}=1\right)\mathbb{P}(\overline{X}=1) \;-\; \frac{1}{1-p}\,\mathbb{E}\left(Y-\mu/\overline{X}=-1\right)\mathbb{P}(\overline{X}=-1)\\ &= q\left(\mathbb{P}\left\{1\mid 1\right\}+\mathbb{P}\left\{-1\mid -1\right\}\right) \;+\; \sigma\,\varphi\left(\frac{S_+-\mu-q}{\sigma}\right) \;-\; \sigma\,\varphi\left(\frac{S_--\mu-q}{\sigma}\right)\\ &\quad -\; \sigma\,\varphi\left(\frac{S_+-\mu+q}{\sigma}\right) \;+\; \sigma\,\varphi\left(\frac{S_--\mu+q}{\sigma}\right)\end{aligned}$$

Ce n'est donc plus un bon estimateur de deux fois l'effet QTL. Cependant, comme l'espérance de $\hat{\delta}$ dépend de q, on va construire une statistique de test basée sur $\hat{\delta}$.
On constate que :

$$\mathbb{E}_{H_0}\left(\hat{\delta}\right) = 0$$

D'où :

$$\mathrm{Var}_{H_0}\left(\hat{\delta}\right) = \mathbb{E}_{H_0}\left(\hat{\delta}^2\right)$$

On a :

$$\hat{\delta}^2 = \frac{1}{p^2}\,(Y-\mu)^2\,1_{\overline{X}=1} \;+\; \frac{1}{(1-p)^2}\,(Y-\mu)^2\,1_{\overline{X}=-1}$$

D'après le lemme 3.5 :

$$\begin{aligned}\mathbb{E}\left(\hat{\delta}^2\right) &= \frac{1}{p^2}\,\mathbb{E}\left\{(Y-\mu)^2/\overline{X}=1\right\}\mathbb{P}(\overline{X}=1) \;+\; \frac{1}{(1-p)^2}\,\mathbb{E}\left\{(Y-\mu)^2/\overline{X}=-1\right\}\mathbb{P}(\overline{X}=-1)\\ &= \frac{1}{p}\,\mathbb{E}\left\{(Y-\mu)^2 1_{Y\notin[S_-,S_+]}/X=1\right\} \;+\; \frac{1}{1-p}\,\mathbb{E}\left\{(Y-\mu)^2 1_{Y\notin[S_-,S_+]}/X=-1\right\}\end{aligned}$$

D'où :

$$\mathbb{E}_{H_0}\left(\hat{\delta}^2\right) = \frac{\mathcal{A}}{p(1-p)}$$

Ainsi, on peut définir la statistique de test T_2 correspondant à la deuxième stratégie. D'après le théorème central limite :

$$T_2 := \frac{\sum_{j=1}^{n} \frac{1}{p}(Y_j-\mu)1_{\overline{X}_j=1} \;-\; \frac{1}{1-p}(Y_j-\mu)1_{\overline{X}_j=-1}}{\sqrt{\frac{n\,\mathcal{A}}{p(1-p)}}} \;\xrightarrow{H_0}\; N(0,\,1) \tag{3.4}$$

On utilise le développement limité à l'ordre 1 de la fonction exponentielle. Ainsi :

$$\varphi\left(\frac{S_- - \mu + q}{\sigma}\right) = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{S_- - \mu}{\sigma}\right)^2}\,\left\{1 - \frac{(S_- - \mu)\,q}{\sigma^2} + o(q)\right\}$$

Et on a également (en travaillant sur les intégrales) :

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

Il en découle que :

$$\mathbb{E}_{H_a}\left(T_2\right) \to \frac{2a}{\sqrt{\frac{\mathcal{A}}{p(1-p)}}}\left\{\gamma - z_{1-\gamma_-}\varphi(z_{1-\gamma_-}) + z_{\gamma_+}\varphi(z_{\gamma_+})\right\}$$

On peut remarquer que cette limite est égale à $\frac{2a}{\sigma^2}\sqrt{\mathcal{A}\,p(1-p)}$.
De plus, $\mathbb{E}_{H_a}\left[\hat{\delta}\right] \to 0$.
En appliquant Portmanteau (car $\forall i \in \{-1,1\}$, $Y/X=i \to N(\mu,\,\sigma^2)$) :

$$\mathbb{E}_{H_a}\left[\hat{\delta}^2\right] \to \frac{\mathcal{A}}{p(1-p)}$$

D'où :

$$\mathrm{Var}_{H_a}\left[\hat{\delta}\right] \to \mathrm{Var}_{H_0}\left[\hat{\delta}\right]$$

Ainsi :

$$T_2 \overset{H_a}{\to} N\left(\frac{2a}{\sigma^2}\sqrt{\mathcal{A}\,p(1-p)},\ 1\right) \tag{3.5}$$

Comme μ et σ sont inconnus, on doit procéder à des ajustements sur la statistique de test T_2.
On peut remplacer μ par $\hat{\mu}$, estimateur dépendant des phénotypes extrêmes. $\hat{\mu}$ peut être obtenu par maximum de vraisemblance ou par la méthode des moments, ces deux estimateurs étant $\sqrt{n}$ consistants.
De plus, comme estimateur consistant de $\mathcal{A}$, on propose $\hat{\mathcal{A}}$:

$$\hat{\mathcal{A}} := \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{\mu})^2 1_{\overline{X}_j \neq 0}$$

Les lois asymptotiques de la statistique T_2 demeurent inchangées.

Remarque 3.7 *Il est possible de choisir comme estimateur $\hat{\mu}$, la moyenne empirique, ce qui rend ce test facile à mettre en oeuvre. Cependant, dans ce cas, on utilise les phénotypes non extrêmes dans l'analyse statistique.*

Efficacité du test

On se propose de calculer l'efficacité de ce test de comparaison de moyenne en selective genotyping, relativement au test oracle (μ, q, σ) où tous les génotypes sont connus.

Soit $n^\star$ le nombre total d'individus considérés pour une expérience en selective genotyping. Jusqu'ici, on avait considéré $n^\star = n$. Si désormais, on pose $\zeta = \frac{n^\star}{n}$, alors :

$$T_2 = \frac{\sum_{j=1}^{n^\star}\frac{1}{p}(Y_j - \hat{\mu})1_{\overline{X}_j=1} \ - \ \frac{1}{1-p}(Y_j - \hat{\mu})1_{\overline{X}_j=-1}}{\sqrt{\frac{n^\star\,\hat{\mathcal{A}}}{p(1-p)}}} \overset{H_0}{\to} N(0,\ 1)$$

Et :

$$T_2 \overset{H_a}{\to} N\left(\frac{2a}{\sigma^2}\sqrt{\zeta\,\mathcal{A}\,p(1-p)},\ 1\right)$$

On s'intéressera plus particulièrement au test unilatéral approprié lorsque a est supérieur à zéro.
Afin de calculer le ζ à partir duquel ce test est plus puissant que le test oracle (μ, q, σ),

on est amené à résoudre l'inéquation suivante (en supposant $a > 0$) :

$$\begin{aligned} & z_\alpha \; - \; \frac{2a}{\sigma^2} \sqrt{\zeta \, \mathcal{A} \, p(1-p)} \; < \; z_\alpha \; - \; \frac{2a \sqrt{p(1-p)}}{\sigma} \\ \Leftrightarrow \quad & \zeta \; > \; \frac{\sigma^2}{\mathcal{A}} \end{aligned}$$

On rappelle que ζ_{eff} désigne la valeur de ζ pour laquelle la puissance du test est égale à celle test oracle (μ, q, σ). Ici :

$$\zeta_{eff} = \frac{\sigma^2}{\mathcal{A}}$$

D'où, l'efficacité κ_2 vérifie :

$$\kappa_2 = \frac{\mathcal{A}}{\sigma^2}$$

Ainsi,

$$\kappa_2 = \gamma \; + \; z_{\gamma_+} \varphi(z_{\gamma_+}) \; - \; z_{1-\gamma_-} \, \varphi(z_{1-\gamma_-}) \tag{3.6}$$

Preuve du i) du théorème

On rappelle tout d'abord le i) du théorème :

Si l'on considère un modèle statistique à trois paramètres (μ, q, σ), alors $\forall p$:

$$i) \;\; \kappa_1 = \kappa_2 = \kappa_3 = \gamma \; + \; z_{\gamma_+} \varphi(z_{\gamma_+}) \; - \; z_{1-\gamma_-} \, \varphi(z_{1-\gamma_-})$$

où κ_1, κ_2 et κ_3 sont les efficacités correspondant aux stratégies une, deux et trois.

Afin de prouver ce point i), on remarque que la statistique T_2 correspondant au test (μ, q, σ) de la deuxième stratégie, et que la statistique de Wald W_1 correspondant au test (μ, q) de la première stratégie présentent les mêmes lois asymptotiques (cf. formule 3.5 page 53 et formule 3.3 page 51). Or par définition, le test de Wald (μ, q, σ) correspondant à la première stratégie est supérieur ou égal en terme de puissance au test (μ, q, σ) de la deuxième stratégie. On en déduit que ces deux tests présentent les mêmes lois asymptotiques.

De la même manière, par définition, le test de Wald (μ, q, σ) correspondant à la troisième stratégie est compris en terme de puissance entre ces deux derniers tests. On conclut donc que les tests correspondant aux trois stratégies présentent les mêmes puissances.

Par conséquent, ces trois tests présentent les mêmes efficacités. D'après la formule (3.6) ci-dessus :

$$\kappa_1 = \kappa_2 = \kappa_3 = \gamma \; + \; z_{\gamma_+} \varphi(z_{\gamma_+}) \; - \; z_{1-\gamma_-} \, \varphi(z_{1-\gamma_-})$$

Preuve du ii) du théorème

On rappelle tout d'abord le ii) du théorème :

Si l'on considère un modèle statistique à trois paramètres (μ, q, σ), alors $\forall p$:

ii) κ_1, κ_2 *et* κ_3 *atteignent leur maximum,* M, *pour* $\gamma_+ = \gamma_- = \dfrac{\gamma}{2}$, *où*

$$M = \gamma \; + \; 2\, z_{\gamma/2}\, \varphi(z_{\gamma/2})$$

$\varphi(x)$ et z_α désignant respectivement la densité d'une loi normale centrée réduite prise au point x, et le quantile d'ordre $1-\alpha$ d'une loi normale centrée réduite.

Afin de prouver ce point ii), on cherche à répondre à la question suivante : comment choisir γ_+ et γ_- afin d'obtenir une efficacité maximale ? On rappelle que l'on a la relation $\gamma_+ + \gamma_- = \gamma$.

Soit $g(.)$ la fonction telle que : $g(z_{\gamma_+}) = \Phi^{-1}\left\{\gamma - 1 + \Phi(z_{\gamma_+})\right\}$. Ainsi, $z_{1-\gamma_-} = g(z_{\gamma_+})$.

Afin de maximiser κ_1, il faut maximiser la fonction $k_1(.)$ définie de la manière suivante :

$$k_1(z_{\gamma_+}) = z_{\gamma_+}\varphi(z_{\gamma_+}) - g(z_{\gamma_+})\, \varphi\left\{g(z_{\gamma_+})\right\}$$

On a :

$$k_1'(z_{\gamma_+}) = \varphi(z_{\gamma_+}) \; + \; z_{\gamma_+}\varphi'(z_{\gamma_+}) \; - \; g'(z_{\gamma_+})\, \varphi\left\{g(z_{\gamma_+})\right\} \; - \; g(z_{\gamma_+})\, g'(z_{\gamma_+})\, \varphi'\left\{g(z_{\gamma_+})\right\}$$

$$g'(z_{\gamma_+}) = \frac{\varphi(z_{\gamma_+})}{\varphi(z_{1-\gamma_-})}$$

Alors :

$$k_1'(z_{\gamma/2}) = \varphi(z_{\gamma/2}) \; - \; \left\{z_{\gamma/2}\right\}^2 \, \varphi(z_{\gamma/2}) \; - \; \varphi(z_{1-\gamma/2}) \; + \; \left\{z_{1-\gamma/2}\right\}^2 \, \varphi(z_{1-\gamma/2}) = 0$$

On conclut que l'efficacité κ_1 est maximale lorsque $\gamma_+ = \gamma_- = \frac{\gamma}{2}$. ■

Lemme présentant les différentes statistiques de test

Lemme 3.8 *Si l'on considère un modèle statistique à trois paramètres (μ, q, σ), la statistique de Wald W_1, la statistique de comparaison de moyenne T_2, la statistique de Wald W_3, correspondant respectivement aux stratégies une, deux et trois :*

$$W_1 := \frac{2\sqrt{n}}{\hat{\sigma}^2}\, \sqrt{\hat{\mathcal{A}}\, p(1-p)}\; \hat{q}_1$$

$$T_2 := \sqrt{p(1-p)}\, \left\{ \frac{\sum_{j=1}^{n} \frac{1}{p}(Y_j - \overline{Y}) 1_{\overline{X}_j = 1} \; - \; \frac{1}{1-p}(Y_j - \overline{Y}) 1_{\overline{X}_j = -1}}{\sqrt{n\, \hat{\mathcal{A}}}} \right\}$$

$$W_3 := \frac{2\sqrt{n}}{\hat{\sigma}^2}\, \sqrt{\hat{\mathcal{A}}\, p(1-p)}\; \hat{q}_3$$

présentent les mêmes lois asymptotiques sous H_0 et sous H_a, à savoir :

$$N(0,\ 1) \quad et \quad N\left(\frac{2a\ \sqrt{\mathcal{A}\ p(1-p)}}{\sigma^2},\ 1\right)$$

$\hat{q}_1$ et $\hat{q}_3$ désignent les EMV de q pour les stratégies une et trois alors que $\mathcal{A}$, $\hat{\mathcal{A}}$, $\overline{Y}$ sont définis de la manière suivante :

$$\hat{\mathcal{A}} = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \overline{Y})^2 1_{\overline{X}_j \neq 0} \quad , \quad \overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

$$\mathcal{A} = \sigma^2\ \left\{\gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-})\right\} \quad , \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2$$

Preuve.
cf. preuve du théorème 3.3. A noter que les estimateurs $\hat{\sigma}^2$ et $\hat{\mathcal{A}}$ sont consistants également sous H_a par contiguïté. ■

On peut remarquer que dans l'expression de la statistique T_2, figure la quantité $\overline{Y}$. Or par définition, la stratégie deux repose uniquement sur les phénotypes extrêmes. Cependant, il est possible de remplacer $\overline{Y}$ dans l'expression de T_2 par $\hat{\mu}$, estimateur de μ dépendant uniquement des phénotypes extrêmes. $\hat{\mu}$ peut être obtenu par maximum de vraisemblance ou par la méthode des moments, ces deux estimateurs étant $\sqrt{n}$ consistants. Dans ce lemme, le choix comme estimateur de μ a été porté sur $\overline{Y}$, afin de proposer une statistique de test la plus simple possible.

L'EMV $\hat{q}_1$ de q pour la stratégie une, pourra être obtenu par l'algorithme EM. Une version de cet algorithme est présentée en annexe 3.4.4 pour un modèle statistique (μ, q).

L'EMV $\hat{q}_3$ de q pour la stratégie trois, pourra être calculé au moyen d'une méthode de Newton.

Illustration graphique

La figure 3.1 présente l'efficacité en fonction de γ. Cette figure a été réalisée sous l'hypothèse que l'on génotypait asymptotiquement autant d'individus à gauche qu'à droite ($\gamma_+ = \gamma_- = \frac{\gamma}{2}$).

On constate que l'efficacité croît avec γ. De plus, lorsque $\gamma = 0$, l'efficacité est nulle. En effet, la loi sous l'alternative est la même que celle sous l'hypothèse nulle (cf. lemme 3.8). Bien au contraire, lorsque $\gamma = 1$, l'efficacité est égale à un : les tests coïncident avec le test oracle.

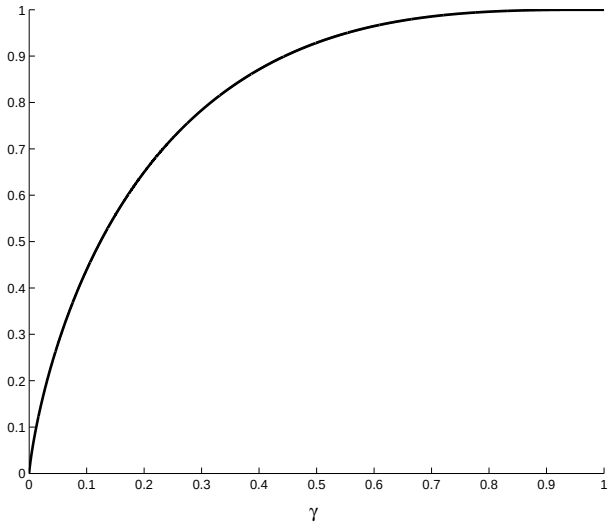


FIG. 3.1 – Efficacité en fonction de γ ($\gamma_+ = \gamma_- = \frac{\gamma}{2}$)

3.3.3 Résultats secondaires

On présente, dans cette section, des résultats à propos des différentes stratégies pour les modèles statistiques (μ, q) et (q).

Résultat théorique pour le modèle (q)

Corollaire 3.9 *Si l'on considère un modèle statistique à un paramètre (q), alors :*

i) $\kappa_1 = \gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-}) + (2p-1)^2\left\{1-\gamma-z_{\gamma_+}\varphi(z_{\gamma_+})+z_{1-\gamma_-}\varphi(z_{1-\gamma_-})\right\}$

ii) $\kappa_2 = 4\,p\,(1-p)\left\{\gamma - z_{1-\gamma_-}\varphi(z_{1-\gamma_-}) + z_{\gamma_+}\varphi(z_{\gamma_+})\right\}$

iii) $\kappa_3 = \gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-}) + \dfrac{(2p-1)^2}{1-\gamma}\left\{\varphi(z_{1-\gamma_-}) - \varphi(z_{\gamma_+})\right\}^2 \quad \forall\gamma \neq 1$

iv) $\kappa_1 = \kappa_2 = \kappa_3 \Leftrightarrow p = \dfrac{1}{2}$

v) $\forall p$ κ_1, κ_2 *et* κ_3 *sont maximum pour* $\gamma_+ = \gamma_- = \dfrac{\gamma}{2}$

Preuve. La preuve est donnée en annexe 3.4.5. ■

Contrairement au résultat obtenu pour un modèle (μ, q, σ), les trois stratégies sont ici équivalentes uniquement lorsque l'on considère une population backcross ($p = 1/2$).

Comme tous les résultats établis dans ce corollaire sont des résultats asymptotiques, on illustre en annexe 3.4.6 la convergence vers l'asymptotique. Les résultats théoriques

établis sont très bons lorsque l'effet QTL est très petit. Ils se comportent également très bien lorsque l'effet du QTL grossit et ce, quelle que soit la valeur de γ.

Illustration graphique du modèle (q)

Toutes les figures de cette section ont été réalisées sous l'hypothèse que l'on génotypait asymptotiquement autant d'individus à gauche qu'à droite ($\gamma_+ = \gamma_- = \frac{\gamma}{2}$).

La figure 3.2 présente les différentes efficacités en fonction de γ et pour quelques valeurs de p choisies.
On remarque tout d'abord que lorsque $\gamma = 1$ et $p = \frac{1}{4}$, on a une efficacité égale à un, pour les stratégies une et trois. Cela signifie que ces tests coïncident avec le test oracle. Ce n'est pas étonnant car l'ensemble des individus est génotypé. Cependant, si l'on s'intéresse à la stratégie deux, on remarque que l'efficacité est d'environ 70% lorsque tous les individus sont génotypés. Ceci s'explique par le fait que le test coïncide alors, avec un test basé sur une statistique T (cf. page 66) et qui est aussi puissant que le test oracle uniquement lorsque $p = \frac{1}{2}$.
Si désormais on observe le cas $\gamma = 0$ et $p = \frac{1}{4}$, les puissances des tests associés aux stratégies deux et trois sont nulles. Cependant, l'efficacité de la stratégie une est de $(2p-1)^2 = 0.25$.
Enfin, lorque $p = \frac{1}{2}$, les trois stratégies présentent la même efficacité. Bien évidemment, lorsque $p \neq \frac{1}{2}$, on a $\kappa_1 > \kappa_3 > \kappa_2$.

La figure 3.3 présente les efficacités en fonction de p, et pour $\gamma = 0.3$. On observe le rôle symétrique que joue p pour chacune des stratégies. Les trois stratégies coïncident évidemment pour $p = \frac{1}{2}$. On constate que la situation $p = \frac{1}{2}$ correspond à :

- la pire des situations pour la stratégie une
- la meilleure des situations pour la stratégie deux

L'efficacité liée à la stratégie trois est constante en raison du contexte de symétrie ($\gamma_+ = \gamma_- = \frac{\gamma}{2}$).

Pour finir, les figures 3.4 représentent les efficacités, en fonction de p et γ, des tests correspondant aux stratégies une et deux. On peut facilement en déduire l'évolution de la figure 3.3 avec γ.

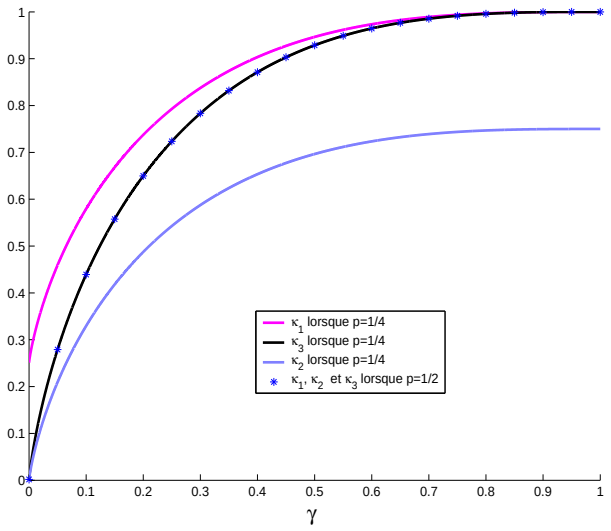


FIG. 3.2 – Efficacité des tests, correspondant aux différentes stratégies, en fonction de γ ($\gamma_+ = \gamma_- = \frac{\gamma}{2}$)

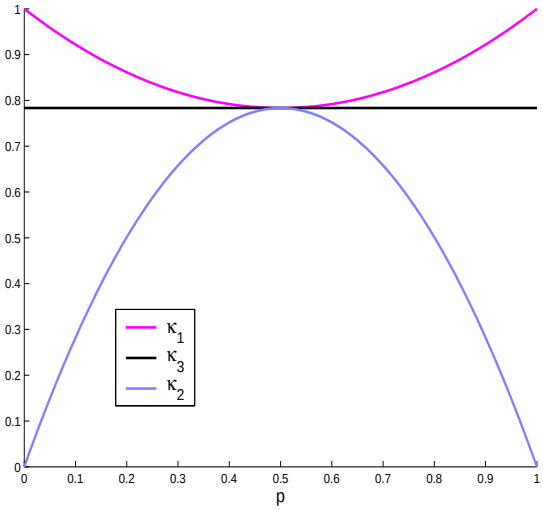
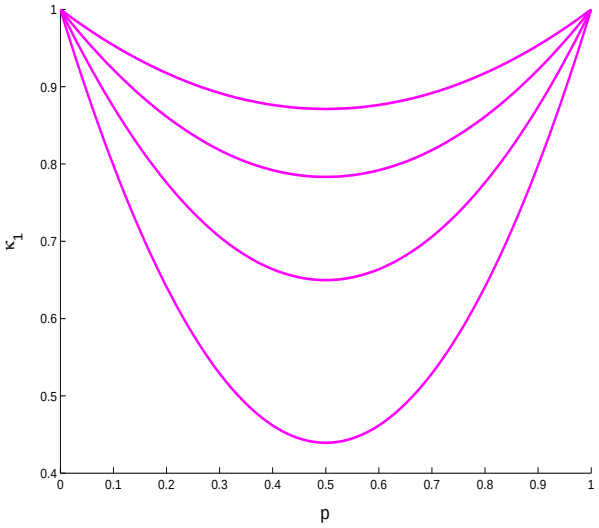
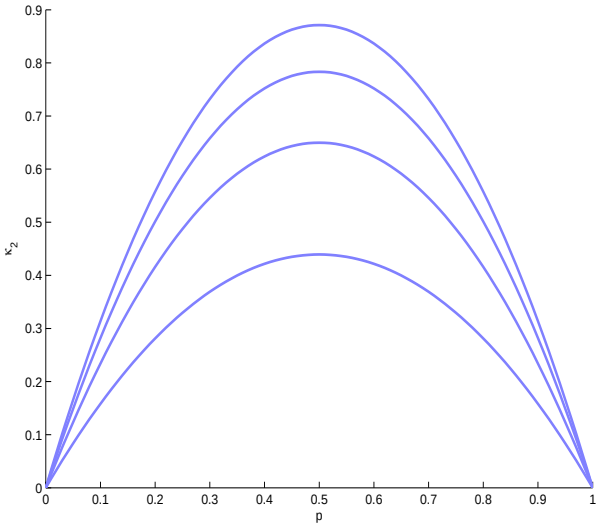


FIG. 3.3 – Efficacité des tests, correspondant aux différentes stratégies, en fonction de p ($\gamma = 0.3$, $\gamma_+ = \gamma_- = \frac{\gamma}{2}$)



Première stratégie



Deuxième stratégie

FIG. 3.4 – Efficacité, en fonction de p et γ, des tests correspondant aux stratégies une et deux. De bas en haut, $\gamma = 0.1$, 0.2 , 0.3 , 0.4 ($\gamma_+ = \gamma_- = \frac{\gamma}{2}$)

Résultat théorique pour le modèle (μ, q)

Corollaire 3.10 *Si l'on considère un modèle statistique à deux paramètres* (μ, q) *alors les résultats sont identiques à ceux énoncés en théorème 3.3.*

Preuve. La preuve est évidente au vu de la preuve du théorème 3.3. A noter que le test oracle (μ, q, σ) et le test oracle (μ, q) présentent les mêmes puissances (cf. annexe 3.4.3). ■

3.3.4 Résumé des différents résultats

Les tables 3.1 et 3.2, résument les différents résultats obtenus à propos des différentes stratégies.

Modèle	Stratégie	Décentrement
(μ, q, σ) et (μ, q)	1, 2 et 3	$\frac{2a\sqrt{\mathcal{A}\ p(1-p)}}{\sigma^2}$
(q)	1	$\frac{a\sqrt{\mathcal{A}+(2p-1)^2(\sigma^2-\mathcal{A})}}{\sigma^2}$
	2	$\frac{2a\sqrt{\mathcal{A}\ p(1-p)}}{\sigma^2}$
	3	$\frac{a}{\sigma}\sqrt{\frac{\mathcal{A}}{\sigma^2}+\frac{(2p-1)^2\left\{\varphi\left(z_{\gamma_+}\right)-\varphi\left(z_{1-\gamma_-}\right)\right\}^2}{1-\gamma}}$

TAB. 3.1 – Table de décentrement

Modèle	Stratégie	Efficacité
(μ, q, σ) et (μ, q)	1, 2 et 3	$\gamma \ + \ z_{\gamma_+}\varphi(z_{\gamma_+}) \ - \ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})$
(q)	1	$\gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-}) + (2p-1)^2\ \left\{1 - \gamma - z_{\gamma_+}\varphi(z_{\gamma_+}) + z_{1-\gamma_-}\varphi(z_{1-\gamma_-})\right\}$
	2	$4\,p\,(1-p)\ \left\{\gamma \ - \ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-}) \ + \ z_{\gamma_+}\ \varphi(z_{\gamma_+})\right\}$
	3	$\gamma \ + \ z_{\gamma_+}\varphi(z_{\gamma_+}) \ - \ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-}) \ + \ \frac{(2p-1)^2}{1\ -\gamma}\left\{\varphi(z_{1-\gamma_-}) \ - \ \varphi(z_{\gamma_+})\right\}^2$

TAB. 3.2 – Table d'efficacité

3.3.5 Remarques générales sur la modélisation du selective genotyping

Afin de modéliser le selective genotyping, deux seuils fixes S_- et S_+ ont été considérés. Un individu est génotypé uniquement si la valeur du caractère quantitatif Y n'appartient pas à l'intervalle $[S_-,\ S_+]$. En choisissant S_+ et S_- de telle sorte que $\mathbb{P}_{H_0}(Y \notin [S_-, S_+]) = \gamma$, par la loi des grands nombres, le pourcentage d'individus génotypés tend asymptotiquement vers γ, que l'on se trouve sous l'hypothèse nulle ou sous l'alternative locale. Ainsi cette modélisation est en accord avec la définition usuelle du selective genotyping : le selective genotyping consiste à génotyper uniquement les $\gamma\%$ de la population présentant des phénotypes extrêmes.

Dans Darvasi and Soller (1992), les auteurs considérent le cas d'une population backcross (ie. $p = 1/2$) et s'intéressent à une comparaison de moyenne sur les phénotypes extrêmes. La différence avec notre approche est qu'ils considèrent des seuils qui varient avec l'effet du QTL. En effet, les auteurs imposent $\mathbb{P}(Y \notin [S_-, S_+]) = \gamma$. Ils utilisent par la suite une approximation à propos des seuils qui n'a pas lieu d'être (cf. formule 1 et 2 de l'article) puis des résultats sur des tailles échantillonales (cf. formule 24) non applicables pour un modèle avec une alternative locale.

Il faudrait reprendre la preuve de cet article tout d'abord en utilisant une méthode de Newton pour le calcul des seuils et non pas l'approximation proposée. L'effet du QTL étant tel que $q = \frac{a}{\sqrt{n}}$, les seuils S_- et S_+ dépendent donc du nombre d'observations n, on ne peut dès lors utiliser le théorème central classique, on se doit au contraire d'utiliser le théorème central limite à la Lindeberg-Feller.

Néanmoins, le choix d'une telle modélisation des seuils n'est pas du tout justifié :

- on ne génotype qu'asymptotiquement un pourcentage γ de la population (par la loi des grands nombres), ce qui est inchangé par rapport à notre modélisation.
- si l'on souhaite étudier les stratégies une et trois, on est dans l'incapacité de calculer les tests de Wald associés. En effet, le support de la vraisemblance change avec l'effet du QTL q, l'information de Fisher est par conséquent inexistante.

A noter que dans leur article, Darvasi et Soller se placent dans le cas particulier où $\mathbb{P}(Y > S_+) = \mathbb{P}(Y < S_-) = \gamma/2$. Si l'on se replace dans ce contexte particulier, nous obtenons malgré tout le même résultat final que celui présenté en formule (27) de l'article, à savoir que l'on doit multiplier le nombre d'individus par $\zeta_{eff} = \gamma + 2\, z_{\gamma/2}\, \varphi(z_{\gamma/2})$ pour que le test sur les extrêmes et le test oracle aient même puissance.

3.4 Annexe

3.4.1 Test de Wald (μ, q, σ) (situation oracle)

On considére tout d'abord le test de Wald dans un modèle statistique à deux paramètres (μ, q). On pose $\theta = (\mu, q)$ et $\theta_0 = (\mu, 0)$.
La vraisemblance L pour une observation (X, Y) s'écrit :

$$L = \frac{p}{\sigma}\,\varphi\left(\frac{y-\mu-q}{\sigma}\right)1_{X=1} \;+\; \frac{1-p}{\sigma}\,\varphi\left(\frac{y-\mu+q}{\sigma}\right)1_{X=-1}$$

$$\frac{\partial logL}{\partial \mu} = \left(\frac{y-\mu-q}{\sigma^2}\right)1_{X=1} \;+\; \left(\frac{y-\mu+q}{\sigma^2}\right)1_{X=-1}$$

$$\frac{\partial^2 logL}{\partial \mu^2} = -\frac{1}{\sigma^2} \quad \text{d'où} \quad I_{11}(\theta_0) = \frac{1}{\sigma^2}$$

$$\frac{\partial logL}{\partial q} = \left(\frac{y-\mu-q}{\sigma^2}\right)1_{X=1} \;-\; \left(\frac{y-\mu+q}{\sigma^2}\right)1_{X=-1}$$

$$\frac{\partial^2 logL}{\partial q^2} = -\frac{1}{\sigma^2} \quad \text{d'où} \quad I_{22}(\theta_0) = \frac{1}{\sigma^2}$$

$$\frac{\partial logL}{\partial \mu\,\partial q} = -\frac{1}{\sigma^2}1_{X=1} \;+\; \frac{1}{\sigma^2}1_{X=-1}$$

$$I_{12}(\theta_0) = \frac{1}{\sigma^2}\,(2p-1)$$

On a donc :

$$I_{\theta_0} = \begin{pmatrix} \frac{1}{\sigma^2} & \frac{1}{\sigma^2}\,(2p-1) \\ \frac{1}{\sigma^2}\,(2p-1) & \frac{1}{\sigma^2} \end{pmatrix}$$

L'EMV $\hat{\theta}$ est tel que $\hat{\theta} = (\hat{\mu}, \hat{q})$ où $\hat{\mu}$ et $\hat{q}$ sont les EMV respectifs de μ et q.
On souhaite tester :

$$H_0 :\; q = 0 \;\; vs \;\; H_1 :\; q \neq 0$$

Le modèle étudié étant régulier, on applique le théorème 2.9. Ainsi, le test de Wald est le suivant :

$$\sqrt{n}\,\hat{q} \overset{H_0}{\rightarrow} N(\,0\,,\, I_{22}^{-1}(\theta_0)\,)$$

On a :

$$I_{\theta_0}^{-1} = \frac{\sigma^2}{4(1-p)p}\begin{pmatrix} 1 & 1-2p \\ 1-2p & 1 \end{pmatrix}$$

D'où la statistique de test W :

$$W := \sqrt{n}\,\frac{2\,\sqrt{(1-p)p}}{\sigma}\,\hat{q} \overset{H_0}{\rightarrow} N(0,\,1)$$

Le théorème 2.9, basé principalement sur les résultats du premier et troisième lemme de Le Cam, nous permet de calculer trés facilement le phénomène de translation qui s'observe lorsque l'on passe de la loi asymptotique sous l'hypothèse nulle à celle sous une alternative contiguë. Ici, le h_n du théorème 2.9 est tel que : $h_n = h = (0, a)$. On en déduit,

$$W \overset{H_a}{\rightarrow} N\left(\frac{2a\,\sqrt{p(1-p)}}{\sigma},\,1\right)$$

Par définition, le test de Wald (μ, q, σ) est supérieur ou égal en terme de puissance au test oracle (μ, q, σ) présenté en section 3.2.2. Or on constate que le test de Wald (μ, q) possède les mêmes lois asymptotiques que le test oracle (μ, q, σ). On en déduit donc que le test de Wald (μ, q, σ) et le test oracle (μ, q, σ) possèdent les mêmes lois asymptotiques.

3.4.2 Test statistique oracle (q)

On suppose ici que μ et σ sont connus. On a donc désormais un modèle statistique à un paramètre (q). Au vu de la section 3.2.2, on est tenté d'utiliser l'analogue de la statistique de test T (cf. formule 3.2 page 44) mais à μ connu. On a dès lors :

$$T = \frac{\sum_{j=1}^{n} \frac{1}{p}(Y_j - \mu)\,1_{X_j=1} \;-\; \frac{1}{1-p}(Y_j - \mu)\,1_{X_j=-1}}{\sigma\,\sqrt{\frac{n}{p(1-p)}}}$$

$$T \overset{H_0}{\rightarrow} N(0,\,1)$$

$$T \overset{H_a}{\rightarrow} N\left(\frac{2a\,\sqrt{p(1-p)}}{\sigma},\,1\right)$$

Cependant, ce test n'est pas le meilleur test à effectuer. Lorsque l'on utilise un test de Wald, les calculs de l'annexe 3.4.1 permettent d'obtenir immédiatement l'EMV $\hat{q}$ de q :

$$\hat{q} = \frac{1}{n} \sum_{j=1}^{n} (Y_j - \mu) 1_{X_j=1} \; - \; (Y_j - \mu) 1_{X_j=-1}$$

En utilisant $I_{22}(\theta_0)$ de l'annexe 3.4.1, la statistique de Wald W s'écrit :

$$W = \frac{\sum_{j=1}^{n} (Y_j - \mu) 1_{X_j=1} \; - \; (Y_j - \mu) 1_{X_j=-1}}{\sigma \sqrt{n}}$$

Soit $\hat{q}_1$ l'EMV de q si l'on dispose uniquement d'une observation.

$$\hat{q}_1 = (Y - \mu) 1_{X=1} \; - \; (Y - \mu) 1_{X=-1}$$

$$\mathbb{E}_{H_0} (\hat{q}_1) = 0 \quad \text{et} \quad \mathbb{E}_{H_0} \left(\hat{q}_1^2 \right) = \sigma^2$$

D'où par le thèorème central limite :

$$W \overset{H_0}{\rightarrow} N(0,\, 1)$$

De plus :

$$\mathbb{E}_{H_a} (\hat{q}_1) = q \quad \text{et} \quad \mathrm{Var}_{H_a} (\hat{q}_1) \rightarrow \sigma^2$$

D'où par le thèorème central limite :

$$W \quad \overset{H_a}{\rightarrow} N \left(\tfrac{a}{\sigma},\, 1 \right)$$

On remarque que :

$$\{W = T\} \; \Leftrightarrow \; \left\{ p = \frac{1}{2} \right\}$$

Et le test à partir de W est toujours au moins plus puissant que celui à partir de T car on a toujours $p(1-p) \leqslant \frac{1}{4}$. Ceci n'est pas étonnant car l'estimateur du maximum de vraisemblance est optimal sous certaines conditions vérifiées dans le cadre de notre modèle (Van der Vaart, 1998).

3.4.3 Test statistique oracle (μ, q)

On considère un modèle statistique à deux paramètres (μ, q). Le test statistique basé sur la statistique de test T (cf. formule 3.2 page 44) représente le test oracle (μ, q). Ce test présente les mêmes lois asymptotiques que le test de Wald (μ, q) (cf. annexe 3.4.1).

3.4.4 Algorithme EM pour la première stratégie

On considère un modèle statistique à deux paramètres (μ, q). On définit μ_{-1} et μ_1 de telle sorte que :

$$\mu_{-1} = \mu - q \quad , \quad \mu_1 = \mu + q$$

Afin de calculer $\hat{\mu}_{-1}$ et $\hat{\mu}_1$ EMV respectivement de μ_{-1} et μ_1, on utilisera l'algorithme EM : l'expression de la vraisemblance ci dessus n'étant pas facilement maximisable. Naturellement, $\hat{\mu} = (\hat{\mu}_{-1} + \hat{\mu}_1)/2$ et $\hat{q} = (\hat{\mu}_1 - \hat{\mu}_{-1})/2$.

On observe $\overline{X}$ et non pas X : on définit par conséquent l'ensemble des données complètes par $R = (R_1, ..., R_n)$ où $R_j = (X_j, Y_j)$. A défaut d'utiliser le vecteur de paramètres $\theta = (\mu, q)$, on introduit le vecteur de paramètres $\Psi = (\mu_{-1}, \mu_1)$ par soucis de clarté pour la présentation de l'algorithme EM.

La vraisemblance $L(R; \Psi)$ des données complètes s'écrit :

$$L(R;\Psi) = \prod_{j=1}^{n} \left\{ \frac{1-p}{\sigma}\, \varphi\left(\frac{y_j - \mu_{-1}}{\sigma}\right)\, 1_{X_j=-1} \right\} \left\{ \frac{p}{\sigma}\, \varphi\left(\frac{y_j - \mu_1}{\sigma}\right)\, 1_{X_j=1} \right\}$$

$$\log L(R;\Psi) = \sum_{j=1}^{n} \log\left\{ \frac{1-p}{\sigma}\varphi\left(\frac{y_j - \mu_{-1}}{\sigma}\right) \right\} 1_{X_j=-1} \; + \; \log\left\{ \frac{p}{\sigma}\varphi\left(\frac{y_j - \mu_1}{\sigma}\right) \right\} 1_{X_j=1}$$

Sachant la valeur courante Ψ^N à l'itération N, la phase dite **expectation (E)** consiste en la détermination de la fonction $Q(\Psi, \Psi^N)$, définie de la manière suivante :

$$Q(\Psi, \Psi^N) = \mathbb{E}_{X/Y=y, \overline{X}, \Psi=\Psi^N} \left\{ \log L(R, \Psi) \right\}$$

On a :

$$\mathbb{P}(X = 1/Y = y, \overline{X}, \Psi = \Psi^N) = 1_{\overline{X}=1} \; + \; \frac{p\, \varphi(\frac{y-\mu_1^N}{\sigma})}{p\, \varphi(\frac{y-\mu_1^N}{\sigma}) + (1-p)\, \varphi(\frac{y-\mu_{-1}^N}{\sigma})} 1_{\overline{X}=0}$$

$$\mathbb{P}(X = -1/Y = y, \overline{X}, \Psi = \Psi^N) = 1_{\overline{X}=-1} \; + \; \frac{(1-p)\, \varphi(\frac{y-\mu_{-1}^N}{\sigma})}{p\, \varphi(\frac{y-\mu_1^N}{\sigma}) + (1-p)\, \varphi(\frac{y-\mu_{-1}^N}{\sigma})} 1_{\overline{X}=0}$$

D'où,

$$Q(\Psi,\Psi^N) = \sum_{j=1}^{n} \log\left\{\frac{1-p}{\sigma}\varphi\left(\frac{y_j-\mu_{-1}}{\sigma}\right)\right\} 1_{\overline{X}_j=-1} \; + \; \log\left\{\frac{p}{\sigma}\varphi\left(\frac{y_j-\mu_1}{\sigma}\right)\right\} 1_{\overline{X}_j=1}$$
$$+ \frac{p\,\varphi(\frac{y_j-\mu_1^N}{\sigma})}{p\,\varphi(\frac{y_j-\mu_1^N}{\sigma}) + (1-p)\,\varphi(\frac{y_j-\mu_{-1}^N}{\sigma})} \; \log\left\{\frac{p}{\sigma}\varphi\left(\frac{y_j-\mu_1}{\sigma}\right)\right\} 1_{\overline{X}_j=0}$$
$$+ \frac{(1-p)\,\varphi(\frac{y_j-\mu_{-1}^N}{\sigma})}{p\,\varphi(\frac{y_j-\mu_1^N}{\sigma}) + (1-p)\,\varphi(\frac{y_j-\mu_{-1}^N}{\sigma})} \; \log\left\{\frac{1-p}{\sigma}\varphi\left(\frac{y_j-\mu_{-1}}{\sigma}\right)\right\} 1_{\overline{X}_j=0}$$

Lors de la phase dite **maximisation (M)**, on actualise la valeur courante du paramètre en maximisant la fonction obtenue à la phase E par rapport à Ψ, soit :

$$\Psi^{N+1} = \text{argmax}_{\Psi}\, Q(\Psi,\Psi^N)$$

$$\frac{\partial Q}{\partial \mu_{-1}} = \sum_{j=1}^{n} \frac{y_j-\mu_{-1}}{\sigma^2}\left\{1_{\overline{X}_j=-1} \; + \; \frac{(1-p)\,\varphi(\frac{y_j-\mu_{-1}^N}{\sigma})}{p\,\varphi(\frac{y_j-\mu_1^N}{\sigma}) + (1-p)\,\varphi(\frac{y_j-\mu_{-1}^N}{\sigma})} 1_{\overline{X}_j=0}\right\}$$

$$\frac{\partial Q}{\partial \mu_1} = \sum_{j=1}^{n} \frac{y_j-\mu_1}{\sigma^2}\left\{1_{\overline{X}_j=1} \; + \; \frac{p\,\varphi(\frac{y_j-\mu_1^N}{\sigma})}{p\,\varphi(\frac{y_j-\mu_1^N}{\sigma}) + (1-p)\,\varphi(\frac{y_j-\mu_{-1}^N}{\sigma})} 1_{\overline{X}_j=0}\right\}$$

En annulant les dérivées ci-dessus, on obtient les estimateurs correspondant à la $N+1$ ème itération :

$$\mu_{-1}^{N+1} = \frac{\sum_{j=1}^{n} y_j\left\{1_{\overline{X}_j=-1} + \frac{(1-p)\,\varphi(\frac{y_j-\mu_{-1}^N}{\sigma})}{p\,\varphi(\frac{y_j-\mu_1^N}{\sigma})+(1-p)\,\varphi(\frac{y_j-\mu_{-1}^N}{\sigma})} 1_{\overline{X}_j=0}\right\}}{\sum_{j=1}^{n} 1_{\overline{X}_j=-1} + \frac{(1-p)\,\varphi(\frac{y_j-\mu_{-1}^N}{\sigma})}{p\,\varphi(\frac{y_j-\mu_1^N}{\sigma})+(1-p)\,\varphi(\frac{y_j-\mu_{-1}^N}{\sigma})} 1_{\overline{X}_j=0}}$$

$$\mu_1^{N+1} = \frac{\sum_{j=1}^{n} y_j\left\{1_{\overline{X}_j=1} + \frac{p\,\varphi(\frac{y_j-\mu_1^N}{\sigma})}{p\,\varphi(\frac{y_j-\mu_1^N}{\sigma})+(1-p)\,\varphi(\frac{y_j-\mu_{-1}^N}{\sigma})} 1_{\overline{X}_j=0}\right\}}{\sum_{j=1}^{n} 1_{\overline{X}_j=1} + \frac{p\,\varphi(\frac{y_j-\mu_1^N}{\sigma})}{p\,\varphi(\frac{y_j-\mu_1^N}{\sigma})+(1-p)\,\varphi(\frac{y_j-\mu_{-1}^N}{\sigma})} 1_{\overline{X}_j=0}}$$

Remarque 3.11 *Cet algorithme est présenté ici dans le cadre d'un modèle statistique à deux paramètres. On peut facilement le généraliser pour un modèle statistique à un paramètre ou encore trois paramètres lorsque la variance est inconnue.*

3.4.5 Preuve du corollaire 3.9

Première stratégie

D'après les calculs de l'information de Fisher, présents dans la preuve du théorème 3.3, le test de Wald pour un modèle (q) et qui correspond à la première stratégie est le suivant :

$$W_1 := \frac{\sqrt{n}}{\sigma^2}\sqrt{\mathcal{A}+(2p-1)^2(\sigma^2-\mathcal{A})}\ \hat{q} \overset{H_0}{\to} N(0,\ 1)$$

Afin d'étudier la puissance de ce test, on applique le théorème 2.9 avec $h_n = h = a$. Ainsi,

$$W_1 \overset{H_a}{\to} N\left(\frac{a}{\sigma^2}\sqrt{\mathcal{A}+(2p-1)^2(\sigma^2-\mathcal{A})},\ 1\right)$$

En utilisant comme test de référence, le test statistique oracle (q) (cf. annexe 3.4.2), on obtient facilement l'efficacité κ_1 correspondant à la première stratégie :

$$\kappa_1 = \gamma\ +\ z_{\gamma_+}\varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})\ +\ (2p-1)^2\ \left\{1-\gamma-z_{\gamma_+}\varphi(z_{\gamma_+})+z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})\right\}$$

Deuxième stratégie

D'après la preuve du théorème 3.3 (cf. formule 3.4 page 52), la statistique T_2 pour un modèle (q) est la suivante :

$$T_2 = \frac{\sum_{j=1}^{n}\frac{1}{p}(Y_j-\mu)1_{\overline{X}_j=1}\ -\ \frac{1}{1-p}(Y_j-\mu)1_{\overline{X}_j=-1}}{\sqrt{\frac{n\ \mathcal{A}}{p(1-p)}}} \overset{H_0}{\to} N(0,\ 1)$$

On avait trouvé :

$$T_2 \overset{H_a}{\to} N(0,\ 1)\quad ,\quad T_2 \overset{H_a}{\to} N\left(\frac{2a}{\sigma^2}\ \sqrt{\mathcal{A}\,p(1-p)},\ 1\right)$$

En utilisant comme test de référence, le test statistique oracle (q) (cf. annexe 3.4.2), on obtient facilement l'efficacité κ_2 correspondant à la deuxième stratégie :

$$\kappa_2 = 4\,p\,(1-p)\ \left\{\gamma\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})\ +\ z_{\gamma_+}\ \varphi(z_{\gamma_+})\right\}$$

Troisième stratégie

On présente ici le test de Wald dans un modèle statistique à un paramètre (q). On note $\theta = (q)$ et $\theta_0 = (0)$. On suppose $\gamma \neq 1$. La vraisemblance L à partir d'une observation s'écrit :

$$L = \frac{1-p}{\sigma}\,\varphi\left(\frac{y-\mu+q}{\sigma}\right)\ 1_{\overline{X}=-1}\ +\ \frac{p}{\sigma}\,\varphi\left(\frac{y-\mu-q}{\sigma}\right)\ 1_{\overline{X}=1}\ +\ \mathbb{P}(\overline{X}=0)\ 1_{\overline{X}=0}$$

$$\frac{\partial logL}{\partial q}\,|_{\theta_0} = -\frac{y-\mu}{\sigma^2}\,1_{\overline{X}=-1} \;+\; \frac{y-\mu}{\sigma^2}\,1_{\overline{X}=1} + \frac{(1-2p)\;\varphi\left(z_{\gamma_+}\right) \;+(2p-1)\;\varphi\left(z_{1-\gamma_-}\right)}{\sigma(1\;-\gamma)}\,1_{\overline{X}=0}$$

$$\left(\frac{\partial logL}{\partial q}\,|_{\theta_0}\right)^2 = \frac{(y-\mu)^2}{\sigma^4}\,1_{\overline{X}=-1} \;+\; \frac{(y-\mu)^2}{\sigma^4}\,1_{\overline{X}=1} + \left[\frac{(1-2p)\;\varphi\left(z_{\gamma_+}\right)+(2p-1)\;\varphi\left(z_{1-\gamma_-}\right)}{\sigma(1\;-\gamma)}\right]^2 1_{\overline{X}=0}$$

D'où

$$I_{\theta_0} = \frac{\mathcal{A}}{\sigma^4} \;+\; \frac{\left\{(1-2p)\;\varphi\left(z_{\gamma_+}\right) \;+\; (2p-1)\;\varphi\left(z_{1-\gamma_-}\right)\right\}^2}{\sigma^2(1\;-\gamma)}$$

On notera $\hat{q}$ l'EMV de q. Il pourra être calculé par la méthode de Newton. Le test de Wald est le suivant :

$$\sqrt{n}\,\hat{q} \;\overset{H_0}{\rightarrow}\; N(\,0\,,\,I_{\theta_0}^{-1}\,)$$

D'où la statistique de test W_3 correspondant à la troisième stratégie :

$$W_3 := \sqrt{n}\;\left[\frac{\mathcal{A}}{\sigma^4} \;+\; \frac{\left\{(1-2p)\;\varphi\left(z_{\gamma_+}\right) \;+\; (2p-1)\;\varphi\left(z_{1-\gamma_-}\right)\right\}^2}{\sigma^2(1\;-\gamma)}\right]^{1/2} \hat{q} \;\overset{H_0}{\rightarrow}\; N(0,\,1)$$

Alors, en appliquant le théorème 2.9 avec $h_n = h = a$, on a :

$$W_3 \;\overset{H_a}{\rightarrow}\; N(a\sqrt{I_{\theta_0}},\,1)$$

En utilisant comme test de référence, le test statistique oracle (q) (cf. annexe 3.4.2), on obtient facilement l'efficacité κ_3 correspondant à la troisième stratégie :

$$\kappa_3 = \gamma \;+\; z_{\gamma_+}\varphi(z_{\gamma_+}) \;-\; z_{1-\gamma_-}\,\varphi(z_{1-\gamma_-}) \;+\; \frac{(2p-1)^2}{1\;-\gamma}\left\{\varphi(z_{1-\gamma_-}) \;-\; \varphi(z_{\gamma_+})\right\}^2 \quad \forall\gamma \neq 1$$

Optimisation du génotypage

On cherche à répondre à la question suivante : comment choisir γ_+ et γ_- afin d'obtenir des efficacités maximales ? On rappelle que l'on a la relation $\gamma_+ + \gamma_- = \gamma$.

On étudie ici les différentes stratégies. Afin de faciliter les calculs, on traite tout d'abord la deuxième stratégie.

On cherche tout d'abord à maximiser l'efficacité κ_2. On rappelle que :

$$\kappa_2 = 4\,p\,(1-p)\;\left\{\gamma \;-\; z_{1-\gamma_-}\,\varphi(z_{1-\gamma_-}) \;+\; z_{\gamma_+}\,\varphi(z_{\gamma_+})\right\}$$

Soit $g(.)$ la fonction telle que : $g(z_{\gamma_+}) = \Phi^{-1}\left\{\gamma - 1 + \Phi(z_{\gamma_+})\right\}$. Ainsi, $z_{1-\gamma_-} = g(z_{\gamma_+})$.
Afin de maximiser κ_2, il faut maximiser la fonction $k_2(.)$ définie de la manière suivante :

$$k_2(z_{\gamma_+}) = z_{\gamma_+}\varphi(z_{\gamma_+}) - g(z_{\gamma_+})\,\varphi\left\{g(z_{\gamma_+})\right\}$$

On a :

$$k_2'(z_{\gamma_+}) = \varphi(z_{\gamma_+}) + z_{\gamma_+}\varphi'(z_{\gamma_+}) - g'(z_{\gamma_+})\,\varphi\left\{g(z_{\gamma_+})\right\} - g(z_{\gamma_+})\,g'(z_{\gamma_+})\,\varphi'\left\{g(z_{\gamma_+})\right\}$$

$$g'(z_{\gamma_+}) = \frac{\varphi(z_{\gamma_+})}{\varphi(z_{1-\gamma_-})}$$

Alors :

$$k_2'(z_{\gamma/2}) = \varphi(z_{\gamma/2}) - \left\{z_{\gamma/2}\right\}^2\,\varphi(z_{\gamma/2}) - \varphi(z_{1-\gamma/2}) + \left\{z_{1-\gamma/2}\right\}^2\,\varphi(z_{1-\gamma/2}) = 0$$

On conclut que l'efficacité κ_2 de la stratégie deux est maximum lorsque $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.
On cherche désormais à maximiser l'efficacité κ_1. On rappelle que :

$$\kappa_1 = \gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\,\varphi(z_{1-\gamma_-}) + (2p-1)^2\left\{1 - \gamma - z_{\gamma_+}\varphi(z_{\gamma_+}) + z_{1-\gamma_-}\,\varphi(z_{1-\gamma_-})\right\}$$

Il faut maximiser la fonction $k_1(.)$ définie de la manière suivante :

$$k_1(z_{\gamma_+}) = k_2(z_{\gamma_+}) - k_2(z_{\gamma_+})\,(2p-1)^2$$

Dés lors :

$$k_1'(z_{\gamma_+}) = k_2'(z_{\gamma_+}) - k_2'(z_{\gamma_+})\,(2p-1)^2$$

Comme $k_2'(z_{\gamma/2}) = 0$, on a $k_1'(z_{\gamma/2}) = 0$.
On conclut que l'efficacité κ_1 de la première stratégie est maximum lorsque $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.

On cherche désormais à maximiser l'efficacité κ_3. On rappelle que :

$$\kappa_3 = \gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\,\varphi(z_{1-\gamma_-}) + \frac{(2p-1)^2}{1-\gamma}\left\{\varphi(z_{1-\gamma_-}) - \varphi(z_{\gamma_+})\right\}^2 \quad \forall\gamma \neq 1$$

Il faut maximiser la fonction $k_3(.)$ définie de la manière suivante :

$$k_3(z_{\gamma_+}) = k_2(z_{\gamma_+}) + \frac{(2p-1)^2}{1-\gamma}\left[\,\varphi\left\{g(z_{\gamma_+})\right\} - \varphi(z_{\gamma_+})\,\right]$$

On a :

$$k_3'(z_{\gamma_+}) = k_2'(z_{\gamma_+}) + \frac{(2p-1)^2}{1-\gamma}\,2\left[\,g'(z_{\gamma_+})\varphi'\left\{g(z_{\gamma_+})\right\} - \varphi'(z_{\gamma_+})\,\right]\left[\,\varphi\left\{g(z_{\gamma_+})\right\} - \varphi(z_{\gamma_+})\,\right]$$

Alors $k_3'(z_{\gamma/2}) = 0$.
On conclut que l'efficacité κ_3 de la troisième stratégie est maximum lorsque $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.■

3.4.6 Convergence vers l'asymptotique

On considère ici un modèle (q) et la deuxième stratégie.

Comparaison puissance théorique/puissance par Monte Carlo

En utilisant un test unilatéral de niveau α, la puissance théorique du test, β_2, vérifie (cf. annexe 3.4.5) :

$$\beta_2 = \mathbb{P}\left(Z > z_\alpha \; - \; \frac{2a}{\sigma^2}\sqrt{\mathcal{A}\, p(1-p)} \right) \quad \text{où } Z \sim N(0,1)$$

Afin de valider ce résultat asymptotique, on se propose de comparer la puissance théorique établie ci-dessus, β_2, avec la puissance calculée par une méthode de Monte-Carlo, notée β_{MC}. On présentera également un intervalle de confiance à 95% pour la vraie valeur de la puissance :

$$IC \;=\; \left[\beta_{MC} - 1.96\,\sqrt{\frac{\beta_{MC}(1-\beta_{MC})}{nb_{ech}}} \;\; ; \;\; \beta_{MC} + 1.96\,\sqrt{\frac{\beta_{MC}(1-\beta_{MC})}{nb_{ech}}} \right]$$

où nb_{ech} désigne le nombre d'échantillons

Les paramètres suivants ont été considérés : $p = \frac{1}{2}$, $a = 2$, $\mu = 0$, $\sigma = 1$.
On rappelle que $q = \frac{a}{\sqrt{n}}$. Les simulations ont été réalisées sous la condition $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.
La table 3.3 page 73 présente pour $n = 10000$ la puissance théorique et la puissance par Monte-Carlo, en fonction du pourcentage d'individus génotypés γ. Afin de calculer la puissance par Monte-Carlo, 10000 échantillons de taille n ont été considérés. On observe sur cette table que quel que soit γ, la puissance théorique calculée appartient à l'intervalle de confiance à 95% de la vraie valeur de la puissance.

On établit les mêmes conclusions lorque l'on considère $n = 100$ (cf table 3.4), $n = 50$ (cf table 3.5), $n = 30$ (cf table 3.6), qui correspondent respectivement à $q = 0.22$, $q = 0.2828$, $q = 0.3651$.

En conclusion, les résultats théoriques établis sont très bons pour des effets du QTL très petits. Lorsque l'effet grossit, ils se comportent également très bien et ce quelle que soit la valeur du paramètre γ.

γ	β_{MC}	β_2	IC en %
0.1	37.4%	37.45%	[36.45 ; 38.35]
0.2	48.69%	48.61%	[47.70 ; 49.67]
0.3	54.62%	54.77%	[53.64 ; 55.60]
0.4	58.69%	58.58%	[57.72 ; 59.66]
0.5	61.54%	60.93%	[60.59 ; 62.49]
0.6	62.77%	62.33%	[61.82 ; 63.72]
0.7	63.38%	63.13%	[62.44 ; 64.32]
0.8	63.85%	63.52%	[62.91 ; 64.79]
0.9	64.06%	63.68%	[63.12 ; 65.00]
1	64.11%	63.68%	[63.17 ; 65.05]

TAB. 3.3 – Puissance théorique (β_2) et puissance par Monte-Carlo (β_{MC}) en fonction du pourcentage de génotypés γ ($nb_{ech} = 10000$, $n = 10000$, $q = \frac{2}{\sqrt{10000}} = 0.02$)

γ	β_{MC}	β_2	IC en %
0.1	37.81%	37.45%	[36.86 ; 38.76]
0.2	48.14%	48.61%	[47.16 ; 49.12]
0.3	54.90%	54.77%	[53.92 ; 55.88]
0.4	58.02%	58.58%	[57.05 ; 58.99]
0.5	59.81%	60.93%	[58.85 ; 60.77]
0.6	62.84%	62.33%	[61.89 ; 63.79]
0.7	62.73%	63.13%	[61.78 ; 63.68]
0.8	63.36%	63.52%	[62.42 ; 64.30]
0.9	63.69%	63.68%	[62.75 ; 64.63]
1	63.36%	63.68%	[62.42 ; 64.30]

TAB. 3.4 – Puissance théorique (β_2) et puissance par Monte-Carlo (β_{MC}) en fonction du pourcentage de génotypés γ ($nb_{ech} = 10000$, $n = 100$, $q = \frac{2}{\sqrt{100}} = 0.2$)

γ	β_{MC}	β_2	IC en %
0.1	38.38%	37.45%	[37.43 ; 39.33]
0.2	48.22%	48.61%	[47.24 ; 49.20]
0.3	55.07%	54.77%	[54.10 ; 56.04]
0.4	57.71%	58.58%	[56.74 ; 58.68]
0.5	60.68%	60.93%	[59.72 ; 61.64]
0.6	63.30%	62.33%	[62.36 ; 64.24]
0.7	63.28%	63.13%	[62.34 ; 64.22]
0.8	64.18%	63.52%	[63.24 ; 65.12]
0.9	63.20%	63.68%	[62.25 ; 64.15]
1	63.64%	63.68%	[62.70 ; 64.58]

TAB. 3.5 – Puissance théorique (β_2) et puissance par Monte-Carlo (β_{MC}) en fonction du pourcentage de génotypés γ ($nb_{ech} = 10000$, $n = 50$, $q = \frac{2}{\sqrt{50}} = 0.2828$)

γ	β_{MC}	β_2	IC en %
0.1	38.27%	37.45%	[37.32 ; 39.22]
0.2	48.80%	48.61%	[47.82 ; 49.78]
0.3	54.64%	54.77%	[53.66 ; 55.62]
0.4	58.60%	58.58%	[57.63 ; 59.57]
0.5	61.48%	60.93%	[60.53 ; 62.43]
0.6	61.73%	62.33%	[60.78 ; 62.68]
0.7	63.21%	63.13%	[62.26 ; 64.16]
0.8	63.27%	63.52%	[62.33 ; 64.21]
0.9	63.79%	63.68%	[62.85 ; 64.73]
1	63.56%	63.68%	[62.62 ; 64.50]

TAB. 3.6 – Puissance théorique (β_2) et puissance par Monte-Carlo (β_{MC}) en fonction du pourcentage de génotypés γ ($nb_{ech} = 10000$, $n = 30$, $q = \frac{2}{\sqrt{30}} = 0.3651$)

Comparaison puissance théorique / puissance par Monte-Carlo en fonction du ratio ζ

Afin de valider les résultats asymptotiques au sujet de l'efficacité, on se propose de comparer la puissance théorique et la puissance par une méthode de Monte-Carlo, tout en faisant varier le ratio ζ.
Les valeurs des paramètres étudiés sont les suivantes : $\gamma = 32\%$, $a = 2$, $\sigma = 1$. On suppose également que $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.
Pour cette configuration, $\zeta_{eff} = 1.2439$, ce qui correspond à une efficacité théorique $\kappa_2 = 80.39\%$. La table 3.7 présente pour $n = 500$ observations, les puissances théoriques (β_2) et empiriques (β_{MC}) en fonction du ratio ζ. On remarque que β_2 se situe quel que soit ζ dans l'intervalle de confiance à 95% de la vraie valeur de la puissance.

On étudie désormais un effet QTL $q = 0.1$. Si on considère toujours $n = 500$, comme $q = \frac{a}{\sqrt{n}}$, alors $a = 2.24$. Cependant, ζ_{eff} est inchangé car il ne dépend ni de a et ni de n. On étudie également le cas $a = 2.5$ et $n = 500$.

Les tables 3.8 et 3.9 traduisent le bon comportement de β_2 quel que soit ζ.

ζ	β_{MC}	β_2	IC en %
1	55.51%	55.69%	[55.07 ; 55.95]
1.24	63.59%	63.56%	[63.17 ; 64.01]
1.3	65.33%	65.34%	[64.91 ; 65.75]
1.5	70.96%	70.76%	[70.56 ; 71.36]

TAB. 3.7 – Puissance théorique (β_2) et puissance par Monte-Carlo (β_{MC}) en fonction du ratio ζ ($nb_{ech} = 50000$, $n = 500$, $n^\star = \zeta n$, $q = \frac{2}{\sqrt{500}} = 0.089$)

ζ	β_{MC}	β_2	IC en %
1	63.98%	64.00%	[63.56 ; 64.40]
1.24	72.29%	72.12%	[71.90 ; 72.68]
1.3	73.76%	73.89%	[73.37 ; 74.15]
1.5	79.38%	79.10%	[79.03 ; 79.73]

TAB. 3.8 – Puissance théorique (β_2) et puissance par Monte-Carlo (β_{MC}) en fonction du ratio ζ ($nb_{ech} = 50000$, $n = 500$, $n^\star = \zeta n$, $q = \frac{2.24}{\sqrt{500}} = 0.1$)

ζ	β_{MC}	β_2	IC en %
1	72.36%	72.29%	[71.97 ; 72.75]
1.24	80.15%	80.12%	[79.80 ; 80.50]
1.3	81.65%	81.75%	[81.31 ; 81.99]
1.5	86.55%	86.33%	[86.25 ; 86.85]

TAB. 3.9 – Puissance théorique (β_2) et puissance par Monte-Carlo (β_{MC}) en fonction du ratio ζ ($nb_{ech} = 50000$, $n = 500$, $n^\star = \zeta n$, $q = \frac{2.5}{\sqrt{500}} = 0.1118$)

Chapitre 4

Selective genotyping en présence de deux caractères quantitatifs corrélés

4.1 Introduction

Dans ce chapitre, on s'intéresse à deux caractères quantitatifs corrélés Y et Z. La motivation est de savoir s'il existe un QTL affectant le caractère Z.

Les contraintes sont les suivantes :

- le phénotype Z est difficile à mesurer pour des raisons biologiques
- les coûts dus au génotypage sont relativement élevés

Le phénotype Y étant bien plus facile à mesurer, un selective genotyping est effectué sur Y. Pour chaque Y extrême, on génotype l'individu et on mesure son phénotype Z.

A l'instar du chapître 3, on se place sur un marqueur génétique. Deux stratégies différentes pour l'analyse statistique en selective genotyping sont étudiées :

- la première consiste à conserver dans l'analyse statistique, tous les phénotypes, même les phénotypes qui ne sont pas considérés comme extrêmes et pour lesquels nous ne disposons pas du génotype
- la deuxième est basée sur la conservation uniquement des phénotypes extrêmes

Les tests correspondant aux deux stratégies seront des tests de Wald et tous ces tests seront comparés en terme d'efficacité au test oracle, celui où tous les génotypes ainsi que tous les phénotypes Z sont connus.

A travers cette étude théorique, on cherche non seulement à proposer différents tests pour la détection de QTL mais également à quantifier l'apport des phénotypes non extrêmes dans l'analyse statistique.

On s'intéresse par la suite à la question de l'optimisation du génotypage en selective

genotyping.

Les principaux résultats obtenus lors de cette étude du selective genotyping sont énoncés en théorème 4.3 page 80.

4.2 Test statistique en l'absence de censure

4.2.1 Modèle en l'absence de censure

Comme dans le chapitre précédent, X est la v.a. correspondant au génotype au QTL. On observe désormais non plus un seul phénotype Y mais deux phénotypes, Y et Z. Le modèle pour ces deux v.a. s'écrit :

$$\begin{pmatrix} \mathrm{Y} \\ \mathrm{Z} \end{pmatrix} = \begin{pmatrix} \mu_Y + q_Y X \\ \mu_Z + q_Z X \end{pmatrix} + \varepsilon$$

où

$$\varepsilon \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & r\,\sigma^2 \\ r\,\sigma^2 & \sigma^2 \end{pmatrix} \right)$$

On supposera $r \notin \{-1, 1\}$. De plus, r et σ^2 seront supposés connus.
On notera μ_{YX} et μ_{ZX} les quantités telles que : $\mu_{YX} = \mu_Y + q_Y X$ et $\mu_{ZX} = \mu_Z + q_Z X$.
Enfin, On considèrera un échantillon de n observations (X_j, Y_j, Z_j) iid.

Remarque 4.1 *q_Z (resp. q_Y) désigne l'effet du QTL sur le phénotype Z (resp. Y).*

4.2.2 Test statistique oracle (μ_Z, q_Z)

Afin de tester la présence d'un QTL affectant le phénotype Z, on confronte les 2 hypothèses suivantes :

$$H_{0Z} : \; q_Z = 0 \;\; vs \;\; H_{1Z} : \; q_Z \neq 0$$

Plus précisément, on utilisera une alternative locale $H_{bZ} : q_Z = \frac{b}{\sqrt{n}}$ où b est une constante.
D'après l'annexe 3.4.3 du chapitre 3, la statistique de test est la suivante :

$$T := \frac{\sum_{j=1}^{n} \frac{1}{p}(Z_j - \overline{Z}) 1_{X_j=1} \; - \; \frac{1}{1-p}(Z_j - \overline{Z}) 1_{X_j=-1}}{\sigma \sqrt{\frac{n}{p(1-p)}}}$$

Les lois asymptotiques sont les suivantes :

$$T \overset{H_{0Z}}{\rightarrow} N(0,\, 1) \qquad T \overset{H_{bZ}}{\rightarrow} N\left(\frac{2\, b\, \sqrt{p(1-p)}}{\sigma},\, 1 \right)$$

où $\overline{Z} = \frac{1}{n} \sum_{j=1}^{n} Z_j$

4.3 Etude des différentes stratégies en selective genotyping

4.3.1 Modèle

On dispose de l'ensemble des phénotypes Y. Cependant, pour un individu donné présentant un phénotype Y, on ne dispose du deuxième phénotype Z et du génotype au QTL X, uniquement si le phénotype Y est extrême (i.e. $Y \notin [S_- , S_+]$ où S_- et S_+ sont deux seuils quelconques). Ainsi, on observe non pas la v.a. X mais la v.a. $\overline{X}$ (cf. section 3.3.1 du chapitre 3).

4.3.2 Efficacités et puissances des tests correspondant aux différentes stratégies (résultats principaux)

On rappelle tout d'abord brièvement les différentes stratégies énoncées en introduction (cf. section 4.1) :

- la stratégie une est basée sur la conservation de l'ensemble des phénotypes
- la stratégie deux est basée sur la conservation uniquement des phénotypes extrêmes

Pour chacune des stratégies, afin de tester la présence de QTL, on confronte les deux hypothèses suivantes :

$$H_{0Z} : \ q_Z = 0 \ \ vs \ \ H_{1Z} : \ q_Z \neq 0$$

Plus précisément, on utilise comme hypothèse alternative, une alternative locale H_{bZ} : $q_Z = \frac{b}{\sqrt{n}}$ où b est une constante.

Le théorème 4.3 énoncé dans cette section, résume les principaux résultats obtenus en terme d'efficacité quant aux différentes stratégies.

Le test oracle sert de test de référence. Tous les tests considérés sont des tests unilatéraux.

Notation 4.2 *Les hypothèses sur q_Y sont les suivantes :*

- H_{0Y} *:* $q_Y = 0$
- H_{aY} *:* $q_Y = \frac{a}{\sqrt{n}}$ *où a est une constante*

Théorème présentant les différentes efficacités

Théorème 4.3 *Soient $\tilde{\kappa}_1$ et $\tilde{\kappa}_2$ les efficacités correspondant respectivement aux stratégies une et deux énoncées en section 4.1.*
Soient γ, γ_+ et γ_-, les quantités respectives $\mathbb{P}_{H_{0Y}}(Y \notin [S_-, S_+])$, $\mathbb{P}_{H_{0Y}}(Y > S_+)$ et $\mathbb{P}_{H_{0Y}}(Y < S_-)$.
Alors, si l'on considère un modèle statistique à quatre paramètres (μ_Z, q_Z, μ_Y, q_Y), $\forall p$:

i) $\tilde{\kappa}_1 = \tilde{\kappa}_2 = \left\{ \dfrac{1-r^2}{\gamma} + \dfrac{r^2}{\kappa_1} \right\}^{-1}$

ii) *$\tilde{\kappa}_1$ et $\tilde{\kappa}_2$ atteignent leur maximum, $\tilde{M}$, pour $\gamma_+ = \gamma_- = \dfrac{\gamma}{2}$, où*

$$\tilde{M} = \left\{ \frac{1-r^2}{\gamma} + \frac{r^2}{M} \right\}^{-1}$$

où κ_1 et M désignent les quantités présentées en théorème 3.3 page 46.

Ainsi, les deux stratégies présentent la même efficacité : il n'y a donc aucun gain de puissance à considérer les phénotypes non extrêmes dans l'analyse statistique et ce, quel que soit p. On rappelle que dans le modèle étudié, le cas $p = 1/2$ correspond à une population backcross.

De plus, comme dans le cas d'un selective genotyping en présence d'un seul caractère quantitatif, les efficacités correspondant aux différentes stratégies sont maximum lorque l'on génotype asymptotiquement le même pourcentage d'individus à droite qu'à gauche : $\gamma_+ = \gamma_- = \frac{\gamma}{2}$.

Enfin, on remarquera dans la preuve du théorème ci-dessous que les différents tests de détection de QTL affectant Z présentent des puissances indépendantes de l'effet QTL sur Y, pour des raisons de contiguïté.

Preuve du théorème

Préliminaires

On suppose tout d'abord l'absence de censure : on a connaissance de Z et X quelle que soit la valeur de Y.
Afin d'effectuer la régression linéaire de Z/X sur Y/X que l'on nommera $\widetilde{Z}/X$, on définit le produit scalaire suivant, pour deux v.a. U_1 et U_2 à valeurs dans $\mathbb{R}$:

$$< U_1 , U_2 > = \mathbb{E}[U_1 U_2]$$

On a :

$$\begin{aligned}\widetilde{Z}/X &= < Z/X\ ,\ \frac{Y/X-\mu_{YX}}{\sigma} > \frac{Y/X-\mu_{YX}}{\sigma} \quad + \quad < Z/X\ ,\ 1 > 1 \\ &= \ r\,Y/X \ - \ r\,\mu_{YX} \ + \ \mu_{ZX}\end{aligned}$$

On pose :

$$Z^\star := \frac{Z \ - \ r\,Y}{\sigma\sqrt{1-r^2}} \quad \text{et} \quad \mu^\star_{ZX} := \frac{\mu_{ZX} \ - \ r\,\mu_{YX}}{\sigma\sqrt{1-r^2}}$$

Ainsi, $Z^\star/X \sim N(\mu^\star_{ZX}, 1)$. Par construction, $(Z-\widetilde{Z})/X$ et $\widetilde{Z}/X$ sont indépendants. D'où, $Z^\star/X$ et Y/X sont indépendants.

Si l'on se replace désormais dans le cadre du modèle censuré, on disposera de $Z^\star$ uniquement lorsque Y sera extrême. Néanmoins comme $Z^\star/X$ et Y/X sont indépendants, $Z^\star/X$ n'est pas affecté par le fait que Y soit extrême.

Première stratégie (test de Wald utilisant l'ensemble des phénotypes Y)

Notation 4.4 *On notera :*

- $L^\star(\mu^\star_{Z-1}, \mu^\star_{Z1}, \mu_Y, q_Y)$ *la vraisemblance pour une observation* $(\overline{X}, Y, Z^\star)$
- $L(\mu_Z, q_Z, \mu_Y, q_Y)$ *la vraisemblance pour une observation* $(\overline{X}, Y, Z)$

Naturellement, on a la relation $L^\star(\mu^\star_{Z-1}, \mu^\star_{Z1}, \mu_Y, q_Y) = L(\mu_Z, q_Z, \mu_Y, q_Y)$.

On a :

$$L^\star(\mu^\star_{Z-1}, \mu^\star_{Z1}, \mu_Y, q_Y) = \left\{\frac{1-p}{\sigma}\varphi\left(\frac{y-\mu_Y+q_Y}{\sigma}\right) \ + \ \frac{p}{\sigma}\varphi\left(\frac{y-\mu_Y-q_Y}{\sigma}\right)\right\} 1_{\overline{X}=0}$$
$$+ \frac{p}{\sigma}\varphi\left(\frac{y-\mu_Y-q_Y}{\sigma}\right)\varphi(z^\star-\mu^\star_{Z1})\,1_{\overline{X}=1} \ + \ \frac{1-p}{\sigma}\varphi\left(\frac{y-\mu_Y+q_Y}{\sigma}\right)\varphi(z^\star-\mu^\star_{Z-1})\,1_{\overline{X}=-1}$$

Les EMV de μ_Y et q_Y, notés $\hat{\mu}_Y$ et $\hat{q}_Y$, pourront être obtenus par l'algorithme EM présenté en annexe 3.4.4 du chapitre 3.

De plus, comme

$$\frac{\partial \log L^\star}{\partial \mu^\star_{Z1}} = (z^\star-\mu^\star_{Z1})1_{\overline{X}=1} \quad \text{et} \quad \frac{\partial \log L^\star}{\partial \mu^\star_{Z-1}} = (z^\star-\mu^\star_{Z-1})1_{\overline{X}=-1}\ ,$$

on obtient facilement $\hat{\mu}^\star_{Z-1}$ et $\hat{\mu}^\star_{Z1}$ EMV respectifs de $\mu^\star_{Z-1}$ et $\mu^\star_{Z1}$ pour n observations :

$$\hat{\mu}^\star_{Z1} = \frac{1}{\sum_{j=1}^n 1_{\overline{X}_j=1}}\sum_{j=1}^n z^\star_j 1_{\overline{X}_j=1} \quad \text{et} \quad \hat{\mu}^\star_{Z-1} = \frac{1}{\sum_{j=1}^n 1_{\overline{X}_j=-1}}\sum_{j=1}^n z^\star_j 1_{\overline{X}_j=-1}\ .$$

On note $\theta = (\mu_Z, q_Z, \mu_Y, q_Y)$ et $\theta^\star = \left(\mu^\star_{Z-1}, \mu^\star_{Z1}, \mu_Y, q_Y\right)$. Ainsi, θ correspond aux paramètres de L et $\theta^\star$ à ceux de $L^\star$.

On a la relation :

$$\begin{aligned} q_Z &= \frac{\sigma}{2}\sqrt{1-r^2}\,(\mu^\star_{Z1} - \mu^\star_{Z-1}) \;+\; r\,q_Y \\ \mu_Z &= \frac{\sigma}{2}\sqrt{1-r^2}\,(\mu^\star_{Z1} + \mu^\star_{Z-1}) + \;r\,\mu_Y \end{aligned}$$

Soit M la matrice telle que $\theta = M\theta^\star$:

$$M = \begin{pmatrix} \frac{\sigma}{2}\sqrt{1-r^2} & \frac{\sigma}{2}\sqrt{1-r^2} & r & 0 \\ -\frac{\sigma}{2}\sqrt{1-r^2} & \frac{\sigma}{2}\sqrt{1-r^2} & 0 & r \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

L'inverse de M, notée M^{-1}, vérifie :

$$M^{-1} = \begin{pmatrix} \frac{1}{\sigma\sqrt{1-r^2}} & -\frac{1}{\sigma\sqrt{1-r^2}} & -\frac{r}{\sigma\sqrt{1-r^2}} & \frac{r}{\sigma\sqrt{1-r^2}} \\ \frac{1}{\sigma\sqrt{1-r^2}} & \frac{1}{\sigma\sqrt{1-r^2}} & -\frac{r}{\sigma\sqrt{1-r^2}} & -\frac{r}{\sigma\sqrt{1-r^2}} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

On note $\theta_{00} = (\mu_Z, 0, \mu_Y, 0)$ et $\theta^\star_{00} = M^{-1}\theta_{00}$. Ainsi,

$$\theta^\star_{00} = \left(\frac{\mu_Z}{\sigma\sqrt{1-r^2}} - \frac{r\mu_Y}{\sigma\sqrt{1-r^2}}\,,\; \frac{\mu_Z}{\sigma\sqrt{1-r^2}} - \frac{r\mu_Y}{\sigma\sqrt{1-r^2}}\,,\; \mu_Y\,,\; 0\right)$$

Notation 4.5 *On note I_θ (resp. $I^\star_{\theta^\star}$) la matrice d'information de Fisher relative à la vraisemblance L (resp. $L^\star$) et prise au point θ (resp. $\theta^\star$).*

Le calcul de $I^\star_{\theta^\star_{00}}$ est le suivant :

$$\frac{\partial \log L^\star}{\partial \mu_Y}\,|_{\theta^\star_{00}} = \frac{y - \mu_Y}{\sigma}$$

$$\frac{\partial \log L^\star}{\partial \mu^\star_{Z-1}}\,|_{\theta^\star_{00}} = (z^\star - \frac{\mu_Z}{\sigma\sqrt{1-r^2}} + \frac{r\mu_Y}{\sigma\sqrt{1-r^2}})\;\; 1_{\overline{X}=-1}$$

$$\frac{\partial \log L^\star}{\partial \mu^\star_{Z1}}\,|_{\theta^\star_{00}} = (z^\star - \frac{\mu_Z}{\sigma\sqrt{1-r^2}} + \frac{r\mu_Y}{\sigma\sqrt{1-r^2}})\;\; 1_{\overline{X}=1}$$

$$\frac{\partial \log L^\star}{\partial q_Y} \mid_{\theta^\star_{00}} = -\left(\frac{y-\mu_Y}{\sigma^2}\right) 1_{\overline{X}=-1} + \left(\frac{y-\mu_Y}{\sigma^2}\right) 1_{\overline{X}=1} + \left(\frac{y-\mu_Y}{\sigma^2}\right)(2p-1)\, 1_{\overline{X}=0}$$

D'où

$$I^\star_{11}(\theta^\star_{00}) = (1-p)\,\gamma \ , \ I^\star_{22}(\theta^\star_{00}) = p\,\gamma \text{ et } I^\star_{33}(\theta^\star_{00}) = \frac{1}{\sigma^2}$$

En utilisant les résultats du chapitre précédent, et en notant comme dans le chapitre précédent :

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

alors,

$$I^\star_{44}(\theta^\star_{00}) = \frac{\mathcal{A}}{\sigma^4} + \frac{(2p-1)^2}{\sigma^4}(\sigma^4 - \mathcal{A}) \quad \text{et} \quad I^\star_{34}(\theta^\star_{00}) = \frac{2p-1}{\sigma^2} \ .$$

D'autre part, tous les autre termes de $I^\star_{\theta^\star_{00}}$ sont nuls.
Si on note $\hat{\theta}$ et $\hat{\theta}^\star$ les vecteurs d'EMV, alors on a la relation $\hat{\theta} = M\hat{\theta}^\star$. Comme le modèle étudié est régulier :

$$\text{Var}\left\{\sqrt{n}\,(\hat{\theta}^\star - \theta^\star_{00})\right\} \overset{H_{0Y}H_{0Z}}{\rightarrow} I^{\star\,-1}_{\theta^\star_{00}}$$

Or $\sqrt{n}\,(\hat{\theta} - \theta_{00}) = \sqrt{n}\,M\,(\hat{\theta}^\star - \theta^\star_{00})$, d'où :

$$\text{Var}\left\{\sqrt{n}\,(\hat{\theta} - \theta_{00})\right\} \overset{H_{0Y}H_{0Z}}{\rightarrow} M\, I^{\star\,-1}_{\theta^\star_{00}}\, M^t$$

Par conséquent :

$$I^{-1}_{\theta_{00}} = M\, I^{\star\,-1}_{\theta^\star_{00}}\, M^t$$

Après calcul, on obtient :

$$I^{-1}_{22}(\theta_{00}) = \frac{\sigma^2\,(1-r^2)}{4\,p\,(1-p)\,\gamma} + \frac{\sigma^4\,r^2}{4\,p\,(1-p)\,\mathcal{A}}$$

On définit la statistique de Wald de la manière suivante :

$$W_1 := \frac{\sqrt{n}}{\sqrt{I^{-1}_{22}(\theta_{00})}}\ \hat{q}_Z$$

L'EMV $\hat{q}_Z$ pourra facilement être obtenu à l'aide des EMV $\hat{\mu}^\star_{Z-1}$, $\hat{\mu}^\star_{Z1}$, et $\hat{q}_Y$.
Comme le modèle est régulier :

$$W_1 \overset{H_{0Z}H_{0Y}}{\rightarrow} N(0,\,1)$$

On applique le théorème 2.9 successivement avec $h_n = h = (0, 0, 0, a)$, $h_n = h = (0, b, 0, 0)$, $h_n = h = (0, b, 0, a)$. Alors, on a :

$$\begin{aligned} W_1 &\overset{H_{0Z}H_{aY}}{\rightarrow} N(0,\ 1) \\ W_1 &\overset{H_{bZ}H_{0Y}}{\rightarrow} N\left(\frac{b}{\sqrt{I_{22}^{-1}(\theta_{00})}},\ 1\right) \\ W_1 &\overset{H_{bZ}H_{aY}}{\rightarrow} N\left(\frac{b}{\sqrt{I_{22}^{-1}(\theta_{00})}},\ 1\right) \end{aligned}$$

Ainsi, que l'on considère l'hypothèse nulle ou l'alternative locale pour le phénotype Y, on a toujours :

$$\begin{aligned} W_1 &\overset{H_{0Z}}{\rightarrow} N(0,\ 1) \\ W_1 &\overset{H_{bZ}}{\rightarrow} N\left(\frac{b}{\sqrt{I_{22}^{-1}(\theta_{00})}},\ 1\right) \end{aligned}$$

L'efficacité $\tilde{\kappa}_1$ de ce test, en prenant pour référence le test oracle (μ_Z, q_Z) s'obtient aisément :

$$\tilde{\kappa}_1 = \left\{ \frac{1-r^2}{\gamma} + \frac{r^2}{\gamma\ +\ z_{\gamma_+}\varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})} \right\}^{-1}$$

On remarque :

$$\tilde{\kappa}_1 = \left\{ \frac{1-r^2}{\gamma} + \frac{r^2}{\kappa_1} \right\}^{-1}$$

où κ_1 provient du théorème 3.3 page 46.

D'après le théorème 3.3, κ_1 est maximale pour $\gamma_+ = \gamma_- = \gamma/2$, il en est de même pour $\tilde{\kappa}_1$.

Deuxième stratégie (test de Wald utilisant uniquement les phénotypes Y extrêmes)

La vraisemblance pour une observation $(\overline{X}, Y, Z^\star)$ s'écrit :

$$\begin{aligned} L^\star(\mu^\star_{Z-1}, \mu^\star_{Z1}, \mu_Y, q_Y) = \mathbb{P}(\overline{X}=0)\ 1_{\overline{X}=0}\ +\ \frac{p}{\sigma}\varphi\left(\frac{y-\mu_Y-q_Y}{\sigma}\right)\varphi(z^\star-\mu^\star_{Z1})\ 1_{\overline{X}=1} \\ +\ \frac{1-p}{\sigma}\varphi\left(\frac{y-\mu_Y+q_Y}{\sigma}\right)\varphi(z^\star-\mu^\star_{Z-1})\ 1_{\overline{X}=-1} \end{aligned}$$

$I^{\star}_{11}(\theta^{\star}_{00})$ et $I^{\star}_{22}(\theta^{\star}_{00})$ sont inchangés par rapport à ce qui précède :

$$I^{\star}_{11}(\theta^{\star}_{00}) = (1-p)\,\gamma \;\;,\;\; I^{\star}_{22}(\theta^{\star}_{00}) = p\,\gamma$$

Le calcul des autres composantes de l'information de Fisher est le suivant :

$$\frac{\partial \log L^{\star}}{\partial \mu_Y} \mid_{\theta^{\star}_{00}} = \frac{y-\mu_Y}{\sigma^2}\,\left\{1_{\overline{X}=-1} + 1_{\overline{X}=1}\right\} \;+\; \frac{\varphi\left(z_{1-\gamma_-}\right) - \varphi\left(z_{\gamma_+}\right)}{\sigma\,(1-\gamma)}\,1_{\overline{X}=0}$$

$$I^{\star}_{33}(\theta^{\star}_{00}) = \frac{\mathcal{A}}{\sigma^4} \;+\; \frac{\left\{\varphi\left(z_{1-\gamma_-}\right) - \varphi\left(z_{\gamma_+}\right)\right\}^2}{\sigma^2\,(1-\gamma)}$$

D'après l'annexe 3.4.5 du chapitre 3 :

$$I^{\star}_{44}(\theta^{\star}_{00}) = \frac{\mathcal{A}}{\sigma^4} \;+\; (2p-1)^2\,\frac{\left\{\varphi\left(z_{1-\gamma_-}\right) - \varphi\left(z_{\gamma_+}\right)\right\}^2}{\sigma^2\,(1-\gamma)}$$

De plus :

$$\begin{aligned}\frac{\partial \log L^{\star}}{\partial \mu_Y \partial q_Y} \mid_{\theta^{\star}_{00}} &= \frac{1}{\sigma^2}(1_{\overline{X}=-1} - 1_{\overline{X}=1}) \;+\; \frac{2p-1}{\sigma^2(1-\gamma)}\left\{z_{1-\gamma_-}\varphi(z_{1-\gamma_-}) - z_{\gamma_+}\varphi(z_{\gamma_+})\right\}1_{\overline{X}=0} \\ &\quad - \frac{2p-1}{\sigma^2(1-\gamma)^2}\left\{\varphi(z_{1-\gamma_-}) - \varphi(z_{\gamma_+})\right\}^2\,1_{\overline{X}=0}\end{aligned}$$

Par conséquent :

$$I^{\star}_{34}(\theta^{\star}_{00}) = (1-2p)\left[\frac{\mathcal{A}}{\sigma^4} + \frac{\left\{\varphi(z_{1-\gamma_-}) - \varphi(z_{\gamma_+})\right\}^2}{\sigma^2(1-\gamma)}\right]$$

Tous les autres termes de l'information de Fisher sont nuls. On cherche désormais à inverser cette matrice. Pour cela, on calcule les inverses par blocs.
On obtient :

$$\begin{aligned}{I^{\star}_{11}}^{-1}(\theta^{\star}_{00}) &= \frac{1}{(1-p)\,\gamma} \quad , \quad {I^{\star}_{22}}^{-1}(\theta^{\star}_{00}) = \frac{1}{p\,\gamma} \\ {I^{\star}_{44}}^{-1}(\theta^{\star}_{00}) &= \frac{\sigma^4}{4\,\mathcal{A}\,p(1-p)}\end{aligned}$$

On note

$$\Lambda := \left[\frac{4\,\mathcal{A}\,p(1-p)}{\sigma^4}\left\{\frac{\mathcal{A}}{\sigma^4} + \frac{\left\{\varphi(z_{\gamma_+}) - \varphi(z_{1-\gamma_-})\right\}^2}{\sigma^2\,(1-\gamma)}\right\}\right]^{-1}$$

Alors d'après l'annexe 3.4.5 du chapitre 3,

$$I_{33}^{\star\,-1}(\theta_{00}^{\star}) = \frac{\Lambda}{\sigma^4}\left\{\mathcal{A} + (2p-1)^2\frac{\left\{\varphi(z_{\gamma_+}) - \varphi(z_{1-\gamma_-})\right\}^2}{1-\gamma}\right\}$$

$$I_{34}^{\star\,-1}(\theta_{00}^{\star}) = \Lambda\,(2p-1)\,\left[\frac{\mathcal{A}}{\sigma^4} + \frac{\left\{\varphi(z_{\gamma_+}) - \varphi(z_{1-\gamma_-})\right\}^2}{\sigma^2\,(1-\gamma)}\right]$$

De la même manière que précédemment :

$$I_{\theta_{00}}^{-1} = M\; I_{\theta_{00}^{\star}}^{\star\,-1}\; M^t$$

Après calcul, on obtient :

$$I_{22}^{-1}(\theta_{00}) = \frac{\sigma^2(1-r^2)}{4\,\gamma\,p(1-p)} \;+\; \frac{r^2\sigma^4}{4\,\mathcal{A}\,p(1-p)}$$

On en déduit le test de Wald (mêmes arguments que pour la première stratégie) :

$$W_2 := \frac{\sqrt{n}}{\sqrt{I_{22}^{-1}(\theta_{00})}}\;\hat{q}_Z \;\overset{H_{0Z}}{\rightarrow}\; N(0,\,1)$$

$$W_2 \;\overset{H_{bZ}}{\rightarrow}\; N\left(\frac{b}{\sqrt{I_{22}^{-1}(\theta_{00})}},\,1\right)$$

L'EMV $\hat{q}_Z$ sera obtenu à l'aide de $\hat{\mu}_{Z-1}^{\star}$, $\hat{\mu}_{Z1}^{\star}$ et $\hat{q}_Y$ ($\hat{q}_Y$ pourra être obtenu par une méthode de Newton).

On remarque que ce test possède même puissance que celui présenté pour la première stratégie. Ce qui conclut la preuve.■

La convergence vers l'asymptotique sera illustrée en annexe 4.4.1.

Illustration graphique

La figure 4.1 présente l'efficacité en fonction de γ et de r ($\gamma_+ = \gamma_- = \frac{\gamma}{2}$). Naturellement, l'efficacité augmente avec r et γ. On remarque également que lorsque $\gamma = 1$, comme on dispose alors de l'ensemble des phénotypes Z, l'efficacité est égale à un.

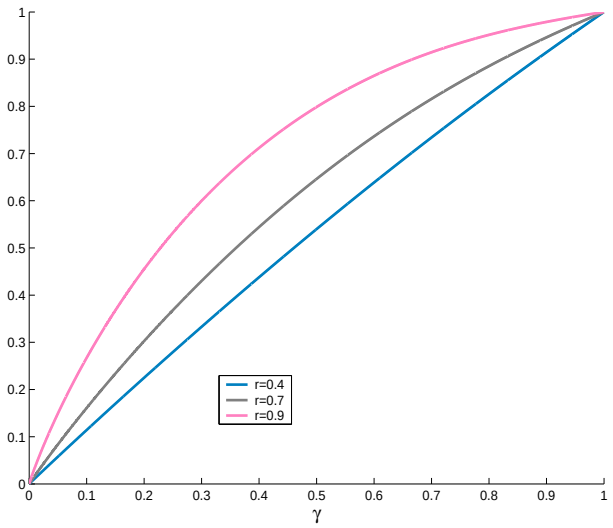


FIG. 4.1 – Efficacité en fonction de γ et de r ($\gamma_+ = \gamma_- = \frac{\gamma}{2}$)

4.3.3 Résultats secondaires

On présente dans cette section, des résultats à propos des différentes stratégies pour les modèles statistiques (q_Z) et (q_Z, q_Y).

Résultat théorique pour le modèle (q_Z, q_Y)

Corollaire 4.6 *Si l'on considère un modèle statistique à deux paramètres* (q_Z, q_Y)*, alors :*

$$i)\quad \tilde{\kappa_1} = \left\{ \frac{1-r^2}{\gamma} + \frac{r^2}{\kappa_1} \right\}^{-1}$$

$$ii)\quad \tilde{\kappa}_2 = \left\{ \frac{1-r^2}{\gamma} + \frac{r^2}{\kappa_3} \right\}^{-1}$$

$$iii)\quad \tilde{\kappa}_1 = \tilde{\kappa}_2 \Leftrightarrow p = \frac{1}{2}$$

$$iv)\quad \forall p \quad \tilde{\kappa}_1 \text{ et } \tilde{\kappa}_2 \text{ sont maximum pour } \gamma_+ = \gamma_- = \frac{\gamma}{2}$$

κ_1 et κ_3 désignent les quantités présentées en corollaire 3.9 page 57.

Contrairement au résultat obtenu pour un modèle (μ_Z, q_Z, μ_Y, q_Y), les deux stratégies sont ici équivalentes uniquement lorsque l'on considère une population backcross ($p = 1/2$).

La preuve du corollaire 4.6 est largement inspirée de celle du théorème 4.3 et de celle du corollaire 3.9.

Illustration graphique du modèle (q_Z, q_Y)

La figure 4.2 présente les efficacités correspondant aux différentes stratégies en fonction de p lorsque $\gamma = 0.3$ et $r = 0.7$. $\tilde{\kappa}_2$ est constante en raison du contexte de symétrie imposé ($\gamma_+ = \gamma_- = \frac{\gamma}{2}$). Le cas $p = 1/2$ correspond à la pire des situations pour la première stratégie. Bien évidemment, lorsque $p \neq 1/2$, on a $\tilde{\kappa}_1 > \tilde{\kappa}_2$.

Enfin, la figure 4.3 présente $\tilde{\kappa}_1$ en fonction de p et r. On peut facilement en déduire l'évolution de la figure 4.2 avec r.

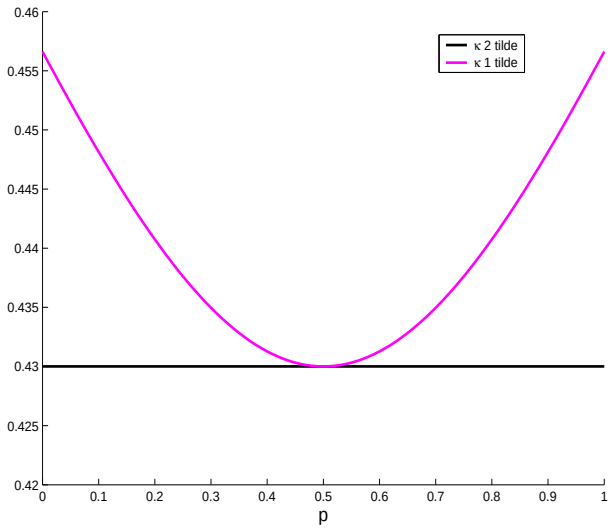


FIG. 4.2 – Efficacité des tests correspondant aux différentes stratégies, en fonction de p ($\gamma = 0.3$, $\gamma_+ = \gamma_- = \frac{\gamma}{2}$, $r = 0.7$)

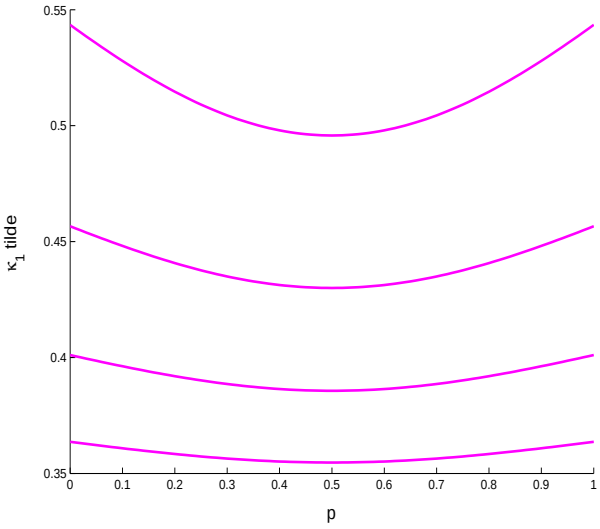


FIG. 4.3 – Efficacité du test correspondant à la première stratégie, en fonction de p et r ($\gamma = 0.3$, $\gamma_+ = \gamma_- = \frac{\gamma}{2}$). De bas en haut, $r = 0.5, 0.6, 0.7, 0.8$.

Résultat théorique pour le modèle (q_Z)

Corollaire 4.7 *Si l'on considère un modèle statistique à un paramètre (q_Z), alors $\forall p$:*

$$\tilde{\kappa}_1 = \tilde{\kappa}_2 = \frac{\mathbb{P}\left(Y \notin [S_-, S_+]\right)}{1 - r^2}$$

Ici, l'effet QTL q_Y est une constante connue. La quantité $\mathbb{P}(Y \notin [S_-, S_+])$ dépend naturellement de q_Y. $\mathbb{P}(Y \notin [S_-, S_+])$ désigne asymptotiquement le pourcentage d'individus génotypés.

On remarque que les efficacités dépendent uniquement du pourcentage total d'individus génotypés et non pas du pourcentage d'individus génotypés à droite et à gauche, comme précédemment. On constate également que le selective genotyping s'avère plus intéressant que le test oracle lorsque l'on a $\mathbb{P}(Y \notin [S_-, S_+]) > 1 - r^2$.

La preuve du corollaire 4.7 est largement inspirée des preuves précédentes.

4.3.4 Résumé des différents résultats

Les tables 4.1 et 4.2, résument les différents résultats obtenus à propos des différentes stratégies.

Modèle	Stratégie	Décentrement
(μ_Z, q_Z, μ_Y, q_Y)	1 et 2	$\frac{b}{\sigma}\left(\sqrt{\frac{(1-r^2)}{4\ p\ (1-p)\ \gamma} + \frac{\sigma^2\ r^2}{4\ p\ (1-p)\ \mathcal{A}}}\right)^{-1}$
(q_Z, q_Y)	1	$\frac{b}{\sigma}\left(\sqrt{\frac{(1-r^2)}{\gamma} + \frac{\sigma^2\ r^2}{\mathcal{A} + (2p-1)^2(\sigma^2-\mathcal{A})}}\right)^{-1}$
	2	$b\,\sigma\left(\sqrt{\frac{\mathcal{A}}{\sigma^2} + \frac{(2p-1)^2\ \left\{\varphi(z_{\gamma_+})-\varphi(z_{1-\gamma_-})\right\}^2}{1-\gamma}}\right)^{-1}$
(q_Z)	1 et 2	$\frac{b}{\sigma}\sqrt{\mathbb{P}(Y \notin [S_-, S_+])/(1-r^2)}$

TAB. 4.1 – Table de décentrement

Modèle	Stratégie	Efficacité
(μ_Z, q_Z, μ_Y, q_Y)	1 et 2	$\left[\frac{1-r^2}{\gamma} + \frac{r^2}{\kappa_1}\right]^{-1}$ où $\kappa_1 = \gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-})$
(q_Z, q_Y)	1	$\left[\frac{1-r^2}{\gamma} + \frac{r^2}{\kappa_1}\right]^{-1}$
	2	$\left[\frac{1-r^2}{\gamma} + \frac{r^2}{\kappa_3}\right]^{-1}$ où $\kappa_1 = \gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-}) + (2p-1)^2\left\{1-\gamma - z_{\gamma_+}\varphi(z_{\gamma_+}) + z_{1-\gamma_-}\varphi(z_{1-\gamma_-})\right\}$ et $\kappa_3 = \gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-}) + \frac{(2p-1)^2}{1-\gamma}\left\{\varphi(z_{1-\gamma_-}) - \varphi(z_{\gamma_+})\right\}^2$
(q_Z)	1 et 2	$\mathbb{P}(Y \notin [S_-, S_+])/(1-r^2)$

TAB. 4.2 – Table d'efficacité

4.4 Annexe

4.4.1 Convergence vers l'asymptotique

On s'intéresse au modèle statistique (μ_Z, q_Z, μ_Y, q_Y). Afin de valider les résultats asymptotiques, on présente ici le résultat de simulations. On s'intéresse au test unilatéral de niveau 5% basé sur la statistique W_1 (première stratégie) énoncée dans la preuve du théorème 4.3. .

La table 4.3 présente le pourcentage de faux positifs en fonction de a et de r ($\gamma = 0.3$). 10000 échantillons de taille $n = 50$ ont été simulés. Un intervalle de confiance à 95% pour la vraie valeur du niveau (ie. le pourcentage de faux positifs) a été calculé. On remarque, que même si le vrai niveau est parfois significativement différent de 5%, les résultats semblent globalement satisfaisants.

La table 4.4 présente les puissances théoriques (basées sur les résultats asymptotiques), notées β_1, et les puissances empiriques, notées β_{MC}, obtenues pour $n = 50$. Un effet $q_Z = 0.5657$ a été considéré. Un intervalle de confiance pour la vraie valeur de la puissance a été calculé. On constate que les puissances théoriques β_1 se situent toujours dans l'intervalle de confiance, ce qui signifie que lorsque $n = 50$, n est suffisamment grand pour que l'on puisse considérer les résultats asymptotiques.

a	$q_Y = \frac{a}{\sqrt{n}}$	r	faux positifs	IC en %
0	0	0.4	5.81 %	[5.35 ; 6.27]
0	0	0.7	5.76 %	[5.30 ; 6.22]
0	0	0.9	5.62 %	[5.17 ; 6.07]
2	0.2828	0.4	4.87 %	[4.45 ; 5.29]
2	0.2828	0.7	5.13 %	[4.70 ; 5.56]
2	0.2828	0.9	4.71 %	[4.29 ; 5.13]

TAB. 4.3 – Pourcentage de faux positifs en fonction de a et de r ($b = 0$, $q_Z = 0$, $n = 50$, 10000 échantillons, $\sigma = 1$, $\mu_Y = 0$, $\mu_Z = 0$, $\gamma = 0.30$)

a	$q_Y = \frac{a}{\sqrt{n}}$	r	β_{MC}	β_1	IC en %
0	0	0.4	73.96 %	74.47 %	[73.10 ; 74.82]
0	0	0.7	82.23 %	82.31 %	[81.48 ; 82.98]
0	0	0.9	92.24 %	92.61 %	[91.72 ; 92.76]
2	0.2828	0.4	74.72 %	74.47 %	[73.87 ; 75.57]
2	0.2828	0.7	83.18 %	83.47 %	[82.45 ; 83.91]
2	0.2828	0.9	92.21 %	92.61 %	[91.68 ; 92.74]

TAB. 4.4 – Puissance théorique (β_1) et puissance par Monte-Carlo (β_{MC}) ($b = 4$, $q_Z = 0.5657$, $n = 50$, 10000 échantillons, $\sigma = 1$, $\mu_Y = 0$, $\mu_Z = 0$, $\gamma = 0.30$)

Deuxième partie

Génome Scan

On se propose dans cette partie d'étudier la technique qui consiste à scanner le génome, afin de détecter et de localiser des QTL. Après avoir introduit le contexte de l'étude, nous développons dans différents chapitres la méthodologie statistique sous-jacente au génome scan.

Notation

Y	phenotype
n	number of observations
j	index of an observation
$[0,T]$	segment representing a chromosome
t	index of a location on $[0,T]$
$t^\star$	location of the QTL
X_t	genotype at location t
$X_{t^\star}$	genotype at the QTL
H_0	null hypothesis (no QTL on $[0,T]$)
θ	parameter of the statistical model
θ_0	true value of the parameter under H_0
I_{θ_0}	Fisher information matrix taken at the point θ_0
$I_{ij}(\theta_0)$	element ij of I_{θ_0}
$I_{\theta_0}^{-1}$	inverse of I_{θ_0}
$I_{ij}^{-1}(\theta_0)$	element ij of $I_{\theta_0}^{-1}$
$\Lambda_{(.)}$	LRT process

Chapter 2 and 3

One family

q	QTL effect
$\left\{Z_{(.)}\right\}^2$	limiting process of $\Lambda_{(.)}$
$V_{(.)}$	linear interpolated process
$H_{\lambda t^\star}$	local alternative hypothesis (there is a QTL of effect $q = \frac{\lambda}{\sqrt{n}}$ at position $t^\star$)
K	number of genetic markers
t_k	location of marker k
t^ℓ	closest marker on the left-side of t
t^r	closest marker on the right-side of t

Several families

I	number of families
i	index of the family
q_i	QTL effect inside family i
$\sum_{i=1}^{I} \left\{ Z_{(.)}^{i} \right\}^2$	limiting process of $\Lambda_{(.)}$
$H_{\lambda t^\star}$	local alternative hypothesis (there is at least one $q_i = \frac{\lambda_i}{\sqrt{n}}$ with $\lambda_i \in \mathbb{R}^\star$ at the position $t^\star$)
K^i	number of genetic markers in family i
t_k^i	location of marker k in family i
$t^{\ell,i}$	closest marker on the left-side of t in family i
$t^{r,i}$	closest marker on the right-side of t in family i

Chapitre 1

Introduction

1.1 Contexte

L'étude porte sur une population de descendants d'un père. La problèmatique se veut la détection d'un QTL sur un chromosome. Le phénotype est observé sur n individus. On note Y_j, $j = 1, .., n$, les observations que l'on supposera indépendamment et identiquement distribuées (iid). Le mécanisme de la génétique implique que parmi les deux chromosomes de chaque individu, l'un est herité de la mère (son effet sera négligé) et l'autre du père. Celui transmis par le père est constitué, en raison de crossing-over, de parties provenant du chromosome 1 du père et de parties provenant du chromosome 2 du père.
Une population backcross, $C \times (C \times D)$, où C et D sont de pures lignées homozygotes, est un cas particulier de la population étudiée. A l'aide de la distance et de la modélisation de Haldane (1919), on représente chaque chromosome par un segment $[0, T]$. La distance sur $[0, T]$ est appelée distance génétique. On la mesure en Morgans. Le point clé est que, si la vraie position du QTL est $t = t^\star$, le phénotype Y obéit à un modèle de mélange dont les poids sont connus :

$$p_t f_{(\mu+q,\sigma)}(.) \; + \; (1 - p_t) f_{(\mu-q,\sigma)}(.) \tag{1.1}$$

où $f_{(\mu,\sigma)}(.)$ désigne une densité Gaussienne de moyenne μ et de variance σ^2. (μ, q, σ) sont les paramètres inconnus. A chaque position $t \in [0, T]$, est effectué un test du rapport de vraisemblance (LRT), testant l'hypothèse "q=0" dans la formule (1.1), en considérant n observations Y_1, ..., Y_n. La quantité obtenue est notée Λ_t. Le processus $\Lambda_{(.)}$ est appelé processus de tests de rapport de vraisemblance. Le choix comme statistique de test du maximum de ce processus, revient à effectuer un LRT dans un modèle où la localisation du QTL est un paramètre supplémentaire. Cette technique qui consiste à scanner le génome a été proposée par Lander and Botstein (1989). On la nomme "Interval Mapping".

Dans le cas particulier où les poids du mélange sont 0 ou 1 en fonction des individus,

Lander and Botstein (1989) ont affirmé que la distribution asymptotique du processus de LRT, $\Lambda_{(.)}$, sur l'intervalle $[0, T]$, était le carré d'un processus d'Ornstein-Uhlenbeck. Ce résultat a été prouvé par Cierco (1998). Des bornes pour la distribution du maximum d'un processus d'Ornstein Uhlenbeck régularisé ont été proposées par Cierco (1996), Azaïs and Wschebor (2009). Des résultats sur la distribution asymptotique du processus de LRT sous l'hypothèse nulle sont présentés dans Rebaï and al. (1994) pour une modélisation particulière des poids du mélange. Ces derniers résultats s'appuient sur les bornes données par Davies (1977), Davies (1987) pour le maximum de processus Gaussiens suffisamment réguliers et pour des processus de chi-deux.

1.2 Feuille de route

Dans notre étude, sont considérés les poids correspondant exactement à la modélisation de Haldane (1919), ce qui nous démarque des travaux déjà existants. Ces poids se distinguent de part leur complexité. On s'attarde également sur les poids considérés par Rebaï afin de généraliser les processus étudiés dans Rebaï and al. (1994), Rebaï and al. (1995). On rappelle que dans la modélisation de Haldane (1919), les crossing-over sont indépendants et que leur nombre suit un processus de poisson d'intensité 1. Dans Rebaï and al. (1994) et Rebaï and al. (1995), les auteurs autorisent uniquement une seule recombinaison entre les marqueurs génétiques (phénomène d'interférence). Par conséquent, la notion de processus de Poisson à accroissements indépendants est perdue. Néanmoins, nous prouvons que les poids utilisés par Rebaï correspondent à une approximation au premier ordre des poids correspondant exactement à la modélisation de Haldane.

Le chapitre 2 présente des résultats asymptotiques sur la distribution du processus $\Lambda_{(.)}$ sous l'hypothèse nulle d'absence de QTL et sous l'alternative, où il existe un seul QTL sur l'intervalle $[0, T]$. Les démonstrations sont uniquement données dans leur grandes lignes. L'accent est davantage porté sur l'illustration, notamment à travers de graphiques représentant des trajectoires des différents processus étudiés. Le résultat théorique est généralisé au cas de populations structurées en familles de pères possédant leurs propres marqueurs informatifs. En effet, la motivation est d'obtenir des résultats asymptotiques sur le processus $\Lambda_{(.)}$ pour les familles outbred dont l'informativité des marqueurs diffèrent à la fois entre les familles de pères et aussi entre descendants issus d'un même père (cf. page 24). Analytiquement, nous avons été en mesure de faire varier l'informativité des marqueurs entre les familles. Cependant, la variation d'informativité au sein d'une même famille semble être une condition trop lourde pour une approche paramétrique théorique. Les calculs présentés ont nécessité des notations particulières afin de pouvoir tenir compte que chaque famille de pères possède ses propres marqueurs informatifs.

Le chapitre 2 se conclut par deux articles soumis. L'un présente des résultats théoriques

et l'autre la partie application des résultats théoriques du premier article.

L'article théorique (cf. page 127) se distingue de la première partie du chapitre 2 : sa motivation première n'est pas d'obtenir des résultats pour les populations outbred mais de présenter des résultats théoriques sur le processus $\Lambda_{(.)}$ notamment lorsque plusieurs QTL sont présents sur l'intervalle $[0, T]$. Une hypothèse d'additivé des effets des QTL a été faite. Cependant, ces résultats sont généralisables au cas d'intéractions entre les QTL : l'épistasie. L'accent est porté ici sur la théorie : les démonstrations mathématiques présentées sont plus rigoureuses (on s'intéresse notamment à la convergence faible du processus).

Dans l'article applicatif (cf. page 128), sont proposées de nombreuses méthodes, adaptées à la carte génétique, et permettant d'obtenir le quantile d'ordre 95% du maximum du processus de LRT sous l'hypothèse nulle. De plus, ces méthodes, basées sur des résultats théoriques asymptotiques, sont validées au moyen de populations simulées. On s'attarde également sur la puissance de la méthode de détection de QTL en comparant notamment une procédure tests multiples et un test global.

Le chapitre 3 propose des résultats sur le maximum d'un processus d'interpolation linéaire qui s'avère très proche du processus $\Lambda_{(.)}$. Ce processus d'interpolation linéaire est la généralisation du processus présent dans Rebaï and al. (1994), Rebaï and al. (1995). Dans ce chapitre, on considère une seule famille. On cherche principalement à répondre à un article de Walling où l'auteur faisait l'observation à travers de simulations, que la loi de l'argmax du processus n'était pas une loi uniforme sur $[0, T]$. Bien au contraire, le maximum du processus était bien plus fréquemment obtenu sur les marqueurs génétiques qu'entre les marqueurs.

Dans notre étude, nous prouvons qu'il s'avère inutile d'effectuer des tests sur tout l'intervalle $[0, T]$. Nous conseillons plus précisément où tester.

Dans **le chapitre 4**, on présente un travail en cours, en collaboration avec Alan Genz (Université de Washington). On y propose notamment une formule théorique, basée sur Delong (1981), et permettant d'obtenir le quantile du maximum du processus de Chi-Deux d'Ornstein-Uhlenbeck à I degrés de libertés sous l'hypothèse nulle. En effet, ce processus est le processus limite du processus $\Lambda_{(.)}$ lorsque l'on considère I familles, et que la distance entre les marqueurs génétiques tend vers zéros.

A noter que les notations sont propres à ce chapitre. Elles sont différentes de celles des chapitres 2 et 3.

Chapitre 2

Asymptotic process for QTL detection

2.1 Introduction

In this chapter, is studied the asymptotic distribution of the LRT process, $\Lambda_{(.)}$, under the null hypothesis (no QTL on the interval $[0, T]$), and under the alternative that there is one QTL located at position $t^\star$ on $[0, T]$. The theoretical result is generalized to families of sires with their own informative markers. These results will be largely illustrated.

At the end of the chapter, two submitted articles are presented : the first is a theoretical one whereas the second deals with applications.

In the first article, the motivation is to obtain asymptotic results about $\Lambda_{(.)}$ under the general alternative that there are m QTL on $[0, T]$. We suggest then to estimate the number of QTL, their effects and their positions using penalized likelihood method (for instance lasso Tibshirani (1996)).

In the second article, we propose several new methods to calculate threshold and power for QTL detection. The methods proposed are fast and easy to implement. A comparison of power between a multiple testing procedure and a global test is realized, showing far better performances of the global test for the detection of a QTL.

2.2 Model

K genetic markers are located on the chromosome, one at each extremity. $0 = t_1 < t_2 < ... < t_K = T$ are the locations of the markers. The goal is to test if there is a QTL lying on the interval $[0, T]$.

Let $N_{(.)}$ be a standard Poisson process and let δ be the Dirac measure. X_t will refer to the "genome information". The law of X_{t_1} is $\frac{1}{2}(\delta_1 + \delta_{-1})$ and $X_t = (-1)^{N_t} X_{t_1}$. However, the "genome information" is available only at locations of genetic markers, that is to say

at $t_1, t_2, ..., t_K$. Let define the Haldane (1919)'s function $r : [0,\ T]^2 \longmapsto \left[0,\ \frac{1}{2}\right]$ such as :

$$r(t,t') = P(X_t X_{t'} = -1) = P(|N_t - N_{t'}| \text{ is odd}) = \frac{1}{2}\left(1 - e^{-2|t-t'|}\right)$$

$\bar{r}(t,t')$ will be the function equal to $1 - r(t,t')$.
We are interested in a quantitative trait Y which depends on the value of X_t at $t^\star \in [0,T]$ which is the location of the QTL. The quantitative trait verifies :

$$Y = \mu \ + \ X_{t^\star}\, q \ + \ \varepsilon$$

where q is the effect of the QTL and ε is a Gaussian white noise.
The goal of this study is to test whether q is equal to zero. The challenge is that $t^\star$ is unknown. A likelihood ratio test (LRT) will be performed at each position $t \in [0,T]$. The asymptotic process generated by all these likelihood ratio tests will be studied under the null hypothesis that there is no QTL on $[0,T]$, and under contiguous alternatives. The study on this process will allow us to obtain some results about the law of the supremum of the process. Indeed, when an Interval Mapping is performed, the supremum of all the likelihood ratio tests on $[0,T]$ is used as a unique test statistic.

2.3 Only two genetic markers

2.3.1 Likelihood Ratio Test

In this section, are considered only two genetic markers located respectively at 0 and $T : 0 = t_1 < t_2 = T$. As explained previously, we are looking for a QTL lying at a position $t^\star \in [t_1, t_2]$. Let $t \in [t_1, t_2]$. Let p_t the quantity equal to $P(X_t = 1 / X_{t_1}\ X_{t_2})$.
It verifies $\forall t \in]t_1, t_2[$:

$$p_t = Q_t^{1,1}\, 1_{X_{t_1}=1} 1_{X_{t_2}=1} \ + \ Q_t^{1,-1}\, 1_{X_{t_1}=1} 1_{X_{t_2}=-1} \ + \ Q_t^{-1,1}\, 1_{X_{t_1}=-1} 1_{X_{t_2}=1} \ + \ Q_t^{-1,-1}\, 1_{X_{t_1}=-1} 1_{X_{t_2}=-1} \tag{2.1}$$

where :

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

We can remark that we have :

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

Besides, $p_{t_1} = 1_{X_{t_1}=1}$ and $p_{t_2} = 1_{X_{t_2}=1}$. So, the weights p_t are continuous at t_1 and t_2. Since the increments of a Poisson process are independent, the density of $Y/X_{t_1}X_{t_2}$ at a position $t \in [t_1, t_2]$ verifies :

$$p_t \, f_{(\mu+q,\sigma)}(y) \; + \; (1 - p_t) \, f_{(\mu-q,\sigma)}(y)$$

where $f_{(\mu,\, \sigma)}(y)$ refers to the density of the normal distribution with mean μ, variance σ^2, at the point y.

Let $\theta = (q,\, \mu,\, \sigma)$ be the parameter of the model at t fixed. θ_0 will be the true value of the parameter under $H_0 : \theta_0 = (0,\, \mu,\, \sigma)$. Let $n \in \mathbb{N}^\star$ be the number of observations. The likelihood $L_n(\theta, t)$ under the alternative, at a position $t \in [t_1, t_2]$, and for n observations j iid $\left(X^j_{t_1},\, X^j_{t_2},\, Y_j\right)$ is :

$$L_n(\theta, t) = \prod_{j=1}^{n} \left[p^j_t f_{(\mu+q,\sigma)}(y_j) \; + \; (1 - p^j_t) f_{(\mu-q,\sigma)}(y_j)\right] \; g_j(t)$$

where

$$\begin{aligned} g_j(t) = \frac{1}{2} \left\{ \bar{r}(t_1, t_2) \, 1_{X^j_{t_1}=1} 1_{X^j_{t_2}=1} \; + \; r(t_1, t_2) \, 1_{X^j_{t_1}=1} 1_{X^j_{t_2}=-1} \right\} \\ + \frac{1}{2} \left\{ r(t_1, t_2) \, 1_{X^j_{t_1}=-1} 1_{X^j_{t_2}=1} \; + \; \bar{r}(t_1, t_2) \, 1_{X^j_{t_1}=-1} 1_{X^j_{t_2}=-1} \right\} \end{aligned}$$

Let $\hat{\theta}$ be the maximum likelihood estimator (MLE) of θ and $\tilde{\theta}$ the MLE under H_0 :

$$\begin{aligned} \hat{\theta} &= (\hat{q},\, \hat{\mu},\, \hat{\sigma}) \\ \tilde{\theta} &= (0,\, \tilde{\mu},\, \tilde{\sigma}) \end{aligned}$$

Let Λ_t be the likelihood ratio test (LRT) performed at the position $t \in [t_1, t_2]$:

$$\Lambda_t = 2 \; \left[\log L_n(\hat{\theta}, t) - \log L_n(\tilde{\theta}, t) \right] \tag{2.2}$$

Notation 2.1 *I_θ will be the Fisher information matrix taken at the point θ . $I_{ij}(\theta)$ refers to the element ij of I_θ. $I^{-1}_{ij}(\theta)$ refers to the element ij of I^{-1}_θ, the inverse of I_θ.*

$$\frac{\partial \log L_1}{\partial q} \, |_{\theta_0} = \frac{y - \mu}{\sigma^2} \; (2p_t - 1)$$

$$\frac{\partial \log L_1}{\partial \mu} \, |_{\theta_0} = \frac{y - \mu}{\sigma^2} \quad , \quad \frac{\partial \log L_1}{\partial \sigma} \, |_{\theta_0} = -\frac{1}{\sigma} \; + \; \frac{(y - \mu)^2}{\sigma^3}$$

We remind that :

$$I_\theta = \mathbb{E}\left[\left(\frac{\partial \log L_1}{\partial \theta} \mid_\theta\right)^2\right]$$

It comes :

$$I_{11}(\theta_0) = \frac{\mathbb{E}\left[(2p_t - 1)^2\right]}{\sigma^2} \quad , \quad I_{22}(\theta_0) = \frac{1}{\sigma^2}$$

As the fourth-order moment of a standard normal distribution is equal to three :

$$I_{33}(\theta_0) = \frac{2}{\sigma^2}$$

After some calculations, we find : $I_{12}(\theta_0) = I_{13}(\theta_0) = I_{23}(\theta_0) = 0$. So,

$$I_{\theta_0} = Diag\left(\frac{\mathbb{E}\left[(2p_t - 1)^2\right]}{\sigma^2}\ ,\ \frac{1}{\sigma^2}\ ,\ \frac{2}{\sigma^2}\right)$$

The formula for $\mathbb{E}\left[(2p_t - 1)^2\right]$ is given in Appendix 2.6.1. This quantity is always different from 0 since t is bounded.

Notation 2.2 *By convention, the notation $o_{P_{\theta_0}}(1)$ is short for a sequence of random vectors that converges to zero in probability under H_0 (i.e. no QTL on the whole interval studied).*

As the model is regular, it comes :

$$\Lambda_t = \left(\sum_{j=1}^{n} \frac{(y_j - \mu)\,(2p_t^j - 1)}{\sigma\,\sqrt{n}\,\sqrt{\mathbb{E}\left[(2p_t - 1)^2\right]}}\right)^2 + o_{P_{\theta_0}}(1) \tag{2.3}$$

As p_t and $\mathbb{E}\left[(2p_t - 1)^2\right]$ are continuous at t_1 and t_2, $\Lambda_{(.)}$ is continuous at t_1 and t_2.

2.3.2 Process under H_0

First, we remind that the score test is based on the score function which is the gradient, with respect to θ, of the logarithm of the likelihood function.
So, let S_t be the score statistic at position t for n observations. $\forall t \in [t_1, t_2]$:

$$S_t = \sum_{j=1}^{n} \frac{(y_j - \mu)\,(2p_t^j - 1)}{\sigma\,\sqrt{n}\,\sqrt{\mathbb{E}\left[(2p_t - 1)^2\right]}}$$

Note that the score test can be obtained, replacing μ by $\hat{\mu}$, according to Prohorov, and replacing σ by $\hat{\sigma}$, according to Slutsky's lemma. Nevertheless, in order to make the reading

easier, the score statistic is defined as above.
According to formula (2.3) :

$$\Lambda_t = (S_t)^2 \ + \ o_{P_{\theta_0}}(1)$$

By the central limit theorem, $S_t \rightarrow N(0,\ 1)$. Besides, $\forall (t, t') \in [t_1, t_2]$:

$$\Gamma(t, t') := \text{Cov}\,(S_t,\ S_{t'}) = \frac{4\mathbb{E}(p_t p_{t'}) - 1}{\sqrt{\mathbb{E}\left[(2p_t - 1)^2\right]}\sqrt{\mathbb{E}\left[(2p_{t'} - 1)^2\right]}}$$

The explicit formula for the covariance is given in appendix 2.6.2. As $|p_t p_{t'}| \leqslant 1$, by dominated convergence theorem, $\mathbb{E}[p_t p_{t'}]$ is continuous at (t_1, t'), (t_2, t') and (t_1, t_2). Then the covariance function is continuous at this points (because the denominator is also continuous). So, the covariance is a continuous function on $[t_1, t_2] \times [t_1, t_2]$.

It comes, under H_0 :

$$\Lambda_{(.)} \overset{F.d.}{\rightarrow} \left\{Z_{(.)}\right\}^2$$

where :
- $\overset{F.d.}{\rightarrow}$ is the convergence of fini-dimensional distributions
- $Z_{(.)}$ is the centered Gaussian process with covariance function $\Gamma(t, t')$

We limit our attention to finite dimensional convergence since for the applications, the interval studied is always discretized, Wu and al. (2007).

2.3.3 Process under $H_{\lambda t^\star}$

The location of the QTL, $t^\star$ ($t^\star \in [t_1, t_2]$), will be added in the definition of H_1. So, the alternative hypothesis can be written :

$$H_{\lambda t^\star} : \text{“ } q = \frac{\lambda}{\sqrt{n}} \textit{ with } \lambda \in \mathbb{R}^\star \textit{ at the position } t^\star \text{ ”}$$

The QTL effect q is such as $q = a/\sqrt{n}$ in order to deal with Le Cam (1986)'s theory.
$\theta_{\lambda, t^\star}$ will be the parameter referring that we are under $H_{\lambda t^\star}$.
Under $H_{\lambda t^\star}$, as the QTL is located at position $t^\star$, the density of $Y/X_{t_1}X_{t_2}$ verifies :

$$p_{t^\star}\ f_{(\mu+q,\sigma)}(y) \ + \ (1 - p_{t^\star}) f_{(\mu+q,\sigma)}(y)$$

Up to know $\forall t \in [t_1, t_2]$:

$$\Lambda_t = (S_t)^2 \ + \ o_{P_{\theta_0}}(1)$$

So, according to Le Cam's first lemma, it comes :

$$\Lambda_t = (S_t)^2 \ + \ o_{P_{\theta_{\lambda,t^\star}}}(1)$$

We have :

$$S_t = S_t^0 \ + \ \sum_{j=1}^{n} \frac{\lambda}{\sigma n} \, X_{t^\star}^j \, h_j(t) \tag{2.4}$$

where

$$h_j(t) = \frac{2p_t^j - 1}{\sqrt{\mathbb{E}\left[(2p_t - 1)^2\right]}}$$

and S_t^0 is the score statistic under H_0.
By the law of large number :

$$\frac{1}{n} \sum_{j=1}^{n} X_{t^\star}^j \, h_j(t) \to \mathbb{E}\left[X_{t^\star} \, h(t)\right]$$

where

$$h(t) = \frac{2p_t - 1}{\sqrt{\mathbb{E}\left[(2p_t - 1)^2\right]}}$$

After some calculations, $\forall (t, t^\star) \in]t_1, t_2[^2$:

$$\begin{aligned} \mathbb{E}\left[X_{t^\star} \, (2p_t - 1)\right] = \bar{r}(t_1, t_2) \ & \left\{2Q_{t^\star}^{1,1} - 1\right\} \ \left\{2Q_t^{1,1} - 1\right\} \\ + \, r(t_1, t_2) \ & \left\{2Q_{t^\star}^{1,-1} - 1\right\} \ \left\{2Q_t^{1,-1} - 1\right\} \end{aligned} \tag{2.5}$$

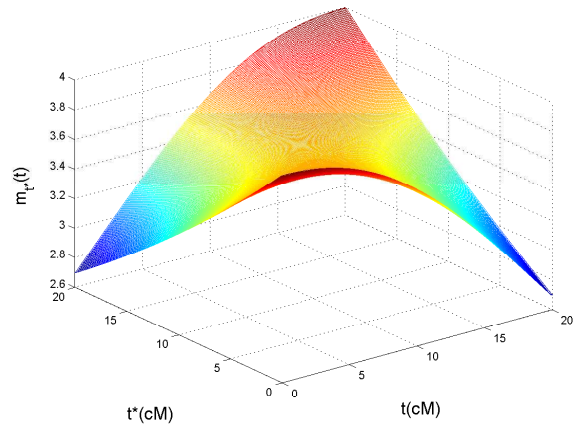
According to dominated convergence theorem, $\mathbb{E}\left[X_{t^\star} \, (2p_t - 1)\right]$ is continuous on $[t_1, t_2]^2$.
Let $m_{t^\star}(t)$ be the quantity such as $\forall (t, t^\star) \in [t_1, t_2]^2$:

$$m_{t^\star}(t) = \frac{\lambda \, \mathbb{E}\left[X_{t^\star} \, (2p_t - 1)\right]}{\sigma \, \sqrt{\mathbb{E}\left[(2p_t - 1)^2\right]}}$$

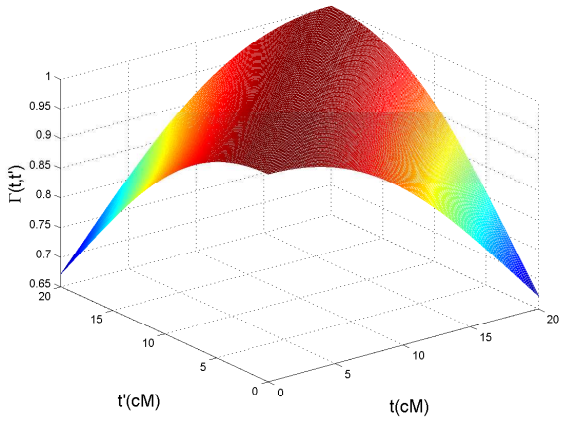
It comes, under $H_{\lambda t^\star}$:

$$\Lambda_{(.)} \overset{F.d.}{\to} \left\{Z_{(.)}\right\}^2$$

where $Z_{(.)}$ is the Gaussian process with covariance function $\Gamma(t, t')$ and expectation $m_{t^\star}(t)$.
Figure 2.1 represents the covariance function $\Gamma(t, t')$ and also the mean function $m_{t^\star}(t)$ ($T = 0.2$M).



Mean function



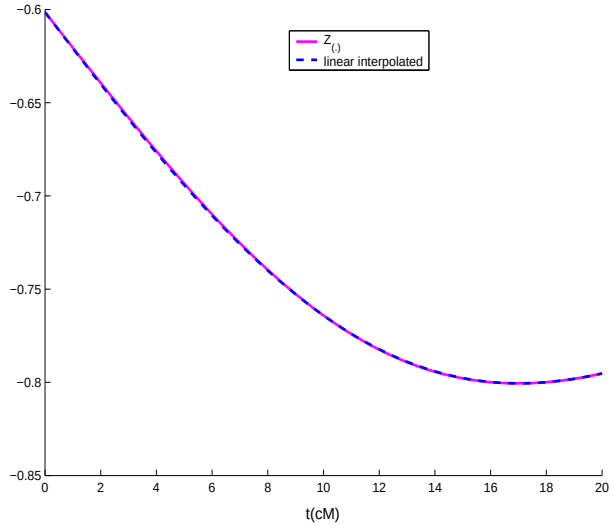
Covariance function

FIG. 2.1 – Mean function and Covariance function of the process $Z_{(.)}$ ($\lambda = 4$, $\sigma = 1$, $T = 0.2$M)

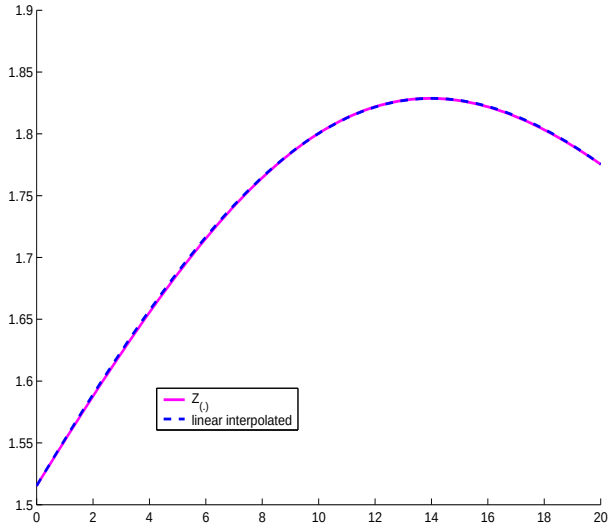
2.3.4 Addendum about the process

As the value of the test statistic between t_1 and t_2 depends only on the "genome information" at t_1 and t_2, we may wonder whether or not the process we are studying is the square (as we deal with LRT process) of an "interpolated process". In the article "Likelihood Ratio Test process for Quantitative Trait Loci detection" joined to this chapter, we proove that the LRT process, $\Lambda_{(.)}$, is asymptotically the square of a non linear interpolated process. We also proove that the square of the "linear interpolated process" renormalized is asymptotically a good approximation of $\Lambda_{(.)}$ provided that the genetic markers are close to each other. This new process is a generalization of the process studied by Rebaï and al. (1995).

In Figure 2.2, a path of the process $Z_{(.)}$ is represented under H_0 ($T = 0.2$M). The path corresponding to the "linear interpolated process" has been added. We can remark that the curves overlap due to the fact the genetic markers are close to each other. In the same way, the mean functions overlap ($t^\star = 14$cM, $\lambda = 2$). So, the paths of the two processes under the alternative are the same (cf Figure 2.3).
Figures 2.4 and 2.5 show the same study now with $T = 1$M. As expected, the different curves don't overlap because genetic markers are no so close to each other.



One path under H_0 for the two processes



Mean functions of the two processes

FIG. 2.2 – Process $Z_{(.)}$ and the linear interpolated process ($\lambda = 2$, $\sigma = 1$, $t^\star = 14$cM, $T = 0.2$M)

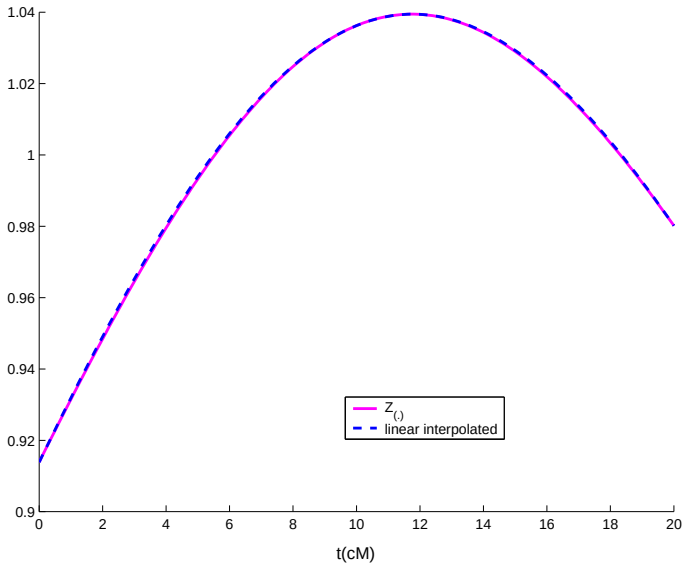
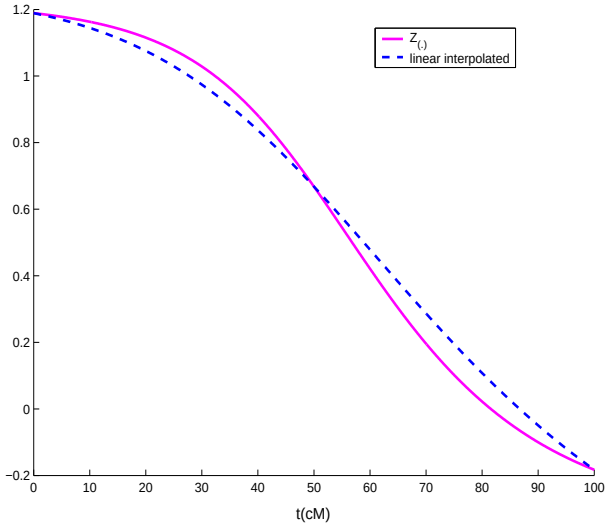
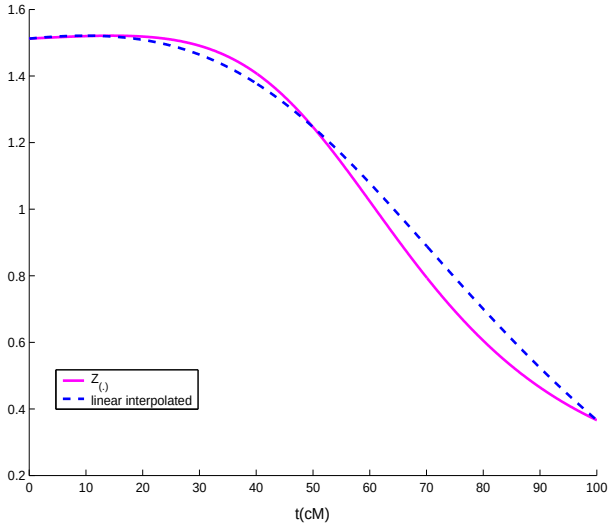


FIG. 2.3 – Same path as under H_0 but under $H_{\lambda t^\star}$ ($\lambda = 2$, $\sigma = 1$, $t^\star = 14$cM, $T = 0.2$M)



One path under H_0 for the two processes



Mean functions of the two processes

FIG. 2.4 – Process $Z_{(.)}$ and the linear interpolated process ($\lambda = 2$, $\sigma = 1$, $t^\star = 14$cM, $T = 1$M)

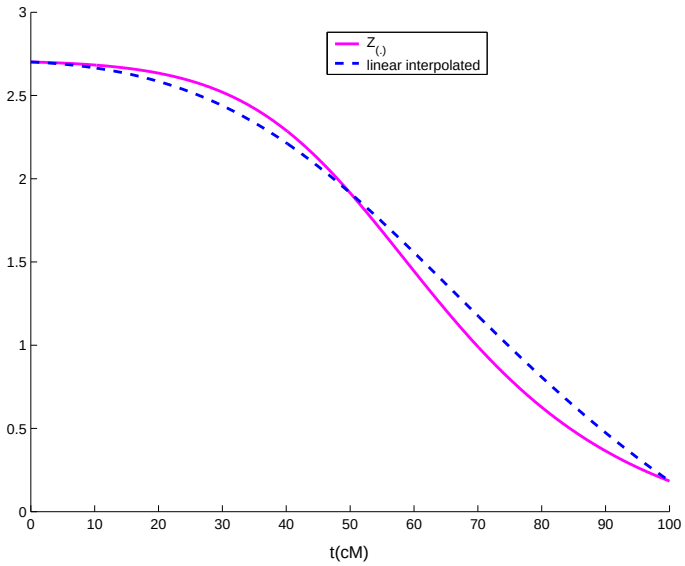


FIG. 2.5 – Same path as under H_0 but under $H_{\lambda t^\star}$ ($\lambda = 2$, $\sigma = 1$, $t^\star = 14$cM, $T = 1$M)

2.4 Several markers : the "Interval Mapping" of Lander and Botstein (1989)

There are now K genetic markers. As explained in Section 2.2, $0 = t_1 < t_2 < ... < t_K = T$ and the the QTL lies at $t^\star \in [t_1, t_K]$. We consider only values t, t' and $t^\star$ of the parameter that are distinct of the markers positions and the results will be prolonged by continuity at markers positions. We define $\mathbb{T}_k = \{t_1, ..., t_K\}$. t^ℓ and t^r are the quantities such 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 p_t the quantity equal to $P(X_t = 1/X_{t^\ell} \ X_{t^r})$. The weights p_t and the quantities $Q_t^{1,1}$, $Q_t^{1,-1}$, $Q_t^{-1,1}$, $Q_t^{-1,-1}$ keep the same expression as in formula (2.1) except that t_1 and t_2 have to be replaced respectively by t^ℓ and t^r.
At a position t, since the increments of a Poisson process are independent, the density of $Y/X_{t_1}X_{t_2}...X_{t_K}$ is the same as the density of $Y/X_{t^\ell}X_{t^r}$ that is to say :

$$p_t \ f_{(\mu+q,\sigma)}(y) \ + \ (1-p_t) \ f_{(\mu-q,\sigma)}(y)$$

As all the information concerning the quantitative trait Y is contained in the markers flanking the position tested, the likelihood can be written in the same way as in Section 2.3.1. So, the likelihood $L_n(\theta, t)$ at a position t for n observations j iid $(X^j_{t^\ell}, X^j_{t^r}, Y_j)$ verifies :

$$L_n(\theta, t) = \prod_{j=1}^{n} \left[p_t^j f_{(\mu+q,\sigma)}(y_j) \ + \ (1-p_t^j) f_{(\mu-q,\sigma)}(y_j)\right] \ g_j(t)$$

where $g_j(t)$ is the same function as in Section 2.3.1 except that t_1 and t_2 has to be replaced respectively by t^ℓ and t^r.

2.4.1 Process under H_0

Formula (2.3) is still suitable. It comes, under H_0 :

$$\Lambda_{(.)} \overset{F.d.}{\to} \left\{Z_{(.)}\right\}^2$$

where $Z_{(.)}$ is the centered Gaussian process with covariance function $\Gamma(t, t')$.
The function $\Gamma(t, t')$ is given in appendix 2.6.3.

2.4.2 Process under $H_{\lambda t^\star}$

In the same way as what has been done in Section 2.3.3, using Le Cam's results, under $H_{\lambda t^\star}$:

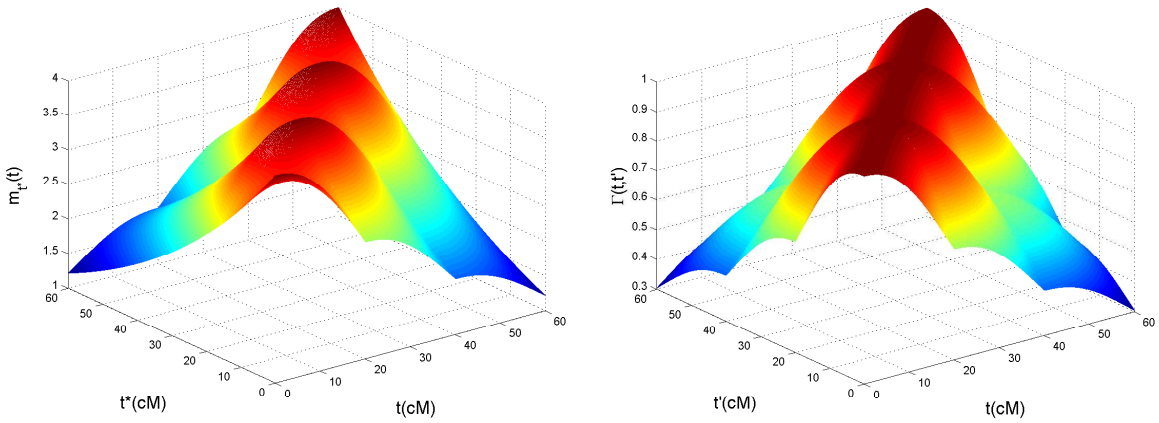
$$\Lambda_{(.)} \overset{F.d.}{\to} \left\{Z_{(.)}\right\}^2$$

where $Z_{(.)}$ is the Gaussian process with covariance function $\Gamma(t,t')$ and expectation $m_{t^\star}(t)$. As previously :

$$m_{t^\star}(t) = \frac{\lambda\,\mathbb{E}\left[X_{t^\star}\,(2p_t-1)\right]}{\sigma\,\sqrt{\mathbb{E}\left[(2p_t-1)^2\right]}}$$

After some calculations :

$$\begin{aligned}\mathbb{E}\left[X_{t^\star}\,(2p_t-1)\right] = \big\{&\bar{r}(t^\ell,t^r)\left(2Q_{t^\star}^{1,1}-1\right)\left(2Q_t^{1,1}-1\right)\\ &+r(t^\ell,t^r)\left(2Q_{t^\star}^{1,-1}-1\right)\left(2Q_t^{1,-1}-1\right)\big\}\;1_{t^\star\in]t^\ell,t^r[}\\ &+e^{-2|t-t^\star|}\,1_{t^\star\notin]t^\ell,t^r[}\end{aligned} \tag{2.6}$$



Mean function Covariance function

FIG. 2.6 – Mean function and Covariance function of the process $Z_{(.)}$ ($\lambda = 4$, $\sigma = 1$, $T = 0.6$M, 4 markers equally spaced every 0.2M)

Figure 2.6 represents the covariance function $\Gamma(t, t')$ and also the mean function $m_{t^\star}(t)$ ($T = 0.6$M, 4 markers equally spaced every 0.2M).

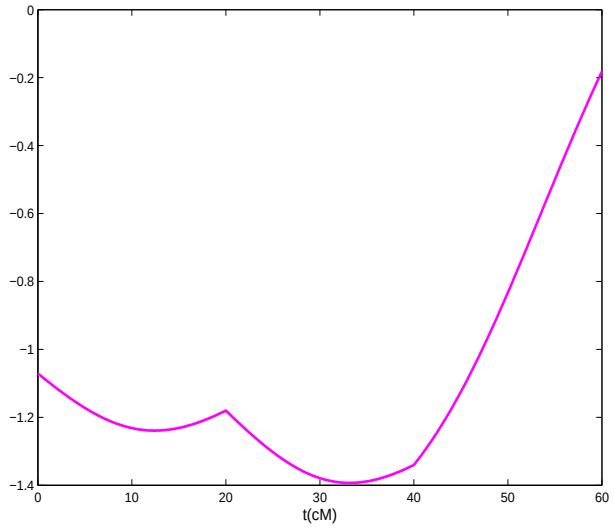
Note that when $(t, t') \in \mathbb{T}_k \times \mathbb{T}_k$, $\Gamma(t, t') = e^{-2|t-t'|}$. It is relative to an Ornstein-Uhlenbeck process, as studied in Lander and Botstein (1989), and Cierco (1998).

2.4.3 Addendum about the process

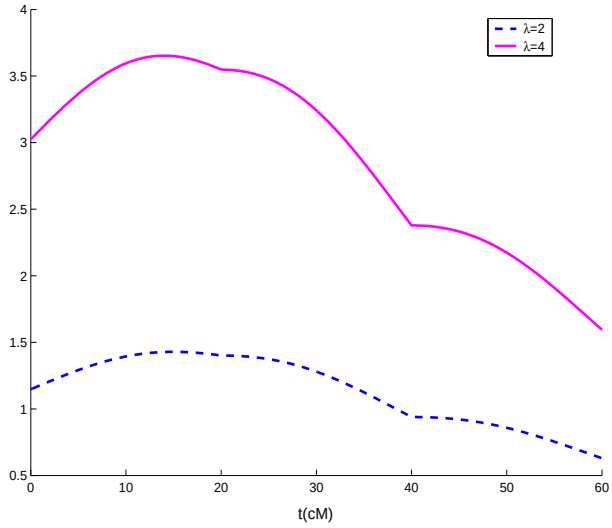
The processes obtained by interpolation, presented in Section 2.3.4, can easily be generalized to the case of several markers. The "linear interpolated process" is still a good approximation if markers are close to each other. This process is a generalization of the process studied by Rebaï and al. (1994).

In Figure 2.7, a path of the process $Z_{(.)}$ is represented under H_0. The map consists of 4 genetic markers equally spaced every 0.2M ($T = 0.6$M). On the right hand-side, is represented the mean function of the process when $\lambda = 2$ and $\lambda = 4$. The QTL is located at $t^\star = 14$cM.

In Figure 2.8, the focus is on the same path of the process $Z_{(.)}$ as under H_0 but under $H_{\lambda t^\star}$. It is noticeable, that when $\lambda = 4$, the signal-to-noise ratio is large enough in order that the supremum of the path is obtained in the area of $t^\star = 14$cM. When $\lambda = 2$, the signal-to-noise ratio is too small.



One path under H_0 for the process $Z_{(.)}$



Mean functions of the process $Z_{(.)}$ as a function of λ

FIG. 2.7 – Process $Z_{(.)}$ ($\sigma = 1$, $t^\star = 14$cM, $T = 0.6$M, 4 markers equally spaced every 0.2M)

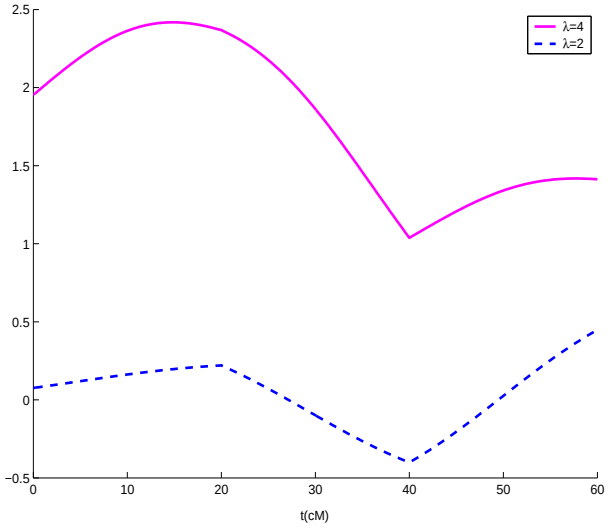


FIG. 2.8 – Same path as under H_0 but under $H_{\lambda t^\star}$ ($\sigma = 1$, $t^\star = 14$cM, $T = 0.6$M, 4 markers equally spaced every 0.2M)

2.5 Generalization to families of sires with their own informative markers

In the previous sections, we were looking for a QTL lying on the interval $[0, T]$ using the concept of Interval Mapping. One population of progenies of a sire has been studied that is to say one family. In order to increase the power of the method, geneticists look for the QTL not in one family but simultanously in several families, each defined by a different sire. It increases the chances to study families whose sires are heterozygote at the QTL. In that case, the Interval Mapping method is also used : LRT are peformed at each position $t \in [0, T]$ and the supremum of these statistics is used as a unique test statistic. Naturally, the putative QTL is supposed to be lying at the same location in each family otherwise the concept of Interval Mapping has no sense.

Let's extend the model presented in Section 2.2 to the case of several families. Let $I \in \mathbb{N}^\star$ be the number of families. Let C be a discrete random variable refering to the family. We note π_i the quantity such as $\pi_i = P(C = i)$. In other words, the individual belongs to family i with probability π_i.

When we deal with different families, the location and the number of informative markers differ in each family. So, K^i will refer to the number of informative markers in family i and $t^i_1, ..., t^i_{K^i}$ will be their locations. We suppose $t^i_1 < t^i_2 < ... < t^i_{K^i}$ and to begin, $\forall i$, $t^i_1 = 0$ and $t^i_{K^i} = T$, that is to say there is one marker at each extremity of the chromosome.

The process $X_{(.)}$ is unchanged. Howewer, the quantitative trait verifies now :

$$(Y/C = i) \ = \mu_i + X_{t^\star} \, q_i + \varepsilon$$

where μ_i and q_i are respectively a polygenic effect and the QTL effect inside family i. ε is a Gaussian white noise.
As the locations of the informative markers differ in each family, some adjustments have to be done. As in Section 2.4, we consider only values t, t' or $t^\star$ of the parameter that are distinct of the marker positions. The result will be prolonged by continuity at the markers positions. We define $\mathbb{T}^I_k = \left\{t^1_1, ..., t^1_{K^1}, ..., t^I_1, ..., t^I_{K^I}\right\}$. $t^{\ell,i}$ and $t^{r,i}$ are the quantities such as :

$$t^{\ell,i} = sup\left\{t^i_k \in \mathbb{T}^I_k : t^i_k < t\right\} \quad , \quad t^{r,i} = inf\left\{t^i_k \in \mathbb{T}^I_k : t < t^i_k\right\}$$

Let $p_{t,i}$ the quantity equal to $\mathbb{P}(X_t = 1/X_{t^{\ell,i}} X_{t^{r,i}})$. It verifies :

$$\begin{aligned} p_{t,i} = Q^{1,1}_{t,i} \, 1_{X_{t^{\ell,i}}=1} 1_{X_{t^{r,i}}=1} \ + \ Q^{1,-1}_{t,i} \, 1_{X_{t^{\ell,i}}=1} 1_{X_{t^{r,i}}=-1} \\ + Q^{-1,1}_{t,i} \, 1_{X_{t^{\ell,i}}=-1} 1_{X_{t^{r,i}}=1} \ + \ Q^{-1,-1}_{t,i} \, 1_{X_{t^{\ell,i}}=-1} 1_{X_{t^{r,i}}=-1} \end{aligned}$$

where :

$$Q_{t,i}^{1,1} = \frac{\bar{r}(t^{\ell,i},t)\,\bar{r}(t,t^{r,i})}{\bar{r}(t^{\ell,i},t^{r,i})} \;,\; Q_{t,i}^{1,-1} = \frac{\bar{r}(t^{\ell,i},t)\,r(t,t^{r,i})}{r(t^{\ell,i},t^{r,i})}$$
$$Q_{t,i}^{-1,1} = \frac{r(t^{\ell,i},t)\,\bar{r}(t,t^{r,i})}{r(t^{\ell,i},t^{r,i})} \;,\; Q_{t,i}^{-1,-1} = \frac{r(t^{\ell,i},t)\,r(t,t^{r,i})}{\bar{r}(t^{\ell,i},t^{r,i})}$$

We can remark that we have :

$$Q_{t,i}^{-1,-1} = 1 - Q_{t,i}^{1,1} \quad \text{and} \quad Q_{t,i}^{-1,1} = 1 - Q_{t,i}^{1,-1}$$

The density of $Y/C = i\; X_{t^{\ell,i}} X_{t^{r,i}}$ verifies at a position t :

$$p_{t,i}\; f_{(\mu_i+q_i,\sigma)}(y) \;+\; (1-p_{t,i})\; f_{(\mu_i+q_i,\sigma)}(y)$$

Let $\theta = (q_1,\; ...,\; q_I,\; \mu_1,\; ...,\; \mu_I,\; \sigma)$ be the parameter of the model at t fixed and θ_0 the true value of the parameter under $H_0 : \theta_0 = (0,\; ...,\; 0,\; \mu_1,\; ...,\; \mu_I,\; \sigma)$.
M_{t^ℓ} and M_{t^r} will be the random variables such as $M_{t^\ell} = \sum_{i=1}^{I} X_{t^{\ell,i}} 1_{C=i}$ and $M_{t^r} = \sum_{i=1}^{I} X_{t^{r,i}} 1_{C=i}$.
The likelihood $L_n(\theta,t)$ under the alternative, at a position t, and for n observations j iid $\left(i(j),\; M_{t^\ell}^j,\; M_{t^r}^j,\; Y_j\right)$ verifies :

$$L_n(\theta,t) = \prod_{j=1}^{n} \sum_{i=1}^{I} \left[p_{t,i}^j f_{(\mu_i+q_i,\sigma)}(y_j) \;+\; (1-p_{t,i}^j) f_{(\mu_i+q_i,\sigma)}(y_j)\right]\; 1_{C_j=i}\; g_{j,i}(t)$$

where

$$g_{j,i}(t) = \frac{\pi_i}{2}\left\{\bar{r}(t^{\ell,i},t^{r,i}) 1_{X_{t^{\ell,i}}^j=1} 1_{X_{t^{r,i}}^j=1} + r(t^{\ell,i},t^{r,i}) 1_{X_{t^{\ell,i}}^j=1} 1_{X_{t^{r,i}}^j=-1}\right\}$$
$$+ \frac{\pi_i}{2}\left\{r(t^{\ell,i},t^{r,i}) 1_{X_{t^{\ell,i}}^j=-1} 1_{X_{t^{r,i}}^j=1} + \bar{r}(t^{\ell,i},t^{r,i}) 1_{X_{t^{\ell,i}}^j=-1} 1_{X_{t^{r,i}}^j=-1}\right\}$$

Let $\hat{\theta}$ be the MLE of θ and $\tilde{\theta}$ the MLE under H_0 :

$$\hat{\theta} = (\hat{q}_1,\; ...,\; \hat{q}_I,\; \hat{\mu}_1, ...,\; \hat{\mu}_I,\; \hat{\sigma})$$
$$\tilde{\theta} = (0,\; ...,\; 0, ...,\; \tilde{\mu}_1, ...,\; \tilde{\mu}_I,\; \tilde{\sigma})$$

A likelihood ratio test will be performed at each position t. In the same way as what has been done in Section 2.3.1 :

$$\frac{\partial log L_1}{\partial q_i}\,|_{\theta_0} = \frac{y-\mu_i}{\sigma^2}\;(2p_{t,i}-1)\;1_{C=i} \quad , \qquad \frac{\partial log L_1}{\partial \mu_i}\,|_{\theta_0} = \frac{y-\mu_i}{\sigma^2}\,1_{C=i}$$

$$\frac{\partial logL_1}{\partial \sigma} \mid_{\theta_0} = -\frac{1}{\sigma} + \sum_{i=1}^{I} \frac{(y-\mu_i)^2}{\sigma^3} 1_{C=i}$$

$$I_{\theta_0} = Diag\left(\frac{\pi_1}{\sigma^2}\ \mathbb{E}\left[(2p_{t,1}-1)^2\right], ..., \frac{\pi_I}{\sigma^2}\ \mathbb{E}\left[(2p_{t,I}-1)^2\right], \frac{\pi_1}{\sigma^2}, ..., \frac{\pi_I}{\sigma^2}, \frac{2}{\sigma^2}\right)$$

The formula for $\mathbb{E}\left[(2p_{t,i}-1)^2\right]$ is given in appendix 2.6.1 (the formula for which $t \in]t_1, t_2[$). Besides, t_1 (resp t_2) has to be replaced by $t^{\ell,i}$ (resp $t^{r,i}$) and the Q_t's by the $Q_{t,i}$'s.

Remarque 2.3 *If the π_i's were unknown, the Fisher information matrix would still be diagonal, it would be the same as above, and the diagonal terms concerning π_i would be equal to $\frac{1}{\pi_i}$. So, the results established further are also true for all the π_i's unknown.*

As the model is regular, it comes :

$$\Lambda_t = \sum_{i=1}^{I} \left(\sum_{j=1}^{n} \frac{(y_j-\mu_i)\,(2p_{t,i}^j-1)}{\sqrt{n\,\pi_i}\,\sigma\,\sqrt{\mathbb{E}\left[(2p_{t,i}-1)^2\right]}} 1_{C_j=i} \right)^2 + o_{P_{\theta_0}}(1) \tag{2.7}$$

2.5.1 Process under H_0

Let $S_{t,i}$ be the score statistic :

$$S_{t,i} = \sum_{j=1}^{n} \frac{(y_j-\mu_i)\,(2p_{t,i}^j-1)}{\sqrt{n\,\pi_i}\,\sigma\,\sqrt{\mathbb{E}\left[(2p_{t,i}-1)^2\right]}} 1_{C_j=i} \tag{2.8}$$

By the central limit theorem, $S_{t,i} \to N(0,1)$.
According to formula (2.7) :

$$\Lambda_t = \sum_{i=1}^{I} (S_{t,i})^2 + o_{P_{\theta_0}}(1)$$

It comes, under H_0 :

$$\Lambda_{(.)} \overset{F.d.}{\to} \sum_{i=1}^{I} \left\{Z_{(.)}^i\right\}^2$$

where the $Z_{(.)}^i$ are independent centered Gaussian processes with covariance function $\Gamma_i(t,t')$. The function $\Gamma_i(t,t')$ is given in appendix 2.6.4.

2.5.2 Process under $H_{\lambda t^\star}$

The alternative hypothesis can be written :

$$H_{\lambda t^\star} : \text{“there is at least one } q_i = \lambda_i/\sqrt{n}, \text{ with } \lambda_i \in \mathbb{R}^\star, \text{ at the position } t^\star \text{ "}$$

$\theta_{\lambda,t^\star}$ will be the parameter refering that we are under $H_{\lambda t^\star}$.
Under $H_{\lambda t^\star}$, as the QTL is located at position $t^\star$, the density of $Y/C = i\ X_{t^{\ell},i} X_{t^{r},i}$ verifies :

$$p_{t^\star,i}\ f_{(\mu_i+q_i,\sigma)}(y)\ +\ (1-p_{t^\star,i})\ f_{(\mu_i-q_i,\sigma)}(y)$$

Up to know :

$$\Lambda_t = \sum_{i=1}^{I} (S_{t,i})^2\ +\ o_{P_{\theta_0}}(1)$$

So, according to Le Cam's first lemma, it comes :

$$\Lambda_t = \sum_{i=1}^{I} (S_{t,i})^2\ +\ o_{P_{\theta_{\lambda,t^\star}}}(1)$$

We have :

$$S_{t,i} = S^0_{t,i}\ +\ \sum_{j=1}^{n} \frac{\lambda_i}{n\ \sigma\ \sqrt{\pi_i}}\ 1_{C_j=i}\ X^j_{t^\star}\ h_{j,i}(t)$$

where

$$h_{j,i}(t) = \frac{2p^j_{t,i} - 1}{\sqrt{\mathbb{E}\left[(2p_{t,i}-1)^2\right]}}$$

and $S^0_{t,i}$ is the score statistic under H_0.
By the law of large number :

$$\frac{1}{n}\sum_{j=1}^{n} X^j_{t^\star}\ h_{j,i}(t)\ 1_{C_j=i} \to \pi_i\ \mathbb{E}\left[X_{t^\star}\ \tilde{h}_i(t)\right]$$

where

$$\tilde{h}_i(t) = \frac{2p_{t,i} - 1}{\sqrt{\mathbb{E}\left[(2p_{t,i}-1)^2\right]}}$$

According to formula (2.6) of Section 2.4.2 :

$$\begin{aligned}\mathbb{E}\left[X_{t^\star}\ (2p_{t,i}-1)\right] = &\left\{\bar{r}(t^{\ell,i},t^{r,i})\left(2Q^{1,1}_{t^\star,i}-1\right)\left(2Q^{1,1}_{t,i}-1\right)\right.\\ &\left.+r(t^{\ell,i},t^{r,i})\left(2Q^{1,-1}_{t^\star,i}-1\right)\left(2Q^{1,-1}_{t,i}-1\right)\right\}\ 1_{t^\star\in]t^{\ell,i},t^{r,i}[}\\ &+\ e^{-2|t-t^\star|}\ 1_{t^\star\notin]t^{\ell,i},t^{r,i}[}\end{aligned}$$

Let $m_{t^\star,i}(t)$ be the quantity such as :

$$m_{t^\star,i}(t) = \frac{\lambda_i \sqrt{\pi_i}\ \mathbb{E}\left[X_{t^\star}\left(2p_{t,i}-1\right)\right]}{\sigma\sqrt{\mathbb{E}\left[(2p_{t,i}-1)^2\right]}}$$

It comes, under $H_{\lambda t^\star}$:

$$\Lambda_{(.)} \overset{F.d.}{\rightarrow} \sum_{i=1}^{I}\left\{Z^i_{(.)}\right\}^2$$

where the $Z^i_{(.)}$ are independent Gaussian processes with covariance function $\Gamma_i(t,t')$ and expectation $m_{t^\star,i}(t)$.

2.5.3 Addendum about the processes

As previously, the processes $Z^i_{(.)}$ are interpolated processes. Indeed, under H_0 and $H_{\lambda t^\star}$:

$$Z^i_t = \frac{\alpha_{t,i}\, Z^i_{t^{\ell,i}} \ + \ \beta_{t,i}\, Z^i_{t^{r,i}}}{\sqrt{\mathbb{E}\left[(2p_{t,i}-1)^2\right]}} \text{ where } \alpha_{t,i} = Q^{1,1}_{t,i} + Q^{1,-1}_{t,i} - 1 \ \ , \ \ \beta_{t,i} = Q^{1,1}_{t,i} - Q^{1,-1}_{t,i}$$
$$\text{and } \mathrm{Cov}(Z^i_{t^{\ell,i}}\ ,\ Z^i_{t^{r,i}}) = e^{-2(t^{r,i}-t^{\ell,i})}$$

In the same way, the mean function, $m_{t^\star,i}(t)$, of the process $Z^i_{(.)}$ under $H_{\lambda t^\star}$ is such as :

$$m_{t^\star,i}(t) = \frac{\alpha_{t,i}\, m_{t^\star,i}(t^{\ell,i}) \ + \ \beta_{t,i}\, m_{t^\star,i}(t^{r,i})}{\sqrt{\mathbb{E}\left[(2p_{t,i}-1)^2\right]}}$$

2.5.4 Relaxing some hypotheses

At the beginning of the Section 2.5, we have supposed that there was one informative marker at each extremity of the chromosome whatever the family is. In fact, this is never the case.

So, let suppose now that $\forall i$, $t^i_1 \geqslant 0$ and $t^i_{K^i} \leqslant T$ (these quantities depend on i now). As previously, only values t, t' or $t^\star$ of the parameter that are distinct of the markers positions will be considered. The result will be prolonged by continuity at marker positions.

Process under H_0 :

We have :

$$\Lambda_{(.)} \overset{F.d.}{\rightarrow} \sum_{i=1}^{I}\left\{Z^i_{(.)}\right\}^2$$

where the $Z^i_{(.)}$ are independent centered Gaussian processes such as :

$$\forall t \in [t^i_1, t^i_{K^i}] \backslash \mathbb{T}^I_k \quad Z^i_t = \frac{\alpha_{t,i}\, Z^i_{t^{\ell,i}} \;+\; \beta_{t,i}\, Z^i_{t^{r,i}}}{\sqrt{\mathbb{E}\left[(2p_{t,i}-1)^2\right]}} \text{ with } \mathrm{Cov}(Z^i_{t^{\ell,i}}\,,\; Z^i_{t^{r,i}}) = e^{-2(t^{r,i}-t^{\ell,i})}$$

and $Z^i_{(.)}$ is constant on $[0, t^i_1] \backslash \mathbb{T}^I_k$ and on $[t^i_{K^i}, T] \backslash \mathbb{T}^I_k$.

Process under $H_{\lambda t^\star}$:

We have :

$$\Lambda_{(.)} \overset{F.d.}{\rightarrow} \sum_{i=1}^{I} \left\{ Z^i_{(.)} \right\}^2$$

where the $Z^i_{(.)}$ are the same processes as above on which a mean function $m_{t^\star,i}(t)$ has been added.

The mean function is such as :

$$\forall (t, t^\star) \in [t^i_1, t^i_{K^i}] \backslash \mathbb{T}^I_k \times [t^i_1, t^i_{K^i}] \backslash \mathbb{T}^I_k \;,\; m_{t^\star,i}(t) = \frac{\alpha_{t,i}\, m_{t^\star,i}(t^{\ell,i}) \;+\; \beta_{t,i}\, m_{t^\star,i}(t^{r,i})}{\sqrt{\mathbb{E}\left[(2p_{t,i}-1)^2\right]}}$$

$m_{t^\star,i}\,(t)$ is equal to $m_{t^\star,i}(t^i_1)$ when $t \in [0, t^i_1] \backslash \mathbb{T}^I_k$

$m_{t^\star,i}\,(t)$ is equal to $m_{t^\star,i}(t^i_{K^i})$ when $t \in [t^i_{K^i}, T] \backslash \mathbb{T}^I_k$

$$\forall (t, t^\star) \in [t^i_1, t^i_{K^i}] \backslash \mathbb{T}^I_k \;\times\; [0, t^i_1] \backslash \mathbb{T}^I_k \bigcup [t^i_{K^i}, T] \backslash \mathbb{T}^I_k \;,\; m_{t^\star,i}\,(t) = \frac{e^{-2|t-t^\star|} \lambda_i \sqrt{\pi_i}}{\sigma}$$

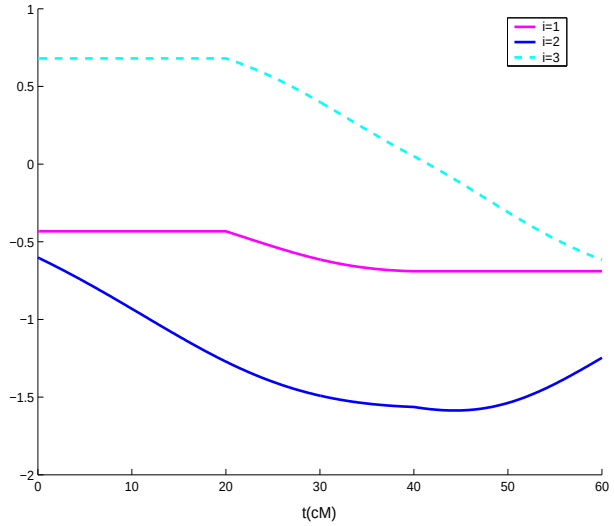
Illustration :

In Figure 2.9, we consider the same map as previously in Figure 2.8, that is to say 4 genetic markers equally spaced every 20cM ($T = 60$cM). $I = 3$ has been choosen.

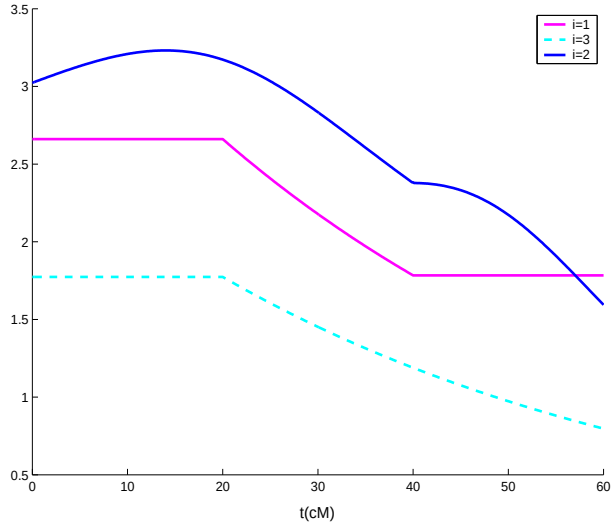
However, since the genetic markers are not always informative, we have considered that for family 1, only the second and the third marker of the map were informative. So, $t^1_1 = 20$cM, $t^1_2 = 40$cM. In the same way, we have choosen $t^2_1 = 0$cM, $t^2_2 = 40$cM, $t^2_3 = 60$cM, $t^3_1 = 20$cM, $t^3_2 = 40$cM, $t^3_3 = 60$cM.

One path for each of the three processes $Z^1_{(.)}$, $Z^2_{(.)}$, $Z^3_{(.)}$, under H_0 has been simulated. Due to the uninformativity, $Z^1_{(.)}$ is constant on $[0\text{cM}, 20\text{cM}]$ and on $[40\text{cM}, 60\text{cM}]$, $Z^3_{(.)}$ is constant on $[0\text{cM}, 20\text{cM}]$. Same remark for the mean functions of these two processes ($t^\star = 14$cM). In family 3, as the first and the second informative genetic marker are respectively located at positions 0cM and 40cM, the process $Z^3_{(.)}$ and its mean function are an interpolation between these two locations.

In Figure 2.10, is represented the corresponding path of the process $\sum_{i=1}^{3} \left\{ Z^i_{(.)} \right\}^2$ under H_0 and under $H_{\lambda t^\star}$.

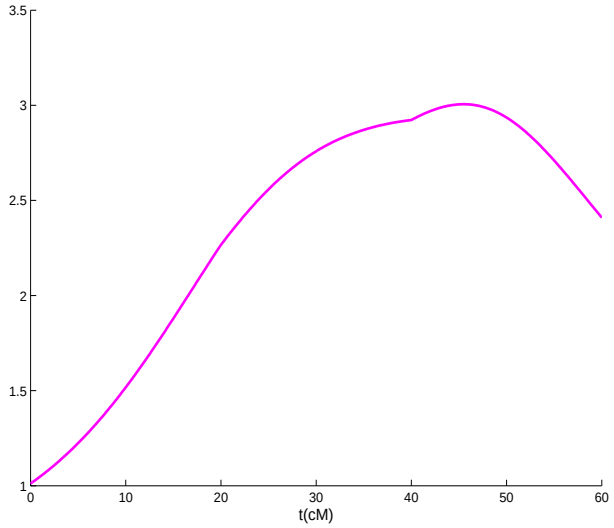


One path under H_0 for the processes $Z^i_{(.)}$

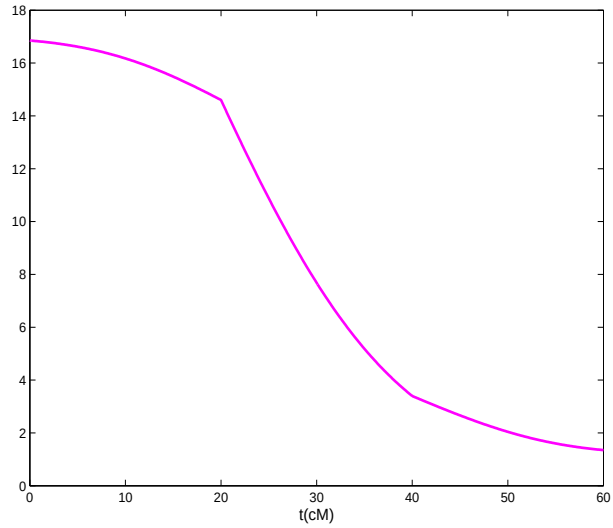


Mean functions of the processes $Z^i_{(.)}$

FIG. 2.9 – Processes $Z^i_{(.)}$ ($I = 3$, $\sigma = 1$, $t^\star = 14$cM, $T = 0.6$M, $\lambda_1\sqrt{\pi_1} = 3$, $\lambda_2\sqrt{\pi_2} = 2$, $\lambda_3\sqrt{\pi_3} = 4$)



Under H_0



Under $H_{\lambda t^\star}$

FIG. 2.10 – path of the process $\sum_{i=1}^{3}\left\{Z^i_{(.)}\right\}^2$ ($\sigma = 1$, $t^\star = 14$cM, $T = 0.6$M, $\lambda_1\sqrt{\pi_1} = 3$, $\lambda_2\sqrt{\pi_2} = 2$, $\lambda_3\sqrt{\pi_3} = 4$)

2.6 Appendix

2.6.1 Formula for $\mathbb{E}\left[(2p_t-1)^2\right]$

$\mathbb{E}\left[(2p_{t_1}-1)^2\right] = \mathbb{E}\left[(2p_{t_2}-1)^2\right] = 1$ and $\forall t \in]t_1, t_2[$:

$$\mathbb{E}\left[(2p_t-1)^2\right] = \bar{r}(t_1,t_2)\left(2Q_t^{1,1}-1\right)^2 + r(t_1,t_2)\left(2Q_t^{1,-1}-1\right)^2$$

2.6.2 Covariances of the process (only 2 markers)

$\forall (t,t') \in]t_1, t_2[^2$:

$$\Gamma(t,t') = \frac{4\mathbb{E}\left[p_t p_{t'}\right] - 1}{\sqrt{\mathbb{E}\left[(2p_t-1)^2\right]}\sqrt{\mathbb{E}\left[(2p_{t'}-1)^2\right]}}$$

$$\mathbb{E}\left[p_t p_{t'}\right]$$
$$= \frac{1}{2}\left\{Q_t^{1,1}Q_{t'}^{1,1}\,\bar{r}(t_1,t_2) + Q_t^{1,-1}Q_{t'}^{1,-1}\,r(t_1,t_2) + Q_t^{-1,1}Q_{t'}^{-1,1}\,r(t_1,t_2) + Q_t^{-1,-1}Q_{t'}^{-1,-1}\,\bar{r}(t_1,t_2)\right\}$$

2.6.3 Covariance of the process (Interval Mapping)

$\forall (t,t') \in [t_1, t_K]\backslash\mathbb{T}_k \times [t_1, t_K]\backslash\mathbb{T}_k$:

$$\Gamma(t,t') = \frac{4\mathbb{E}\left[p_t p_{t'}\right] - 1}{\sqrt{\mathbb{E}\left[(2p_t-1)^2\right]}\sqrt{\mathbb{E}\left[(2p_{t'}-1)^2\right]}}$$

$\mathbb{E}\left[(2p_t-1)^2\right]$ has already been calculated $\forall t \in]t_1, t_2[$ in appendix 2.6.1. This formula is also suitable $\forall t \in [t_1, t_K]\backslash\mathbb{T}_k$.

Besides, according to appendix 2.6.2, $\forall (t,t') \in]t^\ell, t^r[^2$:

$$\mathbb{E}\left[p_t p_{t'}\right]$$
$$= \frac{1}{2}\left\{Q_t^{1,1}Q_{t'}^{1,1}\,\bar{r}(t^\ell,t^r) + Q_t^{1,-1}Q_{t'}^{1,-1}\,r(t^\ell,t^r) + Q_t^{-1,1}Q_{t'}^{-1,1}\,r(t^\ell,t^r) + Q_t^{-1,-1}Q_{t'}^{-1,-1}\,\bar{r}(t^\ell,t^r)\right\}$$

Besides, if $(t,t') \in]t^\ell, t^r[\times[t^r, t_K]\backslash\mathbb{T}_k$:

$$\mathbb{E}\left[p_t p_{t'}\right]$$
$$= \frac{1}{2}\bar{r}(t^\ell,t^r)\left[Q_{t'}^{1,1}\bar{r}\left\{(t')^\ell,(t')^r\right\} + Q_{t'}^{1,-1}r\left\{(t')^\ell,(t')^r\right\}\right]\left[Q_t^{1,1}\bar{r}\left\{t^r,(t')^\ell\right\} + Q_t^{-1,-1}r\left\{t^r,(t')^\ell\right\}\right]$$
$$+ \frac{1}{2}\bar{r}(t^\ell,t^r)\left[Q_{t'}^{-1,1}r\left\{(t')^\ell,(t')^r\right\} + Q_{t'}^{-1,-1}\bar{r}\left\{(t')^\ell,(t')^r\right\}\right]\left[Q_t^{1,1}r\left\{t^r,(t')^\ell\right\} + Q_t^{-1,-1}\bar{r}\left\{t^r,(t')^\ell\right\}\right]$$
$$+ \frac{1}{2}r(t^\ell,t^r)\left[Q_{t'}^{1,1}\bar{r}\left\{(t')^\ell,(t')^r\right\} + Q_{t'}^{1,-1}\,r\left\{(t')^\ell,(t')^r\right\}\right]\left[Q_t^{1,-1}r\left\{t^r,(t')^\ell\right\} + Q_t^{-1,1}\bar{r}\left\{t^r,(t')^\ell\right\}\right]$$
$$+ \frac{1}{2}r(t^\ell,t^r)\left[Q_{t'}^{-1,1}r\left\{(t')^\ell,(t')^r\right\} + Q_{t'}^{-1,-1}\bar{r}\left\{(t')^\ell,(t')^r\right\}\right]\left[Q_t^{1,-1}\bar{r}\left\{t^r,(t')^\ell\right\} + Q_t^{-1,1}r\left\{t^r,(t')^\ell\right\}\right]$$

2.6.4 Covariance of the process (I families of sires)

$\forall (t,t') \in [0,T]\backslash\mathbb{T}_k^I \times [0,T]\backslash\mathbb{T}_k^I$

$$\Gamma_i(t,t') = \frac{4\mathbb{E}\left[p_{t,i}\, p_{t',i}\right] - 1}{\sqrt{\mathbb{E}\left[(2p_{t,i}-1)^2\right]}\sqrt{\mathbb{E}\left[(2p_{t',i}-1)^2\right]}}$$

$\forall (t,t') \in]t^{\ell,i}, t^{r,i}[^2$, $\mathbb{E}\left[p_{t,i}\, p_{t',i}\right]$ has the same expression as in appendix 2.6.3. We just have to replace t^ℓ (resp t^r) by $t^{\ell,i}$ (resp $t^{r,i}$) and the Q_t's (resp. the $Q_{t'}$'s) by $Q_{t,i}$ (resp. $Q_{t',i}$). If $t \in]t^{\ell,i}, t^{r,i}[\times[t^{r,i},T]\backslash\mathbb{T}_k^I$, then $\mathbb{E}\left[p_{t,i}\, p_{t',i}\right]$ has the same expression as in appendix 2.6.3. We just have to proceed to the same adjustments in notations as below. And naturally, $(t')^\ell$ (resp. $(t')^r$) becomes $(t')^{\ell,i}$ (resp. $(t')^{r,i}$).

2.6.5 Proof of "Relaxing some hypotheses"

Let $i \in \{1,...,I\}$. If $t \in [0,t_1^i]\backslash\mathbb{T}_k^I$, then as there is only one flanking marker : $p_{t,i} = \bar{r}(t,t_1^i)\, 1_{X_{t_1^i}=1} \,+\, r(t,t_1^i)\, 1_{X_{t_1^i}=-1}$.
According to formula (2.8) page 120 and by continuity ($t \searrow t_1^i$) :

$$S_{t_1^i,i} = \sum_{j=1}^{n} \frac{(y_j - \mu_i)\,(2\, 1_{X_{t_1^i}^j=1} - 1)}{\sqrt{n\,\pi_i}\,\sigma}\, 1_{C_j=i}$$

Besides, formula (2.8) is also suitable $\forall t \in [0,t_1^i]\backslash\mathbb{T}_k^I$.
If $t \in [0,t_1^i]\backslash\mathbb{T}_k^I$, then $\sqrt{\mathbb{E}\left[(2p_{t,i}-1)^2\right]} = 1 - 2r(t,t_1^i)$ because the recombination rate is between 0 and $\frac{1}{2}$. $1-2r(t,t_1^i)$ will never be equal to zero since t is bounded. As $1_{X_{t_1^i}=-1} = 1 - 1_{X_{t_1^i}=1}$, $2p_{t,i} - 1 = [1-2r(t,t_1^i)][2\, 1_{X_{t_1^i}=1} - 1]$.
It comes :

$$\forall t \in [0,t_1^i]\backslash\mathbb{T}_k^I, \quad S_{t,i} = S_{t_1^i,i}$$

And naturally, by symetry :

$$\forall t \in [t_{K^i}^i, T]\backslash\mathbb{T}_k^I, \quad S_{t,i} = S_{t_{K^i}^i,i}$$

The rest of the proof is straightforward. We just have to use Le Cam's first lemma to obtain the asymptotic process under $H_{\lambda t^\star}$.

2.7 Article submitted “LRT process for QTL detection"

“Likelihood Ratio Test process for Quantitative Trait Loci detection"

Rabier C-E, Azaïs J-M, Delmas C.

Likelihood Ratio Test process for Quantitative Trait Loci detection

Charles-Elie Rabier

Institut de Mathématiques de Toulouse, Toulouse, France.
INRA UR631, Auzeville, France.

Jean-Marc Azaïs

Institut de Mathématiques de Toulouse, Toulouse, France.

Céline Delmas

INRA UR631, Auzeville, France.

Summary. We consider the likelihood ratio test (LRT) process related to the test of the absence of QTL on the interval $[0, T]$ representing a chromosome (a QTL denotes a quantitative trait locus, i.e. a gene with quantitative effect on a trait). We give the asymptotic distribution of this LRT process under the null hypothesis that there is no QTL on $[0, T]$ and under the general alternative that there exist m QTL on $[0, T]$. We propose to estimate the number of QTL, their positions and their effects by penalized likelihood. Our results are extended to the case where individuals are structured into families.

Keywords: Gaussian process, Likelihood Ratio Test, Mixture models, Nuisance parameters present only under the alternative, QTL detection, χ^2 process.

1. Introduction

We study a population of progenies of a sire 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 independent and identically distributed (iid). The mechanism of genetics, or more precisely of meiosis, implies that among the two chromosomes of each individual, one is inherited from the dame (whose effect will be neglected) and the other, inherited from the sire, consists of parts originated from chromosome 1 of the sire and parts originated from chromosome 2 of the sire, due to crossing-overs. Note that the back-cross population, $A \times (A \times B)$, where A and B are purely homozygous lines, is a particular case of such a population. Using the Haldane (1919) distance and modelling, each chromosome will be represented by a segment $[0, T]$. The distance on $[0, T]$ is called the genetic distance (which is measured in Morgans). The key point is that, if the true position of the QTL is $t = t^\star$, the response Y obeys to a mixture model with known weights :

$$p(t) f_{(\mu+q,\sigma)}(.) + \{1 - p(t)\} f_{(\mu-q,\sigma)}(.) \tag{1}$$

where $f_{(\mu,\sigma)}(.)$ denotes a Gaussian density with mean μ and variance σ^2. $(\mu,\ q,\ \sigma)$ are the unknown parameters. At every location $t \in [0, T]$, we perform a likelihood ratio test (LRT) of the hypothesis "$q = 0$" in formula (1) based on n observations $Y_1, ..., Y_n$. We

call $\Lambda_n(t)$ the obtained quantity. The dependence on t of the weights is precisely described in Section 3. We denote $p_j(t)$ the value of the weight $p(t)$ for the jth observation. The process $\{\Lambda_n(t),\ t \in [0,T]\}$ will be called "likelihood ratio test 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.

In the special case where the weights are 0 or 1 depending on the individual, Lander and Botstein (1989) stated that the asymptotic distribution of the LRT process along $[0,T]$ is the square of an Ornstein-Uhlenbeck process. This result has been proved by Cierco (1998). Bounds for the distribution of the maximum of a regularization of an Ornstein-Uhlenbeck process were proposed by Azaïs and Cierco-Ayrolles (2002), Azaïs and Wschebor (2009). Some results about the asymptotic distribution of the LRT process under the null hypothesis are given in Rebaï et al. (1994) for a special modelling of the weights. Their results are inferred from the bounds given by Davies (1977), Davies (1987) for the maximum of sufficiently regular Gaussian and chi-square processes.

In this paper we consider the modelling of the weight used by geneticists to detect QTL, so called Interval Mapping. First we give the asymptotic distribution of the LRT process along the interval $[0,T]$ under the null hypothesis that there is no QTL on $[0,T]$ ($q=0$) and under the alternative that there is one QTL at $t^\star$ on $[0,T]$ which means that the quantitative trait for each individual is distributed as the mixture in formula (1) with $t = t^\star$. Then we compute the asymptotic distribution of the LRT process under the general alternative that there exist m QTL on $[0,T]$ at $t_1^\star, \cdots, t_m^\star$ with additive effects $q^1, \cdots, q^m$. The response is now a mixture of $M = 2^m$ components of the form :

$$\sum_{\alpha=1}^{M} p_\alpha f_{(m_\alpha,\sigma)}(.)$$

where the p_αs and the m_αs are known functions of the unknown parameters μ, m, $t_1^\star$, ..., $t_m^\star$, q^1, ..., q^m. Under this general alternative, the LRT process is shown to converge towards a Gaussian process with mean function depending on these unknown parameters. We propose to estimate the unknown parameters by penalized likelihood.
Besides, we show that the LRT process is asymptotically the square of a "non linear interpolated process" (which means that the LRT statistics at each point can easily be deduced from the Wald or score statistics calculated at the positions where the auxiliary information is available). Note that in some remarks sections we also prove that the LRT process obtained by Rebaï et al. (1994), Rebaï et al. (1995) is asymptotically the square of a "linear interpolated process". Finally, our results are extended to the case where individuals are structured into families of sires. Recently, the law of the LRT process under the null hypothesis has also been obtained by Chang et al. (2009). Our work has been done independently. Technical differences are presented in appendix 8.5.

The originality of our paper is twofold. First we consider the true model used by geneticists to detect QTL whereas the model considered by Rebaï et al. is only an approximation. Then we obtain results not only under the null hypothesis, but also under the general alternative. This last result leads us to propose a new method, based on penalized likelihood, for estimating the number of QTL, their positions and their effects. We refer to the book of Van der Vaart (1998) for element of asymptotic statistics used in proofs. In a future paper, we will present the applications of the theoretical results presented here.

2. Model

The chromosome is the segment $[0,T]$. K genetic markers are located on the chromosome, one at each extremity. $t_1 = 0 < t_2 < ... < t_K = T$ are the locations of the markers. The "genome information" at t will be denoted $X(t)$. The Haldane (1919) model can be written mathematically : let $N(t)$ be a standard Poisson process, the law of $X(t)$ is $\frac{1}{2}(\delta_1 + \delta_{-1})$ and $X(t) = (-1)^{N(t)}X(t_1)$. The Haldane (1919)'s function $r : [0,\ T]^2 \longmapsto \left[0,\ \frac{1}{2}\right]$ is such as :

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

$\bar{r}(t,t')$ will be the function equal to $1 - r(t,t')$.
We are interested in a quantitative trait Y which depends on the value of $X(t)$ at $t^\star \in [t_1, t_K]$ which is the location of the QTL. The quantitative trait verifies :

$$Y_j = \mu\ +\ X(t^\star)\ q\ +\ \sigma\varepsilon$$

where ε is a Gaussian white noise and q the effect of the QTL.
Besides, the "genome information" is available only at locations of genetic markers, that is to say at $t_1, t_2, ..., t_K$. We denote by $X_j(t)$ the value of the variable $X(t)$ for the jth observation. So, in fact, our observation on each individual is (Y_j, $X_j(t_1)$, ..., $X_j(t_K)$). These observations are supposed to be iid. The goal of this study is to test if q is equal to zero. The challenge is that $t^\star$ is unknown.

3. Only 2 genetic markers

To begin, we suppose that there are only two markers ($K = 2$) located at 0 and T : $0 = t_1 < t_2 = T$. As explained previously, we are looking for a QTL lying at a position $t^\star \in [t_1, t_2]$. Let $t \in [t_1, t_2]$. It is clear that the weight $p(t)$ satisfies $p(t) = \mathbb{P}\left\{X(t) = 1 \middle| X(t_1), X(t_2)\right\}$. Consider for example the case $X(t_1) = X(t_2) = 1$, then by the Bayes rule :

$$\mathbb{P}\left\{X(t) = 1 \middle| X(t_1) = 1, X(t_2) = 1\right\} = \frac{(1/2)\ \mathbb{P}\left\{N(t) - N(t_1)\ \text{even}\right\} \mathbb{P}\left\{N(t_2) - N(t)\ \text{even}\right\}}{(1/2)\ \mathbb{P}\left\{N(t_2) - N(t_1)\ \text{even}\right\}}$$

So that, in general $\forall t \in]t_1, t_2[$:

$$\begin{aligned} p(t) = {} & Q_t^{1,1}\ 1_{X(t_1)=1} 1_{X(t_2)=1}\ +\ Q_t^{1,-1}\ 1_{X(t_1)=1} 1_{X(t_2)=-1} \\ & + Q_t^{-1,1}\ 1_{X(t_1)=-1} 1_{X(t_2)=1}\ +\ Q_t^{-1,-1}\ 1_{X(t_1)=-1} 1_{X(t_2)=-1} \end{aligned} \tag{2}$$

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 remark that we have :

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

Besides, $p(t_1) = 1_{X(t_1)=1}$ and $p(t_2) = 1_{X(t_2)=1}$. So, the weights $p(t)$ are continuous at t_1 and t_2.

Let $\theta = (q,\ \mu,\ \sigma)$ be the parameter of the model at t fixed and $\theta_0 = (0,\ \mu,\ \sigma)$ the true value of the parameter under H_0. 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 county measure on $\mathbb{N}$, is $\forall t \in [t_1, t_2]$:

$$L(\theta, t) = \left[p(t) f_{(\mu+q,\sigma)}(y) + \{1 - p(t)\} f_{(\mu-q,\sigma)}(y)\right]\ g(t) \tag{3}$$

where

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

The likelihood $L_n(\theta, t)$ for n observations is obtained by the product of n terms as above. $\hat{\theta} = (\hat{q},\ \hat{\mu},\ \hat{\sigma})$ will be the maximum likelihood estimator (MLE) of θ.

Under H_0, there is no QTL lying on the interval $[t_1, t_2]$. Besides, under H_1, it is supposed that there is only one location where the QTL lies. The location of the QTL, $t^\star$ ($t^\star \in [t_1, t_2]$), will be added in the definition of H_1. So, the alternative hypothesis can be written :

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

The QTL effect q is such as $q = a/\sqrt{n}$ in order to deal with Le Cam (1986)'s theory.

3.1. Results

Theorem 1 *With the previous defined notations,*

$$S_n(.) \Rightarrow Z(.) \quad , \quad \Lambda_n(.) \overset{F.d.}{\rightarrow} \{Z(.)\}^2$$

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

- *$S_n(.)$ is the score process for n observations*
- *$\Rightarrow$ is the weak convergence and $\overset{F.d.}{\rightarrow}$ is the convergence of finite-dimensional distributions*
- *$Z(.)$ is the Gaussian process with covariance function $\forall (t, t') \in [t_1, t_2]^2$:*

$$\Gamma(t, t') = \frac{4\mathbb{E}\left\{p(t)p(t')\right\} - 1}{\sqrt{\mathbb{E}\left[\{2p(t) - 1\}^2\right]}\sqrt{\mathbb{E}\left[\{2p(t') - 1\}^2\right]}}$$

and expectation $\forall (t, t^\star) \in [t_1, t_2]^2$:

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

$$m_{t^\star}(t) = \frac{a\ \mathbb{E}\left[X(t^\star)\left\{2p(t) - 1\right\}\right]}{\sigma\ \sqrt{\mathbb{E}\left[\{2p(t) - 1\}^2\right]}}$$

Another way of characterizing $Z(.)$ *is that* $Z(.)$ *is the non linear interpolated process such as* $\forall t \in [t_1, t_2]$ *:*

$$Z(t) = \{\ \alpha(t)\ Z(t_1)\ +\ \beta(t)\ Z(t_2)\ \} / \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

where $\forall t \in]t_1, t_2[$, $\alpha(t) = Q_t^{1,1} + Q_t^{1,-1} - 1$, $\beta(t) = Q_t^{1,1} - Q_t^{1,-1}$ *and* $\alpha(t_1) = 1$, $\beta(t_1) = 0$, $\alpha(t_2) = 0$, $\beta(t_2) = 1$, $Cov\{Z(t_1), Z(t_2)\} = e^{-2t_2}$.
In the same way, $\forall (t, t^\star) \in [t_1, t_2]^2$ *:*

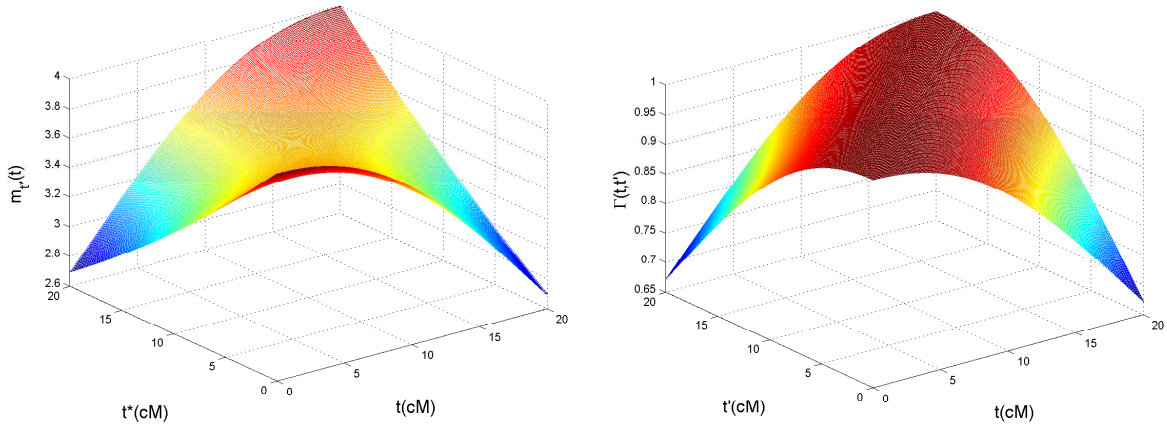
$$m_{t^\star}(t) = \{\ \alpha(t)\ m_{t^\star}(t_1)\ +\ \beta(t)\ m_{t^\star}(t_2)\ \} / \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

The quantity $\mathbb{E}\left[\{2p(t)-1\}^2\right]$ *is given in formula (10) of the proof of the theorem in Section 7.1.* $\mathbb{E}\{p(t)p(t')\}$ *is given in appendix 8.1.* $\mathbb{E}[X(t^\star)\{2p(t)-1\}]$ *is given in formula (15) of the proof in Section 7.1.*

We limit our attention to finite dimensional convergence since for the applications, the interval studied is always discretized, Wu et al. (2007).

Figures 1 represent the covariance function $\Gamma(t, t')$ and also the mean function $m_{t^\star}(t)$. T is equal to 0.2M. We can remark that the covariance function is regular.

Contrary to Azaïs et al. (2006) and Azaïs et al. (2009), the shift at position t is not $\Gamma(t, t^\star)$. The model considered here is more complicated due to the fact that an observation includes the quantitative trait Y and the "genome information", $X(t_1)$ and $X(t_2)$. As it is well



Mean function Covariance function

Fig. 1. Mean function and Covariance function ($a = 4$, $\sigma = 1$, $T = 0.2$M)

known, for regular model, LRT is equivalent to Wald test, and score test in the sense that $\forall t \in [t_1, t_2]$:

$$\Lambda_n(t) = \{W_n(t)\}^2 + o_{P_{\theta_0}}(1) = \{S_n(t)\}^2 + o_{P_{\theta_0}}(1)$$

where $W_n(t)$ and $S_n(t)$ are respectively the Wald and the score test statistic for n observations. We remind that, as in the proof of the theorem in Section 7.1, the notation $o_{P_{\theta_0}}(1)$ is short for a sequence of random vectors that converges to zero in probability under H_0 (i.e. no QTL on the whole interval studied).
According to formula (9), given in Section 7.1 :

$$W_n(t) = \sqrt{n}\ \hat{q}\ \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}/\sigma \quad , \quad S_n(t) = \sum_{j=1}^{n} \frac{(y_j-\mu)\ (2p_j(t)-1)}{\sqrt{n}\ \sigma\ \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}} \tag{4}$$

Note that the Wald test can be obtained, replacing σ by $\hat{\sigma}$, according to Slutsky's lemma. The score test can be obtained, replacing μ by $\hat{\mu}$, according to Prohorov, and replacing σ by $\hat{\sigma}$, according to Slutsky's lemma. Nevertheless, in order to make the reading easier, the Wald and the score test statistic are defined as in formula (4). The score process considered in theorem 1 is based on this formula. Howewer, we have the same result as in theorem 1 for the other score process because the tightness of this process is obvious according to the proof of theorem 1.

After some calculations, we can remark that :

$$S_n(t) = \{\ \alpha(t)\ S_n(t_1)\ +\ \beta(t)\ S_n(t_2)\ \}\ /\sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]} \tag{5}$$

with $\mathrm{Cov}\left\{S_n^0(t_1), S_n^0(t_2)\right\} = e^{-2t_2}$ where $S_n^0(.)$ is the score process under H_0.
It comes :

$$\begin{aligned}\Lambda_n(t) &= \{\ \alpha(t)\ S_n(t_1)\ +\ \beta(t)\ S_n(t_2)\ \}^2\ /\ \mathbb{E}\left[\{2p(t)-1\}^2\right]\ +\ o_{P_{\theta_0}}(1)\\ &= \{\ \alpha(t)\ W_n(t_1)\ +\ \beta(t)\ W_n(t_2)\ \}^2\ /\ \mathbb{E}\left[\{2p(t)-1\}^2\right]\ +\ o_{P_{\theta_0}}(1)\end{aligned}$$

Besides, by contiguity (cf. proof of theorem 1 in Section 7.1), the quantity $o_{P_{\theta_0}}(1)$ converges also to zero under $H_{at^\star}$. That is to say, the LRT statistic at a position t between the two genetic markers is asymptotically equal to the square of a non linear interpolation between the Wald or score test statistics on the markers.
Note that computing the square of this non linear interpolated process will be be quicker than calculating the observed process $\Lambda_n(.)$ on real data, which is time consuming because an EM algorithm is required to calculate the MLE's when the position tested is not on genetic markers.

3.2. Remarks

To construct an approximation of $S_n(.)$ (and $\Lambda_n(.)$), we introduce a new process $V_n(.)$ which is obtained from $S_n(.)$ by :

- linear (or polygonal) interpolation
- renormalization

More precisely :

$$V_n(t) = \left\{ \frac{t_2 - t}{t_2} \, S_n(t_1) \; + \; \frac{t}{t_2} \, S_n(t_2) \right\} / \sqrt{\tau(t)} \tag{6}$$

where

$$\tau(t) = \mathbb{V} \left\{ \frac{t_2 - t}{t_2} \, S_n^0(t_1) \; + \; \frac{t}{t_2} \, S_n^0(t_2) \right\} = \left(\frac{t_2 - t}{t_2} \right)^2 \; + \; 2 \, \frac{t(t_2 - t)}{(t_2)^2} e^{-2t_2} \; + \; \left(\frac{t}{t_2} \right)^2$$

It can be seen easily that $\tau(t) \neq 0$, $\forall t \in [t_1, t_2]$. $V_n(.)$ remains asymptotically a Gaussian process, centered under H_0, with unit variance and $\mathrm{Cov}\left\{ S_n^0(t_1), S_n^0(t_2) \right\} = e^{-2t_2}$.

Some comments about the linear interpolated process $V_n(.)$:

(a) According to formula (11) in Section 7.1 and after some calculations, we can establish that asymptotically, the process $V_n^2(.)$ corresponds to likelihood ratio tests for a mixture model whose weights verify :

$$p(t) = 1_{X(t_1)=1} 1_{X(t_2)=1} \; + \; \frac{t_2 - t}{t_2} \, 1_{X(t_1)=1} 1_{X(t_2)=-1} \; + \; \frac{t}{t_2} \, 1_{X(t_1)=-1} 1_{X(t_2)=1} \tag{7}$$

We can remark that these weights are an approximation at the first order of the weights considered previously in formula (2). So, the linear interpolated process will be a good approximation if and only if the genetic markers are close to each other.

(b) $V_n^2(.)$ is a generalization of the process studied, under H_0, by Rebaï et al. (1995) : the number of individuals in each class is not equal to the expectations (respectively $n\bar{r}(t_1, t_2)/2$, $nr(t_1, t_2)/2$, $nr(t_1, t_2)/2$, $n\bar{r}(t_1, t_2)/2$) but is still random (respectively $\sum_{j=1}^n 1_{X_j(t_1)=1} 1_{X_j(t_2)=1}$, $\sum_{j=1}^n 1_{X_j(t_1)=1} 1_{X_j(t_2)=-1}$, $\sum_{j=1}^n 1_{X_j(t_1)=-1} 1_{X_j(t_2)=1}$ and $\sum_{j=1}^n 1_{X_j(t_1)=-1} 1_{X_j(t_2)=-1}$).

(c) By contiguity (cf. proof of theorem 1 in Section 7.1), under $H_{at^\star}$, $V_n(.)$ is asymptotically the same process as under H_0 on which the mean function $\tilde{m}_{t^\star}(t)$ has been added. $\tilde{m}_{t^\star}(t)$ is such as :

$$\tilde{m}_{t^\star}(t) = \left\{ \frac{t_2 - t}{t_2} \, m_{t^\star}(t_1) \; + \; \frac{t}{t_2} \, m_{t^\star}(t_2) \right\} / \sqrt{\tau(t)}$$

(d) $V_n(.)$ is defined here with $\mathrm{Cov}\left\{ S_n^0(t_1), S_n^0(t_2) \right\} = e^{-2t_2}$. In order to consider other covariances between $S_n^0(t_1)$ and $S_n^0(t_2)$, $\tau(.)$ has to be adapted. It can easily be seen that the new process $V_n^2(.)$ is still a generalization of the process studied by Rebaï et al. (1995) for any covariance between $S_n^0(t_1)$ and $S_n^0(t_2)$ as soon as $\mathbb{E}\left[\{2p(t) - 1\}^2 \right] \neq 0$ ($p(t)$ verifies formula (7)).

4. Several markers : the "Interval Mapping" of Lander and Botstein (1989)

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 :

$$t^\ell = sup \left\{ t_k \in \mathbb{T}_k : t_k < t \right\} \quad , \quad t^r = inf \left\{ t_k \in \mathbb{T}_k : t < t_k \right\}$$

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

Theorem 2 *We have the same result as in theorem 1 except that the following expressions are more complicated :*

$$\mathbb{E}\left[\{2p(t)-1\}^2\right] \quad , \quad \mathbb{E}\{p(t)p(t')\} \quad , \quad \mathbb{E}\left[X(t^\star)\{2p(t)-1\}\right] \quad , \quad \alpha(t) \quad , \quad \beta(t)$$

Besides, $Z(.)$ is now the non linear interpolated process such as :

$$Z(t) = \left\{ \alpha(t)\, Z(t^\ell) \;+\; \beta(t)\, Z(t^r) \right\} / \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

with $\forall k\ \forall k'$, $Cov\{Z(t_k), Z(t_{k'})\} = e^{-2|t_k - t_{k'}|}$.
In the same way, the mean function $m_{t^\star}(t)$ is now such as :

$$m_{t^\star}(t) = \left\{ \alpha(t)\, m_{t^\star}(t^\ell) \;+\; \beta(t)\, m_{t^\star}(t^r) \right\} / \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

All these expressions including a proof are given in appendix 8.2.

Note that $\forall k \ \ \forall k'$, $\Gamma(t_k, t_{k'}) = e^{-2|t_k - t_{k'}|}$. It is relative to an Ornstein-Uhlenbeck process, as studied in Lander and Botstein (1989), and Cierco (1998).
Besides, in the same way as what has been done in Section 3.1, we have :

$$\forall k \quad W_n(t_k) = \sqrt{n}\,\hat{q}/\sigma \quad , \quad S_n(t_k) = \sum_{j=1}^{n} \frac{(y_j - \mu)\,(2\,1_{X_j(t_k)=1} - 1)}{\sigma\ \sqrt{n}}$$

$$\begin{aligned} \Lambda_n(t) &= \left\{ \alpha(t)\, S_n(t^\ell) \;+\; \beta(t)\, S_n(t^r) \right\}^2 / \mathbb{E}\left[\{2p(t)-1\}^2\right] \;+\; o_{P_{\theta_0}}(1) \\ &= \left\{ \alpha(t)\, W_n(t^\ell) \;+\; \beta(t)\, W_n(t^r) \right\}^2 / \mathbb{E}\left[\{2p(t)-1\}^2\right] \;+\; o_{P_{\theta_0}}(1) \end{aligned} \tag{8}$$

Note that $\forall k\ \forall k'$, $\mathrm{Cov}\left\{S_n^0(t_k), S_n^0(t_{k'})\right\} = e^{-2|t_k - t_{k'}|}$.
Besides, by contiguity (cf. appendix 8.2), the quantity $o_{P_{\theta_0}}(1)$ converges also to zero under $H_{at^\star}$.

4.1. Remarks

The linear interpolated process $V_n(.)$ presented in Section 3.2 can easily be generalized to the case of several markers. This process is a generalization of the process studied, under H_0, by Rebaï et al. (1994). The details are given in appendix 8.3.

On the other hand, the problem considered in this article can be viewed as a missing data problem. The auxiliary information $X(t)$ is available only at the location of genetic markers, otherwise the information is missing. Since in absence of missing data, the process is relative to an Ornstein-Uhlenbeck process, the missing observations can be obtained by a kriging method. The process refering to the kriging method will be called $M_n(.)$. After some easy

calculations, we obtain :

$$\begin{aligned} M_n(t) &= \left\{ e^{-2(t-t^\ell)} - \gamma(t)\, e^{-2(t^r-t^\ell)} \right\} S_n(t^\ell) \;\; + \;\; \gamma(t)\, S_n(t^r) \\ &= \left\{ e^{-2(t-t^\ell)} - \gamma(t)\, e^{-2(t^r-t^\ell)} \right\} W_n(t^\ell) \;\; + \;\; \gamma(t)\, W_n(t^r) \;+\; o_{P_{\theta_0}}(1) \end{aligned}$$

where $\gamma(t) = \frac{e^{-2(t^r-t)} - e^{-2(t-2t^\ell+t^r)}}{1-e^{-4(t^r-t^\ell)}}$. Note that this process is asymptotically a Gaussian process, centered under H_0 but with unit variance only at location of genetic markers.

Under $H_{at^\star}$, by contiguity (in the same way of what has been done in the proof of theorem 1), this is asymptotically the same process but the mean function, $\bar{m}_{t^\star}(t)$ is added to the process:

$$\bar{m}_{t^\star}(t) = \left\{ e^{-2(t-t^\ell)} - \gamma(t)\, e^{-2(t^r-t^\ell)} \right\} m_{t^\star}(t^\ell) \;\; + \;\; \gamma(t)\, m_{t^\star}(t^r)$$

If now a rescaling is done in order to obtain a process with unit variance, then by Taylor expansions at the first order, it can be seen that the process obtained by the kriging method is exactly the interpolated process $V_n(.)$ (asymptotic is not required).

4.2. Illustration

Figure 2 represents under H_0, the cumulative distribution function of the square of the sup of four different processes on $[0, 0.6]$ with 4 markers equally spaced every 0.2M : the asymptotic process $Z(.)$, the limiting process of the linear interpolated process $V_n(.)$, the limiting process of the kriging process $M_n(.)$ and the asymptotic process for which the tests are only done on the markers. 100000 sample paths of the different processes have been simulated. Each test is done every cM. The interest is on the quantile of order 95%. It is not surprising that when tests are only on markers, the cumulative density function is above the three other curves. Howewer, the curves of the linear interpolated process and the process $Z(.)$ are pretty close : the estimation of the quantile of order 95% is 6.06 for the linear interpolated and 6.08 for the process $Z(.)$. (5.86 if tests are done only on markers). It is not surprising because markers are close to each other. On the other hand, the kriging process is another way of analyzing data, that's why the corresponding cumulative distribution is so different (the estimation of the quantile is 6.77).

5. Generalization

In the previous sections, we were looking for a QTL lying on the interval $[0, T]$ using the concept of Interval Mapping. One population of progenies of a sire has been studied that is to say one family. In order to increase the power of the method, geneticists look for the QTL not in one family but simultanously in several families, each defined by a different sire. It increases the chances to study families whose sires are heterozygote at the QTL. In that case, the Interval Mapping method is also used : LRT are peformed at each position $t \in [0, T]$ and the supremum of these statistics is used as a unique test statistic. Naturally, the putative QTL is supposed to be lying at the same location in each family otherwise the concept of Interval Mapping has no sense.

Let's extend the model presented in Section 2 to the case of several families. Let $I \in \mathbb{N}^\star$ be the number of families. Let C be a discrete random variable refering to the family :

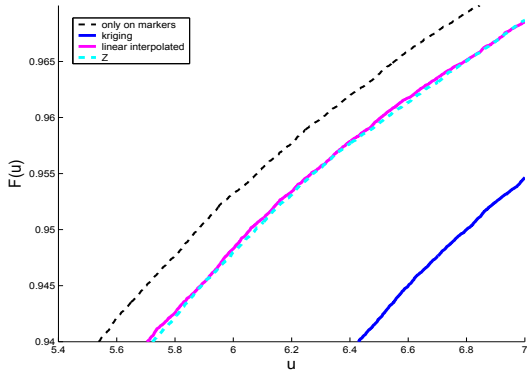


Fig. 2. Cumulative distribution function of the square of four different processes ($T = 0.6$M, 4 markers equaly spaced every 0.2M)

$\pi_i = P(C = i)$. In other words, the individual belongs to family i with probability π_i. When we deal with different families, the location and the number of the genetic markers usually differ in each family. However, in order to make reading easier, we will supposed here, that the location and the number of genetics markers do not differ with the family. The general results are present in Rabier (2009).

In our case, the process $X(.)$ is unchanged. Howewer, the quantitative trait verifies now :

$$\big(Y\big|C = i\big) \;=\; \mu_i \;+\; X(t^\star)\; q_i \;+\; \sigma\; \varepsilon$$

where μ_i and q_i are respectively a polygenic effect and the QTL effect inside family i. ε is a Gaussian white noise.

We denote by C_j the value of the variable C for the jth observation. In fact, our observation on each individual is $(Y_j,\ X_j(t_1),\ ...,\ X_j(t_K),\ C_j)$. These observations are supposed to be iid. The goal of this study is to test if all the q_i are equal to zero. As previously, the challenge is that $t^\star$ is unknown.

The same notations as in Section 4 will be used and as previously, we consider only values t, t' or $t^\star$ of the parameters that are distinct of the marker positions. The result will be prolonged by continuity at the markers positions.
Let $\theta = (q_1,\ ...,\ q_I,\ \mu_1,\ ...,\ \mu_I,\ \sigma)$ be the parameter of the model at t fixed and
$\theta_0 = (0,\ ...,\ 0,\ \mu_1,\ ...,\ \mu_I,\ \sigma)$ the true value of the parameter under H_0. The likelihood of the triplet $\big(Y,\ X(t^\ell),\ X(t^r),\ C\big)$ with respect to the measure $\lambda \otimes N \otimes N \otimes N$, λ being the Lebesgue measure, N the county measure on $\mathbb{N}$, is at a position t :

$$L(\theta, t) = \sum_{i=1}^{I} \left[p(t) f_{(\mu_i+q_i,\sigma)}(y) + \{1-p(t)\}\, f_{(\mu_i-q_i,\sigma)}(y)\right]\ 1_{C=i}\ \frac{\pi_i}{2}\ g(t)$$

where $g(t)$ is the same function as in Section 3 adapted to the Marker interval (t^ℓ, t^r).
The likelihood $L_n(\theta, t)$ for n observations is obtained by the product of n terms as above.
$\hat{\theta} = (\hat{q}_1,\ ...,\ \hat{q}_I,\ \hat{\mu}_1, ...,\ \hat{\mu}_I,\ \hat{\sigma})$ will be the MLE of θ.

The alternative hypothesis can be written :

$$H_{at^\star} : \text{“there is at least one } q_i = a_i/\sqrt{n}, \text{ with } a_i \in \mathbb{R}^\star, \text{ at the position } t^\star \text{ "}$$

5.1. Results

Theorem 3 *With the previous defined notations,*

$$\Lambda_n(.) \stackrel{F.d.}{\rightarrow} \sum_{i=1}^{I} \left\{ Z^i(.) \right\}^2$$

as n tends to infinity, under H_0 and $H_{at^\star}$ where the $Z^i(.)$ are independent Gaussian processes, with covariance function $\Gamma(t,t')$ and with expectation :

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

$$m^i_{t^\star}(t) = \frac{a_i \sqrt{\pi_i}\, \mathbb{E}\left[X(t^\star)\left\{2p(t)-1\right\}\right]}{\sigma \sqrt{\mathbb{E}\left[\left\{2p(t)-1\right\}^2\right]}}$$

Another way of characterizing $Z^i(.)$ is that $Z^i(.)$ is the non linear process such as :

$$Z^i(t) = \left\{ \alpha(t)\, Z^i(t^\ell) \;+\; \beta(t)\, Z^i(t^r) \right\} / \sqrt{\mathbb{E}\left[\left\{2p(t)-1\right\}^2\right]}$$

with $\forall k\, \forall k'$, $Cov\left\{Z^i(t_k), Z^i(t_{k'})\right\} = e^{-2|t_k - t_{k'}|}$.
In the same way :

$$m^i_{t^\star}(t) = \left\{ \alpha(t)\, m^i_{t^\star}(t^\ell) \;+\; \beta(t)\, m^i_{t^\star}(t^r) \right\} / \sqrt{\mathbb{E}\left[\left\{2p(t)-1\right\}^2\right]}$$

The quantities $\Gamma(t,t')$, $\mathbb{E}\left[X(t^\star)\left\{2p(t)-1\right\}\right]$, $\mathbb{E}\left[\left\{2p(t)-1\right\}^2\right]$, $\alpha(t)$ and $\beta(t)$ are the same as in theorem 2.

The proof of this theorem is given in Section 7.2. Note that this theorem is also suitable when the π_i's are unknown.

In the same way as what has been done in the previous Sections, let $W_n(t_k,i)$ and $S_n(t_k,i)$ be respectively the Wald statistic and the score statistic, which corresponds to testing the presence of a QTL in family i at a marker location t_k. According to the proof of theorem 3 in Section 7.2 :

$$W_n(t_k,i) = \sqrt{n\,\pi_i}\,\hat{q}_i/\sigma \quad , \quad S_n(t_k,i) = \sum_{j=1}^{n} \frac{(y_j - \mu_i)\,(2\,1_{X_j(t_k)=1} - 1)}{\sigma\sqrt{n\,\pi_i}} 1_{C_j=i}$$

$$\begin{aligned}
\Lambda_n(t) &= \sum_{i=1}^{I} \left\{ \alpha(t)\, S_n(t^\ell,i) \;+\; \beta(t)\, S_n(t^r,i) \right\}^2 / \mathbb{E}\left[\left\{2p(t)-1\right\}^2\right] \;+\; o_{P_{\theta_0}}(1) \\
&= \sum_{i=1}^{I} \left\{ \alpha(t)\, W_n(t^\ell,i) \;+\; \beta(t)\, W_n(t^r,i) \right\}^2 / \mathbb{E}\left[\left\{2p(t)-1\right\}^2\right] \;+\; o_{P_{\theta_0}}(1)
\end{aligned}$$

Note that $\mathrm{Cov}\left\{S_n^0(t_k,i), S_n^0(t_{k'},i)\right\} = e^{-2|t_k - t_{k'}|}$.
Besides, by contiguity (cf. Section 7.2), the quantity $o_{P_{\theta_0}}(1)$ converges also to zero under $H_{at^\star}$.

5.2. Remarks

The linear interpolated process and the process obtained by kriging can be generalized to the case of several families. The details are presented in appendix 8.4.

6. Extension to several QTL

Until now, it has been supposed that there was only one QTL lying on the interval $[0,T]$. So, the test statistic used was a natural statistic, that is to say the supremum of the process. The interest is now on studying the same processes as previously, $\Lambda_n(.)$ and $S_n(.)$, but under the presence of several QTL on the interval $[0,T]$. In this case, the goal is not to perform a test anymore, but to be able to run a model selection in order to estimate the number of QTL and their locations.
In order to make the reading easier, we will deal with only one family. m will refer to the number of QTL and q^s to the QTL effect of the sth QTL. Its position will be called $t_s^\star$. We impose $t_1^\star < ... < t_m^\star$ and we will suppose that the QTL effects are additives and there is no interaction between them. So, the quantitative trait Y verifies :

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

where ε is a Gaussian white noise.

Let denote $\vec{t^\star}$ the quantity refering to the locations of the QTL. $H_{a\vec{t^\star}}$ will be the following assumption :

$H_{a\vec{t^\star}}$: " there are m QTL located respectively at $t_1^\star$, ..., $t_m^\star$ and with effect
$q^1 = \frac{a^1}{\sqrt{n}}$, ..., $q^m = \frac{a^m}{\sqrt{n}}$ where $(a^1, ..., a^m) \in \mathbb{R}^{m\star}$ "

We will consider values t, t', $t_1^\star$, ..., $t_m^\star$ of the parameters that are distinct of the markers positions, and the result will be prolonged by continuity at the markers positions.

6.1. Results

Theorem 4 *With the previous defined notations,*

$$S_n(.) \Rightarrow Z^\star(.) \quad , \quad \Lambda_n(.) \overset{F.d.}{\rightarrow} \left\{Z^\star(.)\right\}^2$$

as n tends to infinity, under H_0 and $H_{a\vec{t^\star}}$ where the $Z^\star(.)$ is a Gaussian process, with covariance function $\Gamma(t,t')$ and with expectation :

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

$$m_{\vec{t^\star}}(t) = \sum_{s=1}^{m} \frac{a^s\ \mathbb{E}\left[X(t_s^\star)\ \{2p(t)-1\}\right]}{\sigma\ \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}}$$

Another way of characterizing $Z^\star(.)$ *is that* $Z^\star(.)$ *is the non linear process such as :*

$$Z^\star(t) = \left\{ \alpha(t)\ Z^\star(t^\ell)\ +\ \beta(t)\ Z^\star(t^r) \right\} / \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

with $\forall k\ \forall k',\ Cov\{Z^\star(t_k), Z^\star(t_{k'})\} = e^{-2|t_k - t_{k'}|}$.
In the same way :

$$m_{\vec{t}^\star}(t) = \sum_{s=1}^{m} \frac{a^s}{\sigma} \left\{ \alpha(t)\ \mathbb{E}\left[X(t_s^\star)\ \left\{2p(t^\ell)-1\right\}\right]\ +\ \beta(t)\ \mathbb{E}\left[X(t_s^\star)\ \{2p(t^r)-1\}\right] \right\} / \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

The quantities $\Gamma(t,t')$, $\mathbb{E}\left[\{2p(t)-1\}^2\right]$, $\alpha(t)$, $\beta(t)$ *are the same as in theorem 2.*
$\mathbb{E}\left[X(t_s^\star)\ \{2p(t)-1\}\right]$ *is given in formula (21) of the proof of the theorem in Section 7.3.*

As we focus on the same LRT process as previously, formula (8) of Section 4 is still suitable. Besides, by contiguity (cf. Section 7.3), the quantity $o_{P_{\theta_0}}(1)$ converges also to zero under $H_{a\vec{t}^\star}$.

All the results presented in this Section 6 can easily be generalized to the case of several families and also to interactions between the QTL.

6.2. Estimation of the parameters

As for the application, the interval $[0,T]$ is always discretized, let consider only these points of discretization : $0 = s_1 < s_2 < ... < s_d = T$. Without loss of generality, it can be supposed that the QTL are located on these points of discretization.
We estimate the unknown parameters m, a^1,..., a^m and consequently $t_1^\star$,..., $t_m^\star$ by a penalized likelihood method (lasso Tibshirani (1996), elastic net Zou and Hastie (2005), dantzig selector Candes and Tao (2005)) applied to the model :

$$S_n(s_e) = \sum_{i=1}^{d} \frac{a^{s_i}\ \mathbb{E}\left[X(t_{s_i}^\star)\ \{2p(s_e)-1\}\right]}{\sigma \sqrt{\mathbb{E}\left[\{2p(s_e)-1\}^2\right]}}\ +\ \varepsilon_{s_e} \qquad e = 1, ..., d$$

where ε_{s_e} is a Gaussian white noise and $\mathrm{Cov}(\varepsilon_{s_e}, \varepsilon_{s_{e'}}) = \Gamma(s_e, s_{e'})$.
We remind that $S_n(s_e)$ is the score test for n observations performed at the position s_e.
This method will be investigated in a forthcoming paper.

7. Proofs

Notations : I_θ will be the Fisher information matrix taken at the point θ . $I_{ij}(\theta)$ refers to the element ij of I_θ. $I_{ij}^{-1}(\theta)$ refers to the element ij of I_θ^{-1}, the inverse of I_θ.

7.1. Proof of theorem 1

We first compute the score functions and the Fisher information matrix. Let $t \in [t_1, t_2]$.

$$\frac{\partial \log L}{\partial q}\ |_{\theta_0} = \frac{y-\mu}{\sigma^2}\ \{2p(t)-1\}$$

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

$$I_{11}(\theta_0) = \frac{\mathbb{E}\left[\{2p(t)-1\}^2\right]}{\sigma^2} \quad , \quad I_{22}(\theta_0) = \frac{1}{\sigma^2}$$

As the fourth-order moment of a standard normal distribution is equal to three,

$$I_{33}(\theta_0) = \frac{2}{\sigma^2}$$

After some calculations, we find : $I_{12}(\theta_0) = I_{13}(\theta_0) = I_{23}(\theta_0) = 0$. So,

$$I_{\theta_0} = Diag\left[\frac{\mathbb{E}\left[\{2p(t)-1\}^2\right]}{\sigma^2}\ ,\ \frac{1}{\sigma^2}\ ,\ \frac{2}{\sigma^2}\right] \tag{9}$$

where $\mathbb{E}\left[\{2p(t_1)-1\}^2\right] = \mathbb{E}\left[\{2p(t_2)-1\}^2\right] = 1$ and $\forall t \in]t_1, t_2[$:

$$\mathbb{E}\left[\{2p(t)-1\}^2\right] = \bar{r}(t_1,t_2)\left(2Q_t^{1,1}-1\right)^2 + r(t_1,t_2)\left(2Q_t^{1,-1}-1\right)^2 \tag{10}$$

Indeed, $\forall t \in]t_1, t_2[$:

$$\begin{aligned}\mathbb{E}\left[\{2p(t)-1\}^2\right] = {} & 2\left\{\left(Q_t^{1,1}\right)^2\ \bar{r}(t_1,t_2) + \left(Q_t^{1,-1}\right)^2\ r(t_1,t_2)\right\} \\ & + 2\left\{\left(Q_t^{-1,1}\right)^2\ r(t_1,t_2) + \left(Q_t^{-1,-1}\right)^2\ \bar{r}(t_1,t_2)\right\} - 1\end{aligned}$$

As $Q_t^{-1,1} = 1 - Q_t^{1,-1}$, $Q_t^{-1,-1} = 1 - Q_t^{1,1}$ and $\bar{r}(t_1,t_2) + r(t_1,t_2) = 1$, we obtain formula (10).
$\mathbb{E}\left[\{2p(t)-1\}^2\right]$ is always different from zero since the parameter t is bounded. It comes $\forall t \in [t_1, t_2]$:

$$\Lambda_n(t) = \left[\sum_{j=1}^{n} \frac{(y_j-\mu)\ \{2p_j(t)-1\}}{\sigma\ \sqrt{n}\ \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}}\right]^2 + o_{P_{\theta_0}}(1) \tag{11}$$

By convention, the notation $o_{P_{\theta_0}}(1)$ is short for a sequence of random vectors that converges to zero in probability under H_0 (i.e. no QTL on the whole interval studied).

Study under H_0 :
Without loss of generality, we assume that $n = 1$ for the moment and we consider the score function :

$$S(t) = \frac{(y-\mu)\ \{2p(t)-1\}}{\sigma\ \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}} = \frac{y-\mu}{\sigma}\ h(t)$$

where the fact $h(.)$ is a random process independent of y.
It is easy to see that :

$$\mathbb{E}\{S(t)\} = 0 \ , \ \mathbb{V}\{S(t)\} = \mathbb{E}\left[\{h(t)\}^2\right] = 1$$

$\forall (t,t') \in [t_1,t_2]^2$:

$$\begin{aligned}\Gamma(t,t') := \mathrm{Cov}\{S(t),\ S(t')\} = \mathbb{E}\{h(t)h(t')\} &= \frac{\mathbb{E}\left[\{2p(t)-1\}\ \{2p(t')-1\}\right]}{\sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}\sqrt{\mathbb{E}\left[\{2p(t')-1\}^2\right]}} \\ &= \frac{4\mathbb{E}\{p(t)p(t')\}-1}{\sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}\sqrt{\mathbb{E}\left[\{2p(t')-1\}^2\right]}}\end{aligned} \tag{12}$$

The formula for $\mathbb{E}\{p(t)p(t')\}$ is given in appendix 8.1. As $|p(t)p(t')| \leqslant 1$, by dominated convergence theorem, $\mathbb{E}\{p(t)p(t')\}$ is continuous at (t_1,t'), (t_2,t') and (t_1,t_2). Then the covariance function is continuous at this points (because the denominator is also continuous). So, the covariance function is a continuous function on $[t_1,t_2]^2$.
Let $S_n(.)$ be the score process for n observations :

$$S_n(t) = \sum_{j=1}^{n} \frac{(y_j-\mu)\ (2p_j(t)-1)}{\sigma\ \sqrt{n}\ \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}} \tag{13}$$

When n tends to infinity, an application of the Multivariate Central Limit Theorem shows that for $0 \leqslant s_1 < s_2 < ... < s_d \leqslant T$:

$$(S_n(s_1),...,S_n(s_d))' \xrightarrow{\mathcal{L}} N(\underset{\sim}{0},\Sigma)$$

were Σ is the variance covariance matrix, with unit variance and covariance given by formula (12). $\underset{\sim}{0}$ is a column vector of zeros. As $\Lambda_n(t) = S_n^2(t) + o_{P_{\theta_0}}(1)$:

$$(\Lambda_n(s_1),...,\Lambda_n(s_d))' \xrightarrow{\mathcal{L}} \left\{N(\underset{\sim}{0},\Sigma)\right\}^2$$

Study under $H_{at^\star}$:
In this part, we set

$$Y_j = \mu\ +\ \frac{a}{\sqrt{n}}\ X_j(t^\star)\ +\ \sigma\ \varepsilon_j \tag{14}$$

where ε_j is a Gaussian white noise. According to formula (11), $\forall t \in [t_1,t_2]$:

$$\Lambda_n(t) = \{S_n(t)\}^2\ +\ o_{P_{\theta_0}}(1)$$

We remind that $o_{P_{\theta_0}}(1)$ is short for a sequence of random vectors that converges to zero in probability under H_0 (i.e. no QTL on the whole interval studied). Let $o_{P_{\theta_{0,t^\star}}}(1)$ be a

sequence of random vectors that converges to zeros if there is no QTL at position $t^\star$. Then, it is clear that :

$$\Lambda_n(t) = \{S_n(t)\}^2 \ + \ o_{P_{\theta_{0,t^\star}}}(1)$$

Let $\theta_{a,t^\star}$ be the parameter refering that we are under $H_{at^\star}$. Under $H_{at^\star}$, as the QTL is located at position $t^\star$, the density of $Y\big|X(t_1), X(t_2)$ verifies :

$$p(t^\star) f_{(\mu+q,\sigma)}(y) + \{1 - p(t^\star)\} f_{(\mu-q,\sigma)}(y)$$

Let Q_n and P_n two sequences of probability measures defined on the same space $(\Omega_n, \ \mathcal{A}_n)$. Q_n (respectively P_n) is the law corresponding to the density $L_n(\theta_{a,t^\star}, t^\star)$ (resp $L_n(\theta_0, t^\star)$). We will call the log likelihood ratio $\log \frac{dQ_n}{dP_n}$. It verifies : $\log \frac{dQ_n}{dP_n} = \log \left\{ \frac{L_n(\theta_{a,t^\star}, \ t^\star)}{L_n(\theta_0, \ t^\star)} \right\}$.

Notations : $Q_n \triangleleft P_n$ will mean the sequence Q_n is contiguous with the respect to the sequence P_n.

Let $b = (a, \ 0, \ 0)'$. As the model is differentiable in quadratic mean at $\theta_{a,t^\star}$:

$$\log\left(\frac{dQ_n}{dP_n}\right) = \frac{b'}{\sqrt{n}} \nabla \log L_n(\theta_0, t^\star) \ - \frac{1}{2} b' I_{\theta_0} b \ + \ o_{P_{\theta_0}}(1)$$

Then, by the central limit theorem :

$$\log\left(\frac{dQ_n}{dP_n}\right) \xrightarrow[H_0]{\mathcal{L}} N(-\frac{1}{2}\nu^2, \ \nu^2) \ \text{ with } \ \nu^2 = \frac{a^2}{\sigma^2} \, \mathbb{E}\left[\{2p(t^\star) - 1\}^2\right]$$

So, by the iii) of Le Cam's first lemma, we have $Q_n \triangleleft P_n$.

Up to now $\forall t \in [t_1, t_2]$:

$$\Lambda_n(t) = \{S_n(t)\}^2 \ + \ o_{P_{\theta_{0,t^\star}}}(1)$$

As $Q_n \triangleleft P_n$, according to iv) of Le Cam's first lemma :

$$\Lambda_n(t) = \{S_n(t)\}^2 \ + \ o_{P_{\theta_{a,t^\star}}}(1)$$

So, calculations can be done with the score process. According to formula (13) and (14), we have :

$$S_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j h_j(t) \ + \ \sum_{j=1}^{n} \frac{a}{\sigma n} X_j(t^\star) h_j(t) = S_n^0(t) \ + \ \sum_{j=1}^{n} \frac{a}{\sigma n} X_j(t^\star) h_j(t)$$

where $h_j(.)$ is the equivalent of the process $h(.)$ defined above but for the individual j. $S_n^0(.)$ is the process obtained under H_0.
By the law of large number :

$$\frac{1}{n} \sum_{j=1}^{n} X_j(t^\star) h_j(t) \to \mathbb{E}\{X(t^\star) h(t)\}$$

Let suppose $K = 2$ for the moment and, for example $(t, t^\star) \in]t_1, t_2[^2$. Let us compute $\mathbb{E}\left[X(t^\star)\ \{2p(t)-1\}\right]$. We condition on $X(t_1)$ and $X(t_2)$. Consider, for example, the case $X(t_1) = X(t_2) = 1$. In this case, $p(t) = Q_t^{1,1}$ and we have :

$$\begin{aligned}
&\mathbb{E}\left[\ X(t^\star)\ \{2p(t)-1\}\ \middle|\ X(t_1) = X(t_2) = 1\right] = \mathbb{E}\left[\ X(t^\star)\ \left\{2Q_t^{1,1}-1\right\}\ \middle|\ X(t_1) = X(t_2) = 1\right] \\
&= \left\{2Q_t^{1,1}-1\right\}\ \mathbb{E}\left[\ X(t^\star)\ \middle|\ X(t_1) = X(t_2) = 1\right] \\
&= \left\{2Q_t^{1,1}-1\right\}\ \left\{\frac{\bar{r}(t_1,t^\star)\ \bar{r}(t^\star,t_2)}{\bar{r}(t_1,t_2)}\ -\ \frac{r(t_1,t^\star)r(t^\star,t_2)}{\bar{r}(t_1,t_2)}\right\} \\
&= \left\{2Q_t^{1,1}-1\right\}\ \left\{Q_{t^\star}^{1,1} - Q_{t^\star}^{-1,-1}\right\} = \left\{2Q_t^{1,1}-1\right\}\ \left\{2Q_{t^\star}^{1,1}-1\right\}
\end{aligned}$$

Considering the four cases :

$$\begin{aligned}
&\mathbb{E}\left[\ X(t^\star)\ \{2p(t)-1\}\right] \\
&= \left\{2Q_t^{1,1}-1\right\}\ \left\{2Q_{t^\star}^{1,1}-1\right\}\ \frac{1}{2}\ \bar{r}(t_1,t_2)\ +\ \left\{2Q_t^{1,-1}-1\right\}\ \left\{2Q_{t^\star}^{1,-1}-1\right\}\ \frac{1}{2}\ r(t_1,t_2) \\
&+ \left\{2Q_t^{-1,1}-1\right\}\ \left\{2Q_{t^\star}^{-1,1}-1\right\}\ \frac{1}{2}\ r(t_1,t_2)\ +\ \left\{2Q_t^{-1,-1}-1\right\}\ \left\{2Q_{t^\star}^{-1,-1}-1\right\}\ \frac{1}{2}\ \bar{r}(t_1,t_2) \\
&= \bar{r}(t_1,t_2)\ \left\{2Q_{t^\star}^{1,1}-1\right\}\ \left\{2Q_t^{1,1}-1\right\} \\
&+ r(t_1,t_2)\ \left\{2Q_{t^\star}^{1,-1}-1\right\}\ \left\{2Q_t^{1,-1}-1\right\} \qquad (15)
\end{aligned}$$

According to dominated convergence theorem, $\mathbb{E}\left[\ X(t^\star)\ \{2p(t)-1\}\right]$ is continuous on $[t_1,t_2]^2$. As a conclusion, $\forall (t,t^\star) \in [t_1,t_2]^2$:

$$m_{t^\star}(t) = \frac{a\ \mathbb{E}\left[X(t^\star)\ \{2p(t)-1\}\right]}{\sigma\ \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}}$$

A non linear interpolation

After some easy calculations, we can remark that :

$$S_n(t) = \{\ \alpha(t)\ S_n(t_1)\ +\ \beta(t)\ S_n(t_2)\ \}\ /\sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

where $\mathrm{Cov}\left\{S_n^0(t_1), S_n^0(t_2)\right\} = e^{-2t_2}$, $\alpha(t_1) = 1$, $\beta(t_1) = 0$, $\alpha(t_2) = 0$, $\beta(t_2) = 1$ and $\forall t \in]t_1,t_2[$:

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

And it comes :

$$m_{t^\star}(t) = \{\ \alpha(t)\ m_{t^\star}(t_1)\ +\ \beta(t)\ m_{t^\star}(t_2)\ \}\ /\ \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

Weak convergence of the score process

As $p(t)$ and $\mathbb{E}\left[\{2p(t)-1\}^2\right]$ are continuous functions, each trajectory of the process $S_n(.)$ is

a continuous function on $[0,T]$. Let define the modulus of continuity of a continous function x on $[0,T]$:

$$w_x(\delta) = sup_{|t'-t|<\delta} |x(t') - x(t)| \quad \text{where} \quad 0 < \delta \leqslant T$$

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

(a) the sequence $S_n(0)$ is tight.
(b) For each positive ϵ and η, there exist a δ, with $0 < \delta < T$, and an integer n_0 such that $\mathbb{P}\{w_{S_n}(\delta) \geqslant \eta\} \leqslant \epsilon \quad \forall n \geqslant n_0$.

According to Prohorov, the sequence $S_n(0)$ is tight. So, a) is verified.
Let define the functions $\tilde{\alpha}(t)$ and $\tilde{\beta}(t)$ such as :

$$\tilde{\alpha}(t) = \alpha(t)/\sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]} \quad , \quad \tilde{\beta}(t) = \beta(t)/\sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

First, we can remark that $\forall \delta$ such as $0 < \delta \leqslant T$:

$$\begin{aligned} w_{S_n}(\delta) &= sup_{|t'-t|<\delta} |S_n(t') - S_n(t)| \\ &= sup_{|t'-t|<\delta} \left| \{\tilde{\alpha}(t') - \tilde{\alpha}(t)\} S_n(t_1) \; + \; \left\{\tilde{\beta}(t') - \tilde{\beta}(t)\right\} S_n(t_2) \right| \\ &\leqslant max\{|S_n(t_1)|, |S_n(t_2)|\} \quad \left\{w_{\tilde{\alpha}}(\delta) + w_{\tilde{\beta}}(\delta)\right\} \end{aligned}$$

Let $\epsilon > 0$ and $\eta > 0$, as the sequence $max\{|S_n(t_1)|, |S_n(t_2)|\}$ is uniformly tight, $\exists M$ such as $\forall n \geqslant 1 \quad \mathbb{P}[\, max\{|S_n(t_1)|, |S_n(t_2)|\} \geqslant M \,] \leqslant \epsilon$.
It comes, $\mathbb{P}\left[\, max\{|S_n(t_1)|, |S_n(t_2)|\} \left\{w_{\tilde{\alpha}}(\delta) + w_{\tilde{\beta}}(\delta)\right\} \geqslant M \left\{w_{\tilde{\alpha}}(\delta) + w_{\tilde{\beta}}(\delta)\right\} \,\right] \leqslant \epsilon$
As $\tilde{\alpha}(t)$ and $\tilde{\beta}(t)$ are continuous on the compact $[0,T]$, according to Heine's theorem, these functions are uniformly continuous. So, let $\upsilon > 0$, $\exists \delta_1$ with $0 < \delta_1 < T$, such as $w_{\tilde{\alpha}}(\delta_1) < \upsilon/2$ and $\exists \delta_2$ with $0 < \delta_2 < T$ such as $w_{\tilde{\beta}}(\delta_2) < \upsilon/2$. Let $\delta = min(\delta_1, \delta_2)$ then $w_{\tilde{\alpha}}(\delta) + w_{\tilde{\beta}}(\delta) < \upsilon$. If we impose $\upsilon = \eta/M$, then $\forall n \geqslant 1$, $\mathbb{P}\{w_{S_n}(\delta) \geqslant \eta\} \leqslant \epsilon$ which means b) of theorem 8.2 of Billingsley (1999) is fulfilled. So, 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.

7.2. Proof of theorem 3

We first compute the score functions and the Fisher Information matrix. Let $t \in [t_1, t_K] \backslash \mathbb{T}_k$:

$$\frac{\partial logL}{\partial q_i} \,|_{\theta_0} = \frac{y - \mu_i}{\sigma^2} \; \{2p(t) - 1\} \; 1_{C=i} \quad , \quad \frac{\partial logL}{\partial \mu_i} \,|_{\theta_0} = \frac{y - \mu_i}{\sigma^2} \; 1_{C=i}$$

$$\frac{\partial logL}{\partial \sigma} \,|_{\theta_0} = -\frac{1}{\sigma} \; + \; \sum_{i=1}^{I} \frac{(y - \mu_i)^2}{\sigma^3} \; 1_{C=i}$$

$$I_{\theta_0} = Diag\left[\frac{\pi_1}{\sigma^2}\ \mathbb{E}\left[\{2p(t)-1\}^2\right],\ ...,\ \frac{\pi_I}{\sigma^2}\ \mathbb{E}\left[\{2p(t)-1\}^2\right],\ \frac{\pi_1}{\sigma^2}\ ,\ ...\ ,\ \frac{\pi_I}{\sigma^2}\ ,\ \frac{2}{\sigma^2}\right]$$

Remarks : If the π_i's were unknown, after some easy calculations, we find that the Fisher information matrix would still be diagonal, it would be the same as above, and the diagonal terms concerning π_i would be equal to $\frac{1}{\pi_i}$. So, the results established further are also true for all the π_i's unknown.

It comes $\forall t \in [t_1, t_K]\backslash\mathbb{T}_k$:

$$\Lambda_n(t) = \sum_{i=1}^{I}\left[\sum_{j=1}^{n}\frac{(y_j-\mu_i)\ \{2p_j(t)-1\}}{\sqrt{n\ \pi_i}\ \sigma\ \sqrt{\mathbb{E}\left\{\{2p(t)-1\}^2\right\}}}\ 1_{C_j=i}\right]^2\ +\ o_{P_{\theta_0}}(1) \tag{16}$$

Study under H_0 :
Without loss of generality, we assume $n = 1$ for the moment and we consider the score test statistic refering to the test of the presence of a QTL in family i :

$$S(t,i) = \frac{(y-\mu_i)\ \{2p(t)-1\}}{\sqrt{\pi_i}\ \sigma\ \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}}\ 1_{C=i} = \frac{y-\mu_i}{\sigma\ \sqrt{\pi_i}}\ 1_{C=i}\ \ h(t)$$

where $h(.)$ is a random process (the same as in the proof of theorem 1), independent of y and C. It is easy to see that :

$$\mathbb{E}\{S(t,i)\} = 0\ \ ,\ \ \mathbb{V}\{S(t,i)\} = \mathbb{E}\left[\{h(t)\}^2\right] = 1$$

$\forall (t,t') \in [t_1,t_K]\backslash\mathbb{T}_k \times [t_1,t_K]\backslash\mathbb{T}_k$:

$$\mathrm{Cov}\{S(t,i),\ S(t',i)\} = \mathbb{E}\{h(t)h(t')\} = \Gamma(t,t')$$

This function $\Gamma(t,t')$ is the same as in theorem 2.
Let $S_n(.,i)$ be the score process for n observations, related to testing the presence of a QTL in family i on $[0,T]$:

$$S_n(t,i) = \sum_{j=1}^{n}\frac{(y_j-\mu_i)\ \{2p_j(t)-1\}}{\sqrt{n\ \pi_i}\ \sigma\ \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}}\ 1_{C_j=i} \tag{17}$$

When n tends to infinity, an application of the Multivariate Central Limit Theorem shows that for $0 \leqslant s_1 < s_2 < ... < s_d \leqslant T$:

$$(S_n(s_1,i),...,S_n(s_d,i))' \xrightarrow{\mathcal{L}} N(\underset{\sim}{0},\Sigma)$$

were Σ is the variance covariance matrix, with unit variance and covariance given by the function $\Gamma(t,t')$. $\underset{\sim}{0}$ is a column vector of zeros.

According to formula (16), we have $\Lambda_n(t) = \sum_{i=1}^{I}\{S_n(t,i)\}^2 + o_{P_{\theta_0}}(1)$. It comes :

$$(\Lambda_n(s_1),...,\Lambda_n(s_d))' \xrightarrow{\mathcal{L}} \sum_{i=1}^{I}\left\{N(\underset{\sim}{0},\Sigma)\right\}^2$$

Study under $H_{at^\star}$:
In this part, we set

$$Y_j = \mu_i \ + \ \frac{a_i}{\sqrt{n}} \, X_j(t^\star) \ + \ \sigma \, \varepsilon_j \tag{18}$$

where ε_j is a Gaussian white noise.
It is clear that we have also :

$$\Lambda_n(t) = \sum_{i=1}^{I} \{S_n(t,i)\}^2 \ + \ o_{P_{\theta_{0,t^\star}}}(1) \tag{19}$$

Under $H_{at^\star}$, as the QTL is located at position $t^\star$, the density of $Y \big| X(t^\ell), X(t^r), C = i$ verifies :

$$p(t^\star) f_{(\mu_i + q_i, \sigma)}(y) + \{1 - p(t^\star)\} \, f_{(\mu_i - q_i, \sigma)}(y)$$

By the central limit theorem :

$$log\left(\frac{dQ_n}{dP_n}\right) \xrightarrow[H_0]{\mathcal{L}} N(-\frac{1}{2}\nu^2, \ \nu^2) \ \text{ with } \ \nu^2 = \sum_{i=1}^{I} \frac{a_i^2 \ \pi_i}{\sigma^2} \, \mathbb{E}\left[\{2p(t^\star) - 1\}^2\right]$$

By the iii) of Le Cam's first lemma, we have $Q_n \lhd P_n$. According to iv) of Le Cam's first lemma and formula (19) :

$$\Lambda_n(t) = \sum_{i=1}^{I} \{S_n(t,i)\}^2 \ + \ o_{P_{\theta_{a,t^\star}}}(1)$$

So, calculations can be done with the score process. According to formula (17) and (18), we have :

$$\begin{aligned} S_n(t,i) &= \frac{1}{\sqrt{n \ \pi_i}} \sum_{j=1}^{n} \varepsilon_j \ 1_{C_j = i} \ h_j(t) \ + \ \sum_{j=1}^{n} \frac{a_i}{n \ \sigma \ \sqrt{\pi_i}} \ 1_{C_j = i} \ X_j(t^\star) \ h_j(t) \\ &= S_n^0(t,i) \ + \ \sum_{j=1}^{n} \frac{a_i}{n \ \sigma \ \sqrt{\pi_i}} \ 1_{C_j = i} \ X_j(t^\star) \ h_j(t) \end{aligned}$$

where $S_n^0(.,i)$ is the process obtained under H_0. We remind that $h_j(.)$ is the equivalent of the process $h(.)$ for the individual j. By the law of large number :

$$\frac{1}{n} \sum_{j=1}^{n} X_j(t^\star) \ h_j(t) \ 1_{C_j = i} \to \pi_i \, \mathbb{E}\{X(t^\star) h(t)\}$$

It comes : $\forall (t, t^\star) \in [t_1, t_K] \backslash \mathbb{T}_k \times [t_1, t_K] \backslash \mathbb{T}_k$:

$$m_{t^\star}^i(t) = \frac{a_i \ \sqrt{\pi_i} \ \mathbb{E}\left[X(t^\star)\{2p(t) - 1\}\right]}{\sigma \ \sqrt{\mathbb{E}\left[\{2p(t) - 1\}^2\right]}}$$

A non linear interpolation :
We can easily remark that $\forall t \in [t_1, t_K] \backslash \mathbb{T}_k$:

$$S_n(t,i) = \alpha(t)\ S_n(t^\ell, i)\ +\ \beta(t)\ S_n(t^r, i) / \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

where $\alpha(t) = Q_t^{1,1} + Q_t^{1,-1} - 1$, $\beta(t) = Q_t^{1,1} - Q_t^{1,-1}$ and $\forall k\ \forall k'$, Cov $\left\{S_n^0(t_k, i), S_n^0(t_{k'}, i)\right\} = e^{-2|t_k - t_{k'}|}$.
It comes $\forall (t, t^\star) \in [t_1, t_K] \backslash \mathbb{T}_k \times [t_1, t_K] \backslash \mathbb{T}_k$:

$$m_{t^\star}^i(t) = \left\{\ \alpha(t)\ m_{t^\star}^i(t^\ell)\ +\ \beta(t)\ m_{t^\star}^i(t^r)\ \right\} / \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

7.3. Proof of theorem 4

Study under H_0 :
$\Lambda_n(.)$ is the same process as in theorem 2.

Study under $H_{a\vec{t}^\star}$:
In this part, we set

$$Y_j = \mu\ +\ \sum_{s=1}^{m} X_j(t_s^\star)\ q^s\ +\ \sigma\varepsilon_j \tag{20}$$

where ε_j is a Gaussian white noise.
Let's introduce some notations :

- ξ : number of "Marker intervals" which contain the QTL.
  $\gamma = 1, ..., \xi$ will refer to the different intervals.
- m_γ : number of QTL in the interval γ.
  $\tau = 1, ..., m_\gamma$ refers to the τth QTL in the interval γ.
- the sth QTL on $[0, T]$, can be rewritten, $s = (\tau, \gamma) = \left\{\sum_{i=1}^{\gamma-1} m_i\right\}\ +\ \tau$

Let $\theta_{a,\vec{t}^\star} = \left(q^1,\ ...,\ q^m,\ \mu,\ \sigma\right)$ and $\theta_{0,\vec{t}^\star} = (0,\ ...,\ 0,\ \mu,\ \sigma)$.
After some calculations, the likelihood of $\left(Y,\ X\left\{t_{(1,1)}^{\star\ell}\right\},\ X\left\{t_{(1,1)}^{\star r}\right\},\ ...,\ X\left\{t_{(1,\xi)}^{\star\ell}\right\},\ X\left\{t_{(1,\xi)}^{\star r}\right\}\right)$ with respect to the measure $\lambda \otimes N \otimes ... \otimes N$, λ being the Lebesgue measure, N the county measure on $\mathbb{N}$, verifies :

$$L^\star(\theta_{a,\vec{t}^\star}) = \sum_{(u_1,...,u_m)\in\{-1,1\}^m} f_{(\mu+u_1q^1+...+u_mq^m,\sigma)}(y)$$
$$\times \left\{ \left(\prod_{\gamma=1}^{\xi} A\left\{t_{(\tau,\gamma)}^{\star\ell},\ t_{(\tau,\gamma)}^\star\right\} \left[\prod_{\tau=1}^{m_\gamma-1} R\left\{t_{(\tau,\gamma)}^\star,\ t_{(\tau+1,\gamma)}^\star\right\}\right] A\left\{t_{(m_\gamma,\gamma)}^{\star r},\ t_{(m_\gamma,\gamma)}^\star\right\} \right) g^\star(\vec{t}^\star) \right\}$$

where

$$
\begin{aligned}
u_s &= u_{(\tau,\gamma)} \\
A\left\{t\ ,\ t^\star_{(\tau,\gamma)}\right\} &= r\left\{t\ ,\ t^\star_{(\tau,\gamma)}\right\}\ 1_{X(t)u(\tau,\gamma)=-1}\ +\ \bar{r}\left\{t\ ,\ t^\star_{(\tau,\gamma)}\right\}\ 1_{X(t)u(\tau,\gamma)=1} \\
R\left\{t^\star_{(\tau,\gamma)}\ ,\ t^\star_{(\tau+1,\gamma)}\right\} &= \bar{r}\left\{t^\star_{(\tau,\gamma)}\ ,\ t^\star_{(\tau+1,\gamma)}\right\}\ 1_{u(\tau,\gamma)u(\tau+1,\gamma)=1} \\
&\quad + r\left\{t^\star_{(\tau,\gamma)}\ ,\ t^\star_{(\tau+1,\gamma)}\right\}\ 1_{u(\tau,\gamma)u(\tau+1,\gamma)=-1} \\
g^\star(\vec{t}^\star) &= \frac{1}{2}\ \prod_{\gamma=1}^{\xi-1}\ D\left\{t^{\star r}_{(m_\gamma,\gamma)}, t^{\star \ell}_{(1,\gamma+1)}\right\} \\
D(t,t') &= \bar{r}(t,t')\ 1_{X(t)X(t')=1}\ +\ r(t,t')\ 1_{X(t)X(t')=-1}
\end{aligned}
$$

The likelihood $L^\star_n(\theta_{a,\vec{t}^\star})$ for n observations is obtained by the product of n terms as above. Let Q_n and P_n two sequences of probability measures defined on the same space $(\Omega_n,\ \mathcal{A}_n)$. Q_n (respectively P_n) is the law corresponding to the density $L^\star_n(\theta_{a,\vec{t}^\star})$ (resp $L^\star_n(\theta_{0,\vec{t}^\star})$). We will call the log likelihood ratio $\log \frac{dQ_n}{dP_n}$. It verifies : $\log \frac{dQ_n}{dP_n} = \log\left\{\frac{L^\star_n(\theta_{a,\vec{t}^\star})}{L^\star_n(\theta_{0,\vec{t}^\star})}\right\}$.
As the model is differentiable in quadratic mean at $\theta_{a,\vec{t}^\star}$ and according to the central limit theorem :

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

By the iii) of Le Cam's first lemma, we have $Q_n \triangleleft P_n$.
We remind the notation $o_{P_{\theta_0}}(1)$ used in the proof of theorem 1 : $o_{P_{\theta_0}}(1)$ is short for a sequence of random vectors that converges to zeros in probability under H_0 (i.e. no QTL on the whole interval studied).
Besides, we remind the score test statistic for n observations. $\forall t \in [t_1, t_K]\backslash\mathbb{T}_k$, we have :

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

According to the proof of theorem 1 :

$$\Lambda_n(t) = \{S_n(t)\}^2\ +\ o_{P_{\theta_0}}(1)$$

Let $o_{P_{\theta_{0,\vec{t}^\star}}}(1)$ be a sequence of random vectors that converges to zeros if there is no QTL at $t^\star_1$, ..., $t^\star_m$. Then, it is clear that :

$$\Lambda_n(t) = \{S_n(t)\}^2\ +\ o_{P_{\theta_{0,\vec{t}^\star}}}(1)$$

As $Q_n \triangleleft P_n$, according to iv) of Le Cam's first lemma :

$$\Lambda_n(t) = \{S_n(t)\}^2\ +\ o_{P_{\theta_{a,\vec{t}^\star}}}(1)$$

So, calculations can be done with the score process.
According to formula (20) :

$$S_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j h_j(t) \; + \; \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} X_j(t_s^\star) \; a^s \right\} \frac{h_j(t)}{\sigma n}$$
$$= S_n^0(t) \; + \; \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} X_j(t_s^\star) \; a^s \right\} \frac{h_j(t)}{\sigma n}$$

where $h_j(t)$ is the same function as in the proof of theorem 3.
By the law of large number :

$$\frac{1}{n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} X_j(t_s^\star) \; a^s \right\} \; h_j(t) \to \mathbb{E} \left[\left\{ \sum_{s=1}^{m} X(t_s^\star) \; a^s \right\} h(t) \right]$$

We have :

$$\mathbb{E} \left[\left\{ \sum_{s=1}^{m} X(t_s^\star) \; a^s \right\} h(t) \right] = \sum_{s=1}^{m} \frac{a^s \; \mathbb{E} \left[X(t_s^\star) \; \{2p(t) - 1\} \right]}{\sqrt{\mathbb{E} \left[\{2p(t) - 1\}^2 \right]}}$$

According to formula (15) in Section 7.1 and formula (22) in appendix 8.2 :

$$\begin{aligned} &\mathbb{E} \left[X(t_s^\star) \; \{2p(t) - 1\} \right] \\ &= \left\{ \bar{r}(t^\ell, t^r) \left(2Q_{t_s^\star}^{1,1} - 1 \right) \left(2Q_t^{1,1} - 1 \right) \; + \; r(t^\ell, t^r) \left(2Q_{t_s^\star}^{1,-1} - 1 \right) \left(2Q_t^{1,-1} - 1 \right) \right\} \; 1_{t_s^\star \in]t^\ell, t^r[} \\ &+ e^{-2|t - t_s^\star|} \; 1_{t_s^\star \notin]t^\ell, t^r[} \end{aligned} \tag{21}$$

It comes $\forall (t, t_1^\star, ..., t_m^\star) \in \{[t_1, t_K] \backslash \mathbb{T}_k\}^{m+1}$:

$$m_{\vec{t}^\star}(t) = \sum_{s=1}^{m} \frac{a^s \; \mathbb{E} \left[X(t_s^\star) \; \{2p(t) - 1\} \right]}{\sigma \sqrt{\mathbb{E} \left[\{2p(t) - 1\}^2 \right]}}$$

A non linear interpolation :
The process $S_n(.)$ is a non linear interpolation according to formula (23) of appendix 8.2.
It comes $\forall (t, t_1^\star, ..., t_m^\star) \in \{[t_1, t_K] \backslash \mathbb{T}_k\}^{m+1}$:

$$m_{\vec{t}^\star}(t) = \sum_{s=1}^{m} \frac{a^s}{\sigma} \left\{ \; \alpha(t) \; \mathbb{E} \left[X(t_s^\star) \; \left\{ 2p(t^\ell) - 1 \right\} \right] \; + \; \beta(t) \; \mathbb{E} \left[X(t_s^\star) \; \{2p(t^r) - 1\} \right] \; \right\} / \sqrt{\mathbb{E} \left[\{2p(t) - 1\}^2 \right]}$$

Weak convergence of the score process :
Concerning the weak convergence of the score process, the proof is the same as in the proof of theorem 2 in appendix 8.2.

8. Appendix

8.1. Formula for $\mathbb{E}\{p(t)p(t')\}$

$\forall (t,t') \in]t_1,t_2[^2$:

$$\mathbb{E}\{p(t)p(t')\} = \frac{1}{2}\left\{Q_t^{1,1}Q_{t'}^{1,1}\ \bar{r}(t_1,t_2) + Q_t^{1,-1}Q_{t'}^{1,-1}\ r(t_1,t_2)\right\} + \frac{1}{2}\left\{Q_t^{-1,1}Q_{t'}^{-1,1}\ r(t_1,t_2) + Q_t^{-1,-1}Q_{t'}^{-1,-1}\ \bar{r}(t_1,t_2)\right\}$$

This quantity is continuous at t_1 and t_2 (cf. proof of theorem 1 in Section 7.1)

8.2. Sketch of the proof of theorem 2

Let $t \in [t_1,t_K]\backslash\mathbb{T}_k$. As t belongs to the "Marker interval" (t^ℓ,t^r), some adjustments with Section 3 have to be done : t_1 becomes t^ℓ and t_2 becomes t^r. So, $p(t)$ is now the quantity equal to $\mathbb{P}\left\{X(t)=1\middle|X(t^\ell),X(t^r)\right\}$. In the same way, $p(t)$, $Q_t^{1,1}$, $Q_t^{1,-1}$, $Q_t^{-1,1}$ and $Q_t^{-1,-1}$ described in formula (2) have to be adapted to the "Marker interval". The likelihood presented in formula (3), is unchanged except that the focus is on the triplet $\left(Y,\ X(t^\ell),\ X(t^r)\right)$ and the function $g(t)$ has to be adapted to the "Marker interval". Formula (11) of Section 7.1 is also suitable $t \in [t_1,t_K]\backslash\mathbb{T}_k$ because t is bounded. It comes, $\forall (t,t') \in [t_1,t_K]\backslash\mathbb{T}_k \times [t_1,t_K]\backslash\mathbb{T}_k$:

$$\Gamma(t,t') = \frac{4\mathbb{E}\{p(t)p(t')\} - 1}{\sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}\sqrt{\mathbb{E}\left[\{2p(t')-1\}^2\right]}}$$

$\mathbb{E}\left[\{2p(t)-1\}^2\right]$ described in formula (10) of Section 7.1 has to be adapted to the "Marker interval".

$\forall (t,t') \in]t^\ell,t^r[^2$, the expression of $\mathbb{E}\{p(t)p(t')\}$ can be deduced from appendix 8.1 by adapting to the "Marker interval".

Besides, if $(t,t') \in]t^\ell,t^r[\times [t^r,t_K]\backslash\mathbb{T}_k$:

$$\begin{aligned}
&\mathbb{E}\{p(t)p(t')\} \\
&= \frac{1}{2}\bar{r}(t^\ell,t^r)\left[Q_{t'}^{1,1}\bar{r}\left\{(t')^\ell,(t')^r\right\} + Q_{t'}^{1,-1}r\left\{(t')^\ell,(t')^r\right\}\right]\left[Q_t^{1,1}\bar{r}\left\{t^r,(t')^\ell\right\} + Q_t^{-1,-1}r\left\{t^r,(t')^\ell\right\}\right] \\
&+ \frac{1}{2}\bar{r}(t^\ell,t^r)\left[Q_{t'}^{-1,1}r\left\{(t')^\ell,(t')^r\right\} + Q_{t'}^{-1,-1}\bar{r}\left\{(t')^\ell,(t')^r\right\}\right]\left[Q_t^{1,1}r\left\{t^r,(t')^\ell\right\} + Q_t^{-1,-1}\bar{r}\left\{t^r,(t')^\ell\right\}\right] \\
&+ \frac{1}{2}r(t^\ell,t^r)\left[Q_{t'}^{1,1}\bar{r}\left\{(t')^\ell,(t')^r\right\} + Q_{t'}^{1,-1}\ r\left\{(t')^\ell,(t')^r\right\}\right]\left[Q_t^{1,-1}r\left\{t^r,(t')^\ell\right\} + Q_t^{-1,1}\bar{r}\left\{t^r,(t')^\ell\right\}\right] \\
&+ \frac{1}{2}r(t^\ell,t^r)\left[Q_{t'}^{-1,1}r\left\{(t')^\ell,(t')^r\right\} + Q_{t'}^{-1,-1}\bar{r}\left\{(t')^\ell,(t')^r\right\}\right]\left[Q_t^{1,-1}\bar{r}\left\{t^r,(t')^\ell\right\} + Q_t^{-1,1}r\left\{t^r,(t')^\ell\right\}\right]
\end{aligned}$$

In the same way as what has been done in the proof of theorem 1 (cf. Section 7.1), $\forall (t,t^\star) \in [t_1,t_K]\backslash\mathbb{T}_k \times [t_1,t_K]\backslash\mathbb{T}_k$:

$$m_{t^\star}(t) = \frac{a\,\mathbb{E}\left[X(t^\star)\{2p(t)-1\}\right]}{\sigma\sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}}$$

If $(t,t^\star) \in]t^\ell,t^r[^2$, then $\mathbb{E}\left[X(t^\star)\left\{2p(t)-1\right\}\right]$ has the same expression as in formula (15) of Section 7.1 provided that we adapt to the "Marker interval".
Besides, if $(t,t^\star) \in]t^\ell,t^r[\times [t^r,t_K]\backslash\mathbb{T}_k$:

$$\begin{aligned}
&\mathbb{E}\left[X(t^\star)\left\{2p(t)-1\right\}\right] \\
&= 2\,Q_t^{1,1}\,\mathbb{E}\left\{X(t^\star)1_{X(t^\ell)=1}1_{X(t^r)=1}\right\} \;+\; 2\,Q_t^{1,-1}\,\mathbb{E}\left\{X(t^\star)1_{X(t^\ell)=1}1_{X(t^r)=-1}\right\} \\
&+ 2\,Q_t^{-1,1}\,\mathbb{E}\left\{X(t^\star)1_{X(t^\ell)=-1}1_{X(t^r)=1}\right\} \;+\; 2\,Q_t^{-1,-1}\,\mathbb{E}\left\{X(t^\star)1_{X(t^\ell)=-1}1_{X(t^r)=-1}\right\} \\
&= \bar{r}(t^\ell,t)\,\bar{r}(t,t^r)\;\left\{1-2r(t^r,t^\star)\right\} \;+\; \bar{r}(t^\ell,t)\,r(t,t^r)\;\left\{2r(t^r,t^\star)-1\right\} \\
&+ r(t^\ell,t)\,\bar{r}(t,t^r)\;\left\{1-2r(t^r,t^\star)\right\} \;+\; r(t^\ell,t)\,r(t,t^r)\;\left\{2r(t^r,t^\star)-1\right\} \\
&= \left\{1-2r(t,t^r)\right\}\;\left\{1-2r(t^r,t^\star)\right\} = e^{-2(t^\star-t)}
\end{aligned}$$

As we deal with Poisson processes, it is reversible. So, If $(t,t^\star) \in\; [t^{\star r},t_K]\backslash\mathbb{T}_k\times\,]t^{\star\ell},t^{\star r}[$:

$$\mathbb{E}\left[X(t^\star)\left\{2p(t)-1\right\}\right] = \left\{1-2r(t^\star,t^\ell)\right\}\;\left\{1-2r(t^\ell,t)\right\} = e^{-2(t-t^\star)}$$

So, if t and $t^\star$ do not belong to the same "Marker interval" :

$$\mathbb{E}\left[X(t^\star)\left\{2p(t)-1\right\}\right] = e^{-2|t-t^\star|} \tag{22}$$

A non linear interpolation
Concerning the non linear interpolation, we have to adapt formula (5) of Section 3.1 to the "Marker interval". $\forall t \in [t_1,t_K]\backslash\mathbb{T}_k$ we have :

$$S_n(t) = \left\{\;\alpha(t)\;S_n(t^\ell)\;+\;\beta(t)\;S_n(t^r)\;\right\}/\sqrt{\mathbb{E}\left[\left\{2p(t)-1\right\}^2\right]} \tag{23}$$

where $\alpha(t) = Q_t^{1,1}+Q_t^{1,-1}-1$, $\beta(t) = Q_t^{1,1}-Q_t^{1,-1}$ and $\forall k\;\forall k'$, $\mathrm{Cov}\left\{S_n^0(t_k),S_n^0(t_{k'})\right\} = e^{-2|t_k-t_{k'}|}$.
It comes $\forall (t,t^\star) \in [t_1,t_K]\backslash\mathbb{T}_k \times [t_1,t_K]\backslash\mathbb{T}_k$:

$$m_{t^\star}(t) = \left\{\;\alpha(t)\;m_{t^\star}(t^\ell)\;+\;\beta(t)\;m_{t^\star}(t^r)\;\right\}/\;\sqrt{\mathbb{E}\left[\left\{2p(t)-1\right\}^2\right]}$$

Weak convergence of the score process
Each trajectory of the process $S_n(.)$ is a continuous function on $[0,T]$. In the same way as in the proof of theorem 1 in Section 7.1, in order to prove the tightness of the score process, we have to verify that conditions a) and b) of theorem 8.2 of Billingsley (1999) are fulfilled. According to Prohorov, $S_n(0)$ is tight, so a) is fulfilled.
We remind the modulus of continuity of $S_n(t)$:

$$w_{S_n}(\delta) = sup_{|t'-t|<\delta}\,|S_n(t')-S_n(t)| \quad \text{where} \quad 0<\delta\leqslant T$$

Let define $w_{S_n}^k(\delta)$, the modulus of continuity of $S_n(t)$ only between the markers k and $k+1$:

$$w_{S_n}^k(\delta) = sup_{|t'-t|<\delta}\,|S_n(t'+t_k)-S_n(t+t_k)| \quad \text{where} \quad 0<\delta\leqslant t_{k+1}-t_k$$

As the score process is tight when there are only two markers (cf. proof of theorem 1), according to b) of theorem 8.2 of Billingsley (1999), we have for a given k:

$$\forall \epsilon > 0 \; \forall \eta > 0 \; \exists \delta_k \;\text{ with }\; 0 < \delta_k < t_{k+1} - t_k \;\text{ such that }\; \mathbb{P}\left\{ w^k_{S_n}(\delta_k) \geqslant \eta \right\} \leqslant \epsilon$$

So, let $\epsilon > 0$, $\epsilon' = \epsilon/(K-1)$, $\eta > 0$ and we impose $\delta = \min_{k \in \{1,...,K-1\}}(\delta_k)$
then $\forall k \in \{1, ..., K-1\}$ $\mathbb{P}\left\{ w^k_{S_n}(\delta) \geqslant \eta \right\} \leqslant \epsilon'$.
As $w_{S_n}(\delta) \geqslant w^1_{S_n}(\delta) + ... + w^{K-1}_{S_n}(\delta)$, then $\mathbb{P}\left\{ w_{S_n}(\delta) \geqslant \eta \right\} \leqslant \sum_{k=1}^{K-1} \mathbb{P}\left\{ w^k_{S_n}(\delta) \geqslant \eta \right\} \leqslant \epsilon$ which means b) of theorem 8.2 of Billingsley (1999) is fulfilled. So, 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.

8.3. Linear interpolated process under Interval Mapping

In presence of several markers, the linear interpolated process $V_n(.)$ is such as $\forall t \in [t_1, t_K] \backslash \mathbb{T}_k$:

$$\begin{aligned} V_n(t) &= \left\{ \frac{t^r - t}{t^r - t^\ell} \, S_n(t^\ell) \; + \; \frac{t - t^\ell}{t^r - t^\ell} \, S_n(t^r) \right\} / \sqrt{\tau(t)} \\ &= \left\{ \frac{t^r - t}{t^r - t^\ell} \, W_n(t^\ell) \; + \; \frac{t - t^\ell}{t^r - t^\ell} \, W_n(t^r) \right\} / \sqrt{\tau(t)} \; + \; o_{P_{\theta_0}}(1) \end{aligned}$$

where

$$\tau(t) = \left(\frac{t^r - t}{t^r - t^\ell} \right)^2 \; + \; 2 \, \frac{(t^r - t)(t - t^\ell)}{(t^r - t^\ell)^2} \, e^{-2(t^r - t^\ell)} \; + \; \left(\frac{t - t^\ell}{t^r - t^\ell} \right)^2$$

It can be seen easily that $\tau(t) \neq 0, \;\; \forall t \in [t_1, t_K] \backslash \mathbb{T}_k$.
$V_n(.)$ remains asymptotically a Gaussian process with mean equal to 0 under H_0, unit variance, and $\forall k \, \forall k'$, $\mathrm{Cov}\left\{ S^0_n(t_k), S^0_n(t_{k'}) \right\} = e^{-2|t_k - t_{k'}|}$. In the same way as what has been done in Section 3.2, the weights of the model of mixture corresponding to this process verify :

$$p(t) = 1_{X(t^\ell)=1} 1_{X(t^r)=1} \; + \; \frac{t^r - t}{t^r - t^\ell} \, 1_{X(t^\ell)=1} 1_{X(t^r)=-1} \; + \; \frac{t - t^\ell}{t^r - t^\ell} \, 1_{X(t^\ell)=-1} 1_{X(t^r)=1}$$

This weights are an approximation at the first order of the original weights. So, the linear interpolated process will be a good approximation if and only if the genetic markers are close to each other. This process $V_n(.)$ is a generalization of the process studied, under H_0, by Rebaï et al. (1994). By contiguity (in the same way of what has been done in Section 7.1), under $H_{at^\star}$, $V_n(.)$ is asymptotically the same process as under H_0 on which the mean function, $\tilde{m}_{t^\star}(t)$, has been added :

$$\tilde{m}_{t^\star}(t) = \left\{ \frac{t^r - t}{t^r - t^\ell} \, m_{t^\star}(t^\ell) \; + \; \frac{t - t^\ell}{t^r - t^\ell} \, m_{t^\star}(t^r) \right\} / \sqrt{\tau(t)}$$

8.4. Linear interpolated process and kriging process in presence of several families

Let $V_n(.,i)$ be the linear interpolated process for family i. $\forall t \in [t_1, t_K] \backslash \mathbb{T}_k$:

$$\begin{aligned} V_n(t,i) &= \left(\frac{t^r - t}{t^r - t^\ell} \, S_n(t^\ell, i) \; + \; \frac{t - t^\ell}{t^r - t^\ell} \, S_n(t^r, i) \right) / \sqrt{\tau(t)} \\ &= \left(\frac{t^r - t}{t^r - t^\ell} \, W_n(t^\ell, i) \; + \; \frac{t - t^\ell}{t^r - t^\ell} \, W_n(t^r, i) \right) / \sqrt{\tau(t)} \; + \; o_{P_{\theta_0}}(1) \end{aligned}$$

where $\tau(t)$ has the same expression as in appendix 8.3 and $\forall k \; \forall k'$, $\text{Cov}\left\{S_n^0(t_k, i), S_n^0(t_{k'}, i)\right\} = e^{-2|t_k - t_{k'}|}$.
$\sum_{i=1}^{I} \{V_n(.,i)\}^2$ is an approximation of the process $\Lambda_n(.)$ provided that the genetic markers are close to each other.
By contiguity (cf. Section 7.2), under $H_{at^\star}$, $V_n(.,i)$ is the same process as under H_0 on which the mean function, $\tilde{m}_{t^\star}^i(t)$, has been added :

$$\tilde{m}_{t^\star}^i(t) = \left\{ \frac{t^r - t}{t^r - t^\ell} \, m_{t^\star}^i(t^\ell) \; + \; \frac{t - t^\ell}{t^r - t^\ell} \, m_{t^\star}^i(t^r) \right\} / \sqrt{\tau(t)}$$

Let's generalize now the process obtained by kriging. $M_n(.,i)$ will be the kriging process for family i. $\forall t \in [t_1, t_K] \backslash \mathbb{T}_k$:

$$M_n(t,i) = \left\{ e^{-2(t - t^\ell)} - \gamma(t) \, e^{-2(t^r - t^\ell)} \right\} S_n(t^\ell, i) \;\; + \;\; \gamma(t) \, S_n(t^r, i)$$

where $\gamma(t)$ is given in Section 4.1.
$\sum_{i=1}^{I} \{M_n(.,i)\}^2$ will be the kriging process. By contiguity, under $H_{at^\star}$, $M_n(.,i)$ is the same process as under H_0 on which the mean function, $\bar{m}_{t^\star}^i(t)$ has been added :

$$\bar{m}_{t^\star}^i(t_k) = \left\{ e^{-2(t - t^\ell)} - \gamma(t) \, e^{-2(t^r - t^\ell)} \right\} m_{t^\star}^i(t^\ell) \;\; + \;\; \gamma(t) \, m_{t^\star}^i(t^r)$$

8.5. Comparison with Chang et al. (2009)

The law of the LRT process has also been obtained by Chang et al. (2009) under the null hypothesis. We propose here to present technical differences between our work and the work of Chang et al. (2009). As at a location t, the LRT is asymptotically the square of the score test, we will focus only on the score process as in Chang et al. (2009).

The main difference between the two approaches is that we consider the number of individuals in each class as a random variable whereas in Chang et al. (2009), the number of individuals in each class is supposed equal to the expectations (same remark as (b) of Section 3.2).

Our approach allows us to compute the score function $\frac{\partial \log L}{\partial q} \; |_{\theta_0}$ for only one observation and to calculate the Fisher information matrix without approximation.
Anyway, we obtain exactly the same Fisher information matrix as in Chang et al. (2009). However, there are some differences concerning other quantities.

8.5.1. Only two markers :

Let consider that there is only two markers as described in Section 3. Let $t \in]t_1, t_2[$. The result will be prolonged by continuity at the markers positions. According to formula (4)

of Section 3.2 and using the fact that $Q_t^{1,1} = 1 - Q_t^{-1,-1}$ and $Q_t^{1,-1} = 1 - Q_t^{-1,1}$, the score test statistic is :

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

With our notations, the test statistic used in formula (8) of Chang et al. (2009) is :

$$U^\star(t) = \frac{\sqrt{n}}{2} (1 - 2Q_t^{-1,-1}) \frac{\bar{r}(t_1, t_2) (\bar{y}_{11} - \bar{y}_{-1-1})}{\hat{\sigma} \sqrt{\mathbb{E}\left[\{2p(t) - 1\}^2 \right]}} + \frac{\sqrt{n}}{2} (1 - 2Q_t^{-1,1}) \frac{r(t_1, t_2) (\bar{y}_{1-1} - \bar{y}_{-11})}{\hat{\sigma} \sqrt{\mathbb{E}\left[\{2p(t) - 1\}^2 \right]}}$$

where $\bar{y}_{11} = \frac{2}{n\bar{r}(t_1,t_2)} \sum_{j=1}^{n} 1_{X_j(t_1)=1} 1_{X_j(t_2)=1}$, $\bar{y}_{-11} = \frac{2}{nr(t_1,t_2)} \sum_{j=1}^{n} 1_{X_j(t_1)=-1} 1_{X_j(t_2)=1}$
$\bar{y}_{1-1} = \frac{2}{nr(t_1,t_2)} \sum_{j=1}^{n} 1_{X_j(t_1)=1} 1_{X_j(t_2)=-1}$ and $\bar{y}_{-1-1} = \frac{2}{n\bar{r}(t_1,t_2)} \sum_{j=1}^{n} 1_{X_j(t_1)=-1} 1_{X_j(t_2)=-1}$.

We can remark $S_n(t) \neq U^\star(t) + o_{P_{\theta_0}}(1)$. It is due to the approximations done by Chang et al. (2009).
Let $G_n^1(t)$ and $G_n^2(t)$ be the quantities such as :

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

$G_n^1(t)$ and $G_n^2(t)$ are asymptotically standard normal variables under H_0. Besides, $G_n^1(t)$ and $G_n^2(t)$ are independent. Note that $G_n^1(t)$ and $G_n^2(t)$ do not depend on t but we keep t as a parameter in order to adapt these test statistics to the case of several markers in the next Section.
Contrary to formula (9) of Chang et al. (2009) :

$$G_n^1(t) \neq \frac{1}{2} \sqrt{\bar{r}(t_1, t_2) n}\, \frac{\bar{y}_{11} - \bar{y}_{-1-1}}{\hat{\sigma}} + o_{P_{\theta_0}}(1)$$
$$G_n^2(t) \neq \frac{1}{2} \sqrt{r(t_1, t_2) n}\, \frac{\bar{y}_{1-1} - \bar{y}_{-11}}{\hat{\sigma}} + o_{P_{\theta_0}}(1)$$

We have :

$$S_n(t) = \left\{ \sqrt{\bar{r}(t_1, t_2)}\, (1 - 2Q_t^{-1,-1})\, G_n^1(t) + \sqrt{r(t_1, t_2)}\, (1 - 2Q_t^{-1,1})\, G_n^2(t) \right\} / \sqrt{\mathbb{E}\left[\{2p(t) - 1\}^2 \right]} \tag{24}$$

This formula is the corrected version of formula (10) of Chang et al. (2009) without approximations here. According to formula (24), the score at a position t between two markers,

is an interpolation not linear between the test statistic $G_n^1(t)$ and $G_n^2(t)$. Naturally, when t tends to t_1 (resp. t_2), $S_n(t)$ tends to $S_n(t_1)$ (resp. $S_n(t_2)$). It becomes a linear interpolation between $S_n(t_1)$ and $S_n(t_2)$ if a Taylor linearization is done concerning the weights of the model of mixture (cf. Section 3.2).
Finally, we agree with formula (11) of Chang et al. (2009) concerning the covariance of the process, it is exactly the same function as $\Gamma(t,t')$ of theorem 1 of this paper.

Note that the non linear interpolation presented above, in formula (24), is not the same interpolation as presented in formula (5) of Section 3.1 of this paper. Our interpolation is more intuitive, because it is an interpolation between the test statistic on markers. Besides, we will see in the next Section that there is an advantage using our interpolation in terms of simulations.

8.5.2. Several markers : the "Interval Mapping" of Lander and Botstein (1989)
Let consider that there are several markers as described in Section 4. We consider values t, t' of the parameters that are distinct of markers positions. Let $t \in [t_1, t_K]\backslash\mathbb{T}_k$. We have :

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

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

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

This last formula is the corrected version of formula (14) of Chang et al. (2009).
Let $(t,t') \in]t^\ell, t^r[\times [t^r, t_K]\backslash\mathbb{T}_k$. The different covariances under H_0 are :

$$\mathrm{Cov}\left\{G_n^1(t), G_n^1(t')\right\} = \sqrt{\bar{r}(t^\ell,t^r)\;\bar{r}\left\{(t')^\ell,(t')^r\right\}}\;\; e^{-2\left\{(t')^\ell - t^r\right\}}$$

$$\mathrm{Cov}\left\{G_n^1(t), G_n^2(t')\right\} = \sqrt{\bar{r}(t^\ell,t^r)\;r\left\{(t')^\ell,(t')^r\right\}}\;\; e^{-2\left\{(t')^\ell - t^r\right\}}$$

$$\mathrm{Cov}\left\{G_n^2(t), G_n^1(t')\right\} = -\sqrt{r(t^\ell,t^r)\;\bar{r}\left\{(t')^\ell,(t')^r\right\}}\;\; e^{-2\left\{(t')^\ell - t^r\right\}}$$

$$\mathrm{Cov}\left\{G_n^2(t), G_n^2(t')\right\} = -\sqrt{r(t^\ell,t^r)\;r\left\{(t')^\ell,(t')^r\right\}}\;\; e^{-2\left\{(t')^\ell - t^r\right\}}$$

This is exactly the same covariances as in formula (19) of Chang et al. (2009). Besides, we agree with formula (20) of Chang et al. (2009) which establish a relationship between the test statistic G when t and t' belong to 2 consecutive marker interval (as above we suppose $t < t'$):

$$G_n^2(t') = \frac{1}{\sqrt{r(t^r,(t')^r)}}\left\{\sqrt{\bar{r}(t^\ell,t^r)}\,G_n^1(t) - \sqrt{r(t^\ell,t^r)}\,G_n^2(t) - \sqrt{\bar{r}(t^r,(t')^r)}\,G_n^1(t')\right\}$$

To conclude, the non linear interpolation proposed by Chang et al. (2009) is an approximation. We present here their interpolation without approximations. Howewer, their

approximations don't affect the final results concerning the process. Their approach is very interesting because it characterizes the process by interpolation between the test statistic G, and in terms of covariance between the test statistic G.
In this paper, we have calculated the whole covariance function $\Gamma(t, t')$ (cf. theorem 2) in order to see if under the alternative, the shift at a position t was $\Gamma(t, t^\star)$ as in Azaïs et al. (2006) and Azaïs et al. (2009). We have also characterized the process by a non linear interpolation between the test statistic on markers. When tests are performed only on markers, the score process is a Discrete Ornstein Uhlenbeck (DOU) process (cf. Section 4). As it is well known that the DOU process is an AR(1) process, it will be easy to simulate the discrete process on markers, and to obtain the values between markers by interpolation.

9. Acknowledgements

The authors thank Jean-Michel Elsen for having proposed this subject of research and 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.

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.

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

Candes, E. J. and Tao, T., (2005) The Dantzig selector : statistical estimation when p is much larger than n. *Annals of Statistics*, **35**, 2313-2351.

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.

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.

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.

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

Rabier, C-E. (2009) *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.

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

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

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

2.8 Article submitted “Threshold and power for QTL detection"

“Threshold and power for Quantitative Trait Loci detection"

Rabier C-E, Azaïs J-M, Elsen J-M., Delmas C.

Threshold and power for Quantitative Trait Locus detection[☆]

C-E. Rabier[*,a,b], J-M. Azaïs[a], J-M. Elsen[b], C. Delmas[b]

[a] *Université de Toulouse, Institut de Mathématiques de Toulouse, U.P.S., F-31062 Toulouse Cedex 9, France.*
[b] *INRA UR631, Station d'Amélioration Génétique des Animaux, BP 52627-31326 Castanet-Tolosan Cedex, France.*

Abstract

We propose several new methods to calculate threshold and power for Quantitative Trait Locus (QTL) detection. They are based on asymptotic theoretical results presented in Rabier et al. (2009) . The asymptotic validity is checked by simulations. The methods proposed are fast and easy to implement. A comparison of power between a multiple testing procedure and a global test has been realized, showing far better performances of the global test for the detection of a QTL.

Key words: QTL detection, Likelihood Ratio Test, Chi-Square process, Multiple Testing, Threshold, Monte-Carlo methods

1. Introduction

We study 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 in a population of progenies which are structured into sire families. The back-cross population, $A \times (A \times B)$, where A and B are purely homozygous lines, is a particular case of such a population.
A method largely used in order to detect a QTL, is the Interval Mapping proposed by Lander and Botstein (1989). Using the Haldane (1919) distance

[*]Corresponding author. Tel.:+33 5 61 28 52 64; fax.:+33 5 61 28 53 53
Email addresses: charles-elie.rabier@toulouse.inra.fr (C-E. Rabier), azais@cict.fr (J-M. Azaïs), jean-michel.elsen@toulouse.inra.fr (J-M. Elsen), celine.delmas@toulouse.inra.fr (C. Delmas)

and modelling, each chromosome is represented by a segment $[0,T]$. The distance on $[0,T]$ is called the genetic distance (measured in Morgans). At each location $t \in [0,T]$, the presence of a QTL is tested with a Likelihood Ratio Test (LRT). So, multi-testing leads to a LRT process, and taking as test statistic the supremum of this process comes down to perform a LRT in a model when the localisation of the QTL is an extra parameter.
Some theoretical results are present in Rebaï et al. (1994, 1995), in Cierco (1998), and in Azaïs and Cierco-Ayrolles (2002). Howewer, these papers use some approximations. In Rabier et al. (2009), article submitted, the focus is on the exact model. The asymptotic distributions of the LRT process are given under the null hypothesis (no QTL on $[0,T]$), under the alternative that there is one QTL at $t^\star$ on $[0,T]$, and under the general alternative that there are m QTL on $[0,T]$. The focus here is on the null hypothesis and on the particular alternative that there is one QTL on $[0,T]$.

In this paper :

1. we propose methods, as a function of the genetic map, to calculate thresholds for the supremum of the LRT process under H_0.
2. as all these methods are based on asymptotic results, we check the validity of the asymptotic assumption by simulating samples of different sizes.
3. we study the asymptotic power of the Interval Mapping and we give advices on how to optimize the detecting process.

The methods studied are available in a Matlab package with graphical user interface : "imapping.zip".
It can be downloaded at www.math.univ-toulouse.fr/∼rabier .

2. Model

The chromosome is the segment $[0,T]$. K genetic markers are located on the chromosome, one at each extremity. $t_1 = 0 < t_2 < ... < t_K = T$ are the locations of the markers. The "genome information" at t will be denoted $X(t)$. The Haldane (1919) model is the following : the law of $X(t_1)$ is $\frac{1}{2}(\delta_1 + \delta_{-1})$ and $X(t) = (-1)^{N(t)}X(t_1)$ where $N(t)$ is a standard Poisson process. Indeed, the Haldane model assumes that crossovers occur at random and independently of each other. The Haldane (1919)'s function

$r : [0,\ T]^2 \longmapsto \left[0,\ \frac{1}{2}\right]$ is such as :

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

This function links the recombination rate $r(t,t')$ between two loci located respectively at t and t', and the distance $|t-t'|$ between the two loci.

A sire family is defined by a set of progenies. I families will be considered. C is a discrete random variable refering to the family. The individual belongs to family i with probability $\pi_i = P(C = i)$.
The quantitative trait Y depends on the value of $X(t)$ at $t^\star \in [t_1, t_K]$ which is the location of the QTL. It also depends on the family it belongs to. The quantitative trait verifies :

$$\left(Y \middle| C = i\right) \ = \ \mu_i \ + \ X(t^\star)\, q_i \ + \ \sigma\, \varepsilon$$

where μ_i and q_i are respectively a polygenic effect and the QTL effect within family i. ε is a Gaussian white noise.
Besides, the "genome information" is available only at locations of genetic markers, that is to say at $t_1, t_2, ..., t_K$.
n is the number of observations j, (Y_j, $X_j(t_1)$, ..., $X_j(t_K)$, C_j). These observations are supposed to be independent and identically-distributed. We will call one population a sample of n observations.
The goal of this study is to test if there is a QTL on the chromosome with at least one of the sires heterozygous. The challenge is that $t^\star$ is unknown. So, the alternative hypothesis can be written :

$H_{\lambda t^\star}$: "there is at least one $q_i = \lambda_i/\sqrt{n}$, with $\lambda_i \in \mathbb{R}^\star$, at the position $t^\star$ "

In this context, we remind Theorem 3 of Rabier et al. (2009), which gives the asymptotic distribution of the LRT process, $\Lambda_n(.)$, under the null and the alternative hypothesis.

Theorem *With the previous defined notation,*

$$\Lambda_n(.) \overset{F.d.}{\rightarrow} \sum_{i=1}^{I} \left\{Z^i(.)\right\}^2 \tag{1}$$

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

- $\overset{F.d.}{\rightarrow}$ *is the convergence of fini-dimensional distributions*
- *the* $Z^i(.)$ *are independent Gaussian processes. More precisely,* $Z^i(.)$ *is the continuous and the non linear process such as* $\forall t \in]t_k, t_{k+1}[$ *:*

$$Z^i(t) = \left\{ \alpha(t)\, Z^i(t_k) \;+\; \beta(t)\, Z^i(t_{k+1}) \right\} \tag{2}$$

where $Cov\{Z^i(t_k), Z^i(t_{k'})\} = e^{-2|t_k - t_{k'}|}$. *The mean function of* $Z^i(.)$ *verifies :*

  - *under* H_0, $m(t) = 0$
  - *under* $H_{\lambda t^\star}$, $m^i_{t^\star}(t)$ *is proportional to* $\lambda_i \sqrt{\pi_i}$.

We refer to Rabier et al. (2009) for the rather complicated expressions of the functions $m^i_{t^\star}(t)$, $\alpha(t)$, $\beta(t)$ *and the covariance* $\Gamma(t, t')$ *of* $Z^i(.)$.

Note that when the number of genetic markers is infinite, the process $Z^i(.)$ is an Ornstein-Uhlenbeck process. The paths of three processes are presented in Figure 1 (the length of the chromosome is $T = 1$ Morgan):

- the Ornstein-Uhlenbeck process.
- the process $Z^1(.)$ with only 2 markers, located at $t_1 = 0$ and $t_2 = 1$M.
- the process $Z^1(.)$ with markers located every 10cM.

The paths of the last two processes are smooth whereas the paths of the Ornstein-Uhlenbeck process are very jerky. It's not suprising because the Ornstein-Uhlenbeck process can be viewed as a stationnary version of the Brownian motion.

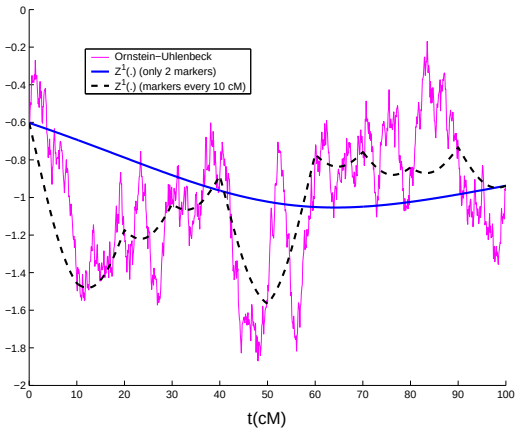


Figure 1: Paths of three different Gaussian processes

3. Different methods to obtain thresholds as a function of the map considered

3.1. Introducing the methods

We propose several new methods as a function of the map considered to calculate thresholds for the supremum of the LRT process under H_0. In particular, two kinds of maps are studied :

- a sparse map : a few markers covering the chromosome
- a dense map : a high density of markers pretty close to each other

We will suppose that when the map is dense, tests are performed only on markers, whereas when the map is sparse, test are also performed between markers.

Under a sparse map, thresholds can be obtained using the most appropriate methods as function of I :

- for $I = 1$, the problem comes down to computing the distribution of the maximum, i-absolute, value of a Gaussian vector. This can be done by a Discrete Monte-Carlo Quasi Monte-Carlo method (DMCQMC) : the method for numerical computation of a multivariate normal probability (Genz, 1992) can be considered. It uses a transformation that simplifies the problem and places it into a form that allows efficient calculation

using MCQMC methods. Note that a Newton's method can been used in order to obtain the threshold. This method is faster than using a simple Monte-Carlo method.

- for $I > 1$, a Discrete Monte-Carlo (DMC) method can be performed. According to the Theorem, when tests are performed only at the markers locations, the asymptotic process is a Discrete Ornstein-Uhlenbeck Chi-Square process with I degrees of freedoms (DOUCS(I)). The definition of such a process is given in the right-hand side of formula (1). When considered at the markers locations, the processes $Z^i(.)$ are simply AR(1) processes and we can interpolate by formula (2). In this situation, the threshold is easily obtained by a Discrete Monte-Carlo method based on a large number of sample paths ($nspaths$) of the asymptotic process.

Under a dense map, we propose theoretical methods to obtain the thresholds. Assuming that the number of genetic markers is infinite, the LRT process is asymptotically an Ornstein-Uhlenbeck Chi Square process with I degrees of freedom (OUCS(I)).
In a paper in progress, Rabier and Genz propose an approximative formula (named DF here) for the threshold of the supremum of such a OUCS(I) process. It is based on Delong (1981)'s work on Brownian motion. This formula is suitable when I and the threshold are large.
Besides, statistical tables given by Estrella (2003), for the threshold of the supremum of the OUCS(I), are also available. In order to obtain its exact tables, Estrella improved Delong's work on hypergeometrics functions. Estrella's method will be denoted ET.

Table 1 is a summary of all the methods proposed for the two kind of maps.

Map	Method
Dense (testing on markers)	ET (table available for $I \leq 20$) DF (formula available for I and threshold large)
Sparse (testing between markers)	$DMCQMC$ (available only for $I = 1$) DMC for $I > 1$

Table 1: Summary of all the methods studied as a function of the map considered and the way of performing tests (DMC for Discrete Monte-Carlo, $DMCQMC$ for Discrete Monte-Carlo Quasi Monte-Carlo, ET for Estrella exact table, DF for Delong approximative formula)

3.2. Applications under the null hypothesis

In this Section, the focus is on thresholds corresponding to the 95% quantile of the supremum of the LRT process under H_0. In order to illustrate the different methods, a sparse map and a dense map are considered. Since all the methods are based on asymptotic results (cf. Theorem), in order to check the validity of this results, some populations of different sizes have been simulated (*npop* denotes the number of populations whereas n denotes the size of a population).

Sparse map

The sparse map consists of a chromosome of length $T = 60$cM with 4 genetic markers equally spaced every 20cM. The presence of a QTL is tested every 5cM.

In Table 2, thresholds are presented as a function of I. In Table 3, the focus is on the number of false positives (NFP) as a function of the number of individuals n (thresholds taken from Table 2). Using Binomial distribution, a 95% confidence interval is calculated (into brackets in the tables) for the true percentage of the number of false positives.

According to Table 3, when there are in mean 200 individuals per family, that is to say $n = 200\,I$, NFP is not significantly different from 5%. When $n = 50\,I$, we can consider that NFP is still fair (even if it is significantly different from 5%) whereas when $n = 30\,I$, NFP is not so nominal.

Dense map

The dense map consists of a chromosome of length $T = 50$cM with 501 genetic markers equally spaced every 0.1cM.

The thresholds are compared in Table 4, and the NFP in Table 5. This aspect suggests fast convergence to asymptotic regime.

3.3. Remark

ET is not appropriate for sparse map for two reasons :

1. *ET* is based on Ornstein-Uhlenbeck (OU) process which is much more irregular than the process $Z^1(.)$ (see Figure 1). When $I = 1$, this can be formalized by the use of Slepian type inequalities, specially lemma (2.1) in Azaïs and Wschebor (2009) which comes from Plackett (1954). It can be proved that the covariances are smaller in the case of OU process than for the process $Z^1(.)$. It implies that the maximum of OU is stochastically greater than the $Z^1(.)$ one. Since $\mathbb{P}\left(\sup|Z^1(.)| > u\right) \approx$

$2\mathbb{P}\left(\sup Z^1(.) > u\right)$, this argument can be approximatively extended to the absolute value.

2. for the sparse map, the focus is not on continuous process but on dicrete process : the maximum of continuous process is always greater than the discrete one.

To sum up, ET will give too large thresholds.

Method	$DMCQMC\ (I=1)$	$DMC\ (I=3)$	$DMC\ (I=5)$
Threshold	6.06	10.76	14.47

Table 2: Thresholds obtained using the appropriate method as a function of the value of I considered ($nspaths = 1000000$). The map consists of 4 genetic markers equally spaced every 20cM (T=60cM). A test is done every 5cM.

Method / n	$DMCQMC\ (I=1)$	$DMC\ (I=3)$	$DMC\ (I=5)$
$n = 200\ I$	5.20% [4.98%; 5.42%]	5.03% [4.82%; 5.24%]	5.22% [5.00%; 5.44%]
$n = 50\ I$	5.78% [5.55%; 6.01%]	5.97% [5.74%; 6.20%]	6.11% [5.88%; 6.34%]
$n = 30\ I$	6.60% [6.36%; 6.84%]	6.77% [6.52%; 7.02%]	7.08% [6.83%; 7.33%]

Table 3: Number of False Positives (NFP) as a function of the number of individuals n and the method considered. The map consists of 4 genetic markers equally spaced every 20cM (T=60cM). A test is done every 5cM ($\sigma = 1$, $\mu_1 = -0.37$, $\mu_2 = 0.03$, $\mu_3 = 0.06$, $\mu_4 = -0.26$, $\mu_5 = 0.27$, $npop = 40000$).

	$I=1$			$I=3$		
Method	ET	DF	DMC	ET	DF	DMC
Threshold	7.84	7.61	7.68	13.09	12.91	12.86

Table 4: Thresholds obtained using theoretical methods ET, DF as function of the value of I considered. DMC for checking ($nspaths = 1000000$). The map consists of 501 genetic markers equally spaced every 0.1cM ($T = 0.5$M). A test is done on each marker.

Method / n	DF	DMC	ET
$n = 1000$	4.78% [4.57%; 4.99%]	5.13% [4.91%; 5.35%]	4.41% [4.21%; 4.61%]
$n = 500$	4.96% [4.75%; 5.17%]	5.15% [4.93%; 5.37%]	4.64% [4.43%; 4.85%]
$n = 150$	5.67% [5.44%; 5.90%]	5.91% [5.68%; 6.14%]	5.34% [5.12%; 5.56%]

Table 5: Number of False Positives (NFP) as a function of the number of individuals n and the method used. $I = 5$ here. The map consists of 501 genetic markers equally spaced every 0.1cM ($T = 0.5$M). A test is done on each marker ($\sigma = 1$, $\mu_1 = -0.37$, $\mu_2 = 0.03$, $\mu_3 = 0.06$, $\mu_4 = -0.26$, $\mu_5 = 0.27$, $npop = 40000$).

4. Study of the statistical power

Motivation

Some of the sires are heterozygotes at the QTL and others homozygous. QTL can only be detected in heterozygous sires families. Thus, two questions arise :

1. is it always profitable to include all the families in the analysis ?
2. do we have to analyze families all together or separately ?

We consider here, the sparse map of Section 3.2. As previously, tests are performed every 5 cM. The level considered is 5%.

About the QTL effects

When we deal with I families, since the total number of individuals is n, the expectation of the number of individuals in family i is only $n\pi_i$. So, in order to see the evolution of the power of the Interval Mapping with the number of families, we will consider $\lambda_i = \frac{\lambda}{\sqrt{\pi_i}}$ (note that if $I = 1$, $\lambda_1 = \lambda$ because $\pi_1 = 1$). As a consequence, the mean function, $m^i_{t^\star}(t)$, of the asymptotic process $Z^i(.)$, is proportional to λ and does not depend on i (cf. Theorem).

How to optimize the QTL detecting process

Only asymptotic results are studied here (cf. Theorem). Figures 2, 3, 4 illustrate question 1 whereas Figures 5, 6, 7 illustrate question 2.
In Figures 2, 3, 4, the power is plotted as a function of $t^\star$, I and the values of the λ_i's. In Figures 5, 6, 7, we compare the power of the approach which

consists of analyzing all families together (as previously), and the power of the approach which consists of analyzing families separately. For all of these Figures, $t^\star$ has been discretized with a step of 5cM.

Figures 2, 3, 4
For all of these three figures, two curves with the same colour on different Figures represent the same quantities.

As expected, the power increases with the number I of families (Figure 2 for $\lambda = 2$) and, for a given I, with the proportion of heterozygous sires (Figure 3 for $I = 5$, $\lambda = 2$ and various number number nz of non zeros λ_i's).

According to Figure 4, it is almost as powerful to consider $I = 1$ with $nz = 1$ (cf. grey curve) as $I = 5$ with $nz = 2$ (cf. green curve). So, it is much more powerful to consider $I = 1$ with $nz = 1$ (cf. grey curve) than $I = 5$ with $nz = 1$ (cf. brown curve). As a consequence, if they could be sorted in advance, it would be more powerful to concentrate the analysis on the families with a segregating QTL. Furthermore, once the families targeted, it would be more powerful to remove the families with very small QTL effects (not illustrated here). Indeed, it is like these families add noise to the model.

Figures 5, 6, 7
Practically, the segregating families are not known before the analysis and the true question is : do we have to analyze all the families together (global approach) or analyze families separately (Bonferroni approach) ? Indeed, since all the results are asymptotic, the variance is not better estimated when the global approach is considered.

Figures 5, 6,7 represent the two approaches. In these Figures, when the curve is :

- blue, $I = 5$
- cyan, $I = 7$
- orange, $I = 12$
- in solid line, it refers to the global approach
- in dashed line, it refers to the Bonferroni approach

We remind the null hypothesis,

$$H_0 : \text{“ } \forall i \in \{1, ..., I\}\,,\ \ q_i = 0 \text{ ”}$$

and the alternative hypothesis,

$$H_{\lambda t^\star} : \text{“ there is at least one } q_i = \lambda_i/\sqrt{n}\text{, with } \lambda_i \in \mathbb{R}^\star\text{, at the position } t^\star \text{ ”}$$

For the global approach, when H_0 is rejected, it only comes out that there is a QTL in at least one family, but this family is not known. For the Bonferroni approach, in order to answer the same question as for the global approach, we define the test statitistic U and the critical region CR, which results from a Bonferroni correction :

$$U = \left(\sup\left\{Z^1(.)\right\}^2,\ ...,\ \sup\left\{Z^I(.)\right\}^2\right)$$

$$CR = \left\{u = (u_1, ..., u_I) \in \mathbb{R}^I \text{ such as there is at least one } u_i \text{ verifying } u_i \geq c\right\}$$

where c is the threshold verifying : $\mathbb{P}\left(\sup\left\{Z_0^1(.)\right\}^2 \geq c\right) = \frac{0.05}{I}$.
$Z_0^1(.)$ is the Gaussian process centered and with covariance function $\Gamma(t, t')$. The Bonferroni correction allows to have $\mathbb{P}_{H_0}(U \in CR) \leq 0.05$. Obviously, the power of the Bonferroni approach is $\mathbb{P}_{H_{\lambda t^\star}}(U \in CR)$.
In Figure 5, the focus is on the particular case where there is only a QTL in family 1. $\lambda = 2$ has been taken. In the Figure, are represented the power of the two approaches as a function of $t^\star$ and I. It is noticeable that the Bonferroni approach is more powerful than the global approach. In Figure 6, the focus is on the particular case where there is a QTL in each family. In that case, the Bonferroni approach is outperformed by the global approach.
Figure 7 represents the mean power of the two approaches. Every alternative hypotheses have been considered (except the null hypothesis), i.e. for a given I, $nz = 1, ..., I$. Equiprobability concerning all these hypotheses has been supposed. According to the Figure, for a given I, there is a mean increase in term of power of at least 15% when the global approach is considered.

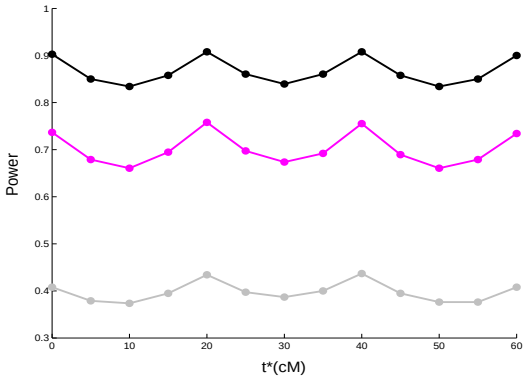


Figure 2: Power as a function of $t^\star$ and I. From top to bottom, $I = 5$, $I = 3$, $I = 1$ ($\lambda = 2$, $\sigma = 1$, $nspaths = 100000$).

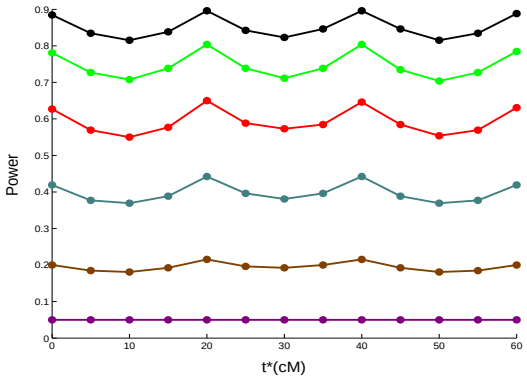


Figure 3: Power as a function of $t^\star$ and the number nz of non zero. From top to bottom, $nz = 5,\ 4,\ 3,\ 2,\ 1,\ 0$ ($I = 5$, $\lambda = 2$, $\sigma = 1$, $nspaths = 100000$).

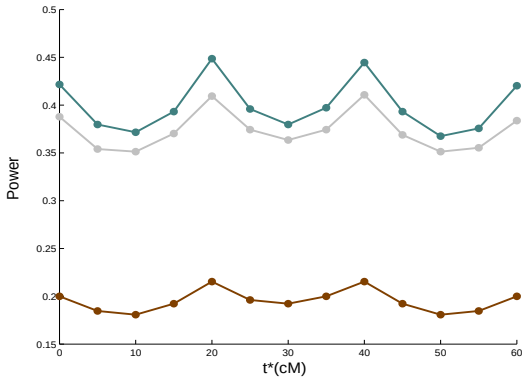


Figure 4: Power as a function of $t^\star$, I and the number nz of non zero. From top to bottom : $I = 5$ with $nz = 2$, $I = 1$ with $nz = 1$, $I = 5$ with $nz = 1$ ($\lambda = 2$, $\sigma = 1$, $nspaths = 100000$).

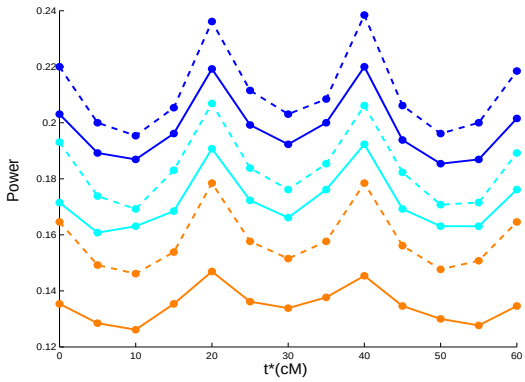


Figure 5: Power of the global approach (in solid line) and power of the Bonferroni approach (in dashed line), as a function of $t^\star$ and in the particular case of $nz = 1$. Orange refers to $I = 12$, cyan to $I = 7$ and blue to $I = 5$ ($\lambda = 2$, $\sigma = 1$, $nspaths = 100000$).

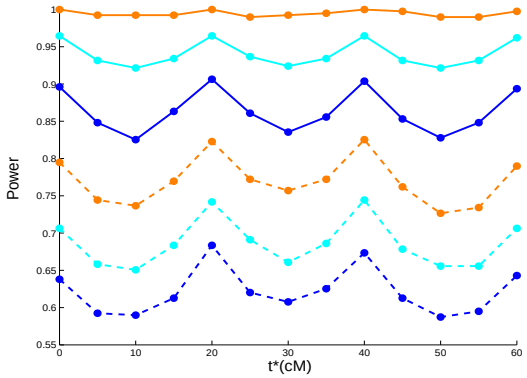


Figure 6: Power of the global approach (in solid line) and power of the Bonferroni approach (in dashed line), as a function of $t^\star$ and in the particular case of $nz = I$. Orange refers to $I = 12$, cyan to $I = 7$ and blue to $I = 5$ ($\lambda = 2$, $\sigma = 1$, $nspaths = 100000$).

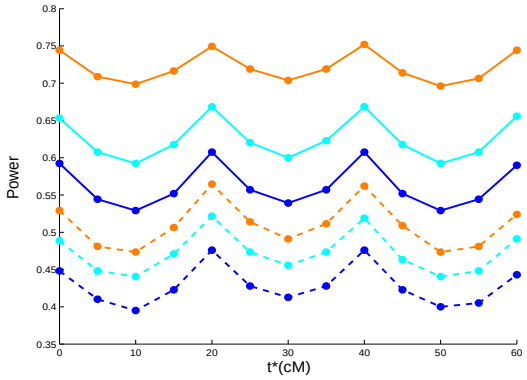


Figure 7: Mean power of the global approach (in solid line) and mean power of the Bonferroni approach (in dashed line), as a function of $t^\star$. Orange refers to $I = 12$, cyan to $I = 7$ and blue to $I = 5$ ($\lambda = 2$, $\sigma = 1$, $nspaths = 100000$).

4.1. Discussion

As all this study of the statistical power is based on asymptotic results, it is important to check the validity of the asymptotic assumption. A numerical evaluation has been performed for $\lambda = 2$ and a QTL located at $t^\star = 25$cM (cf. Table 6). The Theoretical Power, based on results of the Theorem, has been calculated using a DMC method. The Empirical Power (EP) has been computed assuming π_i equal to $1/I$ and a 95% confidence interval for the true value of the power is given into brackets. According to Table 6, the Theoretical Power is always located in the confidence interval whatever the value of n, demonstrating on this example, that the Theoretical Power should also be suitable for moderate values of n.

It can be seen that for all the figures shown here, the method is more powerful when the QTL is located on a marker. This is not surprising since on markers, the distribution from which belongs the quantitative trait Y is exactly known whereas between markers, since this distribution is unknown, a mixture model is used.

According to the Theorem, the LRT process, $\Lambda_n(.)$, is asymptotically the sum of the square of independent interpolated processes. So, it is easy to test every positions between genetic markers. In this article, as far as the sparse map is concerned, tests have been performed only every 5cM. In Figure 8, the focus is on an interval of two genetic markers spaced from 20cM ($I = 3$, $\lambda = 2$). We compare the power of the Interval Mapping when tests are performed every cM and every 5cM, as function of the location of the QTL, $t^\star$ ($I = 3$, $\lambda = 2$). It is noticeable that the two approaches give almost the same power : testing only every 5cM is convenient enough.

	$I = 1$	$I = 3$	$I = 5$
Theoretical Power	37.59%	68.57%	85.58%
EP for $n = 200$ I	38.08% [37.13%; 39.03%]	68.80% [67.89%; 69.71%]	85.00% [84.30%; 85.70%]
EP for $n = 50$ I	37.54% [36.59%; 38.49%]	68.37% [67.46%; 69.28%]	84.74% [84.04%; 85.44%]
EP for $n = 30$ I	37.83% [36.88%; 38.78%]	68.57% [67.66%; 69.48%]	85.15% [84.45%; 85.85%]

Table 6: Theoretical Power and Empirical Power (EP) as a function of I ($\lambda = 2$, $t^\star = 25$cM, $nspaths = 100000$ for the Theoretical Power and $npop = 10000$ for the Empirical Power, $\mu_1 = -0.37$, $\mu_2 = 0.03$, $\mu_3 = 0.06$, $\mu_4 = -0.26$, $\mu_5 = 0.27$, $\sigma = 1$)

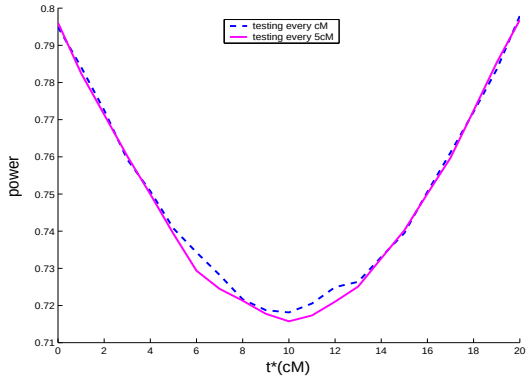


Figure 8: Power as a function of $t^\star$ and the way of testing ($\lambda = 2$, $\sigma = 1$, $nspaths = 100000$, $I = 3$). The map consists of 2 markers spaced from 20cM.

Conclusion

In order to optimize the QTL detecting process, we can advise :

1. to target, whenever possible, families with the biggest QTL effects and then, to analyze all these families together.
2. when it is not possible to target families, to analyze all the families together directly.

5. Acknowledgements

This work has been supported by the Animal Genetic Department of the French National Institute for Agricultural Research (INRA), SABRE, and the National Center for Scientific Research (CNRS).

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.

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

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

Delong, D. M., (1981) Crossing probabilities for a square root boundary by a Bessel process. *Commun. Statist.-Theor. Meth.*, **A10(21)**, 2197-2213.

Estrella, A., (2003) Critical values and p values of bessel process distributions : computation and application to structural break tests. *Econometric Theory*, **19(6)**, 1128-1143.

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

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.

Plackett, R.I. (1954) A reduction formula for normal multivariate integrals. *Biometrika*, **41**, 351-360.

Rabier, C-E., Azaïs, J-M., Delmas, C. (2009) Likelihood Ratio Test for Quantitative Trait Loci detection. *hal-00421215*.

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

Chapitre 3

About the supremum of the linear interpolated process

3.1 Introduction

In chapter 2, we have shown that when the genetic markers were not far from each other, the LRT process was asymptotically close to the square of a linear interpolated process. This is a generalization of the process studied under H_0 by Rebaï and al. (1994), Rebaï and al. (1995).

The aim of this chapter is to study the maximum of the square of this linear interpolated process. The focus is on one family. The results presented here are suitable under the null hypothesis (no QTL on the interval $[0, T]$) and also under the contiguous alternative that there is one QTL with effect $q = \frac{\lambda}{\sqrt{n}}$ at the position $t^\star$.

We proove that it is useless to perform tests at each positions between markers. People should rather perform tests on genetic markers, and only at one position in each marker interval.

At the end of this chapter, we propose a new method, based on these results, to calculate thresholds very quickly.

3.2 Study of the maximum

3.2.1 Only two genetic markers on $[0, 1]$

To begin, as in Section 2.3 of chapter 2, let consider that there are only two genetic markers. We suppose that they are located at $t_1 = 0$ and $t_2 = 1$. Let $\tilde{V}_0$ and $\tilde{V}_1$ be two random variables following a normal distribution with unit variance such as Cov $\left(\tilde{V}_0, \tilde{V}_1\right) = \tilde{\rho}$

with $0 < \tilde{\rho} < 1$. We consider the linear interpolated process $\tilde{V}_{(.)}$ on $[0,1]$:

$$\tilde{V}_t = \frac{(1-t)\,\tilde{V}_0 \;+\; t\,\tilde{V}_1}{\sqrt{(1-t)^2 + t^2 + 2\,\tilde{\rho}\,t\,(1-t)}}$$

The interest is on the supremum of the process $\left\{\tilde{V}_{(.)}\right\}^2$:

$$(\tilde{V}_t)^2 = \frac{(1-t)^2(\tilde{V}_0)^2 \;+\; 2t(1-t)\tilde{V}_0\tilde{V}_1 \;+\; t^2(\tilde{V}_1)^2}{(1-t)^2 \;+\; t^2 \;+\; 2\,\tilde{\rho}\,t\,(1-t)}$$

We will call respectively $N(t)$ and $D(t)$ the numerator and the denominator of the fraction above.

$$\frac{\partial N(t)}{\partial t} = 2\,\left\{(1-t)\,\tilde{V}_0 \;+\; t\,\tilde{V}_1\right\}\,\left\{\tilde{V}_1 - \tilde{V}_0\right\}$$

We can remark that :

$$(1-t)^2 + t^2 + 2\,t\,\tilde{\rho}\,(1-t) = 1 - 2(\,1-\tilde{\rho})\,t\,(1-t) \tag{3.1}$$

It comes :

$$\frac{\partial D(t)}{\partial t} = -2\,(1-\tilde{\rho})\,(1-2t)$$

$$\begin{aligned}\frac{\partial(\tilde{V}_t)^2}{\partial t} = \Big[\, & 2\left\{(1-t)\tilde{V}_0 + t\tilde{V}_1\right\}\left\{\tilde{V}_1 - \tilde{V}_0\right\}\{1-2(1-\tilde{\rho})t(1-t)\} \\ & +\; 2(1-\tilde{\rho})(1-2t)\left\{(1-t)\tilde{V}_0 + t\tilde{V}_1\right\}^2\Big] \,/\, \{D(t)\}^2\end{aligned}$$

$$\begin{aligned}&\frac{\partial(\tilde{V}_t)^2}{\partial t} = 0 \\ &\Leftrightarrow \left\{(1-t)\tilde{V}_0 + t\tilde{V}_1\right\} \\ &\times\; \left[\left\{\tilde{V}_1 - \tilde{V}_0\right\}\{1-2(1-\tilde{\rho})t(1-t)\} +\; (1-\tilde{\rho})(1-2t)\left\{(1-t)\tilde{V}_0 + t\tilde{V}_1\right\}\right] = 0\end{aligned}$$

As $\left\{(1-t)\tilde{V}_0 + t\tilde{V}_1\right\}$ corresponds to a minimum, the focus is on the second term. This second term is equal to zero if :

$$\frac{\tilde{V}_1}{\tilde{V}_0} = \frac{1+\tilde{\rho}}{1+(\tilde{\rho}-1)t} - 1$$

Let define the function $\psi_{\tilde{\rho}}(t)$ such as :

$$\psi_{\tilde{\rho}}(t) = \frac{1+\tilde{\rho}}{1+(\tilde{\rho}-1)t} - 1$$

As $(\tilde{\rho}-1)t$ is a decreasing function on $[0,1]$, then $\psi_{\tilde{\rho}}(t)$ is an increasing function on $[0,1]$ with $\psi_{\tilde{\rho}}(0) = \tilde{\rho}$ and $\psi_{\tilde{\rho}}(1) = \frac{1}{\tilde{\rho}}$.
Let define $\psi_{\tilde{\rho}}^{-1}$ the inverse function of $\psi_{\tilde{\rho}}$. After straightforward calculations, we find :

$$\psi_{\tilde{\rho}}^{-1}(u) = \frac{\tilde{\rho}-u}{(\tilde{\rho}-1)(u+1)}$$

So, the extremum between 0 and 1 is obtained for :

$$\tilde{\xi} = \frac{\tilde{\rho}\,\tilde{V}_0 - \tilde{V}_1}{(\tilde{\rho}-1)(\tilde{V}_1+\tilde{V}_0)}$$

After some calculations, using formula (3.1), we find that :

$$(1-\tilde{\xi})^2 + \tilde{\xi}^2 + 2\,\tilde{\xi}\,\tilde{\rho}\,(1-\tilde{\xi}) = \frac{1+\tilde{\rho}}{1-\tilde{\rho}}\ \frac{(\tilde{V}_0)^2 + (\tilde{V}_1)^2 - 2\,\tilde{\rho}\,\tilde{V}_0\,\tilde{V}_1}{(\tilde{V}_0+\tilde{V}_1)^2}$$

It comes :

$$(\tilde{V}_{\tilde{\xi}})^2 = \frac{(\tilde{V}_0)^2 + (\tilde{V}_1)^2 - 2\,\tilde{\rho}\,\tilde{V}_0\,\tilde{V}_1}{(1+\tilde{\rho})(1-\tilde{\rho})}$$

So :

$$\begin{aligned} sup_{t\in[0,1]}(\tilde{V}_t)^2 = (\tilde{V}_{\tilde{\xi}})^2\, 1_{\frac{\tilde{V}_1}{\tilde{V}_0} \in\,]\,\tilde{\rho}\,,\,\frac{1}{\tilde{\rho}}\,[} \;+\; (\tilde{V}_1)^2\, 1_{\frac{\tilde{V}_1}{\tilde{V}_0} \in\, [\,\frac{1}{\tilde{\rho}}\,,\,+\infty\,[} \\ + (\tilde{V}_0)^2\, 1_{\frac{\tilde{V}_1}{\tilde{V}_0} \in\, [\,0\,,\,\tilde{\rho}\,]} \;+\; (\tilde{V}_0)^2\, 1_{\frac{\tilde{V}_1}{\tilde{V}_0} \in\,]\,-\infty\,,\,0[\,\bigcap\, |\tilde{V}_0|>|\tilde{V}_1|} \\ + (\tilde{V}_1)^2\, 1_{\frac{\tilde{V}_1}{\tilde{V}_0} \in\,]\,-\infty\,,\,0[\,\bigcap\, |\tilde{V}_1|>|\tilde{V}_0|} \end{aligned} \tag{3.2}$$

A concise version of this formula is that the supremum of $\left\{\tilde{V}_{(.)}\right\}^2$ is the maximum of three random variables :

$$sup_{t\in[0,1]}(\tilde{V}_t)^2 = max\left\{(\tilde{V}_0)^2\,,\ (\tilde{V}_{\tilde{\xi}})^2\, 1_{\frac{\tilde{V}_1}{\tilde{V}_0} \in\,]\,\tilde{\rho}\,,\,\frac{1}{\tilde{\rho}}\,[}\,,\ (\tilde{V}_1)^2\right\}$$

3.2.2 Only two genetic markers on $[0,T]$

We propose to generalize the results of the previous Section to the interval $[0,T]$.

We consider that there are only two genetic markers located now at $t_1 = 0$ and $t_2 = T$ with $T \in \mathbb{R}^{+\star}$. V_0 and V_T are two random variables following a normal distribution with unit variance such as Cov $(V_0, V_T) = \rho$ with $0 < \rho < 1$.
We consider the linear interpolated process $V_{(.)}$ on $[0,T]$:

$$V_t = \left\{ \frac{T-t}{T} V_0 \; + \; \frac{t}{T} V_T \right\} / \sqrt{\tau(t)}$$

where

$$\tau(t) = \left(\frac{T-t}{T} \right)^2 \; + \; 2 \, \frac{t(T-t)}{T^2} \rho \; + \; \left(\frac{t}{T} \right)^2$$

We can remark that if $V_0 = \tilde{V}_0$, $V_T = \tilde{V}_1$ and $\tilde{\rho} = \rho$, then $\tilde{V}_{t/T} = V_t$. It comes :

$$sup_{t \in [0,T]} (V_t)^2 = sup_{t \in [0,1]} (\tilde{V}_t)^2$$

According to formula (3.2) :

$$\begin{aligned} sup_{t \in [0,T]} (V_t)^2 = (V_\xi)^2 \, 1_{\frac{V_T}{V_0} \in \,] \, \rho \, , \, \frac{1}{\rho} \, [} \; + \; (V_T)^2 \, 1_{\frac{V_T}{V_0} \in \, [\, \frac{1}{\rho} \, , \, +\infty \, [} \\ + (V_0)^2 \, 1_{\frac{V_T}{V_0} \in \, [\, 0 \, , \, \rho \,]} \; + \; \{V_0\}^2 \, 1_{\frac{V_T}{V_0} \in \,] \, -\infty \, , \, 0 [\, \bigcap \, |V_0| > |V_T|} \\ + (V_T)^2 \, 1_{\frac{V_T}{V_0} \in \,] \, -\infty \, , \, 0 [\, \bigcap \, |V_T| > |V_0|} \end{aligned} \tag{3.3}$$

with

$$\xi = \frac{T \, (\rho \, V_0 - V_T)}{(\rho - 1)(V_0 + V_T)} \quad \text{and} \quad (V_\xi)^2 = \frac{(V_0)^2 + (V_T)^2 - 2 \, \rho \, V_0 \, V_T}{(1+\rho)(1-\rho)}$$

The concise version of this formula is :

$$sup_{t \in [0,T]} (V_t)^2 = max \left\{ (V_0)^2 \, , \, (V_\xi)^2 \, 1_{\frac{V_T}{V_0} \in \,] \, \rho \, , \, \frac{1}{\rho} \, [} \, , \, (V_T)^2 \right\}$$

3.2.3 Several markers : the “Interval Mapping" of Lander and Botstein (1989)

As in Section 2.4 of chapter 2, K genetic markers are now located on the chromosome, one at each extremity. $0 = t_1 < t_2 < ... < t_K = T$ are the locations of the markers.
We consider values of the parameter t that are distinct of the markers positions, and the

result will be prolonged by continuity at the markers positions. We remind the notations of chapter 2 : $\mathbb{T}_k = \{t_1, ..., t_K\}$ and t^ℓ, t^r are the quantities such as :

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

We consider the linear interpolated process $V_{(.)}$ defined such as :

$$V_t = \left\{ \frac{t^r - t}{t^r - t^\ell} V_{t^\ell} + \frac{t - t^\ell}{t^r - t^\ell} V_{t^r} \right\} / \sqrt{\tau(t)}$$

where V_{t^ℓ} and V_{t^r} are two random variables following a normal distribution with unit variance such as Cov $(V_{t^\ell}, V_{t^r}) = \rho(t^\ell, t^r)$ with $0 < \rho(t^\ell, t^r) < 1$. Besides,

$$\tau(t) = \left(\frac{t^r - t}{t^r - t^\ell} \right)^2 + 2 \frac{(t^r - t)(t - t^\ell)}{(t^r - t^\ell)^2} \rho(t^\ell, t^r) + \left(\frac{t - t^\ell}{t^r - t^\ell} \right)^2$$

The interest is still on the supremum of the process $\left\{V_{(.)}\right\}^2$. We can easily adapt the results of the previous Section to the "Marker interval" (t^ℓ, t^r).
ξ becomes $\xi(t^\ell, t^r)$ with :

$$\xi(t^\ell, t^r) = \frac{(t^r - t^\ell) \left\{ \rho(t^\ell, t^r) V_{t^\ell} - V_{t^r} \right\}}{\{\rho(t^\ell, t^r) - 1\} \{V_{t^\ell} + V_{t^r}\}} + t^\ell$$

It comes :

$$\left\{V_{\xi(t^\ell, t^r)}\right\}^2 = \frac{\{V_{t^\ell}\}^2 + \{V_{t^r}\}^2 - 2\,\rho(t^\ell, t^r)\, V_{t^\ell}\, V_{t^r}}{\{1 + \rho(t^\ell, t^r)\} \{1 - \rho(t^\ell, t^r)\}}$$

We adapt formula (3.3) to the "Marker interval" (t^ℓ, t^r) :

$$\begin{aligned} sup_{t \in [t^\ell, t^r]} \{V_t\}^2 &= \left\{V_{\xi(t^\ell, t^r)}\right\}^2 1_{\frac{V_{t^r}}{V_{t^\ell}} \in \,] \rho(t^\ell, t^r) \,,\, \frac{1}{\rho(t^\ell, t^r)} \,[} \\ &+ \{V_{t^r}\}^2 1_{\frac{V_{t^r}}{V_{t^\ell}} \in \,[\, \frac{1}{\rho(t^\ell, t^r)} \,,\, +\infty \,[} + \{V_{t^\ell}\}^2 1_{\frac{V_{t^r}}{V_{t^\ell}} \in \,[\, 0 \,,\, \rho(t^\ell, t^r) \,]} \\ &+ \{V_{t^\ell}\}^2 1_{\frac{V_{t^r}}{V_{t^\ell}} \in \,] -\infty \,,\, 0[\,\cap\, |V_{t^\ell}| > |V_{t^r}|} + \{V_{t^r}\}^2 1_{\frac{V_{t^r}}{V_{t^\ell}} \in \,] -\infty \,,\, 0[\,\cap\, |V_{t^r}| > |V_{t^\ell}|} \end{aligned}$$

So, the concise version of this formula is :

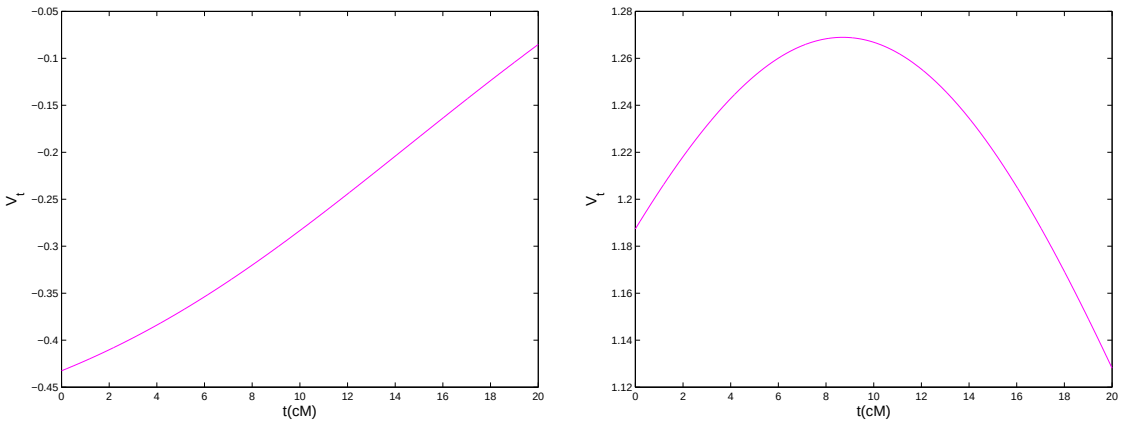
$$sup_{t \in [t^\ell, t^r]} \{V_t\}^2 = max \left[\{V_{t^\ell}\}^2 \,,\, \left\{V_{\xi(t^\ell, t^r)}\right\}^2 1_{\frac{V_{t^r}}{V_{t^\ell}} \in \,] \rho(t^\ell, t^r) \,,\, \frac{1}{\rho(t^\ell, t^r)} \,[} \,,\, \{V_{t^r}\}^2 \right]$$

By generalizing to the whole interval $[0,T]$, we obtain the final result :

$$sup_{t\in[0,T]}\left\{V_t\right\}^2 = max\left[\left\{V_{t_1}\right\}^2,...,\left\{V_{t_K}\right\}^2,\left\{V_{\xi(t_1,t_2)}\right\}^2 1_{\frac{V_{t_2}}{V_{t_1}}\in\,]\,\rho(t_1,t_2)\,,\,\frac{1}{\rho(t_1,t_2)}\,[}\right.$$
$$\left.,...,\left\{V_{\xi(t_{K-1},t_K)}\right\}^2 1_{\frac{V_{t_K}}{V_{t_{K-1}}}\in\,]\,\rho(t_{K-1},t_K)\,,\,\frac{1}{\rho(t_{K-1},t_K)}\,[}\right]$$

Note that this result is obtained for any covariance $\rho(t^\ell,t^r)$. If we want to use the Haldane distance, as in chapter 2, we have to consider $\rho(t^\ell,t^r)=e^{-2(t^r-t^\ell)}$.
As this result is suitable under the null hypothesis and under contiguous alternatives, we can advise geneticists not to test anymore at each positions between markers.

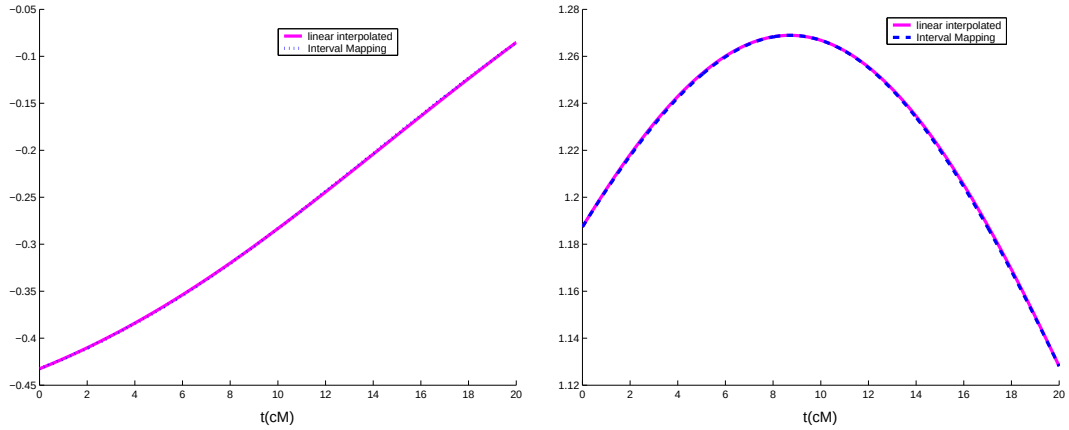
3.2.4 Graphical illustrations under H_0



First path Second path

FIG. 3.1 – Two different paths, under H_0, of the linear interpolated process $V_{(.)}$ with two markers located at $t_1=0$ and $t_2=0.2$M

We consider a map which consists of two genetic markers located at $t_1=0$ and $t_2=0.2$M. Using the Haldane distance, $\rho=0.6703$. Figure 3.1 represents two paths under H_0 of the linear interpolated process $V_{(.)}$. On the first path, we can see that the supremum of $\left\{V_{(.)}\right\}^2$ will be obtained on the markers whereas on the second path, as $\frac{V_{0.2}}{V_0}=\frac{1.1281}{1.1873}=0.9501\in]0.6703,1.4918[$, the supremum of $\left\{V_{(.)}\right\}^2$ will be obtained at $\xi=\frac{0.2\,(0.6703\times1.1281-1.1873)}{(0.6703-1)(1.1281+1.1873)}=0.0870$.
On Figure 3.2, we use the same map. We compare the paths of the linear interpolated process $V_{(.)}$ and those of the process $Z_{(.)}$ corresponding to the Interval Mapping (cf.



First path for the two processes Second path for the two processes

FIG. 3.2 – Comparison between the paths, under H_0, of the linear interpolated process $V_{(.)}$ and those of the process $Z_{(.)}$ corresponding to the Interval Mapping

chapter 2). For $V_{(.)}$, the same paths as presented in figure 3.1, are considered. We can see that the paths overlap which is not surprising because we have shown analytically that the linear interpolated process was a good approximation if the markers are close to each others.

3.2.5 Thresholds

The theoretical result presented in this chapter allow us to propose a new method to obtain the 95% quantile of the supremum of the process $\left\{V_{(.)}\right\}^2$ under H_0. It will use Monte-Carlo Quasi Monte-Carlo methods of Genz (1992) which are very fast (we will use function QSIMVNEF of Genz).

As in the previous section, Haldane's mapping function will be considered. We will consider the performances of this method with other methods usually used in QTL detection.

In Rebaï and al. (1994), we can find an upper bound for the threshold. This bound is the quantity c^2 such as :

$$0.05 = 2\,\Phi(-c) \;+\; \frac{2\,e^{-c^2/2}}{\pi} \sum_{k=1}^{K-1} \arctan\left(\sqrt{\frac{1-e^{-2(t_{k+1}-t_k)}}{1+e^{-2(t_{k+1}-t_k)}}} \right)$$

where Φ is the cumulative distribution of the standardized normal distribution.

This method is based on Davies (1977). However, it is sensitive to the number of genetic

markers. Indeed, the derivative of the process $V_{(.)}$ has a jump at each markers location, and Davies (1977) upper bound is suitable when the derivative of the process has a finite number of jumps.

In Feingold and al. (1993), the authors propose a threshold based on the discrete process resulting from tests only on markers. Besides, they suppose constant the distance between genetic markers. The threshold c^2 is such as :

$$0.05 = 1 \; - \; \Phi(c) \; + \; 2\,T\,c\,\varphi(c)\,\nu(2c\sqrt{\Delta})$$

where φ is the density of a normal standardized, Δ is distance between two consecutive markers.

This method is inspired from Siegmund (1985) where the function ν is fully described. This method requires the number of genetic makers to be not too small.

In Tables 3.1 and 3.2, thresholds corresponding to different methods are computed :

- DMCQMC refers to the method presented in the article “Threshold and power for Quantitative Trait Locus detection" (cf. chapter 2). It is a Discrete Monte-Carlo Quasi Monte-Carlo method, based on the exact asymptotic process $\left\{Z_{(.)}\right\}^2$. We remind that $Z_{(.)}$ is a non linear interpolated process.
- DMCQMClin refers to the method proposed in this chapter.
- Rebaï
- Feingold

Note that the maps considered are the same as in Rebaï and al. (1994).

As expected, Rebaï is very sensitive to the number of genetic markers. We can observe that Feingold, DMCQMClin, and DMCQMC give almost same results.

Howewer DMCQMClin is :

1. faster than DMCQMC
2. appropriate whatever the map, which is not the case for Feingold (cf. Table 3.3).

Method	*DMCQMC*	*DMCQMClin*	*Rebaï*	*Feingold*
Threshold	6.76	6.76	6.92	6.78

TAB. 3.1 – Thresholds as a function of the method considered. The map consists of 6 genetic markers equally spaced every 20cM (T=1M). For *DMCQMC*, a test is done every 5cM.

Method	*DMCQMC*	*DMCQMClin*	*Rebaï*	*Feingold*
Threshold	8.25	8.23	9.09	8.26

TAB. 3.2 – Thresholds as a function of the method considered. The map consists of 51 genetic markers equally spaced every 2cM (T=1M). For *DMCQMC*, a test is done every cM.

Method	*DMCQMC*	*DMCQMClin*	*Feingold*
Threshold	5.42	5.40	5.78

TAB. 3.3 – Thresholds as a function of the method considered. The map consists of 2 genetic markers (T=1M). For *DMCQMC*, a test is done every 5cM.

Chapitre 4

About the supremum of Chi-Square processes

The following is an article in progress, written in collaboration with Professor Alan Genz from Washington State University.

“The supremum of Chi-Square processes"

Rabier C-E., Genz A.

The supremum of Chi-Square processes

C-E. Rabier[a,b], A. Genz*[,c]

[a] *Université de Toulouse, Institut de Mathématiques de Toulouse, U.P.S., F-31062 Toulouse Cedex 9, France.*
[b] *INRA UR631, Station d'Amélioration Génétique des Animaux, BP 52627-31326 Castanet-Tolosan Cedex, France.*
[c] *Department of Mathematics, Washington State University, Pullman, WA 99164-3113, USA.*

Abstract

Using methods, We describe a lower bound for the critical value of the supremum of a Chi-Square process. This bound can be approximated using a MCQMC simulation. We compare numerically this bound with the upper bound given by Davies, only suitable for a regular Chi-Square process. In a second part, we focus a non regular Chi-Square process : the Ornstein-Uhlenbeck Chi-Square process. Recently, Rabier et al. (2009) have shown that this process has an application in genetics : it is the limiting process of the likelihood ratio test process related to the test of a gene on an interval representing a chromosome. Using results from Delong (1981), we propose a theoretical formula for the supremum of such a process and we compare it in particular with our simulated lower bound.

Key words: Chi-Square process, Monte-Carlo Quasi Monte-Carlo, Ornstein-Uhlenbeck process, genetics

1. The Davies Upper bound

In the article of Davies (1987), the focus is on hypothesis testing when a nuisance parameter $t^\star$ is present only under alternative. So, $t^\star$ is meaningless under the null hypothesis. If $t^\star$ were known, the natural way to perform the

*Corresponding author. Tel.:(509)335-2131;; fax.:(509)335-1188
Email addresses: charles-elie.rabier@toulouse.inra.fr (C-E. Rabier), alangenz@wsu.edu (A. Genz)

test is to consider $t = t^\star$. However, as it is only known that t belong to the interval $[\mathcal{L}, \mathcal{U}]$, Davies suggests to use the test statistic :

$$\sup \{S(t) : \mathcal{L} \leq t \leq \mathcal{U}\} \tag{1}$$

where $S(t)$ denotes the test statistic at t.

Besides, Davies considers the case :

$$S(t) = V_1(t)^2 + ... + V_d(t)^2 \tag{2}$$

where the $V_i(t)$ are independent for each t and distributed as a standardized normal under the null hypothesis. The process $S(.)$ is called a Chi-Square process with d degrees of freedom.
The main results of Davies (1987) is the following formula :

$$\mathbb{P}\left(\sup_{t \in [\mathcal{L}, \mathcal{U}]} S(t) > c \right) \leq \mathbb{P}\left(\chi^2_d > c\right) + \int_{\mathcal{L}}^{\mathcal{U}} \Psi(t) dt \tag{3}$$

where

$$\Psi(t) = \mathbb{E}\left(\|\eta(t)\|\right) \, c^{\frac{d-1}{2}} \, e^{-c/2} \, \pi^{-1/2} \, 2^{-d/2} \, / \, \Gamma\left(d/2 + 1/2\right)$$

Γ is the Gamma function and χ^2_d is a random variable which follows a Chi-Square with d degrees of freedom.

We refer to Davies (1987) to obtain the general expression of the quantity $\mathbb{E}\left(\|\eta(t)\|\right)$. The author specifies that formula (3) is suitable when the processes $V_i(.)$ have a derivative with a finite number of jumps.

In what follows, we will call Davies upper bound the right side of formula (3). Besides, we will focus only on Chi-Square processes $S(.)$ where the $V_i(.)$ are independent (particular case of formula 2).

2. Computation of the Discretized Lower Bound

In this section, we present a lower bound for the critical value of the supremum of a Chi-Square process. This is a lower bound because we discretize the process. The probabilities needed for the lower bounds can then be explicitly written as multivariate normal integrals, which can be approximated with simulation methods.

The time interval $[\mathcal{L}, \mathcal{U}]$ for the process is discretized using $t_i = \mathcal{L} + i(\mathcal{U} - \mathcal{L})/m$ for $i = 1, 2, \ldots, m$. Then we define A to be the $m \times m$ covariance matrix for the discretized process, with entries $a_{ij} = r(t_i - t_j)$. If we also define X to be an $d \times m$ matrix, with columns $\mathbf{x}_j$, for $j = 1, \ldots, m$, the integrals that are needed for computation of the lower bound are multivariate Normal probability integrals over a product of m d-dimensional spheres, given by

$$P(u) = \int_{||\mathbf{x}_1||^2 < u^2} \int_{||\mathbf{x}_2||^2 < u^2} \cdots \int_{||\mathbf{x}_m||^2 < u^2} \frac{e^{-\frac{1}{2}\sum_{i=1}^{d}[x_{i1},\ldots,x_{im}]A^{-1}[x_{i1},\ldots,x_{im}]'}}{((2\pi)^m |A|)^{\frac{d}{2}}} \prod_{i=1}^{d}\prod_{j=1}^{m} dx_{ij},$$

for real $u \geq 0$. If we determine u so that $P(u) = 1 - \alpha$, then u will be a lower bound for the critical value of the supremum of the Chi-Square process with covariance function $r(t)$.

In order to describe some simulation methods for approximation of the $P(u)$ integrals, we start with a change of variables designed to simplify the multivariate Normal density. Let L be the $m \times m$ lower triangular Cholesky factor for A (so that $A = LL'$). Now define the change variables to an $d \times m$ matrix of variables Y by $X = YL'$, so that $dX = |A|^{\frac{d}{2}} dY$ and therefore

$$P(u) = \int_{||\mathbf{x}_1(Y)||^2 < u^2} \int_{||\mathbf{x}_2(Y)||^2 < u^2} \cdots \int_{||\mathbf{x}_m(Y)||^2 < u^2} \frac{e^{-\frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{m} y_{ij}^2}}{(2\pi)^{\frac{md}{2}}} \prod_{i=1}^{d}\prod_{j=1}^{m} dy_{ij}, \tag{4}$$

where $x_{ij}(Y) = \sum_{k=1}^{j} l_{jk} y_{ik}$, and $\mathbf{x}_j(Y)$ is a function of $\mathbf{y}_1, \ldots \mathbf{y}_j$.

2.1. Direct Simulation

A direct simulation method for approximating the integrals $I(u)$ uses simulation from the univariate Normal distribution. Let $Y_{ij}^{(k)} \sim N(0, 1)$, and define

$$f(Y^{(k)}) = max_{1 \leq j \leq m}(||\mathbf{x}_j(Y^{(k)}||) \quad \text{and} \quad g(Y^{(k)}) = \begin{cases} 1 \text{ if } f(Y^{(k)}) \leq u \\ 0 \text{ otherwise} \end{cases}.$$

Then

$$P(u) \approx P_N = \frac{1}{N} \sum_{k=1}^{N} g(Y^{(k)})$$

with standard error

$$E_N = (\frac{1}{N(N-1)} \sum_{k=1}^{N} (g(Y^{(k)}) - P_N)^2)^{\frac{1}{2}}.$$

Because the $g(Y^{(k)})$ is 0 or 1, $E_N \approx (\frac{P(u)(1-P(u))}{N})^{\frac{1}{2}} \approx (\frac{P_N(1-P_N)}{N})^{\frac{1}{2}}$ (see Fishman (1996)).

If u_p is the value of u where $P(u) = p$ for a given probability p, an approximate value for u_p can easily be determined from this simulation. Define $\mathbf{F}$ to be the vector of sorted (ascending order) $f(Y^{(k)})$ values and define $[pN]$ to be the value of pN rounded to the nearest integer. Then $u_p \approx F_{[pN]}$.

2.2. Conditional Simulation

The direct simulation method described in the previous section is an "acceptance-rejection" algorithm which can be inefficient for some combinations of u and A, so a potentially more efficient method is considered in this section. This method is a generalization of the method described in the paper by Genz (1992), where MVN probabilities over hyper-rectangular regions were considered. The method in this paper begins with equation (4) for $P(u)$ written in more detail in the form

$$\begin{aligned} P(u) \;=\; & \int\limits_{\sum_{i=1}^{d}(l_{11}y_{i1})^2<u^2} \frac{e^{-\frac{1}{2}\sum_{i=1}^{d} y_{i1}^2}}{(2\pi)^{\frac{d}{2}}} \int\limits_{\sum_{i=1}^{d}(l_{21}y_{i1}+l_{22}y_{i2})^2<u^2} \frac{e^{-\frac{1}{2}\sum_{i=1}^{d} y_{i2}^2}}{(2\pi)^{\frac{d}{2}}} \\ & \cdots \int\limits_{\sum_{i=1}^{d}(l_{m1}y_{i1}+\cdots+l_{mm}y_{im})^2<u^2} \frac{e^{-\frac{1}{2}\sum_{i=1}^{d} y_{im}^2}}{(2\pi)^{\frac{d}{2}}} \prod_{j=m}^{1}\prod_{i=d}^{1} dy_{ij}. \end{aligned} \tag{5}$$

In order to simplify the notation, we assume that A is nonsingular and define a matrix C, determined by scaling the rows of L by the successive diagonal elements of L, so that $c_{ij} = l_{ij}/l_{ii}$, which makes $c_{ii} = 1$. If the scaled sphere radii are defined by $u_i = u/l_{ii}$, then

$$\begin{aligned} P(u) \;=\; & \int\limits_{\sum_{i=1}^{d} y_{i1}^2<u_1^2} \frac{e^{-\frac{1}{2}\sum_{i=1}^{d} y_{i1}^2}}{(2\pi)^{\frac{d}{2}}} \int\limits_{\sum_{i=1}^{d}(c_{21}y_{i1}+y_{i2})^2<u_2^2} \frac{e^{-\frac{1}{2}\sum_{i=1}^{n} y_{i2}^2}}{(2\pi)^{\frac{d}{2}}} \\ & \cdots \int\limits_{\sum_{i=1}^{d}(c_{m1}y_{i1}+\cdots+c_{m-1,1}y_{i,m-1}+y_{im})^2<u_m^2} \frac{e^{-\frac{1}{2}\sum_{i=1}^{n} y_{im}^2}}{(2\pi)^{\frac{d}{2}}} \prod_{j=m}^{1}\prod_{i=d}^{1} dy_{ij}. \end{aligned}$$

The structure of the integration constraints permit the conditional integrations to be completed with many possible orderings for the variables. We describe a method which uses a variable ordering by rows, starting

with $y_{11}, y_{12}, \ldots y_{1m}$, followed by $y_{21}, y_{22}, \ldots y_{2m}$, and so on, finishing with $y_{d1}, y_{d2}, \ldots y_{dm}$, as is indicated by the $\prod_{j=m}^{1} \prod_{i=d}^{1} dy_{ij}$ measure in equation (5).

The "outermost" variable y_{11}, has constraint $-u_1 < y_{11} < u_1$. Given y_{11}, the next variable y_{12}, has constraint $-u_2 - c_{11}y_{11} < y_{12} < u_2 - c_{11}y_{11}$, and so on, with y_{1j} constrained by $-u_j < \sum_{k=1}^{j-1} c_{jk}y_{1k} + y_{1j} < u_j$. If values are given for y_{ik}, $i = 1, \ldots, l-1$, $k = 1, \ldots, m$, and y_{lk}, $k = 1, \ldots, j-1$, then the integral for y_{lj} has the constraint

$$\sum_{i=1}^{l-1} (\sum_{k=1}^{j-1} c_{jk}y_{ik})^2 + (\sum_{k=1}^{j-1} c_{jk}y_{lk} + y_{lj})^2 < u_j^2.$$

Solving for y_{lj}, the constraint for y_{lj} becomes

$$\begin{aligned} -\sqrt{u_j^2 - \sum_{i=1}^{l-1} (\sum_{k=1}^{j-1} c_{jk}y_{ik})^2} - \sum_{k=1}^{j-1} c_{jk}y_{lk} \quad &\leq \\ y_{lj} \quad &\leq \quad \sqrt{u_j^2 - \sum_{i=1}^{l-1} (\sum_{k=1}^{j-1} c_{jk}y_{ik})^2} - \sum_{k=1}^{j-1} c_{jk}y_{1k}. \end{aligned} \tag{6}$$

So the limits for y_{ij}, which depend on

$$\mathbf{y}_{ij} = (y_{11}, y_{12}, \ldots, y_{1,j-1}, y_{21}, y_{22}, \ldots, y_{2,j-1}, \ldots, y_{i1}, \ldots, y_{i,j-1}),$$

are

$$M_{ij}^{\pm}(\mathbf{y}_{ij}) = \pm \sqrt{u_j^2 - \sum_{s=1}^{i-1} (\sum_{k=1}^{j-1} c_{jk}y_{sk})^2} - \sum_{k=1}^{j-1} c_{jk}y_{ik}.$$

Using these limit expressions,

$$\begin{aligned} P(u) \quad = \quad & \int_{M_{11}^-(\mathbf{y}_{11})}^{M_{11}^+(\mathbf{y}_{11})} \phi(y_{11}) \int_{M_{12}^-(\mathbf{y}_{12})}^{M_{12}^+(\mathbf{y}_{21})} \phi(y_{12}) \int_{M_{1m}^-(\mathbf{y}_{1m})}^{M_{1m}^+(\mathbf{y}_{1m})} \phi(y_{1m}) \\ & \int_{M_{21}^-(\mathbf{y}_{21})}^{M_{21}^+(\mathbf{y}_{21})} \phi(y_{21}) \int_{M_{22}^-(\mathbf{y}_{22})}^{M_{22}^+(\mathbf{y}_{22})} \phi(y_{22}) \int_{M_{2m}^-(\mathbf{y}_{n2})}^{M_{2m}^+(\mathbf{y}_{2m})} \phi(y_{2m}) \\ & \cdots \int_{M_{d1}^-(\mathbf{y}_{d1})}^{M_{d1}^+(\mathbf{y}_{d1})} \phi(y_{d1}) \int_{M_{d2}^-(\mathbf{y}_{d2})}^{M_{d2}^+(\mathbf{y}_{d2})} \phi(y_{d2}) \int_{M_{dm}^-(\mathbf{y}_{dm})}^{M_{dm}^+(\mathbf{y}_{dm})} \phi(y_{dm}) \prod_{j=m}^{1} \prod_{i=d}^{1} dy_{ij}, \end{aligned}$$

where $\phi(y) = e^{-\frac{1}{2}y^2}/(2\pi)^{\frac{1}{2}}$, the standard univariate Normal pdf. If the changes of variables $z_{ij} = \Phi(y_{ij})$, where $\Phi(y)$ is the standard univariate Normal cdf, with $dz_{ij} = \phi(y_{ij})dy_{ij}$, are completed, then

$$\begin{aligned} I(u) &= \int_{N_{11}^-(\mathbf{z}_{11})}^{N_{11}^+(\mathbf{z}_{11})} \int_{N_{12}^-(\mathbf{z}_{12})}^{N_{12}^+(\mathbf{z}_{12})} \cdots \int_{N_{1m}^-(\mathbf{z}_{1m})}^{N_{1m}^+(\mathbf{z}_{1m})} \int_{N_{21}^-(\mathbf{z}_{21})}^{N_{21}^+(\mathbf{z}_{21})} \int_{N_{22}^-(\mathbf{z}_{22})}^{N_{22}^+(\mathbf{z}_{22})} \cdots \int_{N_{2m}^-(\mathbf{z}_{2m})}^{N_{2m}^+(\mathbf{z}_{2m})} \\ &\quad \cdots \int_{N_{d1}^-(\mathbf{z}_{d1})}^{N_{d1}^+(\mathbf{z}_{d1})} \int_{N_{d2}^-(\mathbf{z}_{d2})}^{N_{d2}^+(\mathbf{z}_{d2})} \cdots \int_{N_{dm}^-(\mathbf{z}_{dm})}^{N_{dm}^+(\mathbf{z}_{dm})} \prod_{j=m}^{1} \prod_{i=d}^{1} dz_{ij}. \end{aligned}$$

with $N_{ij}^{\pm}(\mathbf{z}_{ij}) = \Phi(M_{ij}^{\pm}(\Phi^{-1}(\mathbf{z}_{ij})))$ and $\mathbf{z}_{ij} = (z_{11}, z_{12}, \ldots, z_{1m}, \ldots, z_{i1}, \ldots, z_{i,j-1})$. Now make the final changes to $(0,1)$ variables using

$$z_{ij} = N_{ij}^-(\mathbf{z}_{ij}) + D_{ij}(\mathbf{z}_{ij})w_{ij}, \quad D_{ij}(\mathbf{z}_{ij}) = N_{ij}^+(\mathbf{z}_{ij}) - N_{ij}^-(\mathbf{z}_{ij}),$$

and then

$$\begin{aligned} P(u) &= \int_0^1 D_{11}(\mathbf{z}_{11}(\mathbf{w}_{11})) \cdots \int_0^1 D_{1m}(\mathbf{z}_{n1}(\mathbf{w}_{1m})) \\ &\quad \int_0^1 D_{21}(\mathbf{z}_{21}(\mathbf{w}_{21})) \cdots \int_0^1 D_{2m}(\mathbf{z}_{2m}(\mathbf{w}_{1m})) \\ &\quad \cdots \int_0^1 D_{d1}(\mathbf{z}_{d1}(\mathbf{w}_{d1})) \cdots \int_0^1 D_{dm}(\mathbf{z}_{nm}(\mathbf{w}_{dm})) \prod_{j=m}^{1} \prod_{i=d}^{1} dw_{ij} \\ &\equiv \int_{\mathbf{0}}^{\mathbf{1}} f(\mathbf{w}) \prod_{j=m}^{1} \prod_{i=d}^{1} dw_{ij}, \end{aligned}$$

with

$$f(\mathbf{w}) = \prod_{i=1}^{d} \prod_{j=1}^{m} (N_{ij}^+(\mathbf{z}_{ij}(\mathbf{w}_{ij})) - N_{ij}^+(\mathbf{z}_{ij}(\mathbf{w}_{ij}))),$$

so that $P(u)$ can be approximated using any numerical integration method for the unit hypercube $H^{(dm)} = [0,1]^{dm}$.

If u_p is the value of u where $P(u) = p$ for a given probability p, an approximate value for u_p can be determined by applying a numerical root-finding method (e.g. the the bisection or secant method) to function $h(u) = P(u) - p$.

2.3. Numerical Integration

A simple Monte Carlo (MC) method for the approximate computation of $P(u)$, which uses $U(0,1)$ random numbers, takes the form

$$P(u) \approx P_N = \frac{1}{N}\sum_{k=1}^{N} f(\mathbf{W}_k),$$

with standard error

$$E_N = (\frac{1}{N(N-1)}\sum_{k=1}^{N}(f(\mathbf{W}_k) - P_N)^2)^{\frac{1}{2}},$$

with all components of $\mathbf{W}_k \sim U(0,1)$.

MC methods using N points have errors that are typically $O(1/N^{\frac{1}{2}})$, so quasi-Monte Carlo (QMC) methods (see Fox (1999)), with asymptotic errors which can be approximately $O(1/N)$ for N points, are often be used to provide improved simulation approximations for these kinds of computations. Given a set of QMC points $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_M$ from $H^{(dm)}$, a typical M-point QMC method for $P(u)$ uses

$$P(u) \approx Q_M = \frac{1}{M}\sum_{s=1}^{M} f(\mathbf{Z}_s).$$

Error estimates for Q_M can be computed if the QMC method is randomized. A simple method for randomization uses random shifts of the QMC approximations in the form

$$Q_M(\mathbf{W}) = \frac{1}{M}\sum_{s=1}^{M} f(\{\mathbf{Z}_s + \mathbf{W}\}),$$

where $\mathbf{W}$ has random $U(0,1)$ components and $\{\mathbf{X}\}$ denotes the vector of fractional parts of the components in $\mathbf{X}$. Then an MCQMC approxmation for $P(u)$ is given by

$$P(u) \approx Q_{N,M} = \frac{1}{N}\sum_{k=1}^{N} Q_M(\mathbf{W}_k),$$

with standard error

$$E_{N,M} = (\frac{1}{N(N-1)} \sum_{k=1}^{N} (Q_M(\mathbf{W}_k) - Q_{N,M})^2)^{\frac{1}{2}}.$$

For these approximations, N is usually chosen to be small (e.g. $N = 12$) relative to M.

3. Illustration of Davies upper bound and the discretized lower bound

In order to illustrate Davies upper bound and the MCQMC lower bound, we have to choose a process suitable for Davies assumptions : we will consider the Chi-Square process where each $V_i(.)$ is a stationnary process with covariance function $r(h) = e^{-h^2/2}$. Besides, as mentioned previously, we consider independent processes $V_i(.)$.

As the processes $V_i(.)$ are independant, we have $\forall j \neq i$:

$$\mathrm{Cov}(V_i'(t), V_j'(t)) = 0 \text{ and } \mathrm{Cov}(V_i'(t), V_j(t)) = 0$$

Since $V_i'(t) = \lim_{h\to 0} \frac{V_i(t+h)-V_i(t)}{h}$, then $\mathbb{E}[V_i'(t)] = 0$. Besides, $r''(0) = -1$, so $\mathrm{Var}(V_i'(t)) = 1$.

On the other hand,

$$\mathbb{E}[V_i(t)V_i'(t)] = \lim_{h\to 0} \frac{\mathbb{E}[V_i(t)V_i(t+h)]}{h} - \lim_{h\to 0} \frac{\mathbb{E}[V_i^2(t)]}{h} = \lim_{h\to 0} \frac{r(h)-1}{h} = 0$$

As a consequence, if we denote $\vec{V}(t) = \begin{pmatrix} V_1(t) \\ \vdots \\ V_d(t) \end{pmatrix}$, $\vec{V}'(t) = \begin{pmatrix} V_1'(t) \\ \vdots \\ V_d'(t) \end{pmatrix}$ and I_{2d} the identity matrix $2d \times 2d$, then

$$\mathrm{Var} \begin{pmatrix} \vec{V}(t) \\ \vec{V}'(t) \end{pmatrix} = I_{2d} \tag{7}$$

As the covariance matrix (7) is the identity, we can use the following result of Davies :

$$\mathbb{E}(\|\eta\|) = 2^{1/2}\, \Gamma(d/2 + 1/2)/\Gamma(d/2)$$

α	10%		5%		1%	
Method / $[\mathcal{L},\, \mathcal{U}]$	MCQMC	DAV	MCQMC	DAV	MCQMC	DAV
(0, 2)	6.28	6.87	8.34	8.39	11.72	11.86
(0, 5)	8.32	8.44	9.90	9.97	13.34	13.45
(0, 10)	9.67	9.81	11.22	11.33	14.76	14.80

Table 1: Bounds for the critical value c of the Chi-Square process with 2 degrees of freedom, as a function of the level of the test α and the interval $[\mathcal{L},\, \mathcal{U}]$ considered. The upper bound refers to Davies' method (DAV) whereas the lower bound to MCQMC.

α	10%		5%		1%	
Method / $[\mathcal{L},\, \mathcal{U}]$	MCQMC	DAV	MCQMC	DAV	MCQMC	DAV
(0, 2)	10.54	10.61	12.33	12.40	16.27	16.36
(0, 5)	12.34	12.48	14.15	14.24	18.03	18.15
(0, 10)	13.92	14.07	15.67	15.79	19.55	19.65

Table 2: Bounds for the critical value c of the Chi-Square process with 4 degrees of freedom, as a function of the level of the test α and the interval $[\mathcal{L},\, \mathcal{U}]$ considered. The upper bound refers to Davies' method (DAV) whereas the lower bound to MCQMC.

Then, according to formula (3) :

$$\mathbb{P}(\sup_{t\in[\mathcal{L},\,\mathcal{U}]} S(t) > c) \leq \mathbb{P}(\chi^2_d > c) \;+\; (\mathcal{U}-\mathcal{L})\, c^{(d-1)/2}\, e^{-c/2}\, \pi^{-1/2}\, 2^{(1-d)/2}/\Gamma(d/2)$$

In order to perform a test at the level α, the challenge is to obtain the critical value c such as $\mathbb{P}(\sup_{t\in[\mathcal{L},\,\mathcal{U}]}\ S(t) > c) = \alpha$: the MCQMC bound and Davies' bound allows us to obtain respectively a lower and an upper bound for the critical value c.

Tables 1, 2 and 3 gives these bounds as a function of the level of the test α, the interval $[\mathcal{L},\, \mathcal{U}]$ and the number of degrees of freedom d of the Chi-Square process. The discretization used for the MCQMC method results in these tables was $h = \frac{1}{4}$ (so $m = 4\,(\mathcal{U}-\mathcal{L})$) for all cases.

α	10%		5%		1%	
Method / $[\mathcal{L},\, \mathcal{U}]$	MCQMC	DAV	MCQMC	DAV	MCQMC	DAV
(0, 2)	15.34	15.45	17.44	17.53	21.94	22.03
(0, 5)	17.47	17.64	19.52	19.65	23.92	24.04
(0, 10)	19.26	19.46	21.27	21.41	25.69	25.71

Table 3: Bounds for the critical value c of the Chi-Square process with 7 degrees of freedom, as a function of the level of the test α and the interval $[\mathcal{L},\, \mathcal{U}]$ considered. The upper bound refers to Davies' method (DAV) whereas the lower bound to MCQMC.

4. About the supremum of a non regular Chi-Square process : the Ornstein-Uhlenbeck process

In this Section, the interest is on the critical value for the supremum of a non regular Chi-Square process : the Ornstein-Uhlenbeck Chi-Square process. It corresponds to the process $S(.)$ of formula (2) with the processes $V_i(.)$ independent and $\mathrm{Cov}(V_i(t), V_i(t')) = e^{-2|t-t'|}$. Such a process has a direct application in genetics, in particular in Quantitative Trait Locus (QTL) detection.

4.1. The Ornstein-Uhlenbeck Chi-Square process in the QTL detection

A QTL denotes a gene with quantitative effect on a trait. The method used by most of geneticists in order to detect a QTL on a chromosome, is the Interval Mapping proposed by Lander and Botstein (1989). Using the Haldane (1919) distance and modelling, each chromosome is represented by a segment $[0,\mathcal{U}]$. The distance on $[0,\mathcal{U}]$ is called the genetic distance (measured in Morgans). At each location $t \in [0,\mathcal{U}]$, using the "genome information" brought by genetic markers, a likelihood ratio test (LRT) is performed, testing the presence of a QTL at this position. So, multi-testing leads to a LRT process, and taking as test statistic the supremum of this process comes down to perform a LRT in a model when the localisation of the QTL is an extra parameter.

In Rabier et al. (2009), is considered a population of progenies which are structured into families of sires. The authors prove that when the number of genetic markers and the number of progenies tends to infinity, the limiting process of the LRT process is an Ornstein-Uhlenbeck Chi-Square process (the number of degrees of freedom corresponds to the number of sires) under the

null hypothesis of the absence of QTL on the interval $[0,\mathcal{U}]$. So, in order to take decision about the presence of a QTL on $[0,\mathcal{U}]$, the critical value for the supremum of an Ornstein-Uhlenbeck Chi-Square process has to be calculated.

4.2. Critical value calculation for the Ornstein-Uhlenbeck Chi-Square process

Let call OU an Ornstein-Uhlenbeck process and OUCS(d) an Ornstein-Uhlenbeck Chi-Square process with d degrees of freedom. We propose here different ways to calculate the critical value of the supremum of an OUCS(d).

Since an OU is an AR(1) process, an OUCS(d) is the sum of d independent AR(1) processes. As a consequence, the critical value can easily be obtained using a Monte-Carlo method.

On the other hand, an upper bound can be obtained using the MCQMC lower bound introduced in Section 2.

Finally, we propose a formula in order to calculate the critical value theoretically.

Let $\vec{W}(t) = \begin{pmatrix} W_1(t) \\ \vdots \\ W_d(t) \end{pmatrix}$ a brownian motion in dimension d and $\vec{X}(t) = \begin{pmatrix} X_1(t) \\ \vdots \\ X_d(t) \end{pmatrix}$ the process such as $\forall t\ \forall i,\ \ X_i(t) = \frac{W_i(e^{2t})}{e^t}$. We can remark that $\mathrm{Cov}(X_i(t), X_i(t')) = e^{-|t-t'|}$ and that reciproquely $W_i(t) = \sqrt{t} X_i(\frac{\log(t)}{2})$.
Besides :

$$\left\|\vec{X}(t)\right\|^2 = e^{-2t}\left\|\vec{W}(e^{2t})\right\|^2$$

We can remark that $\mathrm{Cov}(X_i(2t), X_i(2t')) = e^{-2|t-t'|}$ which corresponds to the covariance of an OU. So, we impose $V_i(t) = X_i(2t)$. Let $T \in \mathbb{R}^{+\star}$, it comes :

$$\sup_{t\in[0,\ \frac{\log(T)}{4}]} S(t) = \sup_{t\in[0,\ \frac{\log(T)}{4}]} \left\|\vec{V}(t)\right\|^2 = \sup_{t\in[0,\ \frac{\log(T)}{2}]} \left\|\vec{X}(t)\right\|^2 = \sup_{t\in[1,\ T]} \left(\frac{\left\|\vec{W}(t)\right\|}{\sqrt{t}}\right)^2 \tag{8}$$

Note that here $\mathcal{L}$ and $\mathcal{U}$ of formula (1) are respectively equal to 0 and $\log(T)/4$. This is convenient to deal with this case in order to relate with Delong (1981)'s work. Indeed, in Delong (1981), there are some important results about :

$$\mathbb{P}\left(\sup_{t\in[1,\,T]} \frac{\left\|\vec{W}(t)\right\|}{\sqrt{t}} \;<\; c\right)$$

In order to calculate this quantity, Delong uses very difficult methods. His results are presented in some exact tables. As a consequence, using formula (8), it is easy to calculate exact critical values for the supremum of the process $S(.)$.
In his article, Delong also gives an approximative formula (cf. page 2205 of the article) suitable for c and T large :

$$\mathbb{P}\left(\sup_{t\in[1,\,T]} \frac{\left\|\vec{W}(t)\right\|}{\sqrt{t}} < c\right) = \frac{(c^2/2)^{d/2}\, e^{-c^2/2}}{\Gamma(d/2)} \left[\log(T)(1-\frac{d}{c^2}) + \frac{2}{c^2} + O(\frac{1}{c^4})\right] \tag{9}$$

According to formulas (8) and (9), we can deduce an approximative formula for the process $S(.)$ suitable for c and T large :

$$\mathbb{P}\left(\sup_{t\in[0,\,\frac{\log(T)}{4}]} S(t) \;<\; c\right) = \frac{(c/2)^{d/2}\, e^{-c/2}}{\Gamma(d/2)} \left[\log(T)(1-\frac{d}{c}) + \frac{2}{c} + O(\frac{1}{c^2})\right] \tag{10}$$

With the help of a Newton's method, it becomes easy to obtain the critical value c corresponding to a test at the α level, that is to say a test such as $\mathbb{P}\left(\sup_{t\in[0,\,\frac{\log(T)}{4}]} S(t) \;>\; c\right) = \alpha$.

A numerical study is presented in Tables 4, 5, 6, and 7. Critical values for the process $S(.)$ are calculated according to the different methods :

- DE refers to Delong Exact table (based on formula (8))
- DF refers to Delong Approximative Formula (based on formula (10))

- MC refers to the Monte-Carlo method using AR(1) processes
- MCQMC refers to the MCQMC lower bound

Note that Davies bound has not been considered since it is only suitable when the processes $V_i(.)$ have a derivative with a finite number of jumps, which is not the case here.

The processes $S(.)$ studied in Tables 4 to 7 are respectively the OUCS(4), the OUCS(5), the OUCS(6) and the OUCS(7). Note that since the exact tables of Delong are available only for $d \leq 4$, DE has only been computed in Table 4. The MCQMC bounds were computed using the discretization stepsize $h = \log(T)/256$ (so $m = 64$).

4.3. Discussion

in progress

α	10%				5%				1%			
Method / $(T,\ \frac{log(T)}{4})$	DE	DF	MC	MCQMC	DE	DF	MC	MCQMC	DE	DF	MC	MCQMC
(20, 0.75)	14.21	14.09	14.21	13.28	16.16	16.10	16.08	15.16	20.43	20.58	20.16	19.27
(30, 0.85)	14.52	14.44	14.52	13.49	16.48	16.43	16.56	15.41	20.70	20.89	20.34	19.46
(40, 0.92)	14.67	14.66	14.67	13.65	16.65	16.65	16.65	15.58	20.88	21.10	20.70	19.63
(50, 0.98)	14.82	14.83	14.75	13.76	16.81	16.81	16.65	15.64	20.98	21.23	20.88	19.75
(60, 1.02)	14.90	14.95	14.82	13.86	16.89	16.93	16.65	15.71	21.07	21.34	21.25	19.84
(70, 1.06)	15.05	15.05	15.98	13.91	16.97	17.02	16.81	15.79	21.16	21.44	20.79	19.91
(80, 1.10)	15.13	15.14	15.13	13.97	17.06	17.11	17.06	15.85	21.25	21.51	21.44	19.98
(100, 1.15)	15.21	15.28	15.13	14.05	17.14	17.23	17.22	15.94	21.34	21.64	21.07	20.05

Table 4: Critical values for the OUCS(4) as a function of the level α of the test and the interval considered. The critical values c correspond to $\mathbb{P}\left(\sup_{t\in[0,\ \frac{\log(T)}{4}]} S(t)\ >\ c\right) = \alpha$ where the process $S(.)$ is an OUCS(4). These values are calculated by different methods : DE (Delong Exact table), DF (Delong Approximate Formula), MC (Monte-Carlo method using AR(1) processes), MCQMC. For the simulation of the four Autoregressive processes, 10000 paths have been generated with a step of 10^{-5}.

α	10%			5%			1%		
Method / $(T,\ \frac{log(T)}{4})$	DF	MC	MCQMC	DF	MC	MCQMC	DF	MC	MCQMC
(20, 0.75)	16.02	16.00	15.13	18.13	18.06	17.13	22.78	22.75	21.44
(30, 0.85)	16.39	16.48	15.37	18.48	18.58	17.38	23.11	23.23	21.65
(40, 0.92)	16.63	16.65	15.55	18.71	18.84	17.50	23.31	23.14	21.74
(50, 0.98)	16.80	16.81	15.62	18.86	18.92	17.66	23.46	23.14	21.89
(60, 1.02)	16.93	16.89	15.67	18.99	19.01	17.69	23.58	23.33	21.95
(70, 1.06)	17.04	17.06	15.81	19.09	19.10	17.75	23.67	23.52	22.00
(80, 1.10)	17.13	16.97	15.85	19.18	19.10	17.85	23.75	23.62	22.07
(100, 1.15)	17.27	17.22	15.95	19.32	19.18	17.94	23.87	23.62	22.16

Table 5: Critical values for the OUCS(5) as a function of the level α of the test and the interval considered. The critical values c correspond to $\mathbb{P}\left(\sup_{t\in[0,\ \frac{\log(T)}{4}]} S(t) > c\right) = \alpha$ where the process $S(.)$ is an OUCS(5). These values are calculated by different methods : DE (Delong Exact table), DF (Delong Approximate Formula), AR (Monte-Carlo method using AR(1) processes), MCQMC. For the simulation of the five Autoregressive processes, 10000 paths have been generated with a step of 10^{-5}.

α	10%			5%			1%		
Method / $(T,\ \frac{log(T)}{4})$	DF	MC	MCQMC	DF	MC	MCQMC	DF	MC	MCQMC
(20, 0.75)	17.86	17.81	16.90	20.05	19.89	18.99	24.86	24.40	23.44
(30, 0.85)	18.24	18.23	17.17	20.41	20.34	19.22	25.20	25.00	23.65
(40, 0.92)	18.49	18.20	17.29	20.65	20.52	19.36	25.41	24.90	23.85
(50, 0.98)	18.67	18.66	17.41	20.81	20.79	19.50	25.56	25.10	23.92
(60, 1.02)	18.81	18.66	17.52	20.94	20.70	19.61	25.68	25.10	23.99
(70, 1.06)	18.92	18.92	17.60	21.05	20.98	19.64	25.78	25.20	24.06
(80, 1.10)	19.02	18.84	17.69	21.13	21.07	19.77	25.86	25.20	24.12
(100, 1.15)	19.17	19.18	17.76	21.28	21.44	19.80	25.99	25.91	24.20

Table 6: Critical values for the OUCS(6) as a function of the level α of the test and the interval considered. The critical values c correspond to $\mathbb{P}\left(\sup_{t\in[0,\ \frac{\log(T)}{4}]} S(t) > c\right) = \alpha$ where the process $S(.)$ is an OUCS(6). These values are calculated by different methods : DE (Delong Exact table), DF (Delong Approximate Formula), MC (Monte-Carlo method using AR(1) processes), MCQMC. For the simulation of the six Autoregressive processes, 10000 paths have been generated with a step of 10^{-5}.

α	10%			5%			1%		
Method / $(T,\ \frac{log(T)}{4})$	DF	MC	MCQMC	DF	MC	MCQMC	DF	MC	MCQMC
(20, 0.75)	19.62	19.71	18.61	21.88	21.81	20.75	26.85	26.73	25.34
(30, 0.85)	20.03	19.80	18.87	22.27	22.09	21.03	27.20	26.52	25.53
(40, 0.92)	20.28	19.80	19.03	22.51	22.28	21.19	27.42	27.03	25.70
(50, 0.98)	20.47	20.52	19.13	22.68	22.56	21.28	27.57	27.25	25.80
(60, 1.02)	20.61	20.52	19.26	22.81	22.66	21.37	27.70	27.46	25.87
(70, 1.06)	20.72	20.52	19.35	22.92	22.66	21.40	27.79	27.20	25.95
(80, 1.10)	20.82	20.70	19.40	23.01	23.04	21.51	27.88	27.88	25.99
(100, 1.15)	20.98	20.79	19.47	23.16	23.04	21.61	28.02	27.67	26.08

Table 7: Critical values for the OUCS(7) as a function of the level α of the test and the interval considered. The critical values c correspond to $\mathbb{P}\left(\sup_{t\in[0,\ \frac{\log(T)}{4}]} S(t) > c\right) = \alpha$ where the process $S(.)$ is an OUCS(7). These values are calculated by different methods : DE (Delong Exact table), DF (Delong Approximate Formula), MC (Monte-Carlo method using AR(1) processes), MCQMC. For the simulation of the seven Autoregressive processes, 10000 paths have been generated with a step of 10^{-5}.

References

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

Delong, D. M., (1981) Crossing probabilities for a square root boundary by a Bessel process. *Commun. Statist.-Theor. Meth.*, **A10(21)**, 2197-2213.

Fishman, G. S., (1996) *Monte Carlo: Concepts, Algorithms, and Applications*, Springer-Verlag, 68.

Fox, B. L., (1999) *Strategies for Quasi-Monte Carlo* (International Series in Operations Research & Management Science, 22), Kluwer Academic Publishers.

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

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., Azaïs, J-M., Delmas, C. (2009) Likelihood Ratio Test for Quantitative Trait Loci detection. *hal-00421215*.

Sloan, I. H. and Joe, S. (1994) *Lattice Methods for Multiple Integration*. Oxford University Press.

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

Conclusion et perspectives

Dans ce travail, nous nous sommes attachés à étudier les théories statistiques sous-jacentes qui sont utilisées en détection de QTL. On a tout d'abord étudié une technique d'amélioration de protocole : le selective genotyping. Ce dispositif consiste à ne génotyper que les individus extrêmes. Puis, dans un deuxième temps, on a étudié le génome scan où l'on recherche des QTL en balayant le génome.
La puissance asymptotique de ces deux approches a été calculée analytiquement et la convergence vers l'asymptotique testée numériquement. On montre que le régime asymptotique peut être obtenu avec des tailles de populations raisonnables.
La théorie développée dans le cadre du génome scan est valable pour des phénotypes non normaux, tandis que le selective genotyping est sensible à la normalité.

L'orginalité de notre approche concernant le selective genotyping réside dans le modèle lui-même. On considère en effet deux seuils fixes S_- et S_+. Un individu est génotypé uniquement si la valeur du caractère quantitatif Y n'appartient pas à l'intervalle $[S_-,\ S_+]$. En choisissant S_+ et S_- de telle sorte que $\mathbb{P}_{H_0}(Y \notin [S_-, S_+]) = \gamma$, par la loi des grands nombres, le pourcentage d'individus génotypés tend asymptotiquement vers γ, que l'on se trouve sous l'hypothèse nulle ou sous l'alternative locale. Ainsi cette modélisation est en accord avec la définition usuelle du selective genotyping : le selective genotyping consiste à génotyper uniquement les $\gamma\%$ de la population présentant des phénotypes extrêmes.

Darvasi and Soller (1992) considèrent des seuils qui varient avec l'effet du QTL : la théorie qu'ils emploient n'est pas cohérente avec cette hypothèse. Quant à nous, nous ne nous limitons pas à une population backcross et nous ne supposons pas de symétrie quant au génotypage (S_- et S_+ sont quelconques), ce qui nous permet notamment de nous intéresser à la question de l'optimisation du génotypage.

Lorsque l'on considère un modèle statistique à trois paramètres (μ, q, σ) ou deux paramètres (μ, q) (μ est l'espérance des phénotype en l'absence de QTL, q l'effet du QTL, et σ^2 est la variance intra génotype au QTL), alors on montre qu'il n'y a aucun gain de puissance à considérer les phénotypes non extrêmes dans l'analyse statistique quelle que soit la proportion p des deux allèles au QTL dans la population.

Cependant, lorsque l'on considère un modèle à un paramètre (q), il n'y a pas de contribution des phénotypes non extrêmes uniquement pour $p = 1/2$, cas d'une population

backcross. On présente plusieurs tests pour le selective genotyping, notamment un test très simple à mettre en oeuvre car il est basé sur une simple comparaison de moyenne. Si on souhaite génotyper uniquement un pourcentage donné de la population, on montre que l'on doit génotyper le même pourcentage d'individus aux deux extrêmes de la population.

On obtient les mêmes conclusions, quant à la contribution des phénotypes non extrêmes et la stratégie de génotypage, pour un selective genotyping en présence de deux caractères corrélés. Seule différence, si l'effet du QTL sur le premier caractère est connu, il n'est pas nécessaire de génotyper de manière symétrique.

En ce qui concerne les résultats sur le génome scan, ils ont été obtenus pour n'importe quelle carte génétique et pour des populations structurées en familles de pères. On a considéré le processus de tests de rapport de vraisemblance (LRT) en référence au test d'absence de QTL sur un intervalle $[0, T]$ représentant un chromosome. On donne la distribution asymptotique du processus de LRT sous l'hypothèse nulle d'absence de QTL sur $[0, T]$, et sous l'alternative qu'il existe un QTL à une position $t^\star$ appartenant à $[0, T]$. Grâce à ces résultats théoriques, on propose de nouvelles méthodes permettant de calculer les seuils et puissances pour la détection de QTL, en utilisant le supremum du processus. Ces méthodes sont rapides et faciles à implémenter. On montre qu'il est souhaitable d'inclure si possible dans l'analyse statistique seulement les familles avec les plus gros effets QTL. Par la comparaison d'une procédure tests multiples et d'un test global, on montre également qu'il est preférable d'analyser les familles simultanément. Dans le cadre d'une famille, on prouve qu'il s'avère inutile d'effectuer des tests sur l'ensemble du chromosome, mais que l'on doit au contraire considérer uniquement quelques positions bien précises.

Enfin, on propose des résultats asymptotiques pour des populations structurées en familles de pères avec chacune leurs propres marqueurs informatifs. Ces processus pourraient approcher des populations outbred.

Nous obtenons également la distribution asymptotique du processus de LRT sous l'alternative générale qu'il existe m QTL sur $[0, T]$. L'étude de cette alternative nous permettra dans un travail en cours de proposer une nouvelle approche non pas basée sur le supremum du processus de LRT comme habituellement mais sur l'ensemble du processus. On pourra ainsi estimer le nombre de QTL, leurs positions et leur effets, en utilisant des techniques de vraisemblance pénalisée ou en recherchant les zéros d'un vecteur Gaussien. On s'intéressera par la même occasion à la détection de QTL en interaction afin de relacher l'hypothèse d'additivité quant aux effets des QTL.

Ce travail est susceptible de nombreux prolongements. Dans le cadre du génome scan, nous nous sommes restreints à une étude de liaison pour des populations structurées en familles de pères. On pourrait imaginer effectuer une étude d'association en abandonnant la structure de famille. Les diverses modélisations du déséquilibre de liaison présentes

dans la littérature pourraient servir à calculer les covariances du processus de détection. Il est à noter que les résultats sur le selective genotyping peuvent être employés dans d'autres domaines où l'analyse de données est cruciale mais contrainte par des impératifs économiques (aéronautique par exemple).

Bibliographie

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

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

Candes, E. J., Tao, T. (2005). The Dantzig selector : statistical estimation when p is much larger than n. *Annals of Statistics*, **35**, 2313-2351.

Cierco, C. (1996). Problèmes statistiques liés à la détection et à la localisation d'un gène à effet quantitatif. Thèse de Doctorat, Université Paul-Sabatier, Toulouse.

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

Churchill, G.A., Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping *Genetics*, **138**, 963-971.

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.

Delong, D. M. (1981). Crossing probabilities for a square root boundary by a Bessel process. *Commun. Statist.-Theor. Meth.*, **A10(21)**, 2197-2213.

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.

Fisher, R.A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Phil. Trans. Roy. Soc. of Edinburgh* **52**, 399-433.

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

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.

Haley, C., Knott, S.A. (1992). A simple regression method for mapping quantitative trait loci by using molecular markers. *Heredity* **69**, 315-324.

Haley, C., Knott, S.A., Elsen, J.M. (1994). Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*, **136**, 1195-1207.

Hayes, B., Goddard, M.E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *G.S.E.* **33**, 209-229.

Hayes, B. (2005). *Gene detection and marker assisted selection : putting the theory into practice*. Report.

Kosambi, D.D. (1944). The estimation of map distance from recombination values. *Annals of Eugenics*, **12**, 172-175.

Knapp, S.J., Bridges, W.C., Birkes, D. (1990). Mapping quantitative trait loci using molecular marker linkage maps. *Theoretical and Applied Genetics*, **79**, 583-592.

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.

Li, W.V., Shao, Q-M. (2002). A normal comparison inequality and its applications. Probab. Theory Related Fields, **122(4)**, 494-508.

Lynch, M., Walsh, B. (2007). *Genetics and Analysis of Quantitative Traits*, Sinauer.

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.

Piepho, H-P. (2001). A quick method for computing approximate thresholds fo quantitative trait loci detection Genetics **157**, 425-432.

Plackett, R.I. (1954). A reduction formula for normal multivariate integrals. *Biometrika*, **41**, 351-360.

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 anlysis : tests and confidence intervals. *Springer, New York.*

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

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.

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

TITLE :

Statistical techniques for the detection of genes with quantitative effects on traits

ABSTRACT :

This works aims at studying and proposing statistical techniques dedicated to the detection and the localization of loci responsible for the variation of a quantitative character (Quantitative Trait Loci : QTL). A first part is devoted to selective genotyping (a process allowing to reduce genotyping costs) and the second part to the genome scan (a technique allowing to detect QTL when scanning the genome). In the first part, using contiguity arguments and efficiency calculations, it is proved that there is no power benefit to consider non extreme phenotypes in the statistical analysis not only in the frame of a backcross, but also in a more general context. When only a given percentage of the population has to be genotyped, it is shown that the method is optimal when genotyping the same percentage of individuals at the two extrema of the population. The same conclusions hold for a selective genotyping with two correlated characters. As far as genome scan is concerned, using Le Cam's lemmas, asymptotic results are given on the distribution of the detecting process under the null hypothesis of absence of QTL on the chromosome, under the alternative that there is only one QTL and under the general alternative where several QTL are present on the chromosome. These results have been established for populations structured in families of sires. Methods suited to the genetic map are proposed in order to calculate thresholds for the supremum of the process, required for decision aids. Finally it is shown that testing the whole chromosome is not relevant, only some very precise positions need to be tested.

AUTEUR : *RABIER Charles-Elie*

TITRE : *Techniques statistiques pour la détection de gènes à effets quantitatifs*

DIRECTEURS DE THESE : *AZAÏS Jean-Marc / ELSEN Jean-Michel*

LIEU ET DATE DE SOUTENANCE : *Institut de Mathématiques de Toulouse le 16/06/2010*

RESUME en français *:*

Devant l'afflux d'informations moléculaires, la statistique est un outil indispensable à l'analyse des données issues du génome. Dans ce contexte, l'objectif de ce travail est d'étudier et de proposer des techniques statistiques propres à la détection et à la localisation de loci responsables de la variation d'un caractère quantitatif (Quantitative Trait Loci : QTL). Une première partie est consacrée au selective genotyping (dispositif permettant de réduire les coûts dus au génotypage) et une deuxième partie au génome scan (technique permettant de détecter les QTL en scannant le génome). Dans la première partie, en utilisant des arguments de contiguïté et des calculs d'efficacité on prouve qu'il n'y a aucun gain de puissance à considérer les phénotypes non-extrêmes dans l'analyse statistique, dans le cadre d'un backcross, mais aussi dans un contexte plus général. Si on souhaite génotyper uniquement un pourcentage donné de la population, on montre qu'on doit génotyper le même pourcentage d'individus aux deux extrêmes de la population. On obtient les mêmes conclusions pour un selective genotyping en présence de deux caractères corrélés. Dans le cadre du génome scan, utilisant les lemmes de Le Cam, on présente des résultats asymptotiques sur la distribution du processus de détection sous l'hypothèse nulle d'absence de QTL, sous l'alternative où il existe un seul QTL sur le chromosome et sous l'alternative générale où plusieurs QTL sont présents sur le chromosome. On propose également quelques méthodes adaptées à la carte génétique et permettant d'obtenir le quantile du supremum du processus de détection, indispensable dans la prise de décision. Enfin, on prouve qu'il s'avère inutile d'effectuer des tests sur l'ensemble du chromosome, mais que l'on doit au contraire considérer uniquement quelques positions bien précises.

MOTS-CLES :

Modèle de mélange, Processus de Chi-Deux, Processus Gaussien, Modèles statistiques à paramètres de nuisance, Test de rapport de vraisemblance, Monte-Carlo Quasi Monte-Carlo, Statistique génétique, Détection de Quantitative Trait Loci

DISCIPLINE : *Mathématiques-Statistiques*

LABORATOIRES :

Institut de Mathématiques de Toulouse, LSP, Université Paul Sabatier (Toulouse III), UMR 5219, 118 route de Narbonne, F-31062 Toulouse Cedex 9

Station d'Amélioration Génétique des Animaux, INRA Auzeville, UR631, BP 52627, 31326 Castanet Tolosan Cedex