Etat d'avancement de la tâche 4

4 Juin 2010

Contexte

- On modélise un chromosome par un segment $[0, T]$
- Considérons tout d'abord seulement deux marqueurs génétiques A et B présents aux extrémités du chromosome
- A et B possédent chacun deux allèles (A_1 et A_2 pour A et B_1, B_2 pour B)
- Existe-t-il un QTL Q (allèles Q_1 et Q_2) sur $[0, T]$? Si oui, à quelle position ?

On s'intéressera ici au backcross

Modèle

Le phénotype Y d'un individu 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 $\varepsilon \sim N(0, \sigma^2)$

On souhaite tester :

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

L'Interval Mapping de Lander et Botstein (1989)

L'Interval Mapping

- Position du QTL inconnue

$\Rightarrow$ on scanne l'intervalle $[0, T]$.

$\Rightarrow$ tests du rapport de vraisemblance sur tout l'intervalle

Construction du LRT

- Pour chaque position $t \in [0, T]$, génotype au QTL inconnu

$\Rightarrow$ calcul des probabilités du génotype au QTL grâce aux recombinaisons et à la formule de Haldane (1919)

$\Rightarrow$ modèle de mélange

L'Interval Mapping de Lander et Botstein (1989)

- Vraisemblance pour n observations j iid :

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

où :

- $\theta = (q, \mu, \sigma)$
- $f_{(\mu,\sigma)}(.)$ densité Gaussienne de moyenne μ et de variance σ^2
- p_t^j probabilité que l'individu j soit de génotype Q_1Q_1 en t, sachant son génotype aux marqueurs A et B

L'Interval Mapping de Lander et Botstein (1989)

- $\Lambda_n(t)$ LRT à la position t
- les $\Lambda_n(t)$ définissent un processus $\Lambda_n(.)$

On recherche un seul QTL sur l'intervalle $[0, T]$

$\Rightarrow$ statistique naturelle : sup $\Lambda_n(.)$

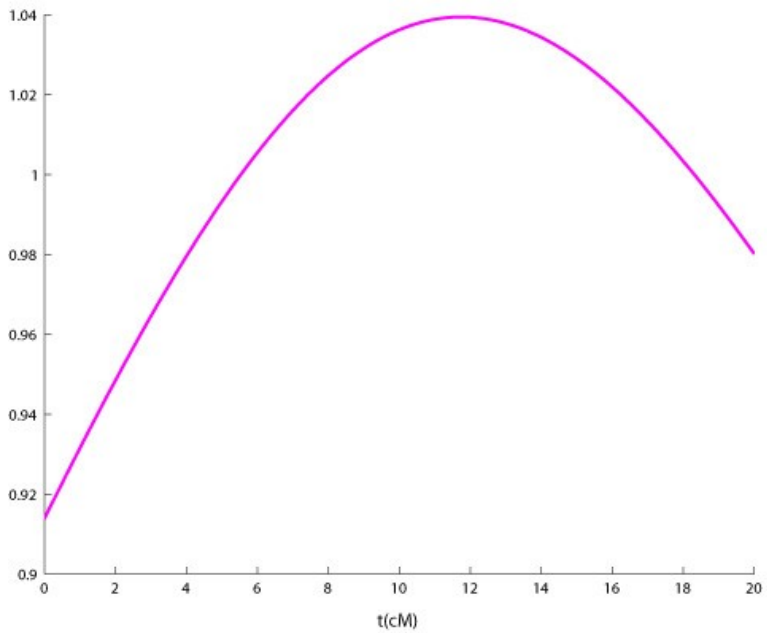


FIG.: Une trajectoire du processus $\Lambda_n(.)$ ($T = 20$cM)

Quelques précisions sur les hypothèses testées

H_0 : “il n’y a pas de QTL sur l’intervalle $[0, T]$”

$H_{at^\star}$: “le QTL est situé en $t^\star \in [0, T]$ avec un effet $q = a/\sqrt{n}$ ”

Une interpolation non linéaire

Théorème

$$\Lambda_n(.) \overset{F.d.}{\to} \{Z(.)\}^2 \qquad où$$

- *$Z(.)$ est le processus d'interpolation non linéaire tel que*

$$\forall t \in [0,T] \quad Z(t) = \frac{\alpha(t)\, Z(0) \;+\; \beta(t)\, Z(T)}{\sqrt{\{\alpha(t)\}^2 + \{\beta(t)\}^2 + 2\alpha(t)\beta(t)e^{-2T}}}$$

$$\alpha(0) = 1, \beta(0) = 0, \alpha(T) = 0, \beta(T) = 1 \quad et \quad Cov\{Z(0), Z(T)\} = e^{-2T}$$

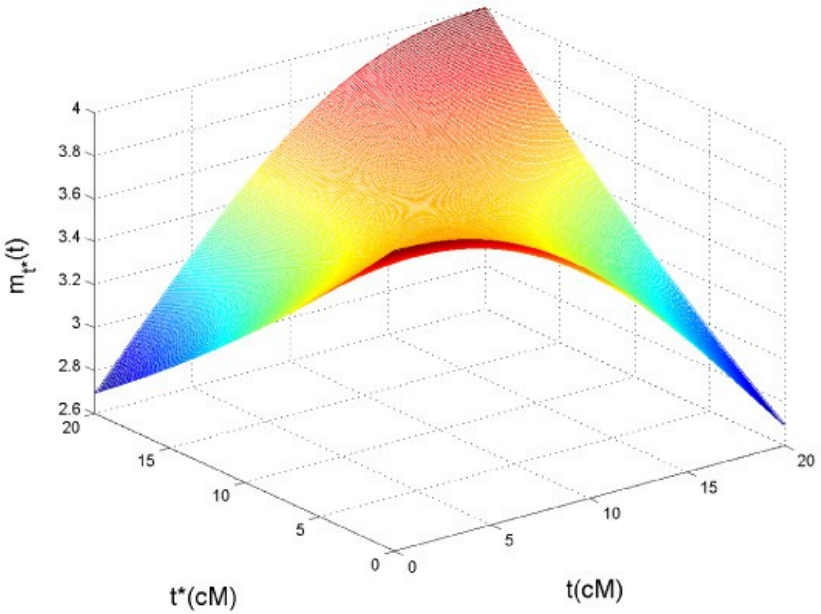
- *$Z(.)$ est un processus Gaussien de variance 1 et de fonction moyenne :*

$$sous\ H_0 : \ m(t) = 0 \ \ \forall t \in [0,T]$$

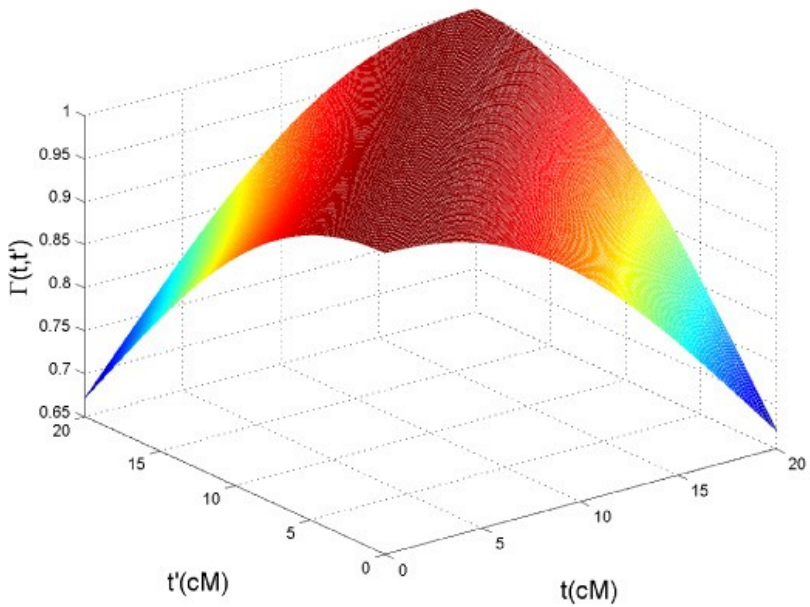
$$sous\ H_{at^\star} : \ m_{t^\star}(0) = \frac{a}{\sigma}\, g(0,t^\star) \ \ , \ \ m_{t^\star}(T) = \frac{a}{\sigma}\, h(T,t^\star)$$

$$\forall t \in [0,T] \quad m_{t^\star}(t) = \frac{\alpha(t)\, m_{t^\star}(0) \;+\; \beta(t)\, m_{t^\star}(T)}{\sqrt{\{\alpha(t)\}^2 + \{\beta(t)\}^2 + 2\alpha(t)\beta(t)e^{-2T}}}$$

Illustrations graphiques



Fonction moyenne



Fonction covariance

FIG.: Fonction moyenne et fonction covariance ($a = 4$, $\sigma = 1$, $T = 20$cM)

Illustrations graphiques

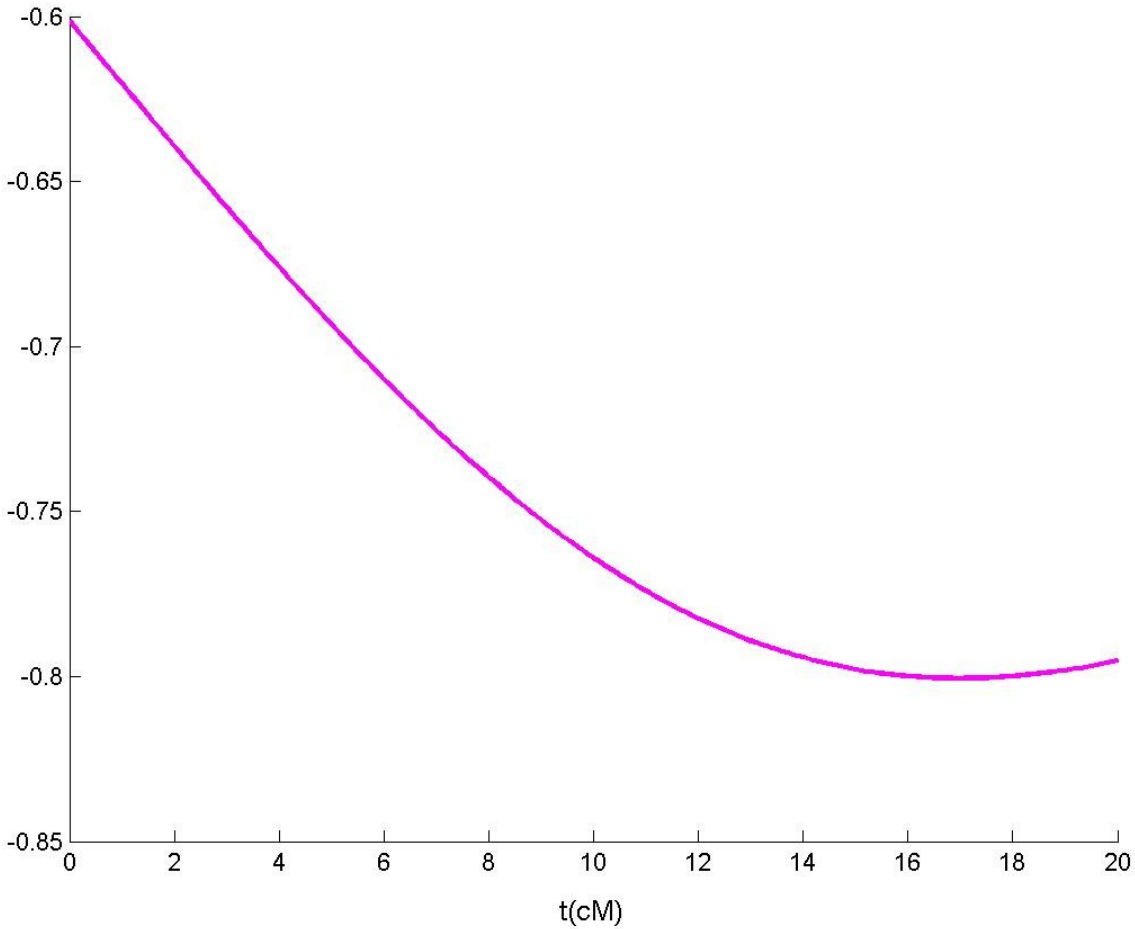


FIG.: Une trajectoire du processus $Z(.)$ sous H_0 ($T = 20$cM)

Illustrations graphiques

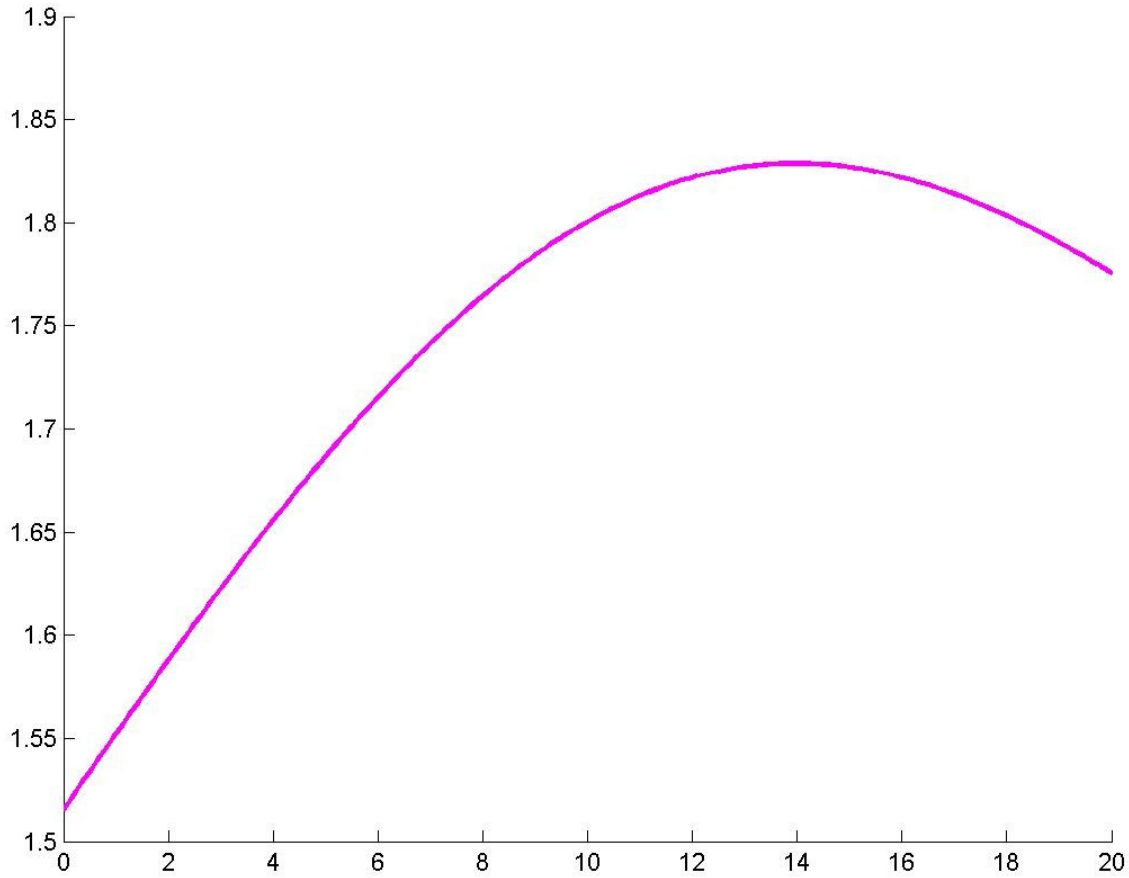


FIG.: Fonction moyenne ($a = 2$, $\sigma = 1$, $t^\star = 14$cM, $T = 20$cM)

Illustrations graphiques

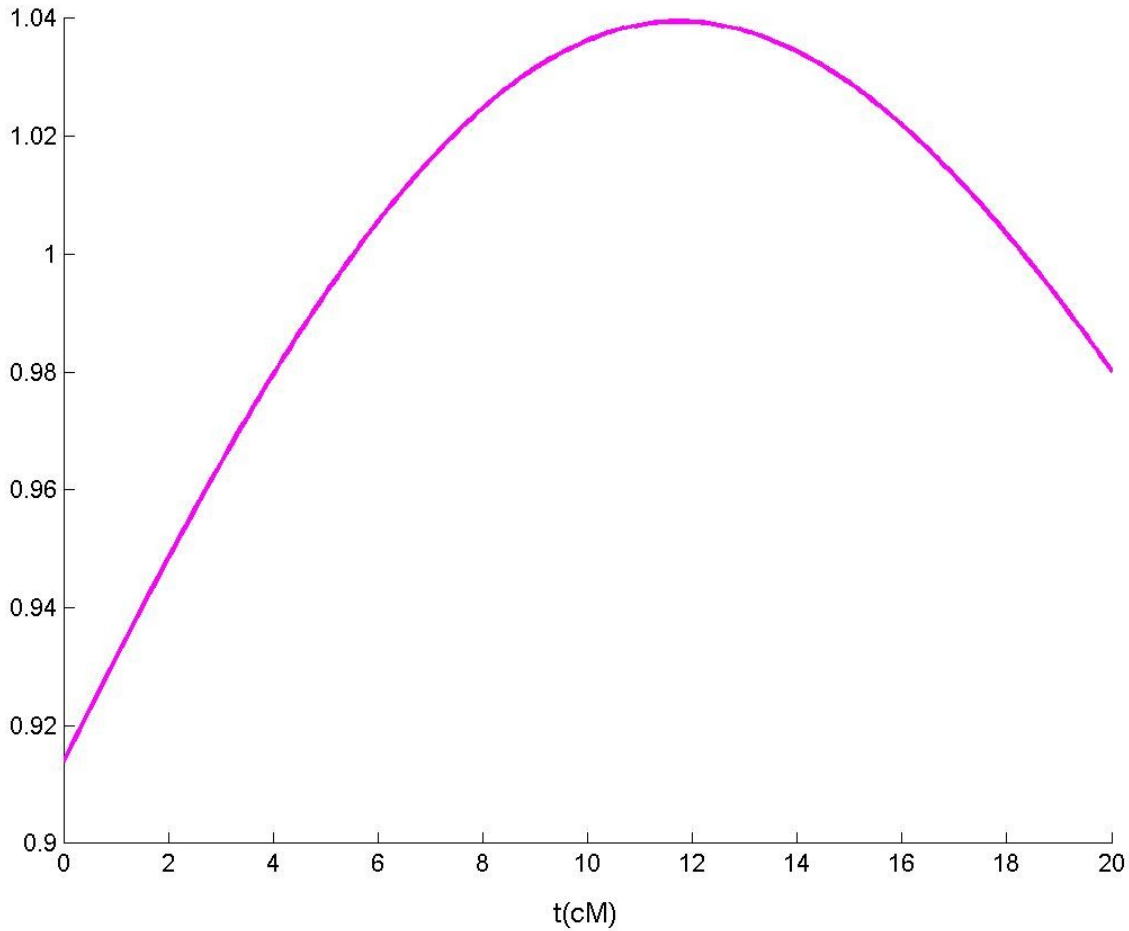


FIG.: Même trajectoire de $Z(.)$ que sous H_0 mais sous $H_{at^\star}$ ($a = 2$, $\sigma = 1$, $t^\star = 14$cM, $T = 20$cM)

A propos des tests multiples

Théorème

$$\textit{Soit} \quad \xi^2 = \frac{\{Z(0)\}^2 + \{Z(T)\}^2 - 2\, e^{-2T}\, Z(0)\, Z(T)}{\left\{1 + e^{-2T}\right\}\left\{1 - e^{-2T}\right\}} \qquad \textit{alors,}$$

$$\sup_{t\in[0,T]} \{Z(t)\}^2 = \max\left[\{Z(0)\}^2 \;,\; \xi^2\, 1_{\frac{Z(T)}{Z(0)} \in \,]\, e^{-2(T)} \,,\, e^{2(T)}\, [}\;,\; \{Z(T)\}^2\right]$$

Inutile d'effectuer des tests partout sur le chromosome ! ! !

L'Interval Mapping lisse les trajectoires

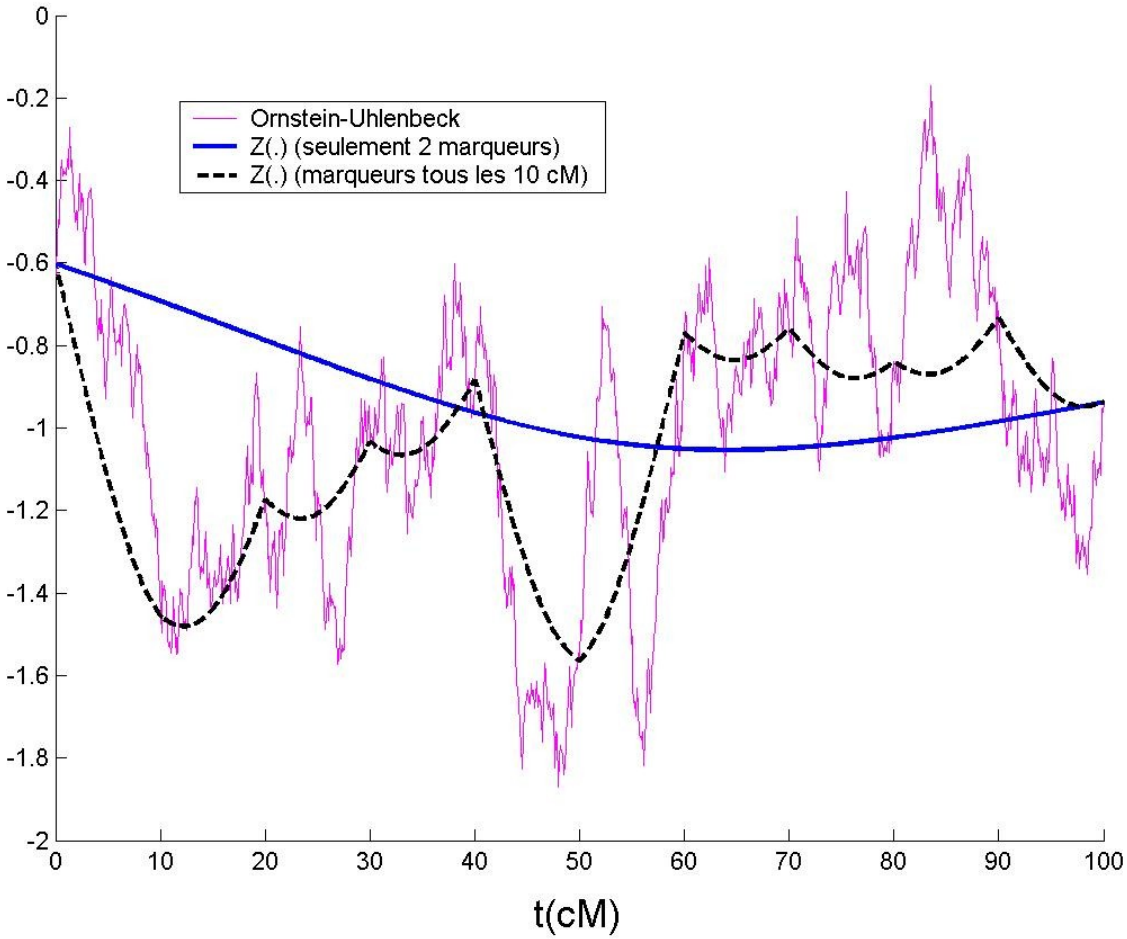


FIG.: 3 processus Gaussiens ($T = 100$cM)

Application au calcul de valeurs critiques

Calcul de la valeur critique c vérifiant

$$P_{H_0}(\sup\{Z(.)\}^2 > c) = 1 - \alpha$$

Méthode	*la notre*	*Rebaï*	*Feingold*
Valeur critique	8.23	9.09	8.26

FIG.: Valeurs critiques en fonction de la méthode considérée (51 marqueurs positionnés tous les 2cM, $T = 1$M, $\alpha = 95\%$)

Méthode	*la notre*	*Feingold*
Valeur critique	5.40	5.78

FIG.: Valeurs critiques en fonction de la méthode considérée (2 marqueurs, $T = 1$M, $\alpha = 95\%$)

Approche multi-QTL

$H_{a\vec{t}^\star}$: "il existe M QTL situés en $t_1^\star, ..., t_M^\star$ avec des effets $q_1 = \frac{a_1}{\sqrt{n}}, ..., q_M = \frac{a_M}{\sqrt{n}}$"

- on supposera les effets QTL additifs
- pour simplifier, on teste uniquement sur les marqueurs et on suppose que les QTL sont présents sur les marqueurs
- on note t_k l'emplacement du marqueur k

$\Rightarrow \Lambda_n(.)$ converge vers le carré d'un processus G sous $H_{a\vec{t}^\star}$

Definition (Processus G)

$$G(t_k) = m_{\vec{t}^\star}(t_k) + Z(t_k)$$

$$\text{où } m_{\vec{t}^\star}(t_k) = \sum_{i=1}^{M} \frac{a_i}{\sigma} exp(-2\left|t_j^\star - t_k\right|)$$

et $Z(.)$ processus d'Ornstein-Uhlenbeck

Approche multi-QTL

$\Rightarrow$ Estimation des paramètres :

- M : nombre de QTL
- $a_1, ..., a_M$: effets des QTL
- $t_1^\star, .., t_M^\star$: positions des QTL

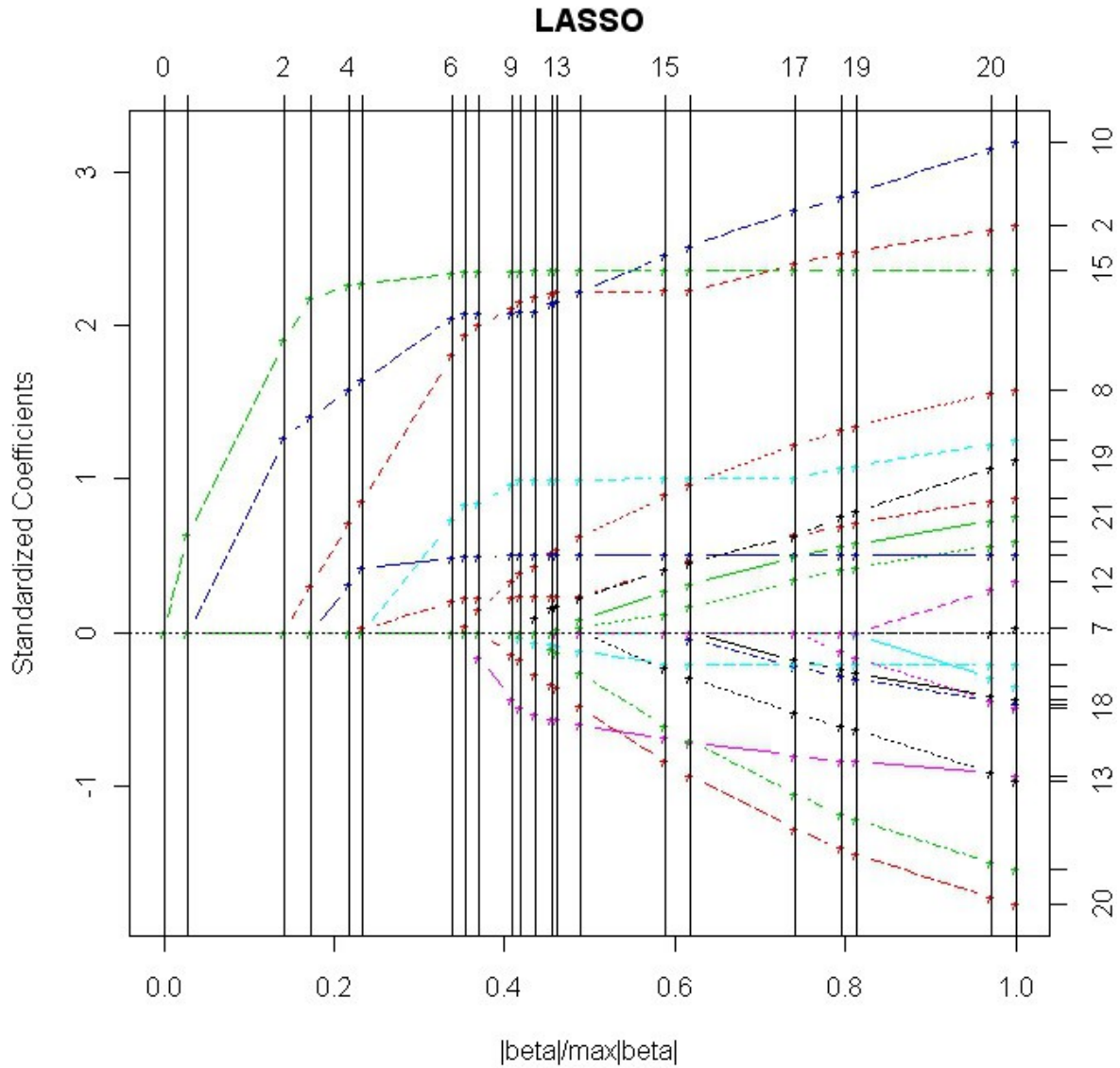
$$argmin \left\| V^{-1/2} G - V^{-1/2} m_{\vec{t}^\star} \right\|^2 + \text{pénalité}$$

avec V matrice de covariance de $Z(.)$

$\Rightarrow$ Exemple :

- $T = 100$cM, 21 marqueurs tous les 5cM
- 1000 individus
- 4 QTL sur les marqueurs 2, 10, 15, 17 avec $q_1 = ... = q_4 = 0.21$

Approche multi-QTL



Modèle

Le phénotype Y d'un individu vérifie :

$$Y = \begin{cases} \mu + \gamma + \varepsilon & \text{si pas de recombinaison entre } t_1 \text{ et } t_2 \\ \mu - \gamma + \varepsilon & \text{si recombinaison entre } t_1 \text{et } t_2 \end{cases}$$

où $\varepsilon \sim N(0, \sigma^2)$

On souhaite tester :

H_0 :γ=0 vs H_1 :$\gamma \neq 0$

pdfpa

Stratégie adoptée

$\Rightarrow$ on scanne le domaine
$\mathcal{D} = \{(t_1, t_2) \in [0, T]^2$ tels que $t_1 < t_2\}$.

$\Rightarrow$ tests du rapport de vraisemblance sur tout le domaine

- Pour chaque position $(t_1, t_2) \in \mathcal{D}$, vraisemblance pour n observations j iid :

$$L_n(\theta, t_1, t_2) = \prod_{j=1}^{n} (1 - r) f_{(\mu+\gamma,\sigma)}(y_j) \; + \; r f_{(\mu-\gamma,\sigma)}(y_j)$$

où :

- $\theta = (\gamma, \mu, \sigma)$
- $f_{(\mu,\sigma)}(.)$ densité Gaussienne de moyenne μ et de variance σ^2
- r probabilité d'avoir une recombinaison entre t_1 et t_2

Stratégie adoptée

- $\Lambda_n(t_1, t_2)$ LRT à la position (t_1, t_2)
- les $\Lambda_n(t_1, t_2)$ définissent un processus bi-dimensionnel $\Lambda_n(.)$

On recherche une interaction significative sur $\mathcal{D}$

$\Rightarrow$ statistique naturelle : sup $\Lambda_n(.)$

Objectifs :

- Caractériser le processus $\Lambda_n(.)$ sous $\mathcal{H}_0$, $\mathcal{H}_{1,t^\star}$, $\mathcal{H}_{1,t_1^\star,\cdots,t_Q^\star}$ en LA, LDLA, LD

Modèle à effets aléatoires

Soit le modèle :

$$y_i = x_i^T \beta + u_i + v_i^p + v_i^m + e_i$$

où :

- y_i est la valeur phénotypique de l'individu i
- x_i^T est un vecteur connu, β est un vecteur d'effets fixes inconnu
- u_i est un effet polygénique
- v_i^p et v_i^m sont les effets à la position testée de l'allèle d'origine paternel et de l'allèle d'origine maternel
- e_i est une erreur résiduelle

Modèle à effets aléatoires

Ce modèle se réecrit :

$$Y = X\beta + Zu + Z_q v + e$$

- Y le vecteur des valeurs phénotypiques
- X, Z et Z_q sont des matrices connues
- β est le vecteur d'effets fixes inconnu
- u est un vecteur gaussien centré de matrice de covariance $\sigma_a^2 A$ où A est une matrice connue et σ_a^2 est une variance inconnue
- v est un vecteur gaussien centré de matrice de covariance $\sigma_q^2 A_q$ où A_q est une matrice connue et σ_q^2 est une variance inconnue
- e est le vecteur des résidus

Modèle à effets aléatoires

On teste :

$$\sigma_q^2 = 0 \text{ vs } \sigma_q^2 > 0 \; \forall t \in [0, T]$$

- On note $S_n(t)$ la statistique de test au point t.
- $\forall t \in [0, T]$, la loi asymtotique de $S_n(t)$, sous l'hypothèse nulle, est $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$

C'est-à-dire :

$$S_n(t) \rightarrow \frac{1}{2} X(t)^2 1_{\{X(t) \geq 0\}}$$

où $X(\cdot)$ est un processus gaussien.

Modèle à effets aléatoires

Objectifs :

- Caractériser le processus limite X sous $\mathcal{H}_0$, $\mathcal{H}_{1,t^\star}$, $\mathcal{H}_{1,t_1^\star,\cdots,t_Q^\star}$ en LA, LDLA, LD