Modèles de mélange et Extrêmes pour la localisation de gènes

Céline Delmas [1] & Charles-Elie Rabier [2]

[1] *INRA, UR 875 MIAT, BP 52627, 31326 Castanet-Tolosan Cedex, France*
Celine.Delmas@toulouse.inra.fr
[2] *INSA, 135 Avenue de Rangueil, 31400 Toulouse*
cerabier@insa-toulouse.fr

Résumé. Nous introduisons une nouvelle méthode de sélection de variables qui nous permet de sélectionner plus de variables que d'observations contrairement aux méthodes classiques telles que le LASSO. Cela est rendu possible par la construction d'un test statistique, une transformation des données et par la connaissance de la corrélation entre les régresseurs. Nous prouvons que le ratio signal sur bruit est largement amélioré en considérant les extrêmes. Cette technique est développée dans le cadre d'un problème de détection de gènes en génétique. Les interactions entre les régresseurs n'influent pas sur la méthode. Des illustrations sont données sur des simulations.

Mots-clés. Détection de gènes, Processus gaussien, Test d'hypothèses, Génotypage sélectif, Extrêmes

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

Keywords. Gene detection, Gaussian process, Hypothesis testing, Selective genotyping, Extremes

1 Contexte

On étudie une population backcross ($A \times B$) où A et B sont deux lignées homozygotes pures. On considère le problème de la détection de loci codant pour un caractère quan-

titatif, aussi appelés QTL (Quantitative Trait Loci), sur un chromosome donné. Le caractère est observé sur n individus et on note Y_j, $j = 1, \cdots, n$, les observations que l'on suppose iid. Le mécanisme de la méiose fait que parmi les deux chromosomes d'un individu, un est purement hérité de A alors que l'autre est formé de morceaux de A et de morceaux de B du fait des crossing-overs. Le chromosome est représenté par le segment $[0, T]$. La distance sur $[0, T]$ est appelée distance génétique et est mesurée en Morgans. Le génome $X(t)$ d'un individu prend la valeur +1 si le chromosome recombiné est originaire de A à la position t et prend la valeur -1 s'il est originaire de B. Le modèle admis pour la structure stochastique de $X(\cdot)$ est dû à Haldane (1919):

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

où $N(\cdot)$ est le processus de Poisson standard sur $[0, T]$ représentant le nombre de crossing-overs. De plus on suppose que m QTL additifs influent sur le caractère quantitatif Y. On note q_s et $t_s^\star$ l'effet et la position du $s$ième QTL, $s = 1 \cdots m$. On suppose un modèle d'analyse de variance pour Y:

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

où ϵ est un bruit blanc gaussien.
En fait "l'information génétique" n'est disponible que sur les marqueurs c'est-à-dire uniquement à certaines positions $t_1, \cdots, t_K$. K est le nombre de marqueurs. Une observation sera donc $(Y, X(t_1), \cdots, X(t_K))$. Nous observons $(Y_j, X_j(t_1), \cdots, X_j(t_K))$ pour $j = 1, \cdots, n$; supposées iid. L'objectif de cette étude est d'estimer le nombre m de QTL, leurs positions $t_1^\star, \cdots, t_m^\star$ et leurs effets $q_1, \cdots, q_m$. Si les QTL étaient positionnés exactement sur les marqueurs on pourrait utiliser l'information génétique sur les marqueurs comme régresseurs dans une méthode de type LASSO (Tibshirani, 1996) pour détecter les QTL. Les QTL pouvant être positionnés n'importe où sur le chromosome nous ne pouvons pas utiliser les méthodes classiques de sélection de variables. La détection de QTL nécessite de travailler sur les modèles de mélange.
Quand il n'y a qu'un seul QTL sur le chromosome positionné en $t_1^\star$ (i.e. $m = 1$); conditionnellement aux informations aux marqueurs, Y suit une loi de mélange:

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

où $p(t_1^\star)$ est un poids connu que nous savons calculer $p(t_1^\star) = P[X(t_1^\star) = 1]$ conditionnellement aux marqueurs et $f_{(m,\sigma)}(\cdot)$ est la densité gaussienne de moyenne m et de variance σ^2. La méthode de l'"interval mapping" proposée par Lander et Botstein (1989) consiste à faire un test de rapport de vraisemblance $q_1 = 0$ vs $q_1 \neq 0$ dans l'équation (1) en tout point t de $[0, T]$. Nous obtenons un processus de test de rapport de vraisemblance $\Lambda_n(t)$, $t \in [0, T]$. Le supremum de $\Lambda_n(t)$, $t \in [0, T]$ est alors la statistique du test de rapport

de vraisemblance de l'hypothèse nulle "il n'y a pas de QTL sur le chromosome" contre l'alternative "il existe un QTL en $t_1^\star$ sur le chromosome". $\arg\sup \Lambda_n(t)$ est un estimateur naturel de la position du QTL. Dans Azaïs et al. (2012) nous avons obtenu la distribution asymptotique exacte de $\Lambda_n(\cdot)$ sous l'hypothèse nulle et sous des hypothèses alternatives contigües. Nous avons montré que le processus de test de rapport de vraisemblance est asymptotiquement le carré d'un "processus non linéaire interpolé" centré sous l'hypothèse nulle et décentré sous l'alternative d'une fonction moyenne qui dépend de l'effet QTL q_1 et de sa position $t_1^\star$. Nous avons également obtenu une formule analytique permettant de calculer le supremum de $\Lambda_n(\cdot)$.
L'utilisation comme statistique de test de $\sup\Lambda_n(\cdot)$ est appropriée pour tester et localiser un QTL sur $[0, T]$ mais ce n'est plus adéquat lorsque plusieurs QTL sont présents sur le chromosome. Lorsque m QTL sont présents sur le chromosome; conditionnellement aux marqueurs, Y suit une loi de mélange:

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

où $w_{\vec{t}^\star}(u_1,\ ...,\ u_m)$ est la probabilité $P\left\{X(t_1^\star)=u_1,...,X(t_m^\star)=u_m\right\}$ conditionnellement aux marqueurs. Dans ce papier nous généralisons les résultats d'Azaïs et al. (2012) pour obtenir la loi asymptotique de $\Lambda_n(\cdot)$ sous l'alternative qu'il existe m QTL sur $[0, T]$ positionnés en $t_1^\star, \cdots, t_m^\star$ d'effets $q_1, \cdots, q_m$. Ce processus est asymptotiquement le carré d'un processus non linéaire interpolé dont la fonction moyenne dépend du nombre de QTL, leurs positions, leurs effets. Ce résultat théorique nous permet de proposer une nouvelle méthode pour estimer le nombre de QTL, leurs positions, leurs effets.
Dans une seconde partie nous généralisons les résultats et la méthode obtenus précédemment au génotypage sélectif. Le génotypage sélectif consiste à génotyper (i.e. obtenir l'information génétique aux marqueurs $X(t_1), \cdots, X(t_K)$), uniquement les individus extrêmes (i.e. les individus dont le phénotype Y est au delà d'un certain seuil: $Y \notin [S_-, S_+]$). Ce dispositif proposé par Lebowitz et al. (1987) s'avère très employé en agronomie, car il permet d'optimiser le génotypage et d'améliorer la puissance de détection.

Si nous notons $\overline{X}(t)$ la variable aléatoire telle que:

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

alors, dans notre problème, une observation sera:

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

Avec nos notations:

- quand $Y \notin [S_-\ ,\ S_+]$, nous avons $\overline{X}(t_1) = X(t_1),\ ..., \overline{X}(t_K) = X(t_K)$.

- quand $Y \in [S_-\ ,\ S_+]$, nous avons $\overline{X}(t_1) = 0,\ ..., \overline{X}(t_K) = 0$, ce qui signifie que l'information génétique est manquante aux marqueurs.

Nous observons dès lors $(Y_j, \overline{X}_j(t_1), \cdots, \overline{X}_j(t_K))$ pour $j = 1, \cdots, n$; supposées iid.

Lorsqu'il n'y a qu'un seul QTL (i.e. $m = 1$) positionné en $t_1^\star$, la loi de $\big(Y,\ \overline{X}(t_1),\ ...,\ \overline{X}(t_K)\big)$ est proportionnelle au mélange

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

et nous avons prouvé (Rabier, 2015), que le processus de rapport de vraisemblance, $\Lambda_n(.)$, est asymptotiquement le carré d'un processus d'interpolation non linéaire, dont la fonction moyenne dépend du génotypage sélectif sous l'alternative contigüe.

Lorsqu'il existe m QTLs, la loi de $\big(Y,\ \overline{X}(t_1),\ ...,\ \overline{X}(t_K)\big)$ est désormais proportionnelle au mélange à 2^m composantes

$$\begin{aligned} & \sum_{(u_1,...,u_m)\in\{-1,1\}^m} w_{\vec{t^\star}}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+...+u_mq_m,\sigma)}(Y)\ 1_{Y\notin[S_-,S_+]} \\ & + v_{\vec{t^\star}}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+...+u_mq_m,\sigma)}(Y)\ 1_{Y\in[S_-,S_+]} \end{aligned}$$

où $v_{\vec{t^\star}}(u_1,\ ...,\ u_m)$ est la probabilité $P(X(t_1^\star) = u_1, X(t_2^\star) = u_2, \ldots, X(t_m^\star) = u_m)$ et où $w_{\vec{t^\star}}(u_1,\ ...,\ u_m)$ est la même quantité que précédemment.

Introduisons les notations suivantes: γ, γ_+, γ_- et $\mathcal{A}$ sont respectivement les quantités $P_{H_0}\left(Y \notin [S_-,\ S_+]\right)$, $P_{H_0}\left(Y > S_+\right)$, $P_{H_0}\left(Y < S_-\right)$ et

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

où $\varphi(x)$ et z_α désignent respectivement la densité d'une loi normale standard prise au point x et le quantile d'ordre $1-\alpha$ d'une loi normale standard.
Nos principaux résultats sont résumés dans la section suivante.

2 Résultats

Théorème 1 *Supposons que les paramètres $(q_1, ..., q_m, \mu, \sigma^2)$ varient dans un compact, que $\exists b > 0$ tel que $\sigma^2 \geq b > 0$, et que m est fini. Soit H_0 l'hypothèse nulle d'absence de QTL sur $[0, T]$, définissons les hypothèses alternatives suivantes:*

$$\begin{aligned} & H_{a\vec{t^\star}} : \text{“il y a } m \text{ QTL localisés respectivement en } t_1^\star,\ ...,\ t_m^\star \\ & \text{d'effets } q_1 = a_1/\sqrt{n}, ..., q_m = a_m/\sqrt{n} \text{ où } a_1 \neq 0, ..., a_m \neq 0\text{”}\ . \end{aligned}$$

Alors le processus de rapport de vraisemblance $\Lambda_n(t)$ et le processus de score $S_n(t)$, $t \in [0, T]$ vérifient:

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

quand n tend vers l'infini, sous H_0 *et* $H_{a\vec{t}^\star}$. $Z(.)$ *est un processus gaussien de variance 1 tel que pour* t *appartenant à l'intervalle de marqueurs* $[t_l, t_r]$*:*

$$Z(t) = \frac{\alpha(t)Z(t_l) + \beta(t)Z(t_r)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(t_l, t_r)}},$$
$$Cov\{Z(t_l), Z(t_r)\} = \rho(t_l, t_r) = e^{-2|t_l - t_r|}$$

où $\alpha(t) = Q_t^{1,1} - Q_t^{-1,1}$, $\beta(t) = Q_t^{1,1} - Q_t^{1,-1}$ *et de fonction moyenne*

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

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

avec

$$m_{\vec{t}^\star}(t_l) = \sum_{s=1}^{m} \sqrt{\mathcal{A}}\, a_s\, \rho(t_l, t_s^\star)/\sigma^2\ , \quad m_{\vec{t}^\star}(t_r) = \sum_{s=1}^{m} \sqrt{\mathcal{A}}\, a_s\, \rho(t_r, t_s^\star)/\sigma^2.$$

A noter que le facteur $\frac{\sqrt{\mathcal{A}}}{\sigma}$, présent dans la fonction moyenne, est dû au génotypage sélectif. Ce facteur devient égal à 1 si l'on se place dans une situation sans génotypage sélectif (i.e. $S_- = S_+$). Cette quantité joue le rôle de coefficient multiplicatif pour les effets QTLs : le signal peut ainsi être augmenté à condition que le nombre d'individus n ait été augmenté lors d'une expérience avec génotypage sélectif.

D'après le théorème précédent, en discrétisant le processus de test de vraisemblance sur la position des marqueurs nous avons quand n est grand:

$$\vec{S}_n = \vec{m}_{\vec{t}^\star} \; + \; \vec{\varepsilon} \; + \; o_P(1)$$

où $\vec{S}_n = (S_n(t_1)\ ,\ S_n(t_2)\ ,\ ...\ ,\ S_n(t_K))'$, $\vec{m}_{\vec{t}^\star} = (m_{\vec{t}^\star}(t_1)\ ,\ m_{\vec{t}^\star}(t_2)\ ,\ ...\ ,\ m_{\vec{t}^\star}(t_K))'$ et $\vec{\varepsilon} \sim N(0, \Sigma)$ où $\Sigma_{kk'} = \rho(t_k, t_{k'})$.
Nous décorrélons les composantes de $\vec{S}_n$ en considérant la décomposition de Cholesky $\Sigma = AA'$:

$$A^{-1}\vec{S}_n = A^{-1}B\left(\frac{a_1\sqrt{\mathcal{A}}}{\sigma^2}\ ,\ ...\ ,\ \frac{a_m\sqrt{\mathcal{A}}}{\sigma^2}\right)' \; + \; A^{-1}\vec{\varepsilon} \; + \; o_P(1)$$

où B est une matrice de taille $K \times m$ telle que $B_{ks} = e^{-2|t_k - t_s^\star|}$.
Puisque le nombre m de QTL et leurs positions $t_1^\star$,...,$t_m^\star$ sont inconnus, considérons une nouvelle discrétisation de $[0, T]$ correspondant aux positions possibles des QTL: $0 \leq t_1' <$

$t'_2 < ... < t'_L \leq T$. $\Delta_1, ..., \Delta_L$ sont les effets correspondants. Ainsi nous pouvons chercher des QTL pas seulement sur les marqueurs. Le modèle se réecrit:

$$A^{-1}\vec{S}_n = A^{-1}C\,(\Delta_1\ ,\ ...\ ,\ \Delta_L)' \ +\ A^{-1}\vec{\varepsilon}\ +\ o_P(1) \tag{3}$$

où C est une matrice de taille $K \times L$ telle que $C_{kl} = e^{-2\left|t_k - t'_l\right|}$.
Enfin pour trouver les Δ_l non nuls, une méthode naturelle est d'utiliser la régression pénalisée L1, appelée LASSO:

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

où $\| \ \|_2$ est la norme L2, $\| \ \|_1$ est la norme L1, $\Delta = (\Delta_1, ..., \Delta_L)'$ et ζ sont des paramètres à calibrer. ζ sera estimé par validation croisée. On pourra montrer que la méthode est sensible à l'intensité du processus de Poisson modélisant le nombre de recombinaisons, à l'ampleur du génotypage sélectif, ainsi qu'à la carte génétique.

Bibliographie

[1] Arias-Castro E., Candes E.J, Plan Y. (2011), Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism, *Annals of Statistics*, 39(5) 2533-2556.
[2] Azaïs J.M., Delmas C., Rabier C.E. (2012), Likelihood ratio test process for Quantitative Trait Locus detection, *Statistics*, 48(4) 787-801.
[3] Donoho D. (2006), For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution, *Comm. Pure Appl. Math.*, 59(6) 797-829.
[4] 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.
[5] Lander E.S., Botstein D. (1989), Mapping mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics*, 138 235-240.
[6] Lebowitz RJ, Soller M, Beckmann, J.S. (1987), Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines, *Theor. Appl. Genet.*, 73 556-562.
[7] Rabier C-E (2015), On stochastic processes for Quantitative Trait Locus mapping under selective genotyping, *Statistics*, 49(1) 19-34.
[8] Rabier C.E., Delmas C. (2016), On gene mapping with the mixture model and the extremes, hal-01273783.
[9] Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society B*, 58(1) 267-288.