L'AdaptSgenoLasso, une variante du SgenoLasso, pour la localisation de gènes et la prédiction génomique à l'aide des extrêmes

Charles-Elie Rabier [1,2] & Céline Delmas [3]

[1] *ISEM, Université de Montpellier, CNRS, EPHE, IRD, France*
[2] *IMAG, Université de Montpellier, CNRS, France*
charles-elie.rabier@umontpellier.fr
[3] *Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France*
celine.delmas.toulouse@inrae.fr

Résumé. Nous présentons l'AdaptSgenoLasso, une nouvelle méthode de vraisemblance pénalisée pour la localisation de gènes, et qui s'avère être une variante du SgenoLasso. L'AdaptSgenoLasso repose sur le concept d'un génotypage sélectif qui est autorisé à varier le long du génome. La version classique du génotypage sélectif, sur laquelle est basé le SgenoLasso, consiste à génotyper uniquement les individus extrêmes, afin d'augmenter le signal lié aux gènes. Cependant, comme le même pourcentage de sélection est appliqué à chaque position du génome, le signal se voit augmenté d'un même facteur proportionnel sur l'ensemble du génome. En considérant un génotypage sélectif qui varie le long du génome grâce à l'AdaptSgenoLasso, nous permettons aux généticiens d'imposer plus de poids à certains loci d'intérêt, connus pour être responsables de la variation de caractères quantitatifs. Le signal est désormais propre à chaque locus.

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

Abstract. We introduce here the AdaptSgenoLasso, a new penalized likelihood method for gene mapping, which is a modified version of the SgenoLasso. AdaptSgenoLasso relies on the concept of a selective genotyping that varies along the genome. The "original version" of the selective genotyping on which the SgenoLasso is built on, consists in genotyping only extreme individuals, in order to increase the signal from genes. However, since the same amount of selection is applied at all genome locations, the signal is increased of the same proportional factor everywhere. By considering a selective genotyping that varies along the genome thanks to the AdaptSgenoLasso, we allow geneticists to impose more weights on some loci of interest, known to be responsible for variation of the quantitative trait. The resulting signal is now dedicated to each locus.

Keywords. Statistical learning, 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 quantitatif, 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 i.i.d. Le mécanisme de la méïose 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 QTLs 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:

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

où ε est un bruit blanc gaussien. Dans le problème classique de détection de QTLs, l'"information génome" est disponible uniquement à des positions fixes $t_1 = 0 < t_2 < ... < t_K = T$, appelées marqueurs génétiques. Ainsi, d'ordinaire, une observation est (Y, $X(t_1)$, ..., $X(t_K)$) et le challenge réside dans le fait que le nombre m de QTLs m et leurs positions $t_1^\star$, ..., $t_m^\star$ sont inconnues. On notera $t^\star = (t_1^\star, ..., t_m^\star)$. Dans cette étude, nous considérons le problème classique, mais afin de réduire les coûts dus au génotypage, un génotypage sélectif qui varie le long du génome est considéré. Décrivons tout d'abord le concept du génotypage sélectif dans sa version originale, celle qui ne varie pas le long du génome. 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: $Y \notin [S_-^1, S_+^1]$). 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. Afin d'introduire le génotypage sélectif qui varie le long du génome, considérons désormais quatre seuils S_-^1, S_-^2, S_+^2, S_+^1 appartenant à $\mathbb{R}$ tels que $S_-^1 \leq S_-^2 \leq S_+^2 \leq S_+^1$. Comme dans le cadre du génotypage sélectif classique, nous observons l'information génome à tous les marqueurs si et seulement si Y est extrême, à savoir si $Y \leq S_-^1$ ou $Y \geq S_+^1$. Cependant, nous considérons

également une carte génétique discrète contenant seulement quelques marqueurs appartenant à la carte dense originale (i.e. la carte comprenant tous les marqueurs), et nous observons à ces quelques positions, l'information génome des individus pour lesquels $Y \leq S_-^2$ ou $Y \geq S_+^2$. En d'autres termes, à ce nombre restreint de marqueurs, nous recueillons l'information génome d'un plus grand nombre d'individus extrêmes. Intuitivement, cela permet d'imposer des poids plus importants à certains loci (cf. Section 2) correspondant à des gènes majeurs bien connus par les généticiens.

Afin de décrire les deux cartes génétiques plus précisement, notons $\mathbb{T}_K^1$ l'ensemble $\{t_1, \ldots, t_K\}$ de marqueurs sur la carte dense, et $\mathbb{T}_K^2$ un sous espace de $\mathbb{T}_K^1$ (i.e. $\mathbb{T}_K^2 \subseteq \mathbb{T}_K^1$), représentant les marqueurs sur la carte discrète. On notera $\#\mathbb{T}_K^2$ le cardinal de $\mathbb{T}_K^2$, et $\sigma(.)$ la fonction injective telle que $\sigma : \{1, \ldots, \#\mathbb{T}_K^2\} \to \{1, \ldots, K\}$. De plus, nous imposons $\sigma(k) < \sigma(k')$ pour $k < k'$. Ainsi, $\mathbb{T}_K^2$ désigne l'ensemble $\left\{t_{\sigma(1)}, t_{\sigma(2)}, \ldots, t_{\sigma(\#\mathbb{T}_K^2)}\right\}$. Nous imposerons également $\sigma(1) = 1$ et $\sigma(\#\mathbb{T}_K^2) = K$, de telle sorte que les marqueurs positionnés en 0 et en T (i.e. aux extrémités du chromosome) soient également localisés sur la carte discrète. Si nous notons $\overline{X}(t)$ et $\widetilde{X}(t)$ les variables aléatoires telles que $\overline{X}(t) = X(t)1_{Y \notin \left[S_-^1\ ,\ S_+^1\right]}$ et $\widetilde{X}(t) = X(t)1_{Y \in \left[S_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ S_+^1\right]}$, alors dans notre problème, une observation correspond à la donnée de

$$\left(Y,\ \overline{X}(t_1),\ \overline{X}(t_2),\ ...,\ \overline{X}(t_K),\ \widetilde{X}(t_{\sigma(1)}), \widetilde{X}(t_{\sigma(2)}),\ ..., \widetilde{X}(t_{\sigma(\#\mathbb{T}_K^2)})\right).$$

Avec nos notations,

- quand $Y \notin \left[S_-^1\ ,\ S_+^1\right]$, nous avons $\overline{X}(t_1) = X(t_1),\ ..., \overline{X}(t_K) = X(t_K)$, ce qui signifie que l'information génome est connue sur la carte dense $\mathbb{T}_K^1$ (et a fortiori sur la carte sparse $\mathbb{T}_K^2$).

- quand $Y \in \left[S_-^1\ ,\ S_-^2\right] \cup \left[S_+^2\ ,\ S_+^1\right]$, nous avons $\widetilde{X}(t_{\sigma(1)}) = X(t_{\sigma(1)}), \widetilde{X}(t_{\sigma(2)}) = X(t_{\sigma(2)}),\ ..., \widetilde{X}(t_{\sigma(\#\mathbb{T}_K^2)}) = X(t_{\sigma(\#\mathbb{T}_K^2)})$, à savoir l'information génome est connue uniquement sur la carte sparse $\mathbb{T}_K^2$.

- quand $Y \in \left[S_-^2\ ,\ S_+^2\right]$, nous avons $\overline{X}(t_1) = 0,\ ..., \overline{X}(t_K) = 0$, et $\widetilde{X}(t_{\sigma(1)}) = 0, \widetilde{X}(t_{\sigma(2)}) = 0,\ ..., \widetilde{X}(t_{\sigma(\#\mathbb{T}_K^2)}) = 0$, ce qui signifie que l'information génome est manquante à tous les marqueurs (i.e. $\mathbb{T}_K^1$).

Nous observons dès lors n observations
$\left(Y_j,\ \overline{X}_j(t_1),\ \overline{X}_j(t_2),\ ...,\ \overline{X}_j(t_K),\ \widetilde{X}_j(t_{\sigma(1)}), \widetilde{X}_j(t_{\sigma(2)}),\ ..., \widetilde{X}_j(t_{\sigma(\#\mathbb{T}_K^2)})\right)$ pour $j = 1, \cdots, n$; supposées iid.

Avant de décrire nos résultats et notre nouvelle méthode de sélection de variables, rappelons le concept de l'Interval Mapping (Lander et Botstein, 1989). Lorsqu'un seul QTL est positionné sur le chromosome (i.e. $m = 1$ dans la formule (1)), l'Interval

Mapping consiste à calculer le Test du Rapport de Vraisemblance (LRT) à chaque position $t \in [0, T]$, confrontant l'hypothèse nulle d'absence de QTL H_0:"$q_1 = 0$," contre l'alternative "$q_1 \neq 0$,". Cela conduit à un processus de LRT $\Lambda_n(.)$ et à un processus de score $S_n(.)$. Ces processus ont été étudiés en détail dans le passé dans la situation de données complètes où tous les individus sont génotypés (e.g. Cierco C., 1998, Azaïs et Wschebor, 2009, Chang et al., 2009, Azaïs et al, 2012), et plus tard dans le cadre du génotypage sélectif classique (e.g Rabier, 2014, 2015). Le maximum de ces processus correspond au LRT sur l'ensemble du chromosome, et la distribution asymptotique de ces processus est désormais bien connue. Cependant la statistique du maximum s'avère inappropriée lorsque $m > 1$. Ainsi, dans cette étude, nous proposons d'étudier, dans le cadre du génotypage sélectif qui varie, la distribution asymptotique des processus de LRT et de score sous l'alternative générale de m QTLs sur le génome. Cela nous permettra d'introduire une nouvelle méthode de sélection de variables, l'AdaptSgenoLASSO, afin d'identifier les nombreux QTLs le long du génome, à l'instar de la méthode du SgenoLasso (Rabier et Delmas, 2021), proposée récemment dans le cadre du génotypage sélectif classique. Contrairement au Lasso (Tibshirani, 1996), le SgenoLasso présente l'avantage de gérer efficacement les données extrêmes. De plus, nous avons montré qu'il présentait de meilleurs performances que le récent RaLasso (Fan et al., 2017) qui modélise pourtant les dépendances entre les erreurs et les regresseurs. Ainsi, l'AdaptSgenoLasso, la nouvelle variante du SgenoLasso permettant d'accorder plus d'importance à certains loci bien connus, s'avère prometteuse.

Introduisons les notations suivantes:

$$\gamma_1 := \mathbb{P}_{\mathcal{H}_0}\left(Y \notin \left[S_-^1,\ S_+^1\right]\right)\ ,\ \gamma_1^+ := \mathbb{P}_{\mathcal{H}_0}\left(Y > S_+^1\right)\ ,\ \gamma_1^- := \mathbb{P}_{\mathcal{H}_0}\left(Y < S_-^1\right)\ , \tag{2}$$

$$\gamma := \mathbb{P}_{\mathcal{H}_0}\left(Y \notin \left[S_-^2,\ S_+^2\right]\right)\ ,\ \gamma^+ := \mathbb{P}_{\mathcal{H}_0}\left(Y > S_+^2\right)\ ,\ \gamma^- := \mathbb{P}_{\mathcal{H}_0}\left(Y < S_-^2\right)\ , \tag{3}$$

$$\mathcal{A}_1 := \sigma^2\left\{\gamma_1 + z_{\gamma_1^+}\varphi(z_{\gamma_1^+}) - z_{1-\gamma_1^-}\varphi(z_{1-\gamma_1^-})\right\}\ , \tag{4}$$

$$\mathcal{B} := \sigma^2\left\{\gamma + z_{\gamma^+}\varphi(z_{\gamma^+}) - z_{1-\gamma^-}\varphi(z_{1-\gamma^-})\right\}\ ,\ \mathcal{A}_2 := \mathcal{B} - \mathcal{A}_1 \tag{5}$$

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

Dans ce qui suit, on considère des valeurs de t distinctes des positions de marqueurs, i.e. $t \in [t_1, t_K]\backslash\mathbb{T}_K^1$. Pour $i = 1, 2$, on définit $t^{\ell,i}$ and $t^{r,i}$ de la manière suivante:

$$t^{\ell,i} = \sup\left\{t_k \in \mathbb{T}_K^i : t_k < t\right\}\ \ ,\ \ t^{r,i} = \inf\left\{t_k \in \mathbb{T}_K^i : t < t_k\right\}. \tag{6}$$

En d'autres termes, en fonction de la carte, t appartient à l'intervalle de marqueurs $(t^{\ell,1}, t^{r,1})$ ou $(t^{\ell,2}, t^{r,2})$.

Theorem 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* $[0, T]$ *et définissons les hypothèses alternatives suivantes:*

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

Alors lorsque n *tend vers l'infini, les processus* $S_n(.)$ *et* $\Lambda_n(.)$ *vérifient :*

$$S_n(.) \Rightarrow Z(.) \ , \ \Lambda_n(.) \overset{F.d.}{\to} Z^2(.) \ , \ \sup \Lambda_n(.) \xrightarrow{\mathcal{L}} \sup Z^2(.) \tag{7}$$

sous $\mathcal{H}_0$ *et* $\mathcal{H}_{a,t^\star}$*, où* $\Rightarrow$ *et F.d. désignent respectivement la convergence faible et la convergence des lois finies-dimensionnelles, et où* $Z(.)$ *est le processus Gaussien de variance* 1 *tel que* $\forall t \in [t_1, t_K] \backslash \mathbb{T}_K^1$ *:*

$$Z(t) = \frac{\sqrt{\mathcal{A}_1} \ \xi_1(t) V_1(t) \ + \ \sqrt{\mathcal{A}_2} \ \xi_2(t) V_2(t)}{\sqrt{\mathcal{A}_1 \ \xi_1^2(t) + \mathcal{A}_2 \ \xi_2^2(t)}} \ .$$

$V_1(.)$ *et* $V_2(.)$ *sont des processus Gaussiens indépendants de variance* 1 *tels que*

$$V_i(t) = \left\{ \alpha_i(t) \ V_i(t^{\ell,i}) \ + \ \beta_i(t) \ V_i(t^{r,i}) \right\} / \xi_i(t)$$
$$\forall (t_k, t_{k'}) \in \mathbb{T}_K^i \times \mathbb{T}_K^i \ \ Cov(V_i(t_k), V_i(t_{k'})) = e^{-2|t_k - t_{k'}|} \ .$$

La fonction moyenne de $Z(.)$ *est nulle* H_0 *et vérifie sous* $H_{a,t^\star}$*:*

$$m_{Z,t^\star}(t) = \frac{\sqrt{\mathcal{A}_1} \ \xi_1(t) m_{V_1,t^\star}(t) \ + \ \sqrt{\mathcal{A}_2} \ \xi_2(t) m_{V_2,t^\star}(t)}{\sqrt{\mathcal{A}_1 \ \xi_1^2(t) + \mathcal{A}_2 \ \xi_2^2(t)}}.$$

Les $\alpha_i(t)$*,* $\beta_i(t)$ *et* $\xi_i(t)$ *des fonctions connues et*

$$m_{V_i,t^\star}(t) = \left\{ \alpha_i(t) \ m_{V_i,t^\star}(t^{\ell,i}) + \beta_i(t) \ m_{V_i,t^\star}(t^{r,i}) \right\} / \xi_i(t)$$
$$\forall t_k \in \mathbb{T}_K^i \ \ m_{V_i,t^\star}(t_k) = \frac{\sqrt{\mathcal{A}_i}}{\sigma^2} \sum_{s=1}^{m} a_s \ e^{-2|t_s^\star - t_k|} \ .$$

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

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

où $\vec{S}_n = (S_n(t_1) \ , \ S_n(t_2) \ , \ ... \ , \ S_n(t_K))' \ , \ \vec{m}_{Z,t^\star} = (m_{Z,t^\star}(t_1) \ , \ m_{Z,t^\star}(t_2) \ , \ ..., m_{Z,t^\star}(t_K))'$ et $\vec{\varepsilon} \sim N(0, \Sigma)$ avec $\Sigma_{kk'} = \text{Cov}\left(Z(t_k), Z(t_{k'})\right)$.

Dans ce qui suit, on se place en déséquilibre de liaison complet, i.e. les m QTLs sont localisés sur certains marqueurs. Ainsi, on recherchera des QTLs seulement aux

emplacements des marqueurs. Grace aux spécificités de la fonction moyenne du processus et après avoir décorrélé les composantes de $\vec{S}_n$ en considérant la décomposition de Cholesky $\Sigma = AA'$, on obtient la relation suivante :

$$A^{-1}\vec{S}_n = A'\left(\Delta_1\ ,\ \ldots\ ,\ \Delta_K\right)' \ +\ A^{-1}\vec{\varepsilon} \ +\ o_P(1) \tag{8}$$

$$\text{où}\ \ \Delta_k = \begin{cases} 0 & \text{si} \quad t_k \notin \{t_1^\star, \ldots, t_m^\star\} \\ \frac{a_s\sqrt{\mathcal{B}}}{\sigma^2} & \text{si } t_k \in \{t_1^\star, \ldots, t_m^\star\} \cap \mathbb{T}_K^2 \text{ avec } s \text{ l'indice tel que } t_s^\star = t_k. \\ \frac{a_s\sqrt{\mathcal{A}_1+\mathcal{A}_2\xi_2^2(t_k)}}{\sigma^2} & \text{si } t_k \in \{t_1^\star, \ldots, t_m^\star\} \cap \mathbb{T}_K^1\backslash\mathbb{T}_K^2 \text{ avec } s \text{ l'indice tel que } t_s^\star = t_k. \end{cases}$$

Les QTLs placés sur les marqueurs de la carte discrète se voient amplifiés d'un facteur $\sqrt{\mathcal{B}}/\sigma$ et d'un facteur $\sqrt{\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_k)}/\sigma$ sur la carte dense. Ainsi, contrairement au génotypage sélectif classique où tous les loci disposent du même coefficient multiplicatif (Rabier et Delmas, 2021), on voit que désormais les facteurs multiplicatifs diffèrent en fonction de l'appartenance des loci aux 2 cartes.

Enfin, afin de trouver les Δ_k non nuls, une méthode naturelle est d'utiliser la régression pénalisée L1, appelée Lasso. Ainsi, en notant $\Delta := (\Delta_1, ..., \Delta_K)'$, l'estimateur AdaptSgeno-Lasso s'écrit

$$\hat{\Delta}_{\text{AdaptSgenoLasso}}(\lambda, \alpha) = \arg\min_{\Delta}\left(\left\|A^{-1}\vec{S}_n - A'\Delta\right\|_2^2 + \lambda\left\|\Delta\right\|_1\right)\ . \tag{9}$$

Notons que l'on pourrait également accorder plus d'importance aux loci gènes majeurs en couplant notre proche avec l'Adaptative Lasso. Cela correspondrait à imposer une pénalité dans la formule (9) de la forme $\|W'\Delta\|_1$ avec des poids W_k égaux à $1/\sqrt{\mathcal{B}}$ sur la carte $\mathbb{T}_K^2$ (i.e. gènes majeurs) et $1/\sqrt{\mathcal{A}_1 + \mathcal{A}_2\xi_2^2(t_k)}$ sur la carte $\mathbb{T}_K^1\backslash\mathbb{T}_K^2$.

Bibliographie

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

[2] Chang, M.N., Wu, R., Wu, S.S., and Casella, G. (2009), Score statistics for mapping quantitative trait loci, *Stat. Appl. Genet. Mol. Biol.*, 8(1) 16.

[3] Fan, J., Li, Q., Wang, Y. (2017), Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1) (2017), pp. 247-265.

[4] 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.

[5] Rabier C-E, Delmas C (2021): The SgenoLasso and its cousins for selective genotyping and extreme sampling: application to association studies and genomic selection, *Statistics*, DOI: 10.1080/02331888.2021.1881785.