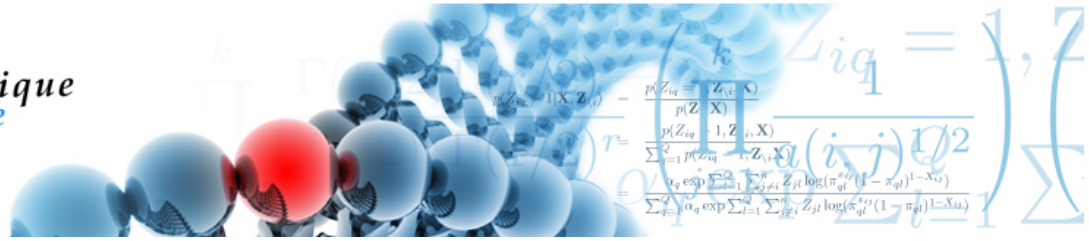


Segmentation conjointe de données génomiques en cancérologie

Morgane Pierre-Jean
Directeur de stage : Pierre Neuvial

Laboratoire Statistique et Génome

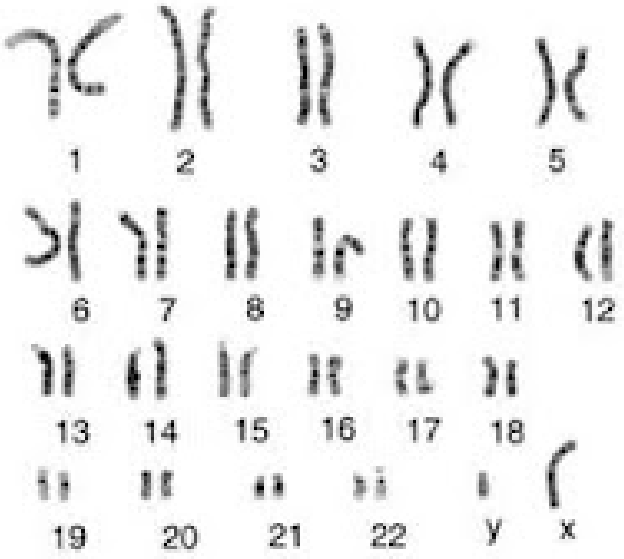
15 Novembre 2012

Sommaire

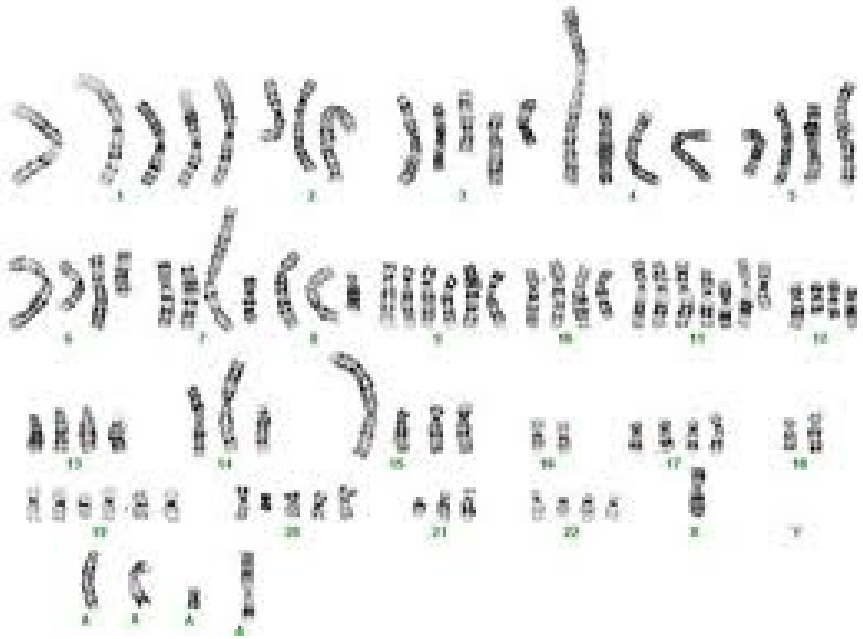
Sommaire

- Les cellules ont normalement deux copies d'ADN
- Cellules cancéreuses : gains et pertes de fragments d'ADN

- Les cellules ont normalement deux copies d'ADN
- Cellules cancéreuses : gains et pertes de fragments d'ADN



Normal



Tumeur

Caractéristiques de données issues des puces de génotypage ou puces SNP :

- Leur grande dimension
- Leur forte structuration

On peut décrire les signaux obtenus par un couple de vecteurs:

$$(\mathbf{c}, \mathbf{b}) = (c_j, b_j)_{j \in \mathcal{J}}$$

où

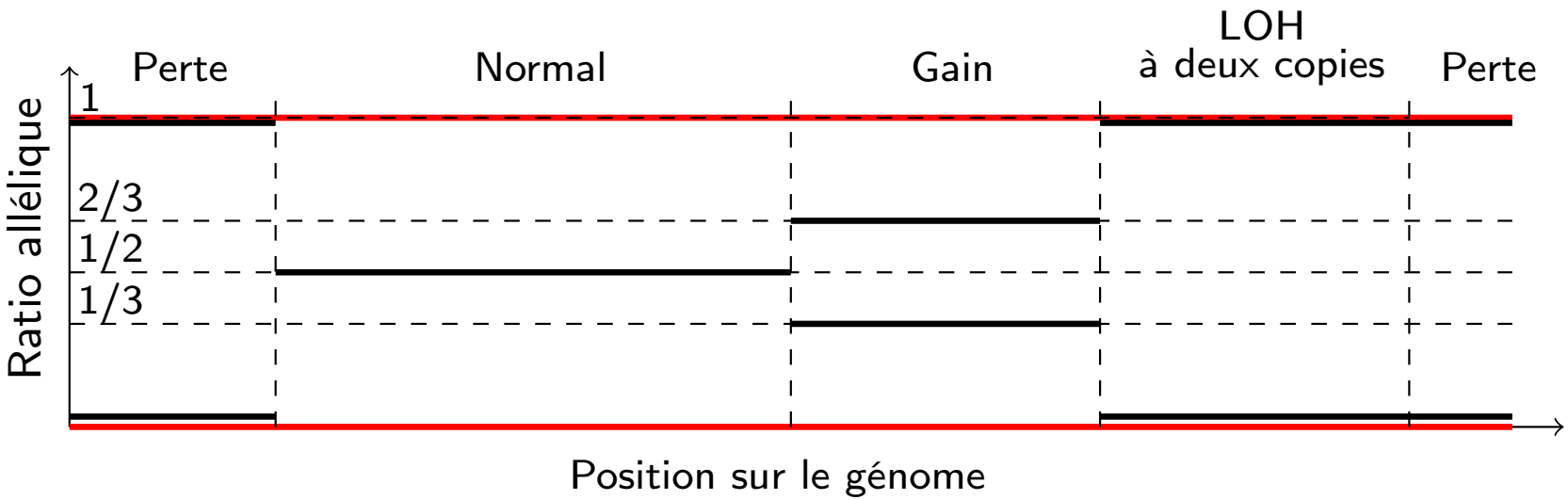
- c_j est le nombre de copies total en j
- b_j la proportion d'allèle B en j

Proportion d'allèle B

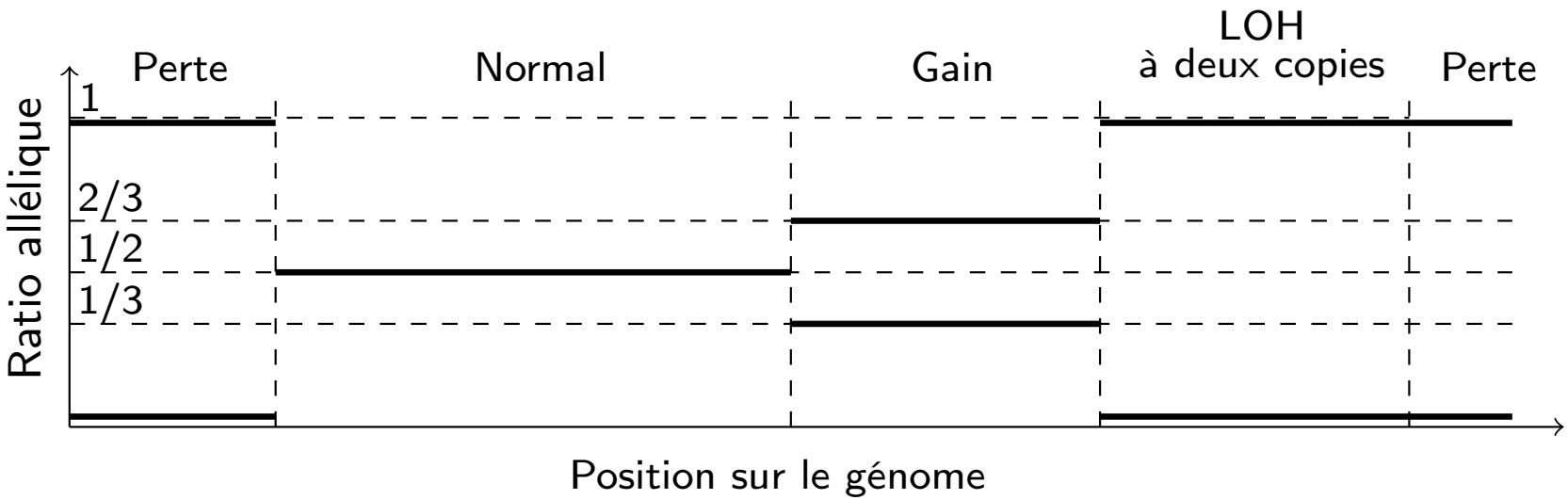
Pour SNP j de la puce à ADN :
Allèle A : AACGTATCG
Allèle B : AACGAATCG
Dans une cellule normale, on a en général deux copies d'ADN si on observe :

- à la fois A et B $\longrightarrow$ la proportion d'allèle B est égale à 1/2
- uniquement A $\longrightarrow$ la proportion d'allèle B est égale à 0/2
- uniquement B $\longrightarrow$ la proportion d'allèle B est égale à 2/2

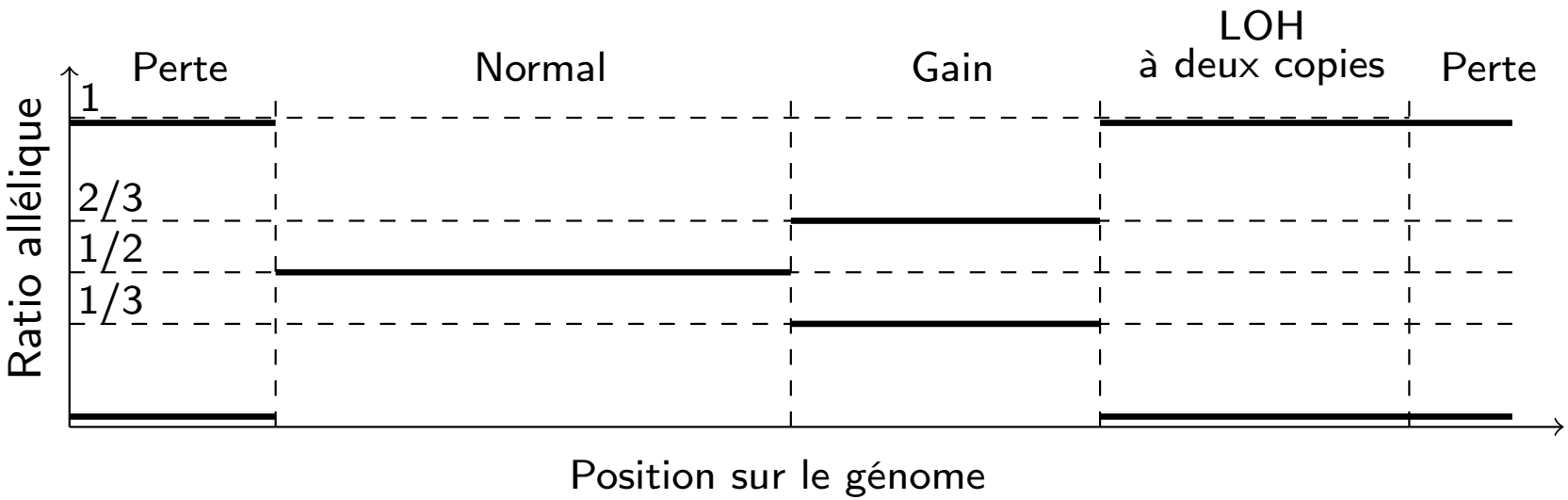
proportion d'allèle B



proportion d'allèle B

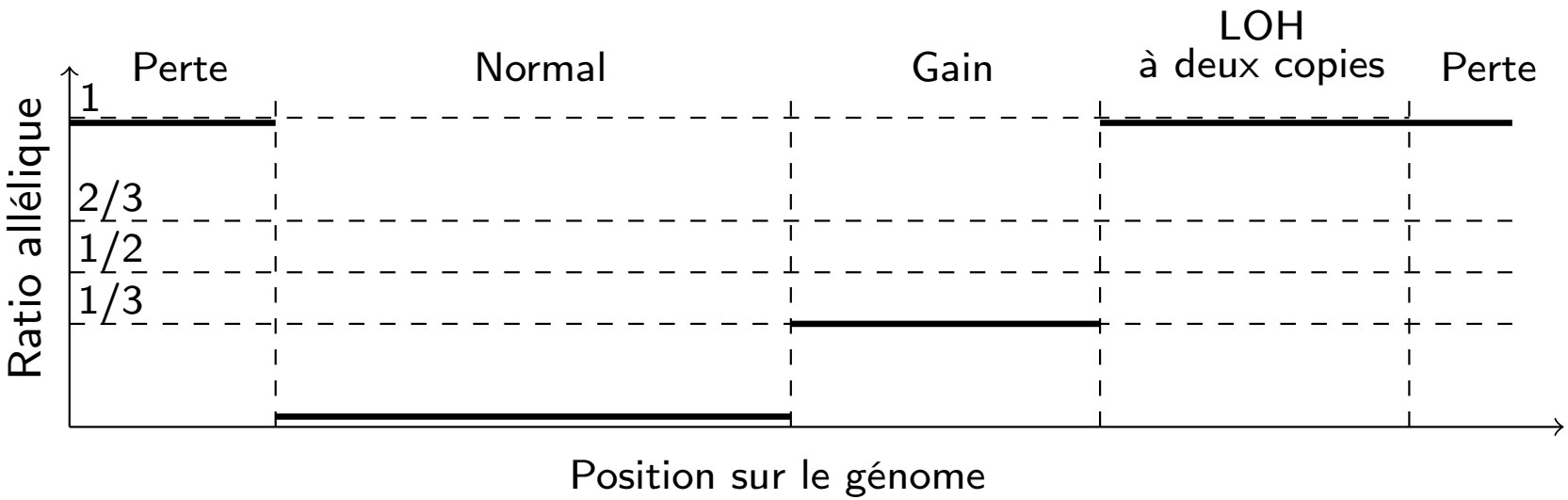


proportion d'allèle B

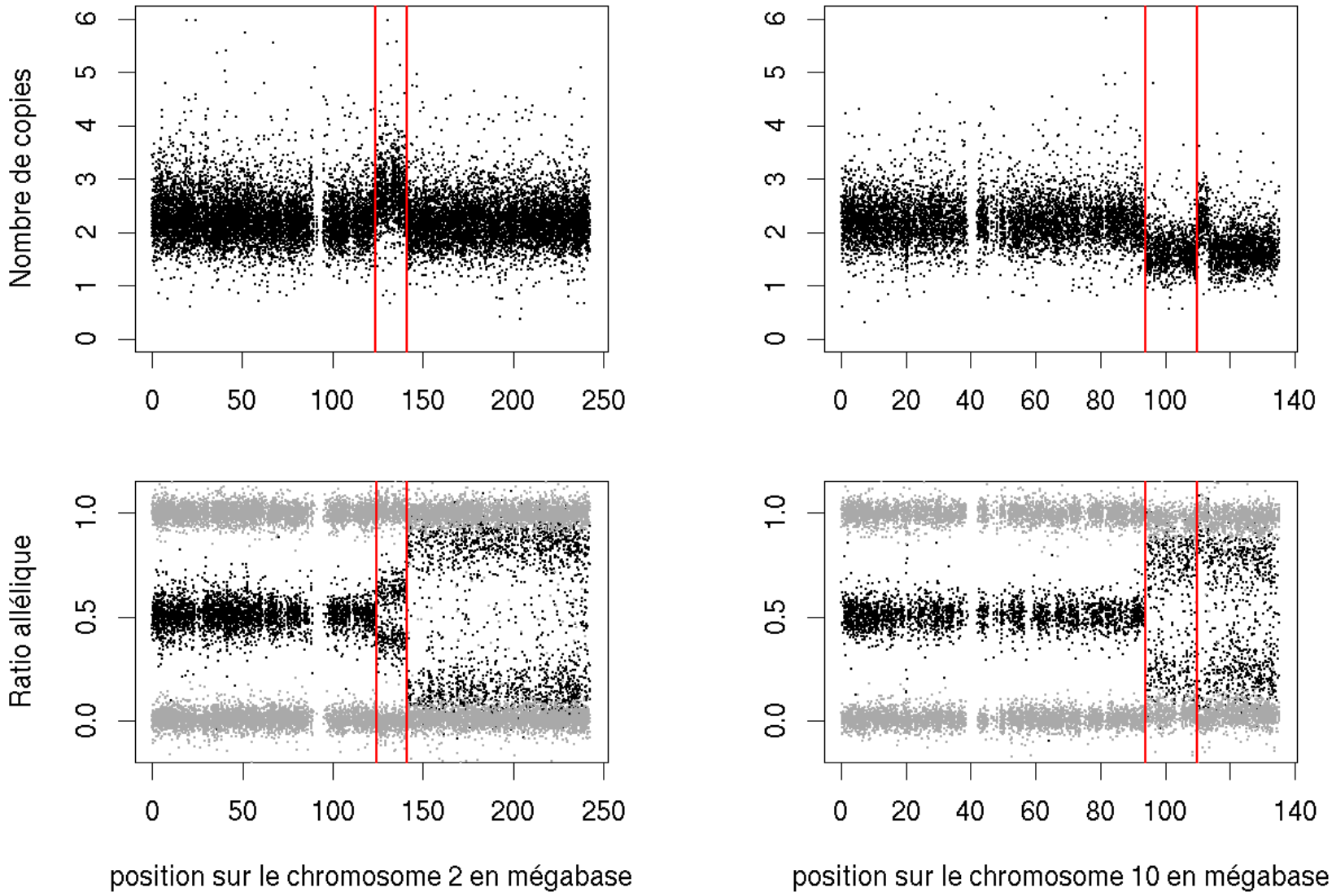


$$2 \times |b_j - 1/2|$$

proportion d'allèle B



$$2 \times |b_j - 1/2|$$


Nombre de copies
0 1 2 3 4 5 6
0 50 100 150 200 250
0 20 40 60 80 100 140
Ratio allélique
0.0 0.5 1.0
position sur le chromosome 2 en mégabase
position sur le chromosome 10 en mégabase

Objectif : Caractériser les régions génomiques altérées

Objectif : Caractériser les régions génomiques altérées
Problèmes sur le plan statistique:

- l'identification des points de rupture
- l'étiquetage des régions en termes de gains, de pertes ou d'amplification

Objectif : Caractériser les régions génomiques altérées
Problèmes sur le plan statistique:

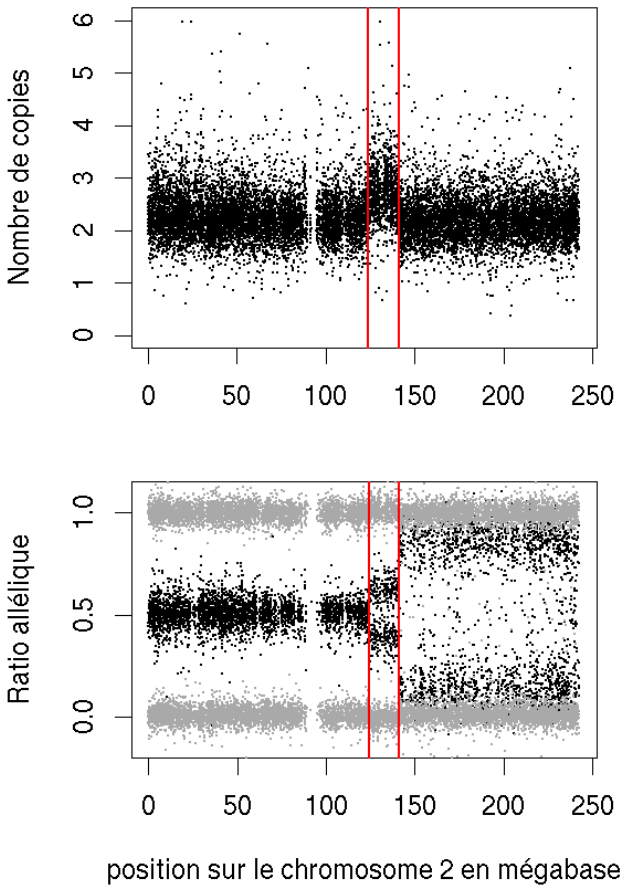
- l'identification des points de rupture
- l'étiquetage des régions en termes de gains, de pertes ou d'amplification

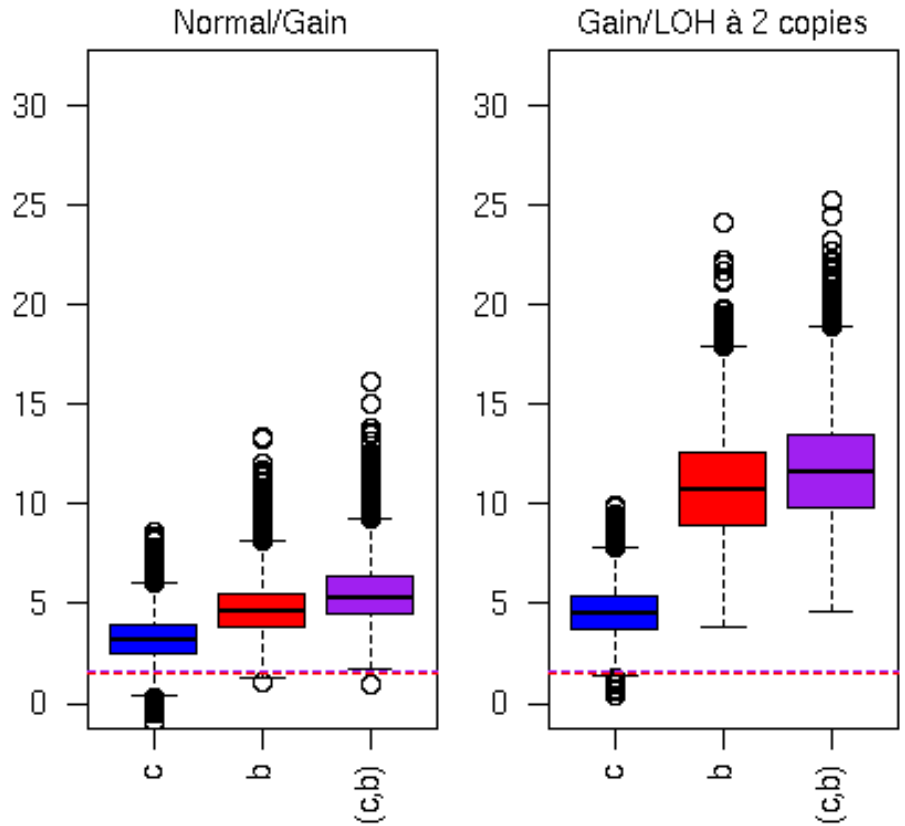
But du stage : mettre en place une méthode de segmentation conjointe des deux dimensions du signal et l'évaluer

Sommaire

Distance entre les états du chromosome 2

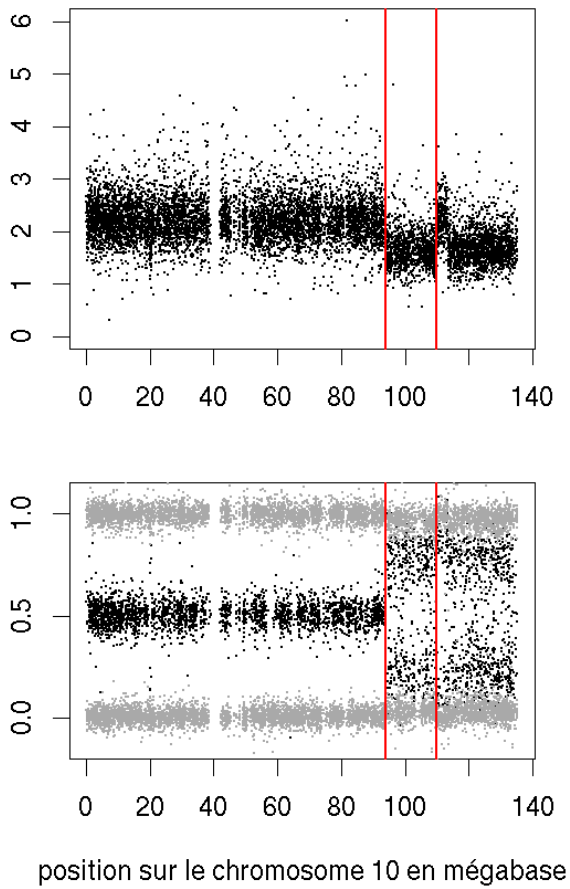
Normal/Gain et Gain/LOH à deux copies

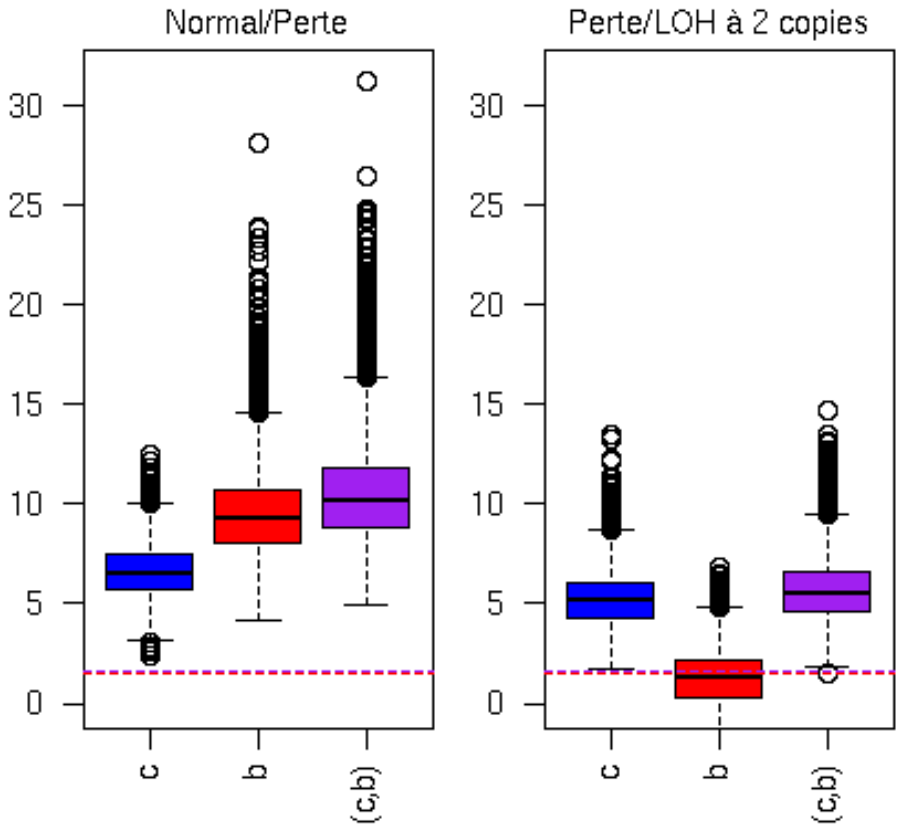




Distance entre les états du chromosome 10

Normal/Perte et Perte/LOH à deux copies





Bilan

- Selon les ruptures ce n'est pas toujours la même statistique qui est la meilleure
- La statistique conjointe permet de faire au moins aussi bien que la meilleure univariée.
- Il devient intéressant d'utiliser les deux dimensions

Sommaire

Modélisation du problème

Point de vue statistique : problème de détection de ruptures dans un signal bivarié

- Modélisation naturelle:

$$c_j = \gamma_j + \epsilon_j$$

où $\forall k \in \{1, K\}, \forall j \in [t_{k-1}, t_k[\qquad \gamma_j = \Gamma_k$

Modélisation du problème

Point de vue statistique : problème de détection de ruptures dans un signal bivarié

- Modélisation naturelle:

$$c_j = \gamma_j + \epsilon_j$$

où $\forall k \in \{1, K\}, \forall j \in [t_{k-1}, t_k[\qquad \gamma_j = \Gamma_k$

- Complexité:

À K fixé, le nombre de positionnements possibles est $C_{n-1}^{K-1} = \mathcal{O}(n^{K-1})$
$\Rightarrow$ **problème combinatoire**

Méthodes existantes

Méthodes exactes par programmation dynamique

[Picard et al. (2005)] : complexité en $\mathcal{O}(Kn^2)$

- Extensible en 2d mais pour des problèmes moins importants

[Rigaill et al.(2010)] : complexité en $\mathcal{O}(nlog(n))$ en moyenne

- Pire des cas $\mathcal{O}(Kn^2)$
- Rapide mais non extensible en 2d

Méthodes existantes

Méthodes exactes par programmation dynamique

[Picard et al. (2005)] : complexité en $\mathcal{O}(Kn^2)$

- Extensible en 2d mais pour des problèmes moins importants

[Rigaill et al.(2010)] : complexité en $\mathcal{O}(nlog(n))$ en moyenne

- Pire des cas $\mathcal{O}(Kn^2)$
- Rapide mais non extensible en 2d

Méthodes approchées par relaxation convexe

[Harchaoui and Lévy-Leduc(2008)] : Complexité en $\mathcal{O}(Kn)$

- Extensible en 2d

[Bleakley and Vert(2011)] : Complexité en $\mathcal{O}(Kn)$

- Méthodes existantes pour segmenter les signaux bivariés sont sous optimales car elles traitent les deux dimensions:
  - soit indépendamment l'une de l'autre [Rigaill et al.(2010)]
  - soit conditionnellement l'une à l'autre [Olshen et al.(2011)]

- Méthodes existantes pour segmenter les signaux bivariés sont sous optimales car elles traitent les deux dimensions:
  - soit indépendamment l'une de l'autre [Rigaill et al.(2010)]
  - soit conditionnellement l'une à l'autre [Olshen et al.(2011)]
- Difficultés majeures:
  - point de vue algorithmique
  - concilier rigueur statistique, pertinence biologique et temps de calcul

Modélisation

La méthode de segmentation conjointe développée est basée sur:

- une méthode de relaxation convexe
- une méthode développée pour segmenter plusieurs profils de patients [Vert and Bleakley (2010)]
- le fait que la localisation des points de rupture est la même dans les deux dimensions

- $\mathbf{Y} = (\mathbf{c} \quad \mathbf{b})$: le nombre de copies d'ADN et le ratio allélique observés.
- $\Gamma = (\Gamma^1 \quad \Gamma^2)$: le vrai nombre de copies et le vrai ratio allélique
- Objectif :

$$\min_{\Gamma \in \mathbb{R}^{n \times 2}} \|\mathbf{Y} - \Gamma\|_2^2 \text{ sous contrainte } K \text{ ruptures dans le signal}$$

- $\mathbf{Y} = (\mathbf{c} \quad \mathbf{b})$: le nombre de copies d'ADN et le ratio allélique observés.
- $\Gamma = (\Gamma^1 \quad \Gamma^2)$: le vrai nombre de copies et le vrai ratio allélique
- Objectif :

$$\min_{\Gamma \in \mathbb{R}^{n \times 2}} \|\mathbf{Y} - \Gamma\|_2^2 \text{ sous contrainte } K \text{ ruptures dans le signal}$$

- Problème de relaxation convexe :

$$\min_{\Gamma \in \mathbb{R}^{n \times 2}} \|\mathbf{Y} - \Gamma\|_2^2 \text{ sous contrainte } \sum_{j=1}^{n-1} \|\Gamma_{j+1} - \Gamma_j\|_2$$

- $\mathbf{Y} = (\mathbf{c} \quad \mathbf{b})$: le nombre de copies d'ADN et le ratio allélique observés.
- $\Gamma = (\Gamma^1 \quad \Gamma^2)$: le vrai nombre de copies et le vrai ratio allélique
- Objectif :

$$\min_{\Gamma \in \mathbb{R}^{n \times 2}} \|\mathbf{Y} - \Gamma\|_2^2 \text{ sous contrainte } K \text{ ruptures dans le signal}$$

- Problème de relaxation convexe :

$$\min_{\Gamma \in \mathbb{R}^{n \times 2}} \|\mathbf{Y} - \Gamma\|_2^2 + \lambda \sum_{j=1}^{n-1} \|\Gamma_{j+1} - \Gamma_j\|_2$$

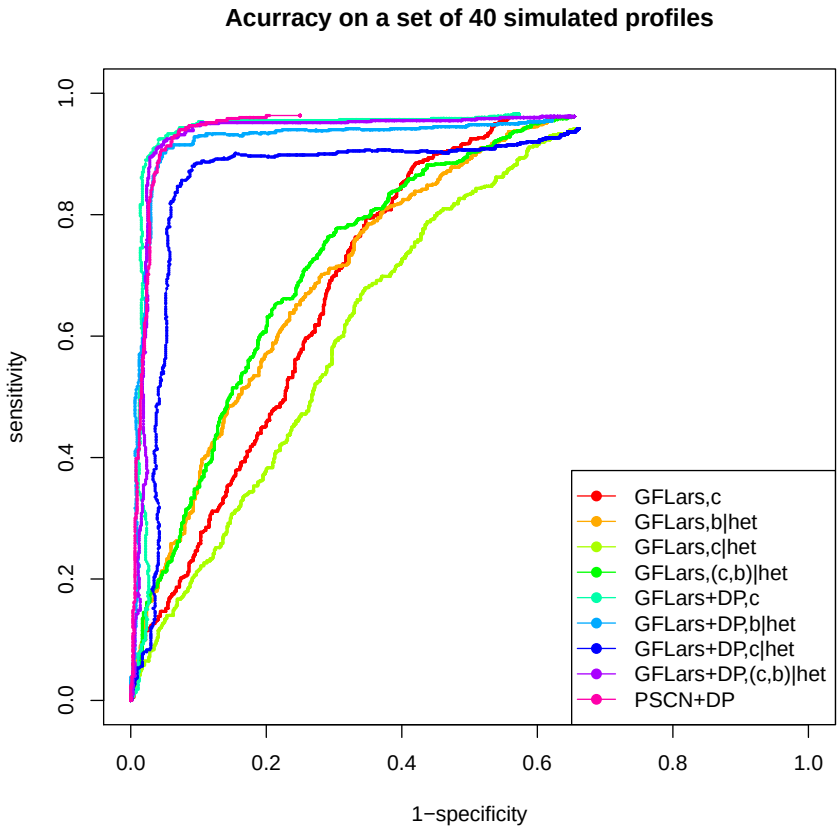
Sélection des meilleures ruptures par programmation dynamique

- Utilisation d'une méthode rapide mais approchée pour trouver des ruptures candidates $\Longrightarrow$ Méthode Group Lasso
- Utilisation d'une méthode exacte mais lente pour sélectionner les meilleures ruptures $\Longrightarrow$ Programmation dynamique

Résultats de la segmentation conjointe : simulations

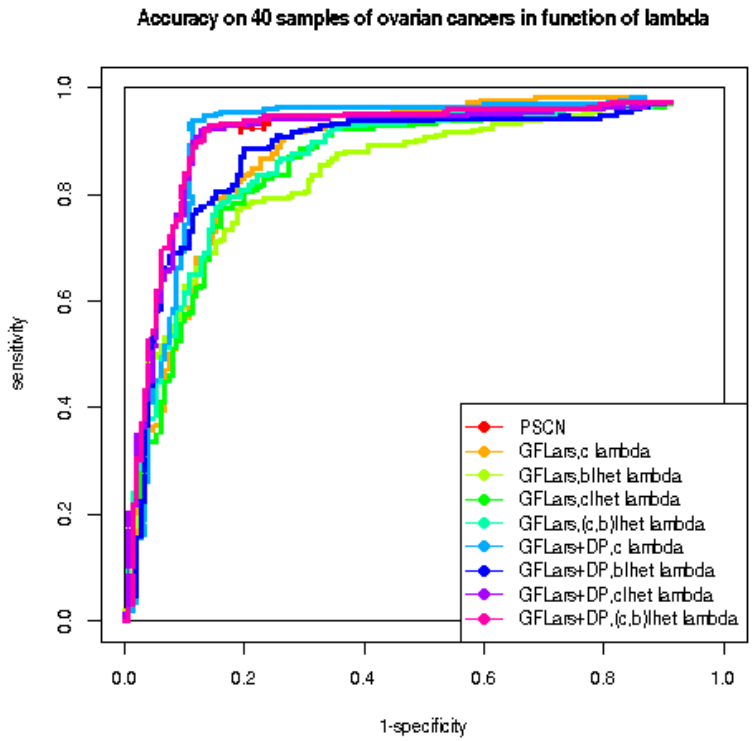
On a simulé 40 profils, et pour un profil

- On simule deux signaux de longueur 20000 à partir de données réelles
- On place 15 ruptures aléatoirement dans le signal
- On simule un état aléatoirement entre deux sauts



Résultats de la segmentation conjointe : données de cancer de l'ovaire

- On dispose de 40 échantillons tumoraux provenant de 40 patientes différentes
- On réalise les segmentations sur chaque profil indépendamment les uns des autres



Sommaire

Conclusion

Résultats obtenus

- Modèle pour la segmentation conjointe du nombre de copies d'ADN et le ratio allélique
- Évaluation de l'efficacité de la méthode sur des simulations et des données réelles, comparaison du modèle avec des méthodes existantes
- Implémentation logicielle (développement d'un package R) rapide
- Utilité de la programmation dynamique mise en évidence

Perspectives

- Amélioration de la méthode
- Segmenter plusieurs chromosomes et aussi plusieurs profils de patients
- Étiquetage des régions

Merci de votre attention

K. Bleakley and J.-P. Vert.

The group fused lasso for multiple change-point detection.
Technical report, Mines ParisTech, 2011.

H. Bondell and B. Reich.

Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR.
Biometrics, 2008.

Z. Harchaoui and C. Lévy-Leduc.

Catching change-points with lasso.
Advances in Neural Information Processing Systems, 2008.

P. Van Loo, S. H. Nordgard, O.C. Lingjaerd, I. H. Russnes, Hege Gand Rye, and V. N. Kristensen.

Allelic-specific copy number analysis of tumors.
Proceedings of the National Academy of Sciences of the United States of America, 2010.

A. B. Olshen, H. Bengtsson, P. Neuvial, P.T. Spellman, R. A. Olshen, and V. E. Seshan.

Parent-specific copy number in paired tumor-normmal studies using circular binary segmentation.
Bioinformatics, 2011.

G. Rigaill.

Pruned dynamic programming for optimal multiple change-point detection.
Technical report, http://arXiv.org/abs/1004.0887, 2010.

G. Rigaill, E. Lebarbier, and S. Robin.

Exact posterior distributions and model selection criteria for multiple change point-criteria.
Statistics and Computing, 2012.

J.-P. Vert and K. Bleakley.

Fast detection of multiple change-points shared by many signals using group LARS.
Advances in Neutral Information Processing Systems, 2010.

Tests utilisés

- Choisir une rupture, on prend deux échantillons (avant et après la rupture)
- Tester si les moyennes des deux échantillons sont égales

$$\mathcal{H}_0 : \mu_1 = \mu_2$$

Tests utilisés

- Choisir une rupture, on prend deux échantillons (avant et après la rupture)
- Tester si les moyennes des deux échantillons sont égales

$$\mathcal{H}_0 : \mu_1 = \mu_2$$

  - Statistiques univariées : Approximation de Welch
  - Statistique bivariée : Test de Welch James

Tests utilisés

- Choisir une rupture, on prend deux échantillons (avant et après la rupture)
- Tester si les moyennes des deux échantillons sont égales

$$\mathcal{H}_0 : \mu_1 = \mu_2$$

  - Statistiques univariées : Approximation de Welch
  - Statistique bivariée : Test de Welch James

$$T_v^2 = \mathbf{y}'\mathbf{S}^{-1}\mathbf{y},$$

où

$$\begin{aligned} \mathbf{S} &= \mathbf{S}_1 + \mathbf{S}_2, \\ \mathbf{S}_i &= \frac{\mathbf{S}_i^*}{n_i}, \qquad i = 1, 2 \end{aligned}$$

et

$$\mathbf{y} = \begin{pmatrix} (\bar{\mathbf{c}}_1 - \bar{\mathbf{c}}_2) & (\bar{\mathbf{b}}_1 - \bar{\mathbf{b}}_2) \end{pmatrix},$$

Distribution asymptotique de la statistique : $T_v^2 \sim_{\mathcal{H}_0} \chi^2(2)$

- On obtient des p-valeurs pour chacun des 3 tests
- On transforme ces p-valeurs en Z-scores : $z = \Phi^{-1}(p)$
- On réalise les étapes précédentes 5000 fois
- On obtient ensuite les distributions empiriques des Z-scores

Modélisation

On veut donc optimiser :

$$\min_{\Gamma \in \mathbb{R}^{n \times 2}} \frac{1}{2}\|\mathbf{Y} - \Gamma\|_2^2 + \lambda \sum_{j=1}^{n-1} \frac{\left((\Gamma_{j+1\,1} - \Gamma_{j\,1})^2 + (\Gamma_{j+1\,2} - \Gamma_{j\,2})^2\right)^{\frac{1}{2}}}{d_j} \tag{1}$$

Modélisation

On veut donc optimiser :

$$\min_{\Gamma \in \mathbb{R}^{n \times 2}} \frac{1}{2}\|\mathbf{Y} - \Gamma\|_2^2 + \lambda \sum_{j=1}^{n-1} \frac{\left((\Gamma_{j+1\,1} - \Gamma_{j\,1})^2 + (\Gamma_{j+1\,2} - \Gamma_{j\,2})^2\right)^{\frac{1}{2}}}{d_j} \tag{1}$$

Reformulation en régression group Lasso :

$$u = \Gamma_{1.} = (\Gamma_{1,1} \quad \Gamma_{1,2}) \in \mathbb{R}^{1 \times 2}$$

$$\beta_{j.} = \frac{(\Gamma_{j+1.} - \Gamma_{j.})}{d_j} = \frac{\left((\Gamma_{j+1,\,1} - \Gamma_{j,\,1}) \quad (\Gamma_{j+1,\,2} - \Gamma_{j,\,2})\right)}{d_j}$$

$\beta_{j.}$ représente le saut entre la j-ème et la $(j+1)$-ème position.

$$\Gamma = \mathbb{1}_{n,1} u + \mathbf{X}\beta. \tag{2}$$

L'équation peut se réécrire de la façon suivante :

$$\min_{\beta \in \mathbb{R}^{(n-1) \times 2}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta - \mathbb{1}_{n,1} u\|_2^2 + \lambda \sum_{j=1}^{n-1} (\beta_{j\,1}^2 + \beta_{j\,2}^2)^{\frac{1}{2}} \tag{3}$$

Modélisation

On veut donc optimiser :

$$\min_{\Gamma \in \mathbb{R}^{n \times 2}} \frac{1}{2}\|\mathbf{Y} - \Gamma\|_2^2 + \lambda \sum_{j=1}^{n-1} \frac{\left((\Gamma_{j+1\,1} - \Gamma_{j\,1})^2 + (\Gamma_{j+1\,2} - \Gamma_{j\,2})^2\right)^{\frac{1}{2}}}{d_j} \tag{1}$$

Reformulation en régression group Lasso :

$$u = \Gamma_{1.} = (\Gamma_{1,1} \quad \Gamma_{1,2}) \in \mathbb{R}^{1 \times 2}$$

$$\beta_{j.} = \frac{(\Gamma_{j+1.} - \Gamma_{j.})}{d_j} = \frac{\left((\Gamma_{j+1,\,1} - \Gamma_{j,\,1}) \quad (\Gamma_{j+1,\,2} - \Gamma_{j,\,2})\right)}{d_j}$$

$\beta_{j.}$ représente le saut entre la j-ème et la $(j+1)$-ème position.

$$\Gamma = \mathbb{1}_{n,1}\, u + \mathbf{X}\beta. \tag{2}$$

L'équation peut se réécrire de la façon suivante :

$$\min_{\beta \in \mathbb{R}^{(n-1) \times 2}} \frac{1}{2}\|\bar{\mathbf{Y}} - \bar{\mathbf{X}}\beta\|_2^2 + \lambda \sum_{j=1}^{n-1} \|\beta_{j,.}\|_2 \tag{3}$$

Algorithme group LARS

Initialisation

- $k = 1$
- $\beta^{[0]} = 0$
- $r^{[0]} = \mathbf{Y}$ est le résidu

On cherche la variable la plus corrélée avec Y

$$\mathcal{A}_1 = \arg \max_j \|\mathbf{X}_j' \mathbf{Y}\|_2^2$$

Algorithme group LARS

Pour k allant de 1 à K

A chaque étape k

- On calcule la direction w qui est une matrice de dimension $(n-1) \times 2$ tel que :

$$w_{\mathcal{A}_k^c} = 0, \text{ et } w_{\mathcal{A}_k} = (X'_{\mathcal{A}_k} X_{\mathcal{A}_k})^{-1} X'_{\mathcal{A}_k} r^{[k-1]}$$

où $X_{\mathcal{A}_k}$ est la matrice composée des colonnes de $\mathbf{X}$ correspondant aux indices de $\mathcal{A}_k$.

Algorithme group LARS

Pour k allant de 1 à K

A chaque étape k

- On calcule la direction w qui est une matrice de dimension $(n-1) \times 2$ tel que :

$$w_{\mathcal{A}_k^c} = 0, \text{ et } w_{\mathcal{A}_k} = (X'_{\mathcal{A}_k} X_{\mathcal{A}_k})^{-1} X'_{\mathcal{A}_k} r^{[k-1]}$$

  où $X_{\mathcal{A}_k}$ est la matrice composée des colonnes de $\mathbf{X}$ correspondant aux indices de $\mathcal{A}_k$.

- Pour chaque j qui n'est pas dans le groupes des indices actifs, on résout l'équation suivante pour $\gamma_j \in [0, 1]$ tel que :

$$\|X'_j(r^{[k-1]} - \gamma_j \mathbf{X} w)\|_2^2 = \|X'_{j'}(r^{[k-1]} - \gamma_j \mathbf{X} w)\|_2^2 \tag{4}$$

Algorithme group LARS

Pour k allant de 1 à K

A chaque étape k

- On calcule la direction w qui est une matrice de dimension $(n-1) \times 2$ tel que :

$$w_{\mathcal{A}_k^c} = 0, \text{ et } w_{\mathcal{A}_k} = (X'_{\mathcal{A}_k} X_{\mathcal{A}_k})^{-1} X'_{\mathcal{A}_k} r^{[k-1]}$$

  où $X_{\mathcal{A}_k}$ est la matrice composée des colonnes de $\mathbf{X}$ correspondant aux indices de $\mathcal{A}_k$.

- Pour chaque j qui n'est pas dans le groupes des indices actifs, on résout l'équation suivante pour $\gamma_j \in [0, 1]$ tel que :

$$\|X'_j(r^{[k-1]} - \gamma_j \mathbf{X} w)\|_2^2 = \|X'_{j'}(r^{[k-1]} - \gamma_j \mathbf{X} w)\|_2^2 \quad (4)$$

- On pose ensuite $\min_{j \notin \mathcal{A}_k} \gamma_j = \gamma_{j*}$, et on met à jour les indices actifs $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{j*\}$

Algorithme group LARS

Pour k allant de 1 à K

A chaque étape k

- On calcule la direction w qui est une matrice de dimension $(n-1) \times 2$ tel que :

$$w_{\mathcal{A}_k^c} = 0, \text{ et } w_{\mathcal{A}_k} = (X'_{\mathcal{A}_k} X_{\mathcal{A}_k})^{-1} X'_{\mathcal{A}_k} r^{[k-1]}$$

  où $X_{\mathcal{A}_k}$ est la matrice composée des colonnes de $\mathbf{X}$ correspondant aux indices de $\mathcal{A}_k$.

- Pour chaque j qui n'est pas dans le groupes des indices actifs, on résout l'équation suivante pour $\gamma_j \in [0, 1]$ tel que :

$$\|X'_j(r^{[k-1]} - \gamma_j \mathbf{X} w)\|_2^2 = \|X'_{j'}(r^{[k-1]} - \gamma_j \mathbf{X} w)\|_2^2 \quad (4)$$

- On pose ensuite $\min_{j \notin \mathcal{A}_k} \gamma_j = \gamma_{j*}$, et on met à jour les indices actifs $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{j*\}$
- On met également à jour $\beta^{[k]} = \beta^{[k-1]} + \gamma_{j*} w$ et $r^{[k]} = \mathbf{Y} - \mathbf{X}\beta^{[k]}$