Détection rapide des frontières des blocs d'une matrice constante par blocs bruitée

Vincent Brault[1,2] &Julien Chiquet[1,3]& Céline Lévy-Leduc[1,2]

[1] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay*
[2] *vincent.brault@agroparistech.fr*
[3] *julien.chiquet@agroparistech.fr*
[4] *celine.levy-leduc@agroparistech.fr*

Résumé. Dans différents contextes, on peut être amené à partitionner les lignes et les colonnes d'une matrice pour former un quadrillage de blocs homogènes sans effectuer de permutations ; c'est notamment le cas pour l'analyse des données Hi-C qui mesurent le degré d'interaction physique entre différentes positions du génome (Dixon et al.). En effet, ces données peuvent être modélisées comme une matrice constante par blocs bruitée. Toutefois, ce problème peut être compliqué pour plusieurs raisons : les méthodes utilisées en segmentation unidimensionnelle comme l'algorithme de programmation dynamique ne s'appliquent pas dans ce cas et la taille importante des données nécessite le développement et la mise en place d'algorithmes performants.

Notre approche est la suivante : nous montrons que ce problème peut être ramené à celui d'un modèle linéaire parcimonieux de grande dimension pour lequel nous proposons une méthode de sélection de variables rapide et efficace.

Dans cet exposé, nous montrerons comment notre méthode fournit un quadrillage pour des matrices de grandes tailles ($10\,000 \times 10\,000$). Nous montrerons également comment la structure bidimensionnelle permet d'obtenir une bonne estimation du nombre et des emplacements des ruptures. Nous illustrerons nos résultats à l'aide de figures et de films et comparerons nos méthodes avec d'autres approches sur des données simulées.

Mots-clés. Ruptures, LASSO, Données en 2 dimensions, Données Hi-C.

Abstract. In this paper, we propose a novel approach for estimating the location of block boundaries (change-points) in a random matrix consisting of a block wise constant matrix observed in white noise. Our contribution consists in rewriting this problem as a variable selection issue in a sparse high-dimensional linear model. We use a penalized least-squares criterion with an ℓ_1-type penalty for dealing with this problem.

We first provide some theoretical results ensuring the consistency of our change-point estimators. Then, we explain how to efficiently implement our method. Finally, we provide some numerical results to illustrate our methodology and apply our approach to Hi-C data which are used for better understanding the chromatin structure.

Keywords. Change-points, LASSO, Two-dimensional data, HiC experiments.

1 Introduction

La technologie Hi-C (High Chromosome Contact map) permet de mesurer le degré d'interaction physique entre différents loci (position le long d'un chromosome). Les données Hi-C sont représentées sous forme de matrices, parfois symétriques, où les indices des lignes et des colonnes correspondent à des loci et dont les éléments peuvent être modélisés par des variables aléatoires ayant des moyennes identiques au sein de blocs homogènes formant un quadrillage. Le but de l'analyse des données Hi-C est de fournir une méthode automatique et efficace d'estimation des frontières des blocs.

Dans nos travaux nous cherchons à segmenter les lignes et les colonnes dans le but d'obtenir un quadrillage. Pour cela, nous proposons une modélisation et un algorithme performant en termes de qualité d'estimation et de temps de calcul.

2 Modèle

Nous supposons qu'il existe des ruptures en ligne $\mathbf{t}_1^\star = (t_{1,1}^\star, \ldots, t_{1,K_1^\star}^\star)$ et en colonne $\mathbf{t}_2^\star = (t_{2,1}^\star, \ldots, t_{2,K_2^\star}^\star)$ telles que la matrice des interactions $\mathbf{Y} = (Y_{i,j})_{1 \leq i,j \leq n}$ puisse se décomposer en somme de deux matrices

$$\mathbf{Y} = \mathbf{U} + \mathbf{E}, \tag{1}$$

où $\mathbf{U} = (U_{i,j})$ est une matrice constante par blocs définie par

$$U_{i,j} = \mu_{k,\ell}^\star \quad \text{si } t_{1,k-1}^\star \leq i \leq t_{1,k}^\star - 1 \text{ et } t_{2,\ell-1}^\star \leq j \leq t_{2,\ell}^\star - 1, \tag{2}$$

avec la convention que $t_{1,0}^\star = t_{2,0}^\star = 1$ et $t_{1,K_1^\star+1}^\star = t_{2,K_2^\star+1}^\star = n+1$. Les coefficients $E_{i,j}$ de la matrice $\mathbf{E} = (E_{i,j})_{1 \leq i,j \leq n}$ sont supposés indépendants, de même loi et de moyenne nulle. Ainsi, les coefficients $Y_{i,j}$ sont supposés être des variables indépendantes avec des moyennes constantes par bloc.

Dans le cas particulier de matrices symétriques, nous pouvons considérer que $\mathbf{t}_1^\star = \mathbf{t}_2^\star$. Notre objectif est de proposer une méthode automatique pour estimer les $\mathbf{t}_1^\star$ et $\mathbf{t}_2^\star$.

3 Estimation

3.1 Réécriture du modèle

Nous commençons par remarquer que la matrice $\mathbf{U}$ peut être réécrite comme le produit d'une matrice $\mathbf{T}$ de taille $n \times n$ triangulaire inférieure ne contenant que des 1 et une matrice $\mathbf{B}$ de taille $n \times n$ où tous les éléments de la matrice sont nuls sauf ceux situés à l'intersection des instants de ruptures à savoir les éléments $\mathbf{B}_{i,j}$ tels que $(i,j) \in \{t_{1,0}^\star, \ldots, t_{1,K_1^\star}^\star\} \times \{t_{2,0}^\star, \ldots, t_{2,K_2^\star}^\star\}$:

$$\mathbf{Y} = \mathbf{T}\mathbf{B}\mathbf{T}^\top + \mathbf{E}, \tag{3}$$

où $\mathbf{T}^\top$ représente la transposée de $\mathbf{T}$. En notant $\mathrm{Vec}(\mathbf{X})$ la vectorisation de la matrice $\mathbf{X}$ formée de ses colonnes mises les unes au dessous des autres pour obtenir un vecteur donnant ainsi $\mathrm{Vec}(\mathbf{Y}) = \mathrm{Vec}(\mathbf{TBT}^\top)+\mathrm{Vec}(\mathbf{E})$. Ainsi, en utilisant le fait que $\mathrm{Vec}(\mathbf{AXC}) = (\mathbf{C}^\top \otimes \mathbf{A})\mathrm{Vec}(\mathbf{X})$, où $\otimes$ représente le produit de Kronecker, le modèle (3) peut-être réécrit comme

$$\mathcal{Y} = \mathcal{X}\mathcal{B} + \mathcal{E}, \tag{4}$$

où $\mathcal{Y} = \mathrm{Vec}(\mathbf{Y})$, $\mathcal{X} = \mathbf{T} \otimes \mathbf{T}$, $\mathcal{B} = \mathrm{Vec}(\mathbf{B})$ et $\mathcal{E} = \mathrm{Vec}(\mathbf{E})$. Grâce à ces transformations, le modèle (1) peut-être vu comme un modèle linéaire sparse où $\mathcal{Y}$ et $\mathcal{E}$ sont des vecteurs de dimension $n^2 \times 1$, $\mathcal{X}$ est une matrice $n^2 \times n^2$ et $\mathcal{B}$ un vecteur de taille $n^2 \times 1$ dont au plus $(K_1^\star + 1)(K_2^\star + 1)$ valeurs sont non nulles avec $K_1^\star$ et $K_2^\star$ très petits devant n.

3.2 Estimation des instants de rupture

L'estimation des instants du modèle (1) est alors équivalent à un problème de sélection de variables pouvant être résolu en minimisant, pour tout $\lambda_n > 0$, le critère suivant :

$$\widehat{\mathcal{B}}(\lambda_n) = \underset{\mathcal{B}\in\mathbb{R}^{n^2}}{\mathrm{Argmin}} \left\{ \|\mathcal{Y} - \mathcal{X}\mathcal{B}\|_2^2 + \lambda_n \|\mathcal{B}\|_1 \right\}, \tag{5}$$

où $\|u\|_2^2$ et $\|u\|_1$ sont définies pour tout vecteur u de $\mathbb{R}^N$ par $\|u\|_2^2 = \sum_{i=1}^N u_i^2$ et $\|u\|_1 = \sum_{i=1}^N |u_i|$. Le critère (5) peut être vu comme la méthode *Least Absolute Shrinkage and Selection Operator* (LASSO) très utilisée dans les modèles linéaires parcimonieux. Grâce à la pénalité ℓ_1, nous nous attendons au fait que l'estimateur $\widehat{\mathcal{B}}$ de $\mathcal{B}$ possède beaucoup de composantes nulles et que les éléments non-nuls soient aux mêmes endroits que ceux de $\mathcal{B}$.

Une fois les positions des éléments non-nuls de $\widehat{\mathcal{B}}$ connus, nous pouvons proposer des estimations de $(t_{1,k}^\star)_{1\le k\le K_1^\star}$ et de $(t_{2,k}^\star)_{1\le k\le K_2^\star}$. Pour cela, notons $\widehat{\mathcal{A}}(\lambda_n)$ l'ensemble des variables actives de $\widehat{\mathcal{B}}$: $\widehat{\mathcal{A}}(\lambda_n) = \left\{ j \in \{1,\dots,n^2\} : \widehat{\mathcal{B}}_j(\lambda_n) \neq 0 \right\}$. Pour chaque a de $\widehat{\mathcal{A}}(\lambda_n)$, considérons la division euclidienne de $(a-1)$ par n, à savoir $(a-1) = nq_a + r_a$ alors

$$\widehat{\mathbf{t}}_1 = (\widehat{t}_{1,k})_{1\le k\le |\widehat{\mathcal{A}}_1(\lambda_n)|} \in \{r_a+1 : a \in \widehat{\mathcal{A}}(\lambda_n)\},\ \widehat{\mathbf{t}}_2 = (\widehat{t}_{2,\ell})_{1\le \ell\le |\widehat{\mathcal{A}}_2(\lambda_n)|} \in \{q_a+1 : a \in \widehat{\mathcal{A}}(\lambda_n)\}$$
$$\text{où } \widehat{t}_{1,1} < \widehat{t}_{1,2} < \cdots < \widehat{t}_{1,|\widehat{\mathcal{A}}_1(\lambda_n)|}, \quad \widehat{t}_{2,1} < \widehat{t}_{2,2} < \cdots < \widehat{t}_{2,|\widehat{\mathcal{A}}_2(\lambda_n)|}. \tag{6}$$

Dans l'équation (6), $|\widehat{\mathcal{A}}_1(\lambda_n)|$ et $|\widehat{\mathcal{A}}_2(\lambda_n)|$ correspondent aux nombres d'éléments distincts dans $\{r_a : a \in \widehat{\mathcal{A}}(\lambda_n)\}$ et $\{q_a : a \in \widehat{\mathcal{A}}(\lambda_n)\}$ respectivement.

4 Résultats théoriques

4.1 Consistance des estimateurs

Pour établir la consistance des estimateurs $\widehat{\mathbf{t}}_1$ et $\widehat{\mathbf{t}}_2$ définis par (6), nous avons besoin de quatre hypothèses (**A1-A4**). Ces hypothèses utilisent les notations suivantes

$$I^\star_{\min} = \min_{0\leq k\leq K_1^\star} |t^\star_{1,k+1} - t^\star_{1,k}| \wedge \min_{0\leq k\leq K_2^\star} |t^\star_{2,k+1} - t^\star_{2,k}|,$$
$$J^\star_{\min} = \min_{1\leq k\leq K_1^\star, 1\leq \ell\leq K_2^\star+1} |\mu^\star_{k+1,\ell} - \mu^\star_{k,\ell}| \wedge \min_{1\leq k\leq K_1^\star+1, 1\leq \ell\leq K_2^\star} |\mu^\star_{k,\ell+1} - \mu^\star_{k,\ell}|,$$

qui correspondent au plus petit écart entre deux ruptures consécutives et à la plus petite différence de valeurs entre deux moyennes de blocs voisins. Ainsi, nous faisons les hypothèses suivantes :

(**A1**) Les variables aléatoires $(E_{i,j})_{1\leq i,j\leq n}$ sont indépendantes, identiquement distribuées, de moyenne nulle et telles qu'il existe une constante positive β telle que pour tout ν dans $\mathbb{R}$, $\mathbb{E}[\exp(\nu E_{1,1})] \leq \exp(\beta\nu^2)$.

(**A2**) La suite des (λ_n) utilisés dans (5) est telle que $(n\delta_n J^\star_{\min})^{-1}\lambda_n \to 0$, quand n tend vers l'infini.

(**A3**) La suite (δ_n) d'éléments positifs est décroissante et tend vers 0 telle que $n\delta_n {J^\star_{\min}}^2/\log(n) \to +\infty$, quand n tend vers l'infini.

(**A4**) $I^\star_{\min} \geq n\delta_n$.

Proposition 1 *Soit $(Y_{i,j})_{1\leq i,j\leq n}$ définie par (1) et $\widehat{t}_{1,k}$, $\widehat{t}_{2,k}$ définis par (6). En supposant (A1–A4) et que $|\widehat{\mathcal{A}}_1(\lambda_n)| = K_1^\star$ et $|\widehat{\mathcal{A}}_2(\lambda_n)| = K_2^\star$, nous avons avec probabilité tendant vers 1*

$$\mathbb{P}\left(\left\{\max_{1\leq k\leq K_1^\star} \left|\widehat{t}_{1,k} - t^\star_{1,k}\right| \leq n\delta_n\right\} \cap \left\{\max_{1\leq k\leq K_2^\star} \left|\widehat{t}_{2,k} - t^\star_{2,k}\right| \leq n\delta_n\right\}\right) \underset{n\to+\infty}{\longrightarrow} 1. \qquad (7)$$

4.2 Implémentation

Pour calculer le vecteur $\widehat{\mathcal{B}}$ du modèle (5), nous utilisons l'algorithme LARS proposé par Efron et al. [1] et Osborne et al. [2]. L'implémentation standard du LARS menant à une complexité en $\mathcal{O}\left(n^4 |\mathcal{A}_{max}| + n^2 |\mathcal{A}_{max}|^2 + |\mathcal{A}_{max}|^3\right)$ où $|\mathcal{A}_{max}|$ est le nombre maximum de variables actives cherchées n'est pas acceptable en vue du traitement de données réelles. Or, comme la matrice $\mathcal{X}$ possède une forme particulière que nous connaissons, nous avons pu optimiser l'algorithme pour obtenir une complexité beaucoup plus faible :

Proposition 2 *Pour une matrice* $\mathbf{Y}$ *de taille* $n \times n$*, la complexité de notre algorithme est*

$$\mathcal{O}\left(n^2 + |\mathcal{A}_{max}|\left(n^2 + |\mathcal{A}_{max}|^2\right)\right).$$

Comme on s'attend à avoir un nombre de variables actives très inférieur au nombre d'éléments de la matrice, l'essentiel de la complexité est de l'ordre de n^2 , c'est-à-dire linéaire en le nombre d'éléments de la matrice.

5 Simulations

Pour montrer empiriquement les résultats des propositions 1 et 2, nous avons simulé des matrices suivant le modèle (1) où $\mathbf{U}$ est une matrice symétrique constante par blocs :

$$\left(\mu^\star_{k,\ell}\right)_{k\in\left\{1,\dots,K_1^\star+1\right\},\ell\in\left\{1,\dots,K_2^\star+1\right\}} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix},$$

où $\mu^\star$ est défini par l'équation (2) et les $E_{i,j}$ sont des variables gaussiennes indépendantes, identiquement distribuées, de moyenne nulle et de variance σ^2 où σ appartient à $\{1, 2, 5\}$. Des exemples de matrices générées avec $n = 500$, $K_1^\star = K_2^\star = 4$ sont présentées sur la ligne supérieure de la figure 1.

Pour évaluer la qualité de l'estimation des instants de ruptures, nous avons présenté sur la ligne inférieure de la figure 1 les courbes ROC (coupées à 0.20 pour la proportion de faux positifs).

Nous voyons que l'estimation des ruptures inclut très rapidement les vraies ruptures.

Pour évaluer la rapidité de notre algorithme proposé dans le package **BlockSeg**, nous avons fait varier le nombre de lignes et de colonnes $n \in \{100, 250, 500, 1000, 2500, 5000\}$ et le degré maximal de sparsité $s \in \{50, 100, 250, 500, 750\}$.

Nous pouvons voir qu'il faut moins de deux minutes pour analyser une matrice de taille $n \times n = 1\,000 \times 1\,000$.

6 Conclusion

Dans cet exposé, nous développerons les résultats obtenus et comparerons nos méthodes avec d'autres méthodes sur des données simulées.

Bibliographie

[1] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004). Least angle regression. *The Annals of statistics* 32 407–499.

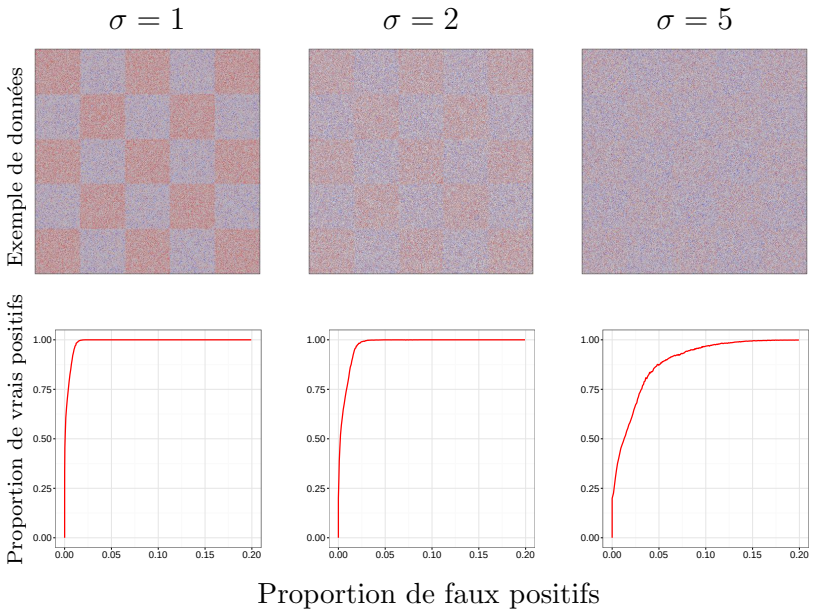

Figure 1: Haut : exemple de matrices $\mathbf{Y}$ générées suivant le modèle (1). Bas : courbe ROC des estimations des ruptures en lignes.

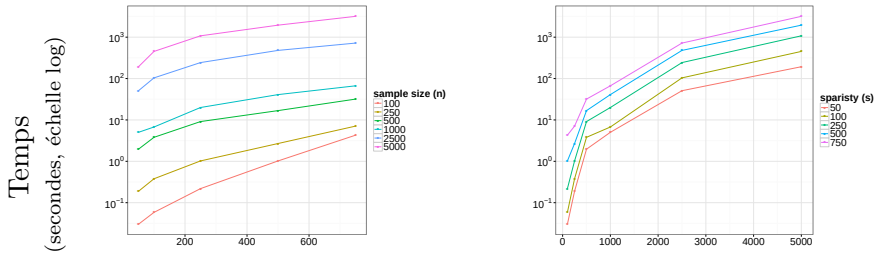

Figure 2: À gauche (resp. à droite) : Estimation du temps utilisé par l'algorithme (en seconde, échelle logarithmique) pour différentes valeurs de nombre de lignes (resp. degrés maximal de sparsité) en fonction du degrés maximal de sparsité (resp. du nombre de lignes).

[2] Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA journal of numerical analysis* 20 389–403.
[3] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.