Vraisemblance auto-pénalisante pour la sélection du nombre de ruptures dans la segmentation bidimensionnelle utilisée pour l'analyse des données Hi-C

Vincent Brault [1,2] & Maud Delattre [1,2] & Émilie Lebarbier [1,2] & Céline Lévy-Leduc [1,2] & Tristan Mary-Huard [1,2,3]

vincent.brault@agroparistech.fr
[1] *INRA, UMR 518 MIA, F-75005 Paris, France*
[2] *AgroParisTech, UMR 518 MIA, F-75005 Paris, France*
[3] *UMR de Génétique Végétale, INRA, Université Paris-Sud, CNRS, Gif-sur-Yvette, France*

Résumé. Nous proposons d'étudier un modèle statistique utilisé pour analyser les données Hi-C. Ces données représentent la mesure du degré d'interaction physique entre différentes positions chromosomiques (voir par exemple Dixon et al. [1]) : les zones de fortes interactions dans le génome forment des blocs diagonaux de valeurs homogènes différentes du reste de la matrice. Dans ce cadre, Lévy-Leduc et al. [3] proposent un modèle de segmentation bidimensionnelle d'une matrice symétrique dont l'objectif est de retrouver les instants de ruptures délimitant ces blocs.

Dans leur article, Lévy-Leduc et al. [3] utilisent un algorithme de programmation dynamique pour estimer les instants de ruptures maximisant la vraisemblance et proposent de sélectionner leur nombre en maximisant cette dernière sans pénalisation.

Dans cet exposé, nous démontrerons que l'estimation obtenue du nombre de ruptures est consistante si la distance minimale entre deux ruptures estimées n'est pas trop petite. Ce résultat sera illustré par des simulations.

Mots-clés. Sélection de modèle, Données Hi-C, Segmentation.

Abstract. A statistic model to analyse the Hi-C data is studied in this talk. These data measure the degree of physical interaction between various positions within a chromosome (see for example Dixon et al. [1]) : zones of strong chromosomic interactions form homogeneous diagonal blocks. Lévy et al. [3] use a technic of bidimensional segmentation for symmetric matrix to aim at finding the boundaries of these blocks.

In their article, Lévy-Leduc et al. [3] use a dynamic programming algorithm to estimate the change-point maximising the likelihood and suggest to estimate the number of change-point maximising the likelihood without using any penalisation.

In this talk, the consistency of this estimation will be demonstrated under the condition that the minimal distance between two change-points is large enough. This result will be illustrated by simulations.

Keywords. Model selection, Hi-C data, Segmentation.

1 Introduction

La technologie Hi-C (High Chromosome Contact map) permet de mettre en évidence des interactions entre des loci (emplacements sur un chromosome) spatialement proches : l'objectif est alors de détecter les régions réagissant entre elles. Les données Hi-C sont représentées sous forme de matrices (symétriques si les mêmes chromosomes sont mis en ligne et en colonne) telles que les zones topologiques induisent des blocs sur la diagonale composés de fortes interactions. Lévy-Leduc et al. [3] proposent un modèle pour identifier les frontières de ces blocs diagonaux.

2 Modèle

Les données sont représentées sous la forme d'une matrice symétrique $\mathbf{y} = (y_{i,j})_{1 \leq i,j \leq n}$ de taille $n \times n$ où chaque case $y_{i,j}$ représente le degré d'interaction entre le locus i et le locus j ; par la suite, seules les cases situées dans la partie supérieure droite (c'est-à-dire vérifiant $1 \leq i \leq j \leq n$) seront étudiées. Lévy-Leduc et al. [3] supposent qu'il existe $K^\star$ blocs $D_k^\star$ inconnus définis uniquement par le vecteur des instants de ruptures $\mathbf{t}^\star = (t_0^\star, \ldots, t_{K^\star}^\star)$:

$$D_k^\star = \left\{ (i,j) \in \{1,\ldots,n\}^2 \middle| \, t_{k-1}^\star \leq i \leq j \leq t_k^\star - 1 \right\}$$

avec les conventions que $t_0^\star = 1$ et $t_{K^\star}^\star = n+1$ (voir la figure 1 pour une représentation graphique de ces notations). Les cases restantes appartiennent à un ensemble noté $E_0^\star$:

$$E_0^\star = \left\{ (i,j) \in \{1,\ldots,n\}^2 \middle| \, 1 \leq i \leq j \leq n \right\} \bigcap \overline{(\cup_{k=1}^{K^\star} D_k^\star)}.$$

Enfin, les variables $Y_{i,j}$ sont supposées indépendantes et, pour chaque zone, de même loi de moyenne $\mu_k^\star$; toutes ces lois appartenant à une même famille paramétrique et les éventuels autres paramètres sont supposés constants pour toute la matrice.

3 Vraisemblance

Dans leur article, Lévy-Leduc et al. [3] utilisent un algorithme de programmation dynamique pour obtenir les estimateurs du maximum de vraisemblance des instants de ruptures et des moyennes. Pour un nombre fixé K de ruptures, nous notons $\mathcal{L}_K(\mathbf{Y}; \boldsymbol{\mu}, \mathbf{t})$ la log-vraisemblance complète.

Les auteurs constatent empiriquement que la fonction

$$K \mapsto \max_{\boldsymbol{\mu}, \mathbf{t}} \mathcal{L}_K(\mathbf{Y}; \boldsymbol{\mu}, \mathbf{t})$$

possède un comportement atypique : elle commence par croître avec le nombre de ruptures puis se stabilise plus ou moins longtemps avant de décroître (voir les exemples de la

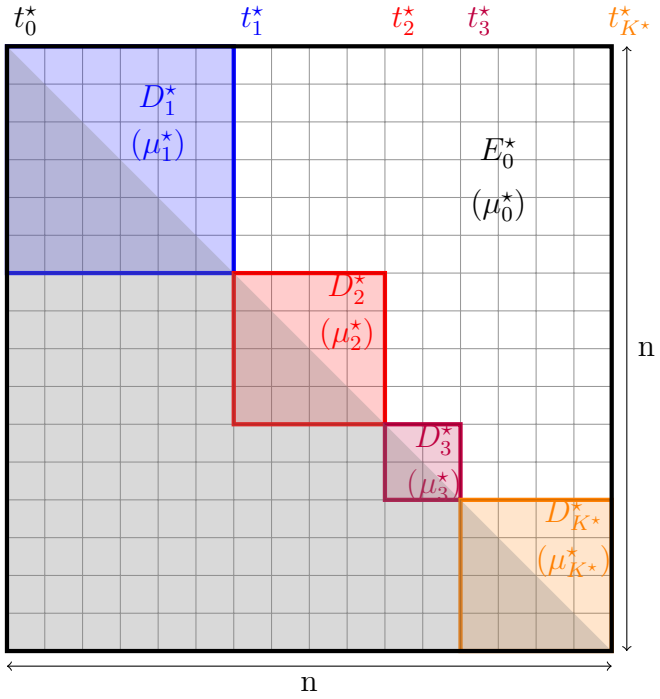

FIGURE 1 – Représentation graphique des notations pour une matrice avec $n = 16$ lignes et $K^\star = 4$ blocs.

figure 2). Ils proposent alors de sélectionner le nombre de ruptures correspondant à ce maximum :

$$\widehat{K} \in \underset{1 \leq K \leq K_{\max}}{\operatorname{argmax}} \max_{\boldsymbol{\mu}, \mathbf{t}} \mathcal{L}_K(\mathbf{Y}; \boldsymbol{\mu}, \mathbf{t}).$$

4 Résultats

4.1 Résultats théoriques

Les résultats sur la consistance de l'estimateur du nombre de ruptures que nous montrerons sont obtenus sous les hypothèses suivantes :

- Chaque variable se décompose par la somme de leur moyenne et d'une variable $\varepsilon_{i,j}$:

  $$Y_{i,j} = \mathbb{E}[Y_{i,j}] + \varepsilon_{i,j}$$

  où les variables $\varepsilon_{i,j}$ sont centrées, indépendantes, symétriques, identiquement distribuées et telles qu'il existe une constante $\beta > 0$ telle que pour tout $\nu \in \mathbb{R}$:

  $$\mathbb{E}[e^{\nu \varepsilon_{11}}] \leq e^{\beta \nu^2}.$$

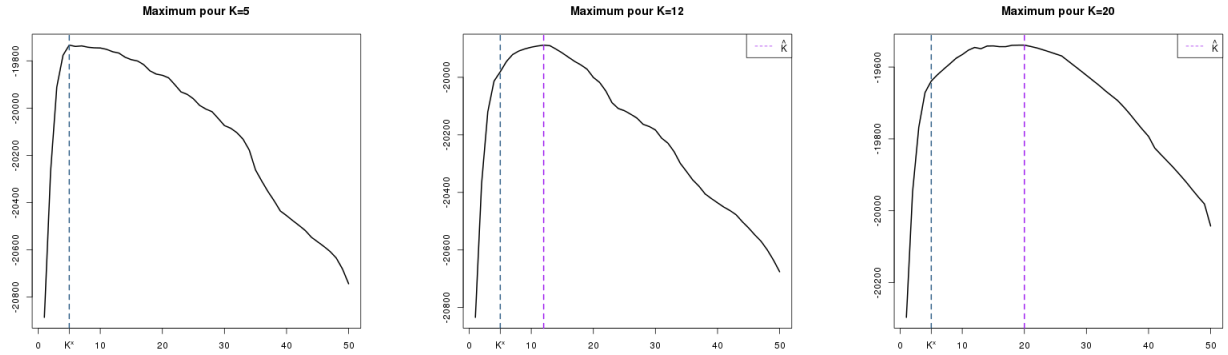

FIGURE 2 – Représentations de différentes log-vraisemblances pour des matrices de tailles 100×100 ayant 5 blocs (l'emplacement étant symbolisé par une ligne bleue) avec l'emplacement du maximum (donc de $\widehat{K}$) à différents endroits (symbolisé par une ligne violette) : 5 à gauche, 12 au milieu et 20 à droite. Les matrices ont été simulées suivant la procédure de la partie 4.2 : les valeurs des cases des 5 blocs sont simulées suivant des lois gaussiennes $\mathcal{N}(1,4)$ et celles de la région $E_0^\star$ suivant une loi $\mathcal{N}(0,4)$. Dans les trois cas, la courbe croît puis décroît.

- La distance minimale entre les moyennes des blocs et celle de $E_0^\star$ définie par :

$$\underline{\lambda}^{(0)} = \min_{1 \leq k \leq K^\star} |\mu_k^\star - \mu_0^\star|$$

  est strictement positive.

Ainsi, nous montrerons les deux théorèmes suivants :

Théorème 4.1. ***Probabilité de sous-estimer le nombre de ruptures***
Sous les hypothèses précédentes, nous avons pour tout $K < K^\star$:

$$\mathbb{P}\left(\widehat{K} = K\right) \underset{n \to +\infty}{\longrightarrow} 0.$$

De plus, si nous notons $\mathcal{A}_{n,K}$ l'ensemble des vecteurs $\mathbf{t}$ possibles pour K ruptures, $\mathcal{A}_{n,K}^{\Delta_n}$ l'ensemble de vecteurs de ruptures ayant un écart plus grand que $n\Delta_n$ pour $\Delta_n > 0$:

$$\mathcal{A}_{n,K}^{\Delta_n} = \{(t_0, \ldots, t_K) \in \mathcal{A}_{n,K} | \forall 1 \leq k \leq K,\ t_k - t_{k-1} \geq n\Delta_n\}$$

et si nous notons $\mathbf{t}_{\Delta_n}^{(n)}$ l'estimateur dans l'ensemble $\mathcal{A}_{n,K}^{\Delta_n}$ et $\widehat{K}_{\Delta_n}^{(n)}$ l'estimateur de $K^\star$ associé alors nous avons le théorème suivant :

Théorème 4.2. ***Probabilité de sur-estimer le nombre de ruptures***
Si Δ_n tend vers 0 lorsque n tend vers l'infini de telle sorte que :

$$\Delta_n \frac{n}{\sqrt{\log n}} \underset{n \to +\infty}{\longrightarrow} +\infty \tag{1}$$

alors nous avons pour tout $K > K^\star$ *:*

$$\mathbb{P}\left(\widehat{K}^{(n)}_{\Delta_n} = K\right) \underset{n\to+\infty}{\longrightarrow} 0.$$

L'hypothèse d'écart minimal entre deux ruptures est classique dans les modèles de ruptures (voir par exemple Lavielle et Moulines [2]) et cohérente avec le problème biologique.

4.2 Applications numériques

Nous avons fait des simulations en prenant $K^\star = 5$ blocs, des lois gaussiennes avec la même moyenne $\mu^\star_k = 1$ pour tous les blocs et valant $\mu^\star_0 = 0$ pour la zone $E^\star_0$. Nous faisons ensuite varier l'écart type (allant de 1 à 10 par pas de 1) et le nombre n de lignes (500, 800 et 1200) ; pour chaque combinaison, nous avons simulé 500 matrices et avons estimé le nombre de ruptures par la procédure proposée ; la figure 3 représentant les diagrammes en boîte ces estimations. Nous voyons que plus le nombre d'observations augmente et plus les estimations sont proches de la vraie valeur.

4.3 Conclusions

Durant cet exposé, nous démontrerons en quoi la structure particulière du modèle permet de ne pas avoir à pénaliser la log-vraisemblance pour obtenir un estimateur consistant du nombre de ruptures et nous illustrerons ces résultats à l'aide de données simulées.

Bibliographie

[1] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, et B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398) : 376380, 2012.
[2] M. Lavielle et E. Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of time series analysis*, 21(1) :3359, 2000.
[3] C. Lévy-Leduc, M. Delattre, T. Mary-Huard, et S. Robin. Two-dimensional segmentation for analyzing hic data. *Bioinformatics*, 30(17) :386392, 2014.

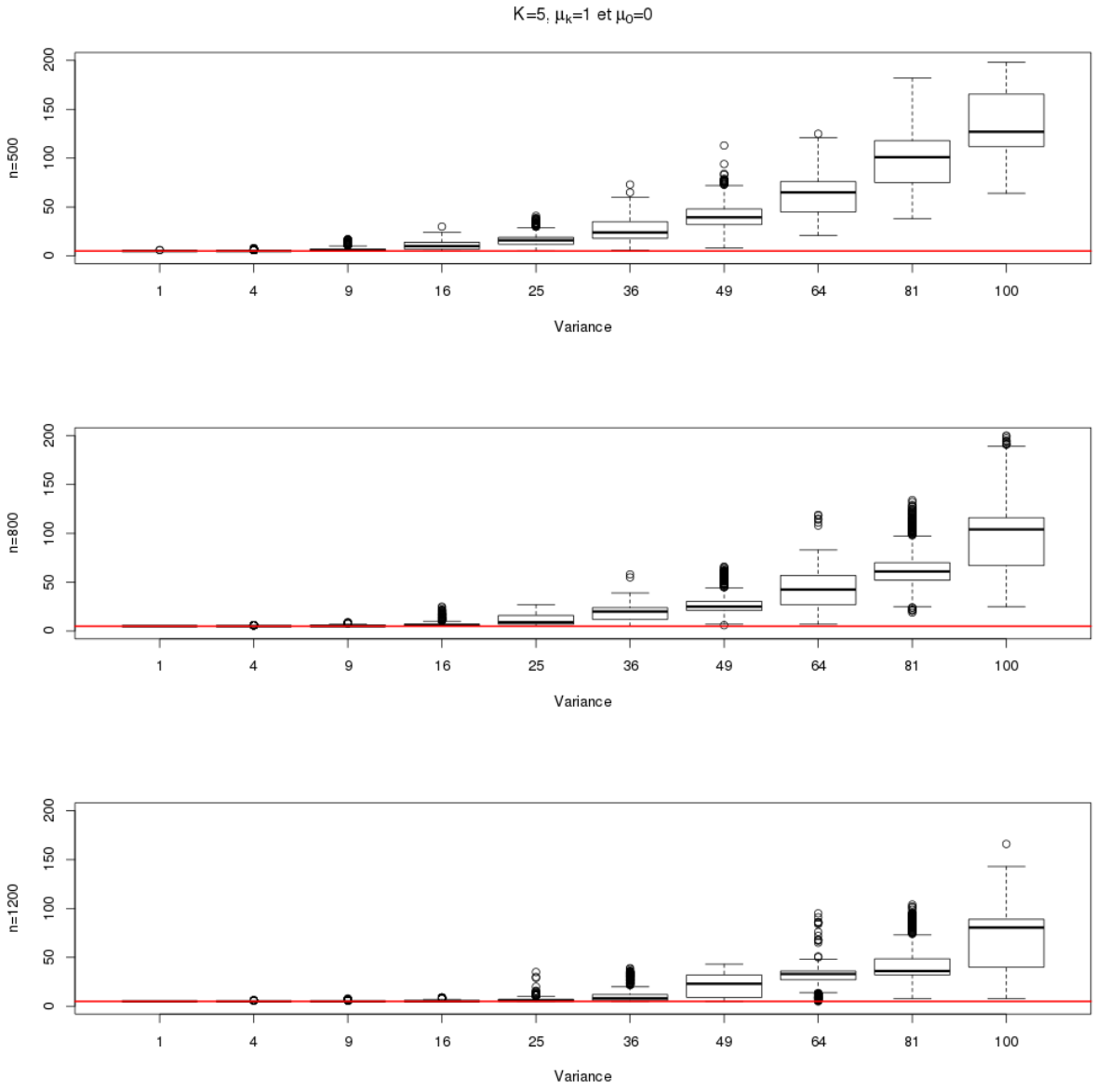

FIGURE 3 – Diagrammes en boîte des estimations du nombre de blocs suivant le nombre n de lignes (en ligne) et la variance (en colonne). Le trait horizontal rouge indique le vrai nombre de blocs.