Package BlockSeg pour la détection rapide des frontières des blocs d'une matrice constante par blocs bruitée

V. Brault[a,b], **J. Chiquet**[a,c] and **C. Lévy-Leduc**[a,d]

[a]UMR MIA-Paris
AgroParisTech, INRA, Université Paris-Saclay
16 rue Claude Bernard, F-75231 Paris Cedex 05

[b]vincent.brault@agroparistech.fr

[c]julien.chiquet@agroparistech.fr

[d]celine.levy-leduc@agroparistech.fr

Mots clefs : Modèle de ruptures, Modèle linéaire sparse en grande dimension, Données Hi-C.

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. [2]). 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.

Pour répondre à cette question, Brault et al. [1] proposent le modèle supposant 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.
Ils montrent que ce problème peut être ramené à celui d'un modèle linéaire parcimonieux de grande dimension pour lequel ils proposent une méthode de sélection de variables rapide et efficace basée sur un critère LASSO (*Least Absolute Shrinkage and Selection Operator*).
Pour sélectionner le nombre de blocs, les auteurs utilisent une adaptation de la *stability selection* proposée par [3] estimant un poids pour chaque indice i en ligne (resp. en colonne) qui sera plus fort si l'indice i correspond à l'emplacement d'une rupture. Ils suggèrent ensuite de conserver comme instant de ruptures les indices associés aux poids plus forts qu'un certain seuil mais ne proposent pas de valeurs universelles pour le choix de ce seuil ; précisant toutefois qu'ils ont observé empiriquement que ce dernier dépend beaucoup de la forme des données.

Les méthodes développées dans [1] sont implémentées dans le package `blockseg`. Dans cet exposé, nous commencerons par un rappel du modèle et de la méthode utilisée puis illustrerons le comportement de l'algorithme à l'aide d'un film explicitant les différentes étapes. Nous terminerons par une présentation des fonctions du package `blockseg`, notamment les sorties graphiques de la fonction `stab.blockSeg` permettant la visualisation des différentes ruptures et des matrices résumées en fonction de différents seuils (voir figure 1).

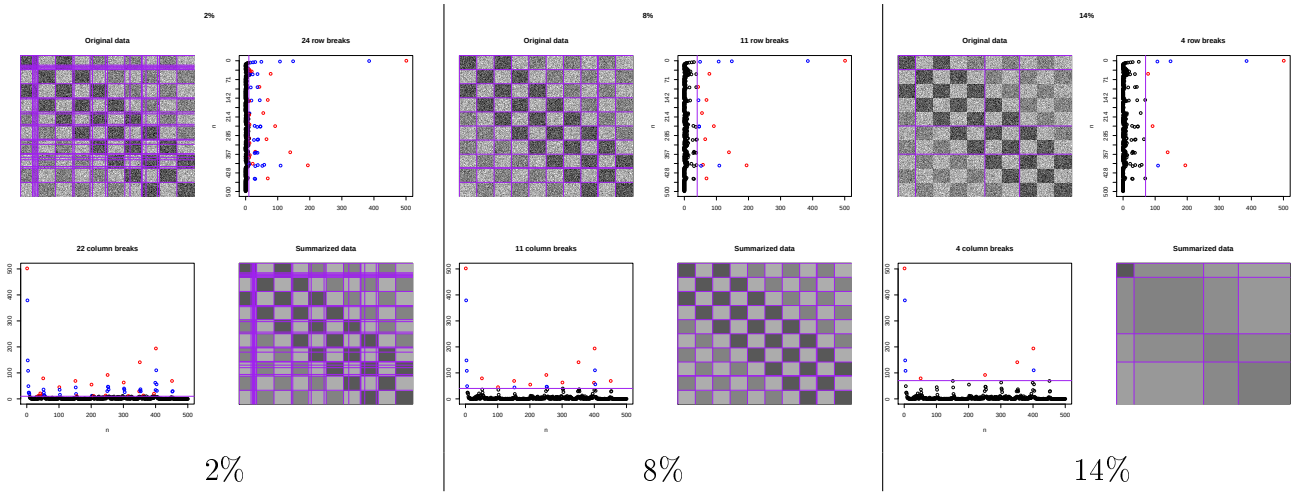

Figure 1: Exemples de sorties de la fonction `stab.blockSeg` pour différents seuils (2%, 8% et 14% du poids maximal attribué à un indice) sur une matrice jouet : si le seuil est trop bas, il y a une sur-estimation du nombre des ruptures et s'il est trop haut, il y a une sous-estimation.

Références

[1] Brault, V. and Chiquet, J. and Lévy-Leduc, C. (2016). Fast Detection of Block Boundaries in Block Wise Constant Matrices: An Application to HiC data. *pre-print*, Version arXiv:1603.03593.

[2] Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. **485** 376–380.

[3] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 417–473.