Modélisation statistique pour détecter des séquences vidéos similaires : application aux véhicules autonomes

Vincent Brault, Adeline Samson, Jean-Charles Quinton

Modélisation statistique pour détecter des séquences vidéos similaires : application aux véhicules autonomes

Vincent Brault[(1)] en collaboration avec Adeline Leclerc-Samson[(1)] et Jean-Charles Quinton[(1)]
à partir d'un travail en collaboration avec Céline Lévy-Leduc[(2)], Sarah Ouadah[(2)] et Laure Sansonnet[(2)]

[(1)]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP [1], LJK, 38000 Grenoble, France
[(2)]UMR MIA-Paris, INRA, AgroParisTech, Université Paris-Saclay, 75005, Paris, France

1 Introduction

Dans la recherche sur les véhicules autonomes, le coût élevé des GPS haute précision est un frein à leur usage exclusif pour la localisation ou navigation. Ceux-ci sont donc réservés à la validation de méthodes utilisant d'autres capteurs. Pour des véhicules réalisant un ensemble de trajets prédéterminés (comme les bus de ville), des caméras peuvent par exemple filmer l'environnement, et la similarité entre les images capturées à différents moments utilisée pour la localisation. Parmi les mesures de similarité proposées dans la littérature, une mesure basée sur les algorithmes de saillance visuelle est utilisée ici (Birem et al., 2014). Les données résumées issues de séquences vidéo réelles (Korrapati et al., 2013) se présentent sous forme de matrices dans lesquelles des lieux différenciés (e.g. ligne droite, intersection...) devraient correspondre à des blocs relativement homogènes.

Sur la figure 1, nous avons représenté à gauche les coordonnées GPS du parcours d'un véhicule dont Korrapati et al. (2013) ont filmé le trajet depuis une caméra située à l'avant ; à droite se trouve la matrice de similarité des images où chaque case représente la ressemblance entre deux images du film (rouge pour une forte similarité et bleu pour une faible similarité). Les blocs rouges sur la diagonale correspondent principalement à des lignes droites sauf au début ou à la fin où ils correspondent au moment où la voiture était arrêtée et que toutes les images étaient donc identiques. À l'opposé, nous voyons vers le centre de la partie supérieure gauche une partie de la diagonale rouge entourée de cases bleues elles-mêmes encadrées par des petits segments rouges parallèles à la diagonale : ceci correspond au moment où la voiture a fait deux tours de rond points donc avec des images successives qui sont très vite différentes (les cases bleues autour de la diagonale) puis les images du deuxième tour qui ressemblent à celles du premier tour (ce qui donne le petit segment rouge).

Le but est alors de proposer une méthode automatique pour estimer les frontières de ces blocs. Or, il existe une problématique similaire développée en biologie pour l'analyse des données Hi-C (Dixon et al., 2012). Dans ce travail, nous avons cherché à utiliser la procédure proposée par Brault et al. (2018b) fondée sur les statistiques de rang afin d'étudier les segmentations obtenues.

Dans cet article, nous présentons dans un premier temps la modélisation basée sur la théorie des tests non paramétriques et expliquons comment nous pouvons estimer les blocs de la matrice. Dans un second temps, nous détaillons les résultats théoriques obtenus par Brault et al. (2018b) et donnons les astuces utilisées dans les démonstrations. Enfin, nous montrons les résultats obtenus sur le trajet présenté sur la figure 1.

2 Contexte statistique

Dans cette partie, nous présentons la modélisation choisie et expliquons comment la théorie des tests permet de répondre à la problématique.

1. Institute of Engineering Univ. Grenoble Alpes

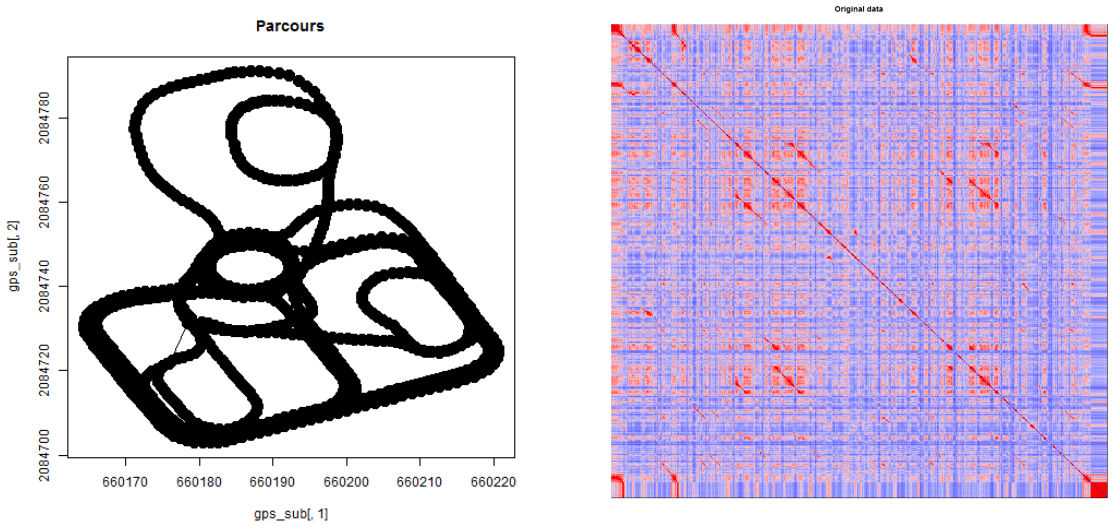

FIGURE 1 – À gauche se trouvent les coordonnées GPS du parcours fait par le véhicule dont le film a été étudié. À droite se trouve la matrice de similarités des images vidéos associées.

2.1 Modélisation

Nous notons $\mathbf{X} = (X_{i,j})_{1 \leq i,j \leq n}$ la matrice de données symétriques, c'est-à-dire que pour tout $(i,j) \in \{1,\dots,n\}^2$, $X_{i,j} = X_{j,i}$. Étant donné un entier $n_1 \in \{1,\dots,n-1\}$ inconnu, nous supposons qu'il existe $2n$ lois de probabilité $\left(\mathbb{P}_1^{(1)}, \mathbb{P}_2^{(1)}, \mathbb{P}_1^{(2)}, \dots, \mathbb{P}_2^{(n)}\right)$ telles que pour tout entier $i \in \{1,\dots,n\}$, les variables $(X_{i,j})_{1 \leq j \leq n}$ sont indépendantes, les variables $(X_{i,j})_{1 \leq j \leq n_1}$ suivent la même loi $\mathbb{P}_i^{(1)}$ et les variables $(X_{i,j})_{n_1+1 \leq j \leq n}$ la même loi $\mathbb{P}_i^{(2)}$ (une représentation schématique des notations est proposée à gauche de la figure 2).

Le but par la suite est de savoir s'il existe au moins une ligne i telle que la loi $\mathbb{P}_i^{(1)}$ soit différente de la loi $\mathbb{P}_i^{(2)}$, dans ce cas, nous dirons alors que n_1 est une *rupture*. Pour cela, nous utilisons la théorie des tests.

2.2 Statistique de test

Le principe d'un test est de mettre en concurrence deux hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ et de décider laquelle serait la plus probable. Dans le cas présenté ici, nous testons la probabilité que n_1 ne soit pas une rupture :

$$\mathcal{H}_0 \ : \forall i \in \{1,\dots,n\},\ \mathbb{P}_i^{(1)} = \mathbb{P}_i^{(2)}$$

contre l'hypothèse que n_1 en soit bien une :

$$\mathcal{H}_1 \ : \exists i \in \{1,\dots,n\},\ \mathbb{P}_i^{(1)} \neq \mathbb{P}_i^{(2)}.$$

Pour pouvoir confronter ces deux hypothèses, nous utilisons une *statistique de rang* inspirée des travaux de Lung-Yut-Fong et al. (2011) dont le principe est de calculer le rang de chaque case et de voir si les rangs sont plutôt mélangés ou, au contraire, que tous les petits rangs sont du même côté ; signifiant qu'une des lois semblent donner des plus petites valeurs que l'autre. Pour cela, nous introduisons la

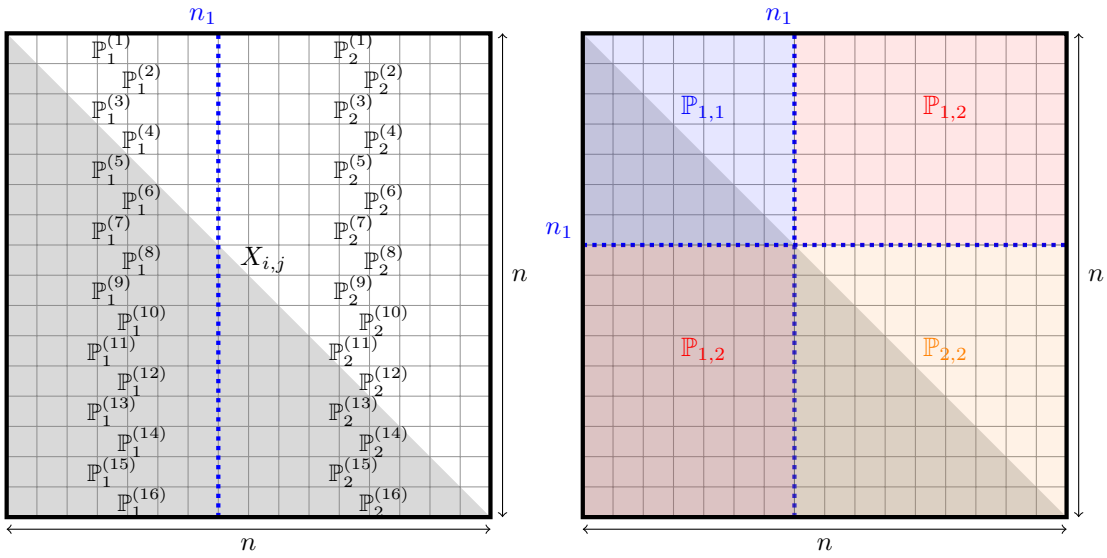

FIGURE 2 – Résumé des notations introduites dans les sections 2.1 et 3.1 : à gauche, les hypothèses faites avec deux lois par ligne ; à droite, la version résumée avec les conséquences de la symétrie. Le triangle grisé rappelle la symétrie de la matrice.

fonction $h : \mathbb{R} \times \mathbb{R} \to \{-1, 0, 1\}$ définie par $h(x, y) = \mathbb{1}_{\{x \leq y\}} - \mathbb{1}_{\{y \leq x\}}$ où $\mathbb{1}$ est la fonction indicatrice valant 1 si la condition est vérifiée et 0 sinon. Nous définissons alors pour tout $i \in \{1, \ldots, n\}$, la statistique :

$$U_{n,i}(n_1) = \frac{1}{\sqrt{nn_1(n - n_1)}} \sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} h(X_{i,j_0}, X_{i,j_1}),$$

qui compare chaque case située à gauche de la potentielle rupture n_1 avec chaque case située à droite.

Sur la figure 3, nous mettons deux exemples de lignes composées de 14 cases avec une rupture à la $6^{\text{ème}}$ ligne en supposant que les cases grises correspondent à des valeurs plus hautes que celles des cases blanches :

- dans le premier cas, les cases sont plutôt mélangées de part et d'autre de la ligne bleue et nous avons :
$$U_{14,i}(6) = \frac{-8}{\sqrt{14 \times 6 \times 8}}$$
donc une valeur proche de zéro ;
- dans le second cas, les cases grises sont plutôt à gauche de la rupture tandis que les cases blanches sont à droite, nous avons alors :
$$U_{14,i}(6) = \frac{35}{\sqrt{14 \times 6 \times 8}}$$
donc une valeur plutôt éloignée de zéro.

Le principe est donc de sommer les effets sur toutes les lignes afin d'obtenir une statistique qui devrait être proche de 0 si les lois avant et après la rupture sont identiques et s'en écarter sinon. Pour cela, nous

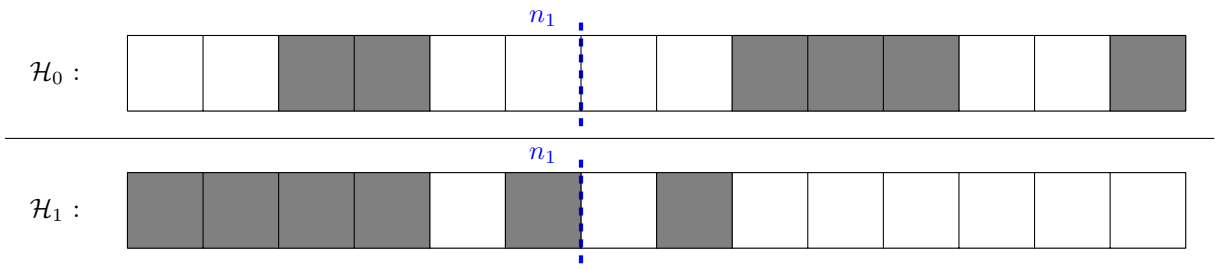

FIGURE 3 – Exemple schématique de deux lignes composées de cases grises et blanches suivant si loi d'apparition est la même sur toute la ligne (ligne du haut) ou si la loi à gauche de la rupture fait apparaître plus de cases grises que la loi à droite de la rupture. La rupture est symbolisée par un trait pointillé bleu.

prenons :

$$S_n(n_1) = \sum_{i=1}^{n} U_{n,i}^2(n_1). \tag{1}$$

2.3 Extension au cas de plusieurs ruptures

Avant d'étendre le cas à plusieurs ruptures, nous commençons par remarquer que nous pouvons réécrire la précédente statistique à l'aide du rang de chaque case :

$$U_{n,i}(n_1) = \frac{2}{\sqrt{nn_1(n-n_1)}} \sum_{j_0=1}^{n_1} \left(\frac{n+1}{2} - R_{j_0}^{(i)} \right) = \frac{2}{\sqrt{nn_1(n-n_1)}} \sum_{j_1=n_1+1}^{n} \left(R_{j_1}^{(i)} - \frac{n+1}{2} \right)$$

où $R_j^{(i)}$ est le rang de la $j^{\text{ème}}$ case au sein de la $i^{\text{ème}}$ ligne :

$$R_j^{(i)} = \sum_{k=1}^{n} \mathbb{1}_{\{X_{i,k} \leq X_{i,j}\}}.$$

Remarquons que la valeur $(n+1)/2$ représente l'espérance de $R_j^{(i)}$ si toutes les cases sont simulées de façon indépendante par la même loi ; la statistique $U_{n,i}$ mesure donc l'écart des cases situées avant la rupture ou après la rupture par rapport à l'espérance que nous pouvons espérer obtenir si les lois sont identiques de chaque côté.

Étant données L ruptures $0 = n_0 < n_1 < \cdots < n_L < n_{L+1} = n$, nous généralisons la statistique $S_n(n_1, \ldots, n_L)$ de la façon suivante :

$$S_n(n_1, \ldots, n_L) = \frac{4}{n^2} \sum_{\ell=0}^{L} (n_{\ell+1} - n_\ell) \sum_{i=1}^{n} \left(\overline{R}_\ell^{(i)} - \frac{n+1}{2} \right)^2 \tag{2}$$

où

$$\overline{R}_\ell^{(i)} = \frac{1}{n_{\ell+1} - n_\ell} \sum_{j=n_\ell+1}^{n_{\ell+1}} R_j^{(i)}$$

est la moyenne des rangs des cases situées entre les ruptures n_ℓ et $n_{\ell+1}$ pour la $i^{\text{ème}}$ ligne. À nouveau, nous remarquons que la statistique sera d'autant plus éloignées de 0 que les moyennes des rangs ne seront pas concentrées près de l'espérance.

Un développement de la statistique (2) pour le cas à une rupture permet de se convaincre que nous retrouvons la statistique de l'équation (1) (voir par exemple la remarque 1, page 145 de Brault et al. (2018b)).

2.4 Estimation des emplacements des ruptures

Bien que nous ayons basé la modélisation sur les tests afin de mieux comprendre l'origine de la formule de S_n, nous nous intéressons surtout à l'emplacement des ruptures $\mathbf{n} = (n_1, \ldots, n_L)$. Comme nous l'avons expliqué jusqu'à présent et dans le cas où $\mathbf{n}$ est une vraie rupture (c'est-à-dire que, pour chaque rupture n_ℓ, il existe au moins une ligne avec deux lois différentes juste avant et juste après la rupture), nous nous attendons à ce que $S_n(\mathbf{n})$ soit éloignée de 0. Pour estimer les ruptures, nous pouvons donc chercher celles qui maximisent la statistique :

$$\widehat{\boldsymbol{n}} \in \underset{0=n_0<n_1<\cdots<n_L<n_{L+1}=n}{\operatorname{argmax}} S_n(n_1, \ldots, n_L).$$

Dans la suite, nous allons nous intéresser principalement aux propriétés de cet estimateur et nous montrons que, sous des hypothèses basiques, cet estimateur a de bonnes propriétés.

3 Résultats théoriques

Dans cette partie, nous explicitons quelques résultats théoriques intéressants que nous pouvons avoir sur ces statistiques. Nous ne donnons ici que les grandes lignes des démonstrations ; ces dernières peuvent être retrouvées dans l'appendice et dans le *supplementary material* de l'article de Brault et al. (2018b).

3.1 Simplification du problème

Pour formaliser la modélisation, nous présentons jusqu'à présent chaque colonne $(X_{1,j}, \ldots, X_{i,j}, \ldots, X_{n,j})^T$ (où A^T représente la transposée de la matrice A) comme un vecteur de $\mathbb{R}^n$ afin de nous appuyer sur la théorie proposée en dimension $\mathbb{R}^N$ où N est fixé par Lung-Yut-Fong et al. (2011). Or, dans le cas étudié ici, les colonnes font partie d'une matrice symétrique ce qui implique que, pour tout couple $(i, j) \in \{1, \ldots, n\}^2$, les variables $X_{i,j}$ et $X_{j,i}$ ont la même loi.

Dans le cas d'une seule rupture, cela signifie qu'il n'y a qu'au plus 3 lois (voir la droite de la figure 2 pour une représentation schématique) :

- $\mathbb{P}_{1,1}$ pour les cases $(X_{i,j})_{1\leq i\leq j\leq n_1}$ donc situées dans la partie supérieure droite du bloc carré situé en haut à gauche de la matrice (puisque chaque case de la partie inférieure est égale à la case symétrique de la partie supérieure) ;
- $\mathbb{P}_{2,2}$ pour les cases $(X_{i,j})_{n_1+1\leq i\leq j\leq n}$ donc situées dans la partie supérieure droite du bloc carré situé en bas à droite de la matrice ;
- $\mathbb{P}_{1,2}$ pour les cases $(X_{i,j})_{1\leq i\leq n_1, n_1+1\leq j\leq n}$ donc situées dans le bloc en haut à droite de la matrice (et, par symétrie, c'est la même loi pour les cases situées en bas à gauche).

Dans ce cadre, le test à étudier revient simplement à comparer l'hypothèse :

$$\mathcal{H}_0 \ : \mathbb{P}_{1,1} \overset{\mathcal{L}}{\sim} \mathbb{P}_{1,2} \text{ et } \mathbb{P}_{2,2} \overset{\mathcal{L}}{\sim} \mathbb{P}_{1,2}$$

contre l'hypothèse que n_1 en soit bien une :

$$\mathcal{H}_1 \ : \mathbb{P}_{1,1} \overset{\mathcal{L}}{\not\sim} \mathbb{P}_{1,2} \text{ ou } \mathbb{P}_{2,2} \overset{\mathcal{L}}{\not\sim} \mathbb{P}_{1,2}.$$

De même, dans le cas où il y a L ruptures, il n'y a qu'au plus $(L+1)(L+2)/2$ lois différentes.

3.2 Consistance de l'estimateur $\widehat{n}$

En statistique, un estimateur $\widehat{\theta}_n$ d'un paramètre $\theta^\star$ appartenant à un ouvert Θ est une fonction des observations indépendantes du paramètre $\theta^\star$. En théorie n'importe quelle fonction des observations peut être un estimateur : par exemple, la fonction constante égale à $\widehat{\theta}_n = 0$ est une estimation ; toutefois, cette estimation est absurde puisque le paramètre d'intérêt $\theta^\star$ n'a pas de raisons de valoir 0.

La première propriété que nous demandons à un estimateur est sa consistance ; c'est-à-dire que nous souhaitons que l'estimateur ne soit pas trop loin du paramètre d'intérêt lorsque le nombre d'observations est suffisamment grand. Nous distinguons deux types de consistance :

- la consistance (simple) impliquant la convergence en probabilité de l'estimateur $\widehat{\theta}_n$ vers le paramètre d'intérêt $\theta^\star$:
$$\forall \delta > 0,\, \mathbb{P}_{\theta^\star}\left(\left\|\widehat{\theta}_n - \theta^\star\right\| > \delta\right) \underset{n\to+\infty}{\longrightarrow} 0$$
où la norme $\|\cdot\|$ sera dans la suite de ce travail la norme infinie $\|\cdot\|_{+\infty}$ définie pour tout $\mathbf{y} \in \mathbb{R}^d$ par :
$$\|\mathbf{y}\|_{+\infty} = \max_{1\leq j\leq p} |y_j|.$$
Dans le cas où Θ est inclus dans un sous-espace de dimension finie, le choix de la norme n'a pas d'importance puisque toutes les normes sont équivalentes.

- la consitance forte impliquant la convergence forte de l'estimateur $\widehat{\theta}_n$ vers le paramètre d'intérêt $\theta^\star$:
$$\mathbb{P}_{\theta^\star}\left(\lim_{n\to+\infty} \widehat{\theta}_n = \theta^\star\right) = 1.$$
Cette propriété est plus forte car la convergence forte implique la convergence en probabilité.

Par exemple, si $Y_1, \ldots, Y_n$ sont des variables indépendantes de paramètre $\theta^\star$ et de même loi admettant un moment d'ordre 1 alors la moyenne $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ est un estimateur fortement consistant de l'espérance $\mathbb{E}_{\theta^\star}[Y_1]$ d'après le théorème de la limite centrale.

Sur la figure 4, nous présentons schématiquement le principe de consistance : l'objectif est de fixer une boule de centre $\theta^\star$ et de rayon $\delta > 0$ et de regarder la probabilité que la suite d'estimateur $\left(\widehat{\theta}_n\right)_{n\in\mathbb{N}}$ soit incluse dans cette boule à partir d'un certain rang.

Pour l'estimation des ruptures, les paramètres étudiés $\boldsymbol{n}^\star = (n_1^\star, \ldots, n_L)$ augmentent avec le nombre d'observation n. Pour contourner ce problème, nous supposons qu'il existe $0 = \tau_0^\star < \tau_1^\star < \cdots < \tau_L^\star < \tau_{L+1}^\star = 1$ tels que pour tout $\ell \in \{0, \ldots, L+1\}$:
$$\frac{n_\ell^\star}{n} \underset{n\to+\infty}{\longrightarrow} \tau_\ell^\star.$$
Nous nous intéressons donc à l'estimation $\widehat{\boldsymbol{\tau}} = (\widehat{\tau}_1, \ldots, \widehat{\tau}_L)$ définie pour tout $\ell \in \{1, \ldots, L\}$ par $\widehat{\tau}_\ell = \widehat{n}_\ell / n$.

Afin de faciliter la lecture, nous présentons le principe dans le cas d'une seule rupture $n_1^\star$ (une démonstration complète est disponible dans le *supplementary material* de l'article de Brault et al. (2018b)). Pour montrer la consistance de $\widehat{n_1}$, nous remarquons que $\widehat{n_1}$ est différent de $n_1^\star$ si et seulement s'il existe $n_1 \in \{1, \ldots, n-1\}$ tel que $S_n(n_1) > S_n(n_1^\star)$ et, étant donné $n_1 \neq n_1^\star$ nous nous intéressons donc à la probabilité suivante :
$$\mathbb{P}\left(S_n(n_1) - S_n(n_1^\star) > 0\right) = \mathbb{P}\left(S_n(n_1) - S_n(n_1^\star) - \mathbb{E}\left[S_n(n_1) - S_n(n_1^\star)\right] > -\mathbb{E}\left[S_n(n_1) - S_n(n_1^\star)\right]\right).$$
L'intérêt de la deuxième forme est que nous étudions la probabilité qu'une variable centrée soit plus grande qu'une valeur déterministe. La probabilité tendra vers 0 si la partie déterministe tend vers $+\infty$ assez vite pour que la variabilité de la variable ne soit pas trop forte.

Pour montrer la consistance de l'estimateur, nous avons besoin de faire trois hypothèses que nous présentons dans le cas d'une seule rupture :

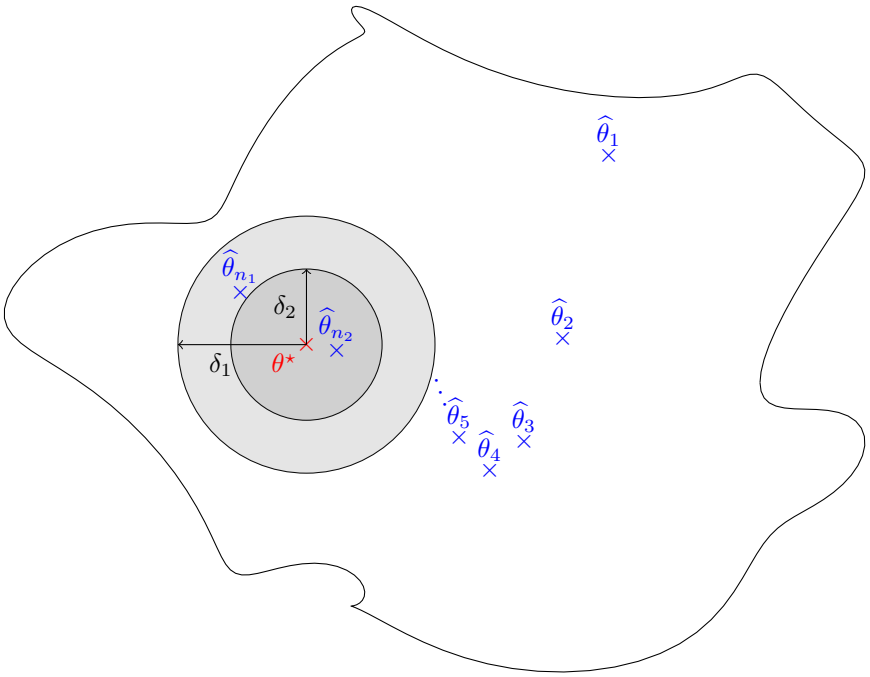

FIGURE 4 – Représentation schématique de la consistance d'un estimateur $\widehat{\theta}_n$ du paramètre $\theta^\star$: l'objectif est de vérifier que, pour chaque boule centrée en $\theta^\star$ de rayon δ, la probabilité que l'estimateur $\widehat{\theta}_n$ soit à l'intérieur tendent vers 1.

(H1) les lois étudiées sont continues ; c'est-à-dire que la probabilité de valoir un singleton est nulle.

(H2) il existe $0 = \tau_0^\star < \tau_1^\star < \cdots < \tau_L^\star < \tau_{L+1}^\star = 1$ tels que pour tout $\ell \in \{0, \ldots, L+1\}$:

$$\frac{n_\ell^\star}{n} \underset{n\to+\infty}{\longrightarrow} \tau_\ell^\star.$$

(H3) si X suit la loi $\mathbb{P}_{1,1}$ (ou $\mathbb{P}_{2,2}$) et Y suit la loi $\mathbb{P}_{1,2}$ alors :

$$\mathbb{E}\left[F_X(Y)\right] \neq \mathbb{E}\left[F_Y(X)\right]$$

où F_X et F_Y sont les fonctions de répartition des variables X et Y respectivement.

Ces conditions sont vraiment minimalistes et sont vérifiées dans la plupart des cas.

De plus, nous notons pour $k \in \{1,2\}$, $\beta^{(k)} = \mathbb{E}\left[h\left(X_k, Y_k\right)\right]$ avec X_k (resp. Y_k) suivant la loi $\mathbb{P}_{k,1}$ (resp. $\mathbb{P}_{k,2}$).

Proposition 3.1 *Dans le cas où les lois étudiées sont continues, c'est-à-dire sous l'hypothèse (H1), nous montrons que :*

$$-\mathbb{E}\left[S_n(n_1) - S_n(n_1^\star)\right] = |n_1 - n_1^\star| \left[\left[\beta^{(1)}\right]^2 n_1^\star + \left[\beta^{(2)}\right]^2 (n - n_1^\star)\right] \underbrace{\left\{\begin{array}{ll} \frac{n-n_1^\star}{n-n_1} & \text{if } n_1 < n_1^\star \\ 1 & \text{if } n_1 = n_1^\star \\ \frac{n_1^\star}{n_1} & \text{if } n_1 > n_1^\star \end{array}\right\}}_{\geq \min\left(\tau_1^\star, 1-\tau_1^\star\right) \text{ sous (H2)}} (1 + o(1)).$$

Les parties en rouge correspondent à celles influencées par l'hypothèse $(H2)$: à l'intérieur du crochet, les valeurs $n_1^\star$ et $n - n_1^\star$ vont être plus grandes que $n \min(\tau_1^\star, 1 - \tau_1^\star)$ donc tendre linéairement vers $+\infty$.

À l'aide de l'hypothèse $(H3)$, nous montrons qu'au moins l'un des $\beta^{(k)}$ est différent de 0 et l'une des parties en bleu est donc non nulle.

Enfin, dans le cas de la consistance et étant donné $\delta > 0$, nous calculons la probabilité que $|\widehat{\tau}_1 - \tau_1^\star|$ soit plus grand que δ et, donc, $|n_1 - n_1^\star|$ est supposé plus grand que $n\delta$ (partie violette de l'équation).

Théorème 3.1 *Sous les trois hypothèses (H1), (H2) et (H3), nous montrons donc que :*

$$\mathbb{P}\left(|\widehat{\tau}_1 - \tau_1^\star| > \delta\right) \underset{n \to +\infty}{\longrightarrow} 0.$$

3.3 Algorithme de programmation dynamique

Estimer les ruptures $\boldsymbol{n}^\star = (n_1^\star, \ldots, n_L^\star)$ consiste à trouver le L-uplet de l'ensemble $\{1, \ldots, n-1\}$ maximisant la statistique $S_n(n_1, \ldots, n_L)$. Or, cela représente $\binom{n-1}{L}$ possibilités soit environ $2,668 \times 10^{35}$ pour $L = 20$ ruptures d'une matrice à $n = 500$ lignes ; en traitant une possibilité par opération, il faudrait à peu près 37 fois l'âge de l'univers au supercalculateur Sequoia d'IBM pour traiter toutes les configurations.

Pour contourner ce problème, nous utilisons la programmation dynamique dont le principe, introduit par Richard (1957), consiste à découper la maximisation en des opérations récurrentes et plus faciles à faire. Ceci est possible dès que nous pouvons découper la statistique à maximiser de la manière suivante :

$$S_n(n_1, \ldots, n_L) = \frac{4}{n^2} \sum_{\ell=0}^{L} \underbrace{(n_{\ell+1} - n_\ell) \sum_{i=1}^{n} \left(\overline{R}_\ell^{(i)} - \frac{n+1}{2}\right)^2}_{=:\Delta(n_\ell : n_{\ell+1})},$$

où les $\Delta(n_\ell : n_{\ell+1})$ ne dépendent que de leurs bornes. Le but est donc de trouver les ruptures $(n_1, \ldots, n_L)$ telles que la somme des $\Delta(n_\ell : n_{\ell+1})$ soit maximale (voir la figure 5 (a)).

Pour ce faire, nous introduisons pour tout $L' \in \{0, \ldots, L\}$ et tout $p \in \{L'+1, \ldots, n\}$ la fonction $I_{L'}(p)$ représentant la valeur maximale de la statistique S_p s'il n'y avait que L' ruptures et p observations (voir la figure 5 (b)) :

$$I_{L'}(p) = \max_{1 < n_1 < \cdots < n_{L'} < n_{L'+1} = p} \sum_{\ell=0}^{L'} \Delta(n_\ell : n_{\ell+1}).$$

Le maximum de la statistique S_n pour L ruptures vaut donc $I_L(n)$. Le but de cette décomposition est d'observer que nous avons une relation de récurrence qui apparaît sur L' :

(L'=0) : Le cas $L' = 0$ correspond au cas où nous cherchons la valeur maximale pour 0 rupture. Nous avons donc directement (voir la figure 5 (c)) :

$$I_0(p) \quad = \quad \max_{1 < n_1 = p} \Delta(1 : n_1) = \Delta(1 : p).$$

(L'=1) : Étant fixé $p > 1$, le cas $L' = 1$ peut être résolu linéairement en regardant toutes les ruptures n_1 possibles (voir la figure 5 (d)) :

$$I_1(p) \quad = \quad \max_{1 < n_1 < n_2 = p} \{\Delta(1 : n_1) + \Delta(n_1 : p)\}$$

ce qui fait une complexité totale de $\mathcal{O}(n^2)$.

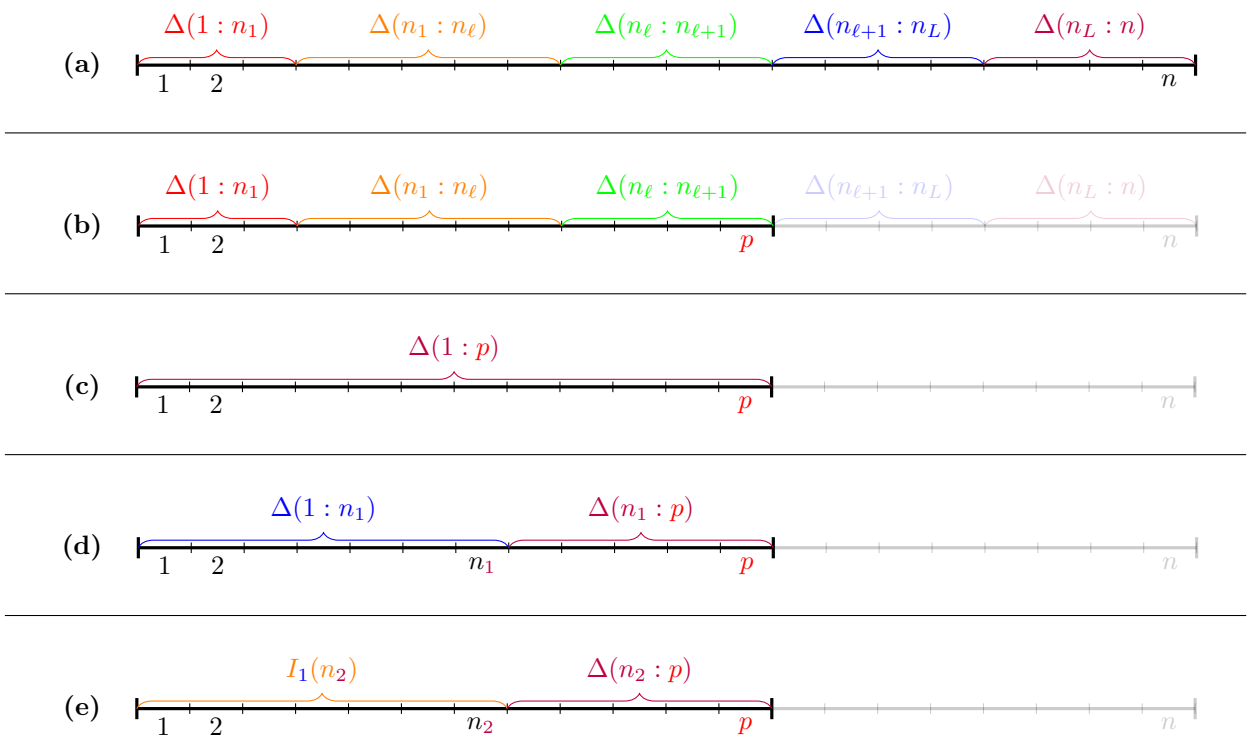

FIGURE 5 – Représentation schématique des maximisations suivant le nombre de ruptures : principe global (a), introduction de la fonction $I_{L'}(p)$ (b), cas avec zéro rupture (c), cas avec une rupture (d) et principe de récurrence avec deux ruptures (e).

(L'=2) : Pour le cas $L' = 2$, nous commençons par remarquer que le maximum peut être fait en deux fois :

$$\begin{aligned} I_2(p) &= \max_{1<n_1<n_2<n_3=p} \{\Delta(1:n_1) + \Delta(n_1:n_2) + \Delta(n_2:p)\} \\ &= \max_{2<n_2<p} \left[\max_{1<n_1<n_2} \{\Delta(1:n_1) + \Delta(n_1+1:n_2)\} + \Delta(n_2:p) \right]. \end{aligned}$$

Donc, si nous connaissons la rupture n_2 optimale, la maximisation consiste simplement à trouver la rupture n_1 optimale. Or, nous connaissons déjà la valeur optimale par l'étape précédente puisqu'elle vaut $I_1(n_2)$ (voir la figure 5 (e)) et la maximisation ne dépend plus que de n_2 :

$$I_2(p) = \max_{2<n_2<p} \left[I_1(n_2) + \Delta(n_2:p) \right].$$

ce qui est fait avec une complexité totale de $\mathcal{O}\left(n^2\right)$ à nouveau.

Proposition 3.2 *Étant donnée la matrice triangulaire supérieure* $(\Delta(n:m))_{1\leq n<m\leq n}$*, le calcul de la matrice* $(I_{L'}(p))_{0\leq L'<p\leq n}$ *se fait de la façon suivante :*

- *Si* $L' = 0$*, nous avons* $I_0(p) = \Delta(1:p)$.

- *Pour tout* $L' \in \{1, \ldots, L\}$ *et pour tout* $p \in \{L'+1, \ldots, n\}$*, nous avons :*

$$I_{L'}(p) = \max_{L' < n_{L'} < p} \left[I_{L'-1}(n_{L'}) + \Delta(n_{L'} : p) \right]$$

À l'aide de cette procédure, nous avons la complexité suivante :

Théorème 3.2 *La complexité pour calculer les ruptures maximisant* S_n *est* $\mathcal{O}\left(n^3\right)$.

Cette complexité est principalement due au calcul de la matrice $(\Delta(n : m))_{1 \leq n < m \leq n}$ qui peut facilement être parallélisé. Une implémentation est disponible dans le package `MuChPoint` de Brault et al. (2018a).

Depuis peu, il existe des algorithmes pour améliorer la complexité de la détection de ruptures proposés notamment par Rigaill (2015). Une extension intéressante de ce travail est de vérifier si nous pouvons utiliser la technique dans ce cas particulier.

4 Application

Le problème de cette procédure est que nous ne possédons pas encore de critère automatique pour choisir le nombre de ruptures. Dans leur article, Brault et al. (2018b) suggèrent d'utiliser l'heuristique de pente développée par Baudry et al. (2012) mais cela n'a pas donné de résultats concluant dans notre cas. Cette question est encore en cours de réflexions.

Pour cet article, nous avons donc choisi le nombre arbitraire de 100 ruptures. Sur la figure 6, nous avons représenté à gauche la matrice originale segmentée et la matrice résumée obtenues à l'aide du package `MuChPoint` de Brault et al. (2018a) : nous retrouvons sur la diagonale les deux blocs rouges du début et de la fin du film correspondants aux moments où la voiture était arrêtée ; nous observons également des blocs rouges sur les lignes et les colonnes de ces blocs qui correspondent aux moments où la voiture est repassée par le point de départ.

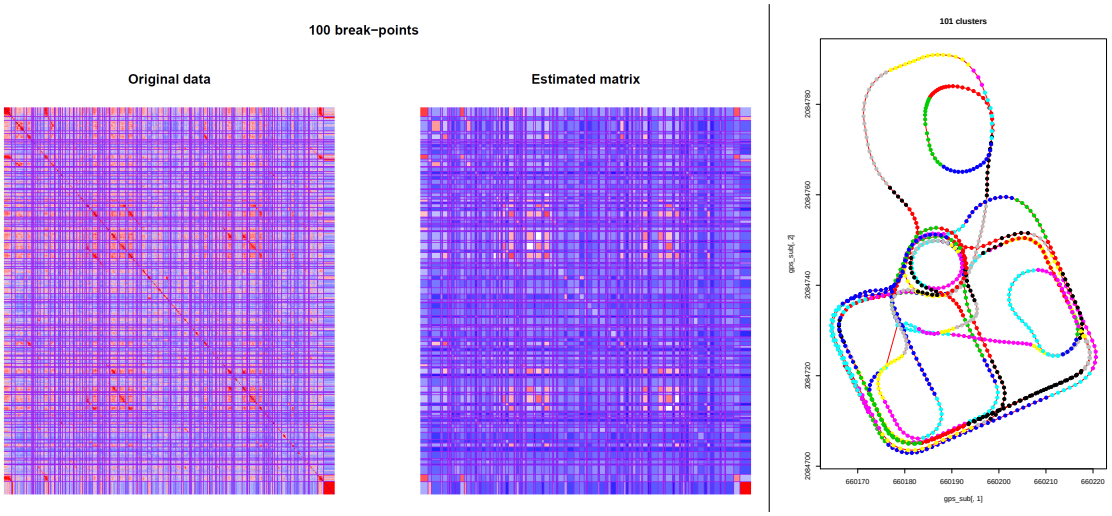

FIGURE 6 – À gauche se trouve la sortie du package `MuChPoint` avec la matrice originale segmentée et la matrice résumée. À droite se trouvent les positions GPS coloriées suivant les groupes formés.

Sur la partie droite de la figure 6, nous avons représenté les coordonnées GPS en coloriant les points par groupe formé (notons qu'à la vue du nombre de groupes, nous avons préféré une palette de 8 couleurs

circulaires plutôt qu'un dégradé afin de mieux voir les ruptures entre deux groupes successifs ; toutefois, deux groupes bleus n'ont pas plus de liens entre eux qu'un groupe bleu avec un groupe rouge). Nous remarquons que la plupart des groupes n'englobent pas de virages. Les grands virages sont également découpés en plusieurs parties.

Pour estimer la cohérence des groupes formés, nous avons cherché à savoir si les blocs avec les valeurs moyennes les plus élevées correspondaient à des coordonnées GPS les plus proches. Pour ce faire, nous calculons la distance de Hausdorff définie pour deux groupes $\mathbf{c}^1 = (c_1^1, \ldots, c_d^1) \in \left(\mathbb{R}^2\right)^d$ et $\mathbf{c}^2 = (c_1^2, \ldots, c_p^2) \in \left(\mathbb{R}^2\right)^p$ par :

$$d_{\mathcal{H}}\left(\mathbf{c}^1, \mathbf{c}^2\right) = \max\left(\max_{1\leq i\leq d}\min_{1\leq j\leq p}\left\|c_i^1 - c_j^2\right\|_2, \max_{1\leq j\leq p}\min_{1\leq i\leq d}\left\|c_i^1 - c_j^2\right\|_2\right)$$

où $\|\cdot\|_2$ est la norme euclidienne. Sur la figure 7, nous avons représenté la distance de Hausdorff calculée entre les coordonnées GPS de chaque croisement de groupes formés en fonction de la valeur moyenne du bloc associé ; la taille des points est proportionnel au nombre total de cases dans le bloc. Nous voyons que pour des valeurs de blocs supérieures à 30, les groupes associés sont géographiquement proches. Nous voyons également que des groupes de coordonnées GPS proches ont pourtant des petites valeurs moyennes des blocs : cela correspond souvent aux moments où le véhicule passe par le même endroit mais dans l'autre sens.

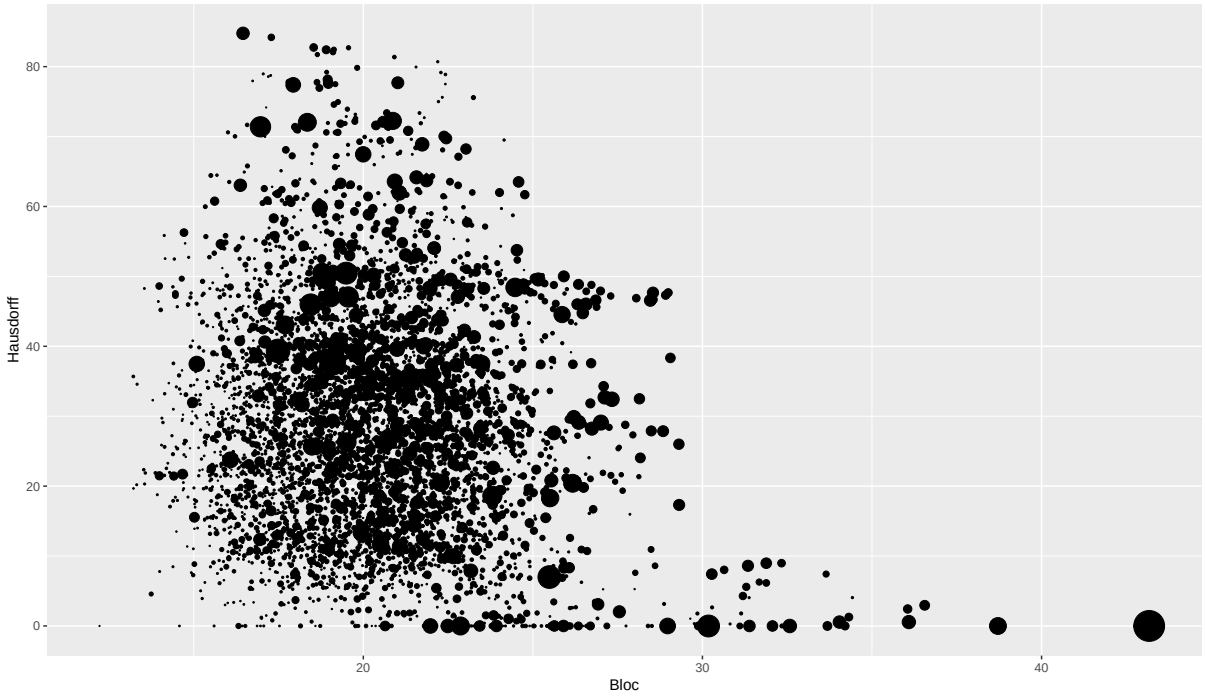

FIGURE 7 – Représentation de la distance de Hausdorff de deux groupes de coordonnées en fonction de la similarité moyenne entre les images associées : la taille de chaque point est proportionnelle aux nombres d'images associées.

5 Perspective

Dans ce travail, nous avons présenté une application d'une méthode utilisée dans l'analyse des données Hi-C et basée sur l'utilisation des statistiques de rang. Nous avons notamment rappelé les résultats théoriques obtenus par Brault et al. (2018b) et présenté l'application. La structure par blocs met bien en évidence les blocs avec des valeurs fortes correspondantes à des coordonnées GPS proches mais il reste des coordonnées GPS proches avec des valeurs faibles ; il serait intéressant de regarder un film où le véhicule fait un même trajet dans le même sens pour comparer.

D'un point de vue théorique, il serait intéressant d'approfondir la problématique de la sélection de modèles pour choisir un nombre de ruptures de façon automatique ; cela se fera notamment en étudiant la loi de la statistique maximisée.

Références

J.-P. Baudry, C. Maugis, et B. Michel. Slope heuristics : overview and implementation. *Statistics and Computing*, 22(2) :455–470, 2012.

M. Birem, J.-C. Quinton, F. Berry, et Y. Mezouar. Sail-map : Loop-closure detection using saliency-based features. Dans *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4543–4548. IEEE, 2014.

V. Brault, G. Cougoulat, S. Ouadah, et L. Sansonnet. Muchpoint : Multiple change point, 2018a. URL `https://CRAN.R-project.org/package=MuChPoint`.

V. Brault, S. Ouadah, L. Sansonnet, et C. Lévy-Leduc. Nonparametric multiple change-point estimation for analyzing large hi-c data matrices. *Journal of Multivariate Analysis*, 165 : 143 – 165, 2018b. ISSN 0047-259X. doi : https ://doi.org/10.1016/j.jmva.2017.12.005. URL `http://www.sciencedirect.com/science/article/pii/S0047259X17307753`.

M. J. Cummins et P. M. Newman. Fab-map : Appearance-based place recognition and mapping using a learned visual vocabulary model. Dans *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 3–10, 2010.

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) :376–380, 2012.

H. Korrapati, J. Courbon, S. Alizon, et F. Marmoiton. " the institut pascal data sets" : un jeu de données en extérieur, multicapteurs et datées avec réalité terrain, données d'étalonnage et outils logiciels. Dans *Orasis, Congrès des jeunes chercheurs en vision par ordinateur*, 2013.

A. Lung-Yut-Fong, C. Lévy-Leduc, et O. Cappé. Homogeneity and change-point detection tests for multivariate data using rank statistics. *arXiv preprint arXiv :1107.1971*, 2011.

B. Richard. Dynamic programming. *Princeton University Press*, 89 :92, 1957.

G. Rigaill. A pruned dynamic programming algorithm to recover the best segmentations with 1 to k_max change-points. *Journal de la Société Française de Statistique*, 156(4) :180–205, 2015.