Estimation robuste de ruptures multiples dans un signal multivarié

Alexandre LUNG-YUT-FONG, Céline LÉVY-LEDUC, Olivier CAPPÉ,

Institut Télécom & CNRS, Télécom ParisTech, LTCI
46, Rue Barrault, 75634 Paris Cédex 13, France
{lung,levyledu,cappe}@telecom-paristech.fr

Résumé – Nous proposons une méthode non-paramétrique de détection de ruptures dans un signal multivarié. La méthode d'estimation que nous proposons découle d'une extension au cas multivarié du test de Kruskal-Wallis permettant de tester l'homogénéité entre les distributions de plusieurs groupes de données. Nous décrivons les propriétés asymptotiques de ce test. Nous proposons également une mise en œuvre rapide de notre méthode utilisant un algorithme de programmation dynamique pour un nombre fixe ou variable de segments. La méthode proposée obtient de très bons résultats, en particulier lorsque la distribution des données est atypique (par exemple en présence de valeurs aberrantes). Nous proposons également une application à des données réelles correspondant à l'étude de profils génomiques.

Abstract – We propose a non-parametric statistical procedure for detecting multiple change-points in multidimensional signals. The method is based on a test statistic that generalizes the well-known Kruskal-Wallis procedure to the multivariate setting. The proposed approach does not require any knowledge about the distribution of the observations and is parameter-free. It is computationally efficient thanks to the use of dynamic programming and can also be applied when the number of change-points is unknown. The method is shown through simulations to be more robust than alternatives, particularly when faced with atypical observations (e.g., with outliers), high noise levels and/or high-dimensional data. We also propose an application to real biological data.

1 Introduction

L'estimation rétrospective de plusieurs ruptures consiste à faire une partition d'une série d'observations en plusieurs segments homogènes de longueurs variables [2]. Cette série d'observations peut par exemple être une série temporelle dans laquelle le niveau du signal peut varier de manière abrupte d'une valeur à une autre à des instants aléatoires appelés ruptures. La segmentation de signaux est une problématique classique qui a des applications dans de nombreux domaines tels que le traitement de la parole, la détection d'intrusions dans les réseaux ou la bioinformatique [1, 11, 15].

Dans le cas paramétrique où la distribution des observations sous-jacente est connue et gaussienne le problème de la segmentation temporelle (ou détection de ruptures) peut être résolu en utilisant le critère des moindres carrés [17] qui peut être optimisé en utilisant l'algorithme de programmation dynamique [6]. Cette approche peut être généralisée au cas de données gaussiennes multivariées [14].

Dans cette contribution, nous considérons des méthodes non-paramétriques *i.e.* qui n'ont pas besoin de connaissance *a priori* sur la distribution sous-jacente des observations et qui peuvent s'appliquer à des signaux multivariés. Il y a, à notre connaissance, peu de méthodes s'attaquant à cette classe de problèmes. On peut par exemple citer l'approche de [4] qui utilise une métrique à base de noyaux pour évaluer les distances entre des observations multivariées. C'est une méthode efficace, mais qui requiert le choix d'un noyau approprié et qui est peu robuste lorsque les données sont bruitées. On peut également citer la méthode de [12] qui permet de détecter la présence d'un changement en étendant le test d'homogénéité de Wilcoxon/Mann-Whitney au cas multivarié.

La méthode que nous proposons, appelée *dynMKW* dans la suite, étend la statistique de [12] à la segmentation multiple, de la même manière que le test de Kruskal-Wallis [7] généralise le test de Mann-Whitney/Wilcoxon à un test d'homogénéité entre plusieurs groupes. L'algorithme de programmation dynamique est ensuite utilisé pour estimer efficacement les instants de rupture.

Nous présentons dans la section 2 une méthode permettant de tester l'homogénéité entre plusieurs groupes d'observations multivariées. La procédure d'estimation de plusieurs ruptures qui découle du test d'homogénéité est ensuite décrite à la section 3, ainsi qu'une méthode permettant de déterminer le nombre de changements dans la section 4. Nous terminons par l'évaluation des méthodes présentées sur des données simulées et réelles en section 5.

2 Test d'homogénéité entre plusieurs groupes de données multivariées

Soient $\mathbf{X}_1, \ldots, \mathbf{X}_n$ n vecteurs aléatoires indépendants de dimension L tels que $\mathbf{X}_j = (X_{j,1}, \ldots, X_{j,L})'$ (où l'on note A' la transposée de la matrice A). On s'intéresse au test de l'hypothèse (H_0) selon laquelle les K groupes, qui ont pour tailles respectives $n_{i+1} - n_i$, $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}; \mathbf{X}_{n_1+1}, \ldots, \mathbf{X}_{n_2}; \ldots; \mathbf{X}_{n_{K-1}+1}, \ldots, \mathbf{X}_{n_K}$ ont la même distribution. On utilisera

dans la suite la convention $n_0 = 0$ et $n_K = n$.

Pour $j \in \{1, \ldots, n\}$ et $\ell \in \{1, \ldots, L\}$, notons $R_j^{(\ell)}$ le rang de $X_{j,\ell}$ parmi $(X_{1,\ell}, \ldots, X_{n,\ell})$, *i.e.* $R_j^{(\ell)} = \sum_{k=1}^{n} \mathbf{1}_{\{X_{k,\ell} \leq X_{j,\ell}\}}$. On définit aussi pour $k \in \{0, \ldots, K-1\}$ le rang moyen du k-ème groupe

$$\bar{R}_k^{(\ell)} = (n_{k+1} - n_k)^{-1} \sum_{j=n_k+1}^{n_{k+1}} R_j^{(\ell)} \ .$$

La statistique de test que nous proposons s'écrit

$$T(n_1, \ldots, n_{K-1}) = \frac{1}{n^2} \sum_{k=0}^{K-1} (n_{k+1} - n_k) \bar{\mathbf{R}}_k' \ \hat{\Sigma}_n^{-1} \ \bar{\mathbf{R}}_k \ , \quad (1)$$

où $\bar{\mathbf{R}}_k = (\bar{R}_k^{(1)} - (n+1)/2, \ldots, \bar{R}_k^{(L)} - (n+1)/2)'$ et $\hat{\Sigma}_n$ est la matrice de taille $L \times L$ dont l'élément (ℓ, ℓ') s'écrit

$$\hat{\Sigma}_{n,\ell\ell'} = \frac{1}{n} \sum_{i=1}^{n} \{\hat{F}_{n,\ell}(X_{i,\ell}) - 1/2\}\{\hat{F}_{n,\ell'}(X_{i,\ell'}) - 1/2\} \ ,$$

où $\hat{F}_{n,\ell}(t) = n^{-1} \sum_{j=1}^{n} \mathbf{1}_{\{X_{j,\ell} \leq t\}}$ est la fonction de répartition empirique de la ℓ-ième coordonnée $X_{1,\ell}$.

À noter que (1) peut être vue comme une extension au cas multivarié du test de Kruskal-Wallis [7]. En effet, lorsque $L = 1$, (1) s'écrit

$$T(n_1, \ldots, n_{K-1}) = \frac{12}{n^2} \sum_{k=0}^{K-1} (n_{k+1} - n_k) \left(\bar{R}_k^{(1)} - n/2 \right)^2 \ , \quad (2)$$

équation dans laquelle $\hat{\Sigma}_{n,11}$ s'écrit $\Sigma_{11} = \mathrm{Var}(F_1(X_{1,1}))$, qui est égale à la variance d'une loi uniforme, c'est à dire 1/12, F_1 étant une fonction de répartition continue. Dans le cas particulier où la partition des données se fait en $K = 2$ groupes (c'est à dire avec un seul changement), (1) n'est autre que la statistique de test proposée dans [12] et qui étend le test classique de Mann-Whitney/Wilcoxon au cas multivarié.

On peut montrer le résultat suivant, qui donne sous l'hypothèse nulle la convergence de la statistique (1), lorsque n tend vers l'infini, vers une distribution limite ne dépendant que de K et de L.

Théorème 1 *Soit $(\mathbf{X}_i)_{1 \leq i \leq n}$ des vecteurs aléatoires i.i.d. à valeurs dans $\mathbb{R}^L$ tels que, pour tout ℓ, la fonction de répartition F_ℓ de $X_{1,\ell}$ est continue. On suppose que pour tout $k = 0, \ldots, K-1$, il existe λ_{k+1} dans $(0,1)$ tel que $(n_{k+1} - n_k)/n \to \lambda_{k+1}$, lorsque n tend vers l'infini. Alors, $T(n_1, \ldots, n_{K-1})$ défini dans* (1) *satisfait*

$$T(n_1, \ldots, n_{K-1}) \xrightarrow{d} \chi^2\left((K-1)L\right) \ , \ \textit{quand } n \to \infty \ , \quad (3)$$

où d désigne la convergence en loi et $\chi^2((K-1)L)$ est la loi du χ^2 à $(K-1)L$ degrés de liberté.

3 Détection de ruptures multiples

La décomposition rétrospective d'une fenêtre d'observations en K (où l'on considère dans un premier temps que ce nombre K est connu) segments homogènes et l'estimation de la position des frontières entre ces segments peut être obtenue en maximisant la statistique précédente (1) :

$$(\hat{n}_1, \ldots, \hat{n}_{K-1}) = \underset{1 \leq n_1 < \cdots < n_{K-1} \leq n-1}{\mathrm{argmax}} T(n_1, \ldots, n_{K-1}) \ . \quad (4)$$

Le calcul de la statistique (1) pour toutes les valeurs possibles des frontières est un problème combinatoire difficile pour de grandes valeurs de K. On peut cependant résoudre ce problème de manière efficace grâce à un algorithme de programmation dynamique. En effet, avec les notations

$$\Delta(n_k + 1 : n_{k+1}) = (n_{k+1} - n_k) \bar{R}_k' \ \hat{\Sigma}_n^{-1} \ \bar{R}_k \ ,$$

et

$$I_K(p) = \max_{1 < n_1 < \cdots < n_{K-1} < n_K = p} \sum_{k=0}^{K-1} \Delta(n_k + 1 : n_{k+1}) \ ,$$

on peut écrire la relation de récurrence suivante :

$$I_K(p) = \max_{n_{K-1}} \{I_{K-1}(n_{K-1}) + \Delta(n_{K-1} + 1 : p)\} \ . \quad (5)$$

Le calcul préalable des $\Delta(i : j)$ pour tous les $1 \leq i < j \leq n$ (il est à noter que le calcul de $\hat{\Sigma}_n^{-1}$ n'est réalisé qu'une seule fois) ainsi que la résolution de cette récurrence mènent à une méthode de complexité proportionnelle à $K \times n^2$ pour la détermination des instants de changement.

4 Sélection du nombre de ruptures

L'estimation du nombre de ruptures est un problème généralement difficile. L'ajout d'une pénalité à un critère d'attache aux données est l'une des méthodes couramment utilisées [9, 10, 8]. On peut aussi citer l'utilisation d'a priori sur la position des ruptures lorsqu'un point de vue bayésien est adopté [13, 3] ou encore d'une pénalité associée à la norme ℓ_1 [5] ou ℓ_1 par blocs [15].

Nous proposons une méthode basée sur une heuristique de pente pour résoudre ce problème, variante d'idées utilisées par exemple dans [8] et souvent préférables aux critères utilisant des pénalités de type AIC ou BIC. La méthode proposée est basée sur le principe qu'en présence de $S^\star \geq 1$ ruptures (pour $K = S^\star + 1$ segments), si l'on trace $I_K(n)$ en fonction de S, pour $S = 0, \ldots, S_{\text{max}}$, le graphe qui en résulte peut se décomposer en deux parties linéaires : une première, pour $S = 0, \ldots, S^\star$ pour laquelle le critère augmente rapidement ; et une deuxième, $S = S^\star, \ldots, S_{\text{max}}$ ou la statistique subit une croissance beaucoup plus faible. Pour chaque valeur de S, $S = 1, \ldots, S_{\text{max}}$, on calcule donc une régression linéaire par la méthode des moindres carrés pour la partie avant et après S ; le nombre estimé de ruptures est la valeur de S pour laquelle la somme des carrés des résidus calculés sur chacune des parties est minimale (voir la Figure 1).

Le cas $S = 0$ est traité séparément, la procédure que nous venons de décrire n'est appliquée que lorsque $T(\hat{n}_1)$ a une valeur significative, au sens où la p-valeur associée au test d'une

unique rupture, obtenue selon la méthode décrite dans [12], est assez petite.

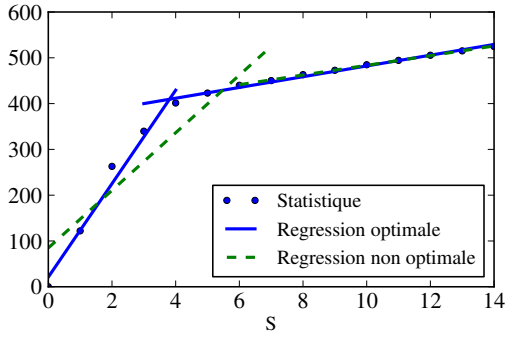

FIG. 1 – Estimation du nombre de ruptures. Dans le cas illustré dans cette figure, le vrai nombre de ruptures est $S^\star = 4$ (régression correspondante représentée en trait plein) ; un mauvais modèle, ici $S = 6$ est également représenté en pointillés

5 Évaluation

5.1 Données simulées

5.1.1 Nombre de ruptures connu

Pour évaluer la méthode proposée, nous avons dans un premier temps simulé un signal de dimension 5 et de longueur 500 qui contient 4 ruptures qui ne se produisent que sur un sous-ensemble des coordonnées, ce qui est caractéristique de nombreuses applications. On ajoute à ce signal un bruit gaussien corrélé de variance marginale σ^2 (voir ainsi la Figure 2 pour un exemple avec un rapport signal à bruit de 16 dB).

L'algorithme proposé (*dynMKW*) est comparé à une méthode à noyau (*Kernel*) s'appuyant sur un noyau gaussien isotrope [4] ainsi qu'à une méthode paramétrique (*Linéaire*) s'appuyant sur le test du maximum de vraisemblance, prenant en compte que les données sont gaussiennes. Les algorithmes sont comparés sur 1000 réplications de Monte-Carlo avec diverses valeurs de σ ainsi qu'avec et sans valeurs aberrantes.

Si les performances des trois algorithmes sont similaires – avec un léger avantage pour l'algorithme paramétrique– lorsque les données ne contiennent pas de valeurs aberrantes, en présence de valeurs aberrantes (5% du signal distribué selon une loi normale multidimensionnelle, aux coordonnées indépendantes et de variance marginale 10 dB plus élevée que la variance marginale du bruit σ^2), *dynMKW* montre sa robustesse, puisque ses performances sont à peine affectées, ce qui n'est pas le cas pour les deux autres méthodes, voir Figure 3.

5.1.2 Nombre de ruptures inconnu

Avec les mêmes signaux que précédemment, on suppose maintenant que le nombre de ruptures est inconnu. On évalue ainsi

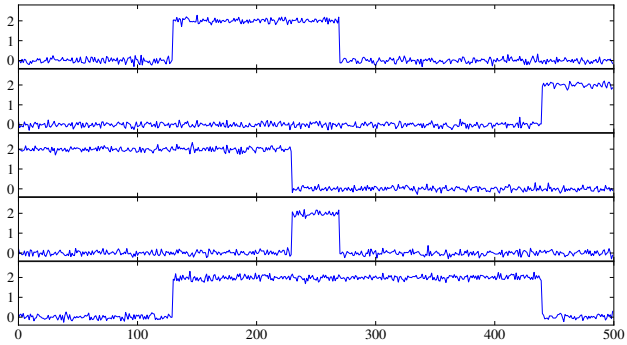

FIG. 2 – Signal déterministe et bruit additif gaussien pour un SNR de 16 dB.

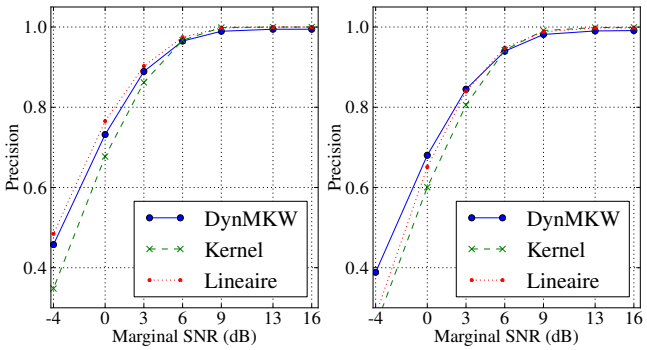

FIG. 3 – Courbes de précision en fonction du rapport signal à bruit marginal du signal pour un signal avec bruit gaussien (droite) et un signal avec bruit gaussien et 5% de valeurs aberrantes (gauche).

les performances de la statistique (1) combinée à l'heuristique présentée à la section 4. Cette méthode d'estimation du nombre de ruptures est comparée avec la méthode de segmentation binaire suggérée par [16] qui consiste à tester de manière récursive la présence d'un changement dans les sous-segments estimés à l'itération précédente. Cette méthode ("*Vost*") récursive est appliquée en utilisant (4) restreinte à la détection d'un seul changement. Une p-valeur asymptotique calculée de la manière décrite dans [12] permet de déterminer si un segment donné doit encore être partitionné (on arrête le processus lorsque la p-valeur est plus grande qu'un certain seuil, par exemple 1% ou 5%).

Les résultats sont ainsi montrés dans la Figure 4 sous forme de courbes de précision et de rappel et mettent en évidence le fait que la méthode d'estimation du nombre de ruptures proposée donne de meilleurs résultats que la méthode récursive. De façon plus qualitative, la méthode *Vost* a tendance à sursegmenter le signal, alors que la méthode que l'on propose, lorsqu'elle produit un résultat erroné, a plutôt tendance à manquer quelques ruptures.

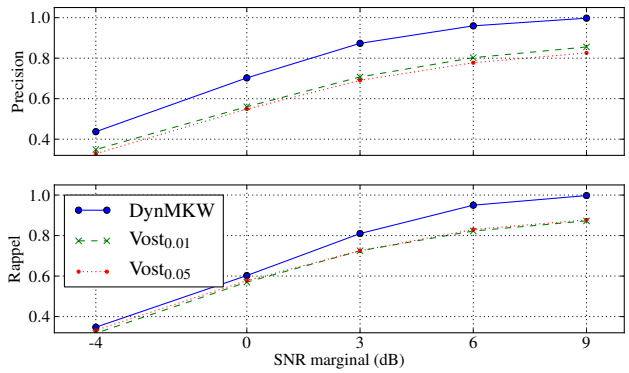

FIG. 4 – Courbes de précision et de rappel pour la méthode utilisant l'heuristique de pente et la segmentation binaire (Vost.) utilisée avec des seuils de 0.01 et 0.05

5.2 Données réelles

Par ailleurs, nous avons testé *dynMKW* en situation réelle, sur les données publiques du nombres de copies d'ADN de patients souffrant du cancer de la vessie[1]. La tâche effectuée consiste à segmenter des données de dimension 9 à 57 (nombre de patients à différents stades du cancer) et de longueur 50 à 200 (ce qui correspond au nombre de sondes placées sur chacun des chromosomes testés) afin de détecter des régions homogènes de délétions ou de réplications d'ADN, ce qui pourrait être une caractéristique du cancer. Les résultats de cette segmentation par dynMKW sur le chromosome 10 de 9 patients sont présentés sur la Figure 5. Nous mettons en évidence le fait que plusieurs coordonnées présentent une rupture (première et seconde rupture) aux mêmes indices, illustrant la pertinence du modèle de segmentation jointe pour ces données. En revanche, on peut remarquer que certaines ruptures ne sont présentes que sur un nombre réduit de coordonnées ; la méthode proposée permet de détecter de tels changements.

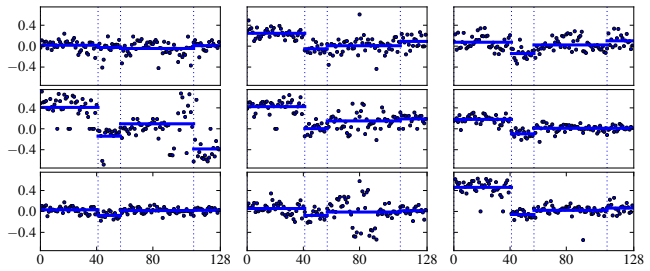

FIG. 5 – Nombre de copies d'ADN dans le chromosome 10 de 9 individus au stade T1 du cancer de la vessie, et superposition de la segmentation obtenue grâce à *dynMKW* ; les lignes verticales désignent les frontières estimées entre deux segments.

[1] http://cbio.ensmp.fr/~frapaport/CGHfusedSVM/index.html

Références

[1] M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes : Theory and Applications*. Prentice-Hall, 1993.

[2] B. E. Brodsky and B. S. Darkhovsky. *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publisher, 1993.

[3] P. Fearnhead. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16 :203–213, 2006.

[4] Z. Harchaoui and O. Cappé. Retrospective multiple change-point estimation with kernels. In *IEEE Workshop on Statistical Signal Processing*, 2007.

[5] Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.*, 105(492) :1480–1493, 2010.

[6] S. Kay. *Fundamentals of statistical signal processing : detection theory*. Prentice-Hall, Inc., 1993.

[7] W. Kruskal and W. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260) :583–621, 1952.

[8] M. Lavielle. Using penalized contrasts for the change-points problems. *Signal Process.*, 85(8) :1501–1510, 2005.

[9] M. Lavielle and E. Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, 21(1) :33–59, 2000.

[10] E. Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Process.*, 85 :717–736, 2005.

[11] C. Lévy-Leduc and F. Roueff. Detection and localization of change-points in high-dimensional network traffic data. *Annals of Applied Statistics*, 3(2) :637–662, 2009.

[12] A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé. Robust changepoint detection based on multivariate rank statistics. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2011.

[13] J. Ruanaidh and W. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer, 1996.

[14] M. S. Srivastava and K. J. Worsley. Likelihood ratio tests for a change in the multivariate normal mean. *J. Amer. Statist. Assoc.*, 81(393) :199–204, 1986.

[15] J. Vert and K. Bleakley. Fast detection of multiple change-points shared by many signals using group LARS. In *Advances in Neural Information Processing Systems 23*, 2010.

[16] L. Y. Vostrikova. Detecting disorder in multidimensional random processes. *Soviet Math. Dokl.*, 24 :55–59, 1981.

[17] Y. Yao and S. T. Au. Least-squares estimation of a step function. *Sankhya : The Indian Journal of Statistics, Series A,*, 51(3) :370–381, 1989.