Détection de Ruptures dans les Signaux Longs

Gérard Biau [1], Kevin Bleakley [2] & David M. Mason [3]

[1] *Sorbonne Universités, UPMC, France & Institut Universitaire de France*
[2] *INRIA Saclay, France & Département de Mathématiques d'Orsay, France*
[3] *University of Delaware, USA*

Résumé. La détection de ruptures dans une suite de données ordonnées dans le temps ou dans l'espace est un problème important dans de nombreux domaines tels que la génétique et la finance. Nous dérivons la distribution asymptotique d'une statistique récemment proposée pour détecter les ruptures, établissant ainsi sa validité. La simulation de sa distribution limite estimée conduit à un nouvel algorithme efficace de détection de ruptures, qui peut être utilisé sur des signaux trés longs. Pour finir, nous évaluons brièvement ce nouvel algorithme sur des données à une ou à plusieurs dimensions.

Mots-clés. Détection de ruptures, Signaux, U-statistique.

Abstract. The detection of change-points in a spatially or time-ordered data sequence is an important problem in many fields such as genetics and finance. We derive the asymptotic distribution of a statistic recently suggested for detecting change-points involving U-statistics, thus establishing its validity. Simulation of its estimated limit distribution leads to a new and computationally efficient change-point detection algorithm, which can be used on very long signals. To finish, we assess this new algorithm on one- and multi-dimensional data.

Keywords. Change-point, Signals, U-statistic.

1 Introduction

The present article builds upon an interesting nonparametric change-point detection method that was recently proposed by Matteson and James (2014). Their method uses U-statistics as the basis of a change-point test. Its interest lies in its ability to detect quite general types of change in distribution, rather than only changes in mean.

Our paper has two main objectives. First, it provides a full theoretical justification of the results in Matteson and James (2014), including a derivation of the limit distribution of the statistic. Second, we provide a method to simulate from an approximate version of the limit distribution. This leads to a new computationally efficient strategy for change-point detection that can be run on much longer signals. Due to space constraints, the computational algorithm and simulations, as well as proofs, can be found in the long version of the article at the author's websites.

2 Theoretical results

2.1 Measuring differences between multivariate distributions

Let us first briefly describe the origins of the nonparametric change-point detection method described in Matteson and James (2014). For random variables Y, Z taking values in $\mathbb{R}^d$, $d \geq 1$, let ϕ_Y and ϕ_Z denote their respective characteristic functions. A measure of the divergence (or "difference") between the distributions of Y and Z is as follows:

$$\mathcal{D}(Y, Z) = \int_{\mathbb{R}} |\phi_Y(t) - \phi_Z(t)|^2 \, w(t) dt, \tag{1}$$

where $w(t)$ is an arbitrary positive weight function for which this integral exists. It turns out that for a particular weight function which depends on a $\beta \in (0, 2)$, one can obtain a very useful result. Let Y, Y' be i.i.d. F_Y and Z, Z' be i.i.d. F_Z, with Y, Y', Z and Z' mutually independent. Denote by $|\cdot|$ the Euclidean norm on $\mathbb{R}^d$. Then, if $\mathbb{E}(|Y|^\beta + |Z|^\beta) < \infty$, Theorem 2 of Székely and Rizzo (2004) yields that

$$\mathcal{D}(Y, Z; \beta) = \mathcal{E}(Y, Z; \beta) := 2\mathbb{E}\,|Y - Z|^\beta - \mathbb{E}\,|Y - Y'|^\beta - \mathbb{E}\,|Z - Z'|^\beta \geq 0, \tag{2}$$

where we have written $\mathcal{D}(Y, Z; \beta)$ instead of $\mathcal{D}(Y, Z)$ to highlight dependence on β. Furthermore, Theorem 2 of Székely and Rizzo (2005) says that $\mathcal{E}(Y, Z; \beta) = 0$ if and only if Y and Z have the same distribution. This remarkable result leads to a simple data-driven divergence measure for distributions. Seen in the context of hypothesizing a change-point in a signal of independent observations $\mathbf{X} = (X_1, \ldots, X_n)$ after the k-th observation X_k, we simply calculate an empirical version of (2):

$$\begin{aligned} \mathcal{E}_{k,n}(\mathbf{X}; \beta) = \frac{2}{k(n-k)} \sum_{i=1}^{k} \sum_{j=k+1}^{n} |X_i - X_j|^\beta - \binom{k}{2}^{-1} \sum_{1 \leq i < j \leq k} |X_i - X_j|^\beta \\ - \binom{n-k}{2}^{-1} \sum_{1+k \leq i < j \leq n} |X_i - X_j|^\beta . \end{aligned} \tag{3}$$

Matteson and James (2014) state without proof that under the null hypothesis of $X_1, \ldots, X_n$ being i.i.d. (no change-points), the sample divergence given in (3) scaled by $\frac{k(n-k)}{n}$ converges in distribution to a non-degenerate random variable as long as $\min\{k, n-k\} \to \infty$. Furthermore, they state that if there is a change-point between two distinct i.i.d. distributions after the k-th point, the sample divergence scaled by $\frac{k(n-k)}{n}$ tends a.s. to infinity as long as $\min\{k, n-k\} \to \infty$. These claims clearly point to a useful statistical test for detecting change-points. However, we cannot find rigorous mathematical arguments to substantiate them in Matteson and James (2014), nor in the earlier work of Székely and Rizzo (2005).

Here, we shall show the existence of the non-degenerate random variable hinted at in Matteson and James (2014) by deriving its distribution. We also show that in the presence of a change-point the correctly-scaled sample divergence indeed tends to infinity with probability 1.

2.2 Main result

Let us first begin in a more general setup. Let $X_1, \ldots, X_n$ be independent $\mathbb{R}^d$-valued random variables. For any symmetric measurable function $\varphi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, whenever the indices make sense we define the following four terms:

$$V_k(\varphi) := \sum_{i=1}^{k} \sum_{j=k+1}^{n} \varphi(X_i, X_j), \tag{4}$$

$$U_n(\varphi) := \sum_{1 \le i < j \le n} \varphi(X_i, X_j), \tag{5}$$

$$U_k^{(1)}(\varphi) := \sum_{1 \le i < j \le k} \varphi(X_i, X_j), \tag{6}$$

$$U_k^{(2)}(\varphi) := \sum_{k+1 \le i < j \le n} \varphi(X_i, X_j). \tag{7}$$

Otherwise, define the term to be zero. Note that in the context of the change-point algorithm we have in mind, $\varphi(x, y) = \varphi_\beta(x, y) := |x - y|^\beta$, $\beta \in (0, 2)$. Next, let us define

$$U_{k,n}(\varphi) := \frac{2}{k(n-k)} V_k(\varphi) - \binom{k}{2}^{-1} U_k^{(1)}(\varphi) - \binom{n-k}{2}^{-1} U_k^{(2)}(\varphi). \tag{8}$$

While $U_{k,n}(\varphi)$ is not a U-statistic, we can express it as a linear combination of U-statistics. Indeed, we find that

$$U_{k,n}(\varphi) = \frac{2(n-1)}{k(n-k)} \left(\frac{U_n(\varphi)}{n-1} - \left(\frac{U_k^{(1)}(\varphi)}{k-1} + \frac{U_k^{(2)}(\varphi)}{n-k-1} \right) \right). \tag{9}$$

Therefore, we now have an expression for $U_{k,n}(\varphi)$ made up of U-statistics. Our aim is to use a test based on $U_{k,n}(\varphi)$ for the null hypothesis $\mathcal{H}_0 : X_1, \ldots, X_n$ have the same distribution, versus the alternative hypothesis $\mathcal{H}_1$ that there is a change-point in the sequence $X_1, \ldots, X_n$, i.e.,

$$\mathcal{H}_1 : \text{There is a } \gamma \in (0,1) \text{ such that } \mathbb{P}(X_1 \le t) = \cdots = \mathbb{P}(X_{\lfloor n\gamma \rfloor} \le t), \tag{10}$$

$$\mathbb{P}(X_{\lfloor n\gamma \rfloor + 1} \le t) = \cdots = \mathbb{P}(X_n \le t), \;\; t \in \mathbb{R}^d, \tag{11}$$

$$\text{and } \mathbb{P}(X_{\lfloor n\gamma \rfloor} \le t_0) \ne \mathbb{P}(X_{\lfloor n\gamma \rfloor + 1} \le t_0) \text{ for some } t_0. \tag{12}$$

For $u,\ v \in \mathbb{R}^d$, $u \leq v$ means that each component of u is less than or equal to the corresponding component of v.

Let us now examine the asymptotic properties of $U_{k,n}(\varphi)$. In the following, we shall denote by F the common (unknown) distribution function of the X_i under $\mathcal{H}_0$, X a generic random variable with distribution function F, and X' an independent copy of X. We assume that

$$\mathbb{E}\varphi^2(X,X') = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \varphi^2(x,y)dF(x)dF(y) < \infty, \tag{13}$$

and set $\Theta = \mathbb{E}\varphi(X,X')$. We also denote $\varphi_1(x) = \mathbb{E}\varphi(x,X')$, and define

$$h(x,y) = \varphi(x,y) - \varphi_1(x) - \varphi_1(y), \quad \tilde{h}_2(x,y) = h(x,y) + \Theta. \tag{14}$$

We define the operator A on $L_2(\mathbb{R}^d, F)$ by

$$Ag(x) := \int_{\mathbb{R}^d} \tilde{h}_2(x,y)g(y)dF(y), \quad x \in \mathbb{R}^d,\ g \in L_2(\mathbb{R}^d, F). \tag{15}$$

Let λ_i, $i \geq 1$, be the eigenvalues of this operator A with corresponding orthonormal eigenfunctions ϕ_i, $i \geq 1$. Since for all $x \in \mathbb{R}^d$,

$$\int_{\mathbb{R}^d} \tilde{h}_2(x,y)dF(y) = 0, \tag{16}$$

we see with $\phi_1 := 1$, $A\phi_1 = 0 =: \lambda_1\phi_1$. Thus $(0,1) = (\lambda_1, \phi_1)$ is an eigenvalue and normalized eigenfunction pair of the operator A. This implies that for every eigenvalue and normalized eigenfunction pair (λ_i, ϕ_i), $i \geq 2$, where λ_i is nonzero,

$$\mathbb{E}\left(\phi_1(X)\phi_i(X)\right) = \mathbb{E}\phi_i(X) = 0. \tag{17}$$

Moreover, we have that in $L_2(\mathbb{R}^d \times \mathbb{R}^d, F \times F)$,

$$\tilde{h}_2(x,y) = \lim_{K\to\infty} \sum_{i=1}^{K} \lambda_i\phi_i(x)\phi_i(y). \tag{18}$$

From this we get that

$$\mathbb{E}\tilde{h}_2^2(X,X') = \sum_{i=1}^{\infty} \lambda_i^2. \tag{19}$$

We shall assume further that

$$\sum_{i=1}^{\infty} |\lambda_i| < \infty. \tag{20}$$

It is crucial for the change-point testing procedure that we shall propose that the function $\tilde{h}_2(x,y)$ defined as in (20) with $\varphi(x,y) = \varphi_\beta(x,y) = |x-y|^\beta$, $\beta \in (0,2)$, satisfies (20) whenever (13) holds. A proof of this can be found in the long version of the article on the author's websites.

Next, for any fixed $\frac{2}{n} \le t < 1 - \frac{2}{n}$, $n \ge 3$, set

$$\begin{aligned}\mathbb{Y}_n(\tilde{h}_2, t) &:= \frac{\left(\lfloor nt \rfloor \left(n - \lfloor nt \rfloor\right)\right)^2}{n^2(n-1)} U_{\lfloor nt \rfloor, n}(\tilde{h}_2) \\ &= \frac{2\lfloor nt \rfloor \left(n - \lfloor nt \rfloor\right)}{n^2} \left(\frac{U_n(\tilde{h}_2)}{n-1} - \left(\frac{U^{(1)}_{\lfloor nt \rfloor}(\tilde{h}_2)}{\lfloor nt \rfloor - 1} + \frac{U^{(2)}_{\lfloor nt \rfloor}(\tilde{h}_2)}{n - \lfloor nt \rfloor - 1} \right) \right).\end{aligned} \tag{21}$$

In the following theorem, $\{\mathbb{B}^{(i)}\}_{i \ge 1}$ denotes a sequence of independent standard Brownian bridges.

Theorem 2.1 *Whenever X_i, $i \ge 1$ are i.i.d. F and φ satisfies (13) and (20), $\mathbb{Y}_n(\varphi, \cdot)$ converges weakly in $D^1[0,1]$ to the tied down mean zero continuous process $\mathbb{Y}$ defined on $[0,1]$ by*

$$\mathbb{Y}(t) := \sum_{i=1}^{\infty} \lambda_i \left(t\,(1-t) - \left(\mathbb{B}^{(i)}\,(t)\right)^2 \right). \tag{22}$$

In particular,

$$\sup_{t \in [0,1]} |\mathbb{Y}_n(\varphi, t)| \xrightarrow{\mathrm{D}} \sup_{t \in [0,1]} |\mathbb{Y}(t)|. \tag{23}$$

Remark 2.1 *Note that a special case of Theorem 2.1 says that for each $t \in (0,1)$,*

$$\frac{\left(\lfloor nt \rfloor \left(n - \lfloor nt \rfloor\right)\right)^2}{n^2(n-1)} U_{\lfloor nt \rfloor, n}(\varphi) \xrightarrow{\mathrm{D}} \mathbb{Y}(t). \tag{24}$$

As suggested in Matteson and James (2014), under the following assumption, a convergence with probability 1 result can be proved for the empirical statistic $\mathcal{E}_{k,n}(\mathbf{X}; \beta)$ in (3). We shall show that this is indeed the case.

Assumption 1 *Let Y_i, $i \ge 1$, and Z_i, $i \ge 1$, be independent i.i.d. sequences, respectively F_Y and F_Z. Also let Y, Y' be i.i.d. F_Y and Z, Z' be i.i.d. F_Z, with Y, Y', Z and Z' mutually independent. Assume that for some $\beta \in (0,2)$, $\mathbb{E}(|Y|^\beta + |Z|^\beta) < \infty$. Choose $\gamma \in (0,1)$. For any given $n > 1/\gamma$, let $X_i = Y_i$, for $i = 1, \ldots, \lfloor n\gamma \rfloor$, and $X_{i + \lfloor n\gamma \rfloor} = Z_i$, for $i = 1, \ldots, n - \lfloor n\gamma \rfloor$.*

Lemma 2.1 *Whenever for a given $\beta \in (0,2)$ Assumption 1 holds, with probability 1 we have:*

$$\mathcal{E}_{\lfloor n\gamma \rfloor, n}(\mathbf{X}; \beta) \to \mathcal{E}(Y, Z; \beta). \tag{25}$$

Next, let $\varphi(x,y)=|x-y|^{\beta}$, $\beta\in(0,2)$. We see that for any $\gamma\in(0,1)$ for all large enough n,

$$\sup_{t\in[0,1]}|\mathbb{Y}_n(\varphi,t)|\geq\frac{\left(\lfloor n\gamma\rfloor\left(n-\lfloor n\gamma\rfloor\right)\right)^2}{n^2(n-1)}\mathcal{E}_{\lfloor n\gamma\rfloor,n}(\mathbf{X};\beta),\tag{26}$$

where it is understood that Assumption 1 holds. Thus by Lemma 2.1, under Assumption 1, whenever $F_Y\neq F_Z$, with probability 1,

$$\sup_{t\in[0,1]}|\mathbb{Y}_n(\varphi,t)|\to\infty.\tag{27}$$

This shows that change-point tests based on the statistic $\sup_{t\in[0,1]}|\mathbb{Y}_n(\varphi,t)|$, under the *sequence of alternatives* of the type given by Assumption 1, are consistent. This also has great practical use when looking for change-points. Intuitively, the $k\in\{1,\ldots,n\}$ that maximizes (3) would be a good candidate for a change-point location.

3 From theory to practice

Theorem 2.1 and the consistency result that follows it lay a firm theoretical foundation to justify the change-point method introduced in Matteson and James (2014). For the present article, since we are not aware of a closed form expression for the distribution function of the limit process, we may imagine that this asymptotic result is of limited practical use. Remarkably, it turns out that we can efficiently approximate via simulation the distribution of its supremum, leading to a new change-point detection algorithm with similar performance to Matteson and James (2014) but much faster for longer signals. For space reasons, this and the simulations that go with it can be found in the longer version on the author's websites.

Bibliographie

[1] Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data, Journal of the American Statistical Association, 109, 334–345.

[2] Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension, InterStat, 5, 1–6.

[3] Székely, G. J. and Rizzo, M. L. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method, Journal of Classification, 22, 151–183.