



HAL
open science

Long signal change-point detection

G rard Biau, Kevin Bleakley, David Mason

► **To cite this version:**

G rard Biau, Kevin Bleakley, David Mason. Long signal change-point detection. 2015. hal-01140119v1

HAL Id: hal-01140119

<https://inria.hal.science/hal-01140119v1>

Preprint submitted on 7 Apr 2015 (v1), last revised 28 Sep 2015 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destin e au d p t et   la diffusion de documents scientifiques de niveau recherche, publi s ou non,  manant des  tablissements d'enseignement et de recherche fran ais ou  trangers, des laboratoires publics ou priv s.

Long signal change-point detection

G rard Biau^{1,2}, Kevin Bleakley^{3,4}, and David M. Mason⁵

¹Sorbonne Universit s, UPMC, France

²Institut Universitaire de France

³INRIA Saclay, France

⁴D partement de Math matiques d'Orsay, France

⁵University of Delaware, USA

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. 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. We assess the algorithm via simulations and on previously benchmarked real-world data sets.

I. INTRODUCTION

When met with a real-world data set ordered by time or space, it is often important to predict when or where something “changed” as we move temporally or spatially through it. In biology, for example, changes in an array Comparative Genomic Hybridization (aCGH) or Chip-Seq data signal as one moves across the genome can represent an event such as a change in genomic copy number, which is extremely important in cancer gene detection [16, 20]. In the financial world, detecting changes in multivariate time-series data is important for decision-making [25]. Change-point detection can also be used to detect financial anomalies [4] and significant changes in a sequence of images [12].

Change-point detection analysis is a well-studied field and there are numerous approaches to the problem. Its extensive literature ranges from parametric methods using log-likelihood functions [5, 14] to nonparametric ones based on Wilcoxon-type statistics, U-statistics and sequential ranks. The reader is referred to the monograph [6] for an in-depth treatment of these methods.

In change-point modeling it is generally supposed that we are dealing with a random process evolving in time or space. The aim is to develop a method to search for a point where possible changes occur in the mean, variance, distribution, etc. of the process. All in all, this comes down to finding ways to decide whether a given signal can be considered homogeneous in a statistical (stochastic) sense.

The present article builds upon an interesting nonparametric change-point detection method that was recently proposed by Matteson and James [15]. It uses U-statistics (see [10]) as the basis of its change-point test. Its interest lies in its ability to detect quite general types of change in distribution. Several theoretical results are presented in [15] to highlight some of the mathematical foundations of their method. These in turn lead to a simple and useful data-driven statistical test for change-point detection. The authors then apply this test successfully to simulated and real-world data.

There are however several weaknesses in [15] both from theoretical and practical points of view. Certain fundamental theoretical considerations are lightly treated, especially the assertion that a limit distribution exists for the important statistic, upon which the rest of the approach hangs. On the practical side, the method is computationally prohibitive for signals of more than a few thousand points, which is unfortunate because real-world signals can be typically much longer.

Our paper has two main objectives. First, it fills in missing theoretical results in [15] including a derivation of the limit distribution of the statistic. This requires the effective application of large sample theory techniques, which were developed to study degenerate U-statistics. 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.

The article is structured as follows. In Section II we provide some context and present the main theoretical results. In Section III we show how to approximate the limit distribution of the statistic, which leads to a new test strategy for change-point detection. We then show how to extend the method to much longer sequences. Simulations and real-data examples are provided in Section IV. A short discussion follows in Section V, and a proof of the paper's main result is given in Section VI. Some important technical results are detailed in the Appendix.

II. THEORETICAL RESULTS

I. Measuring differences between multivariate distributions

Let us first briefly describe the origins of the nonparametric change-point detection method described in [15]. 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:

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

where $w(t)$ is an arbitrary positive weight function for which this integral exists. It turns out that for the specific weight function

$$w(t; \beta) = \left(\frac{2\pi^{1/2}\Gamma(1 - \beta/2)}{\beta 2^\beta \Gamma((d + \beta)/2)} |t|^{d+\beta} \right)^{-1},$$

which depends on a $\beta \in (0, 2)$, one can obtain a not immediately obvious but 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

$$\mathbb{E}(|Y|^\beta + |Z|^\beta) < \infty, \tag{1}$$

Theorem 2 of [23] 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 . Therefore (1) implies that $\mathcal{E}(Y, Z; \beta) \in [0, \infty)$. Furthermore Theorem 2 of [23] 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 $X = X_1, \dots, X_n$ after the k -th observation X_k , we simply calculate an empirical version of (2):

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

Matteson and James [15] state without proof that under the null hypothesis of X_1, \dots, 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 $\min\{k, n-k\} \rightarrow \infty$. 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 $\min\{k, n-k\} \rightarrow \infty$. These claims clearly point to a useful statistical test for detecting change-points. However, we cannot find rigorous mathematical arguments to substantiate them in [15], nor in the earlier work [23].

As this is of fundamental importance to the theoretical and practical validity of this change-point detection method, we shall show the existence of the non-degenerate random variable hinted at in [15] by deriving its distribution. Our approach relies on the asymptotic behavior of U-statistic type processes, which were introduced for the first time for change-point detection in random sequences in [7]; see also Chapter 2 of the book [6]. We also show that in the presence of a change-point the correctly-scaled sample divergence indeed tends to infinity with probability 1.

II. Main result

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

$$\begin{aligned} V_k(\varphi) &:= \sum_{i=1}^k \sum_{j=k+1}^n \varphi(X_i, X_j), \\ U_n(\varphi) &:= \sum_{1 \leq i < j \leq n} \varphi(X_i, X_j), \\ U_k^{(1)}(\varphi) &:= \sum_{1 \leq i < j \leq k} \varphi(X_i, X_j), \\ U_k^{(2)}(\varphi) &:= \sum_{k+1 \leq i < j \leq n} \varphi(X_i, X_j). \end{aligned}$$

Otherwise, define the term to be zero; for instance, $U_1^{(1)}(\varphi) = 0$ and $U_k^{(2)}(\varphi) = 0$ for $k = n-1$ and n . Note that in the context of the change-point algorithm we have in mind, $\varphi(x, y) = \varphi_\beta(x, y) := |x - y|^\beta$, $\beta \in (0, 2)$, but the following results are valid for the more general φ defined above. Notice also that the last three terms are U-statistics absent their normalization constants. Next, let us define

$$U_{k,n}(\varphi) := \frac{2}{k(n-k)} V_k(\varphi) - \binom{k}{2}^{-1} U_k^{(1)}(\varphi) - \binom{n-k}{2}^{-1} U_k^{(2)}(\varphi). \quad (4)$$

Observe that $U_{k,n}(\varphi)$ is a general version of the empirical divergence given in (3). Notice that

$$V_k(\varphi) = U_n(\varphi) - U_k^{(1)}(\varphi) - U_k^{(2)}(\varphi). \quad (5)$$

While $U_{k,n}(\varphi)$ is not a U-statistic, we can use (5) to express it as a linear combination of U-statistics. Indeed, we find that

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

Therefore, we now have an expression for $U_{k,n}(\varphi)$ made up of U-statistics, which will be useful in the following.

Our aim is to use a test based on $U_{k,n}(\varphi)$ for the null hypothesis $\mathcal{H}_0 : X_1, \dots, X_n$ have the same distribution, versus the alternative hypothesis \mathcal{H}_1 that there is a change-point in the sequence X_1, \dots, X_n , i.e.,

$$\begin{aligned} \mathcal{H}_1 : & \text{There is a } \gamma \in (0, 1) \text{ such that } \mathbb{P}(X_1 \leq t) = \dots = \mathbb{P}(X_{\lfloor n\gamma \rfloor} \leq t), \\ & \mathbb{P}(X_{\lfloor n\gamma \rfloor + 1} \leq t) = \dots = \mathbb{P}(X_n \leq t), \quad t \in \mathbb{R}^d, \\ & \text{and } \mathbb{P}(X_{\lfloor n\gamma \rfloor} \leq t_0) \neq \mathbb{P}(X_{\lfloor n\gamma \rfloor + 1} \leq t_0) \text{ for some } t_0. \end{aligned}$$

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 . Also note that for any $z \in \mathbb{R}$, $\lfloor z \rfloor$ stands for its integer part.

Let us now examine the asymptotic properties of $U_{k,n}(\varphi)$. We shall be using notation, methods and results from Section 5.5.2 of monograph [19] to provide the groundwork. 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

$$\mathbb{E}\varphi^2(X, X') = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \varphi^2(x, y) dF(x) dF(y) < \infty, \quad (6)$$

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

$$h(x, y) = \varphi(x, y) - \varphi_1(x) - \varphi_1(y), \quad \tilde{h}_2(x, y) = h(x, y) + \Theta. \quad (7)$$

With this notation, we see that $\mathbb{E}h(X, X') = -\Theta$, and therefore that $\mathbb{E}\tilde{h}_2(X, X') = 0$. Furthermore,

$$U_{k,n}(\varphi) = U_{k,n}(h) = U_{k,n}(\tilde{h}_2), \quad (8)$$

since

$$\frac{U_n(\Theta)}{n-1} - \left(\frac{U_k^{(1)}(\Theta)}{k-1} + \frac{U_k^{(2)}(\Theta)}{n-k-1} \right) = \frac{U_n(\psi)}{n-1} - \left(\frac{U_k^{(1)}(\psi)}{k-1} + \frac{U_k^{(2)}(\psi)}{n-k-1} \right) = 0,$$

where $\psi(x, y) := \varphi_1(x) + \varphi_1(y)$. As in Section 5.5.2 of [19], we then define the operator A on $L_2(\mathbb{R}^d, F)$ by

$$Ag(x) := \int_{\mathbb{R}^d} \tilde{h}_2(x, y) g(y) dF(y), \quad x \in \mathbb{R}^d, \quad g \in L_2(\mathbb{R}^d, F). \quad (9)$$

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 $x \in \mathbb{R}^d$,

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

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,

$$\mathbb{E}(\phi_1(X)\phi_i(X)) = \mathbb{E}\phi_i(X) = 0. \quad (10)$$

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

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

From this we get that

$$\mathbb{E}\tilde{h}_2^2(X, X') = \sum_{i=1}^{\infty} \lambda_i^2. \quad (11)$$

For further details and theoretical justification of these claims, refer to Section 5.5.2 of [19] and both Exercise 44 on pg. 1083 and Exercise 56 on pg. 1087 of [8]. In fact, we shall assume further that

$$\sum_{i=1}^{\infty} |\lambda_i| < \infty. \quad (12)$$

It is crucial for the change-point testing procedure that we shall propose that the function $\tilde{h}_2(x, y)$ defined as in (12) with $\varphi(x, y) = \varphi_\beta(x, y) = |x - y|^\beta$, $\beta \in (0, 2)$, satisfies (12) whenever (6) holds. A proof of this is given in the Appendix.

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

$$\begin{aligned} \mathbb{Y}_n(\tilde{h}_2, t) &:= \frac{(\lfloor nt \rfloor (n - \lfloor nt \rfloor))^2}{n^2(n-1)} U_{\lfloor nt \rfloor, n}(\tilde{h}_2) \\ &= \frac{2\lfloor nt \rfloor (n - \lfloor nt \rfloor)}{n^2} \left(\frac{U_n(\tilde{h}_2)}{n-1} - \left(\frac{U_{\lfloor nt \rfloor}^{(1)}(\tilde{h}_2)}{\lfloor nt \rfloor - 1} + \frac{U_{\lfloor nt \rfloor}^{(2)}(\tilde{h}_2)}{n - \lfloor nt \rfloor - 1} \right) \right). \end{aligned} \quad (13)$$

We define $U_0^{(1)}(\tilde{h}_2) = 0$, $U_0^{(2)}(\tilde{h}_2) = 0$, $U_1^{(1)}(\tilde{h}_2)/0 = 0$, and $U_{n-1}^{(2)}(\tilde{h}_2)/0 = 0$, which gives

$$\begin{aligned} \mathbb{Y}_n(\tilde{h}_2, t) &= 0, \text{ for } t \in \left[0, \frac{1}{n}\right), \\ \mathbb{Y}_n(\tilde{h}_2, t) &= \frac{2(n-1)}{n^2} \left(\frac{U_n(\tilde{h}_2)}{n-1} - \frac{U_1^{(2)}(\tilde{h}_2)}{n-2} \right), \text{ for } t \in \left[\frac{1}{n}, \frac{2}{n}\right), \\ \mathbb{Y}_n(\tilde{h}_2, t) &= \frac{4(n-2)}{n^2} \left(\frac{U_n(\tilde{h}_2)}{n-1} - \frac{U_{n-2}^{(1)}(\tilde{h}_2)}{n-3} - U_{n-2}^{(2)}(\tilde{h}_2) \right), \text{ for } t \in \left[1 - \frac{2}{n}, 1 - \frac{1}{n}\right), \\ \mathbb{Y}_n(\tilde{h}_2, t) &= \frac{2(n-1)}{n^2} \left(\frac{U_n(\tilde{h}_2)}{n-1} - \frac{U_{n-1}^{(1)}(\tilde{h}_2)}{n-2} \right), \text{ for } t \in \left[1 - \frac{1}{n}, 1\right), \text{ and } \mathbb{Y}_n(\tilde{h}_2, 1) = 0. \end{aligned}$$

One can readily check that $\mathbb{Y}_n(\tilde{h}_2, \cdot) \in D^1[0, 1]$, the space of bounded measurable real-valued functions defined on $[0, 1]$ that are right-continuous real valued functions with left-hand limits. Notice that on account of (8) we can also write $\mathbb{Y}_n(\tilde{h}_2, \cdot) = \mathbb{Y}_n(\varphi, \cdot)$, and we will do so from now on. In the following theorem, $\{\mathbb{W}^{(i)}\}_{i \geq 1}$ denotes a sequence of independent standard Wiener processes.

Theorem II.1 Whenever φ satisfies (6) and (12), $\mathbb{Y}_n(\varphi, \cdot)$ converges weakly in $D^1[0, 1]$ to the tied down mean zero continuous process \mathbb{Y} defined on $[0, 1]$ by

$$\mathbb{Y}(t) := \sum_{i=1}^{\infty} \lambda_i \left(t(1-t) \left((\mathbb{W}^{(i)}(1))^2 + 1 \right) - (1-t) (\mathbb{W}^{(i)}(t))^2 - t (\mathbb{W}^{(i)}(1) - \mathbb{W}^{(i)}(t))^2 \right).$$

In particular,

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

The proof of this theorem is deferred to Section VI. Note that a special case of Theorem II.1 says that for each $t \in (0, 1)$,

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

To the best of our knowledge, this is the first time that the limit distribution for this statistic has been proved to exist and been identified. As suggested in [15], with the following assumption a convergence with probability 1 result can also be proved for the empirical statistic $\mathcal{E}_{n,k}(X; \beta)$ in (3).

Assumption 1 Choose $\gamma \in (0, 1)$. For any given $n > 1/\gamma$ let $Y_i, i = 1, \dots, \lfloor n\gamma \rfloor$ be i.i.d. F_Y and $Z_j, j = 1, \dots, n - \lfloor n\gamma \rfloor$ be i.i.d. 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 $\beta \in (0, 2)$, $\mathbb{E}(|Y|^\beta + |Z|^\beta) < \infty$. Let $X_i = Y_i$, for $i = 1, \dots, \lfloor n\gamma \rfloor$, and $X_{i+\lfloor n\gamma \rfloor} = Z_i$, for $i = 1, \dots, n - \lfloor n\gamma \rfloor$.

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

$$\mathcal{E}_{n, \lfloor n\gamma \rfloor}(X; \beta) \rightarrow \mathcal{E}(Y, Z; \beta). \quad (14)$$

The proof of this can be found in the Appendix. 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 ,

$$\sup_{t \in [0,1]} |\mathbb{Y}_n(\varphi, t)| \geq \frac{(\lfloor n\gamma \rfloor (n - \lfloor n\gamma \rfloor))^2}{n^2(n-1)} U_{\lfloor n\gamma \rfloor, n}(\varphi).$$

Thus by Lemma II.1, under Assumption 1, whenever $F_X \neq F_Y$, with probability 1,

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

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 $k \in \{1, \dots, n\}$ that maximizes (3) would be a good candidate for a change-point location.

III. FROM THEORY TO PRACTICE

Theorem II.1 and the consistency result that follows it lay a firm theoretical foundation to justify the change-point method introduced in [15]. 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 [15] but immensely faster for longer signals. For instance, finding and testing the first change-point in a signal of length 5 000 takes eight seconds with our method and eight minutes using [15].

To simulate the process \mathbb{Y} we need true or estimated values of the λ_i . Recall that these are the eigenvalues of the operator A defined in (9). Following [13], the (usually infinite) spectrum of A can be consistently approximated by the (finite) spectrum of the empirical $n \times n$ matrix \tilde{H}_n whose (i, j) -th entry is given by

$$\tilde{H}_n(X_i, X_j) = \frac{1}{n} (\varphi(X_i, X_j) - \mu(i) - \mu(j) + \eta),$$

where μ is the vector of row means (excluding the diagonal entry) of matrix $\varphi(X_i, X_j)$ and η the mean of its upper-diagonal elements.

In our experience, the λ_i estimated in this way tend to be quite accurate for even small n . We assert this because upon simulating longer and longer i.i.d. signals, rapid convergence of the λ_i is clear. Furthermore, as there is an exponential drop-off in their magnitude, working with only a small number (say 20 or 50) of the largest ones appears to be sufficient for obtaining good results. We illustrate these claims in Section IV. Let us now present our basic algorithm for detecting and testing for one potential change-point.

Algorithm for detecting and testing one change-point

1. Given signal X_1, \dots, X_n , $n \geq 4$, find the $2 \leq k \leq n - 2$ that maximizes the original empirical divergence given in (3) multiplied by the correct normalization given in (13), i.e., $k^2(n - k)^2/n^2(n - 1)$, and denote the value of this maximum t^* .
2. Calculate the m largest (in absolute value) eigenvalues of the matrix \tilde{H}_n , where $\varphi(X_i, X_j) = |X_i - X_j|^\beta$ and $\beta \in (0, 2)$.
3. Simulate R times the m -truncated version of $\mathbb{Y}(t)$ using the m eigenvalues from the previous step. Record the R values s_1, \dots, s_R of the (absolute) supremum of the process obtained.
4. Reject the null hypothesis of no distributional change (at level α) if $t_{\text{crit}} \leq \alpha$, where $t_{\text{crit}} := \frac{1}{R} \sum_{r=1}^R \mathbf{1}_{\{s_r > t^*\}}$. In this case, we deduce a change-point at the k at which t^* is found. Typically, we set $\alpha = .05$.

Remark III.1 *The E-divisive algorithm described in [15] follows a similar logic except that t_{crit} is calculated via permutation (see [17]). Instead of steps 2 and 3, the order of the n data is permuted R times and for the r -th permuted signal, $1 \leq r \leq R$, step 1 is performed to obtain the absolute maximum s_r . The same step 4 is then used to accept or reject the change-point.*

Note that the above algorithm can be used to find a putative change-point in any contiguous subset of the original signal. This motivates the following bisection algorithm for finding all change-points in a signal.

Algorithm for detecting and testing many change-points.

1. Run the one change-point algorithm on the whole signal. If the null hypothesis of no distributional change is not rejected, stop. Otherwise, we suppose have a first change-point at k and two resulting subsignals X_1, \dots, X_k and X_{k+1}, \dots, X_n , which we put in a waiting list.
2. While the waiting list is non-empty, run the one change-point algorithm on the longest-waiting subsignal. If the null hypothesis of no distributional change is not rejected, stop. Otherwise, accept a change-point at the location of t^* and add the two resulting subsignals to the waiting list. In either of these cases, return to the waiting list (i.e., repeat the present step until the waiting list is empty).
3. Output the list of change-points found.

The permutation approach (E-divisive) of [15] is effective for short signals. Indeed, [11] showed that if one can perform all possible permutations, the method produces a test that is level α . However, a signal with $n = 10$ points already implies more than three million permutations, so a Monte Carlo strategy (i.e., subsampling permutations with replacement) becomes necessary, typically with $R = 499$. This also gives a test that is theoretically level α (see [17]) but with much-diminished power.

One could propose increasing the value of R but there is an unfortunate computational bottleneck in the approach. Usually, one stores in memory the matrix of $|X_i - X_j|^\beta$ in order to efficiently permute rows/columns and therefore recalculate t^* each time. But for more than a few thousand points, manipulating this matrix is slow if not impossible due to memory constraints. The only alternative to storing and permuting this matrix is simply to recalculate it each time for each permutation, but this is very computationally expensive as n increases. Consequently, the E-divisive approach is only useful for signals up to a few thousand points.

In contrast to this, our algorithm, based on an asymptotic result, risks underperforming on extremely short signals, and its performance will also depend on our ability to estimate well the set of largest λ_i . In reality though, it works quite well, even on short signals. The matrix with entries $|X_i - X_j|^\beta$ needs only to be stored once in memory, and all standard mathematical software (such as Matlab and R) have efficient functions for finding its largest m eigenvalues (the `eigs` function in Matlab and the `eigs` function in the R package `rARPACK`). Each iteration of the algorithm's simulation step requires summing the columns of an $m \times T$ matrix of standard normal variables, where m is the number of λ_i retained and T the number of grid points over which we approximate the Wiener processes between 0 and 1. Each row of the matrix corresponds to one λ_i and a discretized Wiener process across T grid points. For $m = 50$ and $T = 1000$ it takes about one second to perform this $R = 499$ times, and is independent of the number of points in the signal. In contrast, the E-divisive method takes about ten seconds for $n = 1000$, one minute for $n = 2000$, eight minutes for $n = 5000$, etc. One clearly sees the advantage of our approach for longer signals when there are several change-points to find.

IV. EXPERIMENTAL VALIDATION AND ANALYSIS

I. Simulated and real data examples

It is very important to start with the simplest possible case in order to demonstrate the fundamental validity of the new method. A basis for comparison is the E-divisive method from [15]. Here, we consider signals of length $n \in \{10, 100, 1000, 10000\}$ for which either the

whole signal is i.i.d. $\mathcal{N}(0,1)$ or else there is a change-point of height $c \in \{0.1, 0.2, 0.5, 1, 2, 5\}$ after the $(n/2)$ -th point, i.e., the second half of the signal is i.i.d. $\mathcal{N}(c,1)$.

In the former case, we look at the behavior of the Type I error, i.e., the probability of detecting a change-point when there was none. We have fixed $\alpha = .05$ and want to see how close each method is to this as n increases. In the latter case, we look at the power of the test associated to each method, i.e., the probability that an actual change-point is correctly detected as n and c increase. We averaged over 1 000 trials. In the following, unless otherwise mentioned we fix $\beta = 1$. For the asymptotic method, the Wiener processes were simulated 499 times; similarly, for E-divisive we permuted 499 times. Both null distributions were therefore estimated using the same number of repeats. Note that we did not test the E-divisive method for $n = 10\,000$ because each of the 1 000 trials would have taken around two hours to run. All times given in this paper are with respect to a laptop with a 2.13 GHz Intel Core 2 Duo processor with 4Gb of memory. Results are presented in Figure 1. For the Type I error, we

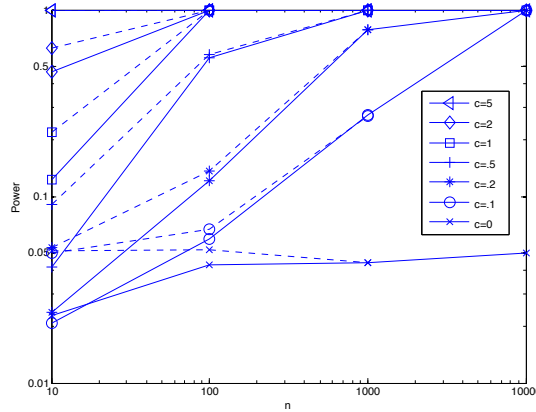


Figure 1: Statistical power of the asymptotic (solid line) and E-divisive (dotted line) methods for detecting change c in mean in a Gaussian signal of length n . The first $n/2$ points are i.i.d. $\mathcal{N}(0,1)$ and the last $n/2$ points i.i.d. $\mathcal{N}(c,1)$. The Type I error is also shown ($c = 0$). Results are averaged over 1 000 trials.

see that both methods hover around the intended value of .05, except for extremely short signals ($n = 10$). As for the statistical power, it increases as n and c increase. Furthermore, the asymptotic method rapidly reaches a similar performance as E-divisive: for $n = 10$, E-divisive is better (but still with quite poor power), for $n = 100$ the asymptotic method has almost caught up, and somewhere between $n = 100$ and $n = 1\,000$ the results become essentially identical; the asymptotic method has a slight edge at $n = 1\,000$.

Let us now see to what extent our method is able to detect changes in variance and tail shape. We considered Gaussian signals of length $n \in \{10, 100, 1\,000, 10\,000\}$ for which there is a change-point after the $(n/2)$ -th point, i.e., the first half of the signal is i.i.d. $\mathcal{N}(0,1)$ and the second half either i.i.d. $\mathcal{N}(0,\sigma^2)$ for $\sigma^2 \in \{2, 5, 10\}$ or i.i.d. Student's t_v distributions with $v \in \{2, 8, 16\}$. Results were averaged over 1 000 trials and are shown in Figure 2. As before, the statistical power tends to increase as n increases and either σ^2 increases or v decreases. The asymptotic method matches or beats the performance of E-divisive starting somewhere between $n = 100$ and $n = 1\,000$.

Next, we considered a real aCGH data set of 57 bladder tumor samples presented in [21] and previously examined in [3] and [15]. Each aCGH profile gives the relative quantity

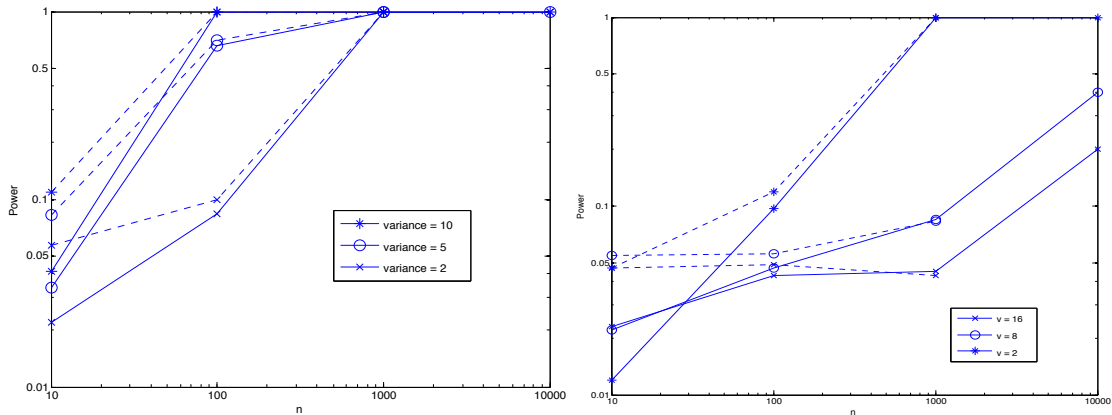


Figure 2: Statistical power of the asymptotic method (solid line) and E-divisive method (dotted line) for detecting change in variance (left) and tail (right) in a signal of length n . The first $n/2$ points are i.i.d. $\mathcal{N}(0,1)$ and the last $n/2$ points either i.i.d. $\mathcal{N}(0,\sigma^2)$, $\sigma^2 \in \{2,5,10\}$ (left) or from a Student's t_v distribution with v degrees of freedom, $v \in \{2,8,16\}$ (right). Results are averaged over 1000 trials.

of DNA for 2215 probes. Samples with more than 7% missing values were removed and remaining missing values imputed as the average of their neighbors as in [15]. This left 43 of the original 57. The first goal is simply to show that our algorithm works meaningfully on multi-dimensional signals (here, 43-dimensional); the biological goal itself is to find locations at which many samples change mean simultaneously, as these may be important copy-number changes in bladder cancer. We found 99 such change-points in a time of four minutes. It seems that several algorithms overestimate the number of useful shared change-points (97 in [3] and [15]). For instance, using our method with $\beta = 1$ seems to find a “shared” change in mean even when only a small number of signals do seem to have that particular change-point; this is fine in theory but not necessarily helpful for the true biological goal mentioned above.

However, we have noticed an intriguing property of our algorithm. If $\beta \rightarrow 0$, the number of change-points selected also decreases to zero, broadly in a way that change-points found later in the algorithm are removed first. Consequently, the choice of a small β seems to act to select “important” variables (here, shared change-points). For instance, with $\beta = .001$, only one change-point is selected, at location 1724. This is where nearly every profile (at least 40 of the 43) has a visually obvious change-point (or amplification—another type of genetic event, here seen as an isolated point far from the signal). Raising β to .002 adds three further change-points (811, 1268 and 1907) and to .005 five more (182, 428, 1141, 1534, and 2044). These are shown for four individuals in Figure 3 by solid, dashed and dotted lines respectively. These change-point locations are indeed shared by many signals, on average around 18/43 from a simple visual check. It is therefore an interesting open question as to whether our algorithm might be used to perform model selection, i.e., selection of the “best” β for a given task, but this is out of the scope here.

II. Algorithm for long signals

Remember that as it currently stands, the longest signal that we can treat depends on the largest matrix that can be stored, which depends in turn on the memory of a given computer (memory problems for simply storing a matrix on a standard PC typically start to occur around

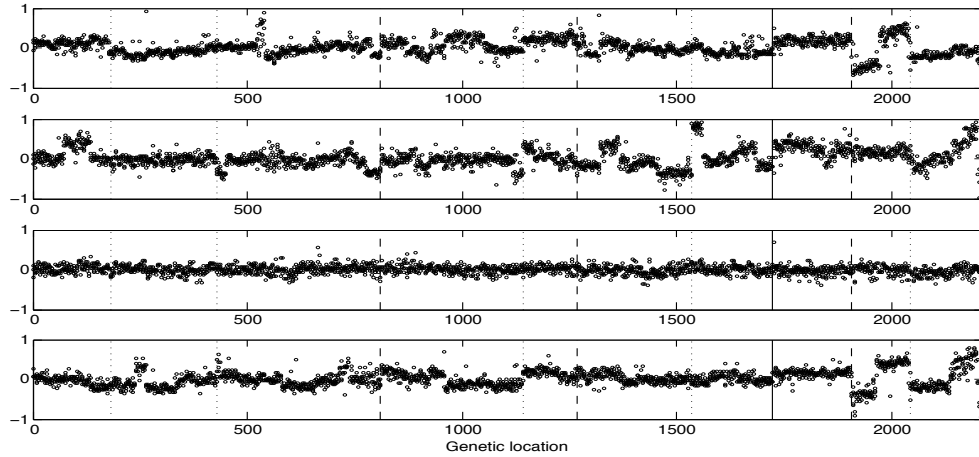


Figure 3: Shared change-point detection in a set of 43 bladder tumor aCGH profiles; four profiles are shown here. The signal is normalized to zero with respect to the normal copy-number of two. Signals above or below zero indicate gains or losses in copy number respectively. Large isolated values indicate amplifications—another important genetic event. The only change-point found for $\beta = .001$ is at location 1724 (solid line). Further shared change-points are found when raising β to .002 (dashed lines) and .005 (dotted lines).

$n = 10$ -15000). For this reason, we now propose a modified algorithm that can treat vastly longer signals. The basic idea is simply to create a sub-signal of equidistant points (here, 2000 points) from the original (long) signal and to run the one change-point algorithm on it. If the sub-signal accurately represents the original signal (essentially if the sub-signal is relatively dense with respect to the true number of change-points), changes in distribution should still be detected. Because of the information loss due to sub-sampling, we then add an extra step to refine the predicted location of the change-point in the original signal (do a local search for the most likely change-point using *all* signal points in an interval around the change-point predicted above). Of course, if the interval currently being treated is short enough, we simply run the normal one change-point algorithm.

We tested this strategy on 3-dimensional simulated signals of length $10^3, 10^4, 10^5$ and 10^6 with ten change-points. After randomly selecting change-point locations, each segment between consecutive change-points was randomly assigned (coin toss) to be either multidimensional Gaussian or Student’s t-distributions, with random means (uniform $\mu \in [-1, 1]^3$), variance-covariance matrices (positive-definite with lower-diagonal elements uniformly generated from $[0, 1]$ and mirrored to the upper-diagonal) and degrees of freedom (one of 2, 8 or 16 with equal probability) depending on the type of distribution randomly selected.

The exact performance of the method in correctly finding change-point locations depends on how much we allow the distribution to change after each change-point; in effect, the method will perform almost perfectly if the data is “easy” enough. For this reason, here we are much more interested in simply measuring its ability to run with extremely long signals in a short period of time. Figure 4 (left) shows a typical trial with 10000 points in a 3-dimensional signal—each plot is one of the three dimensions. Notice that some change-points are easily spotted by eye, others not so. For simplicity, in the top plot we show the location of the true change-points (solid lines) and in the other two plots, the algorithm’s predicted change-points (dotted lines,

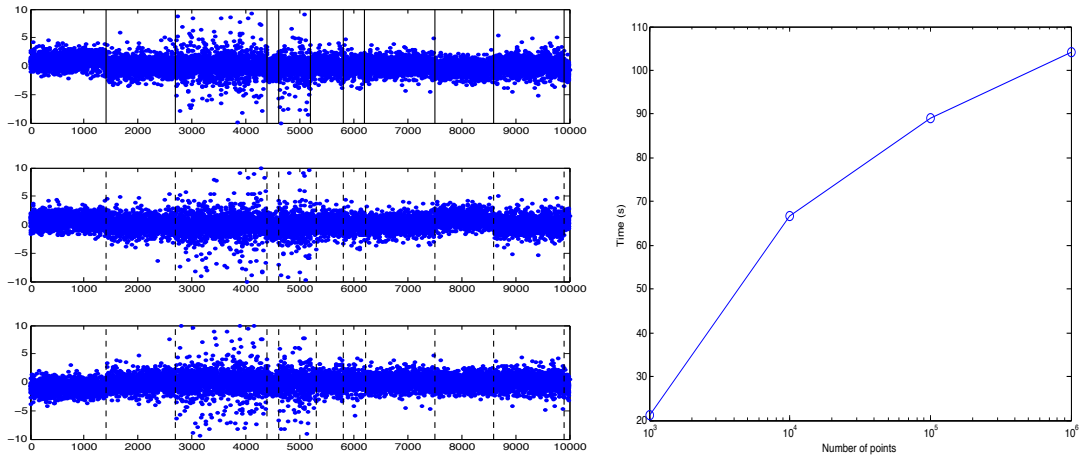


Figure 4: Long-signal change-point detection. Left: typical 3-dimensional signal with 10 000 points and ten random change-point locations. True change-point locations are shown by solid vertical lines, and the predicted change-points (joint across the three dimensions) in dotted lines. Right: time required to run the algorithm for long signals, averaged over 100 trials.

shared of course across the three dimensions/plots). Clearly the actual performance of the algorithm on moderately difficult data is not in question here. Figure 4 (right) shows the time required to run the algorithm for long signals, averaged over 100 trials. We see that even with one million points, we typically have a result in under two minutes.

V. DISCUSSION

We have derived the asymptotic distribution of a statistic that was previously used to build algorithms for finding change-points in signals. Our new result led to a novel way to build a practical algorithm for general change-point detection in long and potentially multidimensional signals that came from the surprising realization that it was possible to approximately simulate from this quite complicated asymptotic distribution. Furthermore, the method appears to have higher power (in the statistical sense) than previous methods based on permutation tests for signals of a thousand points or more. We have tested the algorithm on several simulated and real data sets, as well as a sub-sample variant for dealing with extremely long signals.

An interesting line of future research would be to find ways to segment the original signal without requiring stocking a matrix in memory with the same number of rows and columns as there are points in the signal, currently a bottleneck for our approach and even more so for previous permutation approaches. In addition, the question of a pertinent choice of the power $\beta \in (0, 2)$ remains an open question.

VI. PROOF OF THEOREM II.1

To prove Theorem II.1, we require a useful technical result. Let us begin with some notation. For each integer $K \geq 1$, let $D^K[0, 1]$ denote the space of bounded measurable functions defined on $[0, 1]$ taking values in \mathbb{R}^K that are right-continuous with left-hand limits. For each integer $n \geq 1$, let $\mathbb{V}_n^{(k)}$, $k \geq 1$, be a sequence of processes taking values in $D^1[0, 1]$ such that for some

$M > 0$, uniformly in $k \geq 1$ and $n \geq 1$,

$$\mathbb{E} \left(\sup_{t \in [0,1]} \left| \mathbb{V}_n^{(k)}(t) \right| \right) \leq M. \quad (15)$$

For each integer $K \geq 1$, define the process taking values in $D^K[0,1]$ by

$$\mathbb{V}_{n,K} = \left(\mathbb{V}_n^{(1)}, \dots, \mathbb{V}_n^{(K)} \right).$$

Assume that for each integer $K \geq 1$, $\mathbb{V}_{n,K}$ converges weakly as $n \rightarrow \infty$ to the $D^K[0,1]$ -valued process \mathbb{V}_K defined as

$$\mathbb{V}_K := \left(\mathbb{V}^{(1)}, \dots, \mathbb{V}^{(K)} \right),$$

where $\mathbb{V}^{(k)}$, $k \geq 1$, is a sequence of $D^1[0,1]$ -valued processes such that for some $M > 0$, uniformly in $k \geq 1$,

$$\mathbb{E} \left(\sup_{t \in [0,1]} \left| \mathbb{V}^{(k)}(t) \right| \right) \leq M. \quad (16)$$

We shall establish the following useful result.

Proposition VI.1 *With the notation and assumptions introduced above, for any choice of constants a_m , $m \geq 1$, satisfying $\sum_{m=1}^{\infty} |a_m| < \infty$, the sequence of $D^1[0,1]$ -valued processes*

$$T_n := \sum_{m=1}^{\infty} a_m \mathbb{V}_n^{(m)}$$

converges weakly in $D^1[0,1]$ to the $D^1[0,1]$ -valued process

$$T := \sum_{m=1}^{\infty} a_m \mathbb{V}^{(m)}.$$

Proof. Notice that by (15)

$$\mathbb{E} \left(\sum_{m=1}^{\infty} |a_m| \sup_{t \in [0,1]} \left| \mathbb{V}_n^{(m)}(t) \right| \right) \leq M \sum_{m=1}^{\infty} |a_m| < \infty.$$

From this we get that with probability 1, for each $n \geq 1$,

$$\sum_{m=1}^{\infty} |a_m| \sup_{t \in [0,1]} \left| \mathbb{V}_n^{(m)}(t) \right| < \infty,$$

which in turn implies that with probability 1, for each $n \geq 1$,

$$\lim_{K \rightarrow \infty} \sup_{t \in [0,1]} \left| \bar{T}_n^{(K)}(t) \right| = 0, \quad (17)$$

where

$$\bar{T}_n^{(K)}(t) := \sum_{m=K+1}^{\infty} a_m \mathbb{V}_n^{(m)}(t).$$

Since for each $n \geq 1$ and $K \geq 1$, $T_n^{(K)} \in D^1[0,1]$, where $T_n^{(K)} := \sum_{m=1}^K a_m \mathbb{V}_n^{(m)}$, by completeness of $D^1[0,1]$ in the supremum metric (see page 150 of monograph [2]), we infer that $T_n \in D^1[0,1]$. In the same way we get using (16) that

$$\lim_{K \rightarrow \infty} \sup_{t \in [0,1]} \left| \bar{T}^{(K)}(t) \right| = 0, \quad (18)$$

where

$$\bar{T}^{(K)}(t) := \sum_{m=K+1}^{\infty} a_m \mathbb{V}^{(m)}(t),$$

and thus that $T \in D^1[0,1]$. Also, since by assumption for each integer $K \geq 1$, $\mathbb{V}_{n,K}$ converges weakly as $n \rightarrow \infty$ to the $D^K[0,1]$ -valued process \mathbb{V}_K , we get that $T_n^{(K)}$ converges weakly in $D^1[0,1]$ to $T^{(K)}$, where

$$T_n^{(K)} := \sum_{m=1}^K a_m \mathbb{V}_n^{(m)} \quad \text{and} \quad T^{(K)} := \sum_{m=1}^K a_m \mathbb{V}^{(m)}.$$

We complete the proof by combining this with (17) and (18), and then appealing to Theorem 4.2 of [2]. \square

We are now ready to prove Theorem II.1. It turns out that it is more convenient to prove the result for the following version of the process \mathbb{Y}_n , namely

$$\tilde{\mathbb{Y}}_n(\tilde{h}_2, t) := \frac{2\lfloor nt \rfloor (n - \lfloor nt \rfloor)}{n^3} U_n(\tilde{h}_2) - \frac{2(n - \lfloor nt \rfloor)}{n^2} U_{\lfloor nt \rfloor}^{(1)}(\tilde{h}_2) - \frac{2\lfloor nt \rfloor}{n^2} U_{\lfloor nt \rfloor}^{(2)}(\tilde{h}_2),$$

which is readily shown to be asymptotically equivalent to $\mathbb{Y}_n(\tilde{h}_2, t)$. Following pages 196-197 of [19], we see that

$$\begin{aligned} \frac{2U_n(\tilde{h}_2)}{n} &= \sum_{k=1}^{\infty} \lambda_k \left[\left(\sum_{i=1}^n \phi_k(X_i) / \sqrt{n} \right)^2 - \frac{1}{n} \sum_{i=1}^n \phi_k^2(X_i) \right] =: \sum_{k=1}^{\infty} \lambda_k \Delta_{n,k}, \\ \frac{2U_{\lfloor nt \rfloor, n}^{(1)}(\tilde{h}_2)}{n} &= \sum_{k=1}^{\infty} \lambda_k \left[\left(\sum_{i=1}^{\lfloor nt \rfloor} \phi_k(X_i) / \sqrt{n} \right)^2 - \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} \phi_k^2(X_i) \right] =: \sum_{k=1}^{\infty} \lambda_k \Delta_{n,k}^{(1)}(t), \end{aligned}$$

and

$$\frac{2U_{\lfloor nt \rfloor, n}^{(2)}(\tilde{h}_2)}{n} = \sum_{k=1}^{\infty} \lambda_k \left[\left(\sum_{i=1+\lfloor nt \rfloor}^n \phi_k(X_i) / \sqrt{n} \right)^2 - \frac{1}{n} \sum_{i=1+\lfloor nt \rfloor}^n \phi_k^2(X_i) \right] =: \sum_{k=1}^{\infty} \lambda_k \Delta_{n,k}^{(2)}(t).$$

Thus,

$$\tilde{\mathbb{Y}}_n(\tilde{h}_2, t) = \sum_{k=1}^{\infty} \lambda_k \left(\frac{\lfloor nt \rfloor (n - \lfloor nt \rfloor)}{n^2} \Delta_{n,k} - \frac{(n - \lfloor nt \rfloor)}{n} \Delta_{n,k}^{(1)}(t) - \frac{\lfloor nt \rfloor}{n} \Delta_{n,k}^{(2)}(t) \right) =: \sum_{k=1}^{\infty} \lambda_k \mathbb{V}_n^{(k)}(t). \quad (19)$$

We also write

$$\mathbb{Y}(t) := \sum_{k=1}^{\infty} \lambda_k \mathbb{V}^{(k)}(t),$$

where, for $k \geq 1$,

$$\begin{aligned}\mathbb{V}^{(k)}(t) &= t(1-t) \left(\left(\mathbb{W}^{(k)}(1) \right)^2 - 1 \right) - (1-t) \left(\left(\mathbb{W}^{(k)}(t) \right)^2 - t \right) \\ &\quad - t \left(\left(\mathbb{W}^{(k)}(1) - \mathbb{W}^{(k)}(t) \right)^2 - (1-t) \right) \\ &= t(1-t) \left(\left(\mathbb{W}^{(k)}(1) \right)^2 + 1 \right) - (1-t) \left(\mathbb{W}^{(k)}(t) \right)^2 - t \left(\mathbb{W}^{(k)}(1) - \mathbb{W}^{(k)}(t) \right)^2. \quad (20)\end{aligned}$$

A simple application of Doob's inequality shows that there exists a constant $M > 0$ such that (15) and (16) hold, for $\mathbb{V}_n^{(k)}$ and $\mathbb{V}^{(k)}$ defined as in (19) and (20).

For any integer $K \geq 1$, let \mathbf{U}_1 be the random vector such that $\mathbf{U}_1^T = (\phi_1(X_1), \dots, \phi_K(X_1))$, and for any $n \geq 1$ let $\mathbf{U}_1, \dots, \mathbf{U}_n$ be i.i.d. \mathbf{U}_1 . Clearly, $(\mathbf{U}_1 \mathbf{U}_1^T) = I_K$ and by the multivariate central limit theorem,

$$n^{1/2} \bar{\mathbf{U}}_n \xrightarrow{D} N(0, I_K) \quad (21)$$

as $n \rightarrow \infty$, where $\bar{\mathbf{U}}_n = n^{-1} \sum_{j=1}^n \mathbf{U}_j$. Consider the process defined on $D^K[0, 1]$ by

$$\mathbb{W}_{n,K}(t) := \left(n^{-1/2} \sum_{j \leq \lfloor nt \rfloor} \phi_1(X_j), \dots, n^{-1/2} \sum_{j \leq \lfloor nt \rfloor} \phi_K(X_j) \right) =: \left(\mathbb{W}_n^{(1)}(t), \dots, \mathbb{W}_n^{(K)}(t) \right),$$

where for any $i \geq 1$,

$$\mathbb{W}_n^{(i)}(t) := n^{-1/2} \sum_{j \leq \lfloor nt \rfloor} \phi_i(X_j).$$

By Donsker's theorem the process $\mathbb{W}_{n,K}(t)$ converges weakly as $n \rightarrow \infty$ to the \mathbb{R}^K -valued Wiener process $(\mathbb{W}_K(t))_{0 \leq t \leq 1}$, with mean vector zero and covariance matrix $(t_1 \wedge t_2) I_K$, $t_1, t_2 \in [0, 1]$, where

$$\mathbb{W}_K(t) := \left(\mathbb{W}^{(1)}(t), \dots, \mathbb{W}^{(K)}(t) \right).$$

Using this fact along with the law of large numbers one readily verifies that for each integer $K \geq 1$, $(\mathbb{V}_n^{(1)}, \dots, \mathbb{V}_n^{(K)})$ converges weakly as $n \rightarrow \infty$ to $(\mathbb{V}^{(1)}, \dots, \mathbb{V}^{(K)})$, where $\mathbb{V}_n^{(i)}$ and $\mathbb{V}^{(i)}$ are defined as in (19) and (20). All the conditions for Proposition VI.1 to hold have been verified. Thus the proof of Theorem II.1 is complete. \square

VII. APPENDIX

I. Proof of Lemma II.1

Notice that for each $n > 1$ the $\mathcal{E}_{n, \lfloor n\gamma \rfloor}(X; \beta)$ is equal to the statistic in (3) with $k = \lfloor n\gamma \rfloor$. By the law of large numbers for U-statistics (see Theorem 1 of [18]) for any $\gamma \in (0, 1)$ with probability 1,

$$\binom{\lfloor n\gamma \rfloor}{2}^{-1} \sum_{1 \leq i < j \leq \lfloor n\gamma \rfloor} |Y_i - Y_j|^\beta \rightarrow \mathbb{E} |Y - Y'|^\beta$$

and

$$\binom{n - \lfloor n\gamma \rfloor}{2}^{-1} \sum_{1 \leq i < j \leq n - \lfloor n\gamma \rfloor} |Z_i - Z_j|^\beta \rightarrow \mathbb{E} |Z - Z'|^\beta.$$

Next for any $M > 0$, write

$$|y - z|^\beta = |y - z|^\beta \mathbf{1}\{|y| \leq M, |z| \leq M\} + |y - z|^\beta \mathbf{1}\{|y| \leq M, |z| > M\} \\ + |y - z|^\beta \mathbf{1}\{|y| > M, |z| \leq M\}.$$

By the strong law of large numbers for generalized U-statistics given in Theorem 1 of [18], for any $M > 0$, with probability 1,

$$\frac{2}{\lfloor n\gamma \rfloor (n - \lfloor n\gamma \rfloor)} \sum_{i=1}^{\lfloor n\gamma \rfloor} \sum_{j=1}^{n - \lfloor n\gamma \rfloor} |Y_i - Z_j|^\beta \mathbf{1}\{|Y_i| \leq M, |Z_j| \leq M\} \\ \rightarrow 2\mathbb{E}\left(|Y - Z|^\beta \mathbf{1}\{|Y| \leq M, |Z| \leq M\}\right).$$

Also, observe that

$$\frac{2}{\lfloor n\gamma \rfloor (n - \lfloor n\gamma \rfloor)} \sum_{i=1}^{\lfloor n\gamma \rfloor} \sum_{j=1}^{n - \lfloor n\gamma \rfloor} |Y_i - Z_j|^\beta \mathbf{1}\{|Y_i| \leq M, |Z_j| > M\} \\ \leq \frac{2}{\lfloor n\gamma \rfloor (n - \lfloor n\gamma \rfloor)} \sum_{i=1}^{\lfloor n\gamma \rfloor} \sum_{j=1}^{n - \lfloor n\gamma \rfloor} (M + |Z_j|)^\beta \mathbf{1}\{|Z_j| > M\} \\ = \frac{2}{n - \lfloor n\gamma \rfloor} \sum_{j=1}^{n - \lfloor n\gamma \rfloor} (M + |Z_j|)^\beta \mathbf{1}\{|Z_j| > M\}.$$

By the c_r -inequality and usual law of large numbers for each $M > 0$, with probability 1,

$$\frac{2}{n - \lfloor n\gamma \rfloor} \sum_{j=1}^{n - \lfloor n\gamma \rfloor} (M + |Z_j|)^\beta \mathbf{1}\{|Z_j| > M\} \\ \rightarrow 2\mathbb{E}\left((M + |Z|)^\beta \mathbf{1}\{|Z| > M\}\right) \\ \leq 2^{\beta+2}\mathbb{E}\left(|Z|^\beta \mathbf{1}\{|Z| > M\}\right).$$

Thus for all $M > 0$

$$\limsup_{n \rightarrow \infty} \frac{2}{\lfloor n\gamma \rfloor (n - \lfloor n\gamma \rfloor)} \sum_{i=1}^{\lfloor n\gamma \rfloor} \sum_{j=1}^{n - \lfloor n\gamma \rfloor} |Y_i - Z_j|^\beta \mathbf{1}\{|Y_i| \leq M, |Z_j| > M\} \\ \leq 2^{\beta+2}\mathbb{E}\left(|Z|^\beta \mathbf{1}\{|Z| > M\}\right).$$

In the same way we get that, with probability 1,

$$\limsup_{n \rightarrow \infty} \frac{2}{\lfloor n\gamma \rfloor (n - \lfloor n\gamma \rfloor)} \sum_{i=1}^{\lfloor n\gamma \rfloor} \sum_{j=1}^{n - \lfloor n\gamma \rfloor} |Y_i - Z_j|^\beta \mathbf{1}\{|Y_i| > M, |Z_j| \leq M\} \\ \leq 2\mathbb{E}\left((|Y| + M)^\beta \mathbf{1}\{|Y| > M\}\right) \leq 2^{\beta+2}\mathbb{E}\left(|Y|^\beta \mathbf{1}\{|Y| > M\}\right).$$

Obviously, as $M \rightarrow \infty$,

$$2\mathbb{E}\left(|Y - Z|^\beta \mathbf{1}\{|Y| \leq M, |Z| \leq M\}\right) \rightarrow 2\mathbb{E}|Y - Z|^\beta$$

and

$$2^{\beta+2}\mathbb{E}\left(|Y|^\beta \mathbf{1}\{|Y| > M\}\right) + 2^{\beta+2}\mathbb{E}\left(|Z|^\beta \mathbf{1}\{|Z| > M\}\right) \rightarrow 0.$$

Putting everything together we get that (14) holds. \square

II. A technical result

Let X and X' be i.i.d. F and let φ be a symmetric measurable function from $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\mathbb{E}\varphi^2(X, X') < \infty$. Recall the notation (7). Let A be the operator defined on $L_2(\mathbb{R}^d, F)$ as in (9).

Notice that

$$\mathbb{E}(g(X)\tilde{h}_2(X, X')g(X')) = \int_{\mathbb{R}^d} g(x)Ag(x)dF(x) =: \langle g, Ag \rangle.$$

Let us now introduce some useful definitions. Given $\beta \in (0, 2)$ and $\varphi_\beta(x, y) = |x - y|^\beta$, define as in (7),

$$h_{2,\beta}(x, y) = \varphi_\beta(x, y) - \varphi_{1,\beta}(x) - \varphi_{1,\beta}(y) \quad \text{and} \quad \tilde{h}_{2,\beta}(x, y) = h_{2,\beta}(x, y) + \mathbb{E}\varphi_\beta(X, X').$$

The aim here is to verify that the function $\tilde{h}_{2,\beta}(x, y)$ satisfies the conditions of Theorem II.1 as long as

$$\mathbb{E}|X|^{2\beta} < \infty. \quad (22)$$

Let \tilde{A}_β denote the integral operator

$$\tilde{A}_\beta g(x) = \int_{\mathbb{R}^d} \tilde{h}_{2,\beta}(x, y)g(y)dF(y), \quad x \in \mathbb{R}^d, \quad g \in L_2(\mathbb{R}^d, F).$$

Clearly (22) implies (6) with $\varphi = \varphi_\beta$, which, in turn, by (11) implies

$$\mathbb{E}\tilde{h}_{2,\beta}^2(X, X') = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{h}_{2,\beta}^2(x, y)dF(x)dF(y) = \sum_{i=1}^{\infty} \lambda_i^2 < \infty, \quad (23)$$

where λ_i , $i \geq 1$, are the eigenvalues of the operator \tilde{A}_β , with corresponding orthonormal eigenfunctions ϕ_i , $i \geq 1$.

Next we shall prove that when (22) holds then the eigenvalues λ_i , $i \geq 1$, of this integral operator \tilde{A}_β satisfy (12). This is summarized in the following lemma, whose proof is postponed to the next paragraph.

Lemma VII.1 *Whenever for some $\beta \in (0, 2)$, (22) holds, the eigenvalues λ_i , $i \geq 1$, of the operator \tilde{A}_β satisfy (12).*

The technical results that follow will imply that $\lambda_i \leq 0$ for all $i \geq 1$ and $\sum_{i=1}^{\infty} \lambda_i$ is finite, from which we can infer (12), and thus Lemma VII.1. Let us begin with two definitions.

Definition VII.1 *Let \mathcal{X} be a nonempty set. A symmetric function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called positive definite if*

$$\sum_{i=1}^n \sum_{j=1}^n c_i c_j K(x_i, x_j) \geq 0$$

for all $n \geq 1$, $c_1, \dots, c_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \mathcal{X}$.

Definition VII.2 *Let \mathcal{X} be a nonempty set. A symmetric function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called conditionally negative definite if*

$$\sum_{i=1}^n \sum_{j=1}^n c_i c_j K(x_i, x_j) \leq 0$$

for all $n \geq 1$, $c_1, \dots, c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i = 0$ and $x_1, \dots, x_n \in \mathcal{X}$.

Next, we shall be using part of Lemma 2.1 on page 74 of [1], which we state here for convenience as Lemma VII.2.

Lemma VII.2 *Let K be a symmetric function on $\mathcal{X} \times \mathcal{X}$. Then, for any $x_0 \in \mathcal{X}$, the function*

$$\tilde{K}(x, y) = K(x, x_0) + K(y, x_0) - K(x, y) - K(x_0, x_0)$$

is positive definite if and only if K is conditionally negative definite.

The following lemma can be proved just as Corollary 2.1 in [9].

Lemma VII.3 *Let $H : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a symmetric positive definite function in the sense of Definition VII.1. Assume that H is continuous and $\mathbb{E}H^2(X, X') < \infty$, where X and X' are i.i.d. F . Then $\mathbb{E}(g(X)H(X, X')g(X')) \geq 0$ for all $g \in L_2(\mathbb{R}^d, F)$, i.e., H is L^2 -positive definite in the sense of [9].*

We recall that an operator L on $L_2(\mathbb{R}^d, F)$ is called positive definite if for all $g \in L_2(\mathbb{R}^d, F)$, $\langle g, Lg \rangle \geq 0$.

Proposition VII.1 *Let $\varphi : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a symmetric continuous function that is a conditionally negative definite function in the sense of Definition VII.2. Assume that $\varphi(x, x) = 0$ for all $x \in \mathbb{R}^d$ and $\mathbb{E}\varphi^2(X, X') < \infty$. Then φ defines a positive definite operator L on $L_2(\mathbb{R}^d, F)$ given by*

$$Lg(x) = - \int_{\mathbb{R}^d} h(x, y)g(y)dF(y), \quad x \in \mathbb{R}^d, g \in L_2(\mathbb{R}^d, F),$$

where h is defined as in (7). Furthermore the operator \tilde{L} on $L_2(\mathbb{R}^d, F)$ given by

$$\tilde{L}g(x) = - \int_{\mathbb{R}^d} (h(x, y) + \mathbb{E}\varphi(X, X'))g(y)dF(y), \quad x \in \mathbb{R}^d, g \in L_2(\mathbb{R}^d, F),$$

is also a positive definite operator on $L_2(\mathbb{R}^d, F)$.

Proof. We must show that for all $g \in L_2(\mathbb{R}^d, F)$,

$$\langle g, Lg \rangle = -\mathbb{E}(g(X)h(X, X')g(X')) \geq 0.$$

For any $u \in \mathbb{R}^d$, let us write

$$\begin{aligned} \varphi(x, y, u) &:= \varphi(x, u) + \varphi(y, u) - \varphi(u, u) - \varphi(x, y) \\ &= \varphi(x, u) + \varphi(y, u) - \varphi(x, y). \end{aligned}$$

Since φ is assumed to be conditionally negative definite, by Lemma VII.2 we have that for any fixed $u \in \mathbb{R}^d$, $\varphi(x, y, u)$ is positive definite in the sense of Definition VII.1. Hence, since φ is also assumed to be continuous, by Lemma VII.3 for all $g \in L_2(\mathbb{R}^d, F)$,

$$\mathbb{E}(g(X)\varphi(X, X', u)g(X')) \geq 0.$$

Noting that if U has distribution function F , $\mathbb{E}\varphi(x, y, U) = -h(x, y)$, we get, assuming that X , X' and U are independent, that

$$\mathbb{E}(g(X)\varphi(X, X', U)g(X')) = -\mathbb{E}(g(X)h(X, X')g(X')) \geq 0.$$

Next, notice that for any eigenvalue and normalized eigenfunction $(\tilde{\lambda}_i, \tilde{\phi}_i)$ pair, $i \geq 1$, of the operator \tilde{L} , we have

$$\tilde{\lambda}_i \tilde{\phi}_i(x) = \tilde{L} \tilde{\phi}_i(x) = - \int_{\mathbb{R}^d} (h(x, y) + \mathbb{E} \varphi(X, X')) \tilde{\phi}_i(y) dF(y).$$

Now,

$$\int_{\mathbb{R}^d} (h(x, y) + \mathbb{E} \varphi(X, X')) dF(y) = 0, \text{ for all } x \in \mathbb{R}^d,$$

implies that $(\tilde{\lambda}_1, \tilde{\phi}_1) := (0, 1)$ is an eigenvalue and normalized pair of \tilde{L} . From this we get that whenever $\tilde{\lambda}_i \neq 0$, $\mathbb{E} \tilde{\phi}_i(X) = 0$, $i \geq 2$, which says that for such $\tilde{\lambda}_i$,

$$\tilde{\lambda}_i \tilde{\phi}_i(x) = - \int_{\mathbb{R}^d} h(x, y) \tilde{\phi}_i(y) dF(y).$$

This implies that whenever for some $i \geq 1$, $(\tilde{\lambda}_i, \tilde{\phi}_i)$, with $\tilde{\lambda}_i \neq 0$, is an eigenvalue and normalized eigenfunction pair of the operator \tilde{L} , it is also an eigenvalue and normalized eigenfunction pair of the operator L . Moreover, since the integral operator L is positive definite on $L_2(\mathbb{R}^d, F)$, this implies that for any such nonzero $\tilde{\lambda}_i$ (where necessarily $i \geq 2$)

$$\begin{aligned} & - \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{\phi}_i(x) (h(x, y) + \mathbb{E} \varphi(X, X')) \tilde{\phi}_i(y) dF(x) dF(y) \\ & = - \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{\phi}_i(x) h(x, y) \tilde{\phi}_i(y) dF(x) dF(y) = \tilde{\lambda}_i \geq 0, \end{aligned}$$

which says that the operator \tilde{L} is positive definite on $L_2(\mathbb{R}^d, F)$. \square

III. Proof of Lemma VII.1

A special case of Theorem 3.2.2 in [1] says that the function $\varphi_\beta(x, y) = |x - y|^\beta$, $\beta \in (0, 2)$, is conditionally negative definite. Also see Exercise 3.2.13b in [1] and the discussion after Proposition 3 in [24]. Therefore by Proposition VII.1 the integral operator L_β defined by the function

$$K_\beta(x, y) = -h_{2, \beta}(x, y)$$

is positive definite as well as the integral operator $\tilde{L}_\beta = -\tilde{A}_\beta$ defined by the function

$$\tilde{K}_\beta(x, y) = -h_{2, \beta}(x, y) - \mathbb{E} \varphi_\beta(X, X').$$

Next, as in the proof of Proposition VII.1, any eigenvalue and normalized eigenfunction

$$(\tilde{\lambda}_i, \tilde{\phi}_i) = (-\lambda_i, -\phi_i)$$

pair, with $\tilde{\lambda}_i \neq 0$, $i \geq 1$, of the operator $\tilde{L}_\beta = -\tilde{A}_\beta$ is also an eigenvalue and normalized eigenfunction pair of the operator L_β .

We shall apply Theorem 2 of [22] to show that uniformly on compact subsets D of \mathbb{R}^d ,

$$K_\beta(x, y) = \sum_{i=1}^{\infty} \rho_i \psi_i(x) \psi_i(y), \quad (x, y) \in D \times D,$$

where $\rho_i \geq 0$, $i \geq 1$, are the eigenvalues of the operator $L_\beta = -A_\beta$ with corresponding normalized eigenfunctions ψ_i , $i \geq 1$. In particular

$$K_\beta(x, x) = \sum_{i=1}^{\infty} \rho_i \psi_i^2(x), x \in D,$$

and thus since $\mathbb{E}\psi_i^2(X) = 1$, $i \geq 1$, and $\mathbb{E}K_\beta(X, X) < \infty$, we get

$$\sum_{i=1}^{\infty} \rho_i < \infty.$$

Therefore since, as pointed out above, the eigenvalue and normalized pairs $(-\lambda_i, -\phi_i)$ of $\tilde{L}_\beta = -\tilde{A}_\beta$, with $\lambda_i \neq 0$, are also eigenvalue and normalized pairs of the operator L_β this implies that $\sum_{i=1}^{\infty} |\lambda_i| < \infty$.

Our proof will be complete once we have checked that L_β satisfies the conditions of Theorem 2 of [22].

Since the function $\varphi_\beta(x, y) = |x - y|^\beta$, $\beta \in (0, 2)$, is conditionally negative definite, by Lemma VII.2 the function $K_\beta(x, y)$ is positive definite. To see this note that by Lemma VII.2 for any fixed $u \in \mathbb{R}$ the function

$$\varphi_\beta(x, u) + \varphi_\beta(u, y) - \varphi_\beta(x, y) - \varphi_\beta(u, u) = \varphi_\beta(x, u) + \varphi_\beta(u, y) - \varphi_\beta(x, y)$$

is positive definite. Therefore we readily see that

$$K_\beta(x, y) = \left(\int_{\mathbb{R}^d} \varphi_\beta(x, u) + \varphi_\beta(u, y) - \varphi_\beta(x, y) \right) dF(u)$$

is positive definite. In addition, $K_\beta(x, y)$ is symmetric and continuous, and thus $K_\beta(x, y)$ is a Mercer kernel in the terminology of [22]. We must also verify the following assumptions.

Assumption A. For each $x \in \mathbb{R}^d$, $K_\beta(x, \cdot) \in L_2(\mathbb{R}^d, F)$.

Assumption B. L_β is a bounded and positive definite operator on $L_2(\mathbb{R}^d, F)$ and for every $g \in L_2(\mathbb{R}^d, F)$, the function

$$L_\beta g(x) = \int_{\mathbb{R}^d} K_\beta(x, y) g(y) dF(y)$$

is a continuous function on \mathbb{R}^d .

Assumption C. L_β has at most countably many positive eigenvalues and orthonormal eigenfunctions.

Since φ_β is a symmetric continuous function that is conditionally negative definite in the sense of Definition VII.2 satisfying $\varphi_\beta(x, x) = 0$ for all $x \in \mathbb{R}^d$ and $\mathbb{E}\varphi_\beta^2(X, X') < \infty$, we get by Proposition VII.1 that L_β is a positive definite operator on $L_2(\mathbb{R}^d, F)$. Also (22) obviously implies that Assumption A holds and

$$\mathbb{E}K_\beta^2(X, X') = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} K_\beta^2(x, y) dF(x) dF(y) < \infty, \quad (24)$$

which by Proposition 1 of [22] implies that the operator L_β is bounded and compact. (From Sun's Proposition 1 one can also infer that L_β is positive definite. However, he does not provide a proof. Therefore we invoke our Lemma VII.3 here.) An elementary argument based on the dominated convergence theorem implies that $L_\beta g(x)$ is a continuous function on \mathbb{R}^d . Thus Assumption B is satisfied. Finally, since the operator L_β is compact, Theorem VII.4.5 of [8] implies that Assumption C is fulfilled. Thus the assumptions of Theorem 2 of [22] hold. This completes the proof of Lemma VII.1. \square

REFERENCES

- [1] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, New York, 1984.
- [2] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, New York, 1968.
- [3] K. Bleakley and J.-P. Vert. The group fused lasso for multiple change–point detection. *arXiv:1106.4199*, 2011.
- [4] R. Bolton and D. Hand. Statistical fraud detection: A review. *Statistical Science*, 17:235–255, 2002.
- [5] B. P. Carlin, A. E. Gelfand, and A. F. Smith. Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics*, 41:389–405, 1992.
- [6] M. Csörgő and L. Horváth. *Limit Theorems in Change-Point Analysis*. Wiley, New York, 1997.
- [7] M. Csörgő and L. Horváth. Invariance principles for changepoint problems. *Journal of Multivariate Analysis*, 27:151–168, 1988.
- [8] N. Dunford and J. T. Schwartz. *Linear Operators*. Wiley, New York, 1963.
- [9] J. C. Ferreira and V. A. Menegatto. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, 64:61–81, 2009.
- [10] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19:293–325, 1948.
- [11] W. Hoeffding. The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23:169–192, 1952.
- [12] A. Kim, C. Marzban, D. Percival, and W. Stuetzie. Using labeled data to evaluate change detectors in a multivariate streaming environment. *Signal Processing*, 89:2529–2536, 2009.
- [13] V. Koltchinskii and E. Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6:113–167, 2000.
- [14] M. Lavielle and G. Teyssière. Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46:287–306, 2006.
- [15] D. S. Matteson and N. A. James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109:334–345, 2014.
- [16] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6:27, 2005.
- [17] J. P. Romano and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100:94–108, 2005.
- [18] P.K. Sen. Almost sure convergence of generalized U-statistics. *Ann. Probab.*, 5:287–290, 1977.

-
- [19] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, 1980.
- [20] S.P. Shah, W.L. Lam, R.T. Ng, and K.P. Murphy. Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics*, 23:i450–i458, 2007.
- [21] N. Stransky, C. Vallot, F. Reyat, I. Bernard-Pierrot, S. G. D. de Medina, R. Segraves, Y. de Rycke, P. Elvin, A. Cassidy, C. Spraggon, A. Graham, J. Southgate, B. Asselain, Y. Allory, C. C. Abbou, D. G. Albertson, J.-P. Thiery, D. K. Chopin, D. Pinkel, and F. Radvanyi. Regional copy number-independent deregulation of transcription in cancer. *Nature Genetics*, 38:1386–1396, 2006.
- [22] H. Sun. Mercer theorem for RKHS on concompact sets. *Journal of Complexity*, 21:337–349, 2005.
- [23] G. J. Székely and M. L. Rizzo. Hierarchical clustering via joint between-within distances: Extending Ward’s minimum variance method. *Journal of Classification*, 22(2):151–183, 2005.
- [24] G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143:1249–1272, 2013.
- [25] M. Talih and N. Hengartner. Structural learning with time-varying components: Tracking the cross-section of financial time series. *Journal of the Royal Statistical Society: Series B*, 67:321–341, 2005.