Nonparametric tests for change-point detection à la Gombay and Horváth

Ivan Kojadinovic

Université de Pau et des Pays de l'Adour, France

Based on joint work with Mark Holmes and Jun Yan.

Pau, January 2013

Outline

1. Introduction

2. Theoretical results for change-point detection

3. Tests for change-point detection à la Gombay and Horváth

4. Practical recommendations and illustration

Outline

1. Introduction

2. Theoretical results for change-point detection

3. Tests for change-point detection à la Gombay and Horváth

4. Practical recommendations and illustration

Introduction I

- Let $X_1, \dots, X_n$ be a sequence of independent d-dimensional random vectors for some fixed integer $d \geqslant 1$. The aim of this work is to study, both theoretically and empirically, nonparametric tests for the detection of a change-point in the sequence $X_1, \dots, X_n$.
- The corresponding null hypothesis is

$$H_0 : \exists\, P_0 \text{ such that } X_1, \dots, X_n \text{ have law } P_0. \qquad (1)$$

- As frequently done, the behavior of the derived tests will be investigated under the alternative hypothesis of a single change-point:

$$\begin{aligned} H_1 : \exists \text{ distinct } P_1 \text{ and } P_2, \text{ and } k^\star \in \{1, \dots, n-1\} \text{ such that} \\ X_1, \dots, X_{k^\star} \text{ have law } P_1 \text{ and } X_{k^\star+1}, \dots, X_n \text{ have law } P_2. \end{aligned} \qquad (2)$$

- There exists an abundant literature on nonparametric tests for change-point detection. We shall not review here procedures designed for serially dependent observations.

Introduction II

- The approaches proposed for sequences of independent observations differ, on one hand, according to the test statistic, and on the other hand, according to the resampling technique used to compute an approximate *p*-value for the test statistic.
- In terms of the test statistic, two frequently encountered classes of approaches are those based on *U*-statistics (see e.g. Csörgő and Horváth, 1988; Ferger, 1994; Gombay and Horváth, 2002; Horváth and Hušková, 2005) and those based on empirical c.d.f.s (see e.g. Gombay and Horváth, 1999; Horváth and Shao, 2007).
- As far as the resampling technique is concerned, one finds approaches based on permutations of the original sequence (see e.g. Antoch and Hušková, 2001; Horváth and Hušková, 2005; Horváth and Shao, 2007) and approaches that use a weighted bootstrap based on multiplier central limit theorems (see e.g. Gombay and Horváth, 1999, 2002).

Introduction III

- For a broader presentation of the field of change-point analysis, we refer the reader to the monographs by Brodsky and Darkhovsky (1993) and Csörgő and Horváth (1997).
- In this work, we revisit and extend the approach proposed by Gombay and Horváth (1999) based on the test statistic

$$T_{n,\vee} = \max_{1 \leqslant k \leqslant n-1} \frac{k(n-k)}{n^{3/2}} \sup_{x \in \mathbb{R}^d} \left| F_k(x) - F^{\star}_{n-k}(x) \right|,$$

where

$$F_k(x) = \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}(X_i \leqslant x) \qquad \text{and} \qquad F^{\star}_{n-k}(x) = \frac{1}{n-k} \sum_{i=k+1}^{n} \mathbf{1}(X_i \leqslant x)$$

are the empirical c.d.f.s computed from $X_1, \ldots, X_k$ and $X_{k+1}, \ldots, X_n$, respectively (see also Csörgő and Horváth, 1997, Section 2.6).

Introduction IV

- From a theoretical perspective, we work in the framework of the theory of empirical processes as presented for instance in van der Vaart and Wellner (2000) and Kosorok (2008).
- To obtain results that are valid for many different **classes of functions** (in the sense of empirical process theory), we first extend the multiplier central limit theorem (see e.g. Kosorok, 2008, Theorem 10.1 and Corollary 10.3) to the sequential setting.
- From a more practical perspective, we consider a large number of candidate test statistics based on processes indexed by **lower-left orthants** and by **half-spaces**, and we study the finite-sample performance of the corresponding tests through extensive Monte Carlo experiments involving univariate, bivariate and trivariate data sets.
- As we shall see, in the multivariate case, the tests based on processes indexed by half-spaces appear to be substantially more powerful than more classical tests based on multivariate empirical c.d.f.s (i.e., based on processes indexed by lower-left orthants).

Introduction V

- The last section of this presentation contains practical recommendations and presents an application of the studied tests to trivariate hydrological data.
- Note finally that the code of all the tests studied in this work will be documented and released as an R package whose tentative name is `npcp`.

Outline

1. Introduction

2. Theoretical results for change-point detection

3. Tests for change-point detection à la Gombay and Horváth

4. Practical recommendations and illustration

Notation and setting I

- All the random variables used in this work are defined with respect to the underlying probability space $(\Omega, \mathcal{G}, \mathbf{P})$ and the outer probability measure corresponding to $\mathbf{P}$ is denoted by $\mathbf{P}^*$.
- Let $X_1, \dots, X_n$ be i.i.d. d-dimensional random vectors with law P, and let $\mathcal{F}$ be a class of measurable functions from $\mathbb{R}^d$ to $\mathbb{R}$. The empirical measure is defined to be $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, where δ_x is the measure that assigns a mass of 1 at x and zero elsewhere.
- For $f \in \mathcal{F}$, $\mathbb{P}_n f$ denotes the expectation of f under $\mathbb{P}_n$, and Pf the expectation under P, i.e.,

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^n f(X_i) \qquad \text{and} \qquad Pf = \int f \mathrm{d}P.$$

- The empirical process evaluated at f is then defined as $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf)$.

Notation and setting II

- Saying that $\mathcal{F}$ is P-Donsker means that the sequence of processes $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges weakly to a P-Brownian bridge $\{\mathbb{G}_P f : f \in \mathcal{F}\}$ in the space $\ell^\infty(\mathcal{F})$ of bounded functions from $\mathcal{F}$ to $\mathbb{R}$ equipped with the uniform metric in the sense of Definition 1.3.3 of van der Vaart and Wellner (2000).
- Following usual notational conventions, this weak convergence will simply be denoted by $\mathbb{G}_n \rightsquigarrow \mathbb{G}_P$ in $\ell^\infty(\mathcal{F})$. Furthermore, we say that $F_e : \mathbb{R}^d \to \mathbb{R}$ is an envelope for $\mathcal{F}$ if F_e is measurable and $|f(x)| \leqslant F_e(x)$ for every $f \in \mathcal{F}$ and $x \in \mathbb{R}^d$.
- The advantage of working in this general framework is that the forthcoming results remain valid for many P-Donsker classes $\mathcal{F}$.
- By taking $\mathcal{F}$ to be the class of indicator functions of lower-left orthants in $\mathbb{R}^d$, i.e., $\mathcal{F} = \{y \mapsto \mathbf{1}(y \leqslant x) : x \in \overline{\mathbb{R}}^d\}$ with $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, one recovers the setting studied in Csörgő and Horváth (1997, Section 2.6) and based on empirical cumulative distribution functions (c.d.f.s).

Notation and setting III

- Although this is a natural choice for $\mathcal{F}$, many other choices might be of interest in practice such as the class of indicator functions of closed balls, rectangles or half-spaces (see Romano, 1988, for a related discussion regarding the choice of $\mathcal{F}$).

A multiplier central limit theorem for the sequential empirical process I

- The sequential empirical process is defined as

$$\mathbb{Z}_n(s,f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor ns \rfloor} \{f(X_i) - Pf\} = \sqrt{\lambda_n(s)}\,\mathbb{G}_{\lfloor ns \rfloor} f, \qquad s \in [0,1], f \in \mathcal{F},$$

  where $\lambda_n(s) = \lfloor ns \rfloor / n$ and with the convention that $\mathbb{P}_0 f = 0$ for all $f \in \mathcal{F}$.

- According to Theorem 2.12.1 of van der Vaart and Wellner (2000), $\mathcal{F}$ being P-Donsker is equivalent to $\mathbb{Z}_n \rightsquigarrow \mathbb{Z}_P$ in $\ell^\infty([0,1] \times \mathcal{F})$, where $\mathbb{Z}_P$ is a tight centered mean-zero Gaussian process with covariance function

$$\mathrm{cov}\{\mathbb{Z}_P(s,f), \mathbb{Z}_P(t,g)\} = (s \wedge t)(Pfg - PfPg)$$

  known as a ***P*-Kiefer-Müller** process.

A multiplier central limit theorem for the sequential empirical process II

- Given i.i.d. random variables $\xi_1, \dots, \xi_n$ with mean 0 and variance 1, satisfying $\int_0^\infty \{\mathbf{P}(|\xi_1| > x)\}^{1/2} \mathrm{d}x < \infty$, and independent of the random sample $X_1, \dots, X_n$, we define the following **multiplier** version of $\mathbb{Z}_n$:

$$\widetilde{\mathbb{Z}}_n(s, f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor ns \rfloor} \xi_i \{f(X_i) - Pf\}, \qquad s \in [0, 1], f \in \mathcal{F}.$$

- Notice that the empirical process $\widetilde{\mathbb{Z}}_n$ depends on the unknown map $f \mapsto Pf$ and therefore cannot be computed.

A multiplier central limit theorem for the sequential empirical process III

- With applications in mind, we define two versions of $\widetilde{\mathbb{Z}}_n$ (depending on how $f \mapsto Pf$ is estimated) that can be fully computed. For any $s \in [0,1], f \in \mathcal{F}$, let

$$\widehat{\mathbb{Z}}_n(s,f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor ns \rfloor} \xi_i \{f(X_i) - \mathbb{P}_{\lfloor ns \rfloor} f\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor ns \rfloor} (\xi_i - \bar{\xi}_{\lfloor ns \rfloor}) f(X_i),$$

where $\bar{\xi}_{\lfloor ns \rfloor} = \lfloor ns \rfloor^{-1} \sum_{i=1}^{\lfloor ns \rfloor} \xi_i$ and $\bar{\xi}_0 = 0$ by convention, and let

$$\check{\mathbb{Z}}_n(s,f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor ns \rfloor} \xi_i \{f(X_i) - \mathbb{P}_n f\}.$$

A multiplier central limit theorem for the sequential empirical process IV

- The following result is then a partial extension of the multiplier central limit theorem (see e.g. Kosorok, 2008, Theorem 10.1 and Corollary 10.3) to the sequential setting.

Theorem

Let $\mathcal{F}$ be a P-Donsker class with measurable envelope F_e such that $PF_e^2 < \infty$. Then, $(\mathbb{Z}_n, \tilde{\mathbb{Z}}_n, \hat{\mathbb{Z}}_n, \check{\mathbb{Z}}_n) \rightsquigarrow (\mathbb{Z}_P, \mathbb{Z}'_P, \mathbb{Z}'_P, \mathbb{Z}'_P)$ in $\{\ell^\infty([0,1] \times \mathcal{F})\}^4$, where $\mathbb{Z}'_P$ is an independent copy of $\mathbb{Z}_P$.

- Theorem 1 suggests the following interpretation: when n is large, $\tilde{\mathbb{Z}}_n$ can be regarded as “almost” an independent copy of $\mathbb{Z}_n$, while $\hat{\mathbb{Z}}_n$ and $\check{\mathbb{Z}}_n$ can be regarded as computable copies of $\tilde{\mathbb{Z}}_n$. As we shall see, this interpretation is at the root of the resampling technique considered later.

A multiplier central limit theorem for the sequential empirical process V

- Although each of $\widehat{\mathbb{Z}}_n$ and $\check{\mathbb{Z}}_n$ could be regarded as "almost" an independent copy of $\mathbb{Z}_n$, their behavior for moderate n might differ quite substantially.
- In our Monte Carlo experiments, we empirically investigate which of $\widehat{\mathbb{Z}}_n$ or $\check{\mathbb{Z}}_n$ leads to tests for change-point detection with the best finite-sample properties.

Application to change-point detection I

- Recall that the null and alternative hypotheses under consideration are given in (1) and (2), respectively.
- Let $\mathcal{F}$ be a class of measurable functions. In order to test the aforementioned hypotheses, we extend the approach studied in detail by Csörgő and Horváth (1997, Section 2.6) and compare, for all $k \in \{1, \dots, n-1\}$,

$$\mathbb{P}_k f = \frac{1}{k} \sum_{i=1}^{k} f(X_i) \qquad \text{and} \qquad \mathbb{P}^{\star}_{n-k} f = \frac{1}{n-k} \sum_{i=k+1}^{n} f(X_i), \qquad f \in \mathcal{F}.$$

- Analogous to Csörgő and Horváth (1997, Section 2.6), we define the process

$$\mathbb{D}_n(s,f) = \sqrt{n}\, \lambda_n(s) \{1 - \lambda_n(s)\} \left(\mathbb{P}_{\lfloor ns \rfloor} f - \mathbb{P}^{\star}_{n-\lfloor ns \rfloor} f \right), \; s \in [0,1], f \in \mathcal{F},$$

where $\lambda_n(s) = \lfloor ns \rfloor / n$ and with the convention that $\mathbb{P}_0 f = 0$ and $\mathbb{P}^{\star}_0 f = 0$ for all $f \in \mathcal{F}$.

Application to change-point detection II

- Notice that, under the null hypothesis, for any $s \in [0,1]$ and $f \in \mathcal{F}$, we have

$$\begin{aligned}\mathbb{D}_n(s,f) &= \{1-\lambda_n(s)\}\,\mathbb{Z}_n(s,f) - \lambda_n(s)\{\mathbb{Z}_n(1,f) - \mathbb{Z}_n(s,f)\} \\ &= \mathbb{Z}_n(s,f) - \lambda_n(s)\mathbb{Z}_n(1,f). \qquad (3)\end{aligned}$$

- With resampling in mind, we define two **multiplier** versions of $\mathbb{D}_n$ based on the multiplier versions of $\mathbb{Z}_n$ defined in the previous subsection. For any $s \in [0,1]$ and $f \in \mathcal{F}$, let

$$\begin{aligned}\check{\mathbb{D}}_n(s,f) &= \{1-\lambda_n(s)\}\check{\mathbb{Z}}_n(s,f) - \lambda_n(s)\{\check{\mathbb{Z}}_n(1,f) - \check{\mathbb{Z}}_n(s,f)\} \\ &= \check{\mathbb{Z}}_n(s,f) - \lambda_n(s)\check{\mathbb{Z}}_n(1,f),\end{aligned}$$

and, following Gombay and Horváth (1999), let

$$\hat{\mathbb{D}}_n(s,f) = \{1-\lambda_n(s)\}\hat{\mathbb{Z}}_n(s,f) - \lambda_n(s)\hat{\mathbb{Z}}_n^{\star}(s,f), \qquad (4)$$

Application to change-point detection III

where

$$\widehat{\mathbb{Z}}_n^\star(s,f) = \frac{1}{\sqrt{n}} \sum_{i=\lfloor ns \rfloor+1}^{n} (\xi_i - \bar{\xi}^\star_{n-\lfloor ns \rfloor}) f(X_i) \tag{5}$$

with

$$\bar{\xi}^\star_{n-\lfloor ns \rfloor} = \frac{1}{n-\lfloor ns \rfloor} \sum_{i=\lfloor ns \rfloor+1}^{n} \xi_i,$$

and $\bar{\xi}^\star_0 = 0$ by convention.

- Notice that the process $\widehat{\mathbb{Z}}_n^\star$ defined above is, up to a small error term vanishing as $n \to \infty$, the version of the process $(s,f) \mapsto \widehat{\mathbb{Z}}_n(1-s,f)$ computed from the “reversed” sequence $(\xi_n, X_n), (\xi_{n-1}, X_{n-1}), \dots, (\xi_1, X_1)$.

Application to change-point detection IV

- The following two results extend Theorems 2.1, 2.2 and 2.3 of Gombay and Horváth (1999) and suggest, for large n and under the null hypothesis, to interpret each of $\widehat{\mathbb{D}}_n$ and $\check{\mathbb{D}}_n$ as an "almost" independent copy of $\mathbb{D}_n$.

Theorem

Assume that H_0 holds and that $\mathcal{F}$ is P_0-Donsker with measurable envelope F_e such that $P_0 F_e^2 < \infty$. Then, $(\mathbb{D}_n, \widehat{\mathbb{D}}_n, \check{\mathbb{D}}_n) \rightsquigarrow (\mathbb{D}_{P_0}, \mathbb{D}'_{P_0}, \mathbb{D}'_{P_0})$ in $\{\ell^\infty([0,1] \times \mathcal{F})\}^3$, where $\mathbb{D}_{P_0}(s,f) = \mathbb{Z}_{P_0}(s,f) - s\mathbb{Z}_{P_0}(1,f)$, $s \in [0,1]$, $f \in \mathcal{F}$, and $\mathbb{D}'_{P_0}$ is an independent copy of $\mathbb{D}_{P_0}$.

Application to change-point detection V

- As we continue, for any $J : \mathcal{F} \to \mathbb{R}$, $\sup_{f \in \mathcal{F}} |Jf|$ will be denoted by $\|J\|_{\mathcal{F}}$. Also, for any sequence of maps $Y_1, Y_2, \dots,$ each from Ω to $\mathbb{R}$, we say that the sequence Y_n is bounded in outer probability if, for any $\varepsilon > 0$, there exists $M > 0$ such that $\sup_{n \in \mathbb{N}} \mathbf{P}^* (|Y_n| > M) < \varepsilon$.

Theorem

Assume that H_1 holds with $k^\star = \lfloor nt \rfloor$ for some $t \in (0,1)$ and that $\mathcal{F}$ is P_1 and P_2-Donsker with measurable envelope F_e such that $P_1 F_e^2 < \infty$ and $P_2 F_e^2 < \infty$. Then,

(i) $\sup_{s \in [0,1]} \|n^{-1/2} \mathbb{D}_n(s,f) - K_t(s,f)\|_{\mathcal{F}} \xrightarrow{\mathbf{P}^*} 0$,
where $K_t(s,f) = (P_1 f - P_2 f)(s \wedge t)\{1 - (s \vee t)\}$,

(ii) $\sup_{s \in [0,1]} \|\hat{\mathbb{D}}_n(s,f)\|_{\mathcal{F}}$ *is bounded in outer probability,*

(iii) $\check{\mathbb{D}}_n$ *converges weakly in $\ell^\infty([0,1] \times \mathcal{F})$.*

Application to change-point detection VI

- The previous result will be used in the next section to show that various tests for change-point detection based on $\mathbb{D}_n$ will tend to reject H_0 under H_1 as n increases.

Outline

1. Introduction

2. Theoretical results for change-point detection

3. Tests for change-point detection à la Gombay and Horváth

4. Practical recommendations and illustration

Tests for change-point detection à la Gombay and Horváth I

- The aim of this section is to use the results of the previous section to derive tests for change-point detection in the spirit of those proposed by Gombay and Horváth (1999).
- Among the many possible choices for $\mathcal{F}$, we consider the following two:

  $(\mathcal{C}1)$ the collection $\mathcal{O}$ of indicator functions of lower-left orthants in $\mathbb{R}^d$, where

$$\mathcal{O} = \{f_x(y) = \mathbf{1}(y \leqslant x) : x \in \overline{\mathbb{R}}^d\};$$

Tests for change-point detection à la Gombay and Horváth II

($\mathcal{C}2$) the collection $\mathcal{H}$ of indicator functions of half-spaces in $\mathbb{R}^d$, where

$$\mathcal{H} = \{f_{a,b}(y) = \mathbf{1}(a^\top y \leqslant b) : a \in \mathcal{S}_d, b \in \overline{\mathbb{R}}\}$$

and $\mathcal{S}_d$ is the subset of $\mathbb{R}^d$ composed of vectors with Euclidean norm one.

- It is well-known that lower-left orthants and half-spaces are Vapnik-Čhervonenkis classes of sets. Consequently, $\mathcal{O}$ and $\mathcal{H}$ are P-Donsker for any law P (see e.g. van der Vaart and Wellner, 2000; Kosorok, 2008).
- As we continue, in the case of choice ($\mathcal{C}1$), for any $s \in [0,1]$ and $f_x \in \mathcal{O}$, $\mathbb{D}_n(s, f_x)$, $\widehat{\mathbb{D}}_n(s, f_x)$ and $\check{\mathbb{D}}_n(s, f_x)$ will simply be denoted by $\mathbb{D}_n(s, x)$, $\widehat{\mathbb{D}}_n(s, x)$ and $\check{\mathbb{D}}_n(s, x)$, respectively.

Tests for change-point detection à la Gombay and Horváth III

- Similarly, in the case of choice $(\mathcal{C}2)$, for any $s \in [0,1]$ and $f_{a,b} \in \mathcal{H}$, $\mathbb{D}_n(s, f_{a,b})$, $\hat{\mathbb{D}}_n(s, f_{a,b})$ and $\check{\mathbb{D}}_n(s, f_{a,b})$ will be denoted by $\mathbb{D}_n(s, a, b)$, $\hat{\mathbb{D}}_n(s, a, b)$ and $\check{\mathbb{D}}_n(s, a, b)$, respectively.
- In the framework under consideration, a change in the sequence $X_1, \dots, X_n$ can occur at any point $k \in \{1, \dots, n-1\}$.
- A test for change-point detection could therefore be obtained by first defining a test statistic for any possible change-point $k \in \{1, \dots, n-1\}$, and then by combining the resulting $n-1$ statistics into a global statistic using some function from $\psi : \mathbb{R}^{n-1} \to \mathbb{R}$.

Tests for change-point detection à la Gombay and Horváth IV

- In the case of choice $(\mathcal{C}1)$, two natural possibilities for the $n-1$ change-point statistics are respectively

$$S_{n,k} = \int_{\mathbb{R}^d} \left\{ \mathbb{D}_n \left(\frac{k}{n}, x \right) \right\}^2 \mathrm{d}F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{D}_n \left(\frac{k}{n}, X_i \right) \right\}^2,$$

$k \in \{1, \dots, n-1\}$, where $F_n(x) = \mathbb{P}_n f_x$, $x \in \overline{\mathbb{R}}^d$, is the empirical c.d.f. computed from $X_1, \dots, X_n$, and

$$T_{n,k} = \sup_{x \in \mathbb{R}^d} \left| \mathbb{D}_n \left(\frac{k}{n}, x \right) \right| = \max_{1 \leqslant i \leqslant n} \left| \mathbb{D}_n \left(\frac{k}{n}, X_i \right) \right|, \qquad k \in \{1, \dots, n-1\}.$$

Tests for change-point detection à la Gombay and Horváth V

- Two natural choices for the function ψ are the maximum and the arithmetic mean which leads to the following four global statistics:

$$S_{n,\vee} = \max_{1\leqslant k\leqslant n-1} S_{n,k} = \sup_{s\in[0,1]} \int_{\mathbb{R}^d} \{\mathbb{D}_n(s,x)\}^2 \, \mathrm{d}F_n(x),$$

$$T_{n,\vee} = \max_{1\leqslant k\leqslant n-1} T_{n,k} = \sup_{s\in[0,1]} \sup_{x\in\mathbb{R}^d} |\mathbb{D}_n(s,x)|,$$

$$S_{n,+} = \frac{1}{n}\sum_{k=1}^{n-1} S_{n,k} = \int_0^1 \int_{\mathbb{R}^d} \{\mathbb{D}_n(s,x)\}^2 \, \mathrm{d}F_n(x)\mathrm{d}s,$$

$$T_{n,+} = \frac{1}{n}\sum_{k=1}^{n-1} T_{n,k} = \int_0^1 \sup_{x\in\mathbb{R}^d} |\mathbb{D}_n(s,x)| \, \mathrm{d}s.$$

Note that $T_{n,\vee}$ is the global statistic used in Gombay and Horváth (1999).

Tests for change-point detection à la Gombay and Horváth VI

- In the case of choice $(\mathcal{C}2)$, for any $k \in \{1, \dots, n-1\}$, we first define

$$U_{n,k} = \int_{\mathcal{S}_d^+} \int_{\mathbb{R}} \left\{ \mathbb{D}_n \left(\frac{k}{n}, a, b \right) \right\}^2 \mathrm{d}F_{a,n}(b) \mathrm{d}a$$
$$= \int_{\mathcal{S}_d^+} \frac{1}{n} \sum_{i=1}^n \left\{ \mathbb{D}_n \left(\frac{k}{n}, a, a^\top X_i \right) \right\}^2 \mathrm{d}a,$$

where $\mathcal{S}_d^+ = \{a \in \mathcal{S}_d : a_1 > 0\}$ and, for any $a \in \mathcal{S}_d^+$, $F_{a,n}$ is the (univariate) empirical c.d.f. computed from the projected sample $a^\top X_1, \dots, a^\top X_n$, and

$$V_{n,k} = \sup_{a \in \mathcal{S}_d^+} \sup_{b \in \mathbb{R}} \left| \mathbb{D}_n \left(\frac{k}{n}, a, b \right) \right| = \sup_{a \in \mathcal{S}_d^+} \max_{1 \leqslant i \leqslant n} \left| \mathbb{D}_n \left(\frac{k}{n}, a, a^\top X_i \right) \right|.$$

Tests for change-point detection à la Gombay and Horváth VII

- As in the case of choice $(\mathcal{C}1)$, this leads to four global statistics depending on whether the change-point statistics are combined using the maximum or the arithmetic mean, i.e.,

$$U_{n,\vee} = \max_{1\leqslant k\leqslant n-1} U_{n,k} = \sup_{s\in[0,1]} \int_{\mathcal{S}_d^+} \int_{\mathbb{R}} \{\mathbb{D}_n(s,a,b)\}^2 \,\mathrm{d}F_{a,n}(b)\mathrm{d}a,$$

$$V_{n,\vee} = \max_{1\leqslant k\leqslant n-1} V_{n,k} = \sup_{s\in[0,1]} \sup_{a\in\mathcal{S}_d^+} \sup_{b\in\mathbb{R}} |\mathbb{D}_n(s,a,b)|,$$

$$U_{n,+} = \frac{1}{n}\sum_{k=1}^{n-1} U_{n,k} = \int_0^1 \int_{\mathcal{S}_d^+} \int_{\mathbb{R}} \{\mathbb{D}_n(s,a,b)\}^2 \,\mathrm{d}F_{a,n}(b)\mathrm{d}a\mathrm{d}s,$$

$$V_{n,+} = \frac{1}{n}\sum_{k=1}^{n-1} V_{n,k} = \int_0^1 \sup_{a\in\mathcal{S}_d^+} \sup_{b\in\mathbb{R}} |\mathbb{D}_n(s,a,b)| \,\mathrm{d}s.$$

Tests for change-point detection à la Gombay and Horváth VIII

- In our Monte Carlo experiments, the integral and the supremum over $a \in \mathcal{S}_d^+$ in the definitions of $U_{n,k}$ and $V_{n,k}$, respectively, were approximated numerically based on a uniform discretization of $\mathcal{S}_d^+$ into m points.
- Notice finally that the change-point statistics $S_{n,k}$ and $U_{n,k}$ (resp. $T_{n,k}$ and $V_{n,k}$) coincide when $d = 1$ since $\mathcal{S}_1^+ = \{1\}$.
- Let us now explain how approximate p-values for these statistics can be computed using the multiplier processes $\hat{\mathbb{D}}_n$ and $\check{\mathbb{D}}_n$.
- For the sake of brevity, we present the approach and state the key results only for $S_{n,\vee}$ as the cases of the other seven global statistics are similar.
- Let N be a large integer and let $\xi_i^{(j)}$, $i \in \{1, \dots, n\}$, $j \in \{1, \dots, N\}$, be i.i.d. random variables with mean 0 and variance 1 satisfying $\int_0^\infty \{\mathbf{P}(|\xi_i^{(j)}| > x)\}^{1/2} \mathrm{d}x < \infty$, and independent of $X_1, \dots, X_n$.

Tests for change-point detection à la Gombay and Horváth IX

- Also, for any $j \in \{1, \dots, N\}$, let $\widehat{\mathbb{D}}_n^{(j)}$ (resp. $\check{\mathbb{D}}_n^{(j)}$) denote the version of $\widehat{\mathbb{D}}_n$ (resp. $\check{\mathbb{D}}_n$) computed from $\xi_1^{(j)}, \dots, \xi_n^{(j)}$.
- Moreover, for any $j \in \{1, \dots, N\}$, let

$$\widehat{S}_{n,\vee}^{(j)} = \sup_{s \in [0,1]} \int_{\mathbb{R}^d} \left\{ \widehat{\mathbb{D}}_n^{(j)}(s,x) \right\}^2 \mathrm{d}F_n(x)$$

and

$$\check{S}_{n,\vee}^{(j)} = \sup_{s \in [0,1]} \int_{\mathbb{R}^d} \left\{ \check{\mathbb{D}}_n^{(j)}(s,x) \right\}^2 \mathrm{d}F_n(x).$$

Tests for change-point detection à la Gombay and Horváth X

- The following result is then essentially a corollary of Theorem 2.

Proposition

Under H_0,

$$\left(S_{n,\vee}, \widehat{S}_{n,\vee}^{(1)}, \dots, \widehat{S}_{n,\vee}^{(N)}, \check{S}_{n,\vee}^{(1)}, \dots, \check{S}_{n,\vee}^{(N)}\right) \rightsquigarrow \left(S_{\vee}, S_{\vee}^{(1)}, \dots, S_{\vee}^{(N)}, S_{\vee}^{(1)}, \dots, S_{\vee}^{(N)}\right)$$

in $[0,\infty)^{(2N+1)}$, where

$$S_{\vee} = \sup_{s \in [0,1]} \int_{\mathbb{R}^d} \{\mathbb{D}_{P_0}(s,x)\}^2 \mathrm{d}F_0(x)$$

is the weak limit of $S_{n,\vee}$, F_0 is the c.d.f. corresponding to P_0, and $S_{\vee}^{(1)}, \dots, S_{\vee}^{(N)}$ are independent copies of $S_{\vee}$.

Tests for change-point detection à la Gombay and Horváth XI

- The previous proposition suggests interpreting the $\widehat{S}_{n,\vee}^{(j)}$ (resp. the $\check{S}_{n,\vee}^{(j)}$) under the null hypothesis as N "almost" independent copies of $S_{n,\vee}$ and thus computing an approximate p-value for $S_{n,\vee}$ as

$$\frac{1}{N}\sum_{j=1}^{N}\mathbf{1}\left(\widehat{S}_{n,\vee}^{(j)} \geqslant S_{n,\vee}\right) \qquad \text{or as} \qquad \frac{1}{N}\sum_{j=1}^{N}\mathbf{1}\left(\check{S}_{n,\vee}^{(j)} \geqslant S_{n,\vee}\right). \quad (6)$$

Proposition

Assume that H_1 holds with $k^\star = \lfloor nt \rfloor$ for some $t \in (0,1)$. Then, $S_{n,\vee} \overset{\mathbf{P}^*}{\to} +\infty$ while, for any $j \in \{1,\dots,N\}$, $\widehat{S}_{n,\vee}^{(j)}$ and $\check{S}_{n,\vee}^{(j)}$ are bounded in outer probability.

- A consequence of the previous proposition is that, under H_1, the approximate p-values for $S_{n,\vee}$ will tend to zero in outer probability.

Tests for change-point detection à la Gombay and Horváth XII

- As mentioned earlier, results analogous to the previous can be obtained for $S_{n,+}$, $T_{n,\vee}$, $T_{n,+}$, $U_{n,\vee}$, $U_{n,+}$, $V_{n,\vee}$ and $V_{n,+}$.

Outline

1. Introduction

2. Theoretical results for change-point detection

3. Tests for change-point detection à la Gombay and Horváth

4. Practical recommendations and illustration

Practical recommendations and illustration I

- From the results of our Monte Carlo experiments, the tests based on $\breve{S}_{n,\vee}$ and $\breve{T}_{n,+}$ seem good choices in the univariate case, while the test based on $\breve{V}_{n,+}$ seems to be a good one in the multivariate case.
- As an illustration, we applied the tests based on the processes $\breve{\mathbb{D}}_n^{(j)}$ to the trivariate hydrological data collected at the Ceppo Morelli dam, Italy, studied in Salvadori et al. (2011).
- The data consist of annual maxima for 49 years (in the range 1937–1994) of three variables: L (dam reservoir water level in m), Q (peak flow in $m^3.s^{-1}$) and V (peak volume in 10^6 m^3).
- We first tested for a change in the distribution of (L,Q,V). The approximate p-values of the eight tests based on the processes $\breve{\mathbb{D}}_n^{(j)}$ are given in the first line of the next table.
- For a detailed description of the data, see Section 2 of Salvadori et al. (2011).

Practical recommendations and illustration II

Table: Approximate p-values of the tests based on the processes $\check{\mathbb{D}}_n^{(j)}$ for the trivariate hydrological data under consideration. The trivariate (resp. bivariate) tests based on half-spaces were run with $m = 32$ (resp. $m = 8$). The approximate p-values were computed from $N = 10,000$ multiplier realizations.

Variables	$\check{S}_{n,\vee}$	$\check{S}_{n,+}$	$\check{T}_{n,\vee}$	$\check{T}_{n,+}$	$\check{U}_{n,\vee}$	$\check{U}_{n,+}$	$\check{V}_{n,\vee}$	$\check{V}_{n,+}$
(L,Q,V)	0.114	0.120	0.015	0.028	0.010	0.010	0.004	0.006
L	0.479	0.314	0.510	0.236				
(Q,V)	0.024	0.028	0.012	0.012	0.015	0.014	0.004	0.007

- Since there are both physical and statistical reasons to believe that L is independent of (Q,V) as explained in Salvadori et al. (2011), as a second step, we tested for a change in the distribution of L and in the distribution of (Q,V) separately. The obtained approximate p-values are reported in the second and third lines of Table 1.
- As can be seen from the results of the test based on $\check{V}_{n,+}$, there is strong evidence of a change in the distributions of (L,Q,V) and (Q,V).

Practical recommendations and illustration III

- From the second line of Table 1, we see that, on the contrary, there is no evidence of a change in the distribution of L. The latter finding is completely consistent with the fact that the variability of L is mainly due to the management policy of the reservoir which is constant over time.
- Indeed, as explained in Salvadori et al. (2011), the target of the dam manager is to keep a high water level in order to maximize electricity production.
- As classically done in the literature, under the hypothesis of a single break in a distribution, the change-point can be estimated by one of $\arg\max_{1 \leqslant k \leqslant n-1} \breve{S}_{n,k}$, $\arg\max_{1 \leqslant k \leqslant n-1} \breve{T}_{n,k}$, $\arg\max_{1 \leqslant k \leqslant n-1} \breve{U}_{n,k}$ or $\arg\max_{1 \leqslant k \leqslant n-1} \breve{V}_{n,k}$ depending on which test one wants to consider. For instance, the last estimator gives 31 for both (L, Q, V) and (Q, V), which corresponds to a change after the year 1976.

Practical recommendations and illustration IV

- Finally, let us mention that the approach based on multivariate empirical c.d.f.s considered in Csörgő and Horváth (1997, Section 2.6) and in Gombay and Horváth (1999) has been extended by Inoue (2001) to serially dependent observations, although the latter work is not aware of the former ones.
- A future research direction would be to study generalizations as such proposed in this work in the setting considered by Inoue (2001).

Bibliography I

J. Antoch and M. Hušková. Permutation tests in change point analysis. *Statistics and Probability Letters*, 53:37–46, 2001.

B.E. Brodsky and B.S. Darkhovsky. *Nonparametric methods in change point problems*. Kluwer Academic Publishers, 1993.

M. Csörgő and L. Horváth. Invariance principles for changepoint problems. *Journal of Multivariate Analysis*, 27:151–168, 1988.

M. Csörgő and L. Horváth. *Limit theorems in change-point analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, UK, 1997.

D. Ferger. On the power of nonparametric changepoint-tests. *Metrika*, 41: 277–292, 1994.

E. Gombay and L. Horváth. Change-points and bootstrap. *Environmetrics*, 10(6), 1999.

Bibliography II

E. Gombay and L. Horváth. Rates of convergence for *U*-statistic processes and their bootstrapped versions. *Journal of Statistical Planning and Inference*, 102:247–272, 2002.

L. Horváth and M. Hušková. Testing for changes using permutations of *U*-statistics. *Journal of Statistical Planning and Inference*, 128:351–371, 2005.

L. Horváth and Q.M. Shao. Limit theorems for permutations of empirical processes with applications to change point analysis. *Stochastic Processes and their Applications*, 117:1870–1888, 2007.

A. Inoue. Testing for distributional change in time series. *Econometric Theory*, 17(1):156–187, 2001.

M.R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer, New York, 2008.

J.P. Romano. A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, 83(403):698–708, 1988.

Bibliography III

G. Salvadori, C. De Michele, and F. Durante. On the return period and design in a multivariate framework. *Hydrol. Earth Syst. Sci.*, 15: 3293–3305, 2011.

A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer, New York, 2000. Second edition.