Noname manuscript No.
(will be inserted by the editor)

Ensembles of Change-Point Methods to Estimate the Change Point in Residual Sequences

Cesare Alippi · Giacomo Boracchi · Manuel Roveri

Received: date / Accepted: date

Abstract Change-Point Methods (CPMs) are statistical tests design to assess whether a given sequence comes from an unique, stationary, data-generating process. CPMs eventually estimate the change-point location, i.e., the point where the data-generating process shifted. While there exists a large literature concerning CPMs meant for sequences of independent and identically distributed (i.i.d.) random variables, their use on time-dependent signals has not been properly investigated. In this case, a straightforward solution consists in computing at first the residuals between the observed signal and the output of a suitable approximation model, and then applying the CPM on the residual sequence. Unfortunately, in practical applications, such residuals are seldom i.i.d., and this may prevent the CPMs to operate properly. To counteract this problem, we introduce the ensemble of CPMs, which aggregates several estimates obtained from CPMs executed on different subsequences of residuals, obtained from random sampling. Experiments show that the ensemble of CPMs improves the change-point estimates when the residuals are not i.i.d., as it is often the case in real-world scenarios.

Keywords Ensemble of Methods · Change-Point Methods · Changes in Processes · Residual Sequences

This research has been funded by the European Commissions 7th Framework Program, under grant Agreement INSFO-ICT-270428 (iSense). The final publication is available at link.springer.com

Cesare Alippi · Giacomo Boracchi · Manuel Roveri
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Mlano, Milano, Italy
Tel.: +39-02-23993512
Fax: +39-02-23993411
E-mail: {cesare.alippi, giacomo.boracchi, manuel.roveri}@polimi.it

1 Introduction

The work concerns the estimation of the change point in a given data sequence, i.e., the time instant when the data-generating process shifted from the initial stationary state into a different one. The change-point localization is typically performed by means of offline techniques, which can be eventually triggered by online change-detection mechanisms whenever these perceive a nonstationrity in the data.

The ability to estimate the change-point location in a sequence of data is of paramount importance in several application domains, such as financial time-series [1], climate analysis [2] and process control [3]. In fact, once the change point has been estimated, the causes of the change can be better investigated and any data-processing application can be reconfigured on the post-change conditions. For example, a change may indicate that a fault occurred in a system and the estimation of the fault-time instant is crucial for the fault identification and isolation phases [4]. In a classification scenario, a change point may correspond to concept drift in the data-generating process: availability of the change-time instant allows active classifiers to recognize recurrent concepts and activate the most appropriate one [5,6].

The change-point formulation [7–9] provides a general framework for designing statistical tests to locate a change point within a data sequence. Change-point methods (CPMs) are statistical tests designed to operate on sequences of independent and identically distributed (i.i.d.) random variables. Remarkably, CPMs do not require any preliminary training phase as they can be directly applied to the data, as for traditional statistical hypothesis tests. Unfortunately, CPMs cannot be directly used to locate changes in time-dependent signals, where the i.i.d. assumption is not satisfied.

A straightforward solution to cope with signals characterized by time dependency consists in applying the CPMs to the discrepancy (i.e., the residual) between the acquired

data and the output of a suitable approximation model [10]. In principle, in many situations, a good model would provide residuals that are white noise. In practice, however, this is a rare event because the approximation is typically affected by model bias. As a result, the residual sequences are often correlated and influenced by the dynamic of the original signal. To the best of our knowledge, countermeasures for applying CPMs in this practical scenario have never been properly investigated, though it is well known that correlation in the residuals impairs the performance of CPMs [11]. In particular, there are no solutions to improve the CPM performance on residual sequences.

In this work, we address the specific problem of estimating a change point within a time-dependent signal, which is reformulated as the problem of estimating the change point within a sequence of residuals derived from an approximation model. To this purpose, we propose an ensemble of CPMs that, when the residuals are not i.i.d., is able to locate the change-point better than a single CPM.

Ensemble methods have been successfully applied in many regression and classification problems [12], as well as in specific applications [13] (e.g., face, object and optical character recognition, intrusion detection, medical diagnosis) thanks to their ability to improve the performance of a single model by aggregating different models designed to solve the same problem [14, 12, 13]. In the last years, the interest in ensemble methods for time series prediction has exponentially grown (e.g., [15–17]), corroborating the use of ensemble methods in time-dependent scenarios. However, ensemble methods have never been proposed for estimating the change-time instant, and this paper illustrates a first attempt in this direction.

The proposed ensemble aggregates several individual estimates of the change point, each obtained by running a CPM on a subsequence defined by random sampling. The random sampling is meant to break down the temporal relationship in the residuals, which may hinder the change-point localization. Therefore, the proposed solution is general, since it is possible to build an ensemble of CPMs using any CPM to compute the individual estimates. Our experiments show that the ensemble of CPMs provides reliable estimates of the change point even in situations where a single CPM, executed on the whole residual sequence, would fail.

A preliminary study on this topic was published in [4], where an ensemble of CPMs was used for fault-diagnosis purposes in the specific case of changes in Autoregressive Moving Average (ARMA) models. This paper advances [4] by addressing the more general issue of estimating a change point in processes generating time-dependent signals. In addition, different aggregation mechanisms of the ensemble have been investigated and a detailed and comprehensive experimental analysis is presented, which includes also linear and nonlinear testbeds.

The paper is organized as follows: Section 2 describes the CPM formulation and the related literature. Section 3 formulates the problem of estimating the change-time instant in sequences of residuals, while the proposed ensemble of CPMs is detailed in Section 4. Experimental results are presented and discussed in Section 5, while conclusions are drawn in Section 6.

2 Change-Point Methods

2.1 The Change-Point Formulation

We say that a sequence $\mathcal{X}$

$$\mathcal{X} = \{x(t), 1 \leq t \leq L\},$$

of random variables contains a change-point T^* if $x(t)$ is distributed as

$$x(t) \sim \begin{cases} \mathcal{P}_0, & \text{if } t < T^* \\ \mathcal{P}_1, & \text{if } t \geq T^* \end{cases}, \tag{1}$$

where $\mathcal{P}_0$ and $\mathcal{P}_1$ are two different stationary distributions. The commonly made assumption [8, 9, 18] is that both $\{x(t), t < T^*\}$ and $\{x(t), t \geq T^*\}$ contain i.i.d. realizations of $\mathcal{P}_0$ and $\mathcal{P}_1$, respectively. Fig. 1 illustrates these settings.

CPMs [9] are statistical hypothesis tests, which analyze in an offline manner the sequence $\mathcal{X}$. Their null hypothesis consists in assuming that all data in $\mathcal{X}$ have been generated from the same distribution. When the null hypothesis is rejected, $\mathcal{X}$ is considered to contain a change point, whose location is also estimated.

From the practical point of view a CPM operates as follows: each time instant $S \in \{1, \ldots, L\}$ of $\mathcal{X}$ is considered as a candidate change point, and a test statistic $\mathcal{T}$ is computed to decide whether S is a change-point or not, given a predefined level of confidence α.

More in detail, for each candidate change point S, the sequence $\mathcal{X}$ is partitioned into two nonoverlapping sets

$$\begin{aligned} \mathcal{A}_S &= \{x(t),\ t = 1, \ldots, S\}, \\ \mathcal{B}_S &= \{x(t),\ t = S+1, \ldots, L\}, \end{aligned} \tag{2}$$

and the test statistic

$$\mathcal{T}_S = \mathcal{T}(\mathcal{A}_S, \mathcal{B}_S), \tag{3}$$

is computed to measure the degree of dissimilarity between $\mathcal{A}_S$ and $\mathcal{B}_S$ (test statistics typically used in CPMs are reviewed in Section 2.2).

The values of $\mathcal{T}_S$ are computed for each change-point candidate, yielding $\{\mathcal{T}_S, S = 1, \ldots, L\}$. In what follows, we denote by $\mathcal{T}_{M_\mathcal{X}}$ the maximum value of the statistic $\mathcal{T}$ over all the change-points candidates in $\mathcal{X}$, i.e.,

$$\mathcal{T}_{M_\mathcal{X}} = \max_{S=1,\ldots,L} (\mathcal{T}_S)\,. \tag{4}$$

Then, the value of $\mathcal{T}_{M_\mathcal{X}}$ is compared with a predefined threshold $h_{L,\alpha}$, which depends on the statistic $\mathcal{T}$, the cardinality L of $\mathcal{X}$ and $0 < \alpha < 1$, which sets the percentage of type I errors (i.e., false positives) of the CPM. When $\mathcal{T}_{M_\mathcal{X}}$ exceeds $h_{L,\alpha}$, the CPM rejects the null hypothesis, and $\mathcal{X}$ is claimed to contain a change point at the location maximizing (4)

$$M_\mathcal{X} = \underset{S=1,\ldots,L}{\operatorname{argmax}} \left(\mathcal{T}_S\right). \tag{5}$$

Conversely, when $\mathcal{T}_{M_\mathcal{X}} < h_{L,\alpha}$, there is not enough statistical evidence to reject the null hypothesis, and $\mathcal{X}$ is considered to be stationary. Therefore, the CPM outcome is

$$\begin{cases} \text{The estimated change-point is } M_\mathcal{X} & \text{if } \mathcal{T}_{M_\mathcal{X}} \geq h_{L,\alpha} \\ \text{No change-point identified in } \mathcal{X} & \text{if } \mathcal{T}_{M_\mathcal{X}} < h_{L,\alpha} \end{cases}. \tag{6}$$

2.2 Test Statistics Used in CPMs

A common approach to locate arbitrary changes in a sequence of samples drawn from an unknown distribution consists in analyzing their moments by means of nonparametric test statistics [18]. Several nonparametric statistics are based on the rank computation, such as the Mann-Whitney [19] (to assess changes in the mean), the Mood [20] (to assess changes in the variance) and the Lepage ones [21] (to assess both changes affecting the mean and the variance). A CPM based on the Mann-Whitney statistic was introduced in [8] together with a CPM for Bernoulli random variables. Shifts in the mean of a Gaussian random variable can be found by CPMs based on the two-sample t test statistic [9]. A different approach consists in locating change points by comparing the empirical distributions over two sets of data, as in the CPMs [22] that are based on the Kolmogorov-Smirnov and the Cramer Von Mises [23] statistics.

The change-point formulation has been also used to monitor online and sequentially data streams, by iterating the CPM at each new sample arrival [24]. Approximated solutions have been adopted to bound the computational complexity and memory requirement of CPMs applied to data streams [18,22]. In particular, such a streaming adaptation is required when the test statistic $\mathcal{T}$ is computationally demanding (such as test statistics based on the rank computation). So far we mentioned test statistics for scalars, however, the change-point formulation can be used to analyze multivariate data, such as the CPM in [25], which relies on the on Hotelling T^2 statistic.

It is worth noting that the computation of the thresholds is the major issue when designing CPMs. In fact, even when the distribution of the test statistics $\mathcal{T}$ is known for any partition of $\mathcal{X}$, the distribution of its maximum $\mathcal{T}_{M_\mathcal{X}}$ may be far from being trivial and, often, only asymptotic or approximated expressions are available. Furthermore, there are no obvious analytical expressions for computing these thresholds when the CPMs are used in an online manner. In fact, in this case, one should compute the probability that $\mathcal{T}_{M_\mathcal{X}}$ exceeds $h_{L,\alpha}$ at the L-th sample, conditioned on the fact that $\mathcal{T}$ never exceeded the threshold on the previous $L-1$ samples. Therefore, very often, thresholds $\{h_{L,\alpha}, L > 0\}$ have to be computed by numerical simulations, as in [9].

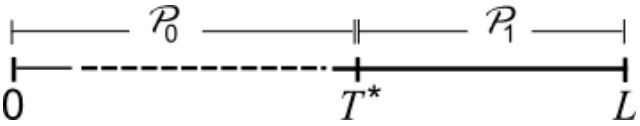

Fig. 1 The considered time-instants: T^* is the time instant when the process abruptly changes, L is the time instant the change is detected by an external change-detection test. During the interval $[0, T^*)$, data are generated from $\mathcal{P}_0$, while during interval $[T^*, L]$ data are generated from $\mathcal{P}_1$, which are both stationary and unknown. Goal of change-point methods is to estimate T^* given the whole sequence $\mathcal{X}$.

3 The Problem Formulation

Let us extend the traditional framework for CPMs (1) by considering $\mathcal{X}$ as a time-dependent signal. In this scenario, $\mathcal{P}_0$ and $\mathcal{P}_1$ in (1) become two stationary processes generating time-dependent signals (and not anymore random variables providing i.i.d. realizations). We always assume that the change corresponds to a shift of $\mathcal{P}_0$, which is permanently replaced by $\mathcal{P}_1$ and that the descriptions of $\mathcal{P}_0$ and $\mathcal{P}_1$ are not provided. Our goal is always estimating T^*, assuming that the change has been safely detected at $L > T^*$ by any change-detection method, e.g., [26–28].

The proposed procedure aims at estimating, in an offline manner, the change point T^* within a given data sequence $\mathcal{X}$ that contains data generated both before and after the change. As commented in Section 1, it is not possible to directly apply the CPM to $\mathcal{X}$ when this is characterized by a temporal dependence, and the CPM should be rather applied to the residuals between the data and a suitable approximation model. We here consider a general class of approximation models such as the multiple input/single output (MISO) time-invariant models in the predictive form

$$\begin{aligned} \widehat{x}(t) = f_\theta(&x(t-1), \ldots, x(t-n_x), \\ &u(t), u(t-1), \ldots, u(t-n_u)), \end{aligned} \tag{7}$$

where $f_\theta(\cdot)$ is a linear/nonlinear function parametrized by vector θ, the system output $x(t)$ and its prediction $\widehat{x}(t)$ are scalar elements, and $u(t) \in \mathbb{R}^m$ represents the (possible) input of (7). The integers $n_x \geq 0$ and $n_u \geq 0$ are parameters representing the order of the output and input (if present), respectively. We assume that f is given and that the parameter vector $\widehat{\theta}_0$ has been estimated on an initial training sequence containing only data generated from $\mathcal{P}_0$. Selection of the best f as well as the estimate of θ are outside the scope of the paper; the interested reader can refer to [29].

The residual at time t is computed as

$$r(t) = x(t) - f_{\hat{\theta}_0}(x(t-1), \ldots, x(t-n_x), u(t), u(t-1), \ldots, u(t-n_u)), \quad (8)$$

and the change-point has to be located within the residual sequence

$$\mathcal{R} = \{r(t), t = 1, \ldots, L\}. \quad (9)$$

Following the change-point formulation described in Section 2, a CPM on the residual sequence is immediately obtained by replacing $x(t)$ with $r(t)$ and $\mathcal{X}$ with $\mathcal{R}$.

Unfortunately, in real world applications, residuals in $\mathcal{R}$ are typically far from being i.i.d., even before T^*, due to model bias. Moreover, we expect a large degree of dependency among residuals after T^*. These circumstances violate the hypothesis required by the CPMs and explain why, more often than in the i.i.d. case, the statistic $\mathcal{T}$ does not properly locate the change-point in $\mathcal{R}$. In the next section we introduce the ensemble of CPM, which has been designed to operate on residuals when these are not i.i.d., and which provides better estimate than a single CPM executed on the whole residual sequence, as discussed in the experiments.

4 Ensemble of CPMs

We denote by $\mathcal{E}_d(\mathcal{R})$ the ensemble of CPMs that aggregates d individual estimates of the change point, namely $\{M_i, i = 1, \ldots, d\}$, which are provided by a CPM executed on different subsequences of $\mathcal{R}$. Peculiarity of the proposed ensemble is that each subsequence is obtained by random sampling to reduce the temporal dependency in $\mathcal{R}$. The way how these individual estimates are computed is described in Section 4.1, the aggregation is presented in Section 4.2, while Section 4.3 summarizes the ensemble of CPMs.

4.1 Computing the Individual Estimates M_i

In what follows we describe the way how each individual estimate $\{M_i, i = 1, \ldots, d\}$ of the ensemble is computed. At first, a subsequence of $n < L$ elements is extracted from $\mathcal{R}$ by means of the operator $\mathrm{D}_n(\cdot)$, which performs a random sampling over a finite sequence of integers. Let $\{1, \ldots, L\}$ be the indexes of $\mathcal{R}$, we denote by

$$\mathcal{I}_n^{(i)} = \mathrm{D}_n(\{1, \ldots, L\}), \quad (10)$$

a sequence containing indexes randomly extracted – without repetition – from $\{1, \ldots, L\}$, in such a way that $\#\mathcal{I}_n^{(i)} = n$ and the indexes in $\mathcal{I}_n^{(i)}$ are monotonically increasing. The sequence $\mathcal{I}_n^{(i)}$ is used to select the elements of $\mathcal{R}$ that are used to compute the i-th individual estimate, namely,

$$\mathcal{R}_i = \{r(t),\ t \in \mathcal{I}_n^{(i)}\}. \quad (11)$$

Therefore, $\mathcal{R}_i$ is a subsequence of $\mathcal{R}$ containing n elements randomly chosen, without repetition, and ordered as they appear in $\mathcal{R}$.

A CPM with a specific statistic $\mathcal{T}$ is executed on $\mathcal{R}_i$, providing m_i that corresponds to the partitioning maximizing the test statistic in $\mathcal{R}_i$, as in (5). Referring to the above notation, the i-th individual estimate of the ensemble, M_i, is given by

$$M_i = \mathcal{I}_n^{(i)}[m_i], \quad (12)$$

where $\mathcal{I}_n^{(i)}[m_i]$ indicates the element at the position m_i in $\mathcal{I}_n^{(i)}$. Note that (12) maps the estimate of the change point from the subsequence $\mathcal{R}_i$ back to $\mathcal{R}$ indexes, thus in the temporal domain. We denote by $\mathcal{T}_{M_i}$ the value that the test statistic $\mathcal{T}$ reaches in M_i. Therefore, when $\mathcal{T}_{M_i} < h_{n,\alpha}$, the CPM executed on $\mathcal{R}_i$ is not able to locate the change point. Equation (12) may provide inaccurate results when the subsequence $\mathcal{R}_i$ does not include the true change point; it is however possible to mitigate this problem by taking $M_i = (\mathcal{I}_n^{(i)}[m_i] + \mathcal{I}_n^{(i)}[m_i + 1])/2$.

4.2 Aggregation

The procedure described in Section 4.1 is repeated d times, yielding the estimates $\{M_i, i = 1, \ldots, d\}$ computed on randomly defined subsequences. In addition, $M_{\mathcal{R}}$ and the corresponding value of the statistic $\mathcal{T}_{M_{\mathcal{R}}}$ are also computed on the whole sequence $\mathcal{R}$, as in (4)-(6). The ensemble $\mathcal{E}_d(\mathcal{R})$ aggregates $\{M_i, i = 1, \ldots, d\}$ together with $M_{\mathcal{R}}$ to obtain a final estimate M of the change point T^*. As it often happens in ensemble methods [13], the aggregation consists in a weighted averages of the individual estimates

$$M = \frac{\sum_{i=1}^{d} \omega_i M_i + \omega_{d+1} M_{\mathcal{R}}}{\sum_{i=1}^{d+1} \omega_i}. \quad (13)$$

The most straightforward solution to define the weights $\{\omega_i,\ i = 1, \ldots, d+1\}$ is to set them as binary values, to consider only change points for which the test statistics exceeds the corresponding threshold, i.e.,

$$\omega_i = \begin{cases} 0, & \text{if } \mathcal{T}_{M_i} < h_{n,\alpha} \quad i = 1, \ldots, d \\ 1, & \text{if } \mathcal{T}_{M_i} \geq h_{n,\alpha} \quad i = 1, \ldots, d \\ 0, & \text{if } \mathcal{T}_{M_{\mathcal{R}}} < h_{L,\alpha} \quad i = d+1 \\ 1, & \text{if } \mathcal{T}_{M_{\mathcal{R}}} \geq h_{L,\alpha} \quad i = d+1 \end{cases}. \quad (14)$$

In (14) all the estimates having test statistic above the threshold are considered as equally relevant. A different solution consists in assigning larger weights to estimates provided by partitions yielding larger values of the test statistic,

such as

$$\omega_i^a = \begin{cases} 0, \text{ if } \mathcal{T}_{M_i} < h_{n,\alpha} & i = 1, \ldots, d \\ \frac{\mathcal{T}_{M_i}}{h_{n,\alpha}}, \text{ if } \mathcal{T}_{M_i} \geq h_{n,\alpha} & i = 1, \ldots, d \\ 0, \text{ if } \mathcal{T}_{M_\mathcal{R}} < h_{L,\alpha} & i = d+1 \\ \frac{\mathcal{T}_{M_\mathcal{R}}}{h_{L,\alpha}}, \text{ if } \mathcal{T}_{M_\mathcal{R}} \geq h_{L,\alpha} & i = d+1 \end{cases}. \quad (15)$$

Another viable option consists in selecting, among all the individual estimates, only the one corresponding to the maximum value of the test statistic, i.e.,

$$\omega_i^s = \begin{cases} 1, \text{ if } \mathcal{T}_{M_i} = \max\limits_{j=1,\ldots,d+1} \left(\mathcal{T}_{M_j}\right) \\ 0, \text{ otherwise} \end{cases}. \quad (16)$$

In the experimental section we evaluate these three aggregation strategies. We point out that when none of the CPMs executed on the subsequences $\mathcal{R}_i, i = 1, \ldots, d$ or on $\mathcal{R}$ determines a change point (i.e., when $\mathcal{T}_{M_i} < h_{n,\alpha}, i = 1, \ldots, d$ and $\mathcal{T}_{\mathcal{R}_i} < h_{L,\alpha}$), none of the individual estimates $\{M_i, i = 1, \ldots, d\}$ and $M_\mathcal{R}$ is defined, and $\mathcal{E}_d$ is not able to locate the change point in $\mathcal{R}$.

4.3 The Algorithm

Algorithm 1 details the proposed ensemble of CPMs. In particular the loop at lines 2 - 12 performs d-times the random sampling of $\mathcal{R}$, computes the individual estimates $\{M_i, i = 1, \ldots, d\}$ and the corresponding statistic $\mathcal{T}_{M_i}$. The values of the test statistic define the aggregation weights $\{\omega_i, i = 1, \ldots, d\}$, as in (14) or (15) (in case of weights performing selection among the individual estimates the lines 10 - 12 and 15 - 17 have to be replaced by (16)). Then, the change-point estimate from the whole residual sequence $\mathcal{R}$ together with its weight are computed at lines 14 - 17. Finally, the $d+1$ estimates are aggregated at line 18 to provide the output M of the ensemble.

Algorithm 1: Ensemble of CPMs

Input: $\mathcal{R} = \{r(t), t = 1, \ldots, L\}$ (the residual sequence), n (the random sampling parameter), $h_{n,\alpha}$ and $h_{L,\alpha}$ (the thresholds of the CPM), d (the number of individual estimates of the ensemble),
Output: M (the estimate of the change-time instant).

1- $i = 1$
2- **while** *($i \leq d$)* **do**
3- $\quad \mathcal{I}_n^{(i)} = \mathrm{D}_n(\{1, \ldots, L\})$,
4- $\quad$ **foreach** $S \in \mathcal{I}_n^{(i)}$ **do**
5- $\quad\quad \mathcal{A}_S = \{r(t),\ t \in \mathcal{I}_n^{(i)}, t \leq S\}$,
6- $\quad\quad \mathcal{B}_S = \{r(t),\ t \in \mathcal{I}_n^{(i)}, t > S\}$,
7- $\quad\quad \mathcal{T}_S = \mathcal{T}(\mathcal{A}_S, \mathcal{B}_S)$,
$\quad$ **end**
8- $\quad m_i = \underset{S \in \mathcal{I}_n^{(i)}}{\operatorname{argmax}} (\mathcal{T}_S)$,
9- $\quad M_i = \mathcal{I}_n^{(i)}[m_i]$, and $\mathcal{T}_{M_i} = \max\limits_{S \in \mathcal{I}_n^{(i)}} (\mathcal{T}_S)$,
10- $\quad$ **if** *($\mathcal{T}_{M_i} \geq h_{n,\alpha}$)* **then**
11- $\quad\quad \omega_i = 1$ (or $\omega_i = \frac{\mathcal{T}_{M_i}}{h_{n,\alpha}}$),
$\quad$ **else**
12- $\quad\quad \omega_i = 0$,
$\quad$ **end**
13- $\quad i = i + 1$;
end
14- Compute $M_\mathcal{R}$ and $\mathcal{T}_{M_\mathcal{R}}$ as in (4)-(6)
15- **if** *($\mathcal{T}_{M_\mathcal{R}} \geq h_{L,\alpha}$)* **then**
16- $\quad \omega_{d+1} = 1$ (or $\omega_{d+1} = \frac{\mathcal{T}_{M_\mathcal{R}}}{h_{L,\alpha}}$),
else
17- $\quad \omega_{d+1} = 0$,
end
18- $M = \frac{\sum_{i=1}^{d} \omega_i M_i + \omega_{d+1} M_\mathcal{R}}{\sum_{i=1}^{d+1} \omega_i}$.

5 Experiments

The proposed ensemble of CPM has been evaluated on a large experimental campaign encompassing both synthetically generated data (sequences from ARMA processes) and two testbeds (i.e., the Hairdryer and the Two-Tanks system, the former representing a linear model, the latter a nonlinear one).

5.1 Considered Solutions

We consider CPMs based the Lepage test statistic [21] to assess simultaneously changes in mean and variance of the residuals. The Lepage test statistic is defined as

$$\mathcal{L} = \mathcal{U}^2 + \mathcal{M}^2 \quad (17)$$

being $\mathcal{U}$ the Mann-Whitney [19] and $\mathcal{M}$ the Mood [20] test statistics. The statistic $\mathcal{U}$ is meant to locate changes in the mean, while $\mathcal{M}$ in the variance; both statistics are based on rank computation. Thresholds $h_{L,\alpha}$ for the Lepage statistics are made available by the `CPM` package [30] implemented in `R` statistical software. In what follows we denote by $\mathrm{CPM}_\mathcal{L}$ the CPM built upon (17).

In particular, we tested the following solutions:

- $\mathrm{CPM}_\mathcal{L}(\mathcal{R})$: the CPM executed on $\mathcal{R}$.
- $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$: the CPM executed on $\mathcal{R}$ with $h_{L,\alpha} = 0$. This CPM always estimates the change-point at the location where the test statistic is maximized, disregarding whether the test statistic is above or below the threshold.
- $\mathrm{CPM}_{\mathcal{L},0}(\partial\mathcal{R})$: the $\mathrm{CPM}_{\mathcal{L},0}$ applied to $\partial\mathcal{R} = \{r(t+1) - r(t), t = 1, \ldots, L-1\}$. Like in [18], such a preprocessing is meant to reduce the temporal correlation in the data.
- $\mathcal{E}_d$: the ensemble of CPMs where the aggregation weights $\{\omega_i, i = 1, \ldots, d+1\}$ are computed as in (14). We consider four cardinalities of the ensemble, $d \in \{10, 25, 50, 100\}$.
- $\mathcal{E}_{100}^a$: the ensemble of 100 CPMs with aggregation weights $\{\omega_i^a, i = 1, \ldots, d+1\}$ computed as in (15).

$\mathcal{E}^s_{100}$: the ensemble of 100 CPM that selects the individual estimate yielding the largest statistical evidence of the change, as in (16).

In all the CPMs we set $\alpha = 0.05$, and in the ensembles we set $n = L/2$.

5.2 Figures of Merit

We evaluate the performance of the ensembles of CPMs by means of the following figures of merit:

- The False Negative Rate (FNR), which is the percentage of experiments where the change-point is not located (i.e., the test statistic never exceeds the threshold).
- The change-point location accuracy, which is evaluated by inspecting (by means of boxplots and histograms) the empirical distribution of the change-point estimates w.r.t. the true location of the change.

We do not consider the false positive rate since every sequence contains a change-point.

5.3 Description of Datasets

5.3.1 Dataset of ARMA Processes

We consider ARMA models as data-generating processes:

$$x(t) = \sum_{i=1}^{p} \phi_i x(t-i) + \sum_{i=1}^{q} \psi_i \epsilon(t-i) + \epsilon(t), \qquad (18)$$

where $\theta = [\phi_1, \ldots, \phi_p, \psi_1, \ldots, \psi_q]$ represents the parameter vector, $\epsilon(t) \sim \mathcal{N}(0, \sigma^2)$ denotes the innovation at time t, $p > 0$ and $q > 0$ correspond to the orders of the autoregressive (AR) and moving-average (MA) terms, respectively.

Each data sequence is composed of 550 samples and contains an abrupt and permanent shift in the parameter vector θ at time $T^* = 500$: before T^* data are generated according to θ_0, after T^* the parameter vector becomes $\theta_1 \neq \theta_0$. We considered three scenarios where $L = \{530, 540, 550\}$, representing situations where the change-detection test triggering the CPMs has a different detection latency, thus the CPM is executed on a different amount of samples generated after the change.

A dataset of 10000 data sequences was prepared by randomly generating the parameter vectors θ_0 and θ_1, including only those yielding stable systems. In each sequence, the orders (p, q) have been also randomly selected within their range $p \in \{1, \ldots, 4\}$ and $q \in \{0, 1, 2\}$. The standard deviation of the innovation is $\sigma = 0.1$. The parameter vector θ_1 is unknown and we estimate θ_0 by classical system identification techniques [31] on an initial training sequence of 400 samples, assuming that the ranges of p and q were known, and selecting $\widehat{\theta}_0$ according to the Akaike information criteria [32].

5.3.2 Linear/Nonlinear Testbed

We considered the Hairdryer and the Two-Tanks system testbeds, both available in Mathworks `Matlab`. The former application refers to a single input/single output (SISO) system described in [31], modeling the relationship between the power of a heating device (the input) and the air temperature (the output). As described in [31], this system can be successfully modeled by means of an autoregressive with exogenous input (ARX) model. The first 700 samples of the sequence have been used to train the ARX model, and two types of abrupt changes have been injected in the sequence at $T^* = 800$: an additive change summing a constant value δ_A, and a multiplicative change that scales the sequence of $1 + \delta_M$. We set $\delta_A = 0.08$ corresponding to approximately 2% of the signal amplitude and $\delta_M = 0.02$. L has been fixed to 850.

The latter application refers to a SISO system modeling two tanks (upper and lower) connected by a pipe. The system takes as input the voltage at the pump of the upper tank, while the output is the liquid level in the lower tank. The system can be modeled by means of a non-linear system (e.g., Nonlinear ARX or Hammerstein-Wiener model). A Nonlinear ARX model (where the nonlinear component is represented by a Wavelet Network) [31] has been trained on the first 2200 samples. We consider both additive and multiplicative changes with parameters $\delta_A = 0.03$ corresponding to approximately 5% of the amplitude of the signal and $\delta_M = 0.2$. Here, the change is injected at time $T^* = 2300$, while $L = 2350$. The parameters δ_A and δ_M in both testbeds have been set such that $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ was not able to locate the change-point, to outline the effectiveness of the ensemble where a single CPM would fail.

5.4 Discussion

Experiments on the ARMA dataset show that the ensemble guarantees low FNR (Fig. 2) and good estimates of the change-point location where $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ fails (Fig. 5). In particular, the distribution of the ensemble estimates conditioned on the fact that $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ never exceeds the threshold shows that the ensembles $\mathcal{E}_d$ may provide better estimates than $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ (Fig. 5). Similar results hold in the comparison with $\mathrm{CPM}_{\mathcal{L},0}(\partial\mathcal{R})$ (Fig. 6). These plots demonstrate that the advantages provided by the ensemble of CPMs cannot be simply achieved by arbitrarily lowering the thresholds or detrending the residuals when a single CPM is used.

Fig. 2 (a-c) show that the $\mathcal{E}_d$s have lower FNR than $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ and that, as expected, the FNR decreases with d. By comparing the FNR values in the three scenarios, we see that FNR decreases when L increases, since the more samples after the change, the easier to estimate a change-point in the sequence.

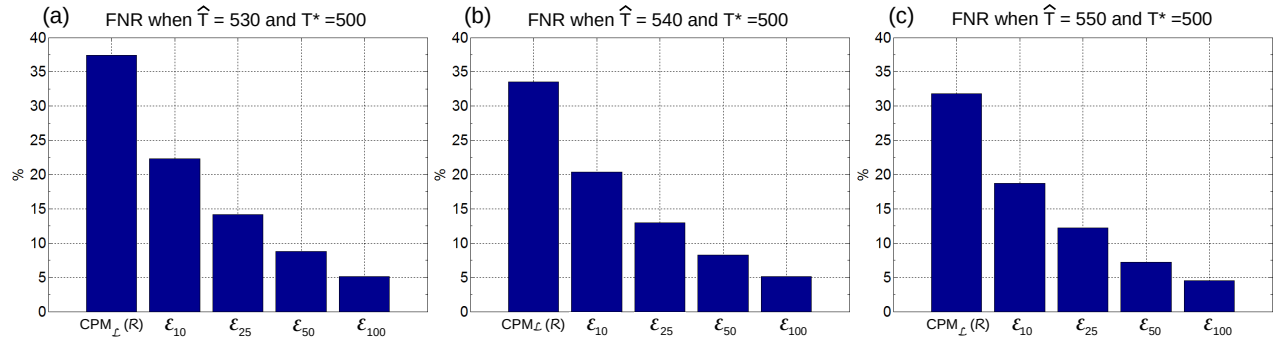

Fig. 2 ARMA dataset: the FNR for the three considered scenarios $L = \{530, 540, 550\}$.

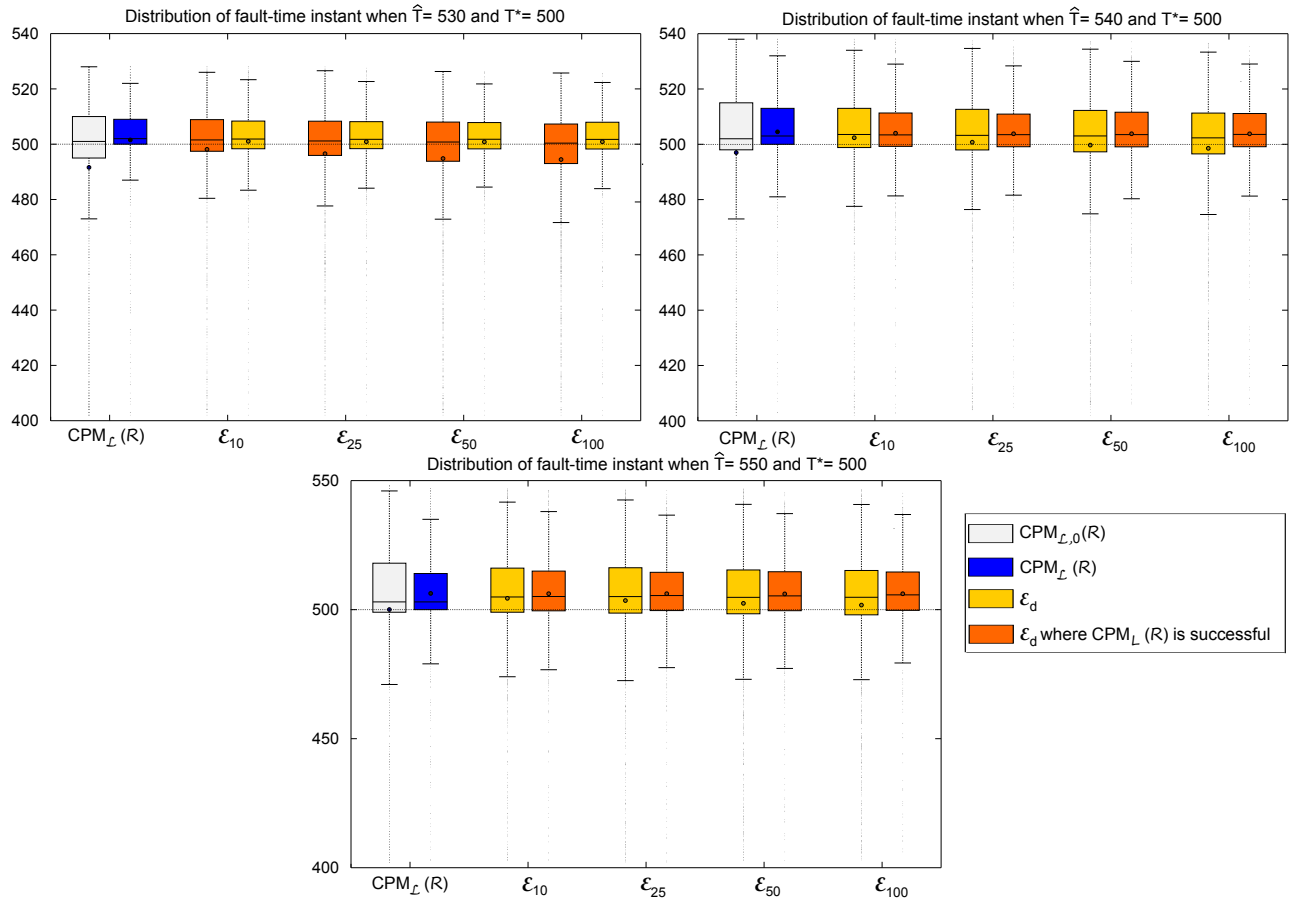

Fig. 3 ARMA dataset: the boxplots provide compact views of the empirical distribution of the change-point estimates. The central line in each box represents the sample median, the bottom and top of each box represents the 25% and 75% quantile, respectively. Each circular marker corresponds to the sample mean, while the whiskers are meant to identify outliers.

The boxplots in Fig. 3 provide a compact view of the empirical distributions of the estimated change-point for $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ and $\mathcal{E}_d$. These boxplots show that all the considered CPMs suffer from a structural delay that increases with L, as the sample mean (represented by a circular marker) of each boxplot is shifted upwards when L increases. If the residuals were i.i.d. after the change this would not happen, since increasing the number of samples consolidates the evidence of the change in the test statistic. In contrast, when the residuals after the change are not i.i.d., and for instance follow a trend, it is very likely that a point after T^* maximizes the test statistic (hence shifting forward the change-point estimate). For this reason, the number of outliers (displayed with small gray dots out of the boxplot whiskers) in the right tail of each distribution also increases with L, and this is particularly evident when $L = 550$.

Fig. 3 shows also that, in all the considered scenarios, $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ (light gray boxes) is characterized by the largest

interquantile range (the extremes of the boxplot are the 25th and 75th percentiles) and a large number of outliers. In particular, because of these outliers, the sample mean falls below the 25% quantile when $L = 530$. Furthermore, the dispersion of the estimates of $\mathcal{E}_d$ (orange boxes) increases with d, and a similar consideration holds for the number of outliers.

It is important to analyze the accuracy of $\mathcal{E}_d$ conditioned to the fact that $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ is able to locate the change point or not. Fig 4 compares $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ with $\mathcal{E}^a_{100}$ only on sequences where $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ was successful when $L = 530$. The two histograms are rather similar though the ensemble has, as expected, a less peaked and smoother distribution. The yellow boxplots in Fig. 3 have been computed from these subsequences, and confirm that $\mathcal{E}_d$ provide estimates that are located very close to that of $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ and characterized by a similar dispersion. Therefore, the aggregation phase can successfully compensate possible inaccurate individual estimates. In contrast, Fig. 5 compares $\mathcal{E}_d$ with $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ on those sequences where $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ was not successful. Here, both distributions of $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ and $\mathcal{E}_d$ have an heavy left tail, which is coherent with the large number of outliers in the left part of the boxplots in Fig. 3. These tails indicate that often, in these sequences, the point maximizing the test statistic was located far before T^*. This is probably due to estimation errors of $\widehat{\theta}_0$, which induce model bias and make $\{r(t), t < T^*\}$ far from being white noise. However, in these situations, estimates from $\mathcal{E}^a_{100}$ are better clustered around T^*, while $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ is often stacked at the extremes of $\mathcal{R}$. This means that on residual sequences, it is often better to use an ensemble of CPMs than lowering the thresholds to ease the change-point localization.

As a further comparison, we contrasted $\mathcal{E}^a_{100}$ with $\mathrm{CPM}_{\mathcal{L},0}(\partial\mathcal{R})$ in Fig. 6, showing that the latter is also affected by the same problems of $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$, and that the improvement provided by the ensemble cannot be achieved by detrending the residuals. Fig. 7 compares the distribution of $\mathcal{E}_{100}$, $\mathcal{E}^a_{100}$ and $\mathcal{E}^s_{100}$ in the whole dataset, and shows that there is no substantial difference when weights (14) and (15) are considered. However, when the aggregation is performed by selection as in (16), the distribution tends to be more peaked at T^*, while suffering from heavy tails as $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ does.

Interestingly, the experimental results on the testbeds described in Section 5.3.2 are in line with those on the ARMA dataset. While the outputs of $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ and $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ are deterministic, i.e., given an specific input sequence they always provide the same output, the output of any ensemble $\mathcal{E}_d$ is stochastic, because of the random sampling. Therefore, in Fig 8 - 11 we plot the empirical distribution of $\mathcal{E}_{100}$ and $\mathcal{E}^a_{100}$ estimates, computed over 500 iterations on the same sequence. In the figure captions we report the percentage of times when the change point was located within these sequences (over the 500 runs), as well as the estimate provided by $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ ($\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ was never able to locate the change point here). From these experiments it emerges that the ensemble can be successfully used to estimate change points in more complex and realistic scenarios, though in Fig. 10, $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ provides a better estimate of the change point.

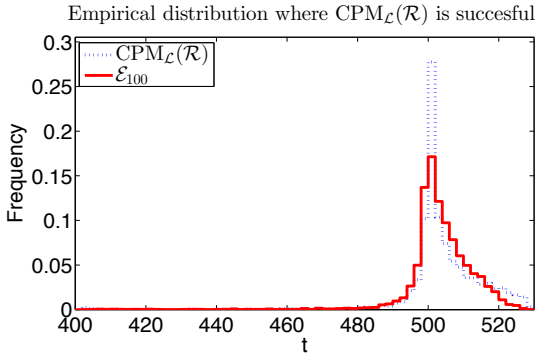

Fig. 4 ARMA dataset: comparison of $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ and $\mathcal{E}_{100}$ on sequences where $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ was successful.

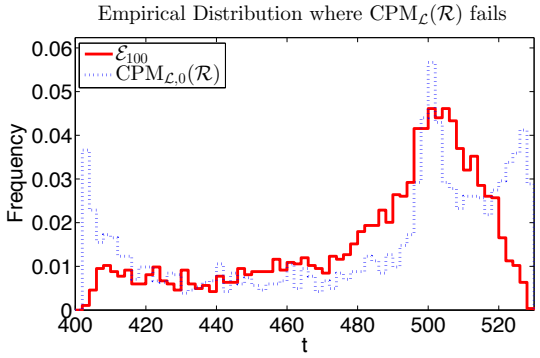

Fig. 5 ARMA dataset: comparison of $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ and $\mathcal{E}_{100}$ on sequences where $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ was not successful (while $\mathcal{E}_{100}$ was successful).

6 Conclusions

We presented an ensemble of CPMs that, combined with a suitable approximation model, represents a viable option for estimating changes in data-generating processes, thus extending the applicability of CPM beyond sequences of i.i.d. random samples. In conclusion, we remark that using the ensemble of CPMs is not beneficial when analyzing sequences of i.i.d. samples. In fact, on i.i.d. sequences, the more data are used, the better the estimate of the change point is, therefore, it is not convenient to perform random sampling. Fur-

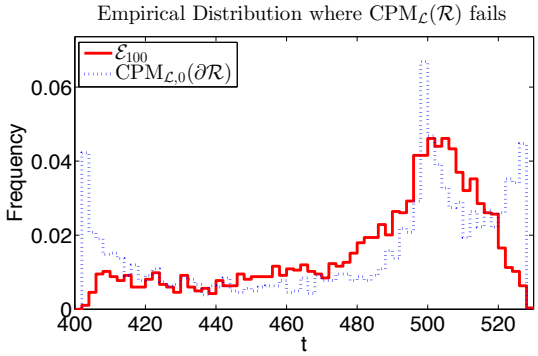

Fig. 6 ARMA dataset: comparison of $\mathrm{CPM}_{\mathcal{L},0}(\partial\mathcal{R})$ and $\mathcal{E}_{100}$ on sequences where $\mathrm{CPM}_{\mathcal{L}}(\mathcal{R})$ was not successful (while $\mathcal{E}_{100}$ was successful).

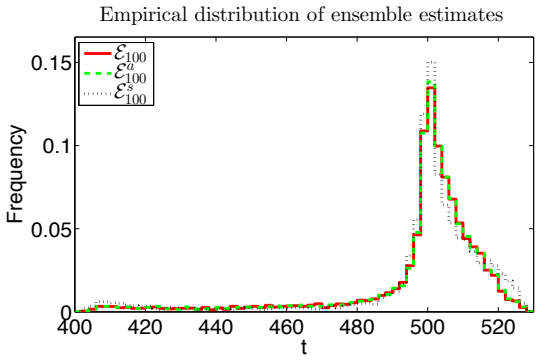

Fig. 7 ARMA dataset: performance of the ensemble when different aggregation strategies are considered.

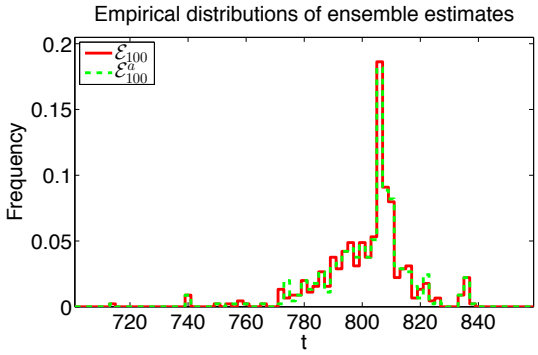

Fig. 8 Hairdryer dataset affected by additive change: distribution of ensemble estimates over 500 runs. The ensemble was able to locate the change point in about 90.2% of runs, while $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ estimated the change point at sample 806.

thermore, CPMs guarantee a controlled false positive rate on i.i.d. sequences, while the ensembles of CPMs do not.

However, when the i.i.d. assumption is not satisfied (as in the case of residuals from approximation models), neither individual CPMs nor the ensemble can control the false positives rate, while the ensemble may provide better estimates, as our experiments show. In light of these considerations, we believe that the ensemble of CPMs is promising and deserves further investigation.

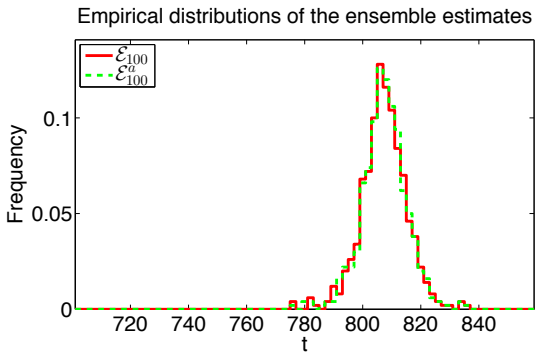

Fig. 9 Hairdryer dataset affected by multiplicative change: distribution of ensemble estimates over 500 runs. The ensemble was able to locate the change point in all the runs, while $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ estimated the change point at sample 806.

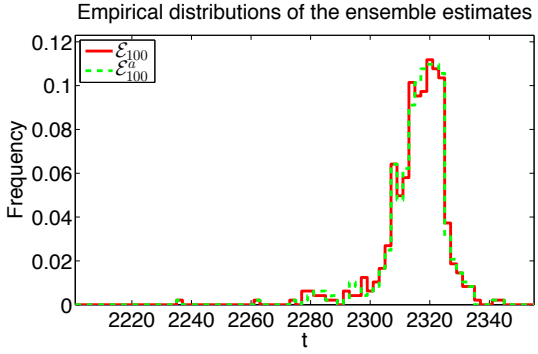

Fig. 10 Two tank dataset affected by an additive change: distribution of ensemble estimates over 500 runs. The ensemble was able to locate the change point in about 98.5% of runs, while $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ estimated the change point at sample 2307.

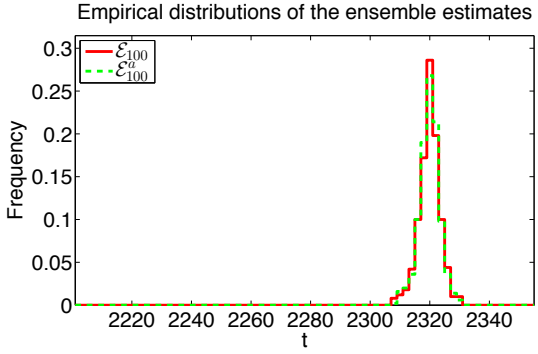

Fig. 11 Two tank dataset affected by a multiplicative change: distribution of ensemble estimates over 500 runs. The ensemble was able to locate the change point in all the runs, while $\mathrm{CPM}_{\mathcal{L},0}(\mathcal{R})$ estimated the change point at sample 2323.

Ongoing works concern the study of other techniques than random sampling for computing the individual estimates, the aggregation of different test statistics in the ensemble, as well as using the ensemble of CPMs to locate change points in features extracted from data streams, as in [33]. In fact, features extracted from time-dependent signals are often not i.i.d., and it can be beneficial to adopt the en-

semble of CPMs to locate change points within feature sequences.

References

1. J. Bai, "Estimation of a change point in multiple regression models," *The Review of Economics and Statistics*, vol. 79, no. 4, pp. 551–563, November 1997.
2. J. Reeves, J. Chen, X. L. Wang, R. Lund, and L. QiQi, "A review and comparison of changepoint detection techniques for climate data," *Journal of Applied Meteorology and Climatology*, vol. 46, no. 6, pp. 900–915, 2007. [Online]. Available: http://journals.ametsoc.org/doi/abs/10.1175/JAM2493.1
3. J. Chen and A. K. Gupta, *Parametric statistical change point analysis*. Birkhauser, 2000.
4. C. Alippi, G. Boracchi, V. Puig, and M. Roveri, "An ensemble approach to estimate the fault-time instant," in *Proceedings of the 4th International Conference on Intelligent Controland Information Processing (ICICIP 2013)*, 2013.
5. C. Alippi, D. Liu, D. Zhao, and L. Bu, "Detecting and reacting to changes in sensing units: The active classifier case," *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2013.
6. C. Alippi, G. Boracchi, and M. Roveri, "Just-in-time classifiers for recurrent concepts," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 24, no. 4, pp. 620–634, April.
7. D. M. Hawkins, "Testing a sequence of observations for a shift in location," *Journal of the American Statistical Association*, vol. 72, no. 357, pp. pp. 180–186, 1977. [Online]. Available: http://www.jstor.org/stable/2286934
8. A. N. Pettitt, "A Non-Parametric Approach to the Change-Point Problem," *Applied Statistics*, vol. 28, no. 2, pp. 126–135, 1979. [Online]. Available: http://dx.doi.org/10.2307/2346729
9. D. M. Hawkins, P. Qiu, and C. W. Kang, "The changepoint model for statistical process control," *Journal of Quality Technology*, vol. Vol. 35, No. 4, pp. 355–366, 2003.
10. F. Gustafsson, *Adaptive Filtering and Change Detection*. Wiley, Oct. 2000. [Online]. Available: http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471492876,descCd-description.html
11. M. B. Perry and J. J. Pignatiello, "Identifying the time of step change in the mean of autocorrelated processes," *Journal of Applied Statistics*, vol. 37, no. 1, pp. 119–136, 2010.
12. T. Dietterich, "Ensemble methods in machine learning," *Multiple classifier systems*, pp. 1–15, 2000.
13. Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall, 2012.
14. A. Krogh and P. Sollich, "Statistical mechanics of ensemble learning," *Physical Review E*, vol. 55, no. 1, p. 811, 1997.
15. J. Wichard and M. Ogorzalek, "Time series prediction with ensemble models," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, vol. 2. IEEE, 2004, pp. 1625–1630.
16. E. Mangalova and E. Agafonov, "Time series forecasting using ensemble of ar models with time-varying structure," in *Evolving and Adaptive Intelligent Systems (EAIS), 2012 IEEE Conference on*. IEEE, 2012, pp. 198–203.
17. C. Alippi, S. Ntalampiras, and M. Roveri, "Model ensemble for an effective on-line reconstruction of missing data in sensor networks," in *International Joint Conference on Neural Networks (IJCNN 2013)*, 2013.
18. G. J. Ross, D. K. Tasoulis, and N. M. Adams, "Nonparametric monitoring of data streams for changes in location and scale," *Technometrics*, vol. 53, no. 4, pp. 379–389, 2011.
19. H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947. [Online]. Available: http://dx.doi.org/10.2307/2236101
20. A. M. Mood, "On the asymptotic efficiency of certain nonparametric two-sample tests," *The Annals of Mathematical Statistics*, vol. Vol. 25, No. 3, pp. 514–522, September 1954.
21. Y. Lepage, "A combination of Wilcoxon's and Ansari-Bradley's statistics," *Biometrika*, vol. Vol. 58, No. 1, pp. 213–217, April 1974.
22. G. Ross and N. M. Adams, "Two nonparametric control charts for detecting arbitrary distribution changes," *Journal of Quality Technology*, vol. Vol 44, No. 22, pp. 102–116, 2012.
23. T. W. Anderson, "On the distribution of the two-sample Cramer-von Mises criterion," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1148–1159, 1962.
24. D. M. Hawkins and K. D. Zamba, "A change-point model for a shift in variance," *Journal of Quality Technology*, vol. 37, no. 1, pp. 21–31, 2005.
25. K. D. Zamba and D. M. Hawkins, "A multivariate change-point model for statistical process control," *Technometrics*, vol. 48, no. 4, pp. pp. 539–549, 2006. [Online]. Available: http://www.jstor.org/stable/25471246
26. M. Basseville and I. V. Nikiforov, *Detection of abrupt changes: theory and application*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.
27. R. Isermann, *Fault-diagnosis systems: an introduction from fault detection to fault tolerance*. Springer Verlag, 2006.
28. C. Alippi, G. Boracchi, and M. Roveri, "A Just-In-Time adaptive classification system based on the Intersection of Confidence Intervals rule," *Neural Networks*, vol. 24, no. 8, pp. 791 – 800, 2011.
29. L. Ljung, *System identification*. Wiley Online Library, 1999.
30. G. J. Ross, "Parametric and nonparametric sequential change detection in R: The cpm package," *Journal of Statistical Software*, Forthcoming.
31. L. Ljung, *System identification: theory for the user*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1986.
32. H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716 – 723, dec 1974.
33. C. Alippi, G. Boracchi, V. Puig, and M. Roveri, "A hierarchy of change-point methods for estimating the time instant of leakages in water distribution networks," in *Proceedings of LEAPS, the 1st Workshop on Learning stratEgies and dAta Processing in nonStationary environments, in 9th AIAI Conference*, September 2013, pp. 1–10.