Pattern Recognition 134 (2023) 109116

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Multi-Source Change-Point Detection over Local Observation Models

Lorena Romero-Medrano[a,b,*], Antonio Artés-Rodríguez[a,b]

[a] Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain
[b] Evidence-Based Behavior (eB^2), Leganés, Spain

ARTICLE INFO

Article history:
Received 16 August 2021
Revised 2 August 2022
Accepted 16 October 2022
Available online 23 October 2022

Keywords:
Change-point detection
Multi-source data
Heterogeneous data
Latent variable models

ABSTRACT

In this work, we address the problem of change-point detection (CPD) on high-dimensional, multi-source, and heterogeneous sequential data with missing values. We present a new CPD methodology based on local latent variable models and adaptive factorizations that enhances the fusion of multi-source observations with different statistical data-type and face the problem of high dimensionality. Our motivation comes from behavioral change detection in healthcare measured by smartphone monitored data and Electronic Health Records. Due to the high dimension of the observations and the differences in the relevance of each source information, other works fail in obtaining reliable estimates of the change-points location. This leads to methods that are not sensitive enough when dealing with interspersed changes of different intensity within the same sequence or partial missing components. Through the definition of local observation models (LOMs), we transfer the local CP information to homogeneous latent spaces and propose several factorizations that weight the contribution of each source to the global CPD. With the presented methods we demonstrate a reduction in both the detection delay and the number of not-detected CPs, together with robustness against the presence of missing values on a synthetic dataset. We illustrate its application on real-world data from a smartphone-based monitored study and add explainability on the degree of each source contributing to the detection.

1. Introduction

The problem of change-point detection (CPD) over a sequence of observations aims to identify abrupt variations in the data distribution, which we refer to as change-points (CPs). This is frequently found in real-world scenarios where tendencies or events need to be detected like finances [12], speech-recognition [8], healthcare [4,15], or nature disaster assessment [20]. The methods that tackle this problem can be categorized as online/offline and model-based/non-parametric. Among the model-based methods, we differentiate between frequentist and Bayesian approaches. A complete review of CPD algorithms can be seen in van den Burg and Williams [5], Truong et al. [21]. This work is based on the online Bayesian approach. There are several works that focus on Bayesian estimation of CPs contingent upon mixture models [7], hidden Markov models (HMMs) [9], and other classification methods [16,18]. We consider the Bayesian Online Change-Point Detection method (BOCPD) presented in Adams and MacKay [1] as the basis of this work.

These methods are generally applied to low-dimensional datasets where all the variables have the same data-type or similar generative properties. However, real-world scenarios are typically composed of high-dimensional heterogeneous variables. Our work is motivated by the use of CPD methods to identify changes in Human Behavioral patterns over both smartphone monitored data and Electronic Health Records (EHR) for healthcare applications. In particular, we are interested in detecting psychiatric crisis in mental health patients using mobility and mobile usage patterns obtained by the preprocessing of digital measurements [2]. In these cases, observations are (i) high-dimensional with information from different sources. This is because we work with data from smartphones, medical tests, diagnoses, questionnaires, demographic information, etc. [3]. Moreover, they are (ii) heterogeneous, i.e., we simultaneously deal with different statistical data-type variables like continuous (weight, lab measurements), binary (medical history questions) or categorical (type of medication), or equal data-type variables with different marginal generative distributions. Also, smartphone data exhibit (iv) missing values due to sensor failures and the need for privacy permissions, among others [13].

[*] Corresponding author at: Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain.

E-mail addresses: lromero@tsc.uc3m.es (L. Romero-Medrano), antonio@eb2.tech (A. Artés-Rodríguez).

https://doi.org/10.1016/j.patcog.2022.109116
0031-3203/

To tackle this problem, the BOCPD algorithm [1] allows to use a Bayesian approach to integrate the presence of missing data but results in noisy detection when dealing with high-dimensional observations because of the need for large statistical evidence to differentiate between noise drifts and real CPs. In [14], a hierarchical extension of this method is introduced where they assume that there is a unique univariate latent representation that simultaneously summarizes the statistical information of every source. This approach solves the high-dimensional data problem. However, the latent variable modeling implies the use of different likelihood functions and entails an optimization problem over a product of functions with different support that results in some variables underrepresented in favor of others, loosing essential generative information for the global detection.

This joint dimensionality reduction has an implicit smoothing effect, making the method not sufficiently sensitive when dealing with interspersed changes of different intensity within the same sequence. Solely the high-intensity CPs are detected in this cases. Even though, the presence of missing temporal data for just a subset of sources can increase this smoothing effect, motivating the search for a more sensitive way to fuse all the sources while taking into account the aforementioned features of the data.

To overcome the limitations of the described setting, we propose a Change-Point Detector based on Local Observation Models (LOM-based CPD) that generalizes and extends the use of latent variables models for change-point detection. The LOM-based CPD tackles the problem in a two-stage modeling method. In the first stage, we propose several Local Observation Models (LOMs) that are based on partitioning the feature space depending on the context-meaning, multi-source and mixed-type nature of the data. This allows the dimensionality reduction of the observations and control over how the local CP information is transfered to homogeneous local spaces, implying technical advantages in the inference process and solving the heterogeneous initial problem. In the second stage, different Factorizations Models for the CP detector are proposed to consider several weighting mechanisms for the homogeneous local latent representations obtained from the first stage, resulting in a generalized hierarchical CPD methodology that holds for any observation model previously introduced.

Our main contributions are:

- We present the LOM-based CPD that allows us to perform online CP detection over multi-source, high-dimensional, and heterogeneous data (Section 2). We propose different LOMs to partition the feature space and transform the heterogeneous original data into homogeneous local latent representations (Section 3).
- We propose several CPD Factorization Models (Section 5) to calibrate the contribution of each local set (independence, mixture-weighting and beta-weighting).
- We evaluate the performance of the methods in terms of precision rate and delay in the detection on a synthetic dataset. We include an experiment on a smartphone-based monitored dataset of a real human behavior study and explainability about the contribution of each source to the global detection (Section 6).

2. CPD over Local Observation Models (LOMs)

Let $\mathbf{x}_{1:t}$ be a sequence of mixed-type high-dimensional observations where $\mathbf{x}_i = [x_{i1}, \dots, x_{iN}]$ is composed of N heterogeneous data sources $\forall i \in \{1, \dots, t\}$. Each observation x_{in} has its own datatype $m \in \{1, \dots, M\}$ and high dimension, d_n, that is constant for every observation. We consider the scenario proposed in Adams and MacKay [1] and assume that $\mathbf{x}_{1:t}$ may be divided into non-overlapping *temporal partitions* $\rho = \{1, 2, \dots\}$ separated by change-points (CPs), where each partition ρ has a surrogate generative distribution $p(\mathbf{x}|\theta_\rho)$ with unknown parameters θ_ρ and observations are i.i.d. within it. Our goal is to find both (i) the location of the CPs that define the beginning and end of each temporal partition and (ii) the unknown parameters of the distribution within them, all this in an online manner.

When dealing with high-dimensional data as in our case, the complexity of the generative distribution $p(\mathbf{x}_t|\theta)$ increases naturally with the dimension of the observations $\mathbf{x}_t$ leading to an extremely large set of parameters θ_t to estimate at each time-step. In these cases, we need proportional statistical evidence to feasibly update the posterior distribution given $\mathbf{x}_{1:t}$. Otherwise, the CPs are typically confounded with noise drifts in the underlying parameters, leading to a noisy detection. Moreover, the original method [1] works over homogeneous data sequences while our setting is composed of different statistical type sources.

To tackle the high-dimensionality problem, latent variable models are frequently used because they allow the transformation of the initial heterogeneous space into a new one where we can make decisions about its dimensionality or nature (continuous, discrete) while keeping the generative characteristics of the original observations. In [14], they apply this strategy to mixed-type data under the hypothesis that there is a low-dimensional latent representation of the data, $\mathbf{z}_{1:t}$, where the true CPs lie. They propose the consideration of heterogeneous likelihoods to obtain a univariate latent representation where the change-point detection is later performed. However, this assumption has an implicit smoother effect that may lose the detection of low-intensity changes when they are interspersed with high-intensity ones. Moreover, the optimization problem over a product of functions with different support usually leads to uncalibrated modeling of the variables. Consequently, information is lost and noise is introduced in the detection.

We generalize and extend the work in Moreno-Muñoz et al. [14], and present a LOM-based CPD methodology that allows to carry out the detection over homogeneous local latent representations of the original data, assuming different factorization models for the local contributions. General flow of is depicted in Fig. 1. We consider that there exist D partitions of the feature space $\mathbf{x}_{1:t} = \{x^d_{1:t}\}^D_{d=1}$ (*local sets*) such that their associated latent representations $\mathbf{z}_{1:t} = \{z^d_{1:t}\}^D_{d=1}$ (*local latent representations*) are independent and the characteristic information of the local CPs is fully contained. With this approach we directly solve the high-dimensionality and mixed-type data problems because we reduce the high dimension of the heterogeneous space of observations $\mathcal{X}$ to a low-dimensional homogeneous space of local latent representations $\mathcal{Z}$. We choose to work with univariate discrete variables to keep a lower number of parameters to estimate and to simplify the interpretability of the algorithm. That is, $z^d_t \in \{1, \dots, K_d\}$ $\forall t, d$, with K_d the number of latent classes defined for local set d. We now deal with multi-source information that is not heterogeneous anymore, allowing an equivalent level of treatment for all of them.

As an example, let us consider that we aim to study behavioral changes of a person. We represent the mobility at each day with a 48-dim real variable that counts the distance walked by half an hour time-slots. Additionally, we consider a 48-dim binary variable representing whether or not the user has been at home. Another one indicates if the user has used or not WhatsApp. We can have a categorical variable expressing the activity of the person (running, walking) in each time-slot. In this case, the feature space could be partitioned by source, resulting in $D = 4$ local sets. Or, we could assume that the distance walked and activity performed conform to a unique one, having $D = 3$ local sets and designing structured relationship between them through the local latent modeling. Different Local Observation Models (LOMs) are presented in Section 3.

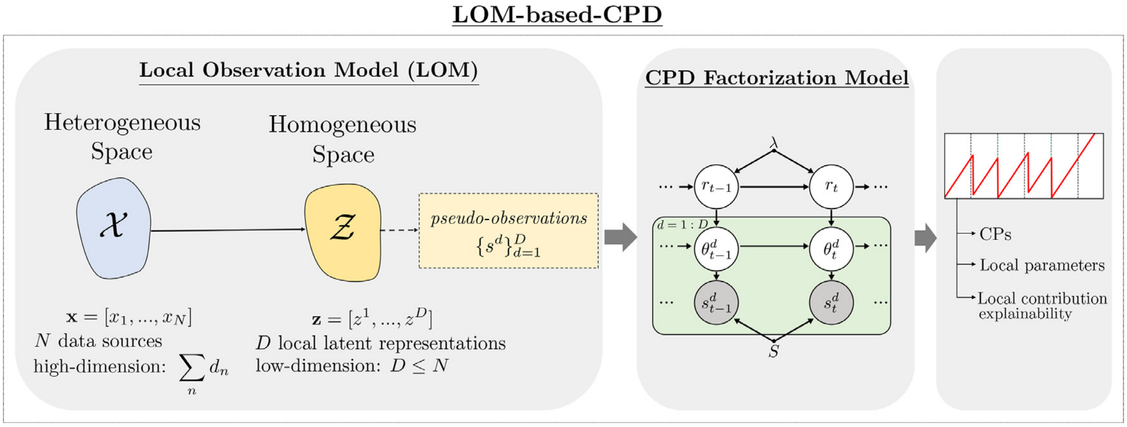

Fig. 1. LOM-based CPD flow.

In all the cases, the result is a set of posterior distributions $\left\{p(z_t^d \mid x_t^d)\right\}_{d=1}^D$ with size equal to D, the number of local sets considered. We follow the approach of Romero-Medrano et al. [17] by drawing S i.i.d. samples of each posterior distribution and transforming them into *counting vectors* $s_t^d \in \mathbb{Z}_+^{K_d}$, where the kth component $s_t^{d,k}$ counts the times that class k has been drawn from local posterior d. We consider this definition because it will allow us to assume multinomial generative distributions that maintain inference analytically tractable while keeping the most information as possible from the posterior distributions.

Finally, we can consider the sequence of counting vectors $\mathbf{s}_{1:t} = \left\{s_{1:t}^d\right\}_{d=1}^D$ as homogeneous input *pseudo-observations* to perform the change-point detection. Up to now, we have not assumed statistical dependence between the local sets for any LOM presented, but there could actually be dependences between them. In the previous example, the variable *WhatsApp usage* could be conditioned by the variable *Being at home*. We can assume independence over the variables or give some flexibility by modeling their relationship through different weighting mechanisms to calibrate their contribution. Each option gives specific benefits and structure complexity. We propose different factorizations models for the CP detector in Section 5. The graphical representations of these relationships are shown in Fig. 3. The first one assumes that every local set distribution conditioned on the generative parameters is independent from the others. The second one is considered a mixture model of every local set distribution, with or without memory of past contributions. The third one poses a Beta prior on the weights, allowing different combinations of variables contributing at each timestep. The last two approaches allow us to obtain explainability on the degree of contribution by each local set to the global detection at each time-step, and to directly cancel its contribution when the whole local set is missed.

3. Local Observation Models (LOMs)

The generative model for each local set d and time-step t is expressed as

$$p(x_t^d|\theta_t^d) = \int p(x_t^d|z_t^d)p(z_t^d|\theta_t^d)dz_t^d \tag{1}$$

where $p(x_t^d|z_t^d)$ is assumed to be fixed and $p(z_t^d|\theta_t^d)$ is the distribution over the latent variable that can be either continuous or discrete, but it is always of the same nature for every local set $\in \{1, ., D\}$. We choose to consider categorical univariate latent variables $z_t^d \in \{1, \ldots, K\}$ $\forall$ d, t, while both the factorization and nature of the likelihoods $\{p(x_t^d|z_t^d)\}_{d=1}^D$ depend on the sources that compose each local set and thus, how we define them. The inference is carried out separately for each local set d via the online Expectation-Maximization algorithm (EM) [6]. With this approach we reduce the dimensionality of each observation at instant t from $\sum_{n=1}^N d_n$ to D, the number of local sets considered.

We present different LOMs that are built through the definition of several partitioning alternatives of the feature space into D sets to consider different scenarios and benefit from the specific characteristics of the original sequence. They are summarized in Table 1.

3.1. Full joint representation (Joint OM)

In this approach, we consider a univariate latent representation z_t for the whole observation $\mathbf{x}_t$, assuming that there is a unique latent representation that holds the generative characteristics of every source simultaneously. This observation model is presented in Moreno-Muñoz et al. [14] and corresponds with the $D = 1$ case. They propose an heterogeneous mixture model with likelihood

$$p(\mathbf{x}_t \mid z_t, \{\phi_{k1}, \ldots, \phi_{kN}\}_{k=1}^K) = \prod_{k=1}^K \prod_{n=1}^N p(x_{tn} \mid \phi_{kn})^{\mathbb{1}\{z_t = k\}} \quad \forall t, \tag{2}$$

where $z_t \in [1, 2, \ldots, K]$ indicates which component of the mixture is active in observation t, K is the total number of components of the mixture model, and ϕ_{kn} are the generative parameters of component k and source n. Given the class z_t and the parameters, the variables $x_{t1}, \ldots, x_{tN}$ are considered independent. Note that, in this observation model, the likelihood for each component k is a product of mixed-type likelihoods, each one defined based on the nature of each source. With this approach, we obtain one posterior probability distribution $p(z_t \mid \mathbf{x}_t)$ for each time-step t.

3.2. Independent source representation (Sources OM)

Given an observation $\mathbf{x}_t$, in this approach we define an observation model based on the assumption that there exists a latent representation for each data source x_{tn}, having $x_t^d := x_{tn}$ and $D = N$, i.e., the number of local sets is equal to the number of sources. The likelihood of each local set d is

$$p(x_t^d \mid z_t^d, \{\phi_k^d\}_{k=1}^{K_d}) = \prod_{k=1}^{K_d} p(x_t^d \mid \phi_k^d)^{\mathbb{1}\{z_t^d = k\}} \quad \forall d, t, \tag{3}$$

where K_d is the dimension of the latent variable for source d. Note that the likelihood of each local set, $p(x_t^d \mid \phi_k^d)$, is no longer heterogeneous, because each set d (each source) has its own data-type. With this approach we not only solve the high-dimensionality problem, but we also avoid the product of mixed-typed likelihoods

Table 1
Summary of local observation models (LOMs).

Local observation models		
LOM	Likelihood	Number of sets
Joint OM	$p(\mathbf{x}_t \mid z_t, \{\phi_{k1}, \ldots, \phi_{kN}\}_{k=1}^{K}) = \prod_{k=1}^{K} \prod_{n=1}^{N} p(x_{tn} \mid \phi_{kn})^{\mathbb{1}\{z_t=k\}}$	$D = 1$
Sources OM	$p(x_t^d \mid z_t^d, \{\phi_k^d\}_{k=1}^{K_d}) = \prod_{k=1}^{K_d} p(x_t^d \mid \phi_k^d)^{\mathbb{1}\{z_t^d=k\}}$	$D = N$, nb. of sources
Grouped OM	$p(x_t^d \mid z_t^d, \{\phi_k^d\}_{k=1}^{K_d}) = \prod_{k=1}^{K_d} p(x_t^d \mid \phi_k^d)^{\mathbb{1}\{z_t^d=k\}}$	$D = M$, nb. of data-types
Context OM	$p(x_t^d \mid z_t^d, \{\phi_{k1}^d, \ldots, \phi_{kN_d}^d\}_{k=1}^{K_d}) = \prod_{k=1}^{K_d} \prod_{n=1}^{N_d} p(x_{tn}^d \mid \phi_{kn}^d)^{\mathbb{1}\{z_t^d=k\}}$	Depends on the prior knowledge

that can bias the resulting posterior for the latent classes. We transform each source to a latent space that has the same nature for all of them, independently of the initial type. Hence, the result in this case is not only one posterior probability distribution at time-step t, but a new set $\left\{p(z_t^d \mid x_t^d)\right\}_{d=1}^{D}$, whose size is equal to the number of sources, N.

3.3. Data-type based representation (Grouped OM)

This approach is based on the previous one and motivated by the technical advantage of avoiding the product of mixed-type likelihoods. We propose a partition of the feature space based on the data-type of the sources, assuming that there is a latent representation for each group. We stack the same statistical type sources having $x_t^d := [x_{tn}$ s.t. source n is of type m] and $D = M$, the number of data-types. The likelihood is expressed in Eq. (3). Like the Sources OM, the likelihood of each local set is no longer heterogeneous. The difference is that, in this case, the set x_t^d based on type m has dimension $\sum d_n$ for every source n of type m. The result is the set $\left\{p(z_t^d \mid x_t^d)\right\}_{d=1}^{D}$ with size equal to the number of data-types, M.

3.4. Prior knowledge based representation (Context OM)

In this approach, we propose to group the sources using contextual information of the data such as external relations between sources due to the collection method or context meaning. For example, in a behavioral modeling scenario, we can benefit from existing correlations between variables like the number of steps walked in a day and the distance traveled, both computed from location traces, reducing the noise in the latent variable modeling. We can also define local sets by life domains like mobility, physical activity or social interactions, to study behavioral changes over domains instead of over variables, that could be more informative in a health analysis context.

In the limit of considering a unique group, $D = 1$, or each source separately, $D = N$, we have the observation models presented in Sections 3.3 and 3.2, respectively. The likelihood is the generalization of Eqs. (2) and (3),

$$p(x_t^d \mid z_t^d, \left\{\phi_{k1}^d, \ldots, \phi_{kN_d}^d\right\}_{k=1}^{K_d}) = \prod_{k=1}^{K_d} \prod_{n=1}^{N_d} p(x_{tn}^d \mid \phi_{kn}^d)^{\mathbb{1}\{z_t^d=k\}} \quad \forall d, t. \tag{4}$$

K_d is the dimension of the latent variable associated with local set d and N_d is the number of sources conforming to that set. The result is the set of posterior distributions $\left\{p(z_t^d \mid x_t^d)\right\}_{d=1}^{D}$ whose size depends on the prior knowledge we considered.

4. LOM-based CPD

We introduce a hierarchical CPD generalization, that we call LOM-based CPD, and extends the original BOCPD, allowing to perform change-point detection over any local observation model presented in Section 3.

4.1. Background: Bayesian online change-point detection

Our work is based on the Bayesian Online Change-Point Detection (BOCPD) method [1], where the change-point detection problem is tackled by the introduction of a new auxiliary discrete variable r_t, the *run-length*, that counts the number of time-steps since the last CP so, at each time-step t: $r_t = 0$ if there is a CP at time t and, $r_t = r_{t-1} + 1$, otherwise. The objective of the method is to obtain the posterior distribution of r_t given the data, $p(r_t|\mathbf{x}_{1:t})$, that provides an uncertainty measure for the last CP location at each time-step. As an example, for $t = 100$, $p(r_{100} = 5|x_{1:100})$ measures the probability that a change-point happened 5 time-steps ago, at $t = 95$. $p(r_t|\mathbf{x}_{1:t})$ can be inferred at each time-step in a recursive manner, based on the normalization of the joint distribution $p(r_t, \mathbf{x}_{1:t})$. For its computation, the underlying generative model $p(\mathbf{x}_t|\theta)$ needs to be defined a priori and the parameters have to be updated at each time-step considering the new arrived observation.

4.2. LOM-based CPD

Given the set $\mathbf{s}_{1:t} = \left\{s_{1:t}^d\right\}_{d=1}^{D}$ with $t \in \{1, \ldots T\}$, we want to find the joint distribution of a global run-length r_t and the given pseudo-observations for every time-step. We can build up the joint distribution based on [1],

$$\begin{aligned} p(r_t, \mathbf{s}_{1:t}) &= \sum_{r_{t-1}} p(r_t, r_{t-1}, \mathbf{s}_{1:t}) = \sum_{r_{t-1}} p(r_t, r_{t-1}, \mathbf{s}_t, \mathbf{s}_{1:t-1}) \\ &= \sum_{r_{t-1}} p(r_t, \mathbf{s}_t | r_{t-1}, \mathbf{s}_{1:t-1}) p(r_{t-1}, \mathbf{s}_{1:t-1}) \\ &= \sum_{r_{t-1}} \underset{p(r_t|r_{t-1})}{p(r_t | \mathbf{s}_t, r_{t-1}, \mathbf{s}_{1:t-1})} \underset{\text{predictive posterior}}{p\left(\mathbf{s}_t \middle| r_{t-1}, \mathbf{s}_{1:t-1}^{(r)}\right)} \underset{\text{recursive term}}{p(r_{t-1}, \mathbf{s}_{1:t-1})}. \end{aligned} \tag{5}$$

This formulation allows the recursive computation of $p(r_t, \mathbf{s}_{1:t})$ at each time-step. The first term works as a prior of the change-point, and is considered as proposed in the original paper: $p(r_t|r_{t-1}) = H(r_{t-1} + 1)$ if $r_t = 0$ and $p(r_t|r_{t-1}) = 1 - H(r_{t-1} + 1)$ if $r_t = r_{t-1} + 1$. $H(\cdot)$ is the *hazard function* that for the choice of the geometric distribution becomes constant $H = \frac{1}{\lambda}$ and dependent on the hyperparameter λ [10]. The third term can be computed recursively due to the nature of Eq. (5). The second term is the joint predictive posterior distribution of the new observation $\mathbf{s}_t$ given the sequence of past observations $\mathbf{s}_{1:t-1}$. The conditioning on r_{t-1} represents the current partition, r, that started r_{t-1} time-steps ago, and $\mathbf{s}_{1:t-1}^{(r)}$ denotes the subsequence of $\mathbf{s}_{1:t-1}$ contained on r. The evaluation of this joint predictive term depends on (i) the underlying generative model assumed for each local set d and (ii) the relationship between all of them. The first point is addressed in Section 4.3, while several structures for (ii) are discussed and presented in Section 5.

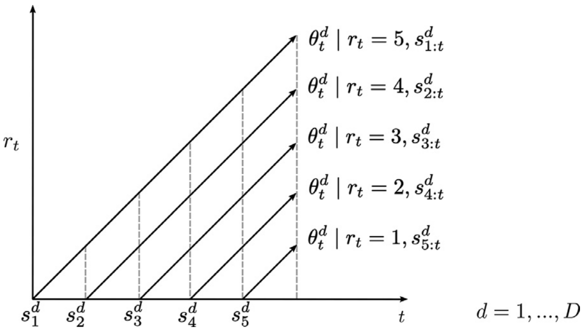
$\theta_t^d \mid r_t = 5, s_{1:t}^d$
$\theta_t^d \mid r_t = 4, s_{2:t}^d$
$\theta_t^d \mid r_t = 3, s_{3:t}^d$
$\theta_t^d \mid r_t = 2, s_{4:t}^d$
$\theta_t^d \mid r_t = 1, s_{5:t}^d$
r_t
s_1^d
s_2^d
s_3^d
s_4^d
s_5^d
t
$d = 1, ..., D$

Fig. 2. Illustration of the inference mechanism developed to estimate the local set parameter distribution in parallel at each possible partition at instant t.

4.3. Local likelihood models and inference

We consider independent generative models for each local set $1, \ldots, D$, resulting in D parallel inference processes to obtain the set of local posterior distributions $\{p(s_t^d | r_{t-1}, s_{1:t-1}^d)\}_{d=1}^D$. Marginalizing out, we can express each local posterior d as

$$p(s_t^d | r_{t-1}, s_{1:t-1}^{d,(r)}) = \int p(s_t^d | \theta_t^{d,(r)}) p(\theta_t^{d,(r)} | r_{t-1}, s_{1:t-1}^{d,(r)}) d\theta_t^{d,(r)}, \quad (6)$$

where $\theta_t^{d,(r)}$ are the generative parameters of the distribution at time-step t and each possible current partition r, that is expressed as the conditioning on the run-length r_{t-1}. We can estimate $p(\theta_t^{d,(r)} | r_{t-1}, s_{1:t-1}^{d,(r)})$ recursively, and, following the thread mechanism depicted in Fig. 2, we can evaluate the likelihood once the new observation is given. For each local set d and instant t we assume that $s_1^d, \ldots, s_t^d$ are independent and multinomial distributed within each temporal partition r. We also consider a prior for the parameters, having

$$s_t^d \sim \text{Multinomial}(\theta_t^d, S), \qquad \theta_t^d \sim \text{Dirichlet}(\alpha^d), \quad (7)$$

where $\boldsymbol{\alpha}^d \in \mathbb{R}_+^K$ and the likelihood expression of $\mathbf{s}_t^d$ is $p(s_{t,1}^d, \ldots, s_{t,K}^d | \theta_t^d, S) = \frac{S!}{\prod_{k=1}^K s_{t,k}^d!} \prod_{k=1}^K (\theta_{t,k}^d)^{s_{t,k}^d}$. The posterior update of the parameters has the following closed form $\tilde{\alpha}^d = \boldsymbol{\alpha}^d + s_t^d$. This allows a direct update when a new sample is observed, and the inference of the posterior distribution of the parameter vector θ_t^d related to the current partition and the data within it. Now, we are able to compute the predictive term, $\Psi_t^{d,(r)} := p(s_t^d | r_{t-1}, s_{1:t-1}^{d,(r)})$, that is a function of the statistics of the model,

$$\Psi_t^{d,(r)} = \prod_{k=1}^K \prod_{j=0}^{s_{t,k}^d - 1} \frac{\alpha_{t-1}^k + j}{S_\alpha + S_c^{(k-1)} + j} \frac{S_c^{(k-1)} + j + 1}{j + 1}, \quad (8)$$

with $S_c^{(k-1)} := \sum_{l=1}^{k-1} s_{t,l}^d \quad \forall k = 1 \ldots K$.

To carry out the full inference for the detection we need the joint predictive distribution conditioned on the current partition and associated data,

$$p(\mathbf{s}_t | r_{t-1}, \mathbf{s}_{1:t-1}) = p(s_t^1, \ldots, s_t^D | r_{t-1}, s_{1:t-1}^1, \ldots, s_{1:t-1}^D). \quad (9)$$

This is direct for the Joint OM where $D = 1$ [17]. In Section 5 we address the problem of how to estimate this joint predictive distribution when $D > 1$ assuming different CPD factorization models over the local predictives $\{p(s_t^d | r_{t-1}, s_{1:t-1}^d)\}_{d=1}^D$.

4.4. Definition of change-points

This method allows us to obtain the posterior distribution of the run-length r_t in a recursive manner. Given the joint distribution $p(r_t, \mathbf{s}_{1:t})$, we can normalize, $p(r_t \mid \mathbf{s}_{1:t}) = \frac{p(r_t, \mathbf{s}_{1:t})}{\sum_{r_t} p(r_t, \mathbf{s}_{1:t})}$. Once the posterior $p(r_t \mid \mathbf{s}_{1:t})$ is obtained, we define the sequence of *maximum-a-posteriori* (MAP) estimates $\{r_{1:t}^*\}$, with $r_t^* = \arg\max p(r_t \mid \mathbf{s}_{1:t}) \ \forall t$, which we use to find the most likely change-points.

Table 2
Summary of CPD factorization models.

CPD factorization models	
Factorization	Expression
Independent CPD	$p(\mathbf{s}_t | r_{t-1}, \mathbf{s}_{1:t-1}) = \prod_{d=1}^D p(s_t^d | r_{t-1}, s_{1:t-1}^d)$
Mixture CPD	$p(\mathbf{s}_t | r_{t-1}, \mathbf{s}_{1:t-1}) = \sum_{d=1}^D \pi_{t,r}^d \cdot p(s_t^d | r_{t-1}, s_{1:t-1}^d)$
	with $\sum_{d=1}^D \pi_{t,r}^d = 1$
Weighted CPD	$p(\mathbf{s}_t | r_{t-1}, \mathbf{s}_{1:t-1}) = \max_{\mathbf{w}_{t,r}} \ p(\mathbf{w}_{t,r}) \cdot \prod_{d=1}^D p(s_t^d | r_{t-1}, s_{1:t-1}^d)^{w_{t,r}^d}$
	with $p(\mathbf{w}_{t,r}) = \prod_{d=1}^D \text{Beta}(w_{t,r}^d; a, b)$

5. CPD Factorization Models

We present three new factorizations of the CP detector to fuse the multi-source information through the construction of the joint predictive distribution $p(\mathbf{s}_t | r_{t-1}, \mathbf{s}_{1:t-1})$ from the homogeneous set of local predictive ones, $\{p(s_t^d | r_{t-1}, s_{1:t-1}^d)\}_{d=1}^D$. Two of these approaches, additionally, allow us to obtain explainability on the degree of each local set contributing to the global detection at each time-step and, to directly cancel its contribution when the whole local set is missed. As detailed in Section 4, each local predictive distribution is computed by taking into account the data related to each temporal partition and local set separately, as expressed in Eq. (8) (Table 2).

5.1. Independent product (Independent CPD)

We assume that parameters θ_t^d of each local set d conditioned to the current temporal partition r and past data of the same set, $s_{1:t-1}^d$, are independent. The relations between variables are graphically represented in Fig. 3(a). Given $\theta = (\theta_t^1, \ldots, \theta_t^D)$, we obtain the following factorization,

$$\begin{aligned} p(\mathbf{s}_t | r_{t-1}, \mathbf{s}_{1:t-1}) &= \int p(\mathbf{s}_t | \theta) p(\theta | r_{t-1}, \mathbf{s}_{1:t-1}) d\theta \\ &= \int \prod_{d=1}^D p(\mathbf{s}_t | \theta_t^d) p(\theta_t^d | r_{t-1}, \mathbf{s}_{1:t-1}) d\theta \\ &= \prod_{d=1}^D \int p(\mathbf{s}_t | \theta_t^d) p(\theta_t^d | r_{t-1}, \mathbf{s}_{1:t-1}) d\theta_t^d \\ &= \prod_{d=1}^D p(s_t^d | r_{t-1}, s_{1:t-1}^d), \end{aligned}$$

where the third equality follows from the independence assumption. The resulting joint predictive is the product of the local predictive distributions and we call this CPD version *Independent CPD*.

5.2. Mixture weighting by temporal partition (Mixture CPD)

We consider that there exists a hidden variable that acts as a weighting mechanism deciding the contribution of each local predictive distribution $p(s_t^d | r_{t-1}, s_{1:t-1}^d)$ to the global detection. In particular, we assume that there is a different variable $h_{t,r}$ for each time-step t and partition r that takes values from 1 to D with

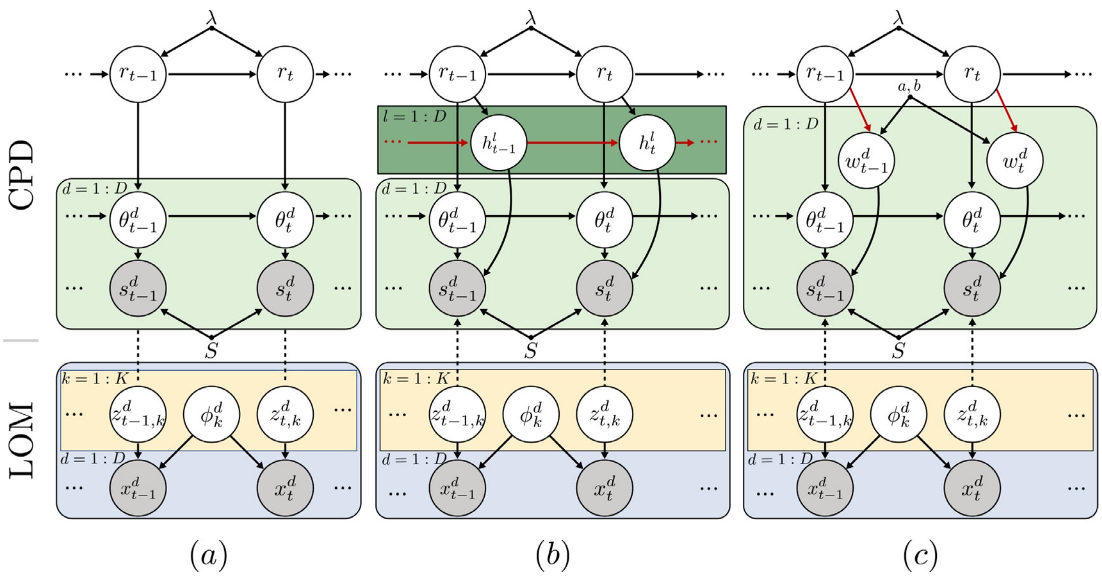

Fig. 3. Graphical representation of LOM (blue boxes) and CPD (upper region) stages, both connected through dashed lines. (a) Graphical representation of Independent CPD. (b) Graphical representation of Mixture CPD. Red lines indicate the Mixture Memory scenario. (c) Graphical representation of Weighted CPD. Red lines indicate the Weighted Partition scenario. Details explained in Appendix D. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

probabilities $\pi_{t,r} = (\pi^1_{t,r}, \dots, \pi^D_{t,r})$, as depicted in the graphical representation shown in Fig. 3(b). Thus, we have a weights matrix $\{\pi^d_{t,r}\}_{r,d}$ of size $D \times t$ at each time-step. The resulting factorization is

$$p(\mathbf{s}_t|r_{t-1}, \mathbf{s}_{1:t-1}) = \sum_{d=1}^{D} \pi^d_{t,r} \cdot p(s^d_t|r_{t-1}, s^d_{1:t-1})$$

$$\text{with} \quad \sum_{d=1}^{D} \pi^d_{t,r} = 1.$$

The model can be seen as a mixture model at each temporal partition, where the d component distribution is the predictive distribution associated to local set d. The parameters of each component have to be previously inferred separately, as explained in Section 5.1. To estimate the weights $\pi^d_{t,r}$, we can solve an optimization problem at each time-step t and partition r, that is to find

$$\tilde{\pi}^1_{t,r}, \dots, \tilde{\pi}^D_{t,r} = \operatorname*{arg\,max}_{\pi^1_{t,r},\dots,\pi^D_{t,r}} \sum_{d=1}^{D} \pi^d_{t,r} \cdot p\left(s^d_t \middle| r_{t-1}, s^d_{1:t-1}\right)$$
$$\text{s.t.} \quad \sum_{d=1}^{D} \pi^d_{t,r} = 1, \quad \pi^d_{t,r} \geq 0 \quad \forall d, \tag{10}$$

where we have noted $p^d_{t,r} := p(s^d_t|r_{t-1}, s^d_{1:t-1})$. We refer to $\{\tilde{\pi}^d_{t,r}\}_d$ as *partial weights*. Since $p^d_{t,r} \geq 0$ $\forall d$, and considering the two constrictions of Eq. (10), we have that

$$0 \leq \sum_{d=1}^{D} \pi^d_{t,r} \cdot p^d_{t,r} \leq \sum_{d=1}^{D} \pi^d_{t,r} \cdot \max_d p^d_{t,r} \leq \max_d p^d_{t,r}. \tag{11}$$

Therefore, the solution to the optimization problem (10) is

$$\tilde{\pi}^{d^*}_{t,r} = 1 \quad \text{for} \quad d^* = \operatorname*{arg\,max}_{d=1,\dots,D} \ p^d_{t,r},$$
$$\tilde{\pi}^{d}_{t,r} = 0 \quad \text{otherwise}. \tag{12}$$

We give the maximum weight to local set d where the new observation is more probable, considering the previous parameter estimations. This is a conservative perspective in the sense that it reduces false alarms, because it needs enough evidence to decide that a change-point is occurring.

Given the partial weights $\tilde{\pi}^1_{t,r}, \dots, \tilde{\pi}^D_{t,r}$ for the new data point we may decide to compute the current weight $\pi^d_{t,r}$ to (i) take into account the weights of the previous time-steps (*memory* approach) or just (ii) the current one (*memoryless* approach), resulting in a memoryless scenario.

We call the first approach *Mixture Memory CPD*. For fixed time-step t and partition r we consider the previous local contributions to the global CP detection and define the current weight of set d as

$$\pi^d_{t,r} = \frac{t-1}{t} \cdot \left(\pi^d_{t-1,r} + \frac{\tilde{\pi}^d_{t,r}}{t-1} \right).$$

This is equivalent to updating the average of past contributions with the current partial one and results in smoother weight functions $\pi^d_{1:t,r}$ in time.

We call the second approach *Mixture Memoryless CPD*. We ignore the past contributions information and consider that the contribution of each local set d is fully determined by the current partial weight, obtaining

$$\pi^d_{t,r} = \tilde{\pi}^d_{t,r}.$$

This option has the advantage that we can totally remove the contribution of a particular local partition at a time-step if we have missing data.

5.3. Likelihood probabilistic weighting (Weighted CPD)

We propose to penalize the predictive distributions by raising each local predictive to a weight in the interval (0,1), placing a prior on them, adapting the idea introduced in Wang et al. [22]. As considered in the Mixture approach, we penalize each temporal partition and local set separately at each time-step, and denote $w^d_{t,r}$ to a weight for temporal partition r and local set d. That is, a matrix $\mathbf{w}_t \in (0,1)^{D \times t}$ of weights for fixed t. The graphical representation of this model is depicted in Fig. 3(c). We propose the following structure for the joint distribution, where we have to solve an optimization problem at each time instant,

$$p(\mathbf{s}_t|r_{t-1}, \mathbf{s}_{1:t-1}) = \max_{\mathbf{w}_{t,r}} \quad p(\mathbf{s}_t, \mathbf{w}_{t,r}|r_{t-1}, \mathbf{s}_{1:t-1}) \tag{13}$$

where

$$p(\mathbf{s}_t, \mathbf{w}_{t,r} | r_{t-1}, \mathbf{s}_{1:t-1}) = p(\mathbf{w}_{t,r}) \cdot \prod_{d=1}^{D} p(s_t^d | r_{t-1}, s_{1:t-1}^d)^{w_{t,r}^d} \cdot Z_{t,r}^{-1}. \tag{14}$$

The term $Z_{t,r}$ is the normalizing factor. $p(\mathbf{w}_{t,r})$ is the prior distribution over the weights that has been considered a product of independent Beta distributions,

$$p(\mathbf{w}_{t,r}) = \prod_{d=1}^{D} \mathrm{Beta}(w_{t,r}^d; a, b), \tag{15}$$

with same hyperparameters $a \in (0, \infty), b \in (0, \infty)$ as suggested in Wang et al. [22]. Note that, the constraint that forces the weights to be defined in the interval (0,1), ensures that none of the predictive distributions become more peaked that it was in the local case. With this approach, which we call *Weighted Partition CPD*, we aim to penalize the local sets d that are less probable, giving more weight to those sets where the new observation s_t^d is more probable. Therefore, making the model more conservative to detect a change-point and more resistant to false alarms.

This approach suggests a second one which we call *Weighted Sum CPD*. This provides an alternative computation of the weights, assuming equal local penalization w_t^d for every temporal partition r and having a vector $\mathbf{w}_t \in \{0, 1\}^D$ of weights at each t instead of a matrix. Under this claim, we propose to compute the weights based on the penalization of the term,

$$\sum_{r_{t-1}} p(s_t^d | r_{t-1}, s_{1:t-1}^d), \tag{16}$$

assuming that the contribution of each local set does not depend on the temporal partition r, but on the general predictive capability of the new data point. Based on the first approach, we want to find the weights $\mathbf{w}_t$ that maximize

$$p(\mathbf{w}_t) \cdot \prod_{d=1}^{D} \left(\sum_{r_{t-1}} p(s_t^d | r_{t-1}, s_{1:t-1}^d) \right)^{w_t^d} \cdot Z_t^{-1}, \tag{17}$$

where $p(\mathbf{w}_t)$ is a product of independent beta distributions as in Eq. (15) and Z_t is the normalizing term. The motivation behind this CPD version is that a data point from a local set d that is not very probable at any of the current possible temporal partitions r could be an outlier that would induce noise in the detection at that time-step, so we are interested in reducing its contribution.

The final expression for the weights $w_{t,r}^d$ and w_t^d and the corresponding mathematical analysis is detailed in Appendix A.

5.4. Missing temporal data

Data can be (i) partially missed within a local set d or (ii) totally missed for several or every local set at time-step t. In the first case, missing data is treated through the construction of the latent representation following the approach presented in Moreno-Muñoz et al. [14]. For the missing temporal case, we consider different approaches for each factorization model that are explained in detail in Appendix B.

6. Experiments

In this section we evaluate the performance of the proposed methods through the metrics of *precision* and *delay* in the detection of different intensity change-points on a synthetic dataset and study their robustness against the presence of missing data (Appendix C). We also compare the LOM-based CPD with other existing CPD mechanisms. Finally, we apply our method to a real multi-source dataset of a human behavior study and analyze the contribution of each source to the detection.

Table 3
Individual observation models. Results over 5 trials for each source separately using Multinomial CPD [17]. Precision: total ratio of detected CPs. Delay: average and standard deviation of the delay just for the detected CPs.

Individual OM				
	Real$_1$	Discrete$_1$	Real$_2$	Discrete$_2$
PRECISION	0.72	0.92	0.36	0.88
DELAY	8.8 ± 9.7	25.3 ± 27.7	36.2 ± 57.8	24.8 ± 23.0

6.1. Synthetic data

We present a comparison of precision and delay in the detection for the same dataset but different observation models (OM): (i) independent sources (Sources OM), (ii) grouping by type of data (Grouped OM) and (iii) joint modeling (Joint OM). We also compare the performance for each factorization and consequent version of the CP detector presented in this work: (i) independent product (Independent CPD), mixture weighting by temporal partition ((ii) Mixture Memory and (iii) Memoryless CPD), likelihood probabilistic weighting ((iv)Weighted Partition and (v) Sum CPD). Details of some approximations considered for the experiments are described in Appendix A.

The generated dataset is composed of four 10-dimensional variables: two Gaussian (Real 1 OM, Real 2 OM) and two Bernoulli (Discrete 1 OM, Discrete 2 OM). Every variable has 5 different intensity CPs periodically defined along $T = 600$ time-steps. Three low-intensity (L) CPs are located at $t = 100, 300, 500$ and two high-intensity (H) CPs at $t = 200, 400$. That is, every variable has the same number of CPs and locations. For the continuous variables, the L-CPs correspond to variations ≈ 0.3 around the mean for two consecutive partitions of the data and the H-CPs correspond to random variations between 3.0 and 6.0. For the discrete variables, the L-CPs correspond to variations ≈ 0.2 around the mean and the H-CPs to variations ≈ 0.7. We have run the methods 5 times for 5 datasets, obtaining the results presented in Tables 4 and 5. We have also included the results of the detection for each variable in Table 3 to know the information contained within each of them separately, and see the gain of the fusion of sets. The precision metric is measured as the ratio of detected CPs over every trial. We consider a CP as detected if there is a decrease higher than 20 time-steps from r_{t-1}^* to r_t^*. The delay is measured as the difference between the moment of the CP detection, given by t, and the real location of the CP, given by r_t. Note that the delay is presented as the average ± the standard deviation over every trial and has been computed just over the detected CPs. Therefore, it is usual to see lower delay values for methods with a lower precision rate, but we consider it as a more informative measure for the performance of the method.

Looking at Table 3 we see that we are not able to detect all the CPs by treating each source separately. Only in the discrete case the precision rate is ≈ 0.9, but then the delay is 25.3 ± 27.7. However, considering the Sources OM, we detect every CP for Independent CPD, Weighted Partition CPD and Weighted Sum CPD, i.e., a precision rate of 1.0. Moreover, the delay rates are really low, with a mean of 8.4 time-steps and a standard deviation of 8.8. The Joint OM gets lower precision rates due to the smoothing effect of the joint modeling. Due to this fact, it is able to detect only the high-intensity CPs although keeping a low delay, similar to the Sources OM. This can be shown in the example of Fig. 5. Red lines indicate the MAP estimates of the run-length, and the dashed green and black lines indicate the location of H-CPs and L-CPs, respectively. The same occurs with the Mixture Memory and Mixture Memoryless CPDs. In this case, we always detect the H-CPs and some of the L-CPs, but this approach is more accurate with the abrupt ones.

Table 4
Sources and grouped observation models. Results over 5 trials for every CPD factorization. Precision: total ratio of detected CPs. Delay: average and standard deviation of the delay just for the detected CPs.

Sources OM					
CPD version	Mixture memory	Mixture memoryless	Independent	Weighted partition	Weighted sum
PRECISION DELAY	0.64 20.1 ± 17.18	0.64 8.6 ± 12.6	1.0 8.4 ± 8.8	1.0 12.1 ± 13.5	1.0 8.7 ± 9.25
Grouped OM					
PRECISION DELAY	0.36 11.3 ± 13.9	0.36 10.0 ± 14.13	0.88 24.2 ± 22.4	0.8 27.7 ± 24.9	0.88 24.9 ± 23.43

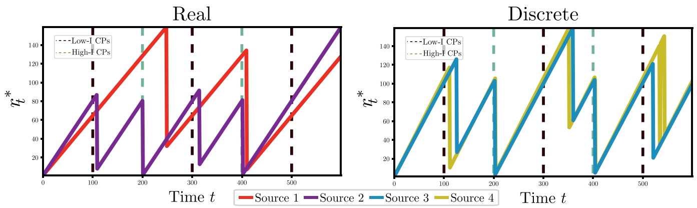

Fig. 4. Detection result of the four sources separately: two Gaussian and two Bernoulli. Red line: MAP estimates of the run-length. Detected CPs are drops of this line. Dashed green and black lines: true high and low intensity CPs, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

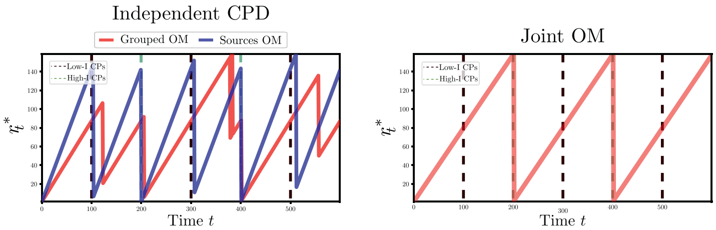

Fig. 5. Detection result for Sources OM and Grouped OM + Independent CPD (Left) and Joint OM (Right). Red line: MAP estimates of the run-length. Detected CPs are drops of this line. Dashed green and black lines: true high and low intensity CPs, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Joint observation nodel. Results over 5 trials using Multinomial CPD [17]. Precision: total ratio of detected CPs. Delay: average and standard deviation of the delay just for the detected CPs.

Joint OM	
PRECISION DELAY	0.48 8.33 ± 12.73

As a conclusion, we can see that the Sources OM achieves the best precision rate and delay results for several CPD adaptations. The method is able to detect every CP, even if we have mixed intensity levels and to keep an average delay of 8.4 time-steps. In particular, the Independent CPD obtains better metrics independently of the observation model chosen. Examples of the output for every observation model and CPD version can be shown in Figs. 4–6.

6.2. Comparison with existing CPD methods

In this subsection, we compare the presented LOM-based CPD with other existing CPD methods in the literature. The variables used are the same as those generated in Section 6.1 and also the metrics definition of precision rate and delay in the detection. We have considered for comparison the (i) Sources OM + Independent CPD, (ii) Grouped OM + Independent CPD, (iii) Multinomial CPD, (iv) Hierarchical CPD, (v) Optimal Partitioning (OP) [11] and (vi) Binary Segmentation (BS) [19]. The methods (i)–(iii) works on samples of the local latent spaces that are considered multinomial distributed. The Multinomial CPD is equivalent to the Joint OM proposed in this work, because we consider only one local set composed of every source. Methods (iv)–(vi) work over univariate representations of the original data. We have performed a dimensionality reduction as proposed for the Joint OM, but we have considered the *maximum-a-posterior (MAP)* estimates as input for the CP detector. Moreover, the OP and BS are offline techniques since they work over the whole sequence of data. We compare them just in terms of precision rate because the delay metric does not make sense in this scenario.

The results are shown in Table 6. We see that the two LOM-based approaches ((i) and (ii)) obtain the best results in terms of precision rate in the detection. For the Sources OM + Independent CPD every CP is detected while we obtain a precision rate of 0.88 for the Grouped OM + Independent CPD. The Multinomial CPD keeps low delay in the detection, but it is computed just over the detected CPs and is still lower than that of the Sources OM. The Hierarchical CPD and the offline methods obtain precision rates of 0.44, which are equivalent to the detection of the high-intensity CPs. This is expected when we work with only one representation of the data (one local set composed of every source) due to the smoothing effect of the dimensionality reduction and the optimization problem of a product of different support likelihoods, as we saw in the performance subsection.

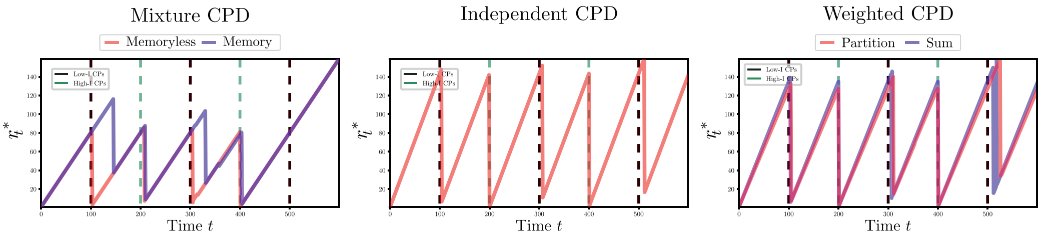

Fig. 6. Detection result for every factorization using the Sources OM. Red line: MAP estimates of the run-length. Detected CPs are drops of this line. Dashed green and black lines: true high and low intensity CPs, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

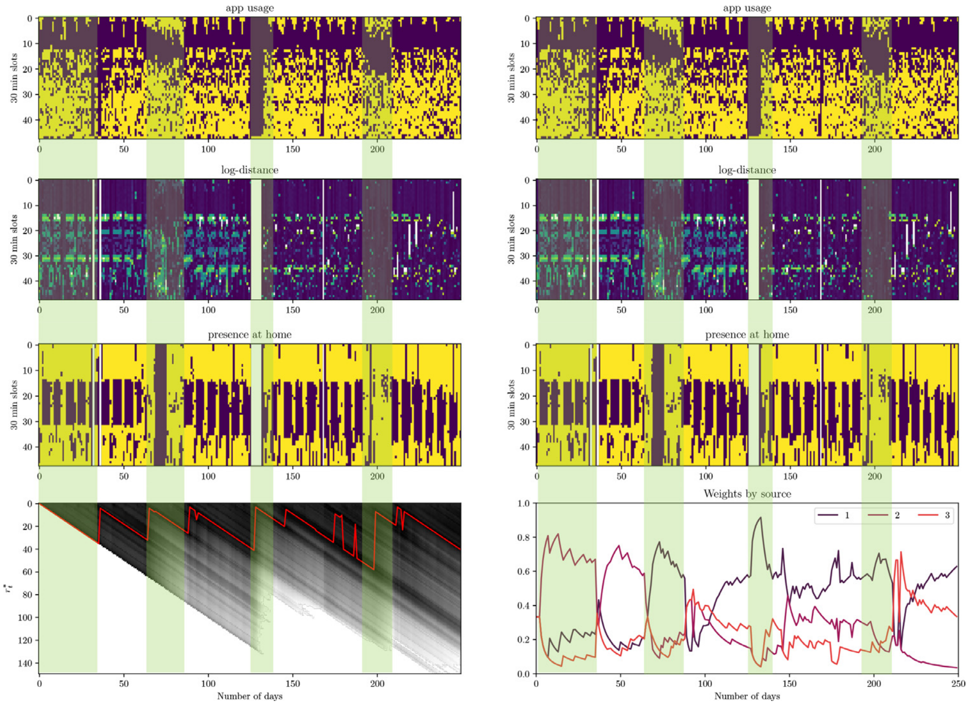

Fig. 7. Result of the *Sources OM with Mixture Memory CPD* performed over a user from the Human Behavior dataset. The green regions separates the different partitions of the data given by the detector output ignoring the delay. **Three upper rows:** of both images: App usage data (1), log-distance data (2) and presence at home data (3) every 30 min, respectively. **Left bottom**: CPD run length probabilities and the MAP (in red) that determines the CPs. **Right bottom**: Weights associated to each source during the detection for each partition determined by the run length MAP. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
LOM-based CPD comparison with existing CPD methods. Results over 5 trials for every CPD method. Precision: total ratio of detected CPs. Delay: average and standard deviation of the delay just for the detected CPs.

Method	Sources OM + Independent CPD	Grouped OM + Independent CPD	Multinomial CPD	Hierarchical CPD	OP	BS
Precision	1.0	0.88	0.48	0.44	0.44	0.44
Delay	8.08 ± 8.56	24.86 ± 23.82	8.0 ± 12.96	2.82 ± 0.94	–	–

6.3. Human behavior data

We examine the presented method through a real-world example. The data are part of a human behavior study with anonymized daily measurements obtained from the passive monitoring of individuals using their personal smartphones. The collection and preprocessing of the data were performed by the Evidence-Based Behavior (eB^2) app between April, 2019 and March, 2020 [2].

Each daily observation $\mathbf{x}_t$ is composed of three variables: *App Usage* (x_t^1), *Log-Distance* (x_t^2) and *Presence at Home* (x_t^3). The daily information is split up into 30 min slots during the day, resulting in 48-dim variables. App Usage is a binary variable that is set to 1 if the phone has been used at a time slot, and 0 otherwise. The Log-Distance is a continuous variable that corresponds to the logarithm of the distance traveled, which is computed as the difference between two consecutive location coordinates. The Presence at Home variable is also binary, and is 1 if the user has been at home at a time slot, and 0 otherwise. The real variable has been modeled as an isotropic Gaussian and the binary variables as multivariate Bernoulli.

The method used for the detection is the Sources OM with the Mixture Memory CPD and the results are shown in Fig. 7. The three upper rows show the three daily variables. In the y-axis we have the 48 time slots and in the x-axis, the number of days with data (250 for this user). In the bottom image of the left figure we can see the output of the detection. The red line indicates the MAP of the run-length, r_t^*, at each time-step, used to define the change-points. The green regions separate the different partitions of the data given by r_t^*, ignoring the delay. The condition considered for the detection is that there is a decrease higher than 20 time-steps from r^*_{t-1} to r^*_t that stays in the same temporal partition for more than 5 time-steps so as to avoid possible outliers. We have detected 7 CPs, thus, 8 partitions with these conditions. Looking at the data, we may sense that the first CP separates two partitions where there is mainly a slight change in the distribution of the binary variables over the course of the day. This is probably what we refer to as a low-intensity change. For instance, we see that there is a reduction in the phone usage between slots 20 and 30. Moreover, there are several days without information from Distance and Home variables where there is not use of the phone during almost the whole day. The second CP is more evident because there is a visible change in the data of the three variables and a clear partition is defined between the second and third CPs. The fifth partition seems to be mainly determined by the absence of phone usage during several days, as there is no information from the other variables and it presents a huge difference related to the App Usage during the previous days. Within this partition and after this marked period, there is a slight reduction in the distance traveled and a slight increase in the usage of the phone. The beginning of the sixth partition is detected with higher delay, probably due to the fact that this is a low-intensity change. We mainly appreciate it in Distance and Home variables. Note that within this partition we could have detected another CP, but it depends on the chosen metric, and this is not the case. The seventh partition is visibly different from partitions 6 and 8, but is not as evident as partition 3. In fact, the change in App Usage variable is not abrupt but gradual, probably the reason for the delay in the detection.

Explainability: contribution of each source to the detection The Mixture Memory CPD gives an intuition about the contribution of each source to the detection and is presented in Fig. 7. In this method, a weight is estimated for each source and for each possible current partition, recursively. In the last row of the right image we can see the weights estimation associated to the MAP run length r^*_t of left image at each time-step. Due to the weights expression, the more probable the new observation, the higher the weight. The method is conservative in this sense, reducing the potential false alarms. In particular, if there is no information on a concrete source, the weight is set to zero, leaving the total contribution to the observed ones. This behavior is reflected in partition 5 and another example occurs in partition 3. We see more variability and blurry data for Distance and Home variables with consequently lower weights, while the App Usage is more constant within the partition, obtaining higher estimates. Note that the gradual change of this variable is also reflected in the gradual decrease of the weight along the partition. We see the same behavior in the last partition, where the Distance variable weight is lower than the ones related to other variables. This is probably due to the variability of the data distribution along several days faced to the more constant data for the two other variables.

7. Conclusion

In this paper we present a new CPD methodology based on adaptive local observation models (LOMs) that works on high-dimensional, multi-source and heterogeneous sequences of data while handling missing observations.

We introduce several LOMs to adapt the detection method to possible situations, partitioning the feature space depending on the context-meaning, multi-source and mixed-type nature of the data. In this way, we control how the local CP information is transferred to homogeneous local latent spaces, which are considered univariate and discrete for every partition avoiding the heterogeneous initial problem. We propose three different factorizations of the CP detector to fuse the local information that holds for any LOM, and include specific mechanisms to deal with missing temporal data. Two of these factorization models assume different weights for every partition, adding explainability about the contribution of each source to the global detection. We compare the performance of every couple LOM-CPD versions between them, and also with respect to the CP detection over each source separately on a synthetic dataset. The result is an adaptive LOM-based CPD method that enhances the fusion of heterogeneous multi-source data with respect to previous works. This method improves the sensitivity in the detection in terms of precision rate and delay when there are different intensity CPs within the sequence, together with higher robustness against missing data presence. We finally illustrate the results on a real-world dataset from a smartphone-based monitoring study for healthcare.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partly supported by Spanish government (AEI/MCI) under grants RTI2018-099655-B-100, PID2021-123182OB-I00, PID2021-125159NB-I00, and TED2021-131823B-I00, by Comunidad de Madrid under grant IND2018/TIC-9649, IND2022/TIC- 23550, by the European Union (FEDER) and the European Research Council (ERC) through the European Union's Horizon 2020 research and innovation program under Grant 714161, and by Comunidad de Madrid and FEDER through IntCARE-CM.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2022.109116.

References

[1] R. P. Adams, D. J. MacKay, Bayesian online changepoint detection, 2007.
[2] S. Berrouiguet, Ramírez, et al., Combining continuous smartphone native sensors data capture and unsupervised data mining techniques for behavioral changes detection: a case series of the evidence-based behavior (eB2) study, JMIR mHealth uHealth 6 (12) (2018) e9472, doi:10.2196/mhealth.9472.
[3] G.S. Birkhead, M. Klompas, N.R. Shah, Uses of electronic health records for public health surveillance to advance public health, Annu. Rev. Public Health 36 (1) (2015) 345–359.
[4] B. Boashash, G. Azemi, et al., Principles of time–frequency feature extraction for change detection in non-stationary signals: applications to newborn EEG abnormality detection, Pattern Recognit. 48 (3) (2015) 616–627.
[5] G.J.J. van den Burg, C.K.I. Williams, An evaluation of change point detection algorithms, 2020. arXiv:2003.06222.
[6] O. Cappé, E. Moulines, On-line expectation-maximization algorithm for latent data models, J. R. Stat. Soc 71 (3) (2009) 593–613.
[7] T. Çelik, Bayesian change detection based on spatial sampling and Gaussian mixture model, Pattern Recognit. Lett. 32 (12) (2011) 1635–1642.
[8] M.F. Chowdhury, S.A. Selouani, D. O'Shaughnessy, Bayesian on-line spectral change point detection: asoft computing approach for on-line ASR, Int. J. Speech Technol. 15 (1) (2012) 5–23.
[9] E. Epaillard, N. Bouguila, Proportional data modeling with hidden Markov models based on generalized Dirichlet and beta-Liouville mixtures applied to anomaly detection in public areas, Pattern Recognit. 55 (2016) 125–136.
[10] M. Evans, N. Hastings, B. Peacock, Statistical Distributions, Wiley Series in Probability and Statistics, Wiley, 2000.
[11] B. Jackson, J.D. Scargle, et al., An algorithm for optimal partitioning of data on an interval, IEEE Signal Process. Lett. 12 (2) (2005) 105–108.
[12] M. Lavielle, G. Teyssière, Adaptive Detection of Multiple Change-Points in Asset Price Volatility, Springer Berlin Heidelberg, 2007, pp. 129–156.
[13] G. Liu, J. Onnela, Bidirectional imputation of spatial GPS trajectories with missingness using sparse online gaussian process, J. Am. Med. Inform. Assoc. 28 (8) (2021) 1777–1784.
[14] P. Moreno-Muñoz, D. Ramírez, et al., Change-point detection in hierarchical circadian models, Pattern Recognit. 113 (2021) 107820.
[15] P. Moreno-Muñoz, L. Romero-Medrano, et al., Passive detection of behavioral shifts for suicide attempt prevention, ML4MH, NeurIPS (2020).
[16] J.A. Quinn, M. Sugiyama, A least-squares approach to anomaly detection in static and sequential data, Pattern Recognit. Lett. 40 (2014) 36–40.

[17] L. Romero-Medrano, P. Moreno-Muñoz, A. Artés-Rodríguez, Multinomial Sampling of Latent Variables for Hierarchical Change-Point Detection, in: Journal of Signal Processing Systems, 94, 2nd, Springer, 2022, pp. 215–227.
[18] H. Sagha, H. Bayati, J.del R. Millán, R. Chavarriaga, On-line anomaly detection and resilience in classifier ensembles, Pattern Recognit. Lett. 34 (15) (2013) 1916–1927.
[19] A. Scott, M. Knott, A cluster analysis method for grouping means in the analysis of variance, Biometrics 30 (1974) 507.
[20] Y. Sun, L. Lei, et al., Nonlocal patch similarity based heterogeneous remote sensing change detection, Pattern Recognit. 109 (2021) 107598.
[21] C. Truong, L. Oudre, N. Vayatis, Selective review of offline change point detection methods, Signal Process. 167 (2020) 107299.
[22] Y. Wang, A. Kucukelbir, D.M. Blei, Robust probabilistic modeling with Bayesian data reweighting, in: ICML, in: Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 3646–3655.

Lorena Romero-Medrano obtained her Degree in Mathematics from Universidad de Zaragoza in 2015 and a Master in Mathematical Modelling (Mathematics Applied to Biological and Medical Sciences Major) from Université Pierre et Marie Curie, Paris VI, in 2017 supported by a PGSM-Inria fellowship. During that year, she held a 6-month traineeship at Inria Research Institute in Paris (MAMBA team) to work on the modelling of liver hemodynamics and continued with this project as research visitor at Universidad Complutense de Madrid. She is currently Ph.D. student at the Dept. of Signal Theory and Communications at Universidad Carlos III de Madrid and Evidence-Based Behavior (eB2). Her research interests include mathematical modelling, machine learning, probabilistic methods, and its applications to human behavior and biomedical sciences.

Antonio Artés-Rodríguez (S'89-M'93-SM'01) was born in Alhama de Alméra, Spain, in 1963. He received the Ingeniero de Telecomunicación and Doctor Ingeniero de Telecomunicación degrees, both from the Universidad Politecnica de Madrid, Madrid, Spain, in 1988 and 1992, respectively. He is a Professor at the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Madrid. Prior to this, he held different teaching positions at Universidad de Vigo, Universidad Politecnica de Madrid, and Universidad de Alcalá, all of them in Spain. He has participated in more than 70 projects and contracts and has coauthored more that 50 journal papers and more than 100 international conference papers. His research interests include signal processing, machine learning, and information theory methods, and its application to health and sensor networks.